Geometric Number Theory

Lenny Fukshansky

Minkowski'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. 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 $\boldsymbol{x}, \boldsymbol{y} \in K$ the line segment connecting them is entirely contained in K, i.e. for any $0 \leq t \leq 1$, $t\boldsymbol{x} + (1-t)\boldsymbol{y} \in K$. Further, K is called **0**-*symmetric* if for any $\boldsymbol{x} \in K$, $-\boldsymbol{x} \in K$.

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

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

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. Norms, sets, and volumes

Throughout this section we will work in the real vector space $\mathbb{R}^n$, where $n \geq 1$.

DEFINITION 1.2.1. A function $F : \mathbb{R}^n \to \mathbb{R}$ is called a *norm* if

(1) $F(\boldsymbol{x}) \geq 0$ with equality if and only if $\boldsymbol{x} = \mathbf{0}$,
(2) $F(a\boldsymbol{x}) = |a|F(\boldsymbol{x})$ for each $a \in \mathbb{R}$, $\boldsymbol{x} \in \mathbb{R}^n$,
(3) *Triangle inequality:* $F(\boldsymbol{x} + \boldsymbol{y}) \leq F(\boldsymbol{x}) + F(\boldsymbol{y})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.

For each positive integer p, we can introduce the *L_p-norm* $\| \ \|_p$ on $\mathbb{R}^n$ defined by

$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p},$$

for each $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$. We also define the *sup-norm*, given by

$$|\boldsymbol{x}| = \max_{1 \leq i \leq n} |x_i|.$$

These indeed are norms on $\mathbb{R}^n$ (Problem 1.1).

Unless stated otherwise, we will regard $\mathbb{R}^n$ as a normed linear space (i.e. a vector space equipped with a norm) with respect to the Euclidean norm $\| \ \|_2$: from now on we will refer to it simply as $\| \ \|$. Recall that for every two points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, Euclidean distance between them is given by

$$d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|.$$

We start with definitions and examples of a few different types of subsets of $\mathbb{R}^n$ that we will often encounter.

DEFINITION 1.2.2. A subset $X \subseteq \mathbb{R}^n$ is called *compact* if it is closed and bounded.

Recall that a set is closed if it contains all of its limit points, and it is bounded if there exists $M \in \mathbb{R}_{>0}$ such that for every two points $\boldsymbol{x}, \boldsymbol{y}$ in this set $d(\boldsymbol{x}, \boldsymbol{y}) \leq M$. For instance, the closed unit ball centered at the origin in $\mathbb{R}^n$

$$\mathbb{B}_n = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| \leq 1\}$$

is a compact set, but its interior, the open ball

$$\mathbb{B}_n^o = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| < 1\}$$

is not a compact set. If we now write

$$\mathbb{S}_{n-1} = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| = 1\}$$

for the unit sphere centered at the origin in $\mathbb{R}^n$, then it is easy to see that $\mathbb{B}_n = \mathbb{S}_{n-1} \cup \mathbb{B}_n^o$, and we refer to $\mathbb{S}_{n-1}$ as the boundary of $\mathbb{B}_n$ (sometimes we will write $\mathbb{S}_{n-1} = \partial\mathbb{B}_n$) and to B_n^o as the interior of $\mathbb{B}_n$.

From here on we will also assume that all our compact sets have no isolated points. Then we can say more generally that every compact set $X \subset \mathbb{R}^n$ has boundary ∂X and interior X^o, and can be represented as $X = \partial X \cup X^o$. To make this notation precise, we say that a point $\boldsymbol{x} \in X$ is a *boundary* point of X if every open neighborhood U of $\boldsymbol{x}$ contains points in X and points not in X; we write ∂X for the set of all boundary points of X. All points $\boldsymbol{x} \in X$ that are not in ∂X are called *interior* points of X, and we write X^o for the set of all interior points of X.

DEFINITION 1.2.3. A compact subset $X \subseteq \mathbb{R}^n$ is called *convex* if whenever $\boldsymbol{x}, \boldsymbol{y} \in X$, then any point of the form

$$t\boldsymbol{x} + (1-t)\boldsymbol{y},$$

where $t \in [0,1]$, is also in X; i.e. whenever $\boldsymbol{x}, \boldsymbol{y} \in X$, then the entire line segment from $\boldsymbol{x}$ to $\boldsymbol{y}$ lies in X.

We now briefly mention a special class of convex sets. Given a set X in $\mathbb{R}^n$, we define the *convex hull* of X to be the set

$$\mathrm{Co}(X) = \left\{ \sum_{\boldsymbol{x} \in X} t_{\boldsymbol{x}} \boldsymbol{x} : t_{\boldsymbol{x}} \geq 0 \ \forall \ \boldsymbol{x} \in X, \ \sum_{\boldsymbol{x} \in X} t_{\boldsymbol{x}} = 1 \right\}.$$

It is easy to notice that whenever a convex set contains X, it must also contain $\mathrm{Co}(X)$. Hence convex hull of a collection of points should be thought of as the *smallest* convex set containing all of them. If the set X is finite, then its convex hull is called a *convex polytope*. Most of the times we will be interested in convex polytopes, but occasionally we will also need convex hulls of infinite sets.

There is an alternative way of describing convex polytopes. Recall that a hyperplane in $\mathbb{R}^n$ is a translate of a co-dimension one subspace, i.e. a subset $\mathbb{H}$ in $\mathbb{R}^n$ is called a *hyperplane* if

$$\mathbb{H} = \left\{ \boldsymbol{x} \in \mathbb{R}^n : \sum_{i=1}^{n} a_i x_i = b \right\}, \tag{1.1}$$

for some $a_1, \ldots, a_n, b \in \mathbb{R}$. Notice that each hyperplane divides $\mathbb{R}^n$ into two halfspaces. More precisely, a closed *halfspace* $\mathcal{H}$ in $\mathbb{R}^n$ is a set of all $\boldsymbol{x} \in \mathbb{R}^n$ such that either $\sum_{i=1}^n a_i x_i \geq b$ or $\sum_{i=1}^n a_i x_i \leq b$ for some $a_1, \ldots, a_n, b \in \mathbb{R}$. Minkowski-Weyl theorem (Problem 1.4) asserts that a set is a convex polytope in $\mathbb{R}^n$ if and only if it is a bounded intersection of finitely many halfspaces. Polytopes form a very nice class of convex sets in $\mathbb{R}^n$, and we will talk more about them later.

There is, of course, a large variety of sets that are not necessarily convex. Among these, ray sets and star bodies form a particularly nice class. In fact, they are among the not-so-many non-convex sets for which many of the methods we develop in this chapter still work, as we will see later.

DEFINITION 1.2.4. A set $X \subseteq \mathbb{R}^n$ is called a *ray set* if for every $\boldsymbol{x} \in X$, $t\boldsymbol{x} \in X$ for all $t \in [0,1]$.

Clearly every ray set must contain $\mathbf{0}$. Moreover, ray sets can be bounded or unbounded. Perhaps the simplest examples of bounded ray sets are convex sets that contain $\mathbf{0}$. Star bodies form a special class of ray sets.

DEFINITION 1.2.5. A set $X \subseteq \mathbb{R}^n$ is called a *star body* if for every $\boldsymbol{x} \in \mathbb{R}^n$ either $t\boldsymbol{x} \in X$ for all $t \in \mathbb{R}$, or there exists $t_0(\boldsymbol{x}) \in \mathbb{R}_{>0}$ such that $t\boldsymbol{x} \in X$ for all $t \in \mathbb{R}$ with $|t| \leq t_0(\boldsymbol{x})$, and $t\boldsymbol{x} \notin X$ for all $|t| > t_0(\boldsymbol{x})$.

REMARK 1.2.1. We will also require all our star bodies to have boundary which is *locally homeomorphic to* $\mathbb{R}^{n-1}$. Loosely speaking, this means that the boundary of a star body can be subdivided into small patches, each of which looks like a ball in $\mathbb{R}^{n-1}$. More precisely, suppose X is a closed star body and ∂X is its boundary. We say that ∂X is *locally homeomorphic* to $\mathbb{R}^{n-1}$ if for every point $\boldsymbol{x} \in \partial X$ there exists an open neighborhood $U \subseteq \partial X$ of $\boldsymbol{x}$ such that U is homeomorphic to $\mathbb{R}^{n-1}$.

See Remark 1.2.2 below for the definition of what it means for two sets to be homeomorphic. Unless explicitly stated otherwise, all star bodies will be assumed to have this property.

Here is an example of a collection of unbounded star bodies:

$$\mathrm{St}_n = \left\{(x, y) \in \mathbb{R}^2 : -\frac{1}{x^n} \leq y \leq \frac{1}{x^n}\right\},$$

where $n \geq 1$ is an integer. There is also an alternative description of star bodies. For this we need to introduce an additional piece of notation.

DEFINITION 1.2.6. A function $F : \mathbb{R}^n \to \mathbb{R}$ is called a *distance function* if

(1) $F(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$,
(2) F is continuous,
(3) *Homogeneity:* $F(a\boldsymbol{x}) = |a|F(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^n$, $a \in \mathbb{R}$.

Let $f(X_1, \ldots, X_n)$ be a polynomial in n variables with real coefficients. We say that f is *homogeneous* if every monomial in f has the same degree. For instance, x^2+xy-y^2 is a homogeneous polynomial of degree 2, while $x^2 - y + xy$ is an inhomogeneous polynomial of degree 2. If $f(X_1, \ldots, X_n)$ be a homogeneous polynomial of degree d with real coefficients, then

$$F(\boldsymbol{x}) = |f(\boldsymbol{x})|^{1/d}$$

is a distance function (Problem 1.5). As expected, distance functions are closely related to star bodies: for a distance function F the set

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

is a bounded star body (Problem 1.6). In fact, a converse is also true.

THEOREM 1.2.1. *Let X be a star body in $\mathbb{R}^n$. Then there exists a distance function F such that*

$$X = \{\boldsymbol{x} \in \mathbb{R}^n : F(\boldsymbol{x}) \leq 1\}.$$

PROOF. Define F in the following way. For every $\boldsymbol{x} \in \mathbb{R}^n$ such that $t\boldsymbol{x} \in X$ for all $t \geq 0$, let $F(\boldsymbol{x}) = 0$. Suppose that $\boldsymbol{x} \in \mathbb{R}^n$ is such that there exists $t_0(\boldsymbol{x}) > 0$ with the property that $t\boldsymbol{x} \in X$ for all $t \leq t_0(\boldsymbol{x})$, and $t\boldsymbol{x} \notin X$ for all $t > t_0(\boldsymbol{x})$; for such $\boldsymbol{x}$ define $F(\boldsymbol{x}) = \frac{1}{t_0(\boldsymbol{x})}$. It is now easy to verify that F is a distance function; this is left as an exercise, or see Theorem I on p. 105 of [**Cas59**]. □

Notice that all our notation above for convex sets, polytopes, and bounded ray sets and star bodies will usually pertain to closed sets; sometimes we will use the terms like "open polytope" or "open star body" to refer to the interiors of the closed sets.

DEFINITION 1.2.7. A subset $X \subseteq \mathbb{R}^n$ which contains $\mathbf{0}$ is called $\mathbf{0}$-*symmetric* if whenever $\boldsymbol{x}$ is in X, then so is $-\boldsymbol{x}$.

It is easy to see that every set $A_n(C)$ of Problem 1.2, as well as every star body, is $\mathbf{0}$-symmetric, although ray sets in general are not. In fact, star bodies are precisely the $\mathbf{0}$-symmetric ray sets. Here is an example of a collection of asymmetric unbounded ray sets:

$$R_n = \left\{(x, y) \in \mathbb{R}^2 : 0 \leq y \leq \frac{1}{x^n}\right\},$$

where $n \geq 1$ is an integer. An example of a bounded asymmetric ray set is a *cone* on L points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L \in \mathbb{R}^n$, i.e. $\mathrm{Co}(\mathbf{0}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L)$. If X is a star body and F its distance function, i.e. $X = \{\boldsymbol{x} \in \mathbb{R}^n : F(\boldsymbol{x}) \leq 1\}$, then X is convex if and only if

$$F(\boldsymbol{x} + \boldsymbol{y}) \leq F(\boldsymbol{x}) + F(\boldsymbol{y}),$$

for all $\boldsymbol{x}, \boldsymbol{y} \in X$ (Problem 1.7). Next we want to introduce the notion of volume for *bounded* sets in $\mathbb{R}^n$.

Definition 1.2.8. *Characteristic function* of a set X is defined by

$$\chi_X(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{x} \in X \\ 0 & \text{if } \boldsymbol{x} \notin X \end{cases}$$

Definition 1.2.9. A bounded set X is said to have *Jordan volume* if its characteristic function is Riemann integrable, and then we define $\mathrm{Vol}(X)$ to be the value of this integral. A set that has Jordan volume is called *Jordan measurable*.

Definition 1.2.10. Let X and Y be two sets. A function $f : X \to Y$ is called *injective* (or one-to-one) if whenever $f(x_1) = f(x_2)$ for some $x_1, x_2 \in X$, then $x_1 = x_2$; f is called *surjective* (or onto) if for every $y \in Y$ there exists $x \in X$ such that $f(x) = y$; f is called a *bijection* if it is injective and surjective.

Remark 1.2.2. In fact, it is also not difficult to prove that $f : X \to Y$ has an inverse if and only if it is a bijection, in which case this inverse is unique. If such a function f between two sets X and Y exists, we say that X and Y are in *bijective correspondence*. Furthermore, if f and f^{-1} are both continuous, then they are called *homeomorphisms* and we say that X and Y are *homeomorphic* to each other. If f and f^{-1} are also differentiable, then they are called *diffeomorphisms*, and X and Y are said to be *diffeomorphic*.

Theorem 1.2.2. *All convex sets and bounded ray sets have Jordan volume.*

Sketch of proof. We will prove this theorem for convex sets; for bounded ray sets the proof is similar. Let X be a convex set. Write ∂X for the boundary of X and notice that $X = \partial X$ if and only if X is a straight line segment: otherwise it would not be convex. Since it is clear that a straight line segment has Jordan volume (it is just its length), we can assume that $X \neq \partial X$, then X has nonempty interior, denote it by X^o, so $X = X^o \cup \partial X$. We can assume that $\mathbf{0} \in X^o$; if not, we can just translate X so that it contains $\mathbf{0}$ - translation does not change measurability properties. Write $\mathbb{S}_{n-1}$ for the unit sphere centered at the origin in $\mathbb{R}^n$, i.e. $S_{n-1} = \partial \mathbb{B}_n$. Define a map $\varphi : \partial X \to \mathbb{S}_{n-1}$, given by

$$\varphi(\boldsymbol{x}) = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}.$$

Since X is a bounded convex set, it is not difficult to see that φ is a homeomorphism. For each $\varepsilon > 0$ there exists a finite collection of points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{k(\varepsilon)} \in \mathbb{S}_{n-1}$ such that if we let $\mathcal{C}_{\boldsymbol{x}_i}(\varepsilon)$ be an $(n-1)$-dimensional cap centered at $\boldsymbol{x}_i$ in $\mathbb{S}_{n-1}$ of radius ε, i.e.

$$\mathcal{C}_{\boldsymbol{x}_i}(\varepsilon) = \{\boldsymbol{y} \in \mathbb{S}_{n-1} : \|\boldsymbol{y} - \boldsymbol{x}_i\|_2 \leq \varepsilon\},$$

then $\mathbb{S}_{n-1} = \bigcup_{i=1}^{k(\varepsilon)} \mathcal{C}_{\boldsymbol{x}_i}(\varepsilon)$, and so $\partial X = \bigcup_{i=1}^{k(\varepsilon)} \varphi^{-1}\left(\mathcal{C}_{\boldsymbol{x}_i}(\varepsilon)\right)$. For each $1 \leq i \leq k(\varepsilon)$, let $\boldsymbol{y}_i, \boldsymbol{z}_i \in \varphi^{-1}\left(\mathcal{C}_{\boldsymbol{x}_i}(\varepsilon)\right)$ be such that

$$\|\boldsymbol{y}_i\| = \max\{\|\boldsymbol{x}\| : \boldsymbol{x} \in \varphi^{-1}\left(\mathcal{C}_{\boldsymbol{x}_i}(\varepsilon)\right)\},$$

and

$$\|\boldsymbol{z}_i\| = \min\{\|\boldsymbol{x}\| : \boldsymbol{x} \in \varphi^{-1}\left(\mathcal{C}_{\boldsymbol{x}_i}(\varepsilon)\right)\}.$$

Let $\delta_1(\varepsilon)$ and $\delta_2(\varepsilon)$ be minimal positive real numbers such that the spheres centered at the origin of radii $\|\boldsymbol{y}_i\|$ and $\|\boldsymbol{z}_i\|$ are covered by caps of radii $\delta_1(\varepsilon)$ and $\delta_2(\varepsilon)$, $\mathcal{C}_{\boldsymbol{x}_i}(\boldsymbol{y}_i, \varepsilon)$ and $\mathcal{C}_{\boldsymbol{x}_i}(\boldsymbol{z}_i, \varepsilon)$, centered at $\boldsymbol{x}_i$. Define cones

$$C_i^1 = \mathrm{Co}(\mathbf{0}, \mathcal{C}_{\boldsymbol{x}_i}(\boldsymbol{y}_i, \varepsilon)),\ C_i^2 = \mathrm{Co}(\mathbf{0}, \mathcal{C}_{\boldsymbol{x}_i}(\boldsymbol{z}_i, \varepsilon)), \tag{1.2}$$

for each $1 \leq i \leq k(\varepsilon)$. Now notice that

$$\bigcup_{i=1}^{k(\varepsilon)} C_i^2 \subseteq X \subseteq \bigcup_{i=1}^{k(\varepsilon)} C_i^1.$$

Since the cones C_i^1, C_i^2 have Jordan volume (Problem 1.10), the same is true about their finite unions. Moreover,

$$\mathrm{Vol}\left(\bigcup_{i=1}^{k(\varepsilon)} C_i^1\right) - \mathrm{Vol}\left(\bigcup_{i=1}^{k(\varepsilon)} C_i^2\right) \to 0,$$

as $\varepsilon \to 0$. Hence X has Jordan volume, which is equal to the common value of

$$\lim_{\varepsilon \to 0} \mathrm{Vol}\left(\bigcup_{i=1}^{k(\varepsilon)} C_i^1\right) = \lim_{\varepsilon \to 0} \mathrm{Vol}\left(\bigcup_{i=1}^{k(\varepsilon)} C_i^2\right).$$

□

This is Theorem 5 on p. 9 of [**GL87**], and the proof is also very similar.

1.3. Lattices

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

DEFINITION 1.3.1. A *lattice* Λ of *rank* r, $1 \leq r \leq n$, spanned by $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r$ in $\mathbb{R}^n$ is the set of all possible linear combinations of the vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r$ with integer coefficients. In other words,

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

The set $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r$ is called a *basis* for Λ. 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 often 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 $\mathbb{R}^n$:

$$\mathbb{Z}^n = \{\boldsymbol{x} = (x_1, \ldots, x_n) : x_i \in \mathbb{Z} \text{ for all } 1 \leq i \leq n\}.$$

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

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

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

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

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

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

where

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

and A is an $n \times r$ *basis matrix* for Λ of the form $A = (\boldsymbol{a}_1\ \ldots\ \boldsymbol{a}_r)$, which has rank r. In other words, a lattice Λ of rank r in $\mathbb{R}^n$ can always be described as $\Lambda = A\mathbb{Z}^r$, 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 $\boldsymbol{x} \in \mathbb{R}^n$ there exists an open set containing $\boldsymbol{x}$ whose closure is compact, for instance take an open unit ball centered at $\boldsymbol{x}$. More generally, every subspace V of $\mathbb{R}^n$ is also locally compact. A subset Γ of V is called *discrete* if for each $\boldsymbol{x} \in \Gamma$ there exists an open set $S \subseteq V$ such that $S \cap \Gamma = \{\boldsymbol{x}\}$. For instance $\mathbb{Z}^n$ is a discrete subset of $\mathbb{R}^n$: for each point $\boldsymbol{x} \in \mathbb{Z}^n$ the open ball of radius $1/2$ centered at $\boldsymbol{x}$ contains no other points of $\mathbb{Z}^n$. We say that a discrete subset Γ is *co-compact*

in V if there exists a compact $\mathbf{0}$-symmetric subset U of V such that the union of translations of U by the points of Γ covers the entire space V, i.e. if

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

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

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

(1) *Identity:* $\mathbf{0} \in G$,
(2) *Closure:* For every $\boldsymbol{x}, \boldsymbol{y} \in G$, $\boldsymbol{x} + \boldsymbol{y} \in G$,
(3) *Inverses:* For every $\boldsymbol{x} \in G$, $-\boldsymbol{x} \in G$.

By Problems 1.13 and 1.14 a lattice Λ of rank r in $\mathbb{R}^n$ is a discrete co-compact subgroup of $V = \operatorname{span}_{\mathbb{R}} \Lambda$. In fact, the converse is also true.

THEOREM 1.3.1. *Let V be an r-dimensional subspace of $\mathbb{R}^n$, and let Γ be a discrete co-compact subgroup of V. Then Γ is a lattice of rank r in $\mathbb{R}^n$.*

PROOF. In other words, we want to prove that Γ has a basis, i.e. that there exists a collection of linearly independent vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r$ in Γ such that $\Gamma = \operatorname{span}_{\mathbb{Z}}\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r\}$. We start by inductively constructing a collection of vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r$, and then show that it has the required properties.

Let $\boldsymbol{a}_1 \neq \mathbf{0}$ be a point in Γ such that the line segment connecting $\mathbf{0}$ and $\boldsymbol{a}_1$ contains no other points of Γ. Now assume $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{i-1}$, $2 \leq i \leq r$, have been selected; we want to select $\boldsymbol{a}_i$. Let

$$H_{i-1} = \operatorname{span}_{\mathbb{R}}\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{i-1}\},$$

and pick any $\boldsymbol{c} \in \Gamma \setminus H_{i-1}$: such $\boldsymbol{c}$ exists, since $\Gamma \not\subseteq H_{i-1}$ (otherwise Γ would not be co-compact in V). Let P_i be the closed parallelotope spanned by the vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{i-1}, \boldsymbol{c}$. Notice that since Γ is discrete in V, $\Gamma \cap P_i$ is a finite set. Moreover, since $\boldsymbol{c} \in P_i$, $\Gamma \cap P_i \not\subseteq H_{i-1}$. Then select $\boldsymbol{a}_i$ such that

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

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

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

Let $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r$ be the collection of points chosen in this manner. Then we have

$$\boldsymbol{a}_1 \neq \mathbf{0},\ \boldsymbol{a}_i \notin \operatorname{span}_{\mathbb{Z}}\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{i-1}\}\ \forall\ 2 \leq i \leq r,$$

which means that $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r$ are linearly independent. Clearly,

$$\operatorname{span}_{\mathbb{Z}}\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r\} \subseteq \Gamma.$$

We will now show that

$$\Gamma \subseteq \operatorname{span}_{\mathbb{Z}}\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r\}.$$

First of all notice that $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r$ is certainly a basis for V, and so if $\boldsymbol{x} \in \Gamma \subseteq V$, then there exist $c_1, \ldots, c_r \in \mathbb{R}$ such that

$$\boldsymbol{x} = \sum_{i=1}^{r} c_i \boldsymbol{a}_i.$$

Notice that

$$\boldsymbol{x}' = \sum_{i=1}^{r} [c_i]\boldsymbol{a}_i \in \mathrm{span}_{\mathbb{Z}}\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r\} \subseteq \Gamma,$$

where [ ] stands for the *integer part function* (i.e. $[c_i]$ is the largest integer which is no larger than c_i). Since Γ is a group, we must have

$$\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{x}' = \sum_{i=1}^{r} (c_i - [c_i])\boldsymbol{a}_i \in \Gamma.$$

Then notice that

$$d(\boldsymbol{z}, H_{r-1}) = (c_r - [c_r])\ d(\boldsymbol{a}_r, H_{r-1}) < d(\boldsymbol{a}_r, H_{r-1}),$$

but by construction we must have either $\boldsymbol{z} \in H_{r-1}$, or

$$d(\boldsymbol{a}_r, H_{r-1}) \leq d(\boldsymbol{z}, H_{r-1}),$$

since $\boldsymbol{z}$ lies in the parallelotope spanned by $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r$, 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 $\boldsymbol{x} \in \mathrm{span}_{\mathbb{Z}}\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r\}$. Since this is true for every $\boldsymbol{x} \in \Gamma$, we are done. □

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 $\Lambda \subset \mathbb{R}^n$ will be of the form $\Lambda = A\mathbb{Z}^n$, where A is a non-singular $n \times n$ basis matrix for Λ.

THEOREM 1.3.2. *Let Λ be a lattice of rank n in $\mathbb{R}^n$, and let A be a basis matrix for Λ. Then B is another basis matrix for Λ if and only if there exists an $n \times n$ integral matrix U with determinant ± 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 $\boldsymbol{b}_i$ of B can be expressed as

$$\boldsymbol{b}_i = \sum_{j=1}^{n} u_{ij}\boldsymbol{a}_j,$$

where $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ 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 $1 \leq i \leq n$

$$\boldsymbol{a}_i = \sum_{j=1}^{n} w_{ij}\boldsymbol{b}_j,$$

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

$$\Lambda \subseteq \mathrm{span}_{\mathbb{Z}}\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n\}.$$

On the other hand, $\boldsymbol{b}_i \in \Lambda$ for each $1 \leq i \leq n$. Thus B is a basis matrix for Λ. $\square$

COROLLARY 1.3.3. *If A and B are two basis matrices for the same lattice Λ, then*

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

DEFINITION 1.3.2. The common determinant value of Corollary 1.3.3 is called the *determinant* of the lattice Λ, and is denoted by $\det(\Lambda)$.

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

DEFINITION 1.3.3. If Λ and Ω are both lattices in $\mathbb{R}^n$, and $\Omega \subseteq \Lambda$, then we say that Ω is a *sublattice* of Λ.

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

(1) A subset Ω of the lattice Λ is a sublattice if and only if it is a subgroup of the abelian group Λ.
(2) For a sublattice Ω of Λ two cosets $\boldsymbol{x} + \Omega$ and $\boldsymbol{y} + \Omega$ are equal if and only if $\boldsymbol{x} - \boldsymbol{y} \in \Omega$. In particular, $\boldsymbol{x} + \Omega = \Omega$ if and only if $\boldsymbol{x} \in \Omega$.
(3) If Λ is a lattice and μ a real number, then the set
$$\mu\Lambda := \{\mu\boldsymbol{x} : \boldsymbol{x} \in \Lambda\}$$
is also a lattice. Further, if μ is an integer then $\mu\Lambda$ is a sublattice of Λ.

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 Λ.

LEMMA 1.3.4. *Let Ω be a subattice of Λ. There exists a positive integer D such that $D\Lambda \subseteq \Omega$.*

PROOF. Recall that Λ and Ω are both lattices of rank n in $\mathbb{R}^n$. Let $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ be a basis for Ω and $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ be a basis for Λ. Then

$$\mathrm{span}_{\mathbb{R}}\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n\} = \mathrm{span}_{\mathbb{R}}\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n\} = \mathbb{R}^n.$$

Since $\Omega \subseteq \Lambda$, there exist integers $u_{11}, \ldots, u_{nn}$ such that

$$\begin{cases} \boldsymbol{a}_1 = u_{11}\boldsymbol{b}_1 + \cdots + u_{1n}\boldsymbol{b}_n \\ \vdots \qquad \vdots \qquad\qquad \vdots \\ \boldsymbol{a}_n = u_{n1}\boldsymbol{b}_1 + \cdots + u_{nn}\boldsymbol{b}_n. \end{cases}$$

Solving this linear system for $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ in terms of $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$, we easily see that there must exist rational numbers $\frac{p_{11}}{q_{11}}, \ldots, \frac{p_{nn}}{q_{nn}}$ such that

$$\begin{cases} \boldsymbol{b}_1 = \frac{p_{11}}{q_{11}}\boldsymbol{a}_1 + \cdots + \frac{p_{1n}}{q_{1n}}\boldsymbol{a}_n \\ \vdots \qquad \vdots \qquad\qquad \vdots \\ \boldsymbol{b}_n = \frac{p_{n1}}{q_{n1}}\boldsymbol{a}_1 + \cdots + \frac{p_{nn}}{q_{nn}}\boldsymbol{a}_n. \end{cases}$$

Let $D = q_{11} \times \cdots \times q_{nn}$, then $D/q_{ij} \in \mathbb{Z}$ for each $1 \leq i, j, \leq n$, and so all the vectors

$$\begin{cases} D\boldsymbol{b}_1 = \frac{Dp_{11}}{q_{11}}\boldsymbol{a}_1 + \cdots + \frac{Dp_{1n}}{q_{1n}}\boldsymbol{a}_n \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ D\boldsymbol{b}_n = \frac{Dp_{n1}}{q_{n1}}\boldsymbol{a}_1 + \cdots + \frac{Dp_{nn}}{q_{nn}}\boldsymbol{a}_n \end{cases}$$

are in Ω. Therefore $\operatorname{span}_{\mathbb{Z}}\{D\boldsymbol{b}_1, \ldots, D\boldsymbol{b}_n\} \subseteq \Omega$. On the other hand,

$$\operatorname{span}_{\mathbb{Z}}\{D\boldsymbol{b}_1, \ldots, D\boldsymbol{b}_n\} = D \operatorname{span}_{\mathbb{Z}}\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n\} = D\Lambda,$$

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 1.3.5. *Let Λ be a lattice, and Ω a sublattice of Λ. For each basis $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ of Λ, there exists a basis $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ of Ω of the form*

$$\begin{cases} \boldsymbol{a}_1 = v_{11}\boldsymbol{b}_1 \\ \boldsymbol{a}_2 = v_{21}\boldsymbol{b}_1 + v_{22}\boldsymbol{b}_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \boldsymbol{a}_n = v_{n1}\boldsymbol{b}_1 + \cdots + v_{nn}\boldsymbol{b}_n, \end{cases}$$

where all $v_{ij} \in \mathbb{Z}$ and $v_{ii} \neq 0$ for all $1 \leq i \leq n$. Conversely, for every basis $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ of Ω there exists a basis $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ of Λ such that the relations as above hold.

PROOF. Let $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ be a basis for Λ. We will first prove the existence of a basis $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ for Ω as claimed by the theorem. By Lemma 1.3.4, there exist integer multiples of $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ in Ω, hence it is possible to choose a collection of vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n \in \Omega$ of the form

$$\boldsymbol{a}_i = \sum_{j=1}^{i} v_{ij}\boldsymbol{b}_j,$$

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 $\boldsymbol{a}_i$ so that $|v_{ii}|$ is as small as possible, but not 0. We will now show that $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ is a basis for Ω. Clearly,

$$\operatorname{span}_{\mathbb{Z}}\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n\} \subseteq \Omega.$$

We want to prove the inclusion in the other direction, i.e. that

$$\Omega \subseteq \operatorname{span}_{\mathbb{Z}}\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n\}. \tag{1.3}$$

Suppose (1.3) is not true, then there exists $\boldsymbol{c} \in \Omega$ which is not in $\operatorname{span}_{\mathbb{Z}}\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n\}$. Since $\boldsymbol{c} \in \Lambda$, we can write

$$\boldsymbol{c} = \sum_{j=1}^{k} t_j\boldsymbol{b}_j,$$

for some integers $1 \leq k \leq n$ and $t_1, \ldots, t_k$. In fact, let us select a $\boldsymbol{c}$ like this with minimal possible k. Since $v_{kk} \neq 0$, we can choose an integer s such that

$$|t_k - sv_{kk}| < |v_{kk}|. \tag{1.4}$$

Then we clearly have

$$\boldsymbol{c} - s\boldsymbol{a}_k \in \Omega \setminus \operatorname{span}_{\mathbb{Z}}\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n\}.$$

Therefore we must have $t_k - sv_{kk} \neq 0$ by minimality of k. But then (1.4) contradicts the minimality of $|v_{kk}|$: we could take $\boldsymbol{c} - s\boldsymbol{a}_k$ instead of $\boldsymbol{a}_k$, since it satisfies all the conditions that $\boldsymbol{a}_k$ was chosen to satisfy, and then $|v_{kk}|$ is replaced by the smaller nonzero number $|t_k - sv_{kk}|$. This proves that $\boldsymbol{c}$ like this cannot exist, and so (1.3) is true, hence finishing one direction of the theorem.

Now suppose that we are given a basis $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n$ for Ω. We want to prove that there exists a basis $\boldsymbol{b}_1, \dots, \boldsymbol{b}_n$ for Λ such that relations in the statement of the theorem hold. This is a direct consequence of the argument in the proof of Theorem 1.3.1. Indeed, at i-th step of the basis construction in the proof of Theorem 1.3.1, we can choose i-th vector, call it $\boldsymbol{b}_i$, so that it lies in the span of the previous $i-1$ vectors and the vector $\boldsymbol{a}_i$. Since $\boldsymbol{b}_1, \dots, \boldsymbol{b}_n$ constructed this way are linearly independent (in fact, they form a basis for Λ by the construction), we obtain that

$$\boldsymbol{a}_i \in \operatorname{span}_{\mathbb{Z}}\{\boldsymbol{b}_1, \dots, \boldsymbol{b}_i\} \setminus \operatorname{span}_{\mathbb{Z}}\{\boldsymbol{b}_1, \dots, \boldsymbol{b}_{i-1}\},$$

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 1.3.5 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 1.19.

Remark 1.3.1. Let the notation be as in Theorem 1.3.5. Notice that if A is any basis matrix for Ω and B is any basis for Λ, then there exists an integral matrix V such that $A = BV$. Then Theorem 1.3.5 implies that for a given B there exists an A such that V is lower triangular, and for for a given A 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 ± 1, Theorem 1.3.5 can be thought of as the construction of *Hermite normal form* for an integral matrix. Problem 1.19 places additional restrictions that make Hermite normal form unique.

Here is an important implication of Theorem 1.3.5.

Theorem 1.3.6. *Let $\Omega \subseteq \Lambda$ be a sublattice. Then $\frac{\det(\Omega)}{\det(\Lambda)}$ is an integer; moreover, the number of cosets of Ω in Λ, i.e. the index of Ω as a subgroup of Λ is*

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

Proof. Let $\boldsymbol{b}_1, \dots, \boldsymbol{b}_n$ be a basis for Λ, and $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n$ be a basis for Ω, so that these two bases satisfy the conditions of Theorem 1.3.5, 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 matix with entries as described in Theorem 1.3.5; 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 $\boldsymbol{c} \in \Lambda$ is contained in the same coset of Ω in Λ as precisely one of the vectors

$$q_1\boldsymbol{b}_1 + \cdots + q_n\boldsymbol{b}_n,\ 0 \le q_i < v_{ii}\ \forall\ 1 \le i \le n,$$

in other words there are precisely $\prod_{i=1}^{n} |v_{ii}|$ cosets of Ω in Λ. This completes the proof. □

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

DEFINITION 1.3.4. A *fundamental domain* of a lattice Λ of full rank in $\mathbb{R}^n$ is a convex set $\mathcal{F} \subseteq \mathbb{R}^n$ containing $\mathbf{0}$, so that

$$\mathbb{R}^n = \bigcup_{\boldsymbol{x} \in \Lambda} (\mathcal{F} + \boldsymbol{x}),$$

and for every $\boldsymbol{x} \neq \boldsymbol{y} \in \Lambda$, $(\mathcal{F} + \boldsymbol{x}) \cap (\mathcal{F} + \boldsymbol{y}) = \emptyset$.

In other words, a fundamental domain of a lattice $\Lambda \subset \mathbb{R}^n$ is a *full set of coset representatives of* Λ *in* $\mathbb{R}^n$ (see Problem 1.20). 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 1.21).

DEFINITION 1.3.5. Let Λ be a lattice, and $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ be a basis for Λ. Then the set

$$\mathcal{F} = \left\{ \sum_{i=1}^{n} t_i \boldsymbol{a}_i : 0 \le t_i < 1,\ \forall\ 1 \le i \le n \right\},$$

is called a *fundamental parallelotope* of Λ with respect to the basis $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$. 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 $\overline{\mathcal{F}}$ 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 1.3.6. The *Voronoi cell* of a lattice Λ is the set

$$\mathcal{V}(\Lambda) = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| \le \|\boldsymbol{x} - \boldsymbol{y}\|\ \forall\ \boldsymbol{y} \in \Lambda\}.$$

It is easy to see that $\mathcal{V}(\Lambda)$ is (the closure of) a fundamental domain for Λ: 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 Λ be a lattice in $\mathbb{R}^n$ of rank $1 \le r \le n$, and let $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r$ be a basis for it. Write $A = (\boldsymbol{a}_1\ \ldots\ \boldsymbol{a}_r)$ for the corresponding $n \times r$ basis matrix of Λ, then A has rank r since its column vectors are linearly independent. For any $r \times r$ integral matrix U with determinant ± 1, AU is another basis matrix for Λ; moreover, if B is any other basis matrix for

Λ, there exists such a U so that $B = AU$. For each basis matrix A of Λ, we define the corresponding *Gram matrix* to be $M = A^\top A$, 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

$$\begin{aligned} \det(B^\top B) &= \det((AU)^\top (AU)) = \det(U^\top (A^\top A)U) \\ &= \det(U)^2 \det(A^\top A) = \det(A^\top A). \end{aligned}$$

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 1.3.7. Let Λ 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 Λ. The *determinant* of Λ is defined to be

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

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 Λ of not necessarily full rank. For instance, in order to define fundamental domains we can view Λ as a lattice inside of the vector space $\mathrm{span}_{\mathbb{R}}(\Lambda)$. The rest works essentially verbatim, keeping in mind that if $\Omega \subseteq \Lambda$ is a sublattice, then index $[\Lambda : \Omega]$, which is the number of cosets of Ω in Λ, is finite (and hence given by the formula of Theorem 1.3.6) if and only if $\mathrm{rk}(\Omega) = \mathrm{rk}(\Lambda)$.

1.4. 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 Λ in $\mathbb{R}^n$, how can we tell if M contains any points of Λ?

THEOREM 1.4.1 (Blichfeldt, 1914). *Let M be a Jordan measurable set in $\mathbb{R}^n$. Suppose that $\mathrm{Vol}(M) > 1$, or that M is closed, bounded and $\mathrm{Vol}(M) \geq 1$. Then there exist $\boldsymbol{x}, \boldsymbol{y} \in M$ such that $\mathbf{0} \neq \boldsymbol{x} - \boldsymbol{y} \in \mathbb{Z}^n$.*

PROOF. First suppose that $\mathrm{Vol}(M) > 1$. Let

$$P = \{\boldsymbol{x} \in \mathbb{R}^n : 0 \leq x_i < 1 \ \forall\ 1 \leq i \leq n\},$$

and let

$$S = \{\boldsymbol{u} \in \mathbb{Z}^n : M \cap (P + \boldsymbol{u}) \neq \emptyset\}.$$

Since M is bounded, S is a finite set, say $S = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{r_0}\}$. Write $M_r = M \cap (P + \boldsymbol{u}_r)$ for each $1 \leq r \leq r_0$. Also, for each $1 \leq r \leq r_0$, define

$$M'_r = M_r - \boldsymbol{u}_r,$$

so that $M'_1, \ldots, M'_{r_0} \subseteq P$. 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 $M_r \subseteq P + \boldsymbol{u}_r$, $M_s \subseteq P + \boldsymbol{u}_s$, and $(P + \boldsymbol{u}_r) \cap (P + \boldsymbol{u}_s) = \emptyset$. This means that

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

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

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

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

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

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 $\boldsymbol{x} \in M'_r \cap M'_s$. Then we have $\boldsymbol{x} + \boldsymbol{u}_r, \boldsymbol{x} + \boldsymbol{u}_s \in M$, and

$$(\boldsymbol{x} + \boldsymbol{u}_r) - (\boldsymbol{x} + \boldsymbol{u}_s) = \boldsymbol{u}_r - \boldsymbol{u}_s \in \mathbb{Z}^n.$$

Now suppose M is closed, bounded, and $\mathrm{Vol}(M) = 1$. 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

$$\boldsymbol{x}_r \neq \boldsymbol{y}_r \in s_r M$$

such that $\boldsymbol{x}_r - \boldsymbol{y}_r \in \mathbb{Z}^n$. Then there are subsequences $\{\boldsymbol{x}_{r_k}\}$ and $\{\boldsymbol{y}_{r_k}\}$ converging to points $\boldsymbol{x}, \boldsymbol{y} \in M$, respectively. Since for each r_k, $\boldsymbol{x}_{r_k} - \boldsymbol{y}_{r_k}$ is a nonzero lattice point, it must be true that $\boldsymbol{x} \neq \boldsymbol{y}$, and $\boldsymbol{x} - \boldsymbol{y} \in \mathbb{Z}^n$. This completes the proof. □

As a corollary of Theorem 1.4.1 we can prove the following version of *Minkowski Convex Body Theorem*.

THEOREM 1.4.2 (Minkowski). *Let $M \subset \mathbb{R}^n$ be a compact convex $\mathbf{0}$-symmetric set with $\mathrm{Vol}(M) \geq 2^n$. Then there exists $\mathbf{0} \neq \boldsymbol{x} \in M \cap \mathbb{Z}^n$.*

PROOF. Notice that the set

$$\frac{1}{2}M = \left\{\frac{1}{2}\boldsymbol{x} : \boldsymbol{x} \in M\right\} = \begin{pmatrix} 1/2 & 0 & \dots & 0 \\ 0 & 1/2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/2 \end{pmatrix} M$$

is also convex, $\mathbf{0}$-symmetric, and by Problem 1.22 its volume is

$$\det \begin{pmatrix} 1/2 & 0 & \dots & 0 \\ 0 & 1/2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/2 \end{pmatrix} \mathrm{Vol}(M) = 2^{-n}\,\mathrm{Vol}(M) \geq 1.$$

Thererfore, by Theorem 1.4.1, there exist $\frac{1}{2}\boldsymbol{x} \neq \frac{1}{2}\boldsymbol{y} \in \frac{1}{2}M$ such that

$$\frac{1}{2}\boldsymbol{x} - \frac{1}{2}\boldsymbol{y} \in \mathbb{Z}^n.$$

But, by symmetry, since $\boldsymbol{y} \in M$, $-\boldsymbol{y} \in M$, and by convexity, since $\boldsymbol{x}, -\boldsymbol{y} \in M$,

$$\frac{1}{2}\boldsymbol{x} - \frac{1}{2}\boldsymbol{y} = \frac{1}{2}\boldsymbol{x} + \frac{1}{2}(-\boldsymbol{y}) \in M.$$

This completes the proof. □

REMARK 1.4.1. This result is sharp: for any $\varepsilon > 0$, the cube

$$C = \left\{\boldsymbol{x} \in \mathbb{R}^n : \max_{1 \leq i \leq n} |x_i| \leq 1 - \frac{\varepsilon}{2}\right\}$$

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

Problem 1.23 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 $\mathrm{Vol}(M) \geq \det \Lambda$, then there exist $\boldsymbol{x}, \boldsymbol{y} \in M$ such that $\mathbf{0} \neq \boldsymbol{x} - \boldsymbol{y} \in \Lambda$.
- If $\Lambda \subset \mathbb{R}^n$ is a lattice of full rank and $M \subset \mathbb{R}^n$ is a compact convex $\mathbf{0}$-symmetric set with $\mathrm{Vol}(M) \geq 2^n \det \Lambda$, then there exists $\mathbf{0} \neq \boldsymbol{x} \in M \cap \Lambda$.

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

THEOREM 1.4.3. *Let $B = (b_{ij})_{1 \leq i,j \leq n} \in \mathrm{GL}_n(\mathbb{R})$, and for each $1 \leq i \leq n$ define a linear form with coefficients $b_{i1}, \dots, b_{in}$ by*

$$L_i(\boldsymbol{X}) = \sum_{j=1}^{n} b_{ij} X_j.$$

Let $c_1, \dots, c_n \in \mathbb{R}_{>0}$ be such that

$$c_1 \dots c_n = |\det(B)|.$$

Then there exists $\mathbf{0} \neq \boldsymbol{x} \in \mathbb{Z}^n$ such that

$$|L_i(\boldsymbol{x})| \leq c_i,$$

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

PROOF. Let us write $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ for the row vectors of B, then

$$L_i(\boldsymbol{x}) = \boldsymbol{b}_i \boldsymbol{x},$$

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

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

where $R = \{\boldsymbol{x} \in \mathbb{R}^n : |x_i| \leq c_i \ \forall\ 1 \leq i \leq n\}$ is the rectangular box with sides of length $2c_1, \ldots, 2c_n$ centered at the origin in $\mathbb{R}^n$. Then by Problem 1.22,

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

and so by Theorem 1.4.2 there exists $\mathbf{0} \neq \boldsymbol{x} \in P \cap \mathbb{Z}^n$. □

1.5. Successive minima

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

COROLLARY 1.5.1. *Let $M \subset \mathbb{R}^n$ be a compact convex $\mathbf{0}$-symmetric and $\Lambda \subset \mathbb{R}^n$ a lattice of full rank. Define the first successive minimum of M with respect to Λ to be*

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

Then

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

PROOF. The fact that λ_1 has to be positive readily follows from Λ being a discrete set. Hence we only have to prove the upper bound. By Theorem 1.4.2 for a general lattice Λ (Problem 1.23), if

$$\mathrm{Vol}(\lambda M) \geq 2^n \det(\Lambda),$$

then λM contains a nonzero point of Λ. On the other hand, by Problem 1.22,

$$\mathrm{Vol}(\lambda M) = \lambda^n \mathrm{Vol}(M).$$

Hence as long as

$$\lambda^n \mathrm{Vol}(M) \geq 2^n \det(\Lambda),$$

the expanded set λM is guaranteed to contain a nonzero point of Λ. The conclusion of the corollary follows. □

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 Λ: this is the meaning of λ_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 Λ, 3 linearly independent points of Λ, etc. To answer this question is the main objective of this section. We start with a definition.

DEFINITION 1.5.1. Let M be a convex, $\mathbf{0}$-symmetric set $M \subset \mathbb{R}^n$ of nonzero volume and $\Lambda \subseteq \mathbb{R}^n$ a lattice of full rank. For each $1 \leq i \leq n$ define the i-th *succesive minimum* of M with respect to Λ, λ_i, to be the infimum of all positive real numbers λ such that the set λM contains at least i linearly independent points of Λ. In other words,

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

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

REMARK 1.5.1. Notice that the n linearly independent vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ 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 $n = 1, 2, 3$, and in dimension $n = 4$ there always exists a basis consisting of vectors corresponding to successive minima, but when $n = 5$ there is a well known counterexample. Let

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 1 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{2} \end{pmatrix} \mathbb{Z}^5,$$

and let $M = \mathbb{B}_5$, the closed unit ball centered at $\mathbf{0}$ in $\mathbb{R}^n$. Then the successive minima of $\mathbb{B}_5$ with respect to Λ is

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

since $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_5 \in \mathbb{B}_5 \cap \Lambda$, and

$$\boldsymbol{x} = \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)^\top \notin \mathbb{B}_5.$$

On the other hand, $\boldsymbol{x}$ cannot be expressed as a linear combination of $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_5$ with integer coefficients, hence

$$\operatorname{span}_{\mathbb{Z}}\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_5\} \subset \Lambda.$$

An immediate observation is that

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

and Corollary 1.5.1 gives an upper bound on λ_1. Can we produce bounds on all the successive minima in terms of $\operatorname{Vol}(M)$ and $\det(\Lambda)$? This question is answered by *Minkowski's Successive Minima Theorem.*

THEOREM 1.5.2. *With notation as above,*

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

PROOF. We present the proof in case $\Lambda = \mathbb{Z}^n$, leaving generalization of the given argument to arbitrary lattices as an excercise. We start with a proof of the lower bound following [**GL87**], which is considerably easier than the upper bound. Let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ be the n linearly independent vectors corresponding to the respective successive minima $\lambda_1, \ldots, \lambda_n$, and let

$$U = (\boldsymbol{u}_1 \ldots \boldsymbol{u}_n) = \begin{pmatrix} u_{11} & \ldots & u_{n1} \\ \vdots & \ddots & \vdots \\ u_{1n} & \ldots & u_{nn} \end{pmatrix}.$$

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

$$\pm\frac{\boldsymbol{u}_1}{\lambda_1}, \ldots, \pm\frac{\boldsymbol{u}_n}{\lambda_n}$$

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 1.24).

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

$$\text{Vol}(P) = \frac{2^n}{n!} \left| \det \begin{pmatrix} \frac{u_{11}}{\lambda_1} & \cdots & \frac{u_{n1}}{\lambda_n} \\ \vdots & \ddots & \vdots \\ \frac{u_{1n}}{\lambda_1} & \cdots & \frac{u_{nn}}{\lambda_n} \end{pmatrix} \right| = \frac{2^n |\det(U)|}{n!\ \lambda_1 \dots \lambda_N} \geq \frac{2^n}{n!\ \lambda_1 \dots \lambda_n}, \tag{1.5}$$

since $\det(U)$ is an integer. Since $P \subseteq M$, $\text{Vol}(M) \geq \text{Vol}(P)$. Combining this last observation with (1.5) 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

$$E_i = \text{span}_{\mathbb{R}}\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_i\},$$

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 $\boldsymbol{u}_1, \dots, \boldsymbol{u}_n$ to be the n linearly independent vectors corresponding to the respective successive minima $\lambda_1, \dots, \lambda_n$. In fact, notice that there exists a matrix $A \in GL_n(\mathbb{Z})$ such that

$$A\,\text{span}_{\mathbb{R}}\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_i\} \subseteq E_i,$$

for each $1 \leq i \leq n$, i.e. we can rotate each $\text{span}_{\mathbb{R}}\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_i\}$ 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

$$A\boldsymbol{u}_i \in \lambda_i' AM \cap E_i,\ \forall\ 1 \leq i \leq n,$$

where $\lambda_1', \dots \lambda_n'$ is the successive minima of AM with respect to $\mathbb{Z}^n$. Hence we can assume without loss of generality that

$$\text{span}_{\mathbb{R}}\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_i\} \subseteq E_i,$$

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

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

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

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

$$C_q^i = C_q^n \cap E_i.$$

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

$$\text{Vol}(C_q^n + M_n) \leq (2q + \gamma)^n, \tag{1.6}$$

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

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

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

$$\boldsymbol{z} \in \text{int}(\boldsymbol{x} + M_1) \cap \text{int}(\boldsymbol{y} + M_1),$$

and so

$$\begin{aligned}(\boldsymbol{z}-\boldsymbol{x})-(\boldsymbol{z}-\boldsymbol{y}) &= \boldsymbol{y}-\boldsymbol{x}\in \operatorname{int}(M_1)-\operatorname{int}(M_1)\\ &= \{\boldsymbol{z}_1-\boldsymbol{z}_2 : \boldsymbol{z}_1,\boldsymbol{z}_2\in M_1\}=\operatorname{int}(\lambda_1 M),\end{aligned} \tag{1.7}$$

which would contradict minimality of λ_1. Therefore

$$\operatorname{Vol}(C_q^n+M_1)=(2q+1)^n\operatorname{Vol}(M_1)=(2q+1)^n\left(\frac{\lambda_1}{2}\right)^n\operatorname{Vol}(M). \tag{1.8}$$

To finish the proof, we need the following lemma.

LEMMA 1.5.3. *For each* $1\le i\le n-1$,

$$\operatorname{Vol}(C_q^n+M_{i+1})\ge\left(\frac{\lambda_{i+1}}{\lambda_i}\right)^{n-i}\operatorname{Vol}(C_q^n+M_i). \tag{1.9}$$

PROOF. If $\lambda_{i+1}=\lambda_i$ the statement is obvious, so assume $\lambda_{i+1}>\lambda_i$. Let $\boldsymbol{x},\boldsymbol{y}\in\mathbb{Z}^n$ be such that

$$(x_{i+1},\dots,x_n)\neq(y_{i+1},\dots,y_n).$$

Then

$$(\boldsymbol{x}+\operatorname{int}(M_{i+1}))\cap(\boldsymbol{y}+\operatorname{int}(M_{i+1}))=\emptyset. \tag{1.10}$$

Indeed, suppose (1.10) is not true, i.e. there exists $\boldsymbol{z}\in(\boldsymbol{x}+\operatorname{int}(M_{i+1}))\cap(\boldsymbol{y}+\operatorname{int}(M_{i+1}))$. Then, as in (1.7) above, $\boldsymbol{x}-\boldsymbol{y}\in\operatorname{int}(\lambda_{i+1}M)$. But we also have

$$\boldsymbol{u}_1,\dots,\boldsymbol{u}_i\in\operatorname{int}(\lambda_{i+1}M),$$

since $\lambda_{i+1}>\lambda_i$, and so $\lambda_i M\subseteq\operatorname{int}(\lambda_{i+1}M)$. Moreover, $\boldsymbol{u}_1,\dots,\boldsymbol{u}_i\in E_i$, meaning that

$$u_{jk}=0\ \forall\ 1\le j\le i,\ i+1\le k\le n.$$

On the other hand, at least one of

$$x_k-y_k,\ i+1\le k\le n,$$

is not equal to 0. Hence $\boldsymbol{x}-\boldsymbol{y},\boldsymbol{u}_1,\dots,\boldsymbol{u}_i$ are linearly independent, but this means that $\operatorname{int}(\lambda_{i+1}M)$ contains $i+1$ linearly independent points, contradicting minimality of λ_{i+1}. This proves (1.10). Notice that (1.10) implies

$$\operatorname{Vol}(C_q^n+M_{i+1})=(2q+1)^{n-i}\operatorname{Vol}(C_q^i+M_{i+1}),$$

and

$$\operatorname{Vol}(C_q^n+M_i)=(2q+1)^{n-i}\operatorname{Vol}(C_q^i+M_i),$$

since $M_i\subseteq M_{i+1}$. Hence, in order to prove the lemma it is sufficient to prove that

$$\operatorname{Vol}(C_q^i+M_{i+1})\ge\left(\frac{\lambda_{i+1}}{\lambda_i}\right)^{n-i}\operatorname{Vol}(C_q^i+M_i). \tag{1.11}$$

Define two linear maps $f_1,f_2:\mathbb{R}^n\to\mathbb{R}^n$, given by

$$f_1(\boldsymbol{x})=\left(\frac{\lambda_{i+1}}{\lambda_i}x_1,\dots,\frac{\lambda_{i+1}}{\lambda_i}x_i,x_{i+1},\dots,x_n\right),$$

$$f_2(\boldsymbol{x})=\left(x_1,\dots,x_i,\frac{\lambda_{i+1}}{\lambda_i}x_{i+1},\dots,\frac{\lambda_{i+1}}{\lambda_i}x_n\right),$$

and notice that $f_2(f_1(M_i))=M_{i+1}$, $f_2(C_q^i)=C_q^i$. Therefore

$$f_2(C_q^i+f_1(M_i))=C_q^i+M_{i+1}.$$

This implies that

$$\mathrm{Vol}(C_q^i + M_{i+1}) = \left(\frac{\lambda_{i+1}}{\lambda_i}\right)^{n-i} \mathrm{Vol}(C_q^i + f_1(M_i)),$$

and so to establish (1.11) it is sufficient to show that

$$\mathrm{Vol}(C_q^i + f_1(M_i)) \geq \mathrm{Vol}(C_q^i + M_i). \tag{1.12}$$

Let

$$E_i^{\perp} = \mathrm{span}_{\mathbb{R}}\{\boldsymbol{e}_{i+1}, \ldots, \boldsymbol{e}_n\},$$

i.e. $E_i^{\perp}$ is the orthogonal complement of E_i, and so has dimension $n-i$. Notice that for every $\boldsymbol{x} \in E_i^{\perp}$ there exists $\boldsymbol{t}(\boldsymbol{x}) \in E_i$ such that

$$M_i \cap (\boldsymbol{x} + E_i) \subseteq (f_1(M_i) \cap (\boldsymbol{x} + E_i)) + \boldsymbol{t}(\boldsymbol{x}),$$

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

$$(C_q^i + M_i) \cap (\boldsymbol{x} + E_i) \subseteq (C_q^i + f_1(M_i)) \cap (\boldsymbol{x} + E_i)) + \boldsymbol{t}(\boldsymbol{x}),$$

and hence

$$\begin{aligned}
\mathrm{Vol}(C_q^i + M_i) &= \int_{\boldsymbol{x} \in E_i^{\perp}} \mathrm{Vol}_i((C_q^i + M_i) \cap (\boldsymbol{x} + E_i))\, d\boldsymbol{x} \\
&\leq \int_{\boldsymbol{x} \in E_i^{\perp}} \mathrm{Vol}_i((C_q^i + f_1(M_i)) \cap (\boldsymbol{x} + E_i))\, d\boldsymbol{x} \\
&= \mathrm{Vol}(C_q^i + f_1(M_i)),
\end{aligned}$$

where Vol_i stands for the i-dimensional volume. This completes the proof of (1.12), and hence of the lemma. □

Now, combining (1.6), (1.8), and (1.9), we obtain:

$$\begin{aligned}
(2q+\gamma)^n &\geq \mathrm{Vol}(C_q^n + M_n) \geq \left(\frac{\lambda_n}{\lambda_{n-1}}\right) \mathrm{Vol}(C_q^n + M_{n-1}) \geq \ldots \\
&\geq \left(\frac{\lambda_n}{\lambda_{n-1}}\right)\left(\frac{\lambda_{n-1}}{\lambda_{n-2}}\right)^2 \cdots \left(\frac{\lambda_2}{\lambda_1}\right)^{n-1} \mathrm{Vol}(C_q^n + M_1) \\
&= \lambda_n \ldots \lambda_1 \frac{\mathrm{Vol}(M)}{2^n}(2q+1)^n,
\end{aligned}$$

hence

$$\lambda_1 \ldots \lambda_n \leq \frac{2^n}{\mathrm{Vol}(M)} \left(\frac{2q+\gamma}{2q+1}\right)^n \to \frac{2^n}{\mathrm{Vol}(M)},$$

as $q \to \infty$, since $q \in \mathbb{Z}_{>0}$ is arbitrary. This completes the proof. □

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

1.6. 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 $\mathbf{0}$-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 Λ. We define the *inhomogeneous minimum* of M with respect to Λ 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 μ. We start with the following result of Jarnik [**Jar41**].

Lemma 1.6.1.

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

Proof. Let us define a function

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

for every $\boldsymbol{x} \in \mathbb{R}^n$. This function is a norm (Problem 1.25). Then

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

can be thought of as the unit ball with respect to this norm. We will say that F is the *norm of* M. Let $\boldsymbol{z} \in \mathbb{R}^n$ be an arbitrary point. We want to prove that there exists a point $\boldsymbol{v} \in \Lambda$ such that

$$F(\boldsymbol{z} - \boldsymbol{v}) \leq \frac{1}{2}\sum_{i=1}^{n} \lambda_i.$$

This would imply that $\boldsymbol{z} \in \left(\frac{1}{2}\sum_{i=1}^{n} \lambda_i\right) M + \boldsymbol{v}$, and hence settle the lemma, since $\boldsymbol{z}$ is arbitrary. Let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ be the linearly independent vectors corresponding to successive minima $\lambda_1, \ldots, \lambda_n$, respectively. Then

$$F(\boldsymbol{u}_i) = \lambda_i, \ \forall\ 1 \leq i \leq n.$$

Since $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ form a basis for $\mathbb{R}^n$, there exist $a_1, \ldots, a_n \in \mathbb{R}$ such that

$$\boldsymbol{z} = \sum_{i=1}^{n} a_i \boldsymbol{u}_i.$$

We can also choose integer $v_1, \ldots, v_n$ such that

$$|a_i - v_i| \leq \frac{1}{2}, \ \forall\ 1 \leq i \leq n,$$

and define $\boldsymbol{v} = \sum_{i=1}^{n} v_i \boldsymbol{u}_i$, hence $\boldsymbol{v} \in \Lambda$. Now notice that

$$\begin{aligned} F(\boldsymbol{z} - \boldsymbol{v}) &= F\left(\sum_{i=1}^{n}(a_i - v_i)\boldsymbol{u}_i\right) \\ &\leq \sum_{i=1}^{n} |a_i - v_i| F(\boldsymbol{u}_i) \leq \frac{1}{2}\sum_{i=1}^{n} \lambda_i, \end{aligned}$$

since F is a norm. This completes the proof. □

Using Lemma 1.6.1 along with Minkowski's successive minima theorem, we can obtain some bounds on μ in terms of the determinant of Λ and volume of M. A nice bound can be easily obtained in an important special case.

COROLLARY 1.6.2. *If $\lambda_1 \geq 1$, then*

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

PROOF. Since

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

Theorem 1.5.2 implies

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

and by Lemma 1.6.1,

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

The result follows by combining these two inequalities. □

A general bound depending also on λ_1 was obtained by Scherk [**Sch50**], once again using Minkowski's successive minima theorem (Theorem 1.5.2) and Jarnik's inequality (Lemma 1.6.1) He observed that if λ_1 is fixed and $\lambda_2, \ldots, \lambda_n$ are subject to the conditions

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

then the maximum of the sum

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

is attained when

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

Hence we obtain Scherk's inequality for μ.

COROLLARY 1.6.3.

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

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

$$\mathrm{Vol}(\mu M) = \mu^n \mathrm{Vol}(M) \geq \mathrm{Vol}(\mathcal{F}) = \det(\Lambda),$$

hence

$$\mu \geq \left(\frac{\det(\Lambda)}{\mathrm{Vol}(M)} \right)^{1/n}. \tag{1.13}$$

In fact, by Theorem 1.5.2,

$$\left(\frac{\det(\Lambda)}{\mathrm{Vol}(M)} \right)^{1/n} \geq \frac{(\lambda_1 \ldots \lambda_n)^{1/n}}{2} \geq \frac{\lambda_1}{2},$$

and combining this with (1.13), we obtain

$$\mu \geq \frac{\lambda_1}{2}. \tag{1.14}$$

Jarnik obtained a considerably better lower bound for μ in [**Jar41**].

LEMMA 1.6.4.

$$\mu \geq \frac{\lambda_n}{2}.$$

PROOF. Let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ be the linearly independent points of Λ corresponding to the successive minima $\lambda_1, \ldots, \lambda_n$ of M with respect to Λ. Let F be the norm of M, then

$$F(\boldsymbol{u}_i) = \lambda_i, \ \forall\ 1 \leq i \leq n.$$

We will first prove that for every $\boldsymbol{x} \in \Lambda$,

$$F\left(\boldsymbol{x} - \frac{1}{2}\boldsymbol{u}_n\right) \geq \frac{1}{2}\lambda_n. \tag{1.15}$$

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

$$F(\boldsymbol{x}) \leq F\left(\boldsymbol{x} - \frac{1}{2}\boldsymbol{u}_n\right) + F\left(\frac{1}{2}\boldsymbol{u}_n\right) < \frac{1}{2}\lambda_n + \frac{1}{2}\lambda_n = \lambda_n,$$

and similarly

$$F(\boldsymbol{u}_n - \boldsymbol{x}) \leq F\left(\frac{1}{2}\boldsymbol{u}_n - \boldsymbol{x}\right) + F\left(\frac{1}{2}\boldsymbol{u}_n\right) < \lambda_n.$$

Therefore, by definition of λ_n,

$$\boldsymbol{x}, \boldsymbol{u}_n - \boldsymbol{x} \in \operatorname{span}_{\mathbb{R}}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{n-1}\},$$

and so $\boldsymbol{u}_n = \boldsymbol{x} + (\boldsymbol{u}_n - \boldsymbol{x}) \in \operatorname{span}_{\mathbb{R}}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{n-1}\}$, which is a contradiction. Hence we proved (1.15) for all $\boldsymbol{x} \in \Lambda$. Further, by Problem 1.26,

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

Then lemma follows by combining this observation with (1.15). □

We define the *inhomogeneous minimum of* Λ to be the inhomogeneous minimum of the closed unit ball $\mathbb{B}_n$ with respect to Λ, since it will occur quite often. This is another invariant of the lattice.

1.7. Problems

PROBLEM 1.1. *Prove that* $\| \ \|_p$ *for each* $p \in \mathbb{Z}_{>0}$ *and* $| \ |$*, as defined in Section 1.2 are indeed norms on* $\mathbb{R}^n$*.*

PROBLEM 1.2. *Let* F *be a norm on* $\mathbb{R}^n$*, and let* $C \in \mathbb{R}$ *be a positive number. Define*

$$A_n(C) = \{\boldsymbol{x} \in \mathbb{R}^n : F(\boldsymbol{x}) \leq C\}.$$

Prove that $A_n(C)$ *is a convex set. What is* $A_n(C)$ *when* $F = \| \ \|_1$*?*

PROBLEM 1.3. *Prove that a hyperplane* $\mathbb{H}$ *as defined in* (1.1) *is a subspace of* $\mathbb{R}^n$ *if and only if* $b = 0$*. Prove that in this case dimension of* $\mathbb{H}$ *is* $n - 1$ *(we define co-dimension of an* ℓ*-dimensional subspace of an* n*-dimensional vector space, where* $1 \leq \ell \leq n$*, to be* $n - \ell$*; thus co-dimension of* $\mathbb{H}$ *here is 1, as indicated above).*

PROBLEM 1.4. *Prove that each convex polytope in* $\mathbb{R}^n$ *can be described as a bounded intersection of finitely many halfspaces, and vice versa.*

PROBLEM 1.5. *Let* $f(X_1, \ldots, X_n)$ *be a homogeneous polynomial of degree* d *with real coefficients. Prove that*

$$F(\boldsymbol{x}) = |f(\boldsymbol{x})|^{1/d}$$

is a distance function.

PROBLEM 1.6. *If* F *is a distance function on* $\mathbb{R}^n$*, prove that the set*

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

is a bounded star body.

PROBLEM 1.7. *Let* X *be a star body, and let* F *be its distance function, i.e.* $X = \{\boldsymbol{x} \in \mathbb{R}^n : F(\boldsymbol{x}) \leq 1\}$*. Prove that*

$$F(\boldsymbol{x} + \boldsymbol{y}) \leq F(\boldsymbol{x}) + F(\boldsymbol{y}),$$

for all $\boldsymbol{x}, \boldsymbol{y} \in X$ *if and only if* X *is a convex set.*

PROBLEM 1.8. *Let* $f : X \to Y$ *be a bijection. Prove that* f *has an inverse* f^{-1}*. In other words, prove that there exists a function* $f^{-1} : Y \to X$ *such that for every* $x \in X$ *and* $y \in Y$*,*

$$f^{-1}(f(x)) = x, \ f(f^{-1}(y)) = y.$$

PROBLEM 1.9. *Let* $\mathbb{R}$ *be the set of all real numbers, and define sets*

$$L_1 = \{(x, x) : x \in \mathbb{R}\},$$

$$L_2 = \{(x, x) : x \in \mathbb{R}, x \geq 0\} \cup \{(x, -x) : x \in \mathbb{R}, x < 0\}.$$

(1) *Prove that* L_1 *is diffeomorphic to* $\mathbb{R}$*.*
(2) *Prove that* L_2 *is homeomorphic to* $\mathbb{R}$ *by explicitly constructing a homeomorphism.*
(3) *Is the homeomorphism you constructed in part (2) a diffeomorphism?*

PROBLEM 1.10. *Prove that cones like C_i^1 and C_i^2 defined in* (1.2) *have Jordan volume.*

PROBLEM 1.11. *Let $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r \in \mathbb{R}^n$ be linearly independent points. Prove that $r \leq n$.*

PROBLEM 1.12. *Prove that if Λ is a lattice of rank r in $\mathbb{R}^n$, $1 \leq r \leq n$, then $\operatorname{span}_{\mathbb{R}} \Lambda$ is a subspace of $\mathbb{R}^n$ of dimension r (by $\operatorname{span}_{\mathbb{R}} \Lambda$ we mean the set of all finite real linear combinations of vectors from Λ).*

PROBLEM 1.13. *Let Λ be a lattice of rank r in $\mathbb{R}^n$. By Problem 1.12, $V = \operatorname{span}_{\mathbb{R}} \Lambda$ is an r-dimensional subspace of $\mathbb{R}^n$. Prove that Λ is a discrete co-compact subset of V.*

PROBLEM 1.14. *Let Λ be a lattice of rank r in $\mathbb{R}^n$, and let $V = \operatorname{span}_{\mathbb{R}} \Lambda$ be an r-dimensional subspace of $\mathbb{R}^n$, as in Problem 1.13 above. Prove that Λ and V are both additive groups, and Λ is a subgroup of V.*

PROBLEM 1.15. *Let Λ be a lattice and Ω a subset of Λ. Prove that Ω is a sublattice of Λ if and only if it is a subgroup of the abelian group Λ.*

PROBLEM 1.16. *Let Λ be a lattice and Ω a sublattice of Λ of the same rank. Prove that two cosets $\boldsymbol{x} + \Omega$ and $\boldsymbol{y} + \Omega$ of Ω in Λ are equal if and only if $\boldsymbol{x} - \boldsymbol{y} \in \Omega$. Conclude that a coset $\boldsymbol{x} + \Omega$ is equal to Ω if and only if $\boldsymbol{x} \in \Omega$.*

PROBLEM 1.17. *Let Λ be a lattice and $\Omega \subseteq \Lambda$ a sublattice. Suppose that the quotient group Λ/Ω is finite. Prove that rank of Ω is the same as rank of Λ.*

PROBLEM 1.18. *Given a lattice Λ and a real number μ, define*

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

Prove that $\mu\Lambda$ is a lattice. Prove that if μ is an integer, then $\mu\Lambda$ is a sublattice of Λ.

PROBLEM 1.19. *Prove that it is possible to select the coefficients v_{ij} in Theorem 1.3.5 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 1.20. *Prove that for every point $\boldsymbol{x} \in \mathbb{R}^n$ there exists uniquely a point $\boldsymbol{y} \in \mathcal{F}$ such that*

$$\boldsymbol{x} - \boldsymbol{y} \in \Lambda,$$

i.e. $\boldsymbol{x}$ lies in the coset $\boldsymbol{y} + \Lambda$ of Λ in $\mathbb{R}^n$. This means that $\mathcal{F}$ is a full set of coset representatives of Λ in $\mathbb{R}^n$.

PROBLEM 1.21. *Prove that volume of a fundamental parallelotope is equal to the determinant of the lattice.*

PROBLEM 1.22. *Let S be a compact convex set in $\mathbb{R}^n$, $A \in \mathrm{GL}_n(\mathbb{R})$, and define*

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

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

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 1.23. *Prove versions of Theorems 1.4.1 - 1.4.2 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 $\Lambda = A\mathbb{Z}^n$ for some $A \in \mathrm{GL}_n(\mathbb{R})$. Then a point $\boldsymbol{x} \in A^{-1}M \cap \mathbb{Z}^n$ if and only if $A\boldsymbol{x} \in M \cap \Lambda$. Now use Problem 1.22 to relate the volume of $A^{-1}M$ to the volume of M.

PROBLEM 1.24. *Prove that a standard simplex in $\mathbb{R}^n$ has volume $1/n!$.*

PROBLEM 1.25. *Let $M \subset \mathbb{R}^n$ be a compact convex $\mathbf{0}$-symmetric set. Define a function $F : \mathbb{R}^n \to \mathbb{R}$, given by*

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

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

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

PROBLEM 1.26. *Let F be a norm like in Problem 1.25. Prove that the inhomogeneous minimum of the corresponding set M with respect to the full-rank lattice $\Lambda \subset \mathbb{R}^n$ satisfies*

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

PROBLEM 1.27. *Let $n \geq 2$, and let $F : \mathbb{R}^n \to \mathbb{R}$ be a continuous function such that*

(1) *$F(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$,*
(2) *$F(a\boldsymbol{x}) = |a|F(\boldsymbol{x})$ for all $a \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^n$.*

Let

$$X = \{\boldsymbol{x} \in \mathbb{R}^n : F(\boldsymbol{x}) \leq 1\}.$$

Assume additionally that F satisfies the triangle inequality:

$$F(\boldsymbol{x} + \boldsymbol{y}) \leq F(\boldsymbol{x}) + F(\boldsymbol{y}), \tag{1.16}$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Let Λ be a lattice of full rank in $\mathbb{R}^n$. Prove that for every real number

$$\mu \geq 2 \left(\frac{\det(\Lambda)}{\mathrm{Vol}(X)} \right)^{1/n}$$

the intersection $\mu X \cap \Lambda$ contains a nonzero vector. Is this statement still true if F does not satisfy the triangle inequality (4.11)? Either prove your answer or give a counter-example.

PROBLEM 1.28. *Let Λ be a lattice of full rank in $\mathbb{R}^n$, and let $M \subset \mathbb{R}^n$ be a compact convex set such that $\mathrm{Vol}(M) < \det(\Lambda)$. Prove that there exists a point $\boldsymbol{x} \in \mathbb{R}^n$ such that the intersection $M \cap (\Lambda + \boldsymbol{x})$ is empty.*

PROBLEM 1.29. *Let a, b be positive real numbers, and suppose that*

$$\Lambda = \begin{pmatrix} a & b & 0 & 0 \\ 0 & 1 & a & b \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \mathbb{Z}^4$$

is a full-rank sublattice of $\mathbb{Z}^4$.

(1) *What are all the possible values of a and b?*
(2) *Suppose that $a = b$ and Ω is a full-rank sublattice of Λ, such that the volume of a fundamental domain of Ω in $\mathbb{R}^4$ is equal to 20. What are all the possible values of a?*
(3) *Assuming part b, can Λ be a sublattice of any of the following two lattices:*

$$L_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 2 \end{pmatrix} \mathbb{Z}^4, \; L_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 3 \end{pmatrix} \mathbb{Z}^4,$$

and if so, which one(s)? What are all the possible values of a in each case?

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 [**CS99**].

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

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

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

$$\operatorname{int}(B_i') \cap \operatorname{int}(B_j') \neq \emptyset,$$

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 Λ of full rank in $\mathbb{R}^n$; such packings are called *lattice packings*. Alhtough 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

$$\begin{aligned} \Delta = \Delta(\Lambda) \ &:= \ \text{proportion of the space occupied by spheres} \\ &= \ \frac{\text{volume of one sphere}}{\text{volume of a fundamental domain of } \Lambda} \\ &= \ \frac{r^n \omega_n}{\det(\Lambda)}, \end{aligned}$$

where r is the *packing radius*, i.e. radius of each sphere in this lattice packing, and ω_n is the volume of a unit ball in $\mathbb{R}^n$, given by

$$\omega_n = \begin{cases} \frac{\pi^k}{k!} & \text{if } n = 2k \text{ for some } k \in \mathbb{Z} \\ \frac{2^{2k+1}k!\pi^k}{(2k+1)!} & \text{if } n = 2k+1 \text{ for some } k \in \mathbb{Z}. \end{cases} \tag{2.1}$$

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 $\mathcal{V}$ of Λ, i.e. the *inradius* of $\mathcal{V}$. 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 nonzero vector in Λ, in other words $r = \frac{\lambda_1}{2}$, where λ_1 is the first successive minimum of Λ. Therefore

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

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*

$$\Delta \geq \frac{\zeta(n)}{2^{n-1}}, \tag{2.2}$$

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

All known proofs of Theorem 7.3.3 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 7.3.3 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 $\mathbf{0}$-symmetric set in $\mathbb{R}^n$ with volume $< 2\zeta(n)$. Then there exists a lattice Λ in $\mathbb{R}^n$ of determinant 1 such that M contains no points of Λ except for $\mathbf{0}$.*

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

$$\chi_M(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{x} \in M \\ 0 & \text{if } \boldsymbol{x} \notin M \end{cases}$$

for every $\boldsymbol{x} \in \mathbb{R}^n$. 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

$$\Lambda = \begin{pmatrix} T & 0 & \dots & 0 & \xi_1 \\ 0 & T & \dots & 0 & \xi_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & T & \xi_{n-1} \\ 0 & 0 & \dots & 0 & T^{-(n-1)} \end{pmatrix} \mathbb{Z}^n. \tag{2.3}$$

Hence determinant of this lattice is 1 independent of the values of the parameters. Points of Λ with $b = 0$ are of the form

$$(Ta_1, \dots, Ta_{n-1}, 0),$$

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 Λ in M have $b \neq 0$. Notice that M contains a nonzero point of Λ if and only if it contains a primitive point of Λ, where we say that $\boldsymbol{x} \in \Lambda$ is primitive if it is not a scalar multiple of another point in Λ. The number of symmetric pairs of primitive points of Λ in M is given by the counting function $\eta_T(\xi_1, \dots, \xi_{n-1}) =$

$$\sum_{b>0} \sum_{\substack{a_1,\dots,a_{n-1} \\ \gcd(a_1,\dots,a_{n-1},b)=1}} \chi_M \left(T(a_1 + \xi_1 b), \dots, T(a_{n-1} + \xi_{n-1} b), T^{-(n-1)} b \right).$$

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 $\eta_T(\xi_1, \dots, \xi_{n-1})$ 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*

$$\Lambda_h := \begin{pmatrix} 1 & \frac{1}{2} \\ 0 & \frac{\sqrt{3}}{2} \end{pmatrix} \mathbb{Z}^2, \tag{2.4}$$

and in 1940 L. Fejes Tóth proved that this indeed is the optimal packing (a previous proof by Axel Thue. Its density is $\frac{\pi\sqrt{3}}{6} \approx 0.9068996821$.

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 ≈ 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 [**CS99**]. 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* Θ of corresponding sphere covering is defined to be

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

where ω_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 $\mathcal{V}$ of Λ, i.e. the *circumradius* of $\mathcal{V}$. Clearly $\Theta \geq 1$.

Notice that the covering radius R is precisely μ, the inhomogeneous minimum of the lattice Λ. Hence combining Lemmas 1.6.1 and 1.6.4 we obtain the following bounds on the covering radius in terms of successive minima of Λ:

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

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 ≈ 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 Λ are very much dependent on its Voronoi cell $\mathcal{V}$. Moreover, to simultaneously optimize packing and covering properties of Λ we want to ensure that the inradius r of $\mathcal{V}$ 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 Λ: 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 Λ is WR, then by Lemma 1.6.4 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 some how 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

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

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

$$2^{0.2075\ldots n(1+o(1))} \leq \tau(n) \leq 2^{0.401n(1+o(1))}$$

are due to Wyner, Kabatianski and Levenshtein; see [**CS99**] 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: al 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.

$$|\Lambda| = \min\left\{\|\boldsymbol{x}\| : \boldsymbol{x} \in \Lambda \setminus \{\mathbf{0}\}\right\}.$$

The *set of minimal vectors* of Λ is then defined as

$$S(\Lambda) = \{\boldsymbol{x} \in \Lambda : \|\boldsymbol{x}\| = |\Lambda|\}.$$

These minimal vectors are the centers of spheres of radius $|\Lambda|/2$ in the sphere packing associated to Λ which touch the ball centered at the origin. Hence the number of these vectors, $|S(\Lambda)|$ is precisely the kissing number of Λ. 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 $U \in \mathrm{GL}_n(\mathbb{R})$ is called *orthogonal* if $U^{-1} = U^\top$, and the subset of all such matrices in $\mathrm{GL}_n(\mathbb{R})$ is

$$\mathcal{O}_n(\mathbb{R}) = \{U \in \mathrm{GL}_n(\mathbb{R}) : U^{-1} = U^\top\}.$$

This is a subgroup of $\mathrm{GL}_n(\mathbb{R})$ (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 $U \in \mathcal{O}_n(\mathbb{R})$ such that $L = \alpha UM$. 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.

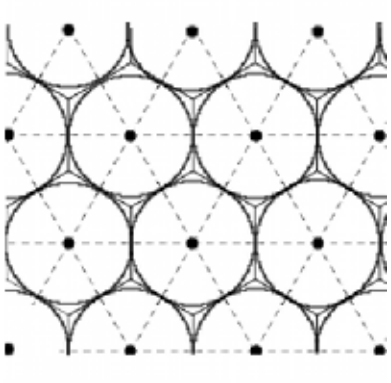


FIGURE 1. Hexagonal lattice with Voronoi cell translates and associated circle packing

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 Λ_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 any only if $L \sim \Lambda_h$.

This result was first obtain 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 Λ be a lattice in $\mathbb{R}^2$ with successive minima $\lambda_1 \leq \lambda_2$ and let $\boldsymbol{x}_1, \boldsymbol{x}_2$ be the vectors in Λ corresponding to λ_1, λ_2, respectively. Then $\boldsymbol{x}_1, \boldsymbol{x}_2$ form a basis for Λ.*

PROOF. Let $\boldsymbol{y}_1 \in \Lambda$ be a shortest vector extendable to a basis in Λ, and let $\boldsymbol{y}_2 \in \Lambda$ be a shortest vector such that $\boldsymbol{y}_1, \boldsymbol{y}_2$ is a basis of Λ. By picking $\pm\boldsymbol{y}_1, \pm\boldsymbol{y}_2$ if necessary we can ensure that the angle between these vectors is no greater than $\pi/2$. Then

$$0 < \|\boldsymbol{y}_1\| \leq \|\boldsymbol{y}_2\|,$$

and for any vector $\boldsymbol{z} \in \Lambda$ with $\|\boldsymbol{z}\| < \|\boldsymbol{y}_2\|$ the pair $\boldsymbol{y}_1, \boldsymbol{z}$ is *not* a basis for Λ. Since $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \Lambda$, there must exist integers a_1, a_2, b_1, b_2 such that

$$(\boldsymbol{x}_1\ \boldsymbol{x}_2) = (\boldsymbol{y}_1\ \boldsymbol{y}_2) \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix}. \tag{2.5}$$

Let θ_x be the angle between $\boldsymbol{x}_1, \boldsymbol{x}_2$, and θ_y be the angle between $\boldsymbol{y}_1, \boldsymbol{y}_2$, then $\pi/3 \leq \theta_x \leq \pi/2$ by Problem 2.6. Moreover, $\pi/3 \leq \theta_y \leq \pi/2$: indeed, suppose

$\theta_y < \pi/3$, then by Problem 2.5,

$$\|\boldsymbol{y}_1 - \boldsymbol{y}_2\| < \|\boldsymbol{y}_2\|,$$

however $\boldsymbol{y}_1, \boldsymbol{y}_1 - \boldsymbol{y}_2$ is a basis for Λ since $\boldsymbol{y}_1, \boldsymbol{y}_2$ is; this contradicts the choice of $\boldsymbol{y}_2$. Define

$$\mathcal{D} = \left| \det \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix} \right|,$$

then $\mathcal{D}$ is a positive integer, and taking determinants of both sides of (2.5), we obtain

$$\|\boldsymbol{x}_1\|\|\boldsymbol{x}_2\| \sin\theta_x = \mathcal{D}\|\boldsymbol{y}_1\|\|\boldsymbol{y}_2\| \sin\theta_y. \tag{2.6}$$

Notice that by definition of successive minima, $\|\boldsymbol{x}_1\|\|\boldsymbol{x}_2\| \leq \|\boldsymbol{y}_1\|\|\boldsymbol{y}_2\|$, and hence (2.6) implies that

$$\mathcal{D} = \frac{\|\boldsymbol{x}_1\|\|\boldsymbol{x}_2\|}{\|\boldsymbol{y}_1\|\|\boldsymbol{y}_2\|} \frac{\sin\theta_x}{\sin\theta_y} \leq \frac{2}{\sqrt{3}} < 2,$$

meaning that $\mathcal{D} = 1$. Combining this observation with (2.5), we see that

$$(\boldsymbol{x}_1\ \boldsymbol{x}_2) \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix}^{-1} = (\boldsymbol{y}_1\ \boldsymbol{y}_2),$$

where the matrix $\begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix}^{-1}$ has integer entries. Therefore $\boldsymbol{x}_1, \boldsymbol{x}_2$ is also a basis for Λ, completing the proof. □

As we know from Remark 1.5.1 in Section 1.5, 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 Δ attains its maximum in $\mathbb{R}^2$ on the set of well-rounded lattices.

LEMMA 2.2.3. *Let Λ and Ω 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 $\boldsymbol{x}_1, \boldsymbol{x}_2$ and $\boldsymbol{y}_1, \boldsymbol{y}_2$ be vectors in Λ and Ω, respectively, corresponding to successive minima. Suppose that $\boldsymbol{x}_1 = \boldsymbol{y}_1$, and angles between the vectors $\boldsymbol{x}_1, \boldsymbol{x}_2$ and $\boldsymbol{y}_1, \boldsymbol{y}_2$ are equal, call this common value θ. Suppose also that*

$$\lambda_1(\Lambda) = \lambda_2(\Lambda).$$

Then

$$\Delta(\Lambda) \geq \Delta(\Omega).$$

PROOF. By Lemma 2.2.2, $\boldsymbol{x}_1, \boldsymbol{x}_2$ and $\boldsymbol{y}_1, \boldsymbol{y}_2$ are minimal bases for Λ and Ω, respectively. Notice that

$$\begin{aligned} \lambda_1(\Lambda) &= \lambda_2(\Lambda) = \|\boldsymbol{x}_1\| = \|\boldsymbol{x}_2\| \\ &= \|\boldsymbol{y}_1\| = \lambda_1(\Omega) \leq \|\boldsymbol{y}_2\| = \lambda_2(\Omega). \end{aligned}$$

Then

$$\begin{aligned} \Delta(\Lambda) &= \frac{\pi\lambda_1(\Lambda)^2}{4\det(\Lambda)} = \frac{\lambda_1(\Lambda)^2\pi}{4\|\boldsymbol{x}_1\|\|\boldsymbol{x}_2\|\sin\theta} = \frac{\pi}{4\sin\theta} \\ &\geq \frac{\lambda_1(\Omega)^2\pi}{4\|\boldsymbol{y}_1\|\|\boldsymbol{y}_2\|\sin\theta} = \frac{\lambda_1(\Omega)^2\pi}{4\det(\Omega)} = \Delta(\Omega). \end{aligned} \tag{2.7}$$

□

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 $\boldsymbol{x}_1, \boldsymbol{x}_2$ be a basis for Λ such that*

$$\|\boldsymbol{x}_1\| = \|\boldsymbol{x}_2\|,$$

and the angle θ between these vectors lies in the interval $[\pi/3, \pi/2]$. Then $\boldsymbol{x}_1, \boldsymbol{x}_2$ is a minimal basis for Λ. In particular, this implies that Λ is WR.

PROOF. Let $\boldsymbol{z} \in \Lambda$, then $\boldsymbol{z} = a\boldsymbol{x}_1 + b\boldsymbol{x}_2$ for some $a, b \in \mathbb{Z}$. Then

$$\|\boldsymbol{z}\|^2 = a^2\|\boldsymbol{x}_1\|^2 + b^2\|\boldsymbol{x}_2\|^2 + 2ab\boldsymbol{x}_1^\top \boldsymbol{x}_2 = (a^2 + b^2 + 2ab\cos\theta)\|\boldsymbol{x}_1\|^2.$$

If $ab \geq 0$, then clearly $\|\boldsymbol{z}\|^2 \geq \|\boldsymbol{x}_1\|^2$. Now suppose $ab < 0$, then again

$$\|\boldsymbol{z}\|^2 \geq (a^2 + b^2 - |ab|)\|\boldsymbol{x}_1\|^2 \geq \|\boldsymbol{x}_1\|^2,$$

since $\cos\theta \leq 1/2$. Therefore $\boldsymbol{x}_1, \boldsymbol{x}_2$ are shortest nonzero vectors in Λ, hence they correspond to successive minima, and so form a minimal basis. Thus Λ is WR, and this completes the proof. □

LEMMA 2.2.5. *Let Λ be a lattice in $\mathbb{R}^2$ with successive minima λ_1, λ_2 and corresponding basis vectors $\boldsymbol{x}_1, \boldsymbol{x}_2$, respectively. Then the lattice*

$$\Lambda_{\mathrm{WR}} = \left(\boldsymbol{x}_1 \ \frac{\lambda_1}{\lambda_2}\boldsymbol{x}_2\right)\mathbb{Z}^2$$

is WR with successive minima equal to λ_1.

PROOF. By Problem 2.6, the angle θ between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is in the interval $[\pi/3, \pi/2]$, and clearly this is the same as the angle between the vectors $\boldsymbol{x}_1$ and $\frac{\lambda_1}{\lambda_2}\boldsymbol{x}_2$. Then by Lemma 2.2.4, Λ_{WR} is WR with successive minima equal to λ_1. □

Now combining Lemma 2.2.3 with Lemma 2.2.5 implies that

$$\Delta(\Lambda_{\mathrm{WR}}) \geq \Delta(\Lambda) \tag{2.8}$$

for any lattice $\Lambda \subset \mathbb{R}^2$, and (2.7) readily implies that the equality in (2.8) occurs if and only if $\Lambda = \Lambda_{\mathrm{WR}}$, which happens if and only if Λ 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 Λ in $\mathbb{R}^2$, (2.7) implies:

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

meaning that $\sin\theta$ is an invariant of Λ, 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 θ is in the interval $[\pi/3, \pi/2]$, it is also an invariant of the lattice, and we call it the *angle of* Λ, denoted by $\theta(\Lambda)$.

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

PROOF. First suppose that Λ and Ω are similar. Let $\boldsymbol{x}_1, \boldsymbol{x}_2$ be the minimal basis for Λ. There exist a real constant α and a real orthogonal 2×2 matrix U such that $\Omega = \alpha U\Lambda$. Let $\boldsymbol{y}_1, \boldsymbol{y}_2$ be a basis for Ω such that

$$(\boldsymbol{y}_1 \ \boldsymbol{y}_2) = \alpha U(\boldsymbol{x}_1 \ \boldsymbol{x}_2).$$

Then $\|\boldsymbol{y}_1\| = \|\boldsymbol{y}_2\|$, and the angle between $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ is $\theta(\Lambda) \in [\pi/3, \pi/2]$. By Lemma 2.2.4 it follows that $\boldsymbol{y}_1, \boldsymbol{y}_2$ is a minimal basis for Ω, and so Ω is WR and $\theta(\Omega) = \theta(\Lambda)$.

Next assume that Ω is WR and $\theta(\Omega) = \theta(\Lambda)$. Let $\lambda(\Lambda)$ and $\lambda(\Omega)$ be the respective values of successive minima of Λ and Ω. Let $\boldsymbol{x}_1, \boldsymbol{x}_2$ and $\boldsymbol{y}_1, \boldsymbol{y}_2$ be the minimal bases for Λ and Ω, respectively. Define

$$\boldsymbol{z}_1 = \frac{\lambda(\Lambda)}{\lambda(\Omega)}\boldsymbol{y}_1, \ \boldsymbol{z}_2 = \frac{\lambda(\Lambda)}{\lambda(\Omega)}\boldsymbol{y}_2.$$

Then $\boldsymbol{x}_1, \boldsymbol{x}_2$ and $\boldsymbol{z}_1, \boldsymbol{z}_2$ 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 2×2 real orthogonal matrix U such that

$$(\boldsymbol{y}_1 \ \boldsymbol{y}_2) = \frac{\lambda(\Lambda)}{\lambda(\Omega)}(\boldsymbol{z}_1 \ \boldsymbol{z}_2) = \frac{\lambda(\Lambda)}{\lambda(\Omega)}U(\boldsymbol{x}_1 \ \boldsymbol{x}_2),$$

and so Λ and Ω are similar lattices. This completes the proof. □

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 Λ, and (2.7) implies that

$$\Delta(\Lambda) = \frac{\pi}{4\sin\theta(\Lambda)}, \tag{2.9}$$

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 Λ is the hexagonal lattice

$$\Lambda_h = \begin{pmatrix} 1 & \frac{1}{2} \\ 0 & \frac{\sqrt{3}}{2} \end{pmatrix} \mathbb{Z}^2,$$

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 Λ, $\Delta(\Lambda) = \Delta(\Lambda_h)$, then by (2.8) and a short argument after it Λ 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 Λ is similar to Λ_h by Lemma 2.2.6. This completes the proof. □

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 Λ in $\mathbb{R}^n$ with minimal vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_m$ is called *eutactic* if there exist positive real numbers $c_1, \dots, c_m$ such that

$$\|\boldsymbol{v}\|^2 = \sum_{i=1}^{m} c_i (\boldsymbol{v}^\top \boldsymbol{x}_i)^2$$

for every vector $\boldsymbol{v} \in \operatorname{span}_{\mathbb{R}} \Lambda$. If $c_1 = \cdots = c_n$, Λ is called *strongly eutactic*. A lattice is called *perfect* if the set of symmetric matrices

$$\{\boldsymbol{x}_i \boldsymbol{x}_i^\top : 1 \le i \le m\}$$

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 $\Lambda = A\mathbb{Z}^n$ is a lattice with basis matrix A, then, as we know, B is another basis matrix for Λ if and only if $B = AU$ for some $U \in \mathrm{GL}_n(\mathbb{Z})$. In this way, the space of full-rank lattices in $\mathbb{R}^n$ can be identified with the set of orbits of $\mathrm{GL}_n(\mathbb{R})$ under the action by $\mathrm{GL}_n(\mathbb{Z})$ by right multiplication. The packing density Δ 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 (such lattices are called *extreme*). 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 (see [**Bac18**] for more general estimates on the number of perfect lattices in a given dimension). Explicit constructions of lattices with good properties, such as perfection or eutaxy often come from various algebraic and combinatorial settings. We refer the reader to the classical books [**CS99**], [**Mar03**], [**TV91**] for some standard constructions, as well as the more recent ones detailed in [**BFE+19**], [**Lad19**], [**BFG+16**], [**BF17**], [**FNPX19**].

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 know 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 $\mathcal{P}$ 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 $A \in \mathrm{GL}_n(\mathbb{R})$ such that $\Lambda = A\mathbb{Z}^n$. There are several questions that can be asked about this setup. We formulate them in algorithmic form.

Shortest Vector Problem (SVP).
Input: A matrix $A \in \mathrm{GL}_n(\mathbb{R})$.
Output: A vector $\boldsymbol{x}_1 \in \Lambda = A\mathbb{Z}^n$ such that $\|\boldsymbol{x}_1\| = \lambda_1$.

Shortest Independent Vector Problem (SIVP).
Input: A matrix $A \in \mathrm{GL}_n(\mathbb{R})$.
Output: Linearly independent vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \Lambda = A\mathbb{Z}^n$ such that

$$\|\boldsymbol{x}_i\| = \lambda_1 \ \forall \ 1 \leq i \leq n.$$

Closest Vector Problem (CVP).
Input: A matrix $A \in \mathrm{GL}_n(\mathbb{R})$ and a vector $\boldsymbol{y} \in \mathbb{R}^n$.
Output: A vector $\boldsymbol{x} \in \Lambda = A\mathbb{Z}^n$ such that

$$\|\boldsymbol{x} - \boldsymbol{y}\| \leq \|\boldsymbol{z} - \boldsymbol{y}\| \ \forall \ \boldsymbol{z} \in \Lambda.$$

Shortest Basis Problem (SBP).
Input: A matrix $A \in \mathrm{GL}_n(\mathbb{R})$.
Output: A basis $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ for $\Lambda = \mathbb{Z}^n$ such that $\|\boldsymbol{b}_i\| =$

$$\min\{\|\boldsymbol{x}\| : \boldsymbol{x} \in \Lambda \text{ is such that } \boldsymbol{b}_1, \ldots, \boldsymbol{b}_{i-1}, \boldsymbol{x} \text{ is extendable to a basis}\}$$

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

Notice that SVP is a special case of CVP where the input vector $\boldsymbol{y}$ is taken to be $\mathbf{0}$: indeed, a vector corresponding to the first successive minimum is precisely a vector that is closer to the origin than any other point of Λ. 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 γ-approximate versions of these problems for some approximation factor γ. 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 γ. For instance, the γ-SVP would ask for a vector $\boldsymbol{x} \in \Lambda$ such that

$$\|\boldsymbol{x}\| \leq \gamma\lambda_1.$$

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

On the other hand, γ-approximate versions of these problems for γ 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 Λ, produces a basis $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ for Λ such that

$$\min_{1 \leq i \leq n} \|\boldsymbol{b}_i\| \leq 2^{\frac{n-1}{2}} \lambda_1,$$

and

$$\prod_{i=1}^{n} \|\boldsymbol{b}_i\| \leq 2^{\frac{n(n-1)}{4}} \det(\Lambda). \tag{2.10}$$

We can compare this to the upper bound given by Minkowski's Successive Minima Theorem (Theorem 1.5.2):

$$\prod_{i=1}^{n} \lambda_i \leq \frac{2^n}{\omega_n} \det(\Lambda). \tag{2.11}$$

For instance, when $n = 2k$ the bound (2.10) gives

$$\prod_{i=1}^{n} \|\boldsymbol{b}_i\| \leq 2^{\frac{k(2k-1)}{2}} \det(\Lambda),$$

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 $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ is such a basis, then for any $a_1, \ldots, a_n \in \mathbb{Z}$,

$$\left\| \sum_{i=1}^{n} a_i \boldsymbol{u}_i \right\|^2 = \sum_{i=1}^{n} a_i^2 \|\boldsymbol{u}_i\|^2,$$

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 $\boldsymbol{b}_1, \boldsymbol{b}_2$ for a planar lattice Λ consists of vectors corresponding to successive minima λ_1, λ_2 of Λ, respectively, if and only if

$$\mu := \frac{\boldsymbol{b}_1^\top \boldsymbol{b}_2}{\|\boldsymbol{b}_1\|^2} \leq \frac{1}{2}.$$

On the other hand, if $|\mu| > 1/2$, then replacing $\boldsymbol{b}_2$ with

$$\boldsymbol{b}_2 - \lfloor \mu \rceil \, \boldsymbol{b}_1,$$

where $\lfloor \mu \rceil$ stands for the nearest integer to μ, 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:

Input: $\boldsymbol{b}_1, \boldsymbol{b}_2 \in \mathbb{R}^2$ such that $\|\boldsymbol{b}_1\| \leq \|\boldsymbol{b}_2\|$

Compute μ: $\mu = \frac{\boldsymbol{b}_1^\top \boldsymbol{b}_2}{\|\boldsymbol{b}_1\|^2}$

Check μ: if $|\mu| \leq 1/2$, output $\boldsymbol{b}_1, \boldsymbol{b}_2$; else set $\boldsymbol{b}_2 \leftarrow \boldsymbol{b}_2 - \lfloor \mu \rceil \, \boldsymbol{b}_1$ and repeat the algorithm (swapping $\boldsymbol{b}_1, \boldsymbol{b}_2$, if necessary, to ensure $\|\boldsymbol{b}_1\| \leq \|\boldsymbol{b}_2\|$)

This algorithm terminates in a finite number of steps (Problem 2.10).

Let us demonstrate this algorithm on an example. Suppose $\Lambda = \mathrm{span}_{\mathbb{Z}}\{\boldsymbol{b}_1, \boldsymbol{b}_2\}$, where

$$\boldsymbol{b}_1 = \begin{pmatrix} 1 \\ 5 \end{pmatrix}, \ \boldsymbol{b}_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

We notice that $\|\boldsymbol{b}_1\| > \|\boldsymbol{b}_2\|$, so we swap the vectors: $\boldsymbol{b}_1 \leftrightarrow \boldsymbol{b}_2$. We then compute

$$\mu = \frac{\boldsymbol{b}_1^\top \boldsymbol{b}_2}{\|\boldsymbol{b}_1\|^2} = 1 > 1/2.$$

The nearest integer to μ is 1, so we set

$$\boldsymbol{b}_2 \leftarrow \boldsymbol{b}_2 - \boldsymbol{b}_1 = \begin{pmatrix} 0 \\ 5 \end{pmatrix}.$$

We still have $\|\boldsymbol{b}_1\| < \|\boldsymbol{b}_2\|$, so no need to swap the vectors. With the new basis $\boldsymbol{b}_1, \boldsymbol{b}_2$ we again compute μ, which is now equal to $0 < 1/2$. Hence we found a shortest basis for Λ:

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \ \begin{pmatrix} 0 \\ 5 \end{pmatrix}.$$

LLL is based on a generalization of this idea. We can start with a basis $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ for a lattice Λ in $\mathbb{R}^n$ and use the Gram-Schmidt orthogonalization procedure to compute a corresponding orthogonal (but not normalized) basis $\boldsymbol{b}'_1, \ldots, \boldsymbol{b}'_n$ for $\mathbb{R}^n$. For any pair of indices i, j with $1 \le j < i \le i$, let us compute the Gram-Schmidt coefficient

$$\mu_{ij} = \frac{\boldsymbol{b}_i^\top \boldsymbol{b}'_j}{\|\boldsymbol{b}'_j\|^2}.$$

If this coefficient is $> 1/2$ in absolute value, we swap $\boldsymbol{b}_i \leftarrow \boldsymbol{b}_i - \lfloor \mu \rceil \, \boldsymbol{b}_j$: this ensures the length reduction, but one other condition is also needed. Formally speaking, a resulting basis $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ is called LLL reduced if the following two conditions are satisfied:

(1) For all $1 \le j < i \le n$, $|\mu_{ij}| \le 1/2$
(2) For some parameter $\delta \in [1/4, 1)$, for all $1 \le k \le n$,

$$\delta \|\boldsymbol{b}'_{k-1}\|^2 \le \|\boldsymbol{b}'_k\| + \mu_{k,(k-1)}^2 \|\boldsymbol{b}'_{k-1}\|^2.$$

Traditionally, δ 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* $\mathcal{O}_n(\mathbb{R})$ *is a subgroup of* $\mathrm{GL}_n(\mathbb{R})$.

PROBLEM 2.5. *Let* $\boldsymbol{x}_1$ *and* $\boldsymbol{x}_2$ *be nonzero vectors in* $\mathbb{R}^2$ *so that the angle* θ *between them satisfies* $0 < \theta < \frac{\pi}{3}$. *Prove that*

$$\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| < \max\{\|\boldsymbol{x}_1\|, \|\boldsymbol{x}_2\|\}.$$

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* $\boldsymbol{x}_1, \boldsymbol{x}_2$ *be the vectors in* Λ *corresponding to* λ_1, λ_2, *respectively. Let* $\theta \in [0, \pi/2]$ *be the angle between* $\boldsymbol{x}_1$ *and* $\boldsymbol{x}_2$. *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* Λ_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* $\boldsymbol{b}_1, \boldsymbol{b}_2$ *for a planar lattice* Λ *consists of vectors corresponding to successive minima* λ_1, λ_2, *respectively, if and only if*

$$\mu := \frac{\boldsymbol{b}_1^\top \boldsymbol{b}_2}{\|\boldsymbol{b}_1\|^2} \leq \frac{1}{2}.$$

On the other hand, if $|\mu| > 1/2$, *then replacing* $\boldsymbol{b}_2$ *with*

$$\boldsymbol{b}_2 - \lfloor \mu \rceil \, \boldsymbol{b}_1,$$

where $\lfloor \mu \rceil$ *stands for the nearest integer to* μ, *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.*

PROBLEM 2.11. *Let*

$$\Lambda = \begin{pmatrix} 1 & 1 \\ 2 & 3 \end{pmatrix} \mathbb{Z}^2, \ \Omega = \begin{pmatrix} 1 & -2 \\ 3 & 1 \end{pmatrix} \mathbb{Z}^2$$

be two full rank lattice in the plane. Which one of them has higher packing density? Prove your answer.

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

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

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 $B(\boldsymbol{X}, \boldsymbol{Y})$ can also be written as

$$B(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{X}^\top \mathcal{B} \boldsymbol{Y},$$

where

$$\mathcal{B} = \begin{pmatrix} b_{11} & b_{12} & \ldots & b_{1n} \\ b_{12} & b_{22} & \ldots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{1n} & b_{2n} & \ldots & b_{nn} \end{pmatrix},$$

is the corresponding $n \times n$ symmetric coefficient matrix, called the *Gram matrix* of the form, and

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

are the variable vectors. Hence symmetric bilinear forms are in bijective correspondence with symmetric $n \times n$ matrices. It is also easy to notice that

$$B(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{X}^\top \mathcal{B} \boldsymbol{Y} = (\boldsymbol{X}^\top \mathcal{B} \boldsymbol{Y})^\top = \boldsymbol{Y}^\top \mathcal{B}^\top \boldsymbol{X} = \boldsymbol{Y}^\top \mathcal{B} \boldsymbol{X} = B(\boldsymbol{Y}, \boldsymbol{X}), \tag{3.1}$$

since $\mathcal{B}$ is symmetric. We can also define the corresponding quadratic form

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

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).

DEFINITION 3.1.1. 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 $\mathcal{B}$, and will denote it by $\det(B)$ or $\det(Q)$.

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

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

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

$$\lambda(\boldsymbol{x}^\top\overline{\boldsymbol{x}}) = (\lambda\boldsymbol{x})^\top\overline{\boldsymbol{x}} = (\mathcal{B}\boldsymbol{x})^\top\overline{\boldsymbol{x}} = \boldsymbol{x}^\top\mathcal{B}\overline{\boldsymbol{x}} = \boldsymbol{x}^\top(\overline{\lambda}\overline{\boldsymbol{x}}) = \overline{\lambda}(\boldsymbol{x}^\top\overline{\boldsymbol{x}}),$$

meaning that $\lambda = \overline{\lambda}$, since $\boldsymbol{x}^\top\overline{\boldsymbol{x}} \neq 0$. 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 eigenvectors as well. In fact, even more is true.

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

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 $\boldsymbol{x}_1$ be an eigenvector of $\mathcal{B}$ with the corresponding eigenvalue λ_1. We can assume that $\|\boldsymbol{x}_1\| = 1$. Use Gram-Schmidt orthogonalization process to extend $\boldsymbol{x}_1$ to an orthonormal basis for $\mathbb{R}^n$, and write U for the corresponding basis matrix such that $\boldsymbol{x}_1$ is the first column. Then it is easy to notice that $U^{-1} = U^\top$. By Problem 3.2,

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

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

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

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 $\boldsymbol{y}_2, \dots, \boldsymbol{y}_n$. For each $2 \leq i \leq n$, define

$$\boldsymbol{y}_i' = \begin{pmatrix} 0 \\ \boldsymbol{y}_i \end{pmatrix} \in \mathbb{R}^n,$$

and let $\boldsymbol{x}_i = U\boldsymbol{y}_i'$. There exist $\lambda_2, \dots, \lambda_n$ such that $A\boldsymbol{y}_i = \lambda_i \boldsymbol{y}_i$ for each $2 \leq i \leq n$, hence

$$U^\top \mathcal{B} U \boldsymbol{y}_i' = \lambda_i \boldsymbol{y}_i',$$

and so $\mathcal{B}\boldsymbol{x}_i = \lambda_i \boldsymbol{x}_i$. Moreover, for each $2 \leq i \leq n$,

$$\boldsymbol{x}_1^\top \boldsymbol{x}_i = (\boldsymbol{x}_1^\top U) \begin{pmatrix} 0 \\ \boldsymbol{y}_i \end{pmatrix} = 0,$$

by construction of U. Finally notice that for each $2 \leq i \leq n$,

$$\|\boldsymbol{x}_i\| = \left(U \begin{pmatrix} 0 \\ \boldsymbol{y}_i \end{pmatrix} \right)^\top U \begin{pmatrix} 0 \\ \boldsymbol{y}_i \end{pmatrix} = (0, \boldsymbol{y}_i^\top) U^\top U \begin{pmatrix} 0 \\ \boldsymbol{y}_i \end{pmatrix} = \|\boldsymbol{y}_i\| = 1,$$

meaning that $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$ 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 $\mathcal{B}$ is diagonalizable by an orthogonal matrix, i.e. there exists a matrix $U \in \mathcal{O}_n(\mathbb{R})$ such that $U^\top \mathcal{B} U$ is a diagonal matrix.*

Proof. By Lemma 3.1.2, we can pick an orthonormal basis $\boldsymbol{u}_1, \dots, \boldsymbol{u}_n$ for $\mathbb{R}^n$ consisting of eigenvectors of $\mathcal{B}$. Let

$$U = (\boldsymbol{u}_1 \ \dots \ \boldsymbol{u}_n),$$

be the corresponding orthogonal matrix. Then for each $1 \leq i \leq n$,

$$\boldsymbol{u}_i^\top \mathcal{B} \boldsymbol{u}_i = \boldsymbol{u}_i^\top (\lambda_i \boldsymbol{u}_i) = \lambda_i (\boldsymbol{u}_i^\top \boldsymbol{u}_i) = \lambda_i,$$

where λ_i is the corresponding eigenvalue, since

$$1 = \|\boldsymbol{u}_i\|^2 = \boldsymbol{u}_i^\top \boldsymbol{u}_i.$$

Also, for each $1 \leq i \neq j \leq n$,

$$\boldsymbol{u}_i^\top \mathcal{B} \boldsymbol{u}_j = \boldsymbol{u}_i^\top (\lambda_j \boldsymbol{u}_j) = \lambda_j (\boldsymbol{u}_i^\top \boldsymbol{u}_j) = 0.$$

Therefore, $U^\top \mathcal{B} U$ is a diagonal matrix whose diagonal entries are precisely the eigenvalues of $\mathcal{B}$. □

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 $\mathrm{GL}_n(\mathbb{R})$ 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

$$B_1(\sigma \boldsymbol{x}, \sigma \boldsymbol{y}) = B_2(\boldsymbol{x}, \boldsymbol{y}),$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Their associated quadratic forms Q_1 and Q_2 are also said to be isomorphic in this case and σ 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 $\mathrm{GL}_n(\mathbb{R})$ (Problem 3.4): these are precisely the linear maps $\sigma : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$B(\sigma \boldsymbol{X}, \sigma \boldsymbol{Y}) = B(\boldsymbol{X}, \boldsymbol{Y}),$$

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

DEFINITION 3.1.3. A symmetric bilinear form B and its associated quadratic form Q are called *diagonal* if their coefficient matrix $\mathcal{B}$ is diagonal. In this case we can write

$$B(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{n} b_i X_i Y_i, \; Q(\boldsymbol{X}) = \sum_{i=1}^{n} b_i X_i^2,$$

where $b_1, \ldots, b_n$ are precisely the diagonal entries of the matrix $\mathcal{B}$.

With this notation we readily obtain the following result.

THEOREM 3.1.4. *Every real symmetric bilinear form, as well as its associated quadratic form, is isomorphic to a real diagonal form. In fact, there exists such an isomorphism σ whose matrix is in $\mathcal{O}_n(\mathbb{R})$: in this case we call σ an isometry.*

PROOF. This is an immediate consequence of Lemma 3.1.3. □

REMARK 3.1.5. Notice that this diagonalization is not unique, i.e. it is possible for a bilinear or quadratic form to be isomorphic to more than one diagonal form (notice that an isomorphism can come from the whole group $\mathrm{GL}_n(\mathbb{R})$, not necessarily from $\mathcal{O}_n(\mathbb{R})$). This procedure does however yield an invariant for nonsingular real quadratic forms, called signature.

DEFINITION 3.1.4. 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

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

where all coefficients b_{i_j}, b_{k_j} 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 $\mathcal{B}$ be the coefficient matrix of Q, and let U, W be two different matrices that diagonalize $\mathcal{B}$ with column vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ and $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$, respectively, arranged in such a way that

$$Q(\boldsymbol{u}_1), \ldots, Q(\boldsymbol{u}_{r_1}) > 0,\ Q(\boldsymbol{u}_{r_1+1}), \ldots, Q(\boldsymbol{u}_n) < 0,$$

and

$$Q(\boldsymbol{w}_1), \ldots, Q(\boldsymbol{w}_{r_2}) > 0,\ Q(\boldsymbol{w}_{r_2+1}), \ldots, Q(\boldsymbol{w}_n) < 0,$$

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

$$V_1^+ = \operatorname{span}_{\mathbb{R}}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{r_1}\},\ V_1^- = \operatorname{span}_{\mathbb{R}}\{\boldsymbol{u}_{r_1+1}, \ldots, \boldsymbol{u}_n\},$$

and

$$V_2^+ = \operatorname{span}_{\mathbb{R}}\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{r_2}\},\ V_2^- = \operatorname{span}_{\mathbb{R}}\{\boldsymbol{w}_{r_2+1}, \ldots, \boldsymbol{w}_n\}.$$

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

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

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. □

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, $Q(\boldsymbol{x}) > 0$, or $Q(\boldsymbol{x}) < 0$ for each $\boldsymbol{0} \neq \boldsymbol{x} \in \mathbb{R}^n$; Q is called *positive* or *negative semi-definite* if, respectively, $Q(\boldsymbol{x}) \geq 0$, or $Q(\boldsymbol{x}) \leq 0$ for each $\boldsymbol{0} \neq \boldsymbol{x} \in \mathbb{R}^n$. 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 Λ be a lattice of full rank in $\mathbb{R}^n$, and let A be a basis matrix for Λ. Then $\boldsymbol{y} \in \Lambda$ if and only if $\boldsymbol{y} = A\boldsymbol{x}$ for some $\boldsymbol{x} \in \mathbb{Z}^n$. Notice that the Euclidean norm of $\boldsymbol{y}$ in this case is

$$\|\boldsymbol{y}\| = (A\boldsymbol{x})^\top (A\boldsymbol{x}) = \boldsymbol{x}^\top (A^\top A)\boldsymbol{x} = Q_A(\boldsymbol{x}),$$

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 Λ corresponding to the basis matrix A.

Now suppose C is another basis matrix for Λ. Then there must exist $U \in \mathrm{GL}_n(\mathbb{Z})$ 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 $\mathrm{GL}_n(\mathbb{Z})$-*congruent*. Notice in this case that for each $\boldsymbol{x} \in \mathbb{R}^n$

$$Q_C(\boldsymbol{x}) = \boldsymbol{x}^\top U^\top (A^\top A)U\boldsymbol{x} = Q_A(U\boldsymbol{x}),$$

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 $\mathrm{GL}_n(\mathbb{Z})$, we will call them *arithmetically equivalent*. We proved the following statement.

PROPOSITION 3.1.7. *All different norm forms of a lattice Λ 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 $\mathcal{B}$, then there exists $U \in \mathcal{O}_n(\mathbb{R})$ such that

$$U^\top \mathcal{B} U = \mathcal{D},$$

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

$$\mathcal{B} = (\sqrt{\mathcal{D}}U)^\top (\sqrt{\mathcal{D}}U).$$

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

$$(WA)^\top (WA) = A^\top (W^\top W)A = A^\top A = \mathcal{B}.$$

Therefore Q is a norm form for every lattice $WA\mathbb{Z}^n$, where $W \in \mathcal{O}_n(\mathbb{R})$. Let us call two lattices Λ_1 and Λ_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 Λ and a quadratic form Q correspond to each other as described in Theorem 3.1.8, then

$$\det(\Lambda) = \sqrt{|\det(Q)|}. \tag{3.2}$$

Now that we have the bijective correspondence between lattices and positive definite quadratic forms, we end this section with an application of Minkonwski's Convex Body Theorem to the context of quadratic forms: this is Theorem 4 on p. 44 of [**GL87**].

THEOREM 3.1.9. *Let*

$$Q(\boldsymbol{X}) = \sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}X_iX_j = \boldsymbol{X}^\top \mathcal{B}\boldsymbol{X}$$

be a positive definite quadratic form in n variables with Gram matrix $\mathcal{B}$. There exists $\mathbf{0} \neq \boldsymbol{x} \in \mathbb{Z}^n$ such that

$$Q(\boldsymbol{x}) \leq 4 \left(\frac{\det(B)}{\omega_n^2}\right)^{1/n}.$$

PROOF. As in the proof of Theorem 3.1.8 above, we can decompose B as $B = A^\top A$ for some $A \in \mathrm{GL}_n(\mathbb{R})$. Then

$$\det(B) = \det(A)^2.$$

For each $r \in \mathbb{R}_{>0}$, define the set

$$E_r = \{\boldsymbol{x} \in \mathbb{R}^n : Q(\boldsymbol{x}) \leq r\} = \{\boldsymbol{x} \in \mathbb{R}^n : (A\boldsymbol{x})^\top (A\boldsymbol{x}) \leq r\} = A^{-1}S_r,$$

where $S_r = \{\boldsymbol{y} \in \mathbb{R}^n : \|\boldsymbol{y}\|^2 \leq r\}$ 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

$$\mathrm{Vol}(E_r) = |\det(A)|^{-1}\,\mathrm{Vol}(S_r) = \omega_n\sqrt{\frac{r^n}{\det(B)}}.$$

Hence if

$$r = 4 \left(\frac{\det(B)}{\omega_n^2}\right)^{1/n},$$

then $\mathrm{Vol}(E_r) = 2^n$, and so by Theorem 1.4.2 there exists $\mathbf{0} \neq \boldsymbol{x} \in E_r \cap \mathbb{Z}^n$. □

3.2. Minkowski's reduction

Let $M \subseteq \mathbb{R}^n$ be a $\mathbf{0}$-symmetric convex set with positive volume, and let $\Lambda \subseteq \mathbb{R}^n$ be a lattice of full rank, as before. In Section 1.4 we discussed the following question: by how much should M be homogeneously expanded so that it contains n linearly independent points of Λ? We learned however that the resulting set of n minimal linearly independent vectors produced this way is not necessarily a basis for Λ. In this section we want to understand by how much should M be homogeneously expanded so that it contains a basis of Λ? 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.

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

then

$$F(\boldsymbol{x} + \boldsymbol{y}) \leq F(\boldsymbol{x}) + F(\boldsymbol{y}).$$

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

Definition 3.2.1. A basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ of Λ is said to be *Minkowski reduced with respect to* M if for each $1 \leq i \leq n$, $\boldsymbol{v}_i$ is such that

$$F(\boldsymbol{v}_i) = \min\{F(\boldsymbol{v}) : \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1}, \boldsymbol{v} \text{ is extendable to a basis of } \Lambda\}.$$

In the frequently occurring case when M is the closed unit ball $\mathbb{B}_n$ centered at $\mathbf{0}$, we will just say that a corresponding such basis is *Minkowski reduced*. Notice in particular that a Minkowski reduced basis contains a shortest nonzero vector in Λ.

From here on let $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ be a Minkowski reduced basis of Λ with respect to M. Then

$$F(\boldsymbol{v}_1) = \lambda_1, \ F(\boldsymbol{v}_i) \geq \lambda_i \ \forall \ 2 \leq i \leq n.$$

Assume first that $M = \mathbb{B}_n$, then $F = \| \ \|$. Write A for the corresponding basis matrix of Λ, i.e. $A = (\boldsymbol{v}_1 \ldots \boldsymbol{v}_n)$, and so $\Lambda = A\mathbb{Z}^n$. Let Q be the corresponding positive definite quadratic form, i.e. for each $\boldsymbol{x} \in \mathbb{R}^n$

$$Q(\boldsymbol{x}) = \boldsymbol{x}^\top A^\top A \boldsymbol{x}.$$

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

$$Q(\boldsymbol{e}_i) = \|\boldsymbol{v}_i\|^2.$$

Hence for each $1 \leq i \leq n$, $Q(\boldsymbol{e}_i) \leq Q(\boldsymbol{x})$ for all $\boldsymbol{x}$ such that

$$\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1}, A\boldsymbol{x}$$

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

$$Q(\boldsymbol{e}_i) \leq Q(\boldsymbol{x}) \ \forall \ \boldsymbol{x} \in \mathbb{Z}^n, \ \gcd(x_i, \ldots, x_n) = 1. \tag{3.3}$$

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 $M = \mathbb{B}_n$, but preserve the rest of notation as above. We can prove the following analogue of Minkowski's successive

minima theorem; this 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*

$$\lambda_i \leq F(\boldsymbol{v}_i) \leq \nu_i \lambda_i. \tag{3.4}$$

Moreover,

$$\prod_{i=1}^{n} F(\boldsymbol{v}_i) \leq 2^n \left(\frac{3}{2}\right)^{\frac{(n-1)(n-2)}{2}} \frac{\det(\Lambda)}{\mathrm{Vol}(M)}. \tag{3.5}$$

PROOF. It is easy to see that (3.5) follows immediately by combining (3.4) with Theorem 1.5.2, hence we only need to prove (3.4). It is obvious by definition of a reduced basis that $F(\boldsymbol{v}_i) \geq \lambda_i$ for each $1 \leq i \leq n$, and that $F(\boldsymbol{v}_1) = \lambda_1$. Hence we only need to prove that for each $2 \leq i \leq n$

$$F(\boldsymbol{v}_i) \leq \nu_i \lambda_i. \tag{3.6}$$

Let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ be the linearly independent vectors corresponding to successive minima $\lambda_1, \ldots, \lambda_n$, i.e.

$$F(\boldsymbol{u}_i) = \lambda_i, \ \forall\ 1 \leq i \leq n.$$

Then, by linear independence, for each $2 \leq i \leq n$ at least one of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_i$ does not belong to the subspace $\mathrm{span}_{\mathbb{R}}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1}\}$, call this vector $\boldsymbol{u}_j$. If the set $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1}, \boldsymbol{u}_j$ is extendable to a basis of Λ, then by construction of reduced basis we must have

$$\lambda_i \geq \lambda_j = F(\boldsymbol{u}_j) \geq F(\boldsymbol{v}_i),$$

and so it implies that $\lambda_i = F(\boldsymbol{v}_i)$, proving (3.6) in this case.

Next assume that the set $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1}, \boldsymbol{u}_j$ is not extendable to a basis of Λ. Let $\boldsymbol{v} \in \mathrm{span}_{\mathbb{R}}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1}, \boldsymbol{u}_j\}$ be such that the set $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1}, \boldsymbol{v}$ is extendable to a basis of Λ. Then we can write

$$\boldsymbol{u}_j = k_1 \boldsymbol{v}_1 + \cdots + k_{i-1} \boldsymbol{v}_{i-1} \pm m\boldsymbol{v},$$

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

$$\boldsymbol{u}_j \notin \mathrm{span}_{\mathbb{R}}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1}\}.$$

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

$$\boldsymbol{v} \in \mathrm{span}_{\mathbb{Z}}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1}, \boldsymbol{u}_j\},$$

which would imply that $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1}, \boldsymbol{u}_j$ is extendable to a basis. Thus $m \geq 2$, and we can write

$$\boldsymbol{v} = \alpha_1 \boldsymbol{v}_1 + \cdots + \alpha_{i-1} \boldsymbol{v}_{i-1} \pm \frac{1}{m} \boldsymbol{u}_j,$$

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 β_k with $|\beta_k| \leq \frac{1}{2}$ such that

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

Then

$$\boldsymbol{v} = \sum_{k=1}^{i-1} (l_k + \beta_k) \boldsymbol{v}_k \pm \frac{1}{m} \boldsymbol{u}_j = \sum_{k=1}^{i-1} l_k \boldsymbol{v}_k + \boldsymbol{v}',$$

where $\boldsymbol{v}' = \sum_{k=1}^{i-1} \beta_k \boldsymbol{v}_k \pm \frac{1}{m}\boldsymbol{u}_j$. Since $\boldsymbol{v} - \boldsymbol{v}' \in \mathrm{span}_{\mathbb{Z}}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1}\}$, it must be that $\boldsymbol{v}' \in \Lambda$, and the set $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1}, \boldsymbol{v}'$ is extendable to a basis of Λ. Then, by definition of $\boldsymbol{v}_i$, we have

$$\begin{aligned} F(\boldsymbol{v}_i) &\leq F(\boldsymbol{v}') \leq \sum_{k=1}^{i-1} F(\beta_k \boldsymbol{v}_k) + F\left(\frac{1}{m}\boldsymbol{u}_j\right) \\ &= \sum_{k=1}^{i-1} |\beta_k| F(\boldsymbol{v}_k) + \frac{1}{m} F(\boldsymbol{u}_j) \\ &\leq \frac{1}{2}\left(\sum_{k=1}^{i-1} F(\boldsymbol{v}_k) + F(\boldsymbol{u}_j)\right) \leq \frac{1}{2}\left(\sum_{k=1}^{i-1} F(\boldsymbol{v}_k) + \lambda_i\right). \end{aligned}$$

Combining this with the previous case, we conclude that

$$F(\boldsymbol{v}_i) \leq \max\left\{\lambda_i, \frac{1}{2}\left(\sum_{k=1}^{i-1} F(\boldsymbol{v}_k) + \lambda_i\right)\right\}, \ \forall\ 2 \leq i \leq n. \tag{3.7}$$

Hence we obtain

$$F(\boldsymbol{v}_2) \leq \max\left\{\lambda_2, \frac{1}{2}(\lambda_1 + \lambda_2)\right\} = \lambda_2,$$

hence $F(\boldsymbol{v}_2) = \lambda_2$. 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 Λ with respect to a convex $\mathbf{0}$-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 $Q(\boldsymbol{X})$ 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 $\boldsymbol{x} \in \mathbb{Z}^n$ such that $Q(\boldsymbol{x}) = m$? 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

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

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 $Q(\boldsymbol{X})$, call it Λ, so that $Q(\boldsymbol{X})$ is a norm form corresponding to some choice of basis matrix A of Λ. Then for any vector $\boldsymbol{y} = A\boldsymbol{x} \in \Lambda$,

$$\|\boldsymbol{y}\| = Q(\boldsymbol{x}),$$

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

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 $\boldsymbol{u} \in \Lambda$ is of the form

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

for some integers a, b. Then

$$\|\boldsymbol{u}\|^2 = a^2 + (am + bp)^2 = a^2(m^2 + 1) + (2abm + b^2 p)p \equiv 0 \pmod p,$$

hence $p \mid \|\boldsymbol{u}\|^2$ for any $\boldsymbol{u} \in \Lambda$. Let $\varepsilon > 0$ and $\mathbb{B}_2(\sqrt{2p-\varepsilon})$ be a circle of radius $\sqrt{2p-\varepsilon}$ centered at the origin in the plane. For sufficiently small ε, the area of $\mathbb{B}_2(\sqrt{2p-\varepsilon})$ is

$$\pi(2p-\varepsilon) > 2^2 p = 2^2 \det(\Lambda),$$

and hence $\mathbb{B}_2(\sqrt{2p-\varepsilon})$ contains a nonzero point of Λ, by Theorem 1.4.2. Let $\boldsymbol{u} = \begin{pmatrix} x \\ y \end{pmatrix}$ be this point, then

$$p \mid \|\boldsymbol{u}\|^2 = x^2 + y^2 \leq 2p - \varepsilon < 2p.$$

This implies that $x^2 + y^2 = p$, and so we are done. □

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

$$a = x^2 + y^2,\ b = z^2 + t^2$$

for some $x, y, z, t \in \mathbb{Z}$. Then

$$\begin{aligned} ab &= (x^2+y^2)(z^2+t^2) \\ &= (xz+yt)^2 + (xt-yz)^2 \\ &= (xz-yt)^2 + (xt+yz)^2. \end{aligned} \tag{3.9}$$

□

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

$$m = 2^e p_1^{f_1} \cdots p_k^{f_k} q_1^{g_1} \cdots q_n^{g_n} \tag{3.10}$$

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 $\equiv 3 \pmod 4$, $e \geq 0$ 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.

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

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

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

$$x_1^{p-1} = (x_1^2)^{2\ell+1} \equiv (-1)^{2\ell+1}(y_1^2)^{2\ell+1} = -y_1^{p-1} \pmod{p}. \tag{3.11}$$

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:

$$\begin{aligned} & (x_1^2 + x_2^2 + x_3^2 + x_4^2)(y_1^2 + y_2^2 + y_3^2 + y_4^2) \\ = \; & (x_1y_1 - x_2y_2 - x_3y_3 - x_4y_4)^2 + (x_1y_2 + x_2y_1 + x_3y_4 - x_4y_3)^2 \\ + \; & (x_1y_3 - x_2y_4 + x_3y_1 + x_4y_2)^2 + (x_1y_4 + x_2y_3 - x_3y_2 + x_4y_1)^2. \end{aligned} \tag{3.12}$$

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 $\boldsymbol{u} \in \Lambda$, then

$$\|\boldsymbol{u}\|^2 = (px_1 + ax_3 + bx_4)^2 + (px_2 + bx_3 - ax_4)^2 + x_3^2 + x_4^2$$

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

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

thus $p \mid \|\boldsymbol{u}\|^2$ for any $\boldsymbol{u} \in \Lambda$. Let $\varepsilon > 0$ and $\mathbb{B}_4(\sqrt{2p - \varepsilon})$ be a ball of radius $\sqrt{2p - \varepsilon}$ centered at the origin in $\mathbb{R}^4$. For sufficiently small ε, the volume of $\mathbb{B}_4(\sqrt{2p - \varepsilon})$ is

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

and hence $\mathbb{B}_4(\sqrt{2p - \varepsilon})$ contains a nonzero point of Λ, by Theorem 1.4.2. Let $\boldsymbol{u} = (u_1, u_2, u_3, u_4)^\top$ be this point, then

$$p \mid \|\boldsymbol{u}\|^2 = u_1^2 + u_2^2 + u_3^2 + u_4^2 \leq 2p - \varepsilon < 2p.$$

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 $Q(\boldsymbol{X})$ be a quadratic form in n variables. Prove that*

$$B(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{2}(Q(\boldsymbol{X}+\boldsymbol{Y}) - Q(\boldsymbol{X}) - Q(\boldsymbol{Y}))$$

is a symmetric bilinear form.

PROBLEM 3.2. *Let $\mathcal{B}$ be an $n \times n$ real symmetric matrix. Let $\boldsymbol{x}_1$ be an eigenvector of $\mathcal{B}$ with the corresponding eigenvalue λ_1 and $\|\boldsymbol{x}_1\| = 1$. Let $U \in \mathcal{O}_n(\mathbb{R})$ be a matrix whose columns are an orthonormal basis containing $\boldsymbol{x}_1$ with $\boldsymbol{x}_1$ being the first column. Prove that the matrix $U^\top \mathcal{B} U$ is of the form*

$$\begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & a_{11} & \dots & a_{1(n-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{(n-1)1} & \dots & a_{(n-1)(n-1)} \end{pmatrix},$$

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

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

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 σ 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 $\mathrm{GL}_n(\mathbb{R})$.*

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

(1) *B is nonsingular.*
(2) *For every $\boldsymbol{0} \neq \boldsymbol{x} \in \mathbb{R}^n$, there exists $\boldsymbol{y} \in \mathbb{R}^n$ so that $B(\boldsymbol{x},\boldsymbol{y}) \neq 0$.*
(3) *For every $\boldsymbol{0} \neq \boldsymbol{x} \in \mathbb{R}^n$ at least one of the partial derivatives*

$$\frac{\partial Q}{\partial X_i}(\boldsymbol{x}) \neq 0.$$

(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 isomorphic 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 $\boldsymbol{x} \mapsto \sqrt{Q(\boldsymbol{x})}$ 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

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

PROBLEM 3.11. *Let*

$$Q(\boldsymbol{X}) = \sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij} X_i X_j$$

be a Minkowski reduced positive definite quadratic form. Prove that

$$\prod_{i=1}^{n} b_{ii} \leq \frac{4^n}{\omega_n^2} \left(\frac{3}{2}\right)^{\frac{(n-1)(n-2)}{2}} \det(Q), \tag{3.13}$$

where ω_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 $M = \left\{\boldsymbol{x} \in \mathbb{R}^n : \sqrt{Q(\boldsymbol{x})} \leq 1\right\}$; 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.*

PROBLEM 3.14. *Let $n \geq 2$ be even, and define quadratic forms*

$$Q_1(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i^2 - X_1 X_2 - X_3 X_4 - \cdots - X_{n-1} X_n,$$

and

$$Q_2(X_1, \ldots, X_n) = \boldsymbol{X}^t B \boldsymbol{X}$$

for some real symmetric matrix B. Suppose that Q_1 and Q_2 are isometric, i.e. there exists a real orthogonal matrix U such that

$$Q_1(U\boldsymbol{X}) = Q_2(\boldsymbol{X}).$$

(1) *Find eigenvalues of B. Prove your answer.*

(2) *Let*

$$Q_1'(X_1, \ldots, X_n) = \frac{1}{2}\sum_{i=1}^{n} X_i^2 - X_1X_2 - X_3X_4 - \cdots - X_{n-1}X_n.$$

Are Q_1' and Q_2 isometric? Prove your answer.

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

$$(a, b) \sim (c, d) \Leftrightarrow ad = bc. \tag{4.1}$$

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 $a_0, a_1, \dots$ are integers in the interval $[-9, 9]$, either all nonnegative 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 $\{a_0, a_1, \dots\}$ 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*

$$x < z < y.$$

PROOF. Since $y - x > 0$, there must exist $n \in \mathbb{Z}$ such that

$$n(y - x) = ny - nx > 1,$$

so $nx + 1 < ny$. Let $m = [nx + 1]$, then m is an integer such that

$$m \leq nx + 1 < m + 1.$$

Hence we have

$$nx < m \leq nx + 1 < ny,$$

and hence

$$x < \frac{m}{n} < y.$$

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 α is called *algebraic* if there exists a nonzero polynomial $p(x)$ with integer coefficients such that $p(\alpha) = 0$. If α 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 α 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 $f(\alpha) = 0$, $\gcd(a_0, \dots, a_d) = 1$, 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 α, 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 ± 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 $S_1, S_2, \dots$ 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

$$S_n = \{x_{n1}, \dots, x_{na_n}\}.$$

Then we can write

$$S = \{x_{11}, \dots, x_{1a_1}, x_{21}, \dots, x_{2a_2}, \dots\}.$$

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

$$\begin{aligned} \mathbb{N} \times \mathbb{N} &= \{(m,n) : m,n \in \mathbb{N}\} \\ &= \{(m,n) : m,n \in \mathbb{N},\ m \leq n\} \cup \{(m,n) : m,n \in \mathbb{N},\ m > n\} \\ &= \left(\bigcup_{n=1}^{\infty} \{(m,n) : m \leq n\} \right) \cup \left(\bigcup_{m=1}^{\infty} \{(m,n) : n < m\} \right), \end{aligned}$$

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 $S_1, S_2, \dots$ be countable sets, say

$$S_n = \{x_{n1}, x_{n2}, \dots\},$$

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*

$$\mathbb{Z}^m := \{\boldsymbol{a} = (a_1, \dots, a_m) : a_1, \dots, a_m \in \mathbb{Z}\}$$

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

$$\mathbb{Z}^d = \left(\bigcup_{a \in \mathbb{N}_0} \{(\boldsymbol{x}, a) : \boldsymbol{x} \in \mathbb{Z}^{d-1}\} \right) \cup \left(\bigcup_{a \in \mathbb{N}} \{(\boldsymbol{x}, -a) : \boldsymbol{x} \in \mathbb{Z}^{d-1}\} \right),$$

Each set like $\{(\boldsymbol{x}, a) : \boldsymbol{x} \in \mathbb{Z}^{d-1}\}$ or $\{(\boldsymbol{x}, -a) : \boldsymbol{x} \in \mathbb{Z}^{d-1}\}$ 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 $\mathbb{Z}^2_* := \{(a,b) \in \mathbb{Z} : b \neq 0\}$ of $\mathbb{Z}^2$ 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 $\bigcup_{d\in\mathbb{N}_0} \mathbb{Z}^{d+1}$, 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 $I_2 \subsetneq I_1$. Continue in the same manner to obtain a nested sequence of intervals:

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

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

$$A := \lim_{n\to\infty} a_n,\ B := \lim_{n\to\infty} b_n.$$

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. □

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 α. 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 α is transcendental we would need to establish that α 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*

$$\left|\alpha - \frac{p}{q}\right| \leq \frac{1}{q(Q+1)}. \tag{4.2}$$

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

$$\left|\alpha - \frac{p}{q}\right| \leq \frac{1}{q^2}. \tag{4.3}$$

PROOF. If α is a rational number with denominator $\leq Q$, there is nothing to prove. Hence we will assume that either α 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. $\{x\} = x - [x]$. 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

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

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

$$|l\alpha - [l\alpha]| \leq \frac{1}{Q+1}, \tag{4.4}$$

or

$$|l\alpha - [l\alpha] - 1| \leq \frac{1}{Q+1}. \tag{4.5}$$

This means that there exists an integer $1 \leq l \leq Q$ and an integer m equal to either $[l\alpha]$ or $[l\alpha] - 1$, 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

$$|(m\alpha - [m\alpha]) - (l\alpha - [l\alpha])| = |(m-l)\alpha - ([m\alpha] - [l\alpha])| \leq \frac{1}{Q+1}.$$

Once again, let $d = \gcd((m-l), ([m\alpha] - [l\alpha]))$, and let $p = \frac{[m\alpha]-[l\alpha]}{d}$ 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 $q \leq Q$,

$$\left|\alpha - \frac{p}{q}\right| \leq \frac{1}{q(Q+1)} < \frac{1}{q^2}. \tag{4.6}$$

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 α 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 contradiciton. 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:

$$L_1(x,y) = x - \alpha y,\ L_2(x,y) = y.$$

Coefficients of these forms are given by the rows of the 2×2 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 $c_1 c_2 = 1$, then by Theorem 1.4.3 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.

$$|p - \alpha q| \leq \frac{1}{c_2},\ |q| \leq c_2.$$

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 $|p - \alpha q| \leq \frac{1}{Q+1}$ and $|q| \leq Q$. Dividing through by q completes the proof. □

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

$$\left|\alpha - \frac{p}{q}\right| \leq \frac{1}{\sqrt{5}\ q^2}. \tag{4.7}$$

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

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{Aq^2}. \tag{4.8}$$

Proof. We know that α is one of the roots of $f(x)$, then let β 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

$$0 \neq \left|f\left(\frac{p}{q}\right)\right| = \frac{|ap^2 + bpq + cq^2|}{q^2} \geq \frac{1}{q^2},$$

since $0 \neq ap^2 + bpq + cq^2 \in \mathbb{Z}$, hence $|ap^2 + bpq + cq^2| \geq 1$. Therefore

$$\begin{aligned} \frac{1}{q^2} &\leq \left| f\left(\frac{p}{q}\right)\right| = a\left|\alpha - \frac{p}{q}\right|\left|\beta - \frac{p}{q}\right| \\ &< \frac{a}{Aq^2}\left|\beta - \frac{p}{q}\right| = \frac{a}{Aq^2}\left|\left(\alpha - \frac{p}{q}\right) + (\beta - \alpha)\right| \\ &\leq \frac{a}{Aq^2}\left|\alpha - \frac{p}{q}\right| + \frac{a}{Aq^2}|\beta - \alpha| < \frac{a}{A^2q^4} + \frac{\sqrt{D}}{Aq^2}, \end{aligned}$$

and subtracting $\frac{\sqrt{D}}{Aq^2}$ from both sides of the above inequality implies

$$\frac{1}{q^2}\left(1 - \frac{\sqrt{D}}{A}\right) < \frac{a}{A^2q^4}.$$

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. □

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 α there exists a constant $C(\alpha) > 0$ such that for any $\frac{p}{q} \in \mathbb{Q}$

$$\left|\alpha - \frac{p}{q}\right| \geq \frac{C(\alpha)}{q^2}. \tag{4.9}$$

In other words, quadratic irrationals are badly approximable.

DEFINITION 4.4.1. An irrational number α 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 necessarilly badly approximable, are certainly "worse" approximable than transcendental numbers 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 α over $\mathbb{Q}$. Then there exists a positive real constant $C(\alpha)$ such that for any $\frac{p}{q} \in \mathbb{Q}$*

$$\left|\alpha - \frac{p}{q}\right| \geq \frac{C(\alpha)}{q^d}. \tag{4.10}$$

PROOF. Let

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

Then, since $d \geq 2$ means that α 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 $f(\alpha) = 0$,

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

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

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

$$p_k = a^{k!} \sum_{n=1}^{k} \frac{1}{a^{n!}},\ q_k = a^{k!} \in \mathbb{Z}.$$

Then

$$\left|\alpha - \frac{p_k}{q_k}\right| = \sum_{n=k+1}^{\infty} \frac{1}{a^{n!}} = \frac{1}{a^{(k+1)!}} \sum_{n=k+1}^{\infty} \frac{a^{(k+1)!}}{a^{n!}} < \frac{1}{a^{(k+1)!}} \sum_{n=0}^{\infty} \frac{1}{a^n}.$$

Clearly $\sum_{n=0}^{\infty} \frac{1}{a^n}$ is a convergent series, so let

$$\mathcal{C} = \sum_{n=0}^{\infty} \frac{1}{a^n},$$

and then we have

$$\left|\alpha - \frac{p_k}{q_k}\right| < \frac{\mathcal{C}}{a^{(k+1)!}} = \frac{\mathcal{C}}{q_k^{(k+1)}} < \frac{\mathcal{C}}{q_k^k}. \tag{4.11}$$

Suppose that α is rational, say $\alpha = c/d$ for some $c, d, \in \mathbb{Z}$. Then (4.11) implies that

$$|cq_k - dp_k| < \frac{\mathcal{C}d}{q_k^{k-1}}$$

for infinitely many p_k/q_k as above. The expression $\frac{\mathcal{C}d}{q_k^{k-1}}$ 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 α cannot be rational.

Now suppose that α is algebraic of degree d. Then, by Theorem 4.4.1, there exists a constant $C(\alpha)$ such that

$$\left|\alpha - \frac{p_k}{q_k}\right| \geq \frac{C(\alpha)}{q_k^d},$$

for every $k \in \mathbb{Z}_{>0}$. However, if we take k large enough so that

$$\frac{\mathcal{C}}{q_k^k} < \frac{C(\alpha)}{q_k^d},$$

then (4.11) implies a contradiction; more specifically, we just need to take k large enough so that

$$k!(k-d) > \frac{\ln \mathcal{C} - \ln C(\alpha)}{\ln a}.$$

This completes the proof. □

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 π (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

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{q^\mu}. \tag{4.12}$$

Indeed, suppose there were infinitely many rational numbers for which (4.12) holds. Let $C(\alpha)$ be the constant guranteed by Theorem 4.4.1. Let Q be an integer so that $C(\alpha) > \frac{1}{Q^{\mu-d}}$. 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

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

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

For an algebraic number α of degree $d \geq 2$, what is the smallest possible μ 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 α 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 α is algebraic, then the number of sufficiently good rational approximations to α is finite, so perhaps one can actually count them, although we are not quite ready to do this. If α is real, but not necessarily algebraic, there may be infinitely many good rational approximations to α, 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 *γ-set* for a real number $\gamma > 1$ if whenever $m, n \in S$ and $m < n$, then $\gamma m \leq n$.

Notice that a γ-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 γ-set will be called a *(C, γ)-set*. Notice that a (C, γ)-set is always finite. It is possible to estimate the cardinality of a (C, γ)-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, γ)-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

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

Hence

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

and (4.13) follows. □

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 α 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*

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

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

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

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

$$\begin{aligned} \frac{1}{q_1 q_2} &\leq \left|\frac{p_1}{q_1} - \frac{p_2}{q_2}\right| = \left|\left(\frac{p_1}{q_1} - \alpha\right) + \left(\alpha - \frac{p_2}{q_2}\right)\right| \\ &\leq \left|\alpha - \frac{p_1}{q_1}\right| + \left|\alpha - \frac{p_2}{q_2}\right| < \frac{1}{2q_1^{2+\delta}} + \frac{1}{2q_2^{2+\delta}} \leq \frac{1}{q_1^{2+\delta}}, \end{aligned}$$

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 γ-set. On the other hand, if q_1, q_2 are in a window of exponential width C, then

$$\ln q_1 \leq C \ln q_2, \ \ln q_2 \leq C \ln q_1,$$

that is these logarithms also form a C-set, hence they form a (C, γ)-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. □

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 α 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

$$A = e^{\ln A} \leq q \leq B = \left(e^{\ln A}\right)^{\frac{\ln B}{\ln A}},$$

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 *δ-approximation* to α 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 δ-approximations $\frac{p}{q}$ to α 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 α, an expression of the form

$$\alpha = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{a_4 + \dots}}}},$$

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

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

We will call a continued fraction expansion for α finite if there exists some n for which α 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 α is rational.*

PROOF. To prove this result, we present an actual algorithm to compute a continued fraction expansion for α. We do it recursively. Define $a_0 = [\alpha]$, the integer part of α. If $\alpha = [\alpha]$, we are done. If not, then

$$\alpha = [\alpha] + (\alpha - [\alpha]) = a_0 + \frac{1}{r_1},$$

where $r_1 = \frac{1}{\alpha - [\alpha]}$. If r_1 is an integer, we are done. If not, let $a_1 = [r_1]$, and so

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

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

$$\alpha = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots + \cfrac{1}{r_n}}}}.$$

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 α, meaning that $\alpha \in \mathbb{Q}$. Hence irrational numbers must have an infinite continued fraction expansion.

On the other hand, suppose α is rational, then each r_n is a rational number, say $r_n = p_n/q_n$, and $r_n > 1$. Therefore $p_n > q_n$, and

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

and so $p_{n+1} = q_n > q_{n+1}$. This means that the sequence of denominators of $r_1, r_2, \dots$, namely $q_1, q_2, \dots$ 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, $\frac{7}{3} = 2 + \frac{1}{3} = [2; 3]$, as well as

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

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

$$\begin{aligned} \sqrt{2} &= \sqrt{2}\left(\frac{1+\sqrt{2}}{1+\sqrt{2}}\right) = 1 + \frac{1}{1+\sqrt{2}} = 1 + \frac{1}{1 + \left(1 + \frac{1}{1+\sqrt{2}}\right)} \\ &= 1 + \frac{1}{2 + \frac{1}{1+\left(1+\frac{1}{1+\sqrt{2}}\right)}} = 1 + \frac{1}{2 + \frac{1}{2+\frac{1}{2+\frac{1}{2+\frac{1}{2+\dots}}}}} \end{aligned}$$

as well as

$$\pi = 3 + \cfrac{1}{7 + \cfrac{1}{15 + \cfrac{1}{1 + \cfrac{1}{292 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{3 + \cfrac{1}{1 + \dots}}}}}}}}}}$$

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

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

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

PROOF. First consider the rational case, and notice that indeed

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots + \cfrac{1}{a_n}}}} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots + \cfrac{1}{(a_n - 1) + \cfrac{1}{1}}}}}. \tag{4.16}$$

Now suppose that $[a_0; a_1, a_2, \dots]$ is either an infinite continued fraction expansion for a real number α 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 $[b_0; b_1, b_2, \dots]$ so that

$$\alpha = [a_0; a_1, a_2, \dots] = [b_0; b_1, b_2, \dots],$$

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 b_0, since otherwise

$$|a_0 - b_0| \geq 1 > |[0; a_1, \dots] - [0; b_1, \dots]|, \tag{4.17}$$

and so we cannot have $[a_0; a_1, \dots] = [b_0; b_1, \dots]$. Suppose $a_k = b_k$ for all $k \leq m$, then we have

$$[a_0; a_1, \dots, a_m, a_{m+1}, \dots] = [a_0; a_1, \dots, a_m, b_{m+1}, \dots],$$

which implies that there must be equality of the new continued fractions:

$$[a_{m+1}; a_{m+2}, \dots] = [b_{m+1}; b_{m+2}, \dots].$$

By the same argument as above,

$$|a_{m+1} - b_{m+1}| \geq 1 > |[0; a_{m+2}, \dots] - [0; b_{m+2}, \dots]|, \tag{4.18}$$

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. □

Since each n-th convergent α_n of an irrational number α 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 α by its convergents α_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 $n \geq 1$,*

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

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 $n \geq 1$*

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

PROOF. We argue by induction on n. First notice that $\frac{p_0}{q_0} = \frac{a_0}{1}$ and $\frac{p_1}{q_1} = \frac{a_0 a_1 + 1}{a_1}$. If $n = 2$, we have

$$\frac{p_2}{q_2} = a_0 + \frac{1}{a_1 + \frac{1}{a_2}} = \frac{a_0 a_1 a_2 + a_2 + a_0}{a_1 a_2 + 1} = \frac{a_2 p_1 + p_0}{a_2 q_1 + q_0}.$$

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

$$\frac{t_{k-1}}{s_{k-1}} = [a_1; a_2, \dots, a_{k-1}]$$

to be the $(k-1)$-st convergent of the derived continued fraction $[a_1; a_2, \dots]$. Then induction hypothesis applies to t_{k-1}/s_{k-1}, and

$$\begin{aligned} \frac{p_k}{q_k} &= a_0 + \frac{s_{k-1}}{t_{k-1}} = \frac{a_0 t_{k-1} + s_{k-1}}{t_{k-1}} = \frac{a_0(a_k t_{k-1} + t_{k-2}) + a_k s_{k-1} + s_{k-2}}{a_k t_{k-1} + t_{k-2}} \\ &= \frac{a_k(a_0 t_{k-1} + s_{k-1}) + (a_0 t_{k-2} + s_{k-2})}{a_k t_{k-1} + t_{k-2}}. \end{aligned}$$

Now, Problem 4.10 guarantees that for each n,

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

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. □

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

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

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

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

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:

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

□

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

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

for all $k \geq 0$.

PROOF. By Lemma 4.6.5,

$$\begin{aligned} \frac{p_{n-2}}{q_{n-2}} - \frac{p_n}{q_n} &= \left(\frac{p_{n-2}}{q_{n-2}} - \frac{p_{n-1}}{q_{n-1}}\right) + \left(\frac{p_{n-1}}{q_{n-1}} - \frac{p_n}{q_n}\right) \\ &= \frac{(-1)^{n-1}}{q_{n-1}q_{n-2}} + \frac{(-1)^n}{q_n q_{n-1}} = \frac{(-1)^{n-1}}{q_{n-1}}\left(\frac{1}{q_{n-2}} - \frac{1}{q_n}\right). \end{aligned}$$

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 α, the conclusion follows. □

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 α and the even-numbered ones are less or equal than α. Therefore

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

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. □

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 α among all rational numbers with denominators no bigger than q_n.

PROOF. Let $\frac{p}{q} \neq \frac{p_n}{q_n}$ be any rational approximation to α 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

$$p = xp_n + yp_{n+1},\ q = xq_n + yq_{n+1}. \tag{4.19}$$

We cannot have x and y both equal 0. Suppose that $x = 0$, $y \neq 0$, then $q = yq_{n+1} > q_n$ by Lemma 4.6.4, which contradicts the choice of q. Then assume $y = 0$, $x \neq 0$, then $q = xq_n$, so $x = 1$, and hence $p = p_n$, $q = q_n$, again a contradiction. Hence we must have $x, y \neq 0$. Further,

$$0 < q = xq_n + yq_{n+1} \leq q_n < q_{n+1},$$

thus x and y must have different signs. Notice that

$$q_n\alpha - p_n \text{ and } q_{n+1}\alpha - p_{n+1}$$

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:

$$\left|\alpha - \frac{p}{q}\right| \geq \frac{|q\alpha - p|}{q} > \frac{|q_n\alpha - p_n|}{q} \geq \frac{|q_n\alpha - p_n|}{q_n} = \left|\alpha - \frac{p_n}{q_n}\right|.$$

This completes the proof. □

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 α 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 $y = \frac{p_n}{q_n}x$ 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

$$\alpha = \sqrt{2} = [1; 2, 2, 2, 2, \dots] = 1.41421356237\dots,$$

then:

$$\alpha_1 = \frac{p_1}{q_1} = [1; 2] = 1 + \frac{1}{2} = \frac{3}{2} = 1.5$$

$$\alpha_2 = \frac{p_2}{q_2} = [1; 2, 2] = 1 + \frac{1}{2 + \frac{1}{2}} = \frac{7}{5} = 1.4$$

$$\alpha_3 = \frac{p_3}{q_3} = [1; 2, 2, 2] = 1 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2}}} = \frac{17}{12} = 1.4166666666\dots$$

$$\alpha_4 = \frac{p_4}{q_4} = [1; 2, 2, 2, 2] = 1 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2}}}} = \frac{41}{29} = 1.41379310345$$

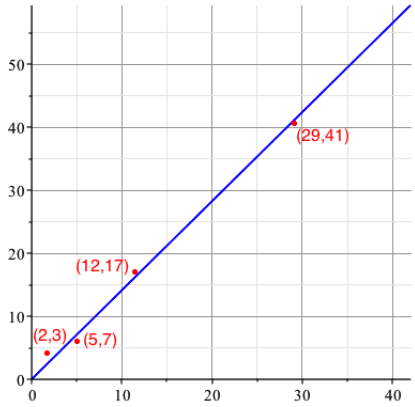


FIGURE 1. Continued fraction convergent approximations to $\sqrt{2}$

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 α.

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 α, hence it cannot be dense. We now prove that in case α 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

$$\|n\alpha - x\| < \varepsilon. \tag{4.20}$$

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

$$0 < \|\alpha\| \leq |\alpha q - p| \leq \frac{1}{q} < \varepsilon.$$

This inequality implies that either

$$0 < \alpha q - p < 1/q \tag{4.21}$$

or

$$-1/q < \alpha q - p < 0. \tag{4.22}$$

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

$$\|\ell q\alpha - x\| \leq |\ell(\alpha q - p) - x| < |\alpha q - p| < \frac{1}{q} < \varepsilon,$$

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 $p - \alpha q$. □

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 α 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

$$(\{n\alpha_1\}, \ldots, \{n\alpha_m\})_{n=1}^{\infty}$$

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 Stanislaw 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 2π 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 α be the angular position of θ, then angular positions of the angles

$$\theta, 2\theta, \ldots, n\theta$$

are given by the sequence of fractional parts

$$(\xi_k)_{k=1}^n = \{k\alpha\}_{k=1}^n.$$

Then the distances between our angular positions on the circle are precisely the gaps between corresponding numbers in this sequence $(\xi_k)_{k=1}^n$. The gap between ξ_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

$$\begin{aligned} s_{k,n} &= \min\left\{(\ell-k)\alpha + n > 0 : (\ell, n) \in \mathbb{Z}^2, 0 < \ell \le n\right\} \\ &= \min\left\{m\alpha + n > 0 : (m, n) \in \mathbb{Z}^2, -k < m \le n-k\right\}, \end{aligned}$$

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

$$s_{k,n} = \min\left\{y > 0 : (x,y)^\top \in A_1\mathbb{Z}^2, -k < x \le n-k\right\}.$$

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

$$F(M,t) = \min\left\{y > 0 : (x,y)^\top \in M\mathbb{Z}^2, -t < x \le 1-t\right\}.$$

This is a function $F : \mathrm{SL}_2(\mathbb{R}) \times (0,1] \to \mathbb{R}_{>0}$, where $\mathrm{SL}_2(\mathbb{R}) = \{M \in \mathrm{GL}_2(\mathbb{R}) : \det(M) = 1\}$ is a subgroup of $\mathrm{GL}_2(\mathbb{R})$ 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

$$A_n = \begin{pmatrix} n^{-1} & 0 \\ 0 & n \end{pmatrix} A_1 \in \mathrm{SL}_2(\mathbb{R}),$$

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 $M \in \mathrm{SL}_2(\mathbb{R})$, 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 $\alpha \in \mathbb{A}$ 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 $\alpha \in \mathbb{A}$. 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 α.

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 μ-approximation to the real number α if*

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

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 α be a rational number, $0 < \alpha < 1$. Prove that $f(\alpha)$ is transcendental.

PROBLEM 4.10. *Let α have the continued fraction expansion $[a_0; a_1, a_2, \dots]$, and let p_n/q_n be its n-th convergent. Let t_n/s_n be the n-th convergent of the number $[a_1; a_2, \dots]$. Prove that*

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

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

$$\phi = 1 + \frac{1}{\phi}. \tag{4.23}$$

PROBLEM 4.12. *Derive a continued fraction expansion for ϕ*

$$[a_0; a_1, a_2, \dots]$$

and use it to prove that ϕ is irrational.

PROBLEM 4.13. *Let us write ϕ_n for the n-th continued fraction approximations to ϕ, i.e.*

$$\phi_n = [a_0; a_1, \dots, a_n].$$

Compute ϕ_1 through ϕ_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*

$$F_n = F_{n-1} + F_{n-2}. \tag{4.24}$$

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 $K \subseteq L \subset \mathbb{C}$ are algebraic fields.

DEFINITION 5.1.2. Let $K \subseteq \mathbb{C}$ and $\alpha \in \mathbb{C}$. We define

$$\begin{aligned} K[\alpha] &:= \operatorname{span}_K \left\{1, \alpha, \alpha^2, \ldots\right\} \\ &= \left\{ \sum_{m=0}^{n} a_m \alpha^m : a_0, \ldots, a_n \in K,\ n \in \mathbb{Z}_{\geq 0} \right\}, \end{aligned}$$

i.e., the set of all finite linear combinations of powers of α 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 α 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 α is transcendental, then $\dim_{\mathbb{Q}} \mathbb{Q}[\alpha] = \infty$.*
(2) *If α is algebraic of degree n, then $\mathbb{Q}[\alpha] = \operatorname{span}_{\mathbb{Q}}\{1, \alpha, \ldots, \alpha^{n-1}\}$, 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 α is algebraic.*
(4) *If α is algebraic, then $\mathbb{Q}(\alpha) = \mathbb{Q}[\alpha]$.*

PROOF. To prove part (1), assume that $\dim_{\mathbb{Q}} \mathbb{Q}[\alpha] = n < \infty$. 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 α 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 α 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

$$\alpha^n = \sum_{m=0}^{n-1} \left(-\frac{a_m}{a_n}\right) \alpha^m. \tag{5.1}$$

Therefore any $\mathbb{Q}$-linear combination of powers of α 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

$$c_0 + c_1\alpha + \ldots c_{n-1}\alpha^{n-1} = 0.$$

In fact, clearing the denominators if necessary, we can assume that $c_0, \ldots, c_{n-1} \in \mathbb{Z}$. But this means that α 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

$$\gamma = \sum_{m=0}^{n-1} c_m \alpha^m \in \mathbb{Q}[\alpha] \tag{5.2}$$

such that $\beta\gamma = 1$. Let γ 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 α^n, lowering the power. After a finite number of such substitutions, we will obtain an expression

$$\beta\gamma = \sum_{l=0}^{n-1} f_l(c_0, \ldots, c_{n-1}) \alpha^l, \tag{5.3}$$

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

$$\begin{aligned} f_0(c_0, \ldots, c_{n-1}) &= 1 \\ f_1(c_0, \ldots, c_{n-1}) &= 0 \\ \vdots \ldots \vdots \ldots \vdots \ldots \vdots & \\ f_{n-1}(c_0, \ldots, c_{n-1}) &= 0. \end{aligned} \tag{5.4}$$

This is a linear system of n equations in n variables, which can be written as $F\boldsymbol{c} = \boldsymbol{e}_1$, where F is the $n \times n$ coefficient matrix of linear polynomials $f_0, \ldots, f_{n-1}$, $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^t \in \mathbb{R}^n$ is the first standard basis vector in $\mathbb{R}^n$, and $\boldsymbol{c} = (c_0, \ldots, c_{n-1})^t$. The matrix F must be nonsingular matrix. Indeed, suppose it is singular, then there exists some $\mathbf{0} \neq \boldsymbol{c} \in \mathbb{Q}^n$ such that $F\boldsymbol{c} = \mathbf{0}$, i.e.

$$f_l(c_0, \ldots, c_{n-1}) = 0 \ \forall \ 0 \leq l \leq n-1.$$

Let γ as in (5.2) be defined with this choice of the coefficient vector $\boldsymbol{c}$. 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 $\boldsymbol{c}$. Let γ be as in (5.2) with this choice of $\boldsymbol{c}$, then $\gamma = \beta^{-1} \in \mathbb{Q}[\alpha]$, and so $\mathbb{Q}[\alpha]$ is a field.

Now suppose α is not algebraic, i.e. it is transcendental. We show that $\mathbb{Q}[\alpha]$ is not a field. Assume it is, then $\alpha^{-1} \in \mathbb{Q}[\alpha]$, which means that

$$\alpha^{-1} = \sum_{m=0}^{n} a_m \alpha^m$$

for some $n \in \mathbb{N}_0$ and $a_0, \ldots, a_n \in \mathbb{Q}$. Hence

$$1 = \alpha\alpha^{-1} = \sum_{m=0}^{n} a_m \alpha^{m+1},$$

and so

$$\sum_{m=0}^{n} a_m \alpha^{m+1} - 1 = 0.$$

This is a polynomial equation over $\mathbb{Q}$ satisfied by α, and multiplying through by the product of denominators of its coefficients we can obtain a polynomial equation over $\mathbb{Z}$ satisfied by α. This contradicts the assumption that α is transcendental. Hence $\mathbb{Q}[\alpha]$ cannot be 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 α must be contained in any field containing $\mathbb{Q}$ and α. To show containment the other way, notice that, by part (3), $\mathbb{Q}[\alpha]$ is a field containing $\mathbb{Q}$ and α, and so it must contain $\mathbb{Q}(\alpha)$. □

EXAMPLE 5.1.1. *We give an example of finding the inverse of an element of* $\mathbb{Q}[\alpha]$ *when* α *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

$$\begin{aligned} 1 &= \beta\beta^{-1} = a3^{4/3} + b3^{3/3} - c3^{2/3} + 2a3^{3/3} + 2b3^{2/3} - 2c3^{1/3} \\ &\quad -2a3^{2/3} - 2b3^{1/3} + 2c \\ &= (2b - 2a - c)3^{2/3} + (3a - 2c - 2b)3^{1/3} + (2c + 6a + 3b), \end{aligned}$$

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 α is transcendental, then $\dim_{\mathbb{Q}} \mathbb{Q}[\alpha] = \infty$ 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 α 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

$$\infty > n = \dim_{\mathbb{Q}} \mathbb{Q}[\alpha] = [\mathbb{Q}(\alpha) : \mathbb{Q}].$$

This is a contradiction, and hence α must be transcendental. □

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

$$\mathbb{A} = \{\alpha_1, \alpha_2, \alpha_3, \dots\},$$

choosing an ordering on $\mathbb{A}$. For each $n \in \mathbb{N}$, define

$$K_n := \mathbb{Q}(\alpha_1, \dots, \alpha_n).$$

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 β 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

$$\beta^{-1}, \gamma^{-1}, \beta \pm \gamma, \beta\gamma \in K_n \subseteq \mathbb{A}.$$

Therefore $\mathbb{A}$ is a field. □

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 α^{-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. □

REMARK 5.1.1. A consequence of Lemma 5.1.4 is that, given one transcendental number β, we can produce infinitely many (but countably many) transcendental numbers:

$$\alpha \pm \beta,\ \alpha\beta,\ \alpha^{-1}\beta \ \forall\ 0 \neq \alpha \in \mathbb{A}.$$

Take, for instance, β 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 $K \subseteq \mathbb{C}$ be a field and $\alpha \in \mathbb{C}$. We say that α 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 α 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 α (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, $m_\alpha(x) \mid f(x)$, 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 α 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, $f(x) \mid f'(x)$. 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. □

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 α is irreducible (Problem 4.3), it must be separable by Lemma 5.2.1, and hence α and its conjugates are all distinct. □

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 α in a number field K is called a *primitive element* if $K = \mathbb{Q}(\alpha)$, i.e. if α 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 $K_1 = \mathbb{Q}(\alpha_1)$, $K_2 = \mathbb{Q}(\alpha_1, \alpha_2) = K_1(\alpha_2)$, $\ldots$, $K = K_n = \mathbb{Q}(\alpha_1, \ldots, \alpha_{n-1}, \alpha_n) = K_{n-1}(\alpha_n)$. We can assume that no K_m equal to K_{m+1}, since otherwise we do not need α_{m+1} in the generating set. Hence we have

$$\mathbb{Q} \subsetneq K_1 \subsetneq K_2 \subsetneq \cdots \subsetneq K_{n-1} \subsetneq K_n = K.$$

Notice that it is sufficient for us to show that there exists $\beta_1 \in K$ such that $K_2 = \mathbb{Q}(\alpha_1, \alpha_2) = \mathbb{Q}(\beta_1)$: 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 α_1 and α_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 α_{1n}'s and all α_{2m}'s are distinct. This means that for each $1 \leq n \leq d$, $1 < m \leq e$ the equation

$$\alpha_{1n} + t\alpha_{2m} = \alpha_{11} + t\alpha_{21} \tag{5.5}$$

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, α_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, α_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 ξ 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 $1 \leq n \leq d$, $1 \leq m \leq e$,

$$\xi = \alpha_{2m} \text{ and } \beta_1 - c\xi = \beta_1 - c\alpha_{2m} = \alpha_{1n},$$

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 α_2 over $\mathbb{Q}(\beta_1)$. Since $p(x)$ and $m_{\alpha_2}(x)$ have coefficients in $\mathbb{Q}(\beta_1)$ and vanish at α_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. □

An algebraic number α 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

$$\mathbb{I} = \{\alpha \in \mathbb{A} : m_\alpha(x) \text{ is monic}\}.$$

Let $\alpha \in \mathbb{I}$ 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 $\alpha \in \mathbb{I}$ 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

$$\beta = \sum_{n=0}^{d-1} b_n \alpha^n, \ \gamma = \sum_{n=0}^{d-1} c_n \alpha^n.$$

Then

$$\beta + \gamma = \sum_{n=0}^{d-1} (b_n + c_n)\alpha^n \in \mathbb{Z}[\alpha].$$

Since $\alpha \in \mathbb{I}$ 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

$$\alpha^d = -\sum_{n=0}^{d-1} a_n \alpha^n. \tag{5.6}$$

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 α higher than $(d-1)$-st as linear combinations of lower powers of α 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. □

Problem 5.13 guarantees that $1, \alpha, \ldots, \alpha^{d-1}$ is a maximal linearly independent collection of powers of α 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 α is in fact finitely generated. Then $\alpha \in \mathbb{I}$.*

PROOF. Let G be the additive abelian group generated by all powers of α, 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 α, there exists a positive integer ℓ which is the maximal power of α 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 α 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 $\alpha \in \mathbb{I}$. □

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

$$\alpha + \beta, \alpha\beta \in G := \operatorname{span}_{\mathbb{Z}}\{\alpha^m \beta^n : m, n \in \mathbb{Z}_{\geq 0}\} \subset \mathbb{C}.$$

This G is a subgroup of $\mathbb{C}$ under the usual addition of complex numbers, and hence is an additive abelian group (Problem 5.14). Since α and β are algebraic integers, we know that $\mathbb{Z}[\alpha]$ and $\mathbb{Z}[\beta]$ are generated by only finitely many powers of α and β, 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 $\mathbb{I}$. □

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 5.15). 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 α/c, where α 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

$$p(x) := a_d^{d-1} m_\beta(x) = \sum_{n=0}^{d} a_n a_d^{d-1} x^n = \sum_{n=0}^{d} a_n a_d^{d-n-1} (a_d x)^n$$

has β 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. □

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 $\beta = \frac{\alpha}{c}$ 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 $\alpha/\beta = \alpha\beta^{-1} \in E$. Since $\alpha, \beta \in \mathcal{O}_K \subset K$, 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 $E = K$. □

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 $\alpha \in \mathcal{O}_K$ 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 $\alpha := c\beta \in \mathcal{O}_K$. Since clearly $\mathbb{Q}(c\beta) = \mathbb{Q}(\beta)$, we are done. □

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

$$K = \mathbb{Q}(\alpha) = \mathbb{Q}[\alpha] = \operatorname{span}_{\mathbb{Q}}\{1, \alpha, \ldots, \alpha^{d-1}\},$$

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 α. For each $1 \leq n \leq d$, define a map $\sigma_n : K \to \mathbb{C}$, given by

$$\sigma_n\left(\sum_{m=0}^{d-1} a_m \alpha^m\right) = \sum_{m=0}^{d-1} a_m \alpha_n^m, \tag{5.7}$$

for each $\sum_{m=0}^{d-1} a_m \alpha^m \in K$. From Problem 5.16 we know that each such σ_n is an injective field homomorphism, so $K \cong \sigma_n(K)$ for each $1 \leq n \leq d$, and

$$\mathbb{Q} = \{\beta \in K : \sigma_n(\beta) = \beta \ \forall \ 1 \leq n \leq d\}.$$

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 τ 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 τ 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

$$K = \mathbb{Q}[\alpha] = \operatorname{span}_{\mathbb{Q}}\{1, \alpha, \ldots, \alpha^{d-1}\}.$$

Let $\alpha = \alpha_1, \alpha_2, \ldots, \alpha_d$ be the algebraic conjugates of α. Let $\beta = \sum_{n=0}^{d-1} c_n \alpha^n \in K$. Since τ 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 α. 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 σ_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 5.17). 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 5.16. 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

$$\Delta(\alpha_1, \ldots, \alpha_d) := \left(\det(\sigma_n(\alpha_k))_{1 \leq n,k \leq d}\right)^2,$$

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 θ 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

$$\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_d \end{pmatrix} = \begin{pmatrix} c_{11} & \ldots & c_{1d} \\ \vdots & \ddots & \vdots \\ c_{d1} & \ldots & c_{dd} \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ \theta^{d-1} \end{pmatrix},$$

i.e. $C = (c_{mn})_{1 \leq m,n \leq d}$ is a rational change of basis matrix. Then by Problem 5.22

$$\Delta(\alpha_1, \ldots, \alpha_d) = \det(C)^2 \Delta(1, \ldots, \theta^{d-1}),$$

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 θ, i.e. $\theta_k = \sigma_k(\theta)$. Then

$$\Delta := \Delta(1, \ldots, \theta^{d-1}) = \left\{ \det \begin{pmatrix} 1 & \theta_1 & \ldots & \theta_1^{d-1} \\ 1 & \theta_2 & \ldots & \theta_2^{d-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \theta_d & \ldots & \theta_d^{d-1} \end{pmatrix} \right\}^2.$$

Notice that each σ_k permutes the conjugates $\theta_1, \ldots, \theta_d$, which means that the action of each σ_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 5.23 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 $\alpha_1, \ldots, \alpha_d \in \mathcal{O}_K$, 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 $\mathbb{I}$, but $\mathbb{Q} \cap \mathbb{I} = \mathbb{Z}$. Thus $\Delta(\alpha_1, \ldots, \alpha_d) \in \mathbb{Z}$. □

We now use the discriminant to prove that the ring of integers $\mathcal{O}_K$ in a number field K of degree d 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 $\mathcal{O}_K$ is an order in K. Notice that if $K = \mathbb{Q}(\theta)$ for some algebraic θ, then $\mathbb{Z}[\theta] \subseteq \mathcal{O}_K$ is also an order in K, however $\mathcal{O}_K$ is a maximal order with respect to inclusion (we will not prove it here), and $\mathbb{Z}[\theta]$ is a maximal order precisely when $\mathcal{O}_K = \mathbb{Z}[\theta]$.

THEOREM 5.2.12. *Let K be a number field of degree d over $\mathbb{Q}$. Then the ring $\mathcal{O}_K$ is a lattice of rank d, i.e. there exists a collection of $\mathbb{Q}$-linearly independent elements $\alpha_1, \ldots, \alpha_d \in \mathcal{O}_K$ such that*

$$\mathcal{O}_K = \left\{ \sum_{n=1}^{d} a_n \alpha_n : a_1, \ldots, a_d \in \mathbb{Z} \right\}.$$

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 $c_1\alpha_1, \ldots, c_d\alpha_d \in \mathcal{O}_K$. Thus the set of linearly independent collections of d elements of $\mathcal{O}_K$ 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

$$\mathcal{O}_K = \left\{ \sum_{n=1}^{d} a_n \beta_n : a_1, \ldots, a_d \in \mathbb{Z} \right\}.$$

Suppose this is not true, then there exists some $x \in \mathcal{O}_K$, 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 $(a_1 - q)\beta_1 \in \mathcal{O}_K$, and since $x \in \mathcal{O}_K$, we have

$$y := x - (a_1 - q)\beta_1 = q\beta_1 + \sum_{k=2}^{d} a_k \beta_k \in \mathcal{O}_K.$$

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 5.22

$$\begin{aligned} \Delta(y, \beta_2, \ldots, \beta_d) &= \det(A)^2 \Delta(\beta_1, \beta_2, \ldots, \beta_d) \\ &= q^2 \Delta(\beta_1, \beta_2, \ldots, \beta_d) < \Delta(\beta_1, \beta_2, \ldots, \beta_d). \end{aligned}$$

This contradicts the choice of $\beta_1, \ldots, \beta_d$, and hence completes our proof. □

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 $A \in \mathrm{GL}_d(\mathbb{Z})$ between them (Problem 5.24). Therefore we have, by Problem 5.22,

$$\Delta(\alpha_1, \ldots, \alpha_d) = \Delta(\beta_1, \ldots, \beta_d).$$

This common value of the discriminant of all the integral bases of a number field K is called the *discriminant of* K, denoted Δ_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*

$$I_1 \subseteq I_2 \subseteq \dots$$

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 $I \not\subseteq J$ 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

$$I_1 \subseteq I_2 \subseteq \dots$$

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 5.25). Hence I must be finitely generated, say $x_1, \dots, x_n$ is a set of generators for I. Then each generator x_k lies in some ideal I_{ℓ_k}, and so $x_1, \dots, x_n \in I_m$, where $m = \max_{1 \leq k \leq n} \ell_k$. Hence

$$I = \langle x_1, \dots, x_n \rangle \subseteq I_m \subseteq I,$$

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 $I \subsetneq J$. Construct an ascending chain of ideals from S

$$I_1 \subseteq I_2 \subseteq \dots$$

by picking each I_n such that $I_{n-1} \subsetneq I_n$. 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 elements. 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 $J \subseteq I$ and J is finitely generated. Suppose I is not finitely generated, then $J \subsetneq I$, i.e. there exists $x \in I \setminus J$. Let $J' = \langle J, x \rangle$, then $J \subsetneq J'$ 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 5.18). 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 $x = x_1 x_2$, 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

$$\cdots \mid y_n \mid y_{n-1} \mid \cdots \mid y_2 \mid y_1 \mid x,$$

which means that there is an infinite ascending chain of ideals

$$\langle x \rangle \subsetneq \langle y_1 \rangle \subsetneq \langle y_2 \rangle \subsetneq \ldots,$$

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 5.19). 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

$$\sqrt{2} - 1, \sqrt{2} + 1 \in \mathcal{O}_{\mathbb{Q}(\sqrt{2})},$$

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. □

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 5.20), however

$$6 = 2 \times 3 = (1 + \sqrt{-5})(1 - \sqrt{-5}), \tag{5.8}$$

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

$$a = q_1 \cdots q_k,\ b = t_1 \cdots t_m,\ c = s_1 \cdots s_n$$

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. $\square$

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 $\sigma_1, \dots, \sigma_d : K \to \mathbb{C}$. Let $\alpha \in K$, then *norm* of α in K is defined as

$$\mathbb{N}_K(\alpha) = \prod_{k=1}^{d} \sigma_k(\alpha),$$

and *trace* of α in K is

$$\mathrm{Tr}_K(\alpha) = \sum_{k=1}^{d} \sigma_k(\alpha).$$

It follows directly from these definitions that norm is a multiplicative function and trace is linear over $\mathbb{Q}$, i.e. for any $\alpha, \beta \in K$,

$$\mathbb{N}_K(c\alpha\beta) = c\,\mathbb{N}_K(\alpha)\mathbb{N}_K(\beta),\ \mathrm{Tr}_K(a\alpha + b\beta) = a\,\mathrm{Tr}_K(\alpha) + b\,\mathrm{Tr}_K(\beta)$$

for any $a, b, c \in \mathbb{Q}$ (Problem 5.26). There are some additional important properties of norm and trace that we will establish here.

LEMMA 5.4.1. *Let $\alpha \in K$, then $\mathbb{N}_K(\alpha), \mathrm{Tr}_K(\alpha) \in \mathbb{Q}$. Further, if $\alpha \in \mathcal{O}_K$, then $\mathbb{N}_K(\alpha), \mathrm{Tr}_K(\alpha) \in \mathbb{Z}$.*

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 $\tau_1, \dots, \tau_n$ be embeddings of L and $\sigma_1, \dots, \sigma_d$ embeddings of K. Then σ_i's restricted to L must be equal to τ_j's, and τ_j's extend to σ_i's on K. In fact, every τ_j extends to the same number of σ_i's, namely to d/n of them, and no two different τ_j's can extend to the same σ_i. Therefore

$$\mathbb{N}_K(\alpha) = \prod_{i=1}^{d} \sigma_i(\alpha) = \prod_{j=1}^{n} \tau_j(\alpha)^{d/n} = \left(\prod_{j=1}^{n} \tau_j(\alpha) \right)^{d/n},$$

$$\mathrm{Tr}_K(\alpha) = \sum_{i=1}^{d} \sigma_i(\alpha) = \sum_{j=1}^{n} \left(\frac{d}{n}\ \tau_j(\alpha) \right) = \frac{d}{n} \sum_{j=1}^{n} \tau_j(\alpha).$$

Let $m_\alpha(x) \in \mathbb{Z}[x]$ be the minimal polynomial of α over $\mathbb{Z}$, then $m_\alpha(x)$ can be factored as

$$m_\alpha(x) = \sum_{j=0}^{n} c_j x^j = c_n(x - \tau_1(\alpha)) \cdots (x - \tau_n(\alpha)),$$

where $c_0, \dots, c_n \in \mathbb{Z}$ and $c_n = 1$ if and only if $\alpha \in \mathcal{O}_K$. Hence we see that

$$c_n \prod_{j=1}^{n} \tau_j(\alpha) = c_0,\ c_n \sum_{j=1}^{n} \tau_j(\alpha) = -c_{n-1}.$$

The result follows. □

COROLLARY 5.4.2. *An element $\alpha \in \mathcal{O}_K$ is a unit if and only if $\mathbb{N}_K(\alpha) = \pm 1$. On the other hand, if $\mathbb{N}(\alpha) = p$, a rational prime, then α 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:

$$\mathbb{N}_K(\alpha\alpha^{-1}) = \mathbb{N}_K(\alpha)\mathbb{N}_K(\alpha^{-1}) = 1.$$

By Lemma 5.4.1, $\mathbb{N}_K(\alpha), \mathbb{N}_K(\alpha^{-1}) \in \mathbb{Z}$, hence we must have

$$\mathbb{N}_K(\alpha) = \mathbb{N}_K(\alpha^{-1}) = \pm 1.$$

On the other hand, suppose $\mathbb{N}_K(\alpha) = \pm 1$. There certainly exists $\alpha^{-1} \in K$: it is our goal to prove that $\alpha^{-1} \in \mathcal{O}_K$. We have that

$$1 = \mathbb{N}_K(\alpha\alpha^{-1}) = \mathbb{N}_K(\alpha)\mathbb{N}_K(\alpha^{-1}) = \mathbb{N}_K(\alpha^{-1}).$$

Since $\alpha \in \mathcal{O}_K$ is of norm ± 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 α^{-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 $\mathbb{N}_K(\alpha) = p$, a rational prime. Suppose $\alpha = xy$ for some $x, y \in \mathcal{O}_K$. Then Lemma 5.4.1 implies that

$$\mathbb{N}_K(x)\mathbb{N}_K(y) = p,$$

meaning that one of x, y has norm ± 1 and the other $\pm p$, hence one of them is a unit. Since this is true for every factorization of α in $\mathcal{O}_K$, we conclude that α 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 α. Notice that

$$\alpha_2 \cdots \alpha_n = \frac{\mathbb{N}_K(\alpha)}{\alpha} \in K.$$

On the other hand, α 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

$$\mathbb{N}_K(\alpha) = \alpha(\alpha_2 \cdots \alpha_n) \in P.$$

Lemma 5.4.1 guarantees that $m := \mathbb{N}_K(\alpha) \in \mathbb{Z}$, 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 $O_K/m\mathcal{O}_K$ is a finitely generated abelian group in which every element has order $\leq m$ (and dividing m): this must be a finite group (Problem 5.27). 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.

$$\mathbb{N}_K(I) = |\mathcal{O}_K/I|.$$

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*

$$\mathbb{N}_K(I) = \left|\frac{\Delta(\alpha_1, \ldots, \alpha_d)}{\Delta_K}\right|^{1/2},$$

where $\alpha_1, \ldots, \alpha_d$ is an integral basis for I and Δ_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 1.17), 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 1.3.6,

$$\mathbb{N}_K(I) = \frac{\det(I)}{\det(\mathcal{O}_K)} = \frac{\det(I)}{\Delta_K},$$

since the discriminant Δ_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 5.22, we have:

$$\Delta(\alpha_1, \ldots, \alpha_d) = \det(A)^2 \Delta(\omega_1, \ldots, \omega_d) = \det(A)^2 \Delta_K,$$

and thus

$$\mathbb{N}_K(I) = |\det(A)| = \left|\frac{\Delta(\alpha_1, \ldots, \alpha_d)}{\Delta_K}\right|^{1/2}.$$

□

COROLLARY 5.4.6. *If $I = \langle\alpha\rangle \subseteq \mathcal{O}_K$ is a principal ideal, then*

$$\mathbb{N}_K(I) = |\mathbb{N}_K(\alpha)|.$$

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

$$\begin{aligned}\Delta(\alpha\omega_1, \ldots, \alpha\omega_d) &= (\det(\sigma_i(\alpha)\sigma_i(\omega_j))_{1\le i,j\le d})^2 \\ &= \left(\prod_{i=1}^{d}\sigma_i(\alpha)\det(\sigma_i(\omega_j))_{1\le i,j\le d}\right)^2 = \mathbb{N}_K(\alpha)^2\Delta_K.\end{aligned}$$

Then, by Lemma 5.4.5,

$$\mathbb{N}_K(I) = \left|\frac{\Delta(\alpha\omega_1, \ldots, \alpha\omega_d)}{\Delta_K}\right|^{1/2} = |\mathbb{N}_K(\alpha)|.$$

□

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

$$B = \alpha^{-1} I = \left\{ \alpha^{-1} x : x \in I \right\},$$

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 $\mathfrak{F}_K$ be the set of all fractional ideals in K. We can define a commutative multiplication operation on $\mathfrak{F}_K$: for $B = \alpha^{-1} I$ and $C = \beta^{-1} J$ in $\mathfrak{F}_K$,

$$BC = \{bc : b \in B, c \in C\} = (\alpha\beta)^{-1} IJ \in \mathfrak{F}_K,$$

since $\alpha\beta \in \mathcal{O}_K$ and IJ is again an ideal in $\mathcal{O}_K$. Notice that for any $B = \alpha^{-1} I \in \mathfrak{F}_K$,

$$B\mathcal{O}_K = \left\{ \alpha^{-1} xy : x \in I, y \in \mathcal{O}_K \right\} = \left\{ \alpha^{-1} z : z \in I \right\} = B.$$

We will now prove an important theorem.

THEOREM 5.5.1. *The set of fractional ideals $\mathfrak{F}_K$ 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

$$I' = \{x \in K : xI \subseteq \mathcal{O}_K\}, \tag{5.9}$$

and for each fractional ideal $B = \alpha^{-1} I \in \mathfrak{F}_K$, define $B' = \alpha I'$. Then each such B' is again a fractional ideal, i.e. $B' \in \mathfrak{F}_K$ for every $B \in \mathfrak{F}_K$ (Probelm 5.29). It is easy to see that $\mathcal{O}_K \subseteq B'$ for each $B' \in \mathfrak{F}_K$. Further, $\mathcal{O}_K' = \mathcal{O}_K$ (Problem 5.28). We will now prove that B' is the inverse of B for every $B \in \mathfrak{F}_K$. 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 $M \subset \mathcal{O}_K$ 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 $P_1, \ldots, P_r \subset \mathcal{O}_K$ 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 $J_1, J_2 \subset \mathcal{O}_K$ such that $J_1 J_2 \subseteq I$ while J_1, J_2 are not in I. Let

$$A_1 = I + J_1, \ A_2 = I + J_2,$$

then $I \subsetneq A_1, A_2$, while

$$A_1 A_2 = (I + J_1)(I + J_2) = I^2 + I(J_1 + J_2) + J_1 J_2 \subseteq I.$$

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 $P_1, \dots, P_r$, $Q_1, \dots, Q_s$ 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. □

Back to our proof, let $0 \neq a \in M$, then the principal ideal $\langle a \rangle \subseteq M$. Let r be the smallest integer for which exists a collection of prime ideals $P_1, \dots, P_r \in \mathcal{O}_K$ with the product $P_1 \cdots P_r \subseteq \langle a \rangle$. Since M is maximal (hence prime) at least one of $P_1, \dots, P_r$ 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,

$$P_2 \cdots P_r \not\subseteq \langle a \rangle,$$

meaning that there exists some $b \in P_2 \cdots P_r$ which is not in $\langle a \rangle$. This being said, $bP_1 = bM \subseteq \langle a \rangle$, i.e. $ba^{-1}M \subseteq \mathcal{O}_K$, hence $ba^{-1} \in M'$. But since $b \notin \langle a \rangle = a\mathcal{O}_K$, 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 $a_1, \dots, a_n$ be a generating set for I, which is also an integral basis. In other words, for every $x \in I$, there exist $b_1, \dots, b_n \in \mathbb{Z}$ such that

$$x = b_1 a_1 + \cdots + b_n a_n.$$

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

$$B\boldsymbol{a} = \theta \boldsymbol{a},$$

where $B = (b_{kj})_{1 \leq k,j \leq n}$ is an integer matrix and $\boldsymbol{a} = (a_1, \dots, a_n)^\top$ is a vector with coordinates in I. Then θ is an eigenvalue of B, which is a root of the monic polynomial

$$\det \begin{pmatrix} b_{11} - x & \dots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{n1} & \dots & b_{nn} - x \end{pmatrix}$$

with integer coefficients. Hence θ 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 II' 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 $II' = \mathcal{O}_K$ for any nonzero ideal $I \subseteq \mathcal{O}_K$. Suppose not, then $II' \subsetneq \mathcal{O}_K$. In fact, let us take I to be a maximal element in the set of all ideals J in $\mathcal{O}_K$ such that $JJ' \subsetneq \mathcal{O}_K$. Let M be a maximal ideal in $\mathcal{O}_K$ containing I. Then we know that

$$\mathcal{O}_K \subsetneq M' \subseteq I',$$

hence

$$I = I\mathcal{O}_K \subsetneq IM' \subseteq II' \subsetneq \mathcal{O}_K.$$

Since IM' is an ideal in $\mathcal{O}_K$, the maximality condition on I implies that

$$(IM')(IM')' = I(M'(IM')') = \mathcal{O}_K.$$

This means that $M'(IM')' \subseteq I'$, hence

$$\mathcal{O}_K = I(M'(IM')') \subseteq II' \subsetneq \mathcal{O}_K,$$

which is a contradiction. Hence $II' = \mathcal{O}_K$.

Step 5. It now easily follows that $BB' = \mathcal{O}_K$ 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

$$BB' = (\alpha^{-1}I)(\alpha I') = II' = \mathcal{O}_K.$$

This completes the argument, showing that $\mathfrak{F}_K$ is the abelian group of fractional ideals in K. □

This result and its proof have some important implications, the first of which is a certain weaker analogue of unique factorization in $\mathcal{O}_K$.

THEOREM 5.5.3. *Every nonzero ideal in $\mathcal{O}_K$ 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 $\mathcal{O}_K$ which do not have such a factorization. Since $\mathcal{O}_K$ is noetherian, S has a maximal element, call it I. Then I itself is not a prime ideal, hence not a maximal ideal in $\mathcal{O}_K$. Let M be a maximal ideal in $\mathcal{O}_K$ such that $I \subset M$. Then by the argument above (Step 4 in the proof of Theorem 5.5.1),

$$I \subsetneq IM' \subsetneq \mathcal{O}_K.$$

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 $P_1, \ldots, P_r \subset \mathcal{O}_K$. Then, since $M'M = \mathcal{O}_K$,

$$P_1 \cdots P_r M = (IM')M = I(M'M) = I,$$

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 $I \subset \mathcal{O}_K$. 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 $P_1 = Q_i$ 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. □

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 5.30). In other words, we want to prove that

$$|\mathcal{O}_K/IP| = |\mathcal{O}_K/I| \cdot |\mathcal{O}_K/P|. \tag{5.10}$$

First let us prove that

$$|\mathcal{O}_K/IP| = |\mathcal{O}_K/I| \cdot |I/IP|. \tag{5.11}$$

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 5.31). By First Isomorphism Theorem for groups,

$$(\mathcal{O}_K/IP) \,/\, \operatorname{Ker}(\phi) \cong \phi(\mathcal{O}_K/IP) = \mathcal{O}_K/I,$$

and since these groups are finite, we have

$$|\mathcal{O}_K/IP| = |\operatorname{Ker}(\phi)| \cdot |\mathcal{O}_K/I|.$$

Notice that $\operatorname{Ker}(\phi)$ consists of all cosets of IP which are in I, i.e. $\operatorname{Ker}(\phi) = I/IP$. This establishes (5.11).

Next we prove that

$$|I/IP| = |\mathcal{O}_K/P|. \tag{5.12}$$

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

$$I'(IP) = P \subseteq I'B \subseteq I'I = \mathcal{O}_K,$$

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 $\mathcal{O}_K/\operatorname{Ker}(f_\alpha) \cong I/IP$. Clearly, $\operatorname{Ker}(f_\alpha) \neq \mathcal{O}_K$, since $I \neq IP$. On the other hand, for every $x \in P$,

$$f_\alpha(x) = \alpha x + IP = 0$$

in I/IP, since $\alpha x \in IP$ (α is in I and x is in P). Therefore $P \subseteq \mathrm{Ker}(f_\alpha)$, and since P is a maximal ideal, we have $P = \mathrm{Ker}(f_\alpha)$. Therefore $\mathcal{O}_K/P \cong I/IP$, 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 $I \subseteq \mathcal{O}_K$ be a nonzero ideal. Then:*

1. *If $\mathbb{N}_K(I)$ is a prime in $\mathbb{Z}$, then I is a prime ideal in $\mathcal{O}_K$.*
2. *$\mathbb{N}_K(I) \in I$.*
3. *If $I \subset \mathcal{O}_K$ is a prime ideal, then $\mathbb{N}_K(I) = p^m$ 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 $A, B \subseteq \mathcal{O}_K$. Then, by multiplicativity of the norm,

$$\mathbb{N}_K(I) = \mathbb{N}_K(A)\mathbb{N}_K(B) = p \in \mathbb{Z},$$

where p is a prime number. Hence either $\mathbb{N}_K(A)$ or $\mathbb{N}_K(B)$ must be equal to 1, say it is $\mathbb{N}_K(A)$. This means that $|\mathcal{O}_K/A| = 1$, and so $A = \mathcal{O}_K$, hence

$$I = AB = \mathcal{O}_K B = B.$$

Therefore I does not have any nontrivial factorization, hence it is a prime ideal.

To prove part (2), let $x + I \in \mathcal{O}_K/I$. Since $\mathcal{O}_K/I$ is an additive abelian group of order $\mathbb{N}_K(I)$, we must have

$$\mathbb{N}_K(I)(x + I) = \mathbb{N}_K(I)x = I,$$

meaning that $\mathbb{N}_K(I)x \in I$ for every $x \in \mathcal{O}_K$. Then take $x = 1$, and we see that $\mathbb{N}_K(I) \in I$.

To prove part (3), assume that

$$\mathbb{N}_K(I) = p_1^{m_1} \cdots p_r^{m_r}$$

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 $I = \mathcal{O}_K$. However, a prime ideal has to be proper. Thus p_i is the only rational integer prime contained in I, which means that $I \mid p_i\mathcal{O}_K$ and $I \nmid q\mathcal{O}_K$ for any prime $q \neq p_i$. Therefore

$$\mathbb{N}_K(I) | \mathbb{N}_K(p_i) = \mathbb{N}_K(p_i) = \prod_{j=1}^{d} \sigma_j(p_i) = p_i^d,$$

where $\sigma_1, \ldots, \sigma_d$ are embeddings of K. Therefore $\mathbb{N}_K(I) = p_i^m$ 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 $\mathcal{O}_K$. The following are true:*

(1) *Let $I \subseteq \mathcal{O}_K$ be an ideal. There exist only finitely many ideals $J \subseteq \mathcal{O}_K$ such that $J \mid I$.*
(2) *Let $m \in \mathbb{Z}$, $m \neq 0$. There exist only finitely many ideals $I \subseteq \mathcal{O}_K$ such that $m \in I$.*
(3) *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 α 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 $\alpha I' \not\subseteq P_i$ for all $1 \leq i \leq r$: if $\alpha I' \not\subseteq P_i$, then

$$P_i \subsetneq \alpha I' + P_i \subseteq \mathcal{O}_K,$$

which by maximality of P_i means that $\alpha I' + P_i = \mathcal{O}_K$. Notice that

$$\alpha I' \not\subseteq P_i \Leftrightarrow P_i \nmid \alpha I' \Leftrightarrow IP_i \nmid \alpha\mathcal{O}_K \Leftrightarrow \alpha \notin IP_i$$

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 $1 \leq i \leq r$, $IP_i \subsetneq I$. 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 $I_m P_m = IJ$. 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 α_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 5.32: 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 $m = \mathbb{N}_K(P)$. 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 β be a nonzero element of I. Then there exists an element $\alpha \in I$ such that I is generated by the pair α, β.*

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 α, β. □

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 $\sigma_1, \dots, \sigma_d : K \hookrightarrow \mathbb{C}$ be its embeddings. We will distinguish between *real* and *complex* embeddings: σ_i is said to be real if the the field $\sigma_i(K)$ is contained in $\mathbb{R}$, and complex otherwise. Notice that complex embeddings come in conjugate pairs: if σ_i is complex, then there is it conjugate embedding $\bar{\sigma}_i$ given by

$$\bar{\sigma}_i(x) := \overline{\sigma_i(x)}$$

for every $x \in K$. Let us order the embeddings

$$\sigma_1, \dots, \sigma_r, \sigma_{r+1}, \bar{\sigma}_{r+1}, \dots, \sigma_{r+s}, \bar{\sigma}_{r+s},$$

where $\sigma_1, \dots, \sigma_r$ are real and $\sigma_{r+1}, \bar{\sigma}_{r+1}, \dots, \sigma_{r+s}, \bar{\sigma}_{r+s}$ are complex. Then $d = r + 2s$, and we can define a map

$$\Sigma := (\sigma_1, \dots, \sigma_r, \sigma_{r+1}, \dots, \sigma_{r+s}) : K \hookrightarrow \mathbb{R}^r \times \mathbb{C}^s,$$

given by $\Sigma(x) = (\sigma_1(x), \dots, \sigma_r(x), \sigma_{r+1}(x), \dots, \sigma_{r+s}(x))$ for every $x \in K$. We can identify $\mathbb{C}^s$ with $\mathbb{R}^{2s}$, thinking of $\Sigma(x)$ as

$$(\sigma_1(x), \dots, \sigma_r(x), \Re(\sigma_{r+1}(x)), \Im(\sigma_{r+1}(x)) \dots, \Re(\sigma_{r+s}(x)), \Im(\sigma_{r+s}(x))).$$

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 $x_1, \dots, x_d$ be a $\mathbb{Z}$-basis for M. Then $\Sigma(M)$ is a lattice of rank d in $\mathbb{R}^d$ with determinant*

$$\det(\Sigma(M)) = 2^{-s} \left|\det(A\ C)\right|, \tag{5.13}$$

where A is an

$$A = (\sigma_i(x_j))_{1 \le i \le r, 1 \le j \le d},\ C = (\sigma_{r+i}(x_j)\ \bar{\sigma}_{r+i}(x_j))_{1 \le i \le s, 1 \le j \le d}. \tag{5.14}$$

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 $d \times 2s$

$$B = (\Re(\sigma_{r+i}(x_j))\ \Im(\sigma_{r+i}(x_j)))_{1 \le i \le s, 1 \le j \le d}$$

For any complex number z,

$$\Re(z) = \frac{1}{2}(z + \bar{z}),\ \Im(z) = \frac{1}{2i}(z - \bar{z}).$$

With this in mind, it is easy to see that the matrix B is column-equivalent to the matrix

$$\left(\frac{1}{2}\left(\sigma_{r+i}(x_j) + \bar{\sigma}_{r+i}(x_j)\right)\ \frac{1}{2i}\left(\sigma_{r+i}(x_j) - \bar{\sigma}_{r+i}(x_j)\right)\right)_{1 \le i \le s, 1 \le j \le d}.$$

This means that

$$\det(\Sigma(M)) = \left|\left(\frac{1}{2}\right)^{2s}\left(\frac{1}{i}\right)^{s}\det(A\ B')\right| = 2^{-2s}\left|\det(A\ B')\right|,$$

where

$$B' = \left((\sigma_{r+i}(x_j) + \bar{\sigma}_{r+i}(x_j)) \ \ (\sigma_{r+i}(x_j) - \bar{\sigma}_{r+i}(x_j))\right)_{1\le i\le s, 1\le j\le d},$$

a matrix column-equivalent to

$$B'' = \left(2\sigma_{r+i}(x_j) \ \ \bar{\sigma}_{r+i}(x_j)\right)_{1\le i\le s, 1\le j\le d}.$$

Thus

$$\det(\Sigma(M)) = 2^{-2s}\left|\det(A\ B'')\right| = 2^{-s}\left|\det(A\ C)\right|,$$

and so we have (5.13). Notice that this determinant cannot be equal to 0 (Problem 5.33), and hence $\Sigma(M)$ is a lattice of full rank in $\mathbb{R}^d$. □

COROLLARY 5.7.2. *If $I \subseteq \mathcal{O}_K$ is an ideal, then*

$$\det(\Sigma(I)) = 2^{-s}|\Delta_K|^{1/2}\mathbb{N}_K(I).$$

In particular,

$$\det(\Sigma(\mathcal{O}_K)) = 2^{-s}|\Delta_K|^{1/2}.$$

PROOF. By Lemma 5.4.5,

$$\mathbb{N}_K(I) = \left|\frac{\Delta(\alpha_1,\dots,\alpha_d)}{\Delta_K}\right|^{1/2},$$

where $\alpha_1,\dots,\alpha_d$ is a $\mathbb{Z}$-basis for I. Then, by Lemma 5.7.1 (and in the notation of that lemma),

$$\det(\Sigma(I)) = 2^{-s}\left|\det(A\ C)\right| = 2^{-s}\left|\Delta(\alpha_1,\dots,\alpha_d)\right|^{1/2} = 2^{-s}\mathbb{N}_K(I)\left|\Delta_K\right|^{1/2}.$$

If $I = \mathcal{O}_K$, then $\mathbb{N}_K(I) = 1$. □

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 $c_1,\dots,c_r,d_1,\dots,d_s \in \mathbb{R}_{>0}$ be such that*

$$c_1\cdots c_r d_1\cdots d_s > \left(\frac{4}{\pi}\right)^s D.$$

Let

$$\begin{aligned} X \ = \ & \big\{\boldsymbol{x} = (x_1,\dots,x_r,y_{11},y_{12},\dots,y_{s1},y_{s2}) \in \mathbb{R}^d : \\ & |x_i| < c_i \ \forall\ 1\le i\le r,\ y_{j1}^2 + y_{j2}^2 < d_j \ \forall\ 1\le j\le s\big\}. \end{aligned}$$

Then there exists $\mathbf{0} \neq \boldsymbol{x} \in X \cap \Lambda$.

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 $1 \le i \le r$ and circles of radius $\sqrt{d_j}$ for $1 \le j \le s$. Therefore X is convex $\mathbf{0}$-symmetric and its volume is

$$\mathrm{Vol}(X) = (2c_1)\cdots(2c_r)\cdot(\pi d_1)\cdots(\pi d_s) > 2^r\pi^s\left(\frac{4}{\pi}\right)^s D = 2^d D.$$

Then Theorem 1.4.2 (see also Problem 1.23) guarantees that there exists a nonzero point $\boldsymbol{x} \in X \cap \Lambda$. □

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*

$$|\mathbb{N}_K(\alpha)| \leq \left(\frac{2}{\pi}\right)^s \mathbb{N}_K(I) \left|\Delta_K\right|^{1/2},$$

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

$$c_1 \cdots c_r \cdot d_1 \cdots d_s = \left(\frac{2}{\pi}\right)^s \mathbb{N}_K(I) \left|\Delta_K\right|^{1/2} + \varepsilon.$$

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

$$\begin{aligned} |\mathbb{N}_K(\alpha)| &= \prod_{i=1}^{r} |\sigma_i(\alpha)| \times \prod_{j=1}^{s} |\sigma_{r+j}(\alpha)\bar{\sigma}_{r+j}(\alpha)| = \prod_{i=1}^{r} |\sigma_i(\alpha)| \times \prod_{j=1}^{s} |\sigma_{r+j}(\alpha)|^2 \\ &< c_1 \cdots c_r \cdot d_1 \cdots d_s = \left(\frac{2}{\pi}\right)^s \mathbb{N}_K(I) |\Delta_K|^{1/2} + \varepsilon. \end{aligned}$$

Since Λ is discrete in $\mathbb{R}^d$, there are only finitely such α 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. □

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 5.34): 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

$$\mathrm{Cl}(K) := \mathfrak{F}_K / \mathfrak{P}_K.$$

Elements of this group are called *ideal classes*, and the order $h_K := |\mathrm{Cl}(K)|$ 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 5.35). 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 $\mathrm{Cl}(K)$ being trivial and hence having order 1. □

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* $\mathrm{Cl}(K)$ *contains an ideal $I \subseteq \mathcal{O}_K$ with*

$$\mathbb{N}_K(I) \leq \left(\frac{2}{\pi}\right)^s |\Delta_K|^{1/2}. \tag{5.15}$$

Proof. Since every ideal class contains an ideal (Problem 5.36), 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

$$|\mathbb{N}_K(\alpha)| \leq \left(\frac{2}{\pi}\right)^s \mathbb{N}_K(M)|\Delta_K|^{1/2}.$$

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,

$$\mathbb{N}_K(I)\mathbb{N}_K(M) = \mathbb{N}_K(\langle\alpha\rangle) = |\mathbb{N}_K(\alpha)| \leq \left(\frac{2}{\pi}\right)^s \mathbb{N}_K(M)|\Delta_K|^{1/2},$$

which means that

$$\mathbb{N}_K(I) \leq \left(\frac{2}{\pi}\right)^s |\Delta_K|^{1/2}.$$

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. □

Proof of Theorem 5.8.2. By Lemma 5.8.3 we know that every ideal class in $\mathrm{Cl}(K)$ 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. □

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. Dirichlet's unit theorem

To finish our discussion of factorization in rings of algebraic integers, we turn our attention to units. While the class number can be used as an indicator of whether the factorization into irreducibles is unique, even when it is it only is unique up to multiplication by units. This prompts a natural question: how big is the unit group $\mathcal{O}_K^\times$ for a given number field K? This question is answered by the famous theorem of Dirichlet, although the proof we present here once again uses techniques from Minkowski's geometry of numbers. Our main arguments in this section are partially based on the exposition of [**Sam70**].

We start with setting up some notation. Define $z_n = e^{2\pi i/n}$ for $n \in \mathbb{Z}_{>1}$: it is an algebraic integer, since it is a root of the monic polynomial $x^n - 1$. Notice that the set

$$\mu_n = \left\{ z_n^k : k \in \mathbb{Z} \right\}$$

contains precisely n distinct elements, specifically $z_n^k = e^{2k\pi i/n}$, $0 \leq k \leq n-1$. Further, μ_n is a cyclic group under multiplication of complex numbers, and hence is isomorphic to $\mathbb{Z}/n\mathbb{Z}$ (Problem 5.37). Thus μ_n is called the group of n-th *roots of unity*, and its generators are called n-th *primitive roots of unity*: clearly, z_n is one of them, and the others are precisely elements of the form z_n^k where $1 \leq k \leq n-1$ and $\gcd(k, n) = 1$ (Problem 5.38).

Let K be a number field of degree $d = r + 2s$, where r is the number of real embeddings and s is the number of conjugate pairs of complex embeddings, as usual. Let $G_K \subset \mathcal{O}_K$ be the set of all roots of unity contained in K. It is easy to see that $G_K \subset \mathcal{O}_K^\times$ is a group (Problem 5.39). In fact, more is true.

FACT 5.9.1. *The group G_K is a finite cyclic group generated by $e^{2\pi i/n}$, where $n = \max\{m \geq 1 : e^{2\pi i/m} \in K\}$.*

We do not prove this fact here. Roots of unity and number fields they generate (called *cyclotomic fields*) play an important role in algebraic number theory; while we do not develop this theory here, we will refer the interested reader to [**Lan94**] or [**ST02**] for some further information. With this notation, we have the following theorem.

THEOREM 5.9.1. *[Dirichlet, 1846] Let K be a number field of degree $d = r + 2s$ as above, and define $t = r + s - 1$. There exist $u_1, \ldots, u_t \in \mathcal{O}_K^\times$ such that for every $u \in \mathcal{O}_K^\times$,*

$$u = z u_1^{n_1} \cdots u_t^{n_t},$$

for some $n_1, \ldots, n_t \in \mathbb{Z}_{\geq 0}$ and $z \in G_K$.

The elements $u_1, \ldots, u_t$ in Theorem 5.9.1 are called a *system of fundamental units* in K. To prove this theorem, we need to introduce the *logarithmic space* corresponding to K. Let

$$\Sigma := (\sigma_1, \ldots, \sigma_r, \sigma_{r+1}, \ldots, \sigma_{r+s}) : K \hookrightarrow \mathbb{R}^r \times \mathbb{C}^s \cong \mathbb{R}^{r+2s} = \mathbb{R}^d$$

be the Minkowski embedding of K, and define the logarithmic map $\ell : K^\times \to \mathbb{R}^{r+s}$ given by

$$\ell(x) = \log|\Sigma(x)| := (\log|\sigma_1(x)|, \ldots, \log|\sigma_r(x)|, \log|\sigma_{r+1}(x)|, \ldots, \log|\sigma_{r+s}(x)|)$$

for every nonzero $x \in K$. Then

$$\ell(xy) = \ell(x) + \ell(y)$$

for all nonzero $x, y \in K$, and thus ℓ is a homomorphism between multiplicative group $(K^\times, \times)$ and additive group $(\mathbb{R}^{r+s}, +)$. Let us determine the kernel of this homomorphism. First notice that $\ell(x) = 0$ if and only if $|\sigma_j(x)| = 1$ for every $1 \leq j \leq r+s$. Now, any complex number x can be written as $x = ae^{i\theta}$ for some $a \in \mathbb{R}_{>0}$ and $\theta \in [0, 2\pi)$, and $|x| = 1$ if and only if $a = 1$. Thus $x \in \mathrm{Ker}(\ell)$ if and only if $x = e^{i\theta}$ for some $\theta \in [0, 2\pi)$ and all of its conjugates are of the same form for some values of θ. Algebraic integers that have this property are precisely roots of unity: a proof of this fact can be found in any abstract algebra book, for instance [**DF03**]. Thus we see that $\mathrm{Ker}(\ell)$ consists precisely of the roots of unity contained in K, i.e.

$$\mathrm{Ker}(\ell) = G_K, \tag{5.16}$$

which is a finite cyclic group by Fact 5.9.1.

Define

$$W := \left\{ \boldsymbol{y} \in \mathbb{R}^{r+s} : \sum_{j=1}^{r} y_j + 2 \sum_{j=r+1}^{r+s} y_j = 0 \right\}.$$

LEMMA 5.9.2. *$\ell(\mathcal{O}_K^\times)$ is a discrete subgroup of W.*

PROOF. First notice that W and $\ell(\mathcal{O}_K^\times)$ are both additive groups in $\mathbb{R}^{r+s}$. By Corollary 5.4.2 we know that $x \in \mathcal{O}_K^\times$ if and only if

$$|\mathbb{N}_K(x)| = \prod_{j=1}^{r+2s} |\sigma_j(x)| = 1,$$

where $|\sigma_j(x)| = |\bar{\sigma}_j(x)|$ for every $r+1 \leq r+2s$: this condition is equivalent to

$$\log |\mathbb{N}_K(x)| = \sum_{j=1}^{r} \log |\sigma_j(x)| + 2 \sum_{j=r+1}^{r+s} \log |\sigma_j(x)| = 0,$$

where the summands are precisely the coordinates of $\ell(x)$. Hence $\ell(\mathcal{O}_K^\times) \subseteq W$.

Let us now prove that $\ell(\mathcal{O}_K^\times)$ is discrete in W. Suppose not, then there exists

$$\boldsymbol{y} = \ell(x) = (\log |\sigma_1(x)|, \ldots, \log |\sigma_r(x)|, \log |\sigma_{r+1}(x)|, \ldots, \log |\sigma_{r+s}(x)|)$$

for some $x \in \mathcal{O}_K^\times$ such that $\|\boldsymbol{y}\| < \varepsilon$ for any $\varepsilon > 0$. Taking sufficiently small ε, this produces an element $x \in \mathcal{O}_K^\times$ with $|\mathbb{N}_K(x)| < 1$. This contradicts the fact that $\mathbb{N}_K(x) \in \mathbb{Z}$. Hence $\ell(\mathcal{O}_K^\times)$ is discrete in W. □

LEMMA 5.9.3. *$\ell(\mathcal{O}_K^\times)$ is a lattice of rank $t = r+s-1$ in W.*

PROOF. Lemma 5.9.2 implies that $\ell(\mathcal{O}_K^\times)$ is a lattice in W. Since $\dim_{\mathbb{R}} W = t$, we only need to prove that $\ell(\mathcal{O}_K^\times)$ has full rank in W, which is equivalent to saying that it does not lie in any proper subspace of W. Without loss of generality, we can identify W with $\mathbb{R}^t$. Since any subspace is cut out by linear forms, we need to show that for any linear form

$$f(\boldsymbol{y}) = c_1 y_1 + \cdots + c_t y_t \tag{5.17}$$

there exists $x \in \mathcal{O}_K^\times$ such that $f(\ell(x)) \neq 0$. Let us pick a real number

$$\alpha \geq \left(\frac{2}{\pi}\right)^s |\Delta_K|^{1/2}.$$

Then for any t-tuple $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_t)$ of positive real numbers, pick $\lambda_{t+1} > 0$ such that

$$\prod_{i=1}^{r} \lambda_i \prod_{j=r+1}^{r+s} \lambda_j^2 = \alpha.$$

Let $X_{\boldsymbol{\lambda}}$ be the set X as in Lemma 5.7.3 with $c_i = \lambda_i$ for $1 \leq i \leq r$ and $d_j = \lambda_j^2$ for $r+1 \leq j \leq r+s$. By Corollary 5.7.4, there exists a nonzero element $x_{\boldsymbol{\lambda}} \in \mathcal{O}_K \cap X_{\boldsymbol{\lambda}}$ such that

$$1 \leq |\mathbb{N}_K(x_{\boldsymbol{\lambda}})| \leq \alpha.$$

On the other hand, let $\lambda_{j+s} = \lambda_j$ for each $r+1 \leq j \leq r+s$. Then for every $1 \leq i \leq d$,

$$|\sigma_i(x_{\boldsymbol{\lambda}})| = |\mathbb{N}_K(x_{\boldsymbol{\lambda}})| \prod_{j \neq i} |\sigma_j(x_{\boldsymbol{\lambda}})|^{-1} \geq \prod_{j \neq i} |\lambda_j|^{-1} = \lambda_i \alpha^{-1}.$$

Thus, for every $1 \leq i \leq d$, we have $\lambda_i \alpha^{-1} \leq |\sigma_i(x_{\boldsymbol{\lambda}})| \leq \lambda_i$, which can be re-written as

$$1 \leq \frac{\lambda_i}{|\sigma_i(x_\lambda)|} \leq \alpha.$$

Taking logarithms, we have

$$0 \leq \log \lambda_i - \log |\sigma_i(x_\lambda)| \leq \log \alpha. \tag{5.18}$$

Now, let f be a linear form as in (5.17) with coefficients $c_1, \ldots, c_t$. Multiplying inequalities in (5.18) by c_i, taking absolute values and summing over all $1 \leq i \leq t$, we obtain

$$\left| f(\ell(x_{\boldsymbol{\lambda}})) - \sum_{i=1}^{t} c_i \log \lambda_i \right| \leq \log \alpha \sum_{i=1}^{t} |c_i|$$

Let $\beta > \log \alpha \sum_{i=1}^{t} |c_i|$, and for each positive integer h pick $\boldsymbol{\lambda}(h) = (\lambda_1(h), \ldots, \lambda_t(h))$ be such that

$$\sum_{i=1}^{t} c_i \log \lambda_i(h) = 2\beta h.$$

Then for the corresponding $x_h := x_{\boldsymbol{\lambda}(h)}$, we have

$$|f(\ell(x_h)) - 2\beta h| < \beta,$$

and so

$$(2h-1)\beta < f(\ell(x_h)) < (2h+1)\beta.$$

Notice that if $h_1 + 1 \leq h_2$, then $(2h_1+1)\beta \leq (2h_2-1)\beta$, and therefore the numbers $f(\ell(x_h))$ are distinct for different values of h. On the other hand, consider the ideals of the form $I_h := \mathcal{O}_K x_h$: for each such ideal,

$$\mathbb{N}_K(I_h) = \mathbb{N}(x_h) \leq \alpha,$$

and so there can be only finitely many such ideals, by Lemma 5.6.3. Hence there must exist some two distinct algebraic integers $x_{h_1}, x_{h_2} \in \mathcal{O}_K$ such that

$$\mathcal{O}_K x_{h_1} = \mathcal{O}_K x_{h_2}.$$

Hence there are elements $u_1, u_2 \in \mathcal{O}_K$ such that $x_{h_1} = u_1 x_{h_2}$ and $x_{h_2} = u_2 x_{h_1}$, i.e.

$$x_{h_1} = u_1 x_{h_2} = u_1 u_2 x_{h_1},$$

and so $u_1, u_2 \in \mathcal{O}_K^\times$. Thus we have

$$f(\ell(u_1)) = f(\ell(x_{h_1})) - f(\ell(x_{h_2})) \neq 0,$$

and this completes the proof. □

We can now prove Dirichlet's theorem.

PROOF OF THEOREM 5.9.1. Notice that any lattice of rank t is isomorphic to $\mathbb{Z}^t$ as abelian groups. Combining Lemma 5.9.3 with (5.16), we see that

$$\mathcal{O}_K^\times \cong G_K \times \mathbb{Z}^t,$$

as abelian groups, where $\mathcal{O}^\times$ and G_K are multiplicatively written and $\mathbb{Z}^t$ is additive. Since every element of $G_K \times \mathbb{Z}^t$ can be written as

$$(z, n_1\boldsymbol{e}_1, \dots, n_t\boldsymbol{e}_t)$$

with $\boldsymbol{e}_1, \dots, \boldsymbol{e}_t$ the standard basis vectors for $\mathbb{Z}^t$, the corresponding element of $\mathcal{O}_K^\times$ mapped to it under this isomorphism can be written as

$$z u_1^{n_1} \cdots u_t^{n_t},$$

where the units $u_1, \dots, u_t \in \mathcal{O}_K^\times$ are preimages of $\boldsymbol{e}_1, \dots, \boldsymbol{e}_t$. This completes the proof. □

5.10. 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 $K \subset \mathbb{C}$ 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 $K \subset \mathbb{C}$ be a field and $\alpha \in \mathbb{C}$. Prove that $K[\alpha] \subset \mathbb{C}$ 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 φ is a ring homomorphism. Describe its kernel and specify under what conditions on α 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 $K \subset \mathbb{C}$ 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 $K = \mathbb{Q}(\sqrt{2}, \sqrt{3})$. Find an integral primitive element for K, i.e. an algebraic integer α so that $K = \mathbb{Q}(\alpha)$.*

PROBLEM 5.12. *Consider the number field $K = \mathbb{Q}(\sqrt{3}, \sqrt{-3})$.*

1. *Determine the degree of K over $\mathbb{Q}$. Prove your answer.*
2. *Find a primitive element for K over $\mathbb{Q}$. Prove your answer.*
3. *Find the minimal polynomial of the primitive element you found in part b. Prove your answer.*

(4) *Is* 2 *a prime in* K*? Prove your answer.*

PROBLEM 5.13. *Let* α *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 5.14. *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 5.15. *Let* A *and* B *be subrings of the same ring* R*. Prove that* $A \cap B$ *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 5.16. *Prove that each* σ_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 5.17. *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 5.18. *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 5.19. *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 5.20. *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 5.21. *Let* $K = \mathbb{Q}(\sqrt{-13})$*. Is* $\mathcal{O}_K$ *a PID (principal ideal domain)?*

PROBLEM 5.22. *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* α *to the* β *basis. Prove that*

$$\Delta(\beta_1, \ldots, \beta_d) = \det(C)^2 \Delta(\alpha_1, \ldots, \alpha_d).$$

PROBLEM 5.23. *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 5.24. *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 $A \in \mathrm{GL}_d(\mathbb{Z})$ between them.*

PROBLEM 5.25. *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 5.26. *Let K be a number field, $\alpha, \beta \in K$ and $a, b, c \in \mathbb{Q}$. Prove that*

$$\mathbb{N}_K(c\alpha\beta) = c\ \mathbb{N}_K(\alpha)\mathbb{N}_K(\beta),\ \mathrm{Tr}_K(a\alpha + b\beta) = a\,\mathrm{Tr}_K(\alpha) + b\,\mathrm{Tr}_K(\beta).$$

PROBLEM 5.27. *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 5.28. *For an ideal $I \subseteq \mathcal{O}_K$, let I' be as in* (5.9)*. Prove that $\mathcal{O}'_K = \mathcal{O}_K$.*

PROBLEM 5.29. *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 5.30. *Complete the proof of Lemma 5.6.1 by showing that if*

$$\mathbb{N}_K(IJ) = \mathbb{N}_K(I)\mathbb{N}_K(J)$$

when J is a prime ideal, then this is true for all ideals J.

PROBLEM 5.31. *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 5.32. *Let R be a principal ideal domain. Prove that every $\alpha \in R$ has a unique factorization into irreducibles.*

PROBLEM 5.33. *Prove that* $\det(\Sigma(M))$ *in* (5.13) *of Lemma 5.7.1 cannot be equal to* 0.

PROBLEM 5.34. *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 5.35. *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 5.36. *Prove that every ideal class in* $\mathrm{Cl}(K)$ *contains an ideal* $I \subseteq \mathcal{O}_K$.

PROBLEM 5.37. *Let* $n \geq 1$ *be an integer, and let*

$$\mu_n = \left\{ e^{2k\pi i/n} : k \in \mathbb{Z} \right\}.$$

Prove that μ_n *contains precisely* n *distinct elements, specifically* $e^{2k\pi i/n}$, $0 \leq k \leq n-1$. *Further, prove that* μ_n *is a cyclic group under multiplication of complex numbers, and hence is isomorphic to* $\mathbb{Z}/n\mathbb{Z}$.

PROBLEM 5.38. *Let notation be as in Problem 5.37 above. Prove that an element* z *is a generator of the cyclic multiplicative group* μ_n *if and only if*

$$z = e^{2k\pi i/n} \text{ for some } 1 \leq k \leq n-1 \text{ such that } \gcd(k,n) = 1.$$

PROBLEM 5.39. *Let* K *be a number field and* $G_K \subset \mathcal{O}_K$ *be the set of all roots of unity contained in* K. *Prove that* $G_K \subset \mathcal{O}_K^\times$, *i.e. every root of unity in* K *is a unit in* $\mathcal{O}_K$. *Further, prove that* G_K *is a group under multiplication of complex numbers.*

PROBLEM 5.40. *Let* K *be a number with the ring of integers* $\mathcal{O}_K$. *Use finiteness of the class number to prove that there exists a number field* L *containing* K *with ring of integer* $\mathcal{O}_L$ *such that for every ideal* I *in* $\mathcal{O}_K$, $\mathcal{O}_L I$ *is a principal ideal.*

PROBLEM 5.41. *Let* D *be a squarefree integer and let* $K = \mathbb{Q}(\sqrt{D})$.

(1) *Determine the ring of integers* $\mathcal{O}_K$ *of* K.
(2) *Let* $p \in \mathbb{Z}$ *be a prime number not dividing* $2D$. *Show that if the ideal* $(p) = p\mathcal{O}_K$ *is prime then the congruence* $x^2 \equiv D \pmod{p}$ *has no solutions.*

CHAPTER 6

Transcendental Number Theory

6.1. Function fields and transcendence

We have already been introduced to transcendental numbers and their basic properties. The goal of this chapter is to further investigate this fascinating topic. In this section, we briefly take a more algebraic look at transcendence and algebraic independence. We start by defining polynomial rings in several variables. A monomial in the variables $x_1, \ldots, x_k$, $k \geq 1$, is an expression of the form

$$x_1^{m_1} x_2^{m_2} \cdots x_k^{m_k}, \tag{6.1}$$

where $m_1, \ldots, m_k \in \mathbb{N}_0$ with $m_1 + \cdots + m_k > 0$. Let R be a commutative ring with 1. Define $R[x_1, \ldots, x_k]$ to be the set of all finite linear combinations of 1 and all possible monomials as in (6.1) with coefficients from R. This is a commutative ring with identity under the standard operations of addition and multiplication on these multivariable polynomials (Problem 6.1).

Let K be a field and $K[x_1, \ldots, x_k]$ be the polynomial ring in $k \geq 1$ variables with coefficients in K. Define

$$K(x_1, \ldots, x_k) = \left\{ \frac{p(x_1, \ldots, x_k)}{q(x_1, \ldots, x_k)} : p, q \in K[x_1, \ldots, x_k],\ q \neq 0 \right\},$$

where we say that

$$p(x_1, \ldots, x_k)/q(x_1, \ldots, x_k) = f(x_1, \ldots, x_k)/g(x_1, \ldots, x_k)$$

if and only if $p(x_1, \ldots, x_k)g(x_1, \ldots, x_k) = f(x_1, \ldots, x_k)q(x_1, \ldots, x_k)$; this can be viewed as an equivalence relation on the set of pairs of polynomials and then $K(x_1, \ldots, x_k)$ is the set of equivalence classes, analogously to construction of $\mathbb{Q}$ from $\mathbb{Z}$. $K(x_1, \ldots, x_k)$ is a field, called the *function field* or *field of rational functions* in k variables over K, and is precisely the quotient field of the polynomial ring $K[x_1, \ldots, x_k]$ (Problem 6.2). We can now give an alternative definition of algebraic independence.

LEMMA 6.1.1. *A collection of numbers $\alpha_1, \ldots, \alpha_k \in \mathbb{C}$ is algebraically independent if and only if there does not exist any nonzero polynomial $p(x_1, \ldots, x_k) \in \mathbb{Q}[x_1, \ldots, x_k]$ such that*

$$p(\alpha_1, \ldots, \alpha_k) = 0. \tag{6.2}$$

PROOF. Suppose that there exists some nonzero polynomial p satisfying (6.2). Define

$$f(x) = p(\alpha_1, \ldots, \alpha_{k-1}, x).$$

Then $f(x) \in \mathbb{Q}(\alpha_1, \ldots, \alpha_{k-1})[x]$ and $f(\alpha_k) = 0$. Let $d = \deg(f(x))$, then $1, \alpha_k, \ldots, \alpha_k^d$ are linearly dependent over $\mathbb{Q}(\alpha_1, \ldots, \alpha_{k-1})$. This means that

$$[\mathbb{Q}(\alpha_1, \ldots, \alpha_k) : \mathbb{Q}(\alpha_1, \ldots, \alpha_{k-1})] \leq d < \infty,$$

and hence $\alpha_1, \ldots, \alpha_k$ are not algebraically independent. Thus, if the numbers $\alpha_1, \ldots, \alpha_k$ are algebraically independent, then no nonzero polynomial p satisfying (6.2) can exist.

Conversely, suppose now that no nonzero polynomial p satisfying (6.2) exists. Suppose, towards a contradiction, that $\alpha_1, \ldots, \alpha_k$ are algebraically dependent. Then, without loss of generality, we can assume that

$$[\mathbb{Q}(\alpha_1, \ldots, \alpha_k) : \mathbb{Q}(\alpha_1, \ldots, \alpha_{k-1})] < \infty.$$

Hence $1, \alpha_k, \ldots, \alpha_k^d$ are linearly dependent over $\mathbb{Q}(\alpha_1, \ldots, \alpha_{k-1})$ for some d. In other words, there exist $a_0, \ldots, a_d \in \mathbb{Q}(\alpha_1, \ldots, \alpha_{k-1})$ such that

$$\sum_{n=0}^{d} a_n \alpha_k^n = 0. \tag{6.3}$$

Notice that $a_0, \ldots, a_d$ are rational functions in $\alpha_1, \ldots, \alpha_{k-1}$, say

$$a_n = \frac{p_n(\alpha_1, \ldots, \alpha_{k-1})}{q_n(\alpha_1, \ldots, \alpha_{k-1})},$$

where $p_n(x_1, \ldots, x_{k-1}), q_n(x_1, \ldots, x_{k-1}) \in \mathbb{Q}[x_1, \ldots, x_{k-1}]$. Write $\boldsymbol{x}$ for $(x_1, \ldots, x_{k-1})$, $\boldsymbol{\alpha}$ for $(\alpha_1, \ldots, \alpha_{k-1})$, and notice by (6.3) we have:

$$\sum_{n=0}^{d} p_n(\boldsymbol{\alpha}) \left(\prod_{m=0, m\neq n}^{d} q_m(\boldsymbol{\alpha}) \right) \alpha_k^n = 0.$$

Then define

$$p(x_1, \ldots, x_k) = \sum_{n=0}^{d} p_n(\boldsymbol{x}) \left(\prod_{m=0, m\neq n}^{d} q_m(\boldsymbol{x}) \right) x_k^n \in \mathbb{Q}[x_1, \ldots, x_k],$$

and notice that $p(\alpha_1, \ldots, \alpha_k) = 0$. This contradicts our assumption, and hence $\alpha_1, \ldots, \alpha_k$ must be algebraically independent. □

Let $\alpha_1, \ldots, \alpha_k \in \mathbb{C}$, and consider a subfield $\mathbb{Q}(\alpha_1, \ldots, \alpha_k) \subseteq \mathbb{C}$ generated by these elements. Let us write $\boldsymbol{\alpha}$ for the k-tuple $(\alpha_1, \ldots, \alpha_k)$, and define the *evaluation map* $\varphi_{\boldsymbol{\alpha}} : \mathbb{Q}(x_1, \ldots, x_k) \to \mathbb{Q}(\alpha_1, \ldots, \alpha_k)$ given by sending $x_n \mapsto \alpha_n$ and extending to the rest of $\mathbb{Q}(x_1, \ldots, x_k)$, i.e., a rational function in $x_1, \ldots, x_k$ will map to its value at the point with $x_1 = \alpha_1, \ldots, x_k = \alpha_k$.

THEOREM 6.1.2. *The following statements are equivalent:*

1. *The map $\varphi_{\boldsymbol{\alpha}}$ is well-defined for all $f \in \mathbb{Q}(x_1, \ldots, x_k)$.*
2. *The map $\varphi_{\boldsymbol{\alpha}}$ is an isomorphism of fields.*
3. *The numbers $\alpha_1, \ldots, \alpha_k$ are algebraically independent.*

PROOF. (1) ⇒ (2): Let $f, g \in \mathbb{Q}(x_1, \ldots, x_k)$, then

$$\varphi_{\boldsymbol{\alpha}}(f + g) = (f + g)(\boldsymbol{\alpha}) = f(\boldsymbol{\alpha}) + g(\boldsymbol{\alpha}) = \varphi_{\boldsymbol{\alpha}}(f) + \varphi_{\boldsymbol{\alpha}}(g),$$

$$\varphi_{\boldsymbol{\alpha}}(fg) = (fg)(\boldsymbol{\alpha}) = f(\boldsymbol{\alpha})g(\boldsymbol{\alpha}) = \varphi_{\boldsymbol{\alpha}}(f)\varphi_{\boldsymbol{\alpha}}(g).$$

Hence $\varphi_{\boldsymbol{\alpha}}$ is a ring homomorphism. Suppose that $f \in \mathrm{Ker}(\varphi_{\boldsymbol{\alpha}})$, then $\varphi_{\boldsymbol{\alpha}}(f) = f(\boldsymbol{\alpha}) = 0$. We can write $f = g/h$, where $g, h \in \mathbb{Q}[x_1, \ldots, x_k]$ are polynomials in k variables with coefficients in $\mathbb{Q}$. Since $f(\boldsymbol{\alpha}) = 0$, we must have

$$g(\alpha_1, \ldots, \alpha_k) = 0.$$

Assume $g \neq 0$, then $1/g \in \mathbb{Q}(x_1, \ldots, x_k)$, however $\varphi_{\boldsymbol{\alpha}}$ is not defined at $1/g$. Hence we must have $g = 0$, meaning that $f = 0$. Therefore $\mathrm{Ker}(\varphi_{\boldsymbol{\alpha}}) = \{0\}$, and so $\varphi_{\boldsymbol{\alpha}}$ is injective. Finally, every element β of $\mathbb{Q}(\alpha_1, \ldots, \alpha_k)$ is a rational function in $\alpha_1, \ldots, \alpha_k$, which means that β is the value of some $f \in \mathbb{Q}(x_1, \ldots, x_k)$ at $\boldsymbol{\alpha}$. This proves surjectivity, and hence $\varphi_{\boldsymbol{\alpha}}$ is a field isomorphism.

(2) $\Rightarrow$ (3): If $\varphi_{\boldsymbol{\alpha}}$ is a field isomorphism, it must be well-defined as a function for each $f = g/h \in \mathbb{Q}(x_1, \ldots, x_k)$, where $g, h \in \mathbb{Q}[x_1, \ldots, x_k]$. This means that cannot exist a polynomial $p(x_1, \ldots, x_k) \in \mathbb{Q}[x_1, \ldots, x_k]$ such that $p(\boldsymbol{\alpha}) = 0$. Hence $\alpha_1, \ldots, \alpha_k$ are algebraically independent by Lemma 6.1.1.

(3) $\Rightarrow$ (1): Since $\alpha_1, \ldots, \alpha_k$ are algebraically independent, Lemma 6.1.1 implies that for any $0 \neq p \in \mathbb{Q}[x_1, \ldots, x_k]$, $p(\boldsymbol{\alpha}) \neq 0$. Then for any $f = g/h \in \mathbb{Q}(x_1, \ldots, x_k)$, where $g, h \in \mathbb{Q}[x_1, \ldots, x_k]$, $\varphi_{\boldsymbol{\alpha}}(f) = g(\boldsymbol{\alpha})/h(\boldsymbol{\alpha})$ is well-defined. □

Hence we have the following immediate characterization of transcendence and algebraic independence.

COROLLARY 6.1.3. *A collection of complex numbers $\alpha_1, \ldots, \alpha_k$ is algebraically independent if and only if $\mathbb{Q}(\alpha_1, \ldots, \alpha_k) \cong \mathbb{Q}(x_1, \ldots, x_k)$. In particular, $\alpha \in \mathbb{C}$ is transcendental if and only if $\mathbb{Q}(\alpha) \cong \mathbb{Q}(x)$.*

6.2. Hermite, Lindemann, Weierstrass

Arguably the two most famous transcendental numbers are e and π. Transcendence of e was originally established by Charles Hermite in 1873, and transcendence of π established in 1882 by Ferdinand von Lindemann by an extension of Hermite's technique. The much more general statement, from which these two results follow, was obtained by Karl Weierstrass in 1885. The most general form of the Hermite-Lindemann-Weierstrass Theorem is as follows.

THEOREM 6.2.1. *Let $s \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_s$ be distinct algebraic numbers, and $d_1, \ldots, d_s$ nonzero algebraic numbers. Then*

$$\sum_{k=1}^{s} d_k e^{\alpha_k} \neq 0.$$

In this section we will establish the famous results of Hermite, Lindemann, and Weierstrass. The general idea of the method used is similar in all three cases, however it will be easier to follow the development of this technique starting with transcendence of e, then π, and only then the general Theorem 6.2.1. Throughout this chapter, we freely use the exponential and logarithmic functions, the basic properties of which are briefly recalled in Appendix C. Our exposition here follows [**MR14**]. We start with some preliminary observations.

Let $f(x)$ be a polynomial with complex coefficients, and let $F(x)$ be the polynomial obtained from $f(x)$ by replacing each coefficient of f with its absolute value. For a complex number t, define

$$I(t, f) := \int_0^t e^{t-u} f(u)\, du. \tag{6.4}$$

Then it is easy to see that

$$|I(t, f)| \leq |t| e^{|t|} F(|t|). \tag{6.5}$$

On the other hand, integrating by parts, we see that

$$I(t, f) = e^t f(0) - f(t) + I(t, f').$$

If degree of $f(x)$ is equal to m, then iterating the above procedure m times, we obtain:

$$I(t, f) = e^t \sum_{j=0}^{m} f^{(j)}(0) - \sum_{j=0}^{m} f^{(j)}(t). \tag{6.6}$$

THEOREM 6.2.2 (Hermite, 1873). *The number e is transcendental.*

PROOF. Working towards a contradiction, suppose that e is algebraic. Then there exist some integers $a_0, \ldots, a_n$, $n \geq 1$, such that

$$\sum_{k=0}^{n} a_k e^k = 0, \tag{6.7}$$

where $a_0, a_n \neq 0$. Let $p > |a_0|$ be a prime, and define a polynomial

$$f(x) = x^{p-1}(x-1)^p \cdots (x-n)^p.$$

Then degree of $f(x)$ is $m = (n+1)p - 1$ and each of the roots $x = 1, \ldots, n$ of $f(x)$ has multiplicity p and the root $x = 0$ has multiplicity $p-1$, which implies that

$$f^{(j)}(k) = 0 \ \forall\ 1 \leq k \leq n, 0 \leq j \leq p,\ \ f^{(j)}(0) = 0 \ \forall\ 0 \leq j \leq p-1. \tag{6.8}$$

With this notation, define

$$J := \sum_{k=0}^{n} a_k I(k, f),$$

where $I(k, f)$ is as in (6.4). Then, by (6.6) above, we have

$$\begin{aligned} J &= \sum_{k=0}^{n} \left(a_k e^k \sum_{j=0}^{m} f^{(j)}(0) - a_k \sum_{j=0}^{m} f^{(j)}(k) \right) \\ &= \sum_{j=0}^{m} \left(f^{(j)}(0) \sum_{k=0}^{n} a_k e^k \right) - \sum_{j=0}^{m} \sum_{k=0}^{n} a_k f^{(j)}(k) \\ &= -\sum_{j=0}^{m} \sum_{k=0}^{n} a_k f^{(j)}(k) = - \sum_{j=p-1}^{m} \sum_{k=0}^{n} a_k f^{(j)}(k), \end{aligned}$$

where the last line follows by (6.7) and (6.8). For $j = p - 1$, the contribution from f is

$$f^{(p-1)}(0) = (p-1)!(-1)^{np}(n!)^p,$$

hence, if $n < p$, then $f^{(p-1)}(0)$ is divisible by $(p-1)!$, but not by p. Now, for every $j \geq p$, then $f^{(j)}(0)$ and $f^{(j)}(k)$ for every $1 \leq k \leq n$ are divisible by $p!$. In other words, J is a nonzero integer divisible by $(p-1)!$, and so

$$|J| \geq (p-1)! \tag{6.9}$$

On the other hand, let $A = \max_{0 \leq k \leq n} |a_k|$, then (6.5) implies that

$$|J| \leq (n+1)A|I(k, f)| \leq n(n+1)Ae^n \max_{1 \leq k \leq n} F(k).$$

Notice that

$$\max_{1 \leq k \leq n} F(k) = (2n)^{p-1} \left(2(n-1)!\right)^p = \frac{1}{n}((2n)!)^p,$$

and so

$$|J| \leq A(n+1)e^n((2n)!)^p. \tag{6.10}$$

Combining (6.9) and (6.10), we obtain:

$$(p-1)! \leq A(n+1)e^n((2n)!)^p,$$

which is certainly not true for sufficiently large p, and so we have a contradiction. □

To attempt the proof of transcendence of π, we need the notion of symmetric polynomials. Let $n \geq 1$ and define S_n to be the set of all permutations of the set of n elements $\{1, \ldots, n\}$: this is a group under the operation of function composition, called the *symmetric group* on n letters (Problem 6.3). A polynomial $f(x_1, \ldots, x_n) \in \mathbb{Q}[x_1, \ldots, x_n]$ is called *symmetric* if for every $\tau \in S_n$,

$$f(x_1, \ldots, x_n) = f(x_{\tau(1)}, \ldots, x_{\tau(n)}).$$

The following property of symmetric polynomials we state without proof.

FACT 6.2.1. *Let $\alpha \in \mathbb{A}$ be of degree n and let $\alpha = \alpha_1, \ldots, \alpha_n$ be algebraic conjugates of α. Let $f(x_1, \ldots, x_n) \in \mathbb{Q}[x_1, \ldots, x_n]$ be a symmetric polynomial. Then*

$$f(\alpha_1, \ldots, \alpha_n) \in \mathbb{Q}.$$

Moreover, if $\alpha \in \mathbb{I}$ and $f(x_1, \ldots, x_n) \in \mathbb{Z}[x_1, \ldots, x_n]$, then

$$f(\alpha_1, \ldots, \alpha_n) \in \mathbb{Z}.$$

Let us also recall that π is half the circumference of a circle of radius 1, which is precisely the angle that the ray emanating from the origin through the point $(-1, 0)$ on the unit circle makes with the ray indicating the positive direction along the x-axis in the Cartesian plane. Hence

$$\cos \pi = -1, \ \sin \pi = 0.$$

THEOREM 6.2.3 (Lindemann, 1882). *The number π is transcendental.*

PROOF. As in the proof of Theorem 6.2.2, suppose π is algebraic. Since we know that $i \in \mathbb{A}$ and $\mathbb{A}$ is a field, $\alpha = \pi i$ must also be algebraic. Let $d = \deg(\alpha)$ and let $\alpha = \alpha_1, \ldots, \alpha_d$ be conjugates of α. Let N be the leading coefficient of $m_\alpha(x)$, then Lemma 5.2.7 implies that $N\alpha$ is an algebraic integer. By Euler's formula,

$$e^{\pi i} = -1,$$

and hence

$$(1 + e^{\alpha_1}) \cdots (1 + e^{\alpha_d}) = 0. \tag{6.11}$$

This product can be written as a sum of 2^d terms of the form e^θ, where

$$\theta = \varepsilon_1 \alpha_1 + \cdots + \varepsilon_d \alpha_d, \ \varepsilon_k = 0, 1 \ \forall \ 1 \leq k \leq d.$$

Suppose that exactly n of these numbers are nonzero, denote them $\beta_1, \ldots, \beta_n$. Let

$$h(x) = \prod_{\varepsilon_1 = 0}^{1} \cdots \prod_{\varepsilon_d = 0}^{1} (x - (\varepsilon_1 \alpha_1 + \cdots + \varepsilon_d \alpha_d))$$

and notice that $h(x)$ is symmetric in $\alpha_1, \ldots, \alpha_d$. Then Fact 6.2.1 implies that $h(x) \in \mathbb{Q}[x]$. Notice that the roots of $h(x)$ are $\beta_1, \ldots, \beta_d$ and 0, which has multiplicity $a = 2^d - n$. Clearing the denominators, this means that for some $C \in \mathbb{Z}$, $h(x) = Cx^a g(x)$, where $g(x) \in \mathbb{Z}[x]$ is the polynomial of degree n with roots $\beta_1, \ldots, \beta_n$. Now (6.11) implies that

$$(2^d - n)e^0 + e^{\beta_1} + \cdots + e^{\beta_n} = 0. \tag{6.12}$$

Let

$$f(x) = N^{np} x^{p-1} \prod_{k=1}^{n} (x - \beta_k)^p$$

for some large prime p, and let $I(t, f)$ for this choice of $f(x)$ be as in (6.4) above. Notice that degree of $f(x)$ is $m = (n + 1)p - 1$. Define

$$J := \sum_{k=1}^{n} I(\beta_k, f).$$

Then, by (6.6),

$$\begin{aligned} J &= \sum_{k=1}^{n}\left(e^{\beta_k}\sum_{j=0}^{m} f^{(j)}(0) - \sum_{j=0}^{m} f^{(j)}(\beta_k)\right) \\ &= \left(\sum_{k=1}^{n} e^{\beta_k}\right)\left(\sum_{j=0}^{m} f^{(j)}(0)\right) - \sum_{j=0}^{m}\sum_{k=1}^{n} f^{(j)}(\beta_k) \\ &= -(2^d - n)\left(\sum_{j=0}^{m} f^{(j)}(0)\right) - \sum_{j=0}^{m}\sum_{k=1}^{n} f^{(j)}(\beta_k), \end{aligned}$$

where the last equality follows by (6.12). Notice that $\sum_{k=1}^{n} f^{(j)}(\beta_k)$ is a symmetric polynomial in $N\beta_1, \ldots, N\beta_n$ for each j. Furthermore, each $N\beta_k$ is a linear combination of algebraic integers $\alpha_1, \ldots, \alpha_d$, and hence is an algebraic integer. Therefore, by Fact 6.2.1, for each $1 \leq j \leq m$, $\sum_{k=1}^{n} f^{(j)}(\beta_k) \in \mathbb{Z}$. Further, each β_k is a root of $f(x)$ of multiplicity p, which means that each derivative $f^{(j)}(\beta_k)$ vanishes for all $j < p$. For each $j \geq p$, $\sum_{k=1}^{n} f^{(j)}(\beta_k)$ is divisible by $p!$. Also,

$$f^{(p-1)}(0) = (p-1)!(-N)^{np}(\beta_1 \cdots \beta_n)^p,$$

which is not divisible by p provided that p is large (specifically, when $p > N\beta_1 \cdots \beta_n$). In addition, $f^{(j)}(0)$ is divisible by $p!$ for all $j \geq p$. Therefore K is divisible by $(p-1)!$, and hence

$$|J| \geq (p-1)!$$

On the other hand,

$$|J| \leq \sum_{k=1}^{n} |I(\beta_k, f)| \leq \sum_{k=1}^{n} |\beta_k| e^{|\beta_k|} F(|\beta_k|)$$

by (6.5) and $F(x)$ related to $f(x)$ is as above. Then we have

$$(p-1)! \leq |J| \leq AC^p$$

for some constants A and C. Taking p sufficiently large, we reach a contradiction. □

We are now ready to prove the Lindemann-Weierstrass Theorem.

Proof of Theorem 6.2.1. Towards a contradiction, suppose that there exist some algebraic numbers $d_1, \ldots, d_s$, not all zero, such that

$$\sum_{k=1}^{s} d_k e^{\alpha_k} = 0. \tag{6.13}$$

Multiplying both sides by some N, by Lemma 5.2.7 we can assume that $d_1, \ldots, d_s$ are algebraic integers. Let $K = \mathbb{Q}(d_1, \ldots, d_s)$, $n = [K : \mathbb{Q}]$, and let $\sigma_k : K \to \mathbb{C}$ for $1 \leq k \leq n$ be embeddings of K. Notice that (6.13) implies that

$$\prod_{l=1}^{n}\left(\sum_{k=1}^{s} \sigma_l(d_k) e^{\alpha_k}\right) = 0. \tag{6.14}$$

The equation (6.14) can be written as

$$a_1 e^{\gamma_1} + \cdots + a_m e^{\gamma_m} = 0, \tag{6.15}$$

where each coefficient a_l is a sum of terms of the form $\sigma_m(d_k)$, which is invariant under each of the embeddings σ_m. Then Problem 5.16 above implies that $a_1, \ldots, a_m \in \mathbb{Q}$, and clearing the denominators, if necessary, we can assume that $a_1, \ldots, a_m \in \mathbb{Z}$. Further, we can assume that the set $\gamma_1, \ldots, \gamma_m$ contains all the conjugates of each of the γ_j's: if some of them are not there, they can always be included by choosing the corresponding a_l coefficient to be 0. Notice also that the exponents $\gamma_1, \ldots, \gamma_m$ are distinct algebraic numbers.

Let us write $\gamma_j^{(l)}$ for the l-th conjugate of γ_j. Let t be a real variable and for each l define the conjugate function

$$A_l(t) := \sum_{k=1}^{m} a_k e^{\gamma_k^{(l)} t}.$$

We will use the fact that when the γ_k's are all distinct, the functions $A_l(t)$ are not identically zero. Define

$$B(t) = \prod_l A_l(t) = \sum_{k=1}^{M} b_k e^{\beta_k t},$$

where the product is over all the conjugate functions $A_l(t)$. Notice that $B(1) = 0$ by our original assumption. Since $a_1, \ldots, a_m \in \mathbb{Z}$, the coefficients $b_1, \ldots, b_M$ are also rational integers, not all equal to zero. Since $\beta_1, \ldots, \beta_M$ are algebraic numbers, let $N \in \mathbb{Z}$ be such that $N\beta_1, \ldots, N\beta_M$ are algebraic integers. For each $1 \leq r \leq M$, define a polynomial

$$f_r(x) = \frac{N^{Mp}}{x - \beta_r} \prod_{k=1}^{M} (x - \beta_k)^p,$$

where $p \in \mathbb{Z}$ is a prime. Let

$$f(x) = \sum_{r=1}^{M} f_r(x),$$

then coefficients of $f(x)$ are symmetric polynomials in the algebraic integers $N\beta_1, \ldots, N\beta_M$. On the other hand, this set of numbers contains all of their algebraic conjugates, since $\beta_1, \ldots, \beta_M$ were generated by $\gamma_1, \ldots, \gamma_m$, which included all the algebraic conjugates. Hence coefficients of $f(x)$ must be in $\mathbb{Z}$ by Fact 6.2.1.

Define

$$J_r := \sum_{k=1}^{M} b_k I(\beta_k, f_r)$$

for each $1 \leq r \leq M$ and let $J := J_1 \cdots J_M$. Let $m := \deg(f_r) = Mp - 1$, and notice that by (6.6),

$$\begin{aligned} J_r &= \sum_{k=1}^{M} b_k \left(e^{\beta_k} \sum_{j=0}^{m} f_r^{(j)}(0) - \sum_{j=0}^{m} f_r^{(j)}(\beta_k) \right) \\ &= -\sum_{k=1}^{M} b_k \sum_{j=0}^{m} f_r^{(j)}(\beta_k), \end{aligned}$$

where the last equality follows from the assumption that $B(1) = 0$. Arguing analogously to our proofs of Theorems 6.2.2 and 6.2.3, we conclude that J is an algebraic integer which is fixed by all the embeddings of the number field $\mathbb{Q}(\beta_1, \ldots, \beta_M)$,

hence it must be in $\mathbb{Z}$. Further, J is divisible by $(p-1)!$, but not by p for a sufficiently large p. In the opposite direction, each $|J_r|$ can be bounded by c_r^p for a suitable positive real c_r, and hence $|J|$ can be bounded by C^p for some constant C. Therefore,

$$(p-1)! \leq |J| \leq C^p,$$

which leads to a contradiction for a large enough p. This completes the proof. $\square$

6.3. Beyond Lindemann-Weierstrass

In this section we discuss some consequences of Theorem 6.2.1. First notice that transcendence of e and π follow easily from the Lindemann-Weierstrass Theorem. Although we have already proved these facts separately, it is still worthwhile to see them derived as consequences of the Lindemann-Weierstrass Theorem. We present these derivations here.

COROLLARY 6.3.1. *e is transcendental.*

PROOF. Suppose $e \in \mathbb{A}$. Then there exists some nonzero polynomial

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

such that

$$p(e) = \sum_{k=0}^{n} a_k e^k = 0.$$

This, however, clearly contradicts Theorem 6.2.1. □

COROLLARY 6.3.2. *π is transcendental.*

PROOF. Suppose π is algebraic. We know also that $i \in \mathbb{A}$, and hence $i\pi \in \mathbb{A}$ since $\mathbb{A}$ is a field. By Euler's formula,

$$e^{i\pi} = \cos\pi + i\sin\pi = -1,$$

and hence we have

$$e^{i\pi} + 1 = 0,$$

which clearly contradicts Theorem 6.2.1. □

Theorem 6.2.1 has many other important consequences. Here are some of them.

COROLLARY 6.3.3. *Let $0 \neq \alpha \in \mathbb{A}$. Then the numbers e^α, $\ln\alpha$, $\sin\alpha$ and $\cos\alpha$ are transcendental.*

PROOF. Suppose e^α is algebraic, say $\gamma = e^\alpha \in \mathbb{A}$. Then

$$e^\alpha - \gamma e^0 = 0,$$

which contradicts Theorem 6.2.1. Hence e^α is transcendental. Now assume that $\ln\alpha$ is algebraic, then $e^{\ln\alpha} = \alpha$ would have to be transcendental, which is a contradiction. Furthermore, Euler's formula implies that

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

and so

$$e^0\sin\alpha - \frac{1}{2i}e^{i\alpha} + \frac{1}{2i}e^{-i\alpha} = 0,$$
$$e^0\cos\alpha - \frac{1}{2}e^{i\alpha} - \frac{1}{2}e^{-i\alpha} = 0.$$

Now Theorem 6.2.1 implies that $\sin\alpha$, $\cos\alpha$ cannot be algebraic. □

COROLLARY 6.3.4. *Let $\alpha_1, \ldots, \alpha_n \in \mathbb{A}$ be linearly independent over $\mathbb{Q}$. Then the numbers*

$$e^{\alpha_1}, \ldots, e^{\alpha_n}$$

are algebraically independent.

PROOF. Suppose that $e^{\alpha_1}, \ldots, e^{\alpha_n}$ are algebraically dependent, then there exists some non-constant polynomial

$$p(x_1, \ldots, x_n) \in \mathbb{Q}[x_1, \ldots, x_n]$$

such that

$$p(e^{\alpha_1}, \ldots, e^{\alpha_n}) = \sum_{i_1, \ldots, i_n} a_{i_1, \ldots, i_n} e^{i_1\alpha_1 + \cdots + i_n\alpha_n} = 0,$$

where the coefficients $a_{i_1,\ldots,i_n}$ are rational numbers, not all zero. Then Theorem 6.2.1 implies that the exponents

$$i_1\alpha_1 + \cdots + i_n\alpha_n$$

cannot be all distinct. Hence there exist some two distinct families of indices $i_1, \ldots, i_n$ and $j_1, \ldots, j_n$ such that

$$i_1\alpha_1 + \cdots + i_n\alpha_n = j_1\alpha_1 + \cdots + j_n\alpha_n,$$

in other words

$$\sum_{k=1}^{n} c_k\alpha_k = 0,$$

where not all of $c_k := i_k - j_k \in \mathbb{Z}$ are equal to zero. This contradicts the assumption that $\alpha_1, \ldots, \alpha_n$ are linearly independent over $\mathbb{Q}$. □

In fact, it is easy to see that Corollary 6.3.4 is equivalent to the Lindemann-Weierstrass Theorem, i.e. it is a convenient reformulation of the famous result. A substantial strengthening of Corollary 6.3.4 is arguably the most important open problem in transcendental number theory.

CONJECTURE 6.3.1 (Schanuel's Conjecture). *Let $\alpha_1, \ldots, \alpha_n \in \mathbb{C}$ be linearly independent over $\mathbb{Q}$. Then*

$$\operatorname{trdeg}(\mathbb{Q}(\alpha_1, \ldots, \alpha_n, e^{\alpha_1}, \ldots, e^{\alpha_n})) \geq n.$$

We now discuss some of the many remarkable implications of this conjecture. First we mention (a weak form of) the famous theorem of Alan Baker (1966) on linear independence of logarithms of algebraic numbers, for which he received a Fields Medal in 1970.

THEOREM 6.3.5 (Baker's Theorem, 1966). *Let*

$$\Lambda = \left\{\ell \in \mathbb{C} : e^{\ell} \in \mathbb{A}\right\}.$$

If $\ell_1, \ldots, \ell_n \in \Lambda$ are linearly independent over $\mathbb{Q}$, then they are algebraically independent (and hence linearly independent over $\mathbb{A}$).

PROOF. Baker's theorem has been proved unconditionally, however the proof is quite complicated. Here we will only show how this result follows from Schanuel's Conjecture. Indeed, Schanuel's Conjecture implies that

$$\operatorname{trdeg}(\mathbb{Q}(\ell_1, \ldots, \ell_n, e^{\ell_1}, \ldots, e^{\ell_n})) \geq n.$$

Since $\ell_1, \ldots, \ell_n \in \Lambda$, we know that $e^{\ell_1}, \ldots, e^{\ell_n} \in \mathbb{A}$, which implies that

$$\operatorname{trdeg}(\mathbb{Q}(\ell_1, \ldots, \ell_n)) = \operatorname{trdeg}(\mathbb{Q}(\ell_1, \ldots, \ell_n, e^{\ell_1}, \ldots, e^{\ell_n})) \geq n.$$

Hence $\ell_1, \ldots, \ell_n$ are algebraically independent. □

In fact, a strong version of Baker's Theorem establishes transcendence of any nonzero linear combination of $\ell_1, \ldots, \ell_n$ with algebraic coefficients, which, in its turn, is a generalization and strengthening of the celebrated Gelfond-Schneider Theorem, established independently in 1934 by Alexander Gelfond and Theodor Schneider. Their theorem presented a solution to Hilbert's 7th Problem.

THEOREM 6.3.6 (Gelfond-Schneider Theorem, 1934). *Let $a, b \in \mathbb{A}$ be such that $a \neq 0, 1$ and $b \notin \mathbb{Q}$. Then $a^b \in \mathbb{T}$.*

Furthermore, Schanuel's Conjecture implies algebraic independence of e and π, which is currently an open problem, as well as a wide variety of other known results and open problems in transcendental number theory. We conclude with yet another famous open problem, which would follow from Schanuel's Conjecture.

CONJECTURE 6.3.2 (Schneider's Four Exponentials Conjecture). *Let x_1, x_2 and y_1, y_2 be pairs of complex numbers linearly independent over $\mathbb{Q}$. Then at least one of the four numbers $e^{x_j y_k}$ where $1 \leq j, k \leq 2$ is transcendental.*

If the linearly independent pair y_1, y_2 in the conjecture above is replaced with the linearly independent triple y_1, y_2, y_3, then the conjecture becomes a theorem, known as the Six Exponentials Theorem. It is our next big goal to prove this result.

6.4. Siegel's Lemma

We now develop an important tool, which will be used to prove another celebrated transcendence result, the Six Exponentials Theorem. This tool is Siegel's Lemma, the simplest version of which was originally observed by Axel Thue in 1909 and then formally proved by Carl Ludwig Siegel in 1929. While Siegel's Lemma originated as a tool used in transcendence arguments, it took on a separate life in the more recent years as a first case of a result on points of bounded height on algebraic varieties: we will talk more about this direction in Section 7.6 below.

Our presentation here partially follows [**Sch91**] and [**MR14**]. Let

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{l1} & \dots & a_{ln} \end{pmatrix}$$

be an $l \times n$ matrix with integer entries and rank equal to $l < n$. Define

$$\Lambda = \{\boldsymbol{x} \in \mathbb{Z}^n : A\boldsymbol{x} = \boldsymbol{0}\}.$$

THEOREM 6.4.1 (Siegel's Lemma, version 1). *With notation as above, there exists* $\boldsymbol{0} \neq \boldsymbol{x} \in \Lambda$ *with*

$$|\boldsymbol{x}| < 2 + (n|A|)^{\frac{l}{n-l}}, \tag{6.16}$$

where $|\boldsymbol{x}| = \max\{|x_i| : 1 \leq i \leq n\}$, $|A| = \max\{|a_{ij}| : 1 \leq i \leq l,\ 1 \leq j \leq n\}$.

PROOF. Let $R \in \mathbb{Z}_{>0}$, and let

$$C_R^n = \{\boldsymbol{x} \in \mathbb{R}^n : |\boldsymbol{x}| \leq R\}$$

be the cube centered at the origin in $\mathbb{R}^n$ with sidelength $2R$. Then

$$|C_R^n \cap \mathbb{Z}^n| = (2R+1)^n.$$

Let $T_A : \mathbb{R}^n \to \mathbb{R}^l$ be a linear map, given by $T_A(\boldsymbol{x}) = A\boldsymbol{x}$ for each $\boldsymbol{x} \in \mathbb{R}^n$. Notice that for every $\boldsymbol{x} \in C_R^n$,

$$|T_A(\boldsymbol{x})| \leq n|A|R,$$

i.e. T_A maps C_R^n into $C_{n|A|R}^l \subseteq \mathbb{R}^l$, since $\mathrm{rk}(A) = l$. Now

$$|C_{n|A|R}^l \cap \mathbb{Z}^l| = (2n|A|R+1)^l.$$

Now let us choose R to be a positive integer satisfying

$$(n|A|)^{\frac{l}{n-l}} \leq 2R < (n|A|)^{\frac{l}{n-l}} + 2.$$

Then

$$\begin{aligned} |C_R^n \cap \mathbb{Z}^n| &= (2R+1)^n = (2R+1)^l (2R+1)^{n-l} \\ &\geq (2R+1)^l (n|A|)^l > (2n|A|R+1)^l \\ &= |C_{n|A|R}^l \cap \mathbb{Z}^l|. \end{aligned}$$

This means that T_A cannot be mapping $C_R^n \cap \mathbb{Z}^n$ into $C_{n|A|R}^l \cap \mathbb{Z}^l$ in a one-to-one manner. Hence, there must exist $\boldsymbol{x} \neq \boldsymbol{y} \in C_R^n \cap \mathbb{Z}^n$ such that $T_A(\boldsymbol{x}) = T_A(\boldsymbol{y})$, i.e.

$$T_A(\boldsymbol{x} - \boldsymbol{y}) = 0,$$

and so $\boldsymbol{x} - \boldsymbol{y} \in \Lambda$. On the other hand,

$$|\boldsymbol{x} - \boldsymbol{y}| \leq |\boldsymbol{x}| + |\boldsymbol{y}| \leq 2R < (n|A|)^{\frac{l}{n-l}} + 2,$$

and this finishes the proof. □

Notice that the main underlying idea in the proof of Siegel's Lemma was the pigeon hole principle. It is remarkable that the exponent $\frac{l}{n-l}$ in the upper bound of (6.16) cannot be improved. To see this, let for instance $l = n-1$ and for a positive integer R consider the $(n-1) \times n$ matrix

$$A = \begin{pmatrix} R & -1 & 0 & \dots & 0 & 0 \\ 0 & R & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & R & -1 \end{pmatrix}.$$

Then $|A| = R$, and every nonzero integer solution of the system of linear equations $A\boldsymbol{x} = \mathbf{0}$ must have $x_n = R^{n-1}x_1$. Therefore, if

$$\Lambda = \{\boldsymbol{x} \in \mathbb{Z}^n : A\boldsymbol{x} = \mathbf{0}\},$$

and $\mathbf{0} \neq \boldsymbol{x} \in \Lambda$, then

$$|\boldsymbol{x}| \geq R^{n-1} = |A|^{\frac{l}{n-l}}.$$

Siegel's Lemma-type results have been proved in a variety of considerably more general settings by a number of authors, employing quite sophisticated machinery from number theory and arithmetic geometry (more on this in Section 7.6). However, the original motivation for Siegel's Lemma came from Diophantine approximation and transcendental number theory.

For our use, we will also need a basic version of Siegel's Lemma over number fields. Let K be a number field of degree d with embeddings $\sigma_1, \dots, \sigma_d$. For each $\alpha \in K$, define its *height*

$$\mathcal{H}(\alpha) := \max\{|\sigma_k(\alpha)| : 1 \leq k \leq d\}.$$

Height functions more generally are devices meant to measure arithmetic complexity of objects, in a certain well-defined sense. This is a somewhat simplified version of a height function, which takes into account only partial information about the arithmetic properties of an algebraic number. We will discuss the theory of height functions and introduce more sophisticated machinery in Section 7.4 below.

As we know, the ring of integers $\mathcal{O}_K$ is a free $\mathbb{Z}$-module of rank d. In other words, $\mathcal{O}_K$ has a $\mathbb{Z}$-basis: there exists a linearly independent collection $\omega_1, \dots, \omega_d \in \mathcal{O}_K$ such that

$$\mathcal{O}_K = \left\{ \sum_{k=1}^{d} a_k \omega_k : a_1, \dots, a_d \in \mathbb{Z} \right\}.$$

Define the corresponding $d \times d$ basis matrix $W := (\sigma_\ell(\omega_k))_{1 \leq \ell, k \leq d}$, which of course is nonsingular. With this notation and information in mind, we can now prove our next result.

THEOREM 6.4.2 (Siegel's Lemma, version 2). *Let K be a number field of degree d, and let $A = (\alpha_{ij})$ be an $l \times n$ matrix of rank $l < n$ with entries $\alpha_{ij} \in \mathcal{O}_K$. Define*

$$\mathcal{H}(A) := \max\{\mathcal{H}(\alpha_{ij}) : 1 \leq i \leq l, 1 \leq j \leq n\}.$$

There exists a solution $\mathbf{0} \neq \boldsymbol{x} = (x_1, \dots, x_n) \in \mathcal{O}_K^n$ *to the homogeneous linear system* $A\boldsymbol{x} = \mathbf{0}$ *with*

$$\max_{1 \le j \le n} \mathcal{H}(x_j) < B_K(l,n)\mathcal{H}(A)^{\frac{l}{n-l}}, \tag{6.17}$$

where $B_K(l,n)$ *is some constant depending only on* l, n *and the number field* K*.*

Proof. Let $\omega_1, \dots, \omega_d \in \mathcal{O}_K$ be a $\mathbb{Z}$-basis for $\mathcal{O}_K$, as described above, and let W be the corresponding basis matrix. Then for each entry α_{ij} of our matrix A, there exist $a_{ijk} \in \mathbb{Z}$, $1 \le k \le d$, such that

$$\alpha_{ij} = \sum_{k=1}^{d} a_{ijk}\omega_k.$$

Applying embeddings $\sigma_1, \dots, \sigma_d$ to the above equation, we obtain

$$\sigma_\ell(\alpha_{ij}) = \sum_{k=1}^{d} a_{ijk}\sigma_\ell(\omega_k)$$

for each $1 \le \ell \le d$, and hence

$$\boldsymbol{\alpha}_{ij} := (\sigma_1(\alpha_{ij}), \dots, \sigma_d(\alpha_{ij}))^t = W(a_{ij1}, \dots, a_{ijd})^t.$$

Since W is invertible, we have

$$\boldsymbol{a}_{ij} := (a_{ij1}, \dots, a_{ijd})^t = W^{-1}\boldsymbol{\alpha}_{ij}.$$

If we write $v_{k\ell}$ for the entries of W^{-1}, then

$$a_{ijk} = \sum_{\ell=1}^{d} v_{k\ell}\sigma_\ell(\alpha_{ij}),$$

and so

$$|a_{ijk}| \le d \max_{1 \le \ell \le d} |v_{k\ell}\sigma_\ell(\alpha_{ij})| \le dC_K\mathcal{H}(A), \tag{6.18}$$

where C_K is a constant depending only on the number field K such that $C_K \ge \max_{1 \le k,\ell \le d} |v_{k\ell}|$.

Now suppose $\boldsymbol{x} \in \mathcal{O}_K^n$ is a nontrivial solution of the system $A\boldsymbol{x} = \mathbf{0}$, and write

$$\boldsymbol{x} = \left(\sum_{\ell=1}^{d} b_{1\ell}\omega_\ell, \dots, \sum_{\ell=1}^{d} b_{n\ell}\omega_\ell \right) \tag{6.19}$$

for some $b_{j\ell} \in \mathbb{Z}$ for $1 \le j \le n$, $1 \le \ell \le d$. Then i-th entry of the vector $A\boldsymbol{x}$ is

$$\sum_{j=1}^{n} \sum_{\ell=1}^{d} \sum_{k=1}^{d} a_{ijk} b_{j\ell} \omega_k \omega_\ell = 0.$$

Since $\omega_k\omega_\ell \in \mathcal{O}_K$, it can also be expressed as a linear combination of ω_m's with $\mathbb{Z}$-coefficients:

$$\omega_k\omega_\ell = \sum_{m=1}^{d} c_{k\ell m}\omega_m$$

for each $1 \le k, \ell \le d$, and hence we have

$$\sum_{m=1}^{d} \sum_{j=1}^{n} \sum_{\ell=1}^{d} \sum_{k=1}^{d} a_{ijk} b_{j\ell} c_{k\ell m} \omega_m = 0.$$

Since $\omega_1, \dots, \omega_d$ are linearly independent over $\mathbb{Z}$, all the coefficients in the above equations must be zero, and hence we have a system of ld homogeneous linear equations with integer coefficients in the nd variables $b_{j\ell}$:

$$\sum_{j=1}^{n} \sum_{\ell=1}^{d} \sum_{m=1}^{d} a_{ijk} b_{j\ell} c_{k\ell m} = 0,$$

for all $1 \le i \le l$, $1 \le m \le d$. Applying Theorem 6.4.1 along with (6.18), we see that there exists a solution with

$$\max_{j,\ell} |b_{j\ell}| \le 2 + (nd^2 C_K \mathcal{H}(A))^{\frac{ld}{nd-ld}},$$

and hence, by (6.19),

$$\max_{1 \le j \le n} \mathcal{H}(x_j) \le d \left(2 + (nd^2 C_K \mathcal{H}(A))^{\frac{l}{n-l}}\right) \max_{1 \le \ell \le d} \mathcal{H}(\omega_\ell).$$

Since the choice of $\omega_1, \dots, \omega_\ell$ depends only on K, the conclusion of the theorem follows. □

Recall that for any $\beta \in K$, there exists $c \in \mathbb{N}$ such that $c\beta \in \mathcal{O}_K$. In fact, for any collection $\beta_1, \dots, \beta_n \in K$, let us define their *common denominator* to be

$$D(\beta_1, \dots, \beta_n) = \min\{c \in \mathbb{N} : c\beta_k \in \mathcal{O}_K \ \forall\ 1 \le k \le n\}.$$

For an $l \times n$ matrix A with entries in K, we will write $D(A)$ for the common denominator of all of its entries, i.e.,

$$D(A) = D(\alpha_{ij} : 1 \le i \le l, 1 \le j \le n).$$

With this notation in mind, we have one more version of Siegel's lemma.

COROLLARY 6.4.3 (Siegel's Lemma, version 3). *Let K be a number field of degree d, and let $A = (\alpha_{ij})$ be an $l \times n$ matrix of rank $l < n$ with entries $\alpha_{ij} \in K$. There exists a solution $\mathbf{0} \ne \boldsymbol{x} = (x_1, \dots, x_n) \in \mathcal{O}_K^n$ to the homogeneous linear system $A\boldsymbol{x} = \mathbf{0}$ with*

$$\max_{1 \le j \le n} H(x_j) < B_K(l, n)(D(A)\mathcal{H}(A))^{\frac{l}{n-l}}, \tag{6.20}$$

where $B_K(l, n)$ is the same constant as in Theorem 6.4.2 above.

PROOF. Let $A' = D(A)A$, then A' is an $l \times n$ matrix with entries in $\mathcal{O}_K$, and $A\boldsymbol{x} = \mathbf{0}$ if and only $A'\boldsymbol{x} = \mathbf{0}$. Then apply Theorem 6.4.2 to the system $A'\boldsymbol{x} = \mathbf{0}$ while keeping in mind that $\mathcal{H}(A') = D(A)\mathcal{H}(A)$. □

6.5. The Six Exponentials Theorem

In this section we use Siegel's Lemma and Maximum Modulus Principle to prove the Six Exponentials Theorem. Our presentation here follows [**MR14**]. Let us set some notation. Let K be a number field of degree d over $\mathbb{Q}$, and let $\sigma_1, \dots, \sigma_d$ be the embeddings of K into $\mathbb{C}$. Recall that for every $\alpha \in K$, the norm of α over K is

$$\mathbb{N}_K(\alpha) = \prod_{k=1}^{d} \sigma_k(\alpha),$$

and we write $\mathbb{N}(\alpha)$ for $\mathbb{N}_{\mathbb{Q}(\alpha)}(\alpha)$. It is not difficult to observe that

$$\mathbb{N}_K(\alpha) = \mathbb{N}(\alpha)^{[K:\mathbb{Q}(\alpha)]}.$$

Notice also that $\mathbb{N}(\alpha)$ is precisely the free coefficient of the minimal polynomial of α over $\mathbb{Q}$, and hence is a rational number. If $\alpha \in \mathcal{O}_K$, then the minimal polynomial of α over $\mathbb{Q}$ is equal to $m_\alpha(x)$, and hence $\mathbb{N}(\alpha) \in \mathbb{Z}$. This in particular implies that for every $\alpha \in \mathcal{O}_K$,

$$1 \leq |\mathbb{N}_K(\alpha)| = |\mathbb{N}(\alpha)|^{[K:\mathbb{Q}(\alpha)]} \leq |\mathbb{N}(\alpha)|^d \leq \mathcal{H}(\alpha)^{d-1}|\alpha|, \tag{6.21}$$

since one of the embeddings $\sigma_1, \dots, \sigma_d$ is the identity map.

THEOREM 6.5.1 (The Six Exponentials Theorem). *Let $x_1, x_2 \in \mathbb{C}$ be linearly independent over $\mathbb{Q}$. Let $y_1, y_2, y_3 \in \mathbb{C}$ also be linearly independent over $\mathbb{Q}$. Then at least one of the six numbers $e^{x_i y_j}$ where $1 \leq i \leq 2$, $1 \leq j \leq 3$ is transcendental.*

PROOF. Suppose that $e^{x_j y_k} \in \mathbb{A}$ for all $1 \leq j \leq 2$, $1 \leq k \leq 3$, and let K be a number field containing all of these numbers. Let $r \in \mathbb{N}$, $a_{ij} \in \mathcal{O}_K$ for all $1 \leq i, j \leq r$, and define

$$F(z) = \sum_{i=1}^{r} \sum_{j=1}^{r} a_{ij} e^{(ix_1 + jx_2)z} \tag{6.22}$$

for a variable $z \in \mathbb{C}$. Let $n \in \mathbb{N}$ and let $k_1, k_2, k_3 \in \mathbb{N}$ range between 1 and n. Then

$$F\left(\sum_{m=1}^{3} k_m y_m\right) = \sum_{i=1}^{r} \sum_{j=1}^{r} a_{ij} \exp\left((ix_1 + jx_2)(k_1 y_1 + k_2 y_2 + k_3 y_3)\right).$$

Since each $\exp\left((ix_1 + jx_2)(k_1 y_1 + k_2 y_2 + k_3 y_3)\right)$ is algebraic, setting each

$$F\left(\sum_{m=1}^{3} k_m y_m\right) = 0 \tag{6.23}$$

yields a system of n^3 equations with algebraic coefficients in the r^2 variables a_{ij}. We want to apply Siegel's Lemma to this system to obtain a small-height solution vector; for this we need $r^2 > n^3$. Let D be the common denominator of the six exponentials

$$\{e^{x_i y_j} : 1 \leq i \leq 2, 1 \leq j \leq 3\},$$

then the common denominator of the coefficients of the system (6.23) is bounded above by D^{6rn}, and heights of these coefficients are bounded above by $e^{c_0 rn}$ for some constant c_0. Now Theorem 6.4.3 guarantees that (6.23) has a solution vector with coordinates $a_{ij} \in \mathcal{O}_K$, not all zero, such that

$$\max_{i,j} \mathcal{H}(a_{ij}) \leq B_K(n^3, r^2) \left(D^{6rn} e^{c_0 rn}\right)^{\frac{n^3}{r^2 - n^3}}.$$

Then, choosing $r = 8n^{3/2}$, we ensure that

$$\max_{i,j} \mathcal{H}(a_{ij}) \leq B_K(n^3, 8n^{3/2}) \left(D^{48n^{5/2}} e^{8c_0 n^{5/2}} \right)^{\frac{1}{63}} \leq e^{c_1 n^{5/2}} \tag{6.24}$$

for some appropriately chosen constant c_1. Then let $a_{ij} \in \mathcal{O}_K$ be a solution to (6.23) with $r = 8n^{3/2}$ satisfying (6.24), and let $F(z)$ be as in (6.22) for this choice of a_{ij}'s. Notice that $F(z)$ is not identically zero, since x_1, x_2 are linearly independent over $\mathbb{Q}$. Also notice that the set

$$S = \{k_1 y_1 + k_2 y_2 + k_3 y_3 : k_1, k_2, k_3 \in \mathbb{N}\}$$

is not discrete, since the numbers y_1, y_2, y_3 are linearly independent over $\mathbb{Q}$. Since $F(z)$ is not identically zero, it cannot vanish on a non-discrete set, and hence there must exist elements of S on which F is not zero. Let

$$s = \max \{t \in \mathbb{N} : F(k_1 y_1 + k_2 y_2 + k_3 y_3) = 0 \ \forall \ 1 \leq k_i \leq t\}.$$

Clearly, $s \geq n$. Define

$$w = k_1 y_1 + k_2 y_2 + k_3 y_3$$

with some $k_i = s + 1$ and all $1 \leq k_i \leq s + 1$ be such that $F(w) \neq 0$. Using (6.24), we can obtain an estimate on the height of $F(w)$:

$$\mathcal{H}(F(w)) \leq C_0^{n^{5/2} + (s+1)r} \leq C_1^{s^{5/2}}$$

for some positive constants C_0, C_1. Observe also that $D^{6r(s+1)} F(w)$ is an algebraic integer. Then, by (6.21) we have:

$$1 \leq \mathbb{N}_K(D^{6r(s+1)} F(w)) \leq \mathcal{H}(D^{6r(s+1)} F(w))^{[K:\mathbb{Q}]-1} |D^{6r(s+1)} F(w)|,$$

and so

$$|F(w)| \geq D^{-6r(s+1)[K:\mathbb{Q}]} \mathcal{H}(F(w))^{-([K:\mathbb{Q}]-1)} \geq C_2^{-s^{5/2}}, \tag{6.25}$$

where C_2 is another constant independent of s.

Our next goal will be to arrive at a contradiction with (6.25) by obtaining an incompatible estimate for $|F(w)|$ from above. Notice that

$$F(w) = \lim_{z \to w} \left\{ F(z) \prod_{1 \leq k_1, k_2, k_3 \leq s} \left(\frac{w - (k_1 y_1 + k_2 y_2 + k_3 y_3)}{z - (k_1 y_1 + k_2 y_2 + k_3 y_3)} \right) \right\}. \tag{6.26}$$

The right hand side of the above identity is a holomorphic function that has s^3 factors in the product. Let R be a real number such that $|w| < R$ and

$$|z - (k_1 y_1 + k2 y_2 + k_3 y_3)| \geq R/2$$

for all z on the circle of radius R. Applying the Maximum Modulus Principle (specifically, Corollary B.2) to the right hand side of (6.26) on the disk of radius R, we conclude that it assumes its maximum value on the boundary, i.e. on the circle of radius R, and hence

$$|F(w)| \leq |F|_R (C_3 s/R)^{s^3},$$

for some constant C_3, where $|F|_R$, the maximum of $F(z)$ on the circle of radius R, can be estimated as follows:

$$|F|_R \leq C_4 e^{c_1 n^{5/2} + c_2 r R} r^2,$$

for some constants C_4, c_2, and with c_1 as above. Taking $R = s^{3/2}$, recalling that $r^2 = 64n^3$, and combining these inequalities yields:

$$|F(w)| \leq 64n^3 C_4 e^{c_1 n^{5/2} + c_2 8(ns)^{3/2}} \left(\frac{C_3}{\sqrt{s}}\right)^{s^3}.$$

Since $s \geq n$, taking n large will cause a contradiction with (6.25), hence completing the proof. $\square$

6.6. Problems

PROBLEM 6.1. *Let R be a commutative ring with 1. Prove that $R[x_1, \ldots, x_k]$ is a commutative ring with identity under the standard operations of addition and multiplication on these multivariable polynomials.*

PROBLEM 6.2. *Let K be a field. Write $\boldsymbol{x}$ for the variable vector $(x_1, \ldots, x_k)$, and prove that $K(\boldsymbol{x})$ is a field under the standard operations of addition and multiplication of rational functions:*

$$\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} + \frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} = \frac{p(\boldsymbol{x})g(\boldsymbol{x}) + f(\boldsymbol{x})q(\boldsymbol{x})}{q(\boldsymbol{x})g(\boldsymbol{x})}, \quad \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \cdot \frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} = \frac{p(\boldsymbol{x})f(\boldsymbol{x})}{q(\boldsymbol{x})g(\boldsymbol{x})}.$$

PROBLEM 6.3. *Prove that S_n, the set of all permutations of the set of n elements, is a group under the operation of function composition.*

PROBLEM 6.4. *Let K be a subfield of $\mathbb{C}$. Recall that an embedding of K into $\mathbb{C}$ is an injective field homomorphism $\tau : K \to \mathbb{C}$.*

1. *Suppose that $\tau : K \to \mathbb{C}$ is a field homomorphism such that for some $a \in K$, $\tau(a) \neq 0$. Prove that τ is an embedding.*
2. *Let $K = \mathbb{Q}(\alpha)$ for some $\alpha \in \mathbb{C}$. Prove that α is transcendental if and only if there exist infinitely many distinct embeddings of K into $\mathbb{C}$.*

PROBLEM 6.5. *Let R be a subring of the ring S, both commutative rings with identity. Let $\alpha \in S$, and define*

$$R[\alpha] = \{f(\alpha) : f(x) \in R[x]\}.$$

More generally, for $\alpha_1, \ldots, \alpha_n \in S$ define recursively

$$R[\alpha_1, \ldots, \alpha_n] = R_{n-1}[\alpha_n],$$

where $R_{n-1} = R[\alpha_1, \ldots, \alpha_{n-1}]$. A subring T of S is called a finitely generated ring extension of R if $T = R[\alpha_1, \ldots, \alpha_n]$ for some $\alpha_1, \ldots, \alpha_n \in S$.

1. *Prove that $R[\alpha]$ is a subring of S.*
2. *Prove that $R[\alpha]$ is the smallest ring containing R and α, with respect to inclusion.*
3. *Let $T = \mathbb{Z}\left[\frac{1}{n}\right]$ for an integer $n > 1$. Compute the group of units $T^\times$ of T.*
4. *Prove that $\mathbb{Q}$ is not a finitely generated ring extension of $\mathbb{Z}$.*

PROBLEM 6.6. *Let $\alpha, \beta \in \mathbb{C}$, both nonzero, let $\gamma = \alpha/\beta$, and suppose $\mathbb{Q}(\alpha) \cong \mathbb{Q}(\beta)$.*

1. *Give necessary and sufficient conditions on α, β so that $[\mathbb{Q}(\gamma) : \mathbb{Q}] < \infty$.*
2. *Suppose α and β are algebraic and $\gamma \in \mathbb{Q}$. Let $f(x) = \sum_{k=0} a_k x^k$ be the minimal polynomial of α. Determine the minimal polynomial $g(x)$ of β.*
3. *Suppose α and β are algebraic and $f(x) = g(x)$, i.e. they have the same minimal polynomial. Is it true that $\gamma \in \mathbb{Q}$?*

CHAPTER 7

Further Topics

In the previous chapters we have given an introduction to the geometric topics in number theory, many of which stem from the pioneering work of Minkowski. In fact, geometric ideas underline many different directions in arithmetic. In this final chapter we will briefly mention several topics for further exploration and provide references to more advanced reading on these subjects. Many results in this chapter will be stated without proof.

7.1. Frobenius problem

Let us start with a simple binary linear Diophantine equation of the form

$$ax + by = c, \tag{7.1}$$

in which a, b, c are nonzero integers. There are always rational solutions to (7.1). For which values of a, b, c does it have solutions in integers x, y? The greatest common divisor provides a criterion for the existence of solutions.

LEMMA 7.1.1. *Let a, b, c be nonzero integers. Then* (7.1) *has a solution in integers x, y if and only if* $\gcd(a, b)|c$.

PROOF. ($\Rightarrow$) Suppose that $ax + by = c$ for some $x, y \in \mathbb{Z}$. Since $\gcd(a, b)$ divides a and b, it divides $ax + by = c$.
($\Leftarrow$) If $\gcd(a, b)|c$, write $c = d\gcd(a, b)$, in which $d \in \mathbb{Z}$. By Euclid's Division Lemma, there exist $x', y' \in \mathbb{Z}$ such that $ax'+by' = \gcd(a, b)$. Thus, $a(dx')+b(dy') = d(ax' + by') = d\gcd(a, b) = c$ and hence (7.1) has integer solutions $x = dx'$ and $y = dy'$. □

In fact, we can classify all integer solutions to (7.1).

THEOREM 7.1.2. *Let a, b, c be nonzero integers, and let $d = \gcd(a, b)$. Assume $d|c$. Then the equation $ax + by = c$ has infinitely many integer solutions. In fact, if x_0, y_0 is one such solution pair, then all solutions are given by*

$$x_t = x_0 - t\frac{b}{d},\ y_t = y_0 + t\frac{a}{d} \tag{7.2}$$

as t ranges over all the integers.

PROOF. First let $t \in \mathbb{Z}$ and x_t, y_t be as in (7.2). Then

$$ax_t + by_t = a\left(x_0 - t\frac{b}{d}\right) + b\left(y_0 + t\frac{a}{d}\right) \quad = (ax_0 + by_0) + t\left(\frac{ab}{d} - \frac{ab}{d}\right) = c,$$

hence our pair x, y is a solution to (7.1) for any $t \in \mathbb{Z}$.

We now show that any solution is of this form. Indeed, suppose x, y is a solution pair, then

$$ax_0 + by_0 = c = ax + by,$$

and so

$$a(x_0 - x) = b(y - y_0).$$

Let us divide both sides of the above equation by d and write $a' = a/d$, $b' = b/d$, then $\gcd(a', b') = 1$ and

$$a'(x_0 - x) = b'(y - y_0).$$

Then Euclid's Lemma implies that $a'|y - y_0$ and $b'|x_0 - x$, say $a' = \frac{y-y_0}{t}$ and $b' = \frac{x_0-x}{s}$ for some integers t and s. Then we have

$$\frac{(y - y_0)(x_0 - x)}{t} = \frac{(x_0 - x)(y - y_0)}{s},$$

and so $s = t$. Therefore we obtain

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

which is precisely what we wanted. □

COROLLARY 7.1.3. *If $\gcd(a, b) = 1$, then for any c the equation $ax + by = c$ has infinitely many solutions. Furthermore, if x_0, y_0 is one such solution pair, then all solutions are of the form*

$$x_t = x_0 - tb,\ y_t = y_0 + ta$$

for $t \in \mathbb{Z}$.

EXAMPLE 7.1.1. *Let $a = 4$, $b = 6$, $c = 9$. Since $\gcd(a, b) = 2 \nmid 9$, the equations $4x + 6y = 9$ has no integer solutions. On the other hand, if $c = 10$, then $\gcd(a, b)|c$, and so the equation $4x + 6y = 10$ has infinitely many integer solutions. Since $x = 1$, $y = 1$ is one such solution, all solutions are of the form*

$$x_t = 1 - 3t, y_t = 1 + 2t$$

as t ranges over all the integers.

These observations also have a simple geometric interpretation. Notice that the set of integer solution pairs to (7.1)

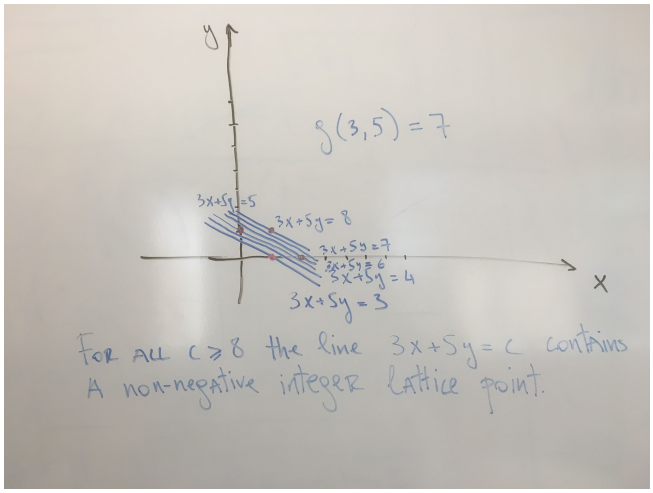
$$\{(x, y) \in \mathbb{Z}^2 : ax + by = c\}$$

is the set of all integer lattice points on the line given by the equation (7.1) in the Euclidean plane. For instance, the set of all such points in the case $a = 4, b = 6, c = 10$ of Example 7.1.1 is $\{(1 - 6t, 1 + 4t) : t \in \mathbb{Z}\}$.

Assume now that $c > 0$ and $\gcd(a, b)$ divides c, so the line $ax + by = c$ contains infinitely many integer lattice points, but does it necessarily contain any such points with nonnegative coordinates? As we will show in the next section, this question has some rather interesting applications. Upon a quick inspection, we can see for instance that the line

$$3x + 5y = c \tag{7.3}$$

contains integer lattice points for any c, but no such points with $x, y \geq 0$ when $c = 1, 2, 4$. For which values of c is our line guaranteed to have nonnegative integer points?



Here is an initial observation, which follows from Theorem 7.1.2 via a geometric argument.

COROLLARY 7.1.4. *Let a, b, c be positive integers with $d := \gcd(a, b)$ dividing c. If $c \geq ab/d$, then the equation (7.1) has integer solution pairs $x, y \geq 0$.*

PROOF. Let $t, s \in \mathbb{Z}$ and consider the solution pairs (x_t, y_t) and (x_s, y_s), as in (7.2), where (x_0, y_0) is some fixed solution pair. Notice that the Euclidean distance between the points (x_t, y_t) and (x_s, y_s) is

$$\sqrt{(x_t - x_s)^2 + (y_t - y_s)^2} = \sqrt{\frac{b^2}{d^2}(t-s)^2 + \frac{a^2}{d^2}(t-s)^2} = \frac{|t-s|\sqrt{a^2+b^2}}{d},$$

which is minimized when $|t - s| = 1$. Let $\ell_{a,b}(c)$ be the line $ax + by = c$ in the Euclidean plane, then the minimal distance between two integer lattice points on $\ell_{a,b}(c)$ is $\frac{\sqrt{a^2+b^2}}{d}$, which is assumed for any neighboring pair of integer lattice points (x_t, y_t) and (x_{t+1}, y_{t+1}). Notice that the intersection of the line $\ell_{a,b}(c)$ with the positive quadrant

$$\{(x, y) \in \mathbb{R}^2 : x, y \geq 0\}$$

is a line segment with endpoints $(c/a, 0)$ and $(0, c/b)$, so the length of this line segment is

$$\sqrt{\frac{c^2}{a^2} + \frac{c^2}{b^2}} = \frac{c\sqrt{a^2+b^2}}{ab}.$$

If the length of this line segment is no less than the distance between the neighboring integer lattice points, then the line segment must contain at least one integer lattice point. This means that when

$$\frac{c\sqrt{a^2+b^2}}{ab} \geq \frac{\sqrt{a^2+b^2}}{d},$$

the equation (7.1) has integer solution pairs $x, y \geq 0$. This happens when $c \geq ab/d$. □

Going back to the example of equation (7.3) and applying Corollary 7.1.4, we are guaranteed that there are nonnegative solutions at least for all $c \geq 15$. Checking by hand, we quickly see that in fact there are nonnegative solutions already for all

$c \geq 8$, suggesting that the bound of Corollary 7.1.4 may not be very good. Indeed, we can obtain more precise results.

Let a, b be relatively prime positive integers, and suppose that we have unlimited supply of coins of denominations a and b. What is the maximal amount of change which we *cannot* give with such coins?

At first it may not seem clear that such a maximal impossible amount of change even exists. Notice, however, that if we use x coins of denomination a and y coins of denomination b, then the total amount of change we are giving is $ax+by$. Hence it is possible to give change in the amount of c if and only if the equation

$$ax + by = c$$

has a nonnegative integer solution x, y. Since $\gcd(a, b) = 1$, we know from Corollary 7.1.4 that is certainly possible at least for all $c \geq ab$. Hence the maximal impossible amount of change must be no bigger than $ab - 1$. But is there an exact formula?

This problem, although possibly in different terms was mentioned in the lectures of a famous German mathematician Ferdinand Georg Frobenius in the late 1800s, although Frobenius himself never published anything in these regards. Nonetheless, this problem became known as the (binary) Frobenius coin exchange problem with the maximal impossible amount of change denoted $g(a, b)$ and called the *Frobenius number* of a and b. Interestingly, closely related problems also appear in recreational mathematical literature under different names, such as postage stamp problem or the chicken McNugget problem. The origins of the latter name are curious: apparently, in the 1980s chicken McNuggets were sold by McDonalds in the UK in boxes of 3, 6 and 20 pieces, prompting a mathematician Henri Picciotto to ask what is the maximal number of nuggets that cannot be purchased (and then answering his own question – it is 43).

Let us now derive a formula for the binary Frobenius number.

THEOREM 7.1.5. *Let* $\gcd(a, b) = 1$*, then*

$$g(a, b) = (a - 1)(b - 1) - 1.$$

In other words, this is the largest number that cannot be represented as $ax+by$ *with* x, y *nonnegative integers.*

PROOF. Since a and b are relatively prime, for every $c \in \mathbb{Z}$ there exist $x, y \in \mathbb{Z}$ such that

$$c = ax + by.$$

We will say that c is *representable* in terms of a and b if there exist such $x, y \geq 0$. Notice in fact that we can assume without loss of generality that $0 \leq x < b$: if $x \geq b$, then $x = nb + x'$ for some $n, x' \in \mathbb{Z}$ with $0 \leq x' < b$, and so

$$c = a(nb + x') + by = ax' + b(an + y),$$

meaning that we can replace x with x' by replacing y with $an + y$, if necessary.

Now, if $0 \leq x < b$, then for every c there is a unique pair (x, y) such that $c = ax + by$, and so c is representable if and only if $y \geq 0$. Notice then that the largest non-representable c corresponds to the largest choice of x (namely, $x = b-1$) and the largest negative choice of y (namely, $y = -1$). This means that the largest non-representable integer is

$$g(a, b) = a(b - 1) + b(-1) = ab - a - b = (a - 1)(b - 1) - 1.$$

□

Theorem 7.1.5 therefore guarantees that for every $c > ab - a - b$ the line $ax + by = c$ contains a nonnegative integer lattice point, however for $c < ab - a - b$ such a point may or may not exist. Revisiting for instance our example (7.3), we see that while $g(3, 5) = 7$, the equation $3x + 5y = c$ has nonnegative integer solutions for $c = 3, 5, 6$, but does not for $c = 1, 2, 4, 7$. Non-representable positive integers with respect to relatively prime a and b are often called *gaps*, so the Frobenius number $g(a, b)$ is the largest gap. Given a and b, how many gaps are there? This natural question was asked as a challenge problem in a journal called Educational Times by James Joseph Sylvester in 1884. Specifically, Sylvester, who has already obtained and published the answer himself in 1882, asked for a proof that this number is equal to $\frac{1}{2}(a-1)(b-1)$; in other words, out of $(a-1)(b-1)-1$ integers between 1 and the Frobenius number $g(a, b)$ about half are non-representable. A clever solution was produced by W. J. Curran Sharp. We prove this result here.

THEOREM 7.1.6. *The number of gaps with respect to a relatively prime pair of positive integers a and b is*

$$\frac{1}{2}(a-1)(b-1).$$

PROOF. Let $0 \le c \le g(a, b)$, and define

$$c' = g(a, b) - c = ab - a - b - c.$$

By our argument in the proof of Theorem 7.1.5, there must exist the unique integers x, y with $0 \le x < b$ such that $c = ax + by$, then

$$c' = ab - a - b - c = ab - a - b - ax - by = ax' + by',$$

where $x' = b - x - 1$ and $y' = -y - 1$. Since $0 \le x' < b$, we see that y' must also be unique.

Suppose that c is representable by a and b (including $c = 0$), then $y \ge 0$, and $y' < 0$, hence c' is not representable. On the other hand, assume that c is not representable, then $y < 0$, and so $y' \ge 0$, meaning that c' is representable. It is clear that c and c' are in a bijection with each other, and $c = c'$ if and only if

$$c = \frac{1}{2}(ab - a - b),$$

but this cannot be an integer, since a and b cannot both be even. Hence precisely a half of $g(a, b) + 1$ integers between 0 and $g(a, b)$ are representable and the rest are gaps, meaning that there are

$$\frac{1}{2}(g(a, b) + 1) = \frac{1}{2}(a-1)(b-1)$$

gaps. □

The Frobenius number has also been defined more generally. Let $n \ge 2$ be an integer and let

$$1 < a_1 < \cdots < a_n \tag{7.4}$$

be relatively prime integers. We say that a positive integer t is *representable* by the n-tuple $\boldsymbol{a} := (a_1, \ldots, a_n)$ if

$$t = a_1 x_1 + \cdots + a_n x_n \tag{7.5}$$

for some nonnegative integers $x_1, \ldots, x_n$, and we call each such solution $\boldsymbol{x} := (x_1, \ldots, x_n)$ of (7.5) a *representation for t in terms of $\boldsymbol{a}$*. Let $s \geq 0$ be an integer, then the *s-Frobenius number* of this n-tuple, $g_s(\boldsymbol{a})$, as defined by Beck and Robins in [**BR04**], is the largest positive integer that has at most s distinct representations in terms of $\boldsymbol{a}$. In the binary case ($n = 2$), Beck and Robins proved the following natural generalization of Theorem 7.1.5.

THEOREM 7.1.7. *Let* $\gcd(a, b) = 1$ *and* $s \geq 0$, *then*

$$g_s(a, b) = (s + 1)ab - (a + b).$$

In the case $s = 0$, the formula of Theorem 7.1.5 is recovered.

This is a generalization of the classical Frobenius number $g_0(\boldsymbol{a})$, i.e., the largest positive integer that has no such representations. The Frobenius number has been studied extensively by a variety of authors, starting as early as late 19th century; see [**Ram05**] for a detailed account and bibliography. The condition

$$\gcd(a_1, \ldots, a_n) = 1 \tag{7.6}$$

implies that $g_s(\boldsymbol{a})$ exists for every s. The algorithmic *Frobenius problem*, known to be NP-hard, is to determine g_0 (or more generally g_s for $s \geq 1$) given n and the relatively prime n-tuple $a_1, \ldots, a_n$ on the input. The hardness of this problem in particular implies that no general closed form formulas for the Frobenius numbers exist, sparking interest in upper and lower bounds.

A geometric approach to the classical Frobenius problem has been pioneered in the influential paper of R. Kannan [**Kan92**], leading to a polynomial-time algorithm to find the Frobenius number for each fixed n. Bounds on the classical Frobenius number stemming from further geometry of numbers applications have been obtained in [**FR07**] and [**AG07**]. These ideas have also been extended to the more general s-Frobenius problem in [**FS11**] and [**AFH12**]. A higher-dimensional analogue of the Frobenius problem has also been considered in the recent years by several authors, notably in [**AH10**], [**AHL13**], and [**ALL16**]. A generalization of this problem to certain number fields has been studied in [**FS20**].

Let us briefly describe Kannan's approach to the Frobenius problem. Let

$$L_{\boldsymbol{a}} = \left\{ \boldsymbol{x} \in \mathbb{Z}^{n-1} : \sum_{i=1}^{n-1} a_i x_i \equiv 0 \pmod{a_n} \right\},$$

then $L_{\boldsymbol{a}}$ is a sublattice of $\mathbb{Z}^{n-1}$ of full rank. Define also a simplex

$$S_{\boldsymbol{a}} = \left\{ \boldsymbol{x} \in \mathbb{R}^{n-1}_{\geq 0} : \sum_{i=1}^{n-1} a_i x_i \leq 1 \right\}.$$

With this notation, Kannan proves the following remarkable identity.

THEOREM 7.1.8.

$$g_0(\boldsymbol{a}) = \mu(S_{\boldsymbol{a}}, L_{\boldsymbol{a}}) - \sum_{i=1}^{n} a_i. \tag{7.7}$$

where $\mu(S_{\boldsymbol{a}}, L_{\boldsymbol{a}})$ is the inhomogeneous minimum (also known as the covering radius) of $S_{\boldsymbol{a}}$ with respect to L, namely

$$\mu(S_{\boldsymbol{a}}, L_{\boldsymbol{a}}) = \inf \left\{ t \in \mathbb{R}_{>0} : tS_{\boldsymbol{a}} + L_{\boldsymbol{a}} = \mathbb{R}^{n-1} \right\}. \tag{7.8}$$

On the other hand, Kannan showed that in every fixed dimension n there is a polynomial-time algorithm to find the covering radius, given $S_{\boldsymbol{a}}$ and $L_{\boldsymbol{a}}$ (which is to say, given $\boldsymbol{a}$). This result, along with his identity (7.7) implies a polynomial-time algorithm for the Frobenius number in fixed dimension. Kannan's Theorem 7.1.8 has been extended to the s-Frobenius numbers in [**AFH12**]. For integer $s \geq 1$, define

$$\mu_s(S_{\boldsymbol{a}}, L_{\boldsymbol{a}}) = \min\{t > 0 : \forall\ \boldsymbol{x} \in \mathbb{R}^n\ \exists\ \boldsymbol{b}_1, \ldots, \boldsymbol{b}_s \in L_{\boldsymbol{a}} \text{ s.t. } \boldsymbol{x} \in \boldsymbol{b}_i + tS_{\boldsymbol{a}}\} \tag{7.9}$$

be the smallest positive number t such that any $\boldsymbol{x} \in \mathbb{R}^n$ is covered by at least s lattice translates of $tS_{\boldsymbol{a}}$: this $\mu_s(S_{\boldsymbol{a}}, L_{\boldsymbol{a}})$ is called the *s-covering radius* of $S_{\boldsymbol{a}}$ with respect to $L_{\boldsymbol{a}}$. If $s = 1$, this is precisely the classical covering radius as in (7.8). With this notation, the following theorem is established in [**AFH12**].

THEOREM 7.1.9.

$$g_s(\boldsymbol{a}) = \mu_{s+1}(S_{\boldsymbol{a}}, L_{\boldsymbol{a}}) - \sum_{i=1}^{n} a_i.$$

Such geometric ideas have also been used by different authors to give expected values of Frobenius numbers with respect to the uniform probability distribution on ensembles of vectors in $\mathbb{Z}^n$ defined with respect to different norms; see [**Arn99**], [**Arn06**], [**AH09**], [**AHH11**], [**BS07**], [**Li15**], [**Mar10**], [**Str12**], [**SSU09**], [**Ust10**], and [**AFH12**] for results on average behavior of Frobenius numbers.

Frobenius numbers and their various generalizations tend to play an important role in several areas of mathematics, including theory of numerical semigroups, commutative algebra, algebraic geometry, number theory, combinatorics, operations research, and theoretical computer science, to name a few. The literature on this subject is vast with a large number of relevant references available in the bibliography to the book [**Ram05**].

7.2. Lattice point counting in homogeneously expanding domains

Another famous optimization problem closely related to the Frobenius problem is known as the *integer knapsack problem*. Let $n \geq 2$ be an integer, $\boldsymbol{a} \in \mathbb{Z}^n_{>0}$ with

$$a_1 < \cdots < a_n, \quad \gcd(a_1, \ldots, a_n) = 1,$$

and $b \in \mathbb{Z}_{>0}$. The corresponding *knapsack polytope* is defined as

$$P(\boldsymbol{a}, b) := \left\{ \boldsymbol{x} \in \mathbb{R}^n_{>0} : \sum_{i=1}^{n} a_i x_i = b \right\}.$$

The problem then is to determine whether $P(\boldsymbol{a}, b) \cap \mathbb{Z}^n = \emptyset$? Here we can think of $a_1, \ldots, a_n$ as weights of different types of objects, then positive integer values of $x_1, \ldots, x_n$ stand for corresponding numbers of objects of each type, and b is the total weight of a knapsack with x_i objects of weight a_i, $1 \leq i \leq n$. Once it is known that for some given values of the total weight b the knapsack is not empty, one can maximize the cost function of this knapsack, provided the objects of different types have assigned prices. This problem comes up frequently in the fields of operations research and resource allocation.

Similar to the Frobenius problem, the integer knapsack problem is also known to be NP-hard. In fact, the Frobenius problem can be stated in terms of the knapsack polytopes as follows: find the smallest positive integer g so that $P(\boldsymbol{a}, b) \cap \mathbb{Z}^n \neq \emptyset$ for every $b > g$.

Notice that we can actually state a more general problem: count the number of integer lattice points in a given knapsack polytope, i.e. determine the cardinality of the set $P(\boldsymbol{a}, b) \cap \mathbb{Z}^n$. Then integer knapsack problem becomes:

$$\text{Is } |P(\boldsymbol{a}, b) \cap \mathbb{Z}^n| = 0?$$

and the Frobenius problem becomes:

$$\text{Find } \min\{g \in \mathbb{Z}_{>0} : |P(\boldsymbol{a}, b) \cap \mathbb{Z}^n| > 0 \ \forall\ b > g\}.$$

Notice also that

$$\sum_{i=1}^{n} a_i x_i = b \iff \sum_{i=1}^{n} a_i \left(\frac{x_i}{b}\right) = 1,$$

i.e. $\boldsymbol{x} \in P(\boldsymbol{a}, b)$ if and only if $\frac{1}{b}\boldsymbol{x} \in P(\boldsymbol{a}, 1)$, meaning that $P(\boldsymbol{a}, b) = bP(\boldsymbol{a}, 1)$, a homogeneous expansion of the polytope. This does not necessarily apply directly to the integer points, since for $\boldsymbol{x} \in \mathbb{Z}^n$ the rescaled points $\frac{1}{b}\boldsymbol{x}$ may no longer be in $\mathbb{Z}^n$. Still, it suggests a natural question: how can we count the number of integer lattice points in homogeneous expansions of polytopes? An area of mathematics that aims to answer this question is Ehrhart theory.

Let $P \subseteq \mathbb{R}^n$ be a convex polytope such that $\mathrm{Vol}(P) > 0$, and vertices of P are points of $\mathbb{Z}^n$: such P is called a *lattice polytope*. Write

$$G_P(t) = |tP \cap \mathbb{Z}^n|.$$

We want to understand the behaviour of $G_P(t)$ for all $t \in \mathbb{Z}_{>0}$; specifically, we will prove a famous theorem of Ehrhart, which states that $G_P(t)$ is a polynomial in t. Our presentation closely follows [**Ewa96**]. First we consider a special case of polytopes, namely simplices.

LEMMA 7.2.1. *Let $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n \in \mathbb{Z}^n$ be linearly independent, and define the simplex*

$$S = \mathrm{Co}(\mathbf{0}, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_n) = \left\{ \sum_{i=1}^{n} t_i \boldsymbol{a}_i : t_i \geq 0 \ \forall \ 1 \leq i \leq n, \ \sum_{i=1}^{n} t_i \leq 1 \right\}.$$

Then there exist $\beta_1, \ldots, \beta_n \in \mathbb{Z}_{\geq 0}$ such that for every $t \in \mathbb{Z}_{>0}$, we have

$$G(tS) = |tS \cap \mathbb{Z}^n| = \binom{n+t}{n} + \sum_{i=1}^{n} \binom{n+t-i}{n} \beta_i.$$

PROOF. Let A be the half-open parallelotope spanned by the vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$, i.e.

$$A = \left\{ \sum_{i=1}^{n} t_i \boldsymbol{a}_i : 0 \leq t_i < 1 \ \forall \ 1 \leq i \leq n \right\}.$$

For every $\boldsymbol{y} \in tS \cap \mathbb{Z}^n$ there exists a unique representation of $\boldsymbol{y}$ of the form

$$\boldsymbol{y} = \boldsymbol{x} + \sum_{i=1}^{n} \alpha_i \boldsymbol{a}_i, \tag{7.10}$$

where $\boldsymbol{x} \in A \cap \mathbb{Z}^n$ and $\alpha_1, \ldots, \alpha_n \in \mathbb{Z}_{\geq 0}$. For each $0 \leq j \leq t$, let H_j be the hyperplane which passes through the points $j\boldsymbol{a}_1, \ldots, j\boldsymbol{a}_n$. We will determine the number of points of $\mathbb{Z}^n$ in $H_j \cap tS$, and the number of points of $\mathbb{Z}^n \cap tS$ in the strips of space bounded by H_{j-1} and H_j for each $1 \leq j \leq t$; notice that $H_0 = \{\mathbf{0}\}$.

First, let $\boldsymbol{x} = \mathbf{0}$ in (7.10). Then $\boldsymbol{y}$ as in (7.10) lies in H_j if and only if

$$\sum_{i=1}^{n} \alpha_i = j, \ 0 \leq \alpha_i \leq j \ \forall \ 1 \leq i \leq n. \tag{7.11}$$

We will prove now that there are precisely $\binom{n+j-1}{n-1}$ possibilities for $\alpha_1, \ldots, \alpha_n$ satisfying (7.11) for each j. We argue by induction on n. If $n = 1$, then there is only $1 = \binom{j}{0}$ possibility. Suppose the claim is true for $n-1$. Then there are $\binom{n+(j-\alpha_n)-2}{n-2}$ possibilities for $\alpha_1, \ldots, \alpha_{n-1}$ such that

$$\sum_{i=1}^{n-1} \alpha_i = j - \alpha_n$$

for each value of $0 \leq \alpha_n \leq j$. Then the number of possibilities for $\alpha_1, \ldots, \alpha_n$ satisfying (7.11) is

$$\sum_{\alpha_n=0}^{j} \binom{n + (j - \alpha_n) - 2}{n-2} = \sum_{i=0}^{j} \binom{n+i-2}{n-2}. \tag{7.12}$$

Then our claim follows by combining (7.12) with the result of Problem 7.2:

$$\sum_{i=0}^{j} \binom{n+i-2}{n-2} = \binom{n+j-1}{n-1}.$$

Now to find the number of points $\boldsymbol{y}$ as in (7.10) with $\boldsymbol{x} = \mathbf{0}$ on $\bigcup_{j=0}^{t} H_j$, we sum over j, using the result of Excercise 7.2 once again:

$$\sum_{j=0}^{t} \binom{n+j-1}{n-1} = \binom{n+t}{n}.$$

If $\boldsymbol{x}$ in (7.10) lies properly between H_0 and H_1, then the number of possible $\boldsymbol{y}$ as given by (7.10) that lie in $\bigcup_{j=0}^{t} H_j$ reduces to $\binom{n+t-1}{n}$. Similarly, the number of possibilities for $\boldsymbol{y}$ as in (7.10) with $\boldsymbol{x}$ lying properly between H_{i-1} and H_i or on H_i is $\binom{n+t-i}{n}$ for each $1 \leq i \leq n$. Therefore, if β_i is the number of points $\boldsymbol{x} \in A \cap \mathbb{Z}^n$ which lie properly between H_{i-1} and H_i or on H_i, then the number of corresponding points $\boldsymbol{y}$ as in (7.10) is

$$\binom{n+t-i}{n}\beta_i.$$

Finally, in the case $t < n$, we let $\beta_i = 0$ for each $t+1 \leq i \leq n$. The statement of the lemma follows.

□

Let $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n \in \mathbb{Z}^n$ be linearly independent, and let

$$S = \mathrm{Co}(\mathbf{0}, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)$$

be the simplex as in Lemma 7.2.1. Define the *pseudo-simplex* associated with S

$$S_0 = S \setminus (\mathrm{Co}(\mathbf{0}, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_{n-1}) \ \cup \ \ldots \ \cup \ \mathrm{Co}(\mathbf{0}, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)).$$

LEMMA 7.2.2. *$G(tS_0)$ is a polynomial in $t \in \mathbb{Z}_{\geq 0}$.*

PROOF. We argue by induction on dimension of S_0. If $\dim(S_0) = 0$, there is nothing to prove, so assume the lemma is true for pseudo-simplices of dimension $< n$. Let $F^{(1)}, \ldots, F^{(s)}$ be proper faces of S which contain $\mathbf{0}$ and satisfy

$$0 < \dim(F^{(i)}) < n, \ \forall \ 1 \leq i \leq s.$$

Then

$$S \setminus S_0 = \{\mathbf{0}\} \ \cup F_0^{(1)} \ \cup \ \ldots \ \cup \ F_0^{(s)}$$

is a disjoint union. By induction hypothesis,

$$G(t(S \setminus S_0)) = 1 + G(tF_0^{(1)}) + \cdots + G(tF_0^{(s)})$$

is a polynomial in t. Hence, by Lemma 7.2.1,

$$G(tS_0) = G(tS) - G(t(S \setminus S_0)) = G(tS) - 1 - G(tF_0^{(1)}) - \cdots - G(tF_0^{(s)})$$

is a polynomial in t. □

We are now ready to prove Ehrhart's theorem.

THEOREM 7.2.3 (Ehrhart). *Let P be a lattice polytope in $\mathbb{R}^n$. Then $G_P(t)$ is a polynomial in $t \in \mathbb{Z}_{\geq 0}$.*

PROOF. We can assume $\mathbf{0}$ to be a vertex of P, since such translation would not change the number of integer lattice points. Notice that each $(n-1)$-dimensional face of P which does not contain $\mathbf{0}$ can be given a decomposition as a simplicial complex whose 0-cells are the vertices of this face. We can then join each simplex, obtained in this manner, to $\mathbf{0}$ resulting in a decomposition of P into a simplicial complex whose 0-cells are precisely the vertices of P. Then P can be represented as a disjoint union

$$P = \{\mathbf{0}\} \ \cup S_0^{(1)} \ \cup \ \ldots \ \cup \ S_0^{(r)},$$

where $S_0^{(1)}, \ldots, S_0^{(r)}$ are precisely the cells of this simplicial complex which contain $\mathbf{0}$, but are not equal to $\{\mathbf{0}\}$. The theorem follows by Lemma 7.2.2. □

$G_P(t)$ as in Theorem 7.2.3 is called *Ehrhart polynomial* of P. An excellent reference on Ehrhart polynomials, their many fascinating properties, and connections to other important mathematical objects is [**BR06**]. For a general lattice polytope P very little is known about the coefficients of its Ehrhart polynomial $G_P(t)$. Let

$$G_P(t) = \sum_{i=0}^{n} c_i(P) t^i,$$

then it is known that the leading coefficient $c_n(P)$ is equal to $\mathrm{Vol}(P)$, and $c_{n-1}(P)$ is $(n-1)$-dimensional volume of the boundary ∂P, which is normalized by the determinants of the sublattices induced by the corresponding faces of P. Also, $c_0(P)$ is the combinatorial *Euler characteristic* $\chi(P)$:

$$\chi(P) = \sum_{i=0}^{n} (-1)^i (\text{number of } i-\text{dimensional faces of } P).$$

The rest of the coefficients of $G_P(t)$ are in general unknown, however there are known relations and identities that they satisfy; see [**BR06**] for further details.

Let us present the first simple example of Ehrhart polynomial. Consider the n-dimensional cube of sidelength 2 centered at the origin:

$$C_n = \{\boldsymbol{x} \in \mathbb{R}^n : |\boldsymbol{x}| \leq 1\}, \tag{7.13}$$

then for each $t \in \mathbb{Z}_{>0}$

$$|tC_n \cap \Lambda| = (2t+1)^n = \sum_{i=0}^{n} 2^i \binom{n}{k} t^i$$

is the corresponding Ehrhart polynomial. We will give two more explicit examples of Ehrhart polynomial. The first one is for an open simplex, which is precisely the interior of the simplex S of Lemma 7.2.1 with $\boldsymbol{a}_i = \boldsymbol{e}_i$ for each $1 \leq i \leq n$; the following observation along with the proof is due to S. I. Sobolev.

PROPOSITION 7.2.4. *Define an open simplex*

$$S^\circ = \left\{ \boldsymbol{x} \in \mathbb{R}^n : x_i > 0 \ \forall\ 1 \leq i \leq n,\ \sum_{i=1}^{n} x_i < 1 \right\}.$$

Then $G_{S^\circ}(t) = 0$ *if* $t \leq n$, *and for every* $t \in \mathbb{Z}_{>n}$,

$$G_{S^\circ}(t) = \binom{t-1}{n}. \tag{7.14}$$

PROOF. Let $t > n$, and notice that the simplex tS° can be mapped by an affine transformation to the simplex

$$tS_1^\circ = \{\boldsymbol{x} \in \mathbb{R}^n : 0 < x_1 < \cdots < x_k < t\}.$$

This transformation is volume-preserving and maps $\mathbb{Z}^n$ to itself. Integral points of tS_1° correspond to increasing sequences of integers $0 < y_1 < \cdots < y_n < t$. The number of such sequences is precisely $\binom{t-1}{n}$, which is the number of all possible n-element subsets of the set $\{1, ..., t-1\}$. □

Notice that (7.14) can be thought of as a geometric interpretation of binomial coefficients. The next example is closely related to the one in Proposition 7.2.4: it has been established in [**BCKV00**].

PROPOSITION 7.2.5. *Let*

$$\mathcal{S}_n = \left\{ \boldsymbol{x} \in \mathbb{R}^n : \sum_{i=1}^{n} |x_i| \leq 1 \right\}.$$

Then for every $t \in \mathbb{Z}_{>0}$

$$G_{\mathcal{S}_n}(t) = \sum_{i=0}^{\min\{t,n\}} 2^i \binom{n}{i} \binom{t}{i}. \tag{7.15}$$

PROOF. Notice that for each $0 \leq i \leq \min\{t, n\}$ the number of points in $t\mathcal{S}_n \cap \mathbb{Z}^n$ with precisely i nonzero coordinates is

$$2^i \binom{n}{i} \binom{t}{i}.$$

Indeed, the number of choices of which coordinates are nonzero is $\binom{n}{i}$; for each such choice there are 2^i choices of $\pm$ signs, and $\binom{t}{i}$ choices of absolute values. Summing over all $0 \leq i \leq \min\{t, n\}$ completes the proof. □

REMARK 7.2.1. A remarkable property of the polynomial in Proposition 7.2.5 is that the right hand side (7.15) is symmetric in t and n. This means that

$$|t\mathcal{S}_n \cap \mathbb{Z}^n| = |n\mathcal{S}_t \cap \mathbb{Z}^t|.$$

All of the lattice point counting results above were specifically for integer points in polytopes, which is a rather special class of convex bodies in $\mathbb{R}^n$ and only one lattice. What can be said for more general convex bodies and lattices? Let $M \subseteq \mathbb{R}^n$ be closed, bounded, and Jordan measurable with $\mathrm{Vol}(M) > 0$, and let $\Lambda \subseteq \mathbb{R}^n$ be a lattice of full rank. Suppose we homogeneously expand M by a positive real parameter t, i.e. for each positive real value of t we will consider the set tM. How many points of Λ are there in tM as t grows? To partially answer this question, we will be interested in the *asymptotic behavior* of the function

$$G_{M,\Lambda}(t) = |tM \cap \Lambda|$$

as $t \to \infty$. In general, this is a very difficult question. We will need to make some additional assumptions on M in order to study $G_{M,\Lambda}(t)$.

DEFINITION 7.2.1. Let S be a subset of some Eucildean space. A map

$$\varphi : S \to \mathbb{R}^n$$

is called a *Lipschitz map* if there exists $\mathcal{C} \in \mathbb{R}_{>0}$ such that for all $\boldsymbol{x}, \boldsymbol{y} \in S$

$$\|\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y})\|_2 \leq \mathcal{C}\|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

We say that $\mathcal{C}$ is the corresponding *Lipschitz constant*.

Let C_n be the cube as in (7.13). We say that a set $S \subseteq \mathbb{R}^n$ is *Lipschitz parametrizable* if there exists a finite number of Lipschitz maps

$$\varphi_j : C_n \to S,$$

such that $S = \bigcup_j \varphi_j(C_n)$.

DEFINITION 7.2.2. Let $f(t)$ and $g(t)$ be two functions defined on $\mathbb{R}$. We will say that

$$f(t) = O(g(t)) \text{ as } t \to \infty$$

if there exists a positive real number $\mathcal{B}$ and a real number t_0 such that for all $t \geq t_0$,

$$|f(t)| \leq \mathcal{B}|g(t)|.$$

We usually use the O-notation to emphasize the fact that $f(t)$ behaves similar to $g(t)$ when t is large. This is quite useful if $g(t)$ is a simpler function than $f(t)$; in this case, such a statement helps us to understand the *asymptotic behavior* of $f(t)$, namely its behavior as $t \to \infty$.

Let ∂M be the boundary of M, and assume that ∂M is $(n-1)$-Lipschitz parametrizable. Notice that for $t \in \mathbb{R}_{>0}$, $\partial(tM) = t\partial M$. The following result is Theorem 2 on p. 128 of [**Lan94**].

THEOREM 7.2.6. *Let $t \in \mathbb{R}_{>0}$, then*

$$G_{M,\Lambda}(t) = \frac{\mathrm{Vol}(M)}{\det(\Lambda)} t^n + O(t^{n-1}),$$

where the constant in O-notation depends on Λ, n, and Lipschitz constants.

PROOF. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be a basis for Λ, and let $\mathcal{F}$ be the corresponding fundamental parallelotope, i.e.

$$\mathcal{F} = \left\{ \sum_{i=1}^{n} t_i \boldsymbol{x}_i : 0 \leq t_i < 1, \ \forall\ 1 \leq i \leq n \right\}.$$

For each point $\boldsymbol{x} \in \Lambda$ we will write $\mathcal{F}_{\boldsymbol{x}}$ for the translate of $\mathcal{F}$ by $\boldsymbol{x}$:

$$\mathcal{F}_{\boldsymbol{x}} = \mathcal{F} + \boldsymbol{x}.$$

Notice that if $\boldsymbol{x} \in tM \cap \Lambda$, then $\mathcal{F}_{\boldsymbol{x}} \cap tM \neq \emptyset$. Moreover, either

$$\mathcal{F}_{\boldsymbol{x}} \subseteq \mathrm{int}(tM),$$

or

$$\mathcal{F}_{\boldsymbol{x}} \cap \partial(tM) \neq \emptyset.$$

Let

$$m(t) = \left| \{ \boldsymbol{x} \in \Lambda : \mathcal{F}_{\boldsymbol{x}} \subseteq \mathrm{int}(tM) \} \right|,$$
$$b(t) = \left| \{ \boldsymbol{x} \in \Lambda : \mathcal{F}_{\boldsymbol{x}} \cap \partial(tM) \neq \emptyset \} \right|.$$

Then clearly

$$m(t) \leq G_{M,\Lambda}(t) \leq m(t) + b(t).$$

Moreover, since $\mathrm{Vol}(\mathcal{F}) = \det(\Lambda)$

$$m(t) \det(\Lambda) \leq \mathrm{Vol}(tM) = t^n \mathrm{Vol}(M) \leq (m(t) + b(t)) \det(\Lambda),$$

hence

$$m(t) \leq \frac{\mathrm{Vol}(M)}{\det(\Lambda)} t^n \leq m(t) + b(t).$$

Therefore to conclude the proof we only need to estimate $b(t)$. Let

$$\varphi : C_{n-1} \to \partial M$$

be one of the Lipschitz paramterizing maps for a piece of the boundary of M, and let $\mathcal{C}$ be the maximum of all Lipschitz constants corresponding to these maps. Then $t\varphi$ parametrizes a corresponding piece of $\partial(tM) = t\partial M$. Cut up each side of C_{n-1}

into segments of length $1/[2t]$, then we can represent C_{n-1} as a union of $[t]^{n-1}$ small cubes with sidelength $1/[2t]$ each, call them $C^1, \ldots, C^{[t]^{n-1}}$. For each such C_i, we have

$$\|\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y})\|_2 \leq \mathcal{C}\|\boldsymbol{x} - \boldsymbol{y}\|_2 \leq \frac{\mathcal{C}\sqrt{n-1}}{[2t]},$$

for each $\boldsymbol{x}, \boldsymbol{y} \in C_i$, i.e. the image of each such C_i under φ has diameter at most $\frac{\mathcal{C}\sqrt{n-1}}{[2t]}$. Hence image of each such C_i under the map $t\varphi$ has diameter at most

$$\mathcal{C}\sqrt{n-1}\ \frac{t}{[2t]} \leq 2\ \mathcal{C}\sqrt{n-1}.$$

Clearly therefore the number of $\boldsymbol{x} \in \Lambda$ such that the corresponding translate $\mathcal{F}_{\boldsymbol{x}}$ has nonempty intersection with $t\varphi(C_i)$, for each $1 \leq i \leq [t]^{n-1}$, is bounded by some constant $\mathcal{C}'$ that depends only on Λ, $\mathcal{C}$, and n. Hence

$$b(t) \leq \mathcal{C}'[t]^{n-1}.$$

This completes the proof. □

Theorem 7.2.6 provides an asymptotic formula for $G_{M,\Lambda}(t)$, demonstrating an important general principle, namely that as $t \to \infty$, $G_{M,\Lambda}(t)$ grows like $\frac{\mathrm{Vol}(M)}{\det(\Lambda)}t^n$, which is what one would expect. However, it does not give any explicit information about the constant in the error term $O(t^{n-1})$. Can this constant be somehow bounded, i.e. what can be said about the quantity

$$\left| G_{M,\Lambda}(t) - \frac{\mathrm{Vol}(M)}{\det(\Lambda)}t^n \right| ?$$

A large amount of work has been done in this direction (see for instance pp. 140 - 147 of [**GL87**] for an overview of results and bibliography). This subject essentially originated in a paper of Davenport [**Dav51**], who used a principle of Lipschitz [**Lip65**]; also see [**Thu93**] for a nice overview of Davenport's result and its generalizations and [**Wid12**] for further recent results. We present here without proof a result of P. G. Spain [**Spa95**], which is a refinement of Davenport's bound, and can be thought of as a continuation of Theorem 7.2.6.

THEOREM 7.2.7. *Let the notation be as in Theorem 7.2.6, and let $\mathcal{C}$ be the maximal Lipschitz constant corresponding to parametrization of ∂M. Then for each $t \in \mathbb{R}_{>0}$,*

$$\left| G_{M,\Lambda}(t) - \frac{\mathrm{Vol}(M)}{\det(\Lambda)}t^n \right| \leq 2^n(\mathcal{C}t + 1)^{n-1}.$$

Finally, for very explicit inequalities in the case of counting lattice points in rectangular boxes see [**Fuk06a**], [**Fuk06b**] and [**FH13**].

7.3. Simultaneous Diophantine approximation

A fundamental problem of Diophantine approximation, which we discussed in Chapter 4 is to approximate a given real number by rational numbers with controlled denominators. Dirichlet's theorem of 1842 (Theorem 4.3.1) is the original result in this direction. In fact, we can ask the same question in higher dimensions: given a point $\boldsymbol{\alpha} \in \mathbb{R}^n$, how closely can we approximate it by points in $\mathbb{Q}^n$ with bounded denominators? This question can be answered by Dirichlet's theorem on simultaneous Diophantine approximation, which also dates back to 1842 (our exposition here follows [**Sch80**]).

THEOREM 7.3.1. *Let $m, n \geq 1$ be integers, and let*

$$(\alpha_{i1}, \ldots, \alpha_{in}) \in \mathbb{R}^n$$

for $1 \leq i \leq m$ be m real points. For any integer $Q > 1$ there exist integers

$$q_1, \ldots, q_n, p_1, \ldots, p_m$$

such that $1 \leq \max_{1 \leq i \leq n} |q_i| < Q^{m/n}$ and

$$|\alpha_{j1} q_1 + \ldots \alpha_{jn} q_n - p_j| \leq \frac{1}{Q},$$

for all $1 \leq j \leq m$.

PROOF. While different proofs of this result exist in the literature, we will present a geometric argument based on Minkowski's Linear Forms Theorem: this is a generalization of the Minkowski-style proof of Theorem 4.3.1 that we presented in Chapter 4. Let $\ell = n + m$ and define the following collection of ℓ linear forms in ℓ variables with real coefficients:

$$L_i(x_1, \ldots, x_\ell) = x_i, \ \forall \ 1 \leq i \leq n$$
$$L_{n+j}(x_1, \ldots, x_\ell) = \alpha_{j1} x_1 + \cdots + \alpha_{jn} x_n - x_{n+j}, \ \forall \ 1 \leq j \leq m.$$

Let B be the matrix of coefficients of these linear forms, i.e.

$$B = \begin{pmatrix} 1 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 1 & 0 & \ldots & 0 \\ \alpha_{11} & \ldots & \alpha_{1n} & -1 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \alpha_{m1} & \ldots & \alpha_{mn} & 0 & \ldots & -1 \end{pmatrix}.$$

This is a lower-triangular matrix with ± 1's on the diagonal, and hence $\det(B) = \pm 1$. Then Theorem 1.4.3 guarantees existence of a point $\mathbf{0} \neq \boldsymbol{x} \in \mathbb{Z}^\ell$ such that

$$|L_i(\boldsymbol{x})| < Q^{m/n}, \ \forall \ 1 \leq i \leq n$$
$$|L_{n+j}(\boldsymbol{x})| \leq Q^{-1}, \ \forall \ 1 \leq j \leq m,$$

since $(Q^{m/n})^n (Q^{-1})^m = 1$. Let $q_i = x_i$ for $1 \leq i \leq n$, and $p_j = x_{n+j}$ for $1 \leq j \leq m$. Then we have

$$\max_{1 \leq j \leq n} |q_j| = \max_{1 \leq j \leq n} |L_i(\boldsymbol{x})| < Q^{m/n},$$

and

$$|\alpha_{j1} q_1 + \ldots \alpha_{jn} q_n - p_j| = |L_{n+j}(\boldsymbol{x})| \leq \frac{1}{Q},$$

for all $1 \leq j \leq m$. We now only need to show that

$$\max_{1 \leq j \leq n} |q_j| \geq 1.$$

Assume not, then $q_1, \ldots, q_n = 0$, and then for all $1 \leq j \leq m$,

$$|L_{n+j}(\boldsymbol{x})| = |p_j| \leq 1/Q < 1,$$

and so all the integers p_j must be equal to 0, and hence $\boldsymbol{x} = \mathbf{0}$, which contradicts Minkowski Linear Forms Theorem. This completes the proof. □

COROLLARY 7.3.2. *Let the notation be as in Theorem 7.3.1. Suppose in addition that*

$$\left(\sum_{j=1}^{n} \alpha_{1j} q_j, \ldots, \sum_{j=1}^{n} \alpha_{mj} q_j \right) \in \mathbb{R}^m \setminus \mathbb{Z}^m, \ \forall \ (q_1, \ldots, q_n) \in \mathbb{Z}^n \setminus \{\mathbf{0}\}. \tag{7.16}$$

Then there exist infinitely many co-prime integer $(n+m)$-tuples

$$q_1, \ldots, q_n, p_1, \ldots, p_m$$

such that

$$|\alpha_{j1} q_1 + \ldots \alpha_{jn} q_n - p_j| < \frac{1}{q^{n/m}}, \ \forall \ 1 \leq j \leq m, \tag{7.17}$$

where $q := \max_{1 \leq i \leq n} |q_i| > 0$.

PROOF. Derivation of this corollary from Theorem 7.3.1 is very similar to the Euclid-style exhaustion argument used in the proof of Theorem 4.3.1 to derive (4.3) from (4.2). Let

$$\boldsymbol{z} := (q_1, \ldots, q_n, p_1, \ldots, p_m) \in \mathbb{Z}^{n+m}$$

be the $(n+m)$-tuple with $q < Q^{m/n}$ guaranteed by Theorem 7.3.1. Then

$$|\alpha_{j1} q_1 + \ldots \alpha_{jn} q_n - p_j| \leq \frac{1}{Q} < \frac{1}{q^{n/m}},$$

for all $1 \leq j \leq m$. Suppose that there are only finitely many such $(n+m)$-tuples, call them

$$\boldsymbol{z}_l := (q_{l1}, \ldots, q_{ln}, p_{l1}, \ldots, p_{lm})$$

for $1 \leq l \leq k$. Let

$$\delta = \min_{1 \leq l \leq k, 1 \leq j \leq m} |\alpha_{j1} q_{l1} + \ldots \alpha_{jn} q_{ln} - p_{lj}|,$$

then $\delta > 0$ by (7.16). Let $Q \in \mathbb{Z}_{>0}$ be such that $\frac{1}{Q} < \delta$. By Theorem 7.3.1, there must exist $\boldsymbol{z}' := (q'_1, \ldots, q'_n, p'_1, \ldots, p'_m)$ such that

$$\left|\alpha_{j1} q'_1 + \ldots \alpha_{jn} q'_n - p'_j\right| \leq \frac{1}{Q} < \delta,$$

for all $1 \leq j \leq m$. Hence $\boldsymbol{z}' \notin \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k\}$, which is a contradiciton. Thus there must be infinitely many such $(n+m)$-tuples. □

Notice that taking $m = n = 1$ in Theorem 7.3.1 and Corollary 7.3.2 reduces to the classical Dirichlet's Theorem 4.3.1: in this case, condition (7.16) means precisely that the real number α_{11} is irrational. On the other hand, taking just $n = 1$ gives simultaneous approximation to m real numbers by rationals with the same bounded denominator, where as taking just $m = 1$ results in producing an integer point of bounded sup-norm at which an irrational linear form is close to an integer. For

instance, taking $n = 1$ and $m = 2$ we see that for an irrational point (x, y) in the plane there exist infinitely many rational points $(p/q, r/q)$ such that

$$\max\left\{\left|x - \frac{p}{q}\right|, \left|y - \frac{r}{q}\right|\right\} < \frac{1}{q^{3/2}}.$$

What if our point (x, y) is on some curve, say a unit circle: can we approximate it by rational points on the same circle? This can be done with the use of the following interesting result of E. Hlawka [**Hla80**] on simultaneous Diophantine approximation by quotients of *Pythagorean triples*, i.e. integer solutions to the equation

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

THEOREM 7.3.3. *Let $x \in (0, 1)$ be a real number. Then there exist infinitely many Pythagorean triples $\left(p, \sqrt{q^2 - p^2}, q\right) \in \mathbb{Z}^3$ such that*

$$\left|x - \frac{p}{q}\right| \leq \frac{2\sqrt{2}}{q}. \tag{7.18}$$

We can use Theorem 7.3.3 to approximate points on a unit circle with rational points on the same circle. The following corollary was established in [**Fuk09a**].

COROLLARY 7.3.4. *Let (x, y) be a point on the unit circle. Then either $x, y \in \{0, \pm 1\}$, or there exist infinitely many rational points $(p/q, r/q)$ on the same circle such that*

$$\max\left\{\left|x - \frac{p}{q}\right|, \left|y - \frac{r}{q}\right|\right\} \leq \frac{2\sqrt{2}}{q}. \tag{7.19}$$

PROOF. First notice that it suffices to prove the statement of this corollary for the case $0 < x, y < 1$, namely the case when the point in question lies in the first quadrant, since any other point on the circle can be obtained from those in the first quadrant by a rational rotation. Let c be an arbitrary real number in the interval $(0, 1)$, then either

$$0 < x \leq \sqrt{1 - c^2} < 1,\ c \leq y < 1, \tag{7.20}$$

or

$$0 < y \leq \sqrt{1 - c^2} < 1,\ c \leq x < 1. \tag{7.21}$$

First assume that (7.20) holds. By Theorem 7.3.3, there exist infinitely many Pythagorean triples (p, r, q) with $r = \sqrt{q^2 - p^2}$ which satisfy (7.18). Then:

$$\begin{aligned} \frac{2\sqrt{2}}{q} \geq \left|x - \frac{p}{q}\right| &= \left|\sqrt{1 - y^2} - \sqrt{1 - \frac{r^2}{q^2}}\right| = \frac{\left|\frac{r^2}{q^2} - y^2\right|}{\sqrt{1 - y^2} + \sqrt{1 - \frac{r^2}{q^2}}} \\ &= \frac{\frac{r}{q} + y}{\sqrt{1 - y^2} + \sqrt{1 - \frac{r^2}{q^2}}}\left|y - \frac{r}{q}\right| \geq \frac{c\left(1 + \frac{n}{n+1}\right)}{2\sqrt{1 - \frac{n^2}{(n+1)^2}c^2}}\left|y - \frac{r}{q}\right|. \end{aligned} \tag{7.22}$$

The last inequality is true because $\frac{w+z}{\sqrt{1-w^2}+\sqrt{1-z^2}}$ is an increasing function in both variables for $0 < z, w < 1$; since $y \geq c$, we can pick q large enough so that r/q would have to be sufficiently close to y so that $r/q \geq \frac{n}{n+1}c$ for some $n \in \mathbb{Z}_{>0}$, then

$r/q + y \geq c\left(1 + \frac{n}{n+1}\right)$, and $\sqrt{1-y^2} + \sqrt{1-\frac{r^2}{q^2}} \leq 2\sqrt{1-\frac{n^2}{(n+1)^2}c^2}$. Then (7.22) implies:

$$\left|y - \frac{r}{q}\right| \leq \frac{\sqrt{1-\frac{n^2}{(n+1)^2}c^2}}{c\left(1+\frac{n}{n+1}\right)} \times \frac{4\sqrt{2}}{q}. \tag{7.23}$$

Since our choice of $c \in (0,1)$ and positive integer n was arbitrary, we can for instance choose

$$c = \frac{2n+2}{\sqrt{8n^2+4n+1}}, \tag{7.24}$$

and take $n = 2$, in which case, combining (7.18), (7.23), and (7.24), we obtain (7.19).

If, on the other hand, (7.21) holds instead of (7.20), simply repeat the above argument interchanging x with y and p/q with r/q. This completes the proof. □

A related result has also been obtained by Kopetzky in [**Kop80**] (also see [**Kop81**]), however his bounds are different in flavor in the sense that the constants in the upper bounds depend on x and y. Notice that the bound of Corollary 7.3.4 can be easily extended to any rational ellipse.

COROLLARY 7.3.5. *Let (x,y) be a point on the ellipse E, given by the equation*

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1,$$

where a, b are positive rational numbers. Then either $(x,y) = (\pm a, 0), (0, \pm b)$, or there exist infinitely many rational points $(p/q, r/q)$ on the same ellipse such that

$$\max\left\{\left|x - \frac{p}{q}\right|, \left|y - \frac{r}{q}\right|\right\} \leq \frac{2\sqrt{2}\max\{a,b\}}{q}. \tag{7.25}$$

PROOF. Problem 7.1. □

Stronger and more general results on approximating points on spheres by rational points on the same spheres (with sharper constants, as well as in higher dimensions) have appeared recently in [**KM15**] and [**Mos16**]. Additionally, [**Cas57**], [**Sch80**] and [**Sch91**] are excellent classical references on Diophantine approximation, including simultaneous and higher dimensional approximations.

Another classical problem on simultaneous approximation that we will briefly mention here is Minkowski's Conjecture. Define a function $\mathbb{N} : \mathbb{R}^n \to \mathbb{R}$ by

$$\mathcal{N}(\boldsymbol{x}) = |x_1 \cdots x_n|.$$

This function is inspired by a number field norm under Minkowski embedding: if K is a number field of degree n with all real embeddings $\sigma_1, \ldots, \sigma_n$, then

$$\Sigma = (\sigma_1, \ldots, \sigma_n) : K \to \mathbb{R}^n$$

is its Minkowski embedding, and for any $x \in K \setminus \{0\}$

$$|\mathbb{N}_K(x)| = \mathcal{N}(\Sigma(x)) \neq 0.$$

Suppose we have a fixed point $\boldsymbol{y} \in \mathbb{R}^n$. How closely can we approximate it with respect to $\mathcal{N}$ by points of $\mathbb{Z}^n$? It is not difficult to see that

$$\sup_{\boldsymbol{y} \in \mathbb{R}^n} \inf_{\boldsymbol{x} \in \mathbb{Z}^n} \mathcal{N}(\boldsymbol{y} - \boldsymbol{x}) = 2^{-n}.$$

In other words, there exist points in $\mathbb{R}^n$ that cannot be approximated by points of $\mathbb{Z}^n$ better than up to $\frac{1}{2^n}$ with respect to $\mathcal{N}$. Indeed, if

$$\boldsymbol{y} = \begin{pmatrix} \frac{1}{2} \\ \vdots \\ \frac{1}{2} \end{pmatrix} \in \mathbb{R}^n,$$

then in fact

$$\inf_{\boldsymbol{x} \in \mathbb{Z}^n} \mathcal{N}(\boldsymbol{y} - \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y} - \boldsymbol{0}) = \mathcal{N}(\boldsymbol{y}) = 2^{-n}.$$

Can we do better if we replace $\mathbb{Z}^n$ with perhaps another *unimodular* lattice, i.e. another lattice of determinant 1? Minkowski conjectured in [**Min00**] that at least we will not do any worth. Define a group of diagonal matrices

$$\mathcal{A} = \left\{ \begin{pmatrix} a_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & a_n \end{pmatrix} : a_i > 0,\ a_1 \cdots a_n = 1 \right\}.$$

Conjecture 7.3.1. *For any unimodular lattice $\Lambda \subset \mathbb{R}^n$, we have:*

$$\sup_{\boldsymbol{y} \in \mathbb{R}^n} \inf_{\boldsymbol{x} \in \Lambda} \mathcal{N}(\boldsymbol{y} - \boldsymbol{x}) \leq 2^{-n}. \tag{7.26}$$

Equality holds if and only if there exists $A \in \mathcal{A}$ such that

$$\Lambda = A\mathbb{Z}^n = \operatorname{span}_{\mathbb{Z}}\{a_1 \boldsymbol{e}_1, \dots, a_n \boldsymbol{e}_n\},$$

where $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$ are the standard basis vectors in $\mathbb{R}^n$, and $a_1, \dots, a_n \in \mathbb{R}$ are positive with $a_1 \dots a_n = 1$.

The significance of this conjecture is perhaps best demonstrated by its implication for the *totally real* number fields, i.e. number fields with all real embeddings as above.

Corollary 7.3.6. *Let K be a totally real number field of degree n and discriminant Δ_K. Then for every $x \in K$ there exists an algebraic integer $y \in \mathcal{O}_K$ such that*

$$\mathbb{N}_K(x - y) \leq 2^{-n} \sqrt{|\Delta_K|}.$$

In other words, every element of a totally real number field can be appropriately approximated with respect to norm by an algebraic integer. Minkowski's conjecture has been proved for $n = 2$ by Minkowski himself [**Min00**], for $n = 3$ by Remak in 1923 [**Rem23**], for $n = 4$ by Dyson in 1948 [**Dys48**], for $n = 5$ by Skubenko in 1972 [**Sku72**], and for $n = 6$ by McMullen in 2005 [**McM05**].

Let us very briefly describe McMullen's geometric approach to this problem. His main contribution was to prove that any lattice can be appropriately "rescaled" to a well-rounded one. More specifically, he prove the following result.

Theorem 7.3.7. *For any lattice $\Lambda \subset \mathbb{R}^n$, there exists $A \in \mathcal{A}$ such that the lattice $A\Lambda$ is well-rounded, i.e. has n linearly independent shortest vectors.*

Recall now that covering radius of a lattice Λ in $\mathbb{R}^n$ is

$$\mu(\Lambda) = \sup_{\boldsymbol{y} \in \mathbb{R}^N} \inf_{\boldsymbol{x} \in \Lambda} \|\boldsymbol{y} - \boldsymbol{x}\|.$$

Notice that this is the same expression as in (7.26), but with respect to the Euclidean norm $\| \ \|$ instead of $\mathcal{N}$. By the inequality between arithmetic and geometric means, we have

$$\mathcal{N}(\boldsymbol{x})^{\frac{1}{n}} \leq \frac{\|\boldsymbol{x}\|}{\sqrt{n}}, \tag{7.27}$$

for all $\boldsymbol{x} \in \mathbb{R}^n$. The following bound on the covering radius has been proved by Woods in 1972 [**Woo72**] for well-rounded lattices when $n \leq 6$.

THEOREM 7.3.8. *Let $n \leq 6$, and let $\Lambda \subset \mathbb{R}^n$ be a well-rounded unimodular lattice. Then*

$$\mu(\Lambda) \leq \frac{\sqrt{n}}{2}. \tag{7.28}$$

Equality holds if and only if $\Lambda = B\mathbb{Z}^n$ is isometric to $\mathbb{Z}^n$.

Now, combining Theorem 7.3.7 with (7.27) and Theorem 7.3.8 yields Conjecture 2.2.2 for $n \leq 6$. Hence McMullen's fundamental contribution to Minkowski's conjecture was the understanding of "distribution" of well-rounded lattices among the orbits of lattices in $\mathbb{R}^n$ under the action of the diagonal group $\mathcal{A}$. In fact, Woods conjectured that the inequality (7.28) holds in any dimension n. Woods' conjecture was proved for $n = 7, 8, 9$ ([**HGRS09**], [**HGRS11**], [**KR16**]) hence also establishing Minkowski's conjecture (now in all dimensions $n \leq 9$), by combination with McMullen's theorem. Interestingly, Woods conjecture has been disproved by a family of explicit counterexamples in all dimensions $n \geq 30$ ([**RSW17**]).

7.4. Absolute values and height functions

In this section we introduce the basic machinery of absolute values and heights, which is used to investigate further questions in Diophantine Approximations and Diophantine Geometry.

DEFINITION 7.4.1. Let K be a field. An *absolute value* on K is a function $|\ |: K \to \mathbb{R}_{\geq 0}$ such that for all $x, y \in K$ we have:

(1) $|x| \geq 0$ with equality if and only if $x = 0$,
(2) $|xy| = |x||y|$,
(3) *Triangle inequality:* $|x + y| \leq |x| + |y|$.

Sometimes (3) can be replaced by the stronger property:

(4) *Ultrametric inequality:* $|x + y| \leq \max\{|x|, |y|\}$.

If $|\ |$ satisfies (1), (2), (3), but fails (4), we say that it is *archimedean* absolute value; if it also satisfies (4), it is called *non-archimedean*.

Here is the most basic example of an absolute value on K: it is called the *trivial* absolute value, and is defined by

$$|x| = \begin{cases} 0 & \text{if } x = 0, \\ 1 & \text{if } x \neq 0. \end{cases}$$

This is the only possible absolute value on a finite field.

We will say that two absolute values $|\ |_1$ and $|\ |_2$ on K are *equivalent* if there exists $\theta \in \mathbb{R}_{>0}$ such that

$$|x|_1 = |x|_2^\theta$$

for all $x \in K$. In this case we will write $|\ |_1 \sim |\ |_2$. It is easy to see that an archimedean absolute value cannot be equivalent to a non-arhimedean one. This relation $\sim$ is an actual equivalence relation (Problem 7.3), and the only absolute value equivalent to the trivial one is itself (Problem 7.4).

Equivalence classes of nontrivial absolute values on K are called *places*. The set of all places of K will be denoted by $M(K)$. Notice that an absolute value $|\ |$ defines a metric on K:

$$(x, y) \to |x - y|$$

for every $x, y \in K$. Therefore $|\ |$ induces a metric topology on K. Moreover, we can talk about the *completion* of K with respect to this topology. K equipped with the metric induced by $|\ |$ is a metric space, we will write $(K, |\ |)$ to mean that we are thinking of K as a metric space with respect to this metric. Recall that a metric space $(K, |\ |)$ is called *complete* if every Cauchy sequence in K converges to a point in K. The *completion* of $(K, |\ |)$ is the set of all equivalence classes of Cauchy sequences on $(K, |\ |)$, where two Cauchy sequences $\{a_n\}$ and $\{b_n\}$ are equivalent if

$$\lim_{n\to\infty} |a_n - b_n| = 0.$$

Notice that $|\ |$ is also defined on the completion of $(K, |\ |)$, and so this completion also has a metric topology induced by $|\ |$. Then $(K, |\ |)$ is complete if and only if it is equal to its completion; by "equal" here we mean isometrically isomorphic as fields: it is a well known fact that completion of a field is also a field, where addition and multiplication on Cauchy sequences are defined component-wise.

Notice that for an absolute value $|\ |$ on K, $x \to |x|$ is a homomorphism from the multiplicative group $K^\times = \{x \in K : x \neq 0\}$ to multiplicative group $\mathbb{R}_{>0}$. Therefore:

(1) $|1| = 1$,
(2) $|\zeta| = 1$ for every root of unity $\zeta \in K$, i.e. for every $\zeta \in K$ such that $\zeta^n = 1$ for some $n \in \mathbb{Z}_{>0}$,
(3) $|-x| = |x|$, for all $x \in K^\times$,
(4) $|x^{-1}| = |x|^{-1}$, for all $x \in K^\times$.

If L/K is an extension of fields and $|\ |$ is an absolute value on L, then its restriction to K is an absolute value on K. It is in general possible that $|\ |$ is non-trivial on L, but is trivial on K.

We will now demonstrate some standard absolute values on $\mathbb{Q}$. The first one is the usual absolute value, which we will denote by $|\ |_\infty$:

$$|x|_\infty = \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x < 0. \end{cases}$$

This is an archimedean absolute value (Problem 7.6), which induces the usual metric topology on $\mathbb{Q}$; the completion of $\mathbb{Q}$ with respect to this topology is $\mathbb{R}$. Sometimes we will write $\mathbb{Q}_\infty$ instead of $\mathbb{R}$ to stress this fact.

Now let $p \in \mathbb{Z}$ be a prime, and define the *p-adic* absolute value $|\ |_p$ on $\mathbb{Q}$ as follows. For each $n \in \mathbb{Z}$, let

$$|n|_p = p^{-\mu(n)},$$

where $p^{\mu(n)}$ is the largest power of p dividing n, hence $|n|_p \leq 1$ for each $n \in \mathbb{Z}$. Now for each $\frac{m}{n} \in \mathbb{Q}$, let

$$\left|\frac{m}{n}\right|_p = \frac{|m|_p}{|n|_p}.$$

This is a non-archimedean absolute value on $\mathbb{Q}$ for every prime p (Problem 7.7). The topology induced by $|\ |_p$ on $\mathbb{Q}$ is called *p-adic topology*; the completion of $\mathbb{Q}$ with respect to this is called the field of *p-adic numbers*, denoted by $\mathbb{Q}_p$. The set

$$\mathbb{Z}_p = \{a \in \mathbb{Q}_p : |a|_p \leq 1\}$$

is a ring, and is called the ring of *p-adic integers*. Problem 7.8 implies that $\mathbb{Z} \subseteq \mathbb{Z}_p$ for every prime $p \in \mathbb{Z}$. Moreover, if we write $\mathcal{P}$ for the set of all primes in $\mathbb{Z}$, then

$$\mathbb{Z} = \bigcap_{p \in \mathcal{P}} \mathbb{Z}_p.$$

The important result classifying all absolute values on $\mathbb{Q}$ is Ostrowski's theorem.

THEOREM 7.4.1 (Ostrowski, 1935). *Any non-trivial absolute value on $\mathbb{Q}$ is equivalent to either $|\ |_\infty$ or $|\ |_p$ for some $p \in \mathcal{P}$.*

PROOF. We start with the following fact, the proof of which is deferred to Problem 7.9.

LEMMA 7.4.2. *An absolute value $|\ |$ on $\mathbb{Q}$ is non-archimedean if and only if $|n| \leq 1$ for every $n \in \mathbb{Z}$. Moreover, for any absolute value $|\ |$ on $\mathbb{Q}$ there exists $\rho \in \mathbb{R}_{>0}$ such that*

$$|n| \leq |n|_\infty^\rho. \tag{7.29}$$

Now suppose $|\ |$ is an absolute value on $\mathbb{Q}$. We will use Lemma 7.4.2 throughout this proof, assuming without loss of generality that $\rho = 1$ in (7.29); indeed, $|\ |^{\frac{1}{\rho}}$ is equivalent to $|\ |$, so it is not important whether we prove that $|\ |^{\frac{1}{\rho}}$ or $|\ |$ is equivalent to $|\ |_\infty$ or $|\ |_p$ for some $p \in \mathcal{P}$.

Let $a, b \in \mathbb{Z}_{>0}$, $a > 1, b > 1$. For any $\nu \in \mathbb{Z}_{>0}$, there exists integers $c_0, \ldots, c_n$ with $0 \leq c_i < a$ and $c_n \neq 0$ such that

$$b^\nu = c_0 + c_1 a + \cdots + c_n a^n.$$

Notice that by Lemma 7.4.2 for each $0 \leq i \leq n$,

$$|c_i| \leq |c_i|_\infty \leq |a|_\infty = a.$$

Also notice that

$$a^n \leq c_n a^n \leq b^\nu,$$

and so $n \leq \frac{\nu \log b}{\log a}$. Then

$$\begin{aligned} |b|^\nu = |b^\nu| &\leq \sum_{i=0}^{n} |c_i||a|^i \leq (n+1)\ a \max\{1, |a|\}^n \\ &\leq \left(1 + \frac{\nu \log b}{\log a}\right) a \max\{1, |a|\}^n. \end{aligned}$$

Therefore

$$|b| \leq \left(1 + \frac{\nu \log b}{\log a}\right)^{1/\nu} a^{1/\nu} \max\{1, |a|\}^{\frac{\log b}{\log a}} \to \max\left\{1, |a|^{\frac{\log b}{\log a}}\right\},$$

as $\nu \to \infty$, in other words

$$|b| \leq \max\left\{1, |a|^{\frac{\log b}{\log a}}\right\}. \tag{7.30}$$

Case 1. Assume $|\ |$ is archimedean. Then by Lemma 7.4.2, there exists $b \in \mathbb{Z}$ such that $|b| > 1$. Then by (7.30), $|a| > 1$ for every $a \in \mathbb{Z}$ except for -1,0,1. Therefore if $a, b \in \mathbb{Z}$, $a, b > 1$, then

$$|b|^{\frac{1}{\log b}} \leq |a|^{\frac{1}{\log a}} \leq |b|^{\frac{1}{\log b}},$$

and so

$$|b|^{\frac{1}{\log b}} = |a|^{\frac{1}{\log a}}.$$

We have

$$1 < |b| \leq |b|_\infty = b,$$

so $|b| = |b|_\infty^\rho = b^\rho$ for some $0 < \rho \leq 1$, and hence

$$|a| = |b|^{\frac{\log a}{\log b}} = b^{\rho \frac{\log a}{\log b}} = a^\rho = |a|_\infty^\rho.$$

Same way therefore $|\alpha| = |\alpha|_\infty^\rho$ for every $\alpha \in \mathbb{Q}$.

Case 2. Assume $|\ |$ is non-archimedean. Then by Lemma 7.4.2, $|n| \leq 1$ for every $n \in \mathbb{Z}$, and since $|\ |$ is non-trivial, there exists $a \in \mathbb{Z}$ such that $|a| < 1$. Let

$$I = \{a \in \mathbb{Z} : |a| < 1\}.$$

This is an ideal in $\mathbb{Z}$ (Problem 7.10). Therefore there exists a prime $p \in \mathbb{Z}$ such that $I = p\mathbb{Z}$. Let $0 \neq \alpha \in \mathbb{Q}$. Write

$$\alpha = p^r \frac{x}{y}$$

with $x, y \in \mathbb{Z}$ such that $p \nmid xy$. Then $x, y \notin I$, hence

$$|x| = |y| = 1,$$

and so

$$|\alpha| = |p^r| = |p|^r.$$

Since $p \in I$, $|p| < 1$, so $|p| = p^{-s}$ for some $s > 0$. Then

$$|\alpha| = p^{-rs} = |r|_p^s.$$

We have shown that $|\ |$ must be equivalent to either $|\ |_\infty$ or $|\ |_p$ for some prime p. This completes the proof. □

Therefore we can write

$$M(\mathbb{Q}) = \{\infty\} \cup \mathcal{P},$$

this way indexing the archimedean place by ∞, and non-archimedean places by p for each $p \in \mathcal{P}$.

THEOREM 7.4.3 (Artin - Whaples Product Formula). *If $0 \neq a \in \mathbb{Q}$, then*

$$|a|_\infty \prod_{p \in \mathcal{P}} |a|_p = 1.$$

PROOF. Problem 7.11. □

Next we discuss absolute values on a number field K. If $|\ |$ is an absolute value on K, its restriction to $\mathbb{Q}$ is an absolute value on $\mathbb{Q}$, and so must belong to either ∞ or one of the p-adic places on $\mathbb{Q}$. Hence absolute values on K are precisely extensions of those on $\mathbb{Q}$. If $v \in M(K)$, we will write $|\ |_v$ for an absolute value that represents it. We know that $|\ |_v$ extends either $|\ |_\infty$ or $|\ |_p$ for some $p \in \mathcal{P}$, and we say that v *lies over* ∞ or p respectively; we denote it by writing $v|\infty$ or $v|p$. The place $v \in M(K)$ is archimedean if and only if $v|\infty$. Sometimes we will write $v \nmid \infty$ to mean that v is non-archimedean, i.e. lies over some p-adic place of $\mathbb{Q}$. For each place $u \in M(\mathbb{Q})$ there may be more than one place $v \in M(K)$ such that $v|u$, however each places $v \in M(K)$ lies over precisely one place $u \in M(\mathbb{Q})$.

First we describe all archimedean places of K. Let $\sigma_1, \ldots, \sigma_r$ be real embeddings of K, and $\tau_1, \overline{\tau}_1, \ldots, \tau_s, \overline{\tau}_s$ conjugate pairs of complex embeddings, then

$$r + 2s = d = [K : \mathbb{Q}].$$

Notice that since $\mathbb{Q}_\infty = \mathbb{R} \subset \mathbb{C}$, the absolute value $|\ |_\infty$ is defined on $\mathbb{R}$ and on $\mathbb{C}$. Also, for each $a \in K$

$$\sigma_i(a) \in \mathbb{R},\ \tau_j(a), \overline{\tau}_j(a) \in \mathbb{C}$$

for each $1 \leq i \leq r$ and $1 \leq j \leq s$. If ρ is one of these embeddings, then we define an absolute value $|\ |_\rho$ on K by

$$|a|_\rho = |\rho(a)|_\infty.$$

It is easy to notice that if $|\ |_{\tau_j} = |\ |_{\overline{\tau}_j}$ for each $1 \leq j \leq s$. However, the absolute values

$$|\ |_{\sigma_1}, \ldots, |\ |_{\sigma_r}, |\ |_{\tau_1}, \ldots, |\ |_{\tau_s}$$

are not equivalent to each other. These represent all the archimedean places of K. For each $v \in M(K)$, we will write K_v for the completion of K at v. If $v|u$ for some $u \in M(\mathbb{Q})$, then $K_v/\mathbb{Q}_u$ is an extension of fields, and we will define the *local degree* of K at v to be the degree of this extension, and denote it by

$$d_v = [K_v : \mathbb{Q}_u].$$

We will also write sometimes $\mathbb{Q}_v$ where $v \in M(K)$ to mean $\mathbb{Q}_u$, where $u \in M(\mathbb{Q})$ is the unique place over which v lies. Notice that if $v \in M(K)$ is archimedean, then K_v is either $\mathbb{R}$ or $\mathbb{C}$, depending on whether v is real or complex, i.e. corresponds to a real or to a complex embedding. Therefore, for each $v|\infty$

$$d_v = [K_v : \mathbb{Q}_\infty] = [K_v : \mathbb{R}] = \begin{cases} 1 & \text{if } v \text{ is real} \\ 2 & \text{if } v \text{ is complex.} \end{cases}$$

Therefore

$$\sum_{v|\infty} d_v = r + 2s = d.$$

Next we describe non-archimedean places of K. Let p be a prime in $\mathbb{Z}$, so that $(p) = p\mathbb{Z}$ is a prime ideal in $\mathbb{Z}$. Recall that $\mathcal{O}_K$, the ring of algebraic integers of K, is a Dedekind domain, which means that there is unique factorization into prime ideals in $\mathcal{O}_K$. Notice that $\mathbb{Z} \in \mathcal{O}_K$, and so $p\mathcal{O}_K$ is an ideal in $\mathcal{O}_K$, although it may no longer be prime. Then there exist prime ideals $P_1, \ldots, P_k$ and positive integers $e_1, \ldots, e_k$ such that

$$p\mathcal{O}_K = P_1^{e_1} \ldots P_k^{e_k},$$

and $\sum_{i=1}^k e_i = d$; each such e_i is called the *ramification degree* of P_i over p. First we define $|0|_{P_i} = 0$. Now let $0 \neq a \in \mathcal{O}_K$, then for each P_i, $1 \leq i \leq k$, define

$$\operatorname{ord}_{P_i} a = \max\{j \in \mathbb{Z} : a \in P_i^j\},$$

and let

$$|a|_{P_i} = p^{-\frac{\operatorname{ord}_{P_i} a}{e_i}}.$$

The number $\operatorname{ord}_{P_i} a$ is well-defined due to unique factorization of ideals into powers of prime ideals: it is precisely the power to which P_i divides $a\mathcal{O}_K$. Notice that K is the field of fractions of $\mathcal{O}_K$, i.e.

$$K = \left\{ \frac{a}{b} : a, b \in \mathcal{O}_K \right\}.$$

Then for each $\alpha = \frac{a}{b} \in K$ with $a, b \in \mathcal{O}_K$, define

$$|\alpha|_{P_i} = \frac{|a|_{P_i}}{|b|_{P_i}}. \tag{7.31}$$

This is an absolute value on K, which restricts to the usual p-adic absolute value on $\mathbb{Q}$ (Problem 7.12). Hence for each prime p in $\mathbb{Z}$, we defined absolute values lying over it; these are all the non-archimedean places of K. Suppose $v \in M(K)$ lies over p, and P_i is the corresponding prime ideal of $\mathcal{O}_K$ with ramification degree e_i over p. In a Dedekind domain every nonzero prime ideal is maximal, hence P_i is a

maximal ideal, and so $\mathcal{O}_K/P_i$ is a field; in fact, it is a finite field of characteristic p, meaning that

$$|\mathcal{O}_K/P_i| = p^{f_i},$$

for some $f_i \in \mathbb{Z}_{>0}$. This f_i is called the *inertia degree* of P_i over p. Its significance for our purposes is that the local degree $d_v = [K_v : \mathbb{Q}_p]$ is equal to $e_i f_i$. A result from algebraic number theory implies that if $P_1, \ldots, P_k$ are prime ideals in $\mathcal{O}_K$ lying over a rational prime p with respective ramification degrees $e_1, \ldots, e_k$ and ramification degrees $f_1, \ldots, f_k$, then

$$\sum_{i=1}^{k} e_i f_i = d.$$

In particular this means that

$$\sum_{v|u} d_v = d$$

is true for any $u \in M(\mathbb{Q})$. The Artin - Whaples product formula works over a number field in a similar way as over $\mathbb{Q}$: we state here without proof.

THEOREM 7.4.4. *If $0 \neq a \in K$, then*

$$\prod_{v \in M(K)} |a|_v^{d_v} = 1.$$

EXAMPLE 7.4.1. *Let $K = \mathbb{Q}(\sqrt{2})$, then $d = 2$. Since K is totally real, there are no complex embeddings. Hence if $v \in M(K)$ is archimedean, then $K_v = \mathbb{R}$, and so $d_v = 1$. Since*

$$\sum_{v|\infty} d_v = 2,$$

K must have two archimedean places. These are precisely the places corresponding to embeddings $\sigma_1, \sigma_2 : K \to \mathbb{R}$, given by

$$\sigma_1(\sqrt{2}) = \sqrt{2},\ \sigma_2(\sqrt{2}) = -\sqrt{2},$$

and fixing $\mathbb{Q}$, hence σ_1 is the identity. Let v_1, v_2 be the archimedean places corresponding to embeddings σ_1, σ_2 respectively. Notice that for every $\alpha \in K$, there exist $a, b \in \mathbb{Q}$ such that $\alpha = a + b\sqrt{2}$, hence

$$|\alpha|_{v_1} = |\sigma_1(a + b\sqrt{2})|_\infty = |a + b\sqrt{2}|_\infty,$$

and

$$|\alpha|_{v_2} = |\sigma_2(a + b\sqrt{2})|_\infty = |a - b\sqrt{2}|_\infty.$$

Now let us look at non-archimedean places of K. Consider for instance all places $v \in M(K)$ lying over 7. Notice that

$$7 = (3 + \sqrt{2})(3 - \sqrt{2}),$$

therefore the ideal $7\mathcal{O}_K$ no longer prime in $\mathcal{O}_K$ splits as the product of these two prime ideals:

$$7\mathcal{O}_K = P_1 P_2,$$

where $P_1 = (3+\sqrt{2})\mathcal{O}_K$ *and* $P_2 = (3-\sqrt{2})\mathcal{O}_K$*. This means that there are two places lying over 7, corresponding to* P_1 *and* P_2*, call them* u_1 *and* u_2 *respectively. Then* $d_{u_1} = d_{u_2} = 1$*. Notice for instance that*

$$3+\sqrt{2} \in P_1,\ 3+\sqrt{2} \notin P_1^2,\ 3+\sqrt{2} \notin P_2,$$
$$3-\sqrt{2} \in P_2,\ 3-\sqrt{2} \notin P_2^2,\ 3-\sqrt{2} \notin P_1,$$

hence

$$|3+\sqrt{2}|_{u_1} = 7^{-1},\ |3-\sqrt{2}|_{u_1} = 7^0,$$
$$|3+\sqrt{2}|_{u_2} = 7^0,\ |3-\sqrt{2}|_{u_2} = 7^{-1}.$$

Recall that prime ideals in $\mathcal{O}_K$ *are maximal. This implies that* $3 \pm \sqrt{2}$ *are not contained in any other prime ideal of* $\mathcal{O}_K$*, hence for every place* $v \in M(K)$ *which is not equal to* v_1*,* v_2*,* u_1*, or* u_2*,* $|3 \pm \sqrt{2}|_v = 1$*. Hence*

$$\prod_{v \in M(K)} |3 \pm \sqrt{2}|_v = |3+\sqrt{2}|_\infty |3-\sqrt{2}|_\infty 7^{-1} = 1.$$

This is a demonstration of the product formula at work.

Remark 7.4.1. The same construction of absolute values as described in this section can be carried out for any field extension of number fields L/K. In this case, we would replace the ground field $\mathbb{Q}$ with K, and talk about places of L lying over places of K in the same precise manner. We will assume this more general construction going forward.

We now introduce *height functions*, which serve as the main tool used to measure arithmetic complexity. We have already seen an example of a height function $\mathcal{H}$ in Section 6.4, however $\mathcal{H}$ only carries archimedean information: it only measured the size of a given algebraic number at the archimedean places. We are now prepared to define more general heights on vectors, which incorporate arithmetic information at all the places of a number field. As above, K is a number field of degree d over $\mathbb{Q}$ and $M(K)$ is its set of places. Let $n \geq 2$ be an integer. For each place v of K we define a *local height* H_v for each vector $\boldsymbol{x} \in K_v^n$ by

$$H_v(\boldsymbol{x}) = \begin{cases} \left(\sum_{i=1}^n |x_i|_v^2\right)^{\frac{1}{2}} & \text{if } v|\infty, \\ \max_{1 \leq i \leq n} |x_i|_v & \text{if } v \nmid \infty. \end{cases}$$

Then for each $\mathbf{0} \neq \boldsymbol{x} \in K^n$, define the *global height* H_K by

$$H_K(\boldsymbol{x}) = \prod_{v \in M(K)} H_v(\boldsymbol{x})^{d_v}. \tag{7.32}$$

Notice that for each $\mathbf{0} \neq \boldsymbol{x} \in K^n$, $H_v(\boldsymbol{x}) = 1$ for all but finitely many places v of K, hence the product in (7.32) is actually finite, therefore convergent, meaning that H_K is well-defined. Also notice that if $0 \neq \alpha \in K$ and $\mathbf{0} \neq \boldsymbol{x} \in K^n$, then

$$\begin{aligned} H_K(\alpha \boldsymbol{x}) &= \prod_{v \in M(K)} |\alpha|_v^{d_v} H_v(\boldsymbol{x})^{d_v} \\ &= \left(\prod_{v \in M(K)} |\alpha|_v^{d_v}\right) \prod_{v \in M(K)} H_v(\boldsymbol{x})^{d_v} = H_K(\boldsymbol{x}) \end{aligned} \tag{7.33}$$

by the product formula. This means that H_K is a *homogeneous* function, and so is *projectively defined*. Indeed, define an equivalence relation on $K^n \setminus \{\mathbf{0}\}$ by writing

$\boldsymbol{x} \sim \boldsymbol{y}$ whenever $\boldsymbol{x} = \alpha\boldsymbol{y}$ for some $0 \neq \alpha \in K$. It is easy to check that this indeed is an equivalence relation, and we write $[x_1 : \cdots : x_n]$ for the equivalence class of the vector $\boldsymbol{x} = (x_1, \ldots, x_n) \in K^n$, which is called the *projective point* corresponding to $\boldsymbol{x}$. The space of all projective points on K^n is called the $(n-1)$-dimensional *projective space* over K, i.e.

$$\mathbb{P}^{n-1}(K) = \{[x_1 : \cdots : x_n] : (x_1, \ldots, x_n) \in K^n \setminus \{\mathbf{0}\}\}.$$

Notice that this is precisely the space of all lines through the origin in K^n, i.e. the space of 1-dimensional subspaces of K^n. This is the simplest example of the more general construction of Grassmannian that we will encounter later. Then (7.33) implies that H_K is well-defined on $\mathbb{P}^{n-1}(K)$, i.e. it can be viewed as a function $H_K : \mathbb{P}^{n-1}(K) \to \mathbb{R}_{>0}$.

Notice that the definition of H_K depends on K. Let L be an extension of K of degree e, hence $[L : \mathbb{Q}] = de$. For each place $v \in M(L)$, we will write $e_v = [L_v : K_v]$, hence $[L_v : \mathbb{Q}_v] = d_v e_v$. Also notice that

$$\sum_{v \in M(L), v|u} e_v = e$$

for each place $u \in M(K)$. Suppose that $\mathbf{0} \neq \boldsymbol{x} \in K^n$, then

$$H_L(\boldsymbol{x}) = \prod_{v \in M(L)} H_v(\boldsymbol{x})^{d_v e_v} = \prod_{u \in M(K)} \prod_{v \in M(L), v|u} H_v(\boldsymbol{x})^{d_u e_v},$$

but since $\boldsymbol{x} \in K^n$, $H_v(\boldsymbol{x}) = H_{v'}(\boldsymbol{x})$ whenever $v, v' \in M(L)$ lie over the same place $u \in M(K)$. Hence:

$$H_L(\boldsymbol{x}) = \prod_{u \in M(K)} H_u(\boldsymbol{x})^{d_u \sum_{v \in M(L), v|u} e_v} = \prod_{u \in M(K)} H_u(\boldsymbol{x})^{d_u e} = H_K(\boldsymbol{x})^e.$$

This suggests that if we want a height function that does not depend on the field of definition, we may want to introduce the normalizing exponent $\frac{1}{[K:\mathbb{Q}]}$.

Definition 7.4.2. Let $\mathbb{A}$ be the field of all algebraic numbers, as before. Define the *absolute height* $H : \mathbb{A}^n \setminus \{\mathbf{0}\} \to \mathbb{R}_{>0}$ by

$$H(\boldsymbol{x}) = H_K(\boldsymbol{x})^{\frac{1}{[K:\mathbb{Q}]}}$$

for every $\mathbf{0} \neq \boldsymbol{x} \in \mathbb{A}^n$, where K is any number field containing the coordinates of $\boldsymbol{x}$. By the discussion above, H does not depend on the choice of this number field. Once again, notice that H is projectively defined. We will also adopt a convention that $H(\mathbf{0}) = 1$.

We also define the *inhomogeneous height* $h_K : K^n \to \mathbb{R}_{>0}$ by

$$h_K(\boldsymbol{x}) = H_K(1, \boldsymbol{x}),$$

for every $\boldsymbol{x} \in K^n$, and the *absolute inhomogeneous height* $h : \mathbb{A}^n \to \mathbb{R}_{>0}$ by

$$h(\boldsymbol{x}) = h_K(\boldsymbol{x})^{\frac{1}{[K:\mathbb{Q}]}},$$

for every $\boldsymbol{x} \in \mathbb{A}^n$, where K is any number field containing the coordinates of $\boldsymbol{x}$. Notice that h_K and h are no longer projectively defined, i.e. if $\alpha \in \mathbb{A}$, then $h(\alpha\boldsymbol{x})$ is not necessarily equal to $h(\boldsymbol{x})$. Also notice that for every $\boldsymbol{x} \in \mathbb{A}^n$,

$$H(\boldsymbol{x}) \leq h(\boldsymbol{x}).$$

For any algebraic number $\alpha \in \mathbb{A}$, we define its *Weil height* to be

$$h(\alpha) = H(1, \alpha).$$

We now briefly outline the basic properties of heights, proofs of which are left to the exercises.

LEMMA 7.4.5. *The following statements are true:*

(1) *If* $\boldsymbol{x} \in \mathbb{Z}$ *is such that* $\gcd(x_1, \ldots, x_n) = 1$, *then*

$$H(\boldsymbol{x}) = \|\boldsymbol{x}\|_2 = \left(x_1^2 + \cdots + x_n^2\right)^{\frac{1}{2}},$$

i.e. height of an integer vector is the Euclidean norm of the corresponding primitive vector.

(2) *If* $0 \neq x_0 \in \mathbb{Z}$, *and*

$$\boldsymbol{x} = \left(\frac{x_1}{x_0}, \ldots, \frac{x_n}{x_0}\right) \in \mathbb{Q}^n,$$

is such that $\gcd(x_0, x_1, \ldots, x_n) = 1$, *then*

$$h(\boldsymbol{x}) = \left(x_0^2 + x_1^2 + \cdots + x_n^2\right)^{\frac{1}{2}},$$

i.e. the inhomogeneous height of a rational vector is the Euclidean norm of the corresponding reduced integer vector $(x_0, x_1, \ldots, x_n)$.

PROOF. Problem 7.13. □

LEMMA 7.4.6. *If* $m_1, \ldots, m_k \in \mathbb{Z}$, *and* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathbb{A}^n$, *then*

$$h\left(\sum_{i=1}^{k} m_i \boldsymbol{x}_i\right) \leq \left(\sum_{i=1}^{k} m_i^2\right)^{\frac{1}{2}} \prod_{i=1}^{k} h(\boldsymbol{x}_i).$$

In particular, if $\alpha_1, \ldots, \alpha_k \in \mathbb{A}$, *then*

$$h\left(\sum_{i=1}^{k} m_i \alpha_i\right) \leq \left(\sum_{i=1}^{k} m_i^2\right)^{\frac{1}{2}} \prod_{i=1}^{k} h(\alpha_i).$$

Additionally, for any $\alpha, \beta \in \mathbb{A}$,

$$h(\alpha\beta) \leq h(\alpha)h(\beta).$$

PROOF. Problem 7.14. □

LEMMA 7.4.7. *Suppose that* K *and* L *are isomorphic number fields with* $\sigma : K \to L$ *an isomorphism, and let us also write* σ *for the isomorphism it induces from* K^n *to* L^n *for each integer* $n \geq 1$. *Then*

$$H(\sigma(\boldsymbol{x})) = H(\boldsymbol{x})$$

for each $\boldsymbol{x} \in K$. *Hence conjugate vectors have the same height. Notice in particular that this implies that conjugate algebraic numbers have the same height.*

PROOF. Problem 7.15. □

The notion of height also extends to polynomials. In particular, if F is a polynomial with coefficients $a_1, \ldots, a_n \in \mathbb{A}$, then we define

$$H(F) = H(a_1, \ldots, a_n).$$

LEMMA 7.4.8. *Let $P(X), Q(X) \in \mathbb{A}[X]$ be polynomials in one variable with coefficients in $\mathbb{A}$ of degrees n_1, n_2 respectively, and let $n = \min\{n_1, n_2\}$. Then*

$$H(PQ) \leq \sqrt{n+1}\ H(P)H(Q).$$

PROOF. Let K be a number field containing coefficients of P and Q, and suppose it has degree d over $\mathbb{Q}$. It is easy to observe that for every $v \in M(K)$ such that $v \nmid \infty$,

$$H_v(PQ) = H_v(P)H_v(Q),$$

where these are precisely the local heights of corresponding coefficient vectors. Let $v \in M(K)$, $v|\infty$, then by Problem 7.16

$$H_v(PQ) \leq \sqrt{n+1}\ H_v(P)H_v(Q).$$

Therefore we have:

$$\begin{aligned} H(PQ) &= \prod_{v\in M(K)} H_v(PQ)^{\frac{d_v}{d}} \\ &\leq \prod_{v\nmid\infty} \left(H_v(P)H_v(Q)\right)^{\frac{d_v}{d}} \prod_{v|\infty} \Big(|n+1|_v^{\frac{1}{2}}\ H_v(P)H_v(Q)\Big)^{\frac{d_v}{d}} \\ &= H(P)H(Q) \prod_{v|\infty} |n+1|_v^{\frac{d_v}{2d}} \\ &= \big(\sqrt{n+1}\big)^{\frac{\sum_{v|\infty} d_v}{d}}\ H(P)H(Q) = \sqrt{n+1}\ H(P)H(Q). \end{aligned}$$

This completes the proof. □

COROLLARY 7.4.9. *Suppose that*

$$P(X) = a_d(X-\alpha_1)\dots(X-\alpha_d),$$

where $a_d, \alpha_1, \dots, \alpha_d \in \mathbb{A}$. Then

$$H(P) \leq 2^{\frac{d-1}{2}} h(\alpha_1)\dots h(\alpha_d). \tag{7.34}$$

PROOF. Notice that here we can view $P(X)$ as a product of d linear polynomials in one variable, hence applying Lemma 7.4.8 $d-1$ times yields (7.34). □

For a vector $\boldsymbol{x} \in \mathbb{A}^n$, we define its *degree* to be

$$\deg(\boldsymbol{x}) = [\mathbb{Q}(x_1, \dots, x_n) : \mathbb{Q}].$$

Also, for a projective point $[\boldsymbol{x}]$ we write $\deg([\boldsymbol{x}])$ to mean the minimum of $\deg(\boldsymbol{x})$ taken over all representatives of $[\boldsymbol{x}]$. We are now ready to prove the fundamental property of heights, which was first established by Northcott in 1949 [**Nor49**]: this result is known as Northcott's theorem, and any height function satisfying this theorem (there are others, not only our H) is said to satisfy *Northcott's finiteness property.*

THEOREM 7.4.10. *Let n, d, B be positive integers. Then the set*

$$S_n(B,d) = \big\{[\boldsymbol{x}] \in \mathbb{P}^{n-1}(\mathbb{A}) : \deg([\boldsymbol{x}]) \leq d,\ H(\boldsymbol{x}) \leq B\big\}$$

is finite.

PROOF. If $\boldsymbol{x} = (x_1, x_2, \dots, x_n) \in \mathbb{A}^n$ with $x_i \neq 0$ for some $1 \leq i \leq n$, then $H(\boldsymbol{x}) = H\left(\frac{\boldsymbol{x}}{x_i}\right)$. Therefore we can always choose a representative $\boldsymbol{x}$ of $[\boldsymbol{x}] \in \mathbb{P}^{n-1}(\mathbb{A})$ with one coordinate equal to 1. Without loss of generality assume $\boldsymbol{x} = (1, x_2, \dots, x_n) \in \mathbb{A}^n$, then

$$H(\boldsymbol{x}) \geq H(1, x_i) = h(x_i), \ \forall \ 2 \leq i \leq n.$$

Therefore it suffices to prove that the set

$$S(B, d) = \{\alpha \in \mathbb{A} : \deg(\alpha) \leq d, \ h(\alpha) \leq B\}$$

is finite. Notice that if $\alpha \in S(B, d)$, then it must be a root of a monic polynomial with rational coefficients of degree at most d

$$P(X) = (X - \alpha_1)(X - \alpha_2) \dots (X - \alpha_d),$$

where $\alpha = \alpha_1, \alpha_2, \dots, \alpha_d$ are conjugates of α. By Lemma 7.4.7, $h(\alpha) = h(\alpha_i)$ for every $1 \leq i \leq d$, and so $h(\alpha_i) \leq B$ for all $1 \leq i \leq d$. By Corollary 7.4.9,

$$H(P) \leq 2^{\frac{d-1}{2}} h(\alpha_1) \dots h(\alpha_d) \leq 2^{\frac{d-1}{2}} B^d. \tag{7.35}$$

Since $P(x)$ is monic, let $\left(\frac{m_0}{m}, \dots, \frac{m_{d-1}}{m}, 1\right) \in \mathbb{Q}$ be the coefficient vector of P, written is such a way that $\gcd(m, m_0, \dots, m_{d-1}) = 1$. Then by Lemma 7.4.5,

$$H(P) = \sqrt{m^2 + m_0^2 + \dots + m_{d-1}^2} = \|\boldsymbol{m}\|_2,$$

where $\boldsymbol{m} = (m, m_0, \dots, m_{d-1}) \in \mathbb{Z}^{d+2}$, and $\| \ \|_2$ stands for the Euclidean norm, as usual. It is now easy to see that there are only finitely many integral vectors $\boldsymbol{m}$ with $\|\boldsymbol{m}\|_2 \leq 2^{\frac{d-1}{2}} B^d$, and so there are only finitely many polynomials P satisfying (7.35). This means that $S(B, d)$ must be finite, and so completes the proof. □

REMARK 7.4.2. The cardinality of $S_n(B, d)$ has been investigated by various authors, starting with a result of Schanuel in 1979. More recently there were upper and lower bounds produced by Schmidt, Gao, Thunder, Masser, Vaaler, and Widmer among others, however there still is no known general asymptotic formula for $|S_n(B, d)|$ (see [**Wid09**] and [**Wid10**] for some recent results and a more detailed bibliogrpahy).

Next we will show how the notion of height can be extended to subspaces of K^n. Let $V \subseteq K^n$ be an ℓ-dimensional subspace, $1 \leq \ell \leq n$. Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_\ell$ be a basis for V, and write $X = (\boldsymbol{x}_1 \ \dots \ \boldsymbol{x}_\ell)$ for the corresponding $n \times \ell$ basis matrix. Let $\mathcal{I}$ be the set of subsets of $\{1, \dots, n\}$ of cardinality ℓ, then

$$|\mathcal{I}| = \binom{n}{\ell}.$$

For each $I \in \mathcal{I}$, let X_I be the $\ell \times \ell$ submatrix of X whose rows are indexed by elements of I. We introduce lexicographic ordering on elements of $\mathcal{I}$, and write

$$\mathcal{I} = \left\{I_1, \dots, I_{\binom{n}{\ell}}\right\}$$

with respect to that order. Then define a vector of *Grassmann coordinates* (also known as *Plücker coordinates*) of V with respect to the basis $\boldsymbol{x}_1, \dots, \boldsymbol{x}_\ell$ to be

$$g(X) = \left(\det(X_{I_1}), \dots, \det\left(X_{I_{\binom{n}{\ell}}}\right)\right) \in K^{\binom{n}{\ell}}.$$

Suppose $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_\ell$ is a different basis for V, and write Y for the corresponding basis matrix. Then there exists a matrix $U \in GL_\ell(K)$ such that

$$Y = XU,$$

and so it is easy to see that

$$g(Y) = \det(U) g(X).$$

As before, we write $[g(X)]$ for the projective point in $\mathbb{P}^{\binom{n}{\ell}-1}$ represented by the vector $g(X)$, hence $[g(X)] = [g(Y)]$, and so we denote this projective point $[g(V)]$ to indicate that it does not depend on the choice of the basis. Define

$$\mathbb{G}_n^\ell(K) = \{[g(V)] : V \subseteq K^n,\ \dim_K(V) = \ell\}.$$

$\mathbb{G}_n^\ell(K)$ is called the $\binom{n}{\ell}$-*Grassmann component* of K^n, and this is the projective space whose points correspond to ℓ-dimensional subspaces of K^n. Notice that this is a generalization of the projective space $\mathbb{P}^{n-1}(K)$, which can be thought of as the space of one-dimensional subspaces of K^n. This is perhaps the simplest example of a parameter space, i.e. of a general type of objects in algebraic geometry which are called *moduli spaces*.

Using this notation, we can now define height of an ℓ-dimensional subspace V of K^n by

$$H(V) = H(g(V)). \tag{7.36}$$

Of course, this works in precisely the same manner for subspaces of $\mathbb{A}^n$. This height function on subspaces of a vector space was originally introduced by W. M. Schmidt in [**Sch67**] and is called the *Schmidt height*. Height can also be defined for more general objects, such as algebraic varieties and intersection cycles; this is done in a manner similar in spirit to the simplest case of linear varieties (namely vector subspaces) that we considered here, namely by parametrizing these objects in an appropriate manner. This, however, is more in the realm of arithmetic geometry, and out of the scope of our exposition.

Let us use Weil height on algebraic numbers to briefly revisit Roth's theorem from Diophantine approximation. Recall that for a positive real number δ, a δ-approximation to an algebraic number α is a rational number $\frac{p}{q}$ with $q > 0$, $\gcd(p, q) = 1$, and

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

Then Roth's theorem (Theorem 4.5.1) can be restated by saying that any algebraic number α has only finitely many δ-approximations for each $\delta > 0$. A natural question to ask is how many δ-approximations are there for a fixed algebraic α and a fixed $\delta > 0$? Recall that in Section 4.5 we were counting the number of δ- and related approximations to α in fixed intervals and windows of exponential width. We can now state a result on the overall number of δ-approximations. There are bounds produced by Davenport, Roth, Luckhardt, Mueller, and Schmidt, among others. A version of the following theorem is due to Bombieri and Van der Poorten.

THEOREM 7.4.11. *Let α be an algebraic number of degree $d \geq 3$, and let $0 < \delta < 1$. Then the number of δ-approximations to α is less than*

$$\frac{\log_+ \log h(\alpha)}{\log(1+\delta)} + c(d, \delta), \tag{7.37}$$

where $\log_+ a = \max\{0, \log a\}$, *and*

$$c(d,\delta) = \frac{10^8}{\delta^5}(\log 2d)^2 \log\left(\left(\frac{50}{\delta}\right)^2 \log 2d\right).$$

The first term in (7.37) is best possible, however $c(d,\delta)$ can likely be improved; see [**Sch91**] for many further details on this subject.

7.5. Mahler measure and Lehmer's problem

Notice that the bound on the number of δ-approximations to an algebraic number α in Theorem 7.4.11 depends on the height of α; in particular, the smaller $h(\alpha)$ is, the fewer such approximations are there. This underlines the meaning of $h(\alpha)$ in the following sense. Height measures arithmetic complexity of an algebraic number, incorporating together its size (i.e. archimedean norms) with its divisibility properties (i.e. non-archimedean norms). Hence, roughly speaking, the larger height is, the more "complicated" the algebraic number is, the "closer" it is to transcendental numbers. We know, on the other hand, that transcendental numbers are better approximable than algebraic numbers, so the dependence of the number of δ-approximations on the height makes sense. But how small can $h(\alpha)$ be? In order to investigate this question, it will be convenient to use a different height h_1 on algebraic numbers, which is closely related to h. For each $\alpha \in \mathbb{A}$, define

$$h_1(\alpha) = \prod_{v \in M(K)} \max\{1, |\alpha|_v\}^{\frac{d_v}{d}},$$

where $K = \mathbb{Q}[\alpha]$, $d = [K : \mathbb{Q}]$, and $d_v = [K_v : \mathbb{Q}_v]$. Then (Problem 7.17) for every $\alpha \in \mathbb{A}$,

$$h_1(\alpha) \le h(\alpha) \le \sqrt{2}\, h_1(\alpha).$$

Hence instead of investigating lower bounds for $h(\alpha)$ we can talk about lower bounds for $h_1(\alpha)$. It is easy to see that

$$h_1(\alpha) \ge 1.$$

Further, a classical theorem of Kronecker guarantees that $h_1(\alpha) = 1$ if and only if α is either 0 or a root of unity. So suppose that $\alpha \in \mathbb{A}$ is of fixed degree d, and $h_1(\alpha) > 1$. How small can $h_1(\alpha)$ be? In other words, is there a gap in values of $h_1(\alpha)$, or is it continuous? In this direction we state a famous conjecture of D. H. Lehmer, dating back to 1932 [**Leh33**].

CONJECTURE 7.5.1. *There exists an absolute constant $C \in \mathbb{R}_{>1}$ such that for any algebraic number α of degree d which is not 0 or a root of unity, we have*

$$h_1(\alpha) \ge C^{\frac{1}{d}}.$$

In this section we briefly review some of the material required to understand this outstanding conjecture and its significance. For more detailed information on the material of this section see [**EW99**], which is an excellent account of this fascinating subject.

Let $\alpha_1, \dots, \alpha_d$ be conjugate algebraic numbers of degree d, and let

$$f(x) = a_d \prod_{i=1}^{d} (x - \alpha_i) = \sum_{i=0}^{d} a_i x^i \in \mathbb{Z}[x]$$

be their minimal polynomial. We define the *Mahler measure* of $f(x)$ to be

$$M(f) = |a_d|_\infty \prod_{i=1}^{d} \max\{1, |\alpha_i|_\infty\}. \tag{7.38}$$

If α is an algebraic number of degree d and $f(x) \in \mathbb{Z}[x]$ is its minimal polynomial, then (Problem 7.18)

$$h_1(\alpha) = M(f)^{\frac{1}{d}}.$$

Hence we can restate Lehmer's conjecture in terms of Mahler measure, which incidentally is how it was originally stated.

CONJECTURE 7.5.2. *There exists an absolute constant $C \in \mathbb{R}_{>1}$ such that for any polynomial $f(x) \in \mathbb{Z}[x]$ which is not a product of cyclotomics and a power of x, we have*

$$M(f) \geq C.$$

Moreover,

$$C = M(g) = 1.1762808...,$$

where

$$g(x) = x^{10} + x^9 - x^7 - x^6 - x^5 - x^4 - x^3 + x + 1$$

is the so-called Lehmer's polynomial.

We can think of Mahler measure as another height function defined on polynomials in $\mathbb{Z}[x]$; in particular it satisfies Northcott's finiteness property as in Theorem 7.4.10. More specifically, let $B, d \in \mathbb{R}_{\geq 1}$, and consider the set

$$S(B,d) = \{f(x) \in \mathbb{Z}[x] : \deg(f) \leq d,\ M(f) \leq B\}.$$

Northcott's theorem (Theorem 7.4.10) implies that it is finite. Indeed, if $f(x) \in S(B,d)$, then its roots $\alpha_1, \ldots, \alpha_d$ must be in the set

$$S'(B,d) = \{\alpha \in \mathbb{A} : \deg(\alpha) \leq d,\ h(\alpha) \leq B^{\frac{1}{d}} \leq B\},$$

which is finite by Theorem 7.4.10.

The definition of Mahler measure immediately implies a few nice properties that it satisfies. First of all, $M(f) = 1$ if and only if $f(x)$ is a product of cyclotomic polynomials and x^n for some $n \in \mathbb{Z}_{>0}$. Also notice that unlike the height H on polynomials, which only satisfies Lemma 7.4.8, M is a multiplicative function, i.e.

$$M(fq) = M(f)M(q),$$

for any two polynomials $f(x), q(x)$. On the other hand, Mahler proved that $M(f)$ is comparable to $H(f)$, specifically that there exist constants $c_1(d)$ and $c_2(d)$, depending only on the degree d of the polynomial f, such that

$$c_1(d)H(f) \leq M(f) \leq c_2(d)H(f). \tag{7.39}$$

Notice that the definition of Mahler measure automatically extends to polynomials in $\mathbb{C}[x]$. Although we can no longer think of it as a height function, since the roots of a polynomial in $\mathbb{C}[x]$ are not necessarily algebraic, the multiplicative property remains true. Mahler measure can also be defined as an integral via an application of the classical Jensen's formula from complex analysis:

THEOREM 7.5.1. *For any $\alpha \in \mathbb{C}$,*

$$e^{\int_0^1 \log|\alpha - e^{2\pi i\theta}|\ d\theta} = \max\{1, |\alpha|\},$$

where $|\ |$ stands for the regular $|\ |_\infty$ absolute value on $\mathbb{C}$.

This is a standard theorem, a proof of which can be found for instance in [**EW99**], Lemma 1.9. An immediate application of this is the following result, which was originally proved by Mahler; in fact, this along with (7.39) is the reason why $M(f)$ is called Mahler measure.

COROLLARY 7.5.2. *For any nonzero $f(x) \in \mathbb{C}[x]$,*

$$M(f) = e^{\int_0^1 \log\left|f\left(e^{2\pi i\theta}\right)\right|\, d\theta}.$$

PROOF. Let $d = \deg(f)$, then there exist $a_d, \alpha_1, \ldots, \alpha_d \in \mathbb{C}$ such that

$$f(x) = a_d \prod_{j=1}^{d} (x - \alpha_j),$$

and so

$$\log\left|f\left(e^{2\pi i\theta}\right)\right| = \log|a_d| + \sum_{j=1}^{d} \log\left|\alpha_j - e^{2\pi i\theta}\right|.$$

Therefore, by Theorem 7.5.1

$$\begin{aligned} e^{\int_0^1 \log\left|f\left(e^{2\pi i\theta}\right)\right|\, d\theta} &= |a_d| \prod_{j=1}^{d} e^{\int_0^1 \log\left|\alpha_j - e^{2\pi i\theta}\right|\, d\theta} \\ &= |a_d| \prod_{j=1}^{d} \max\{1, |\alpha|\} = M(f). \end{aligned}$$

This completes the proof. □

Corollary 7.5.2 hence implies that Mahler measure is a continuous function on $\mathbb{C}[x]$.

Now let us review some of the results in the direction of Lehmer's conjecture. For a polynomial $f(x) \in \mathbb{C}[x]$ of degree d, define its *reciprocal polynomial*

$$f^*(x) = x^d f\left(x^{-1}\right).$$

We will say that $f(x)$ is *reciprocal* if $f(x) = f^*(x)$, and that it is *non-reciprocal* otherwise. The reciprocal condition on $f(x)$ is equivalent to the condition that its coefficients read forward same as backward. The following theorem was proved by C. Smyth in 1971 [**Smy71**].

THEOREM 7.5.3. *If $f(x) \in \mathbb{Z}[x]$ is non-reciprocal such that $f(0)f(1) \neq 0$, then*

$$M(f) \geq M(x^3 - x - 1) = 1.324...$$

Smyth's theorem implies that we can restrict our attention to reciprocal polynomials. However, in this case no absolute lower bound is known. The best unconditional bound known depends on the degree of $f(x)$: it was obtained by E. Dobrowolski in 1979 [**Dob79**] and looks as follows.

THEOREM 7.5.4. *Let $\varepsilon > 0$. Then there exists $d_0(\varepsilon) \in \mathbb{Z}$ such that for every $f(x) \in \mathbb{Z}[x]$ of degree $d > d_0(\varepsilon)$,*

$$M(f) > 1 + (1 - \varepsilon)\left(\frac{\log\log d}{\log d}\right)^3. \tag{7.40}$$

Actually, Cantor, Straus, Loubotin, and Rausch slightly improved Dobrowolski's method in several papers published between 1982 and 1985, which allowed to replace $(1 - \varepsilon)$ in the lower bound of (7.40) with $(9/4 - \varepsilon)$, however getting rid of the dependence on d seems to be a major obstacle. Dobrowolski's theorem stated in

terms of height of algebraic numbers implies that there exists a positive constant C_1 such that for every $\alpha \in \mathbb{A}$ of degree d, we have

$$h_1(\alpha) > \left(1 + C_1 \left(\frac{\log\log d}{\log d}\right)^3\right)^{\frac{1}{d}},$$

which is weaker than the bound of the form $C^{\frac{1}{d}}$ proposed by Conjecture 7.5.1 where the constant C is independent of d. However, if we consider pairs of algebraic numbers α and $1-\alpha$, it turns out to be possible to produce a lower bound on the product of their heights which is completely independent of d. The following Lehmer-type result with an implicit constant was first obtained by Zhang in 1992 with the use of Arakelov theory; in 1995 Zagier obtained an elementary proof of it with an explicit constant.

THEOREM 7.5.5 (Zhang, Zagier). *Let $\alpha \in \mathbb{A}$ be not equal to $0, 1$, or $\frac{1\pm\sqrt{-3}}{2}$, then*

$$h_1(\alpha)h_1(1-\alpha) \geq \sqrt{\frac{1+\sqrt{5}}{2}},$$

with equality if and only if α or $1-\alpha$ is a primitive 10-th root of unity.

Notice that Theorem 7.5.5 can be viewed as a uniform lower bound on the height of algebraic points on the curve

$$x + y = 1.$$

There is also a uniform upper bound on all such points that satisfy one additional condition.

DEFINITION 7.5.1. Two numbers $\alpha, \beta \in \mathbb{A}$ are called *multiplicatively dependent* if there exist $t \in \mathbb{A}$, $a, b \in \mathbb{Z}$, and $\varepsilon_1, \varepsilon_2$ roots of unity such that

$$\alpha = \varepsilon_1 t^a, \beta = \varepsilon_2 t^b.$$

The following bound was obtained in 1998 by P. Cohen and U. Zannier in [**CZ98**].

THEOREM 7.5.6. *Let $0, 1 \neq \alpha \in \mathbb{A}$ be such that α and $1-\alpha$ are multiplicatively dependent, then*

$$h_1(\alpha) \leq 2,$$

with equality if and only if $\alpha = 2$ or $1/2$.

Hence by combining the results of Theorems 7.5.5 and 7.5.6, we conclude that if (x, y) is an algebraic point on the curve

$$x + y = 1$$

with multiplicatively dependent coordinates $\neq 0, 1, \frac{1\pm\sqrt{-3}}{2}$, then

$$\sqrt{\frac{1+\sqrt{5}}{2}} \leq h_1(x)h_1(y) \leq 4.$$

This is a rare example of a uniform bound for the height of a set of algebraic points on an algebraic variety.

7.6. Points of small height

In the previous section we saw an example of uniform bounds on height of points on a simple curve. In a more general situation we cannot hope to obtain such nice results. In fact, on higher dimensional varieties it is usually very difficult to prove existence of even a single point of relatively small height. If however we were to prove existence with an explicit bound on height, it would reduce the search for such points to a finite set due to Northcott's theorem. In other words, we can think of a bound on the height of a point on variety over a fixed number field K as a *search bound*.

We start discussing this topic with the case of a linear variety, i.e. by revisiting Siegel's lemma we introduced in Section 6.4, but this time in a more powerful form. Let us look back at Theorems 6.4.1 and 6.4.2: they provide a bound on the height of a solution to a homogeneous linear system in terms of the height of the coefficient matrix A of this system. Notice, however, that if we multiply A by 2 the solution space does not change, but height of A certainly changes in a way that would affect the upper bounds of these theorems. This problem is circumvented by using Schmidt height (7.36) on the solution space instead of the height of a coefficient matrix. The following version of Siegel's lemma was proved by Bombieri and Vaaler in 1983, see [**BV83**].

THEOREM 7.6.1. *Let V be an ℓ-dimensional subspace of K^n, $\ell < n$. Then there exists a basis $\boldsymbol{x}_1, ..., \boldsymbol{x}_\ell \in \mathcal{O}_K^n$ for V such that*

$$\prod_{i=1}^{\ell} H(\boldsymbol{x}_i) \le \left\{ n|\Delta_K|^{1/d} \right\}^{\ell/2} H(V), \tag{7.41}$$

where Δ_K is the discriminant of K, and $d = [K : \mathbb{Q}]$ as usual.

In other words, Theorem 7.6.1 states that a subspace V of K^n has a basis of relatively small height with coordinates in $\mathcal{O}_K$, where the bound on the height is explicit and depends on the height of V. In particular, it implies the existence of a non-zero point of small height in V, bounded as follows.

COROLLARY 7.6.2. *Let V be an ℓ-dimensional subspace of K^n, $\ell < n$. Then there exists $\mathbf{0} \neq \boldsymbol{x} \in \mathcal{O}_K^n \cap V$ such that*

$$H(\boldsymbol{x}) \le \left\{ n|\Delta_K|^{1/d} \right\}^{1/2} H(V)^{1/\ell}. \tag{7.42}$$

This corollary can be viewed as a generalization of Minkowski's Convex Body Theorem, specifically of Corollary 1.5.1 (Problem 7.19). The dependence on $H(V)$ in (7.41) and (7.42) is sharp. An analogous bound has been proved for a small-height basis of a subspace V of $\mathbb{A}^n$ by Roy and Thunder, see [**RT96**], where the constant in the upper bound does not depend on any number field; this is often desired, since Δ_K can be quite large.

Next we consider the case of a quadratic hypersurface. Namely, let

$$F(\boldsymbol{X}) = \sum_{i=1}^{n} \sum_{j=1}^{n} f_{ij} X_i X_j \in K[X_1, \dots, X_n]$$

be a quadratic form in n variables with coefficients in the number field K of degree d over $\mathbb{Q}$. We say that F is *isotropic* over K if there exists $\mathbf{0} \neq \boldsymbol{x} \in K^n$ such that

$F(\boldsymbol{x}) = 0$. Provided that F is isotropic over K, we are interested in proving the existence of a non-zero point of bounded height in the quadratic variety

$$\mathcal{V}_K(F) = \{\boldsymbol{x} \in K^n : F(\boldsymbol{x}) = 0\}$$

with an explicit bound on height. The following theorem was originally proved by Cassels in 1955 for the case $K = \mathbb{Q}$, and then extended to arbitrary number fields by Raghavan in 1975; see [**Cas55**] and [**Rag75**].

THEOREM 7.6.3. *Let F be a quadratic form, which is isotropic over K as above, then there exists $\mathbf{0} \neq \boldsymbol{x} \in \mathcal{V}_K(F)$ such that*

$$H(\boldsymbol{x}) \leq c_1(K, n) H(F)^{\frac{n-1}{2}},$$

where the constant $c_1(K, n)$ in the upper bound is explicit and depends on K and n.

The dependence of $H(F)$ in the upper bound of Theorem 7.6.3 is best possible, as demonstrated by the following example due to M. Kneser [**Cas56**]. Consider an integral quadratic form

$$F(\boldsymbol{X}) = X_1^2 - \sum_{i=2}^{n} (X_i - cX_{i-1})^2 = (1-c^2)X_1^2 - (1+c^2)\sum_{i=2}^{n-1} X_i^2 - X_n^2 + 2c\sum_{i=2}^{n} X_{i-1}X_i$$

for some large positive integer c. Then $H(F) = 1 + c^2$. Now, if $F(\boldsymbol{x}) = 0$ for some $\mathbf{0} \neq \boldsymbol{x} \in \mathbb{Z}^n$, then it must be true that

$$0 \neq x_1^2 = \sum_{i=2}^{n} (x_i - cx_{i-1})^2 = y_2^2 + \cdots + y_n^2,$$

where $y_i = x_i - cx_{i-1}$ for each $2 \leq i \leq n$. We can express

$$x_n = y_n + cy_{n-1} + \cdots + c^{n-1}y_2 + c^{n-1}x_1.$$

Then the smallest possible absolute value of x_n becomes

$$(c^{n-1} - c^{n-2})|x_1| > \frac{1}{2}c^{n-1} = \frac{1}{2}\left(H(F) - 1\right)^{\frac{n-1}{2}}.$$

Let us now consider the case when instead of being a quadratic form, F is an inhomogeneous quadratic polynomial over K. In other words, let

$$F(\boldsymbol{X}) = \sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij}X_iX_j + \sum_{i=1}^{n} f_{0i}X_i + f_{00} \in K[X_1, \ldots, X_n],$$

and suppose that

$$\mathcal{V}_K(F) = \{\boldsymbol{x} \in K^n : F(\boldsymbol{x}) = 0\}$$

is not empty. We want to prove the existence of a point $\boldsymbol{x} \in \mathcal{V}_K(F)$ of bounded height. Notice that we can "homogenize" F by adding one more variable X_0, i.e. consider the quadratic form in $n+1$ variables

$$F(\boldsymbol{X}) = \sum_{i=0}^{n}\sum_{j=1}^{n} f_{ij}X_iX_j \in K[X_0, \ldots, X_n].$$

Problem 7.20 guarantees that a point $\boldsymbol{x} = (x_0, x_1, \ldots, x_n) \in K^{n+1}$ with $x_0 \neq 0$ is a zero of $F(X_0, \ldots, X_n)$ if and only if the point $\boldsymbol{x}' = (x_1, \ldots, x_n) \in K^n$ is a zero of

$$F_1(X_1, \ldots, X_n) := F(1, X_1, \ldots, X_n).$$

Hence we want to look for small-height zeros of F with additional condition $x_0 \neq 0$. The following theorem was originally proved for the case $K = \mathbb{Q}$ by Masser in 1998 [**Mas98**] and extended over an arbitrary number field in [**Fuk04**].

THEOREM 7.6.4. *Let F be a quadratic form in $n+1 \geq 2$ variables with coefficients in K. Suppose that there exists $\boldsymbol{x} = (x_0, ..., x_n) \in K^{n+1}$ such that $F(\boldsymbol{x}) = 0$ and $x_0 \neq 0$, then there exists such $\boldsymbol{x}$ with*

$$H(\boldsymbol{x}) \leq c_2(K, n) H(F)^{\frac{n+1}{2}}, \tag{7.43}$$

where the constant in the upper bound is explicit, and depends in particular on Δ_K.

This implies that if an inhomogeneous quadratic polynomial in n variables with coefficients in K has a zero over K, then it has such a zero of height bounded as in (7.43). The exponent in the upper bound of (7.43) is best possible as demonstrated by an example of Masser presented in [**Mas98**]: for a fixed integer $a \geq 2$, consider the inhomogeneous quadratic polynomial

$$F(X_1, \ldots, X_n) = 2X_1 - (X_2 - aX_1)^2 - \cdots - (X_n - aX_{n-1})^2 - 2a^2.$$

The height of this polynomial is a constant multiple of a^2. It is not very difficult to show that the "smallest" rational zeros this polynomial has are of the height $\geq c(n)a^{n+1} = c'(n)H(F)^{\frac{n+1}{2}}$ for appropriate dimensional constants $c(n)$, $c'(n)$. See also [**Fuk13**] for a survey of a vast variety of further results on Cassels' theorem and its many generalizations, including the more complicated inhomogeneous situation over the ring of integers instead of a field.

What can be said about bounds on height of solutions of polynomials of degree higher than 2 in an arbitrary number of variables over a fixed number field K? There are some known results in this direction for rational cubic forms in large enough number of variables: the current state of the art in this direction is a rather technical result obtained in [**BDE12**]. For sufficiently general polynomials of higher degree, this problem seems to be out of reach at the present time. In fact, such a bound would provide an algorithm to decide whether a Diophantine equation has an integral solution, and so would imply a positive answer to Hilbert's 10th problem in this case, i.e. this would mean that there exists an algorithm to decide whether such an equation has nontrivial integral solutions. However, by the famous theorem of Matijasevich [**Mat70**] Hilbert's 10th problem is undecidable. This means that in general such bounds do not exist over $\mathbb{Q}$; in fact, they seem unlikely to exist over any fixed number field even for a quartic polynomial (see [**Mas02**] for further details). The problem becomes easier if we allow for solutions to lie over some extension of K of bounded degree. The following basic bound is easy to prove (see [**Fuk09b**]).

PROPOSITION 7.6.5. *Let $d \geq 1$, $n \geq 2$, and $F(X_1, ..., X_n)$ be a homogeneous polynomial in n variables of degree d with coefficients in a number field K. There exists $\mathbf{0} \neq \boldsymbol{z} \in \mathbb{A}^n$ with $\deg_K(\boldsymbol{z}) \leq d$ such that $F(\boldsymbol{z}) = 0$ and*

$$H(\boldsymbol{z}) \leq \sqrt{2}\ H(F)^{1/d}.$$

Here $\deg_K(\boldsymbol{z})$ is the degree $[L : K]$, where L is the number field generated over K by the coordinates of the point $\boldsymbol{z}$.

Further investigations of small-height solutions of polynomial equations have strong connections with arithmetic geometry via the study of points of bounded height on algebraic varieties. This subject requires a more extensive theory of height functions. An excellent source for further reading in this direction is [**BG06**].

7.7. Problems

PROBLEM 7.1. *Prove Corollary 7.3.5.*

PROBLEM 7.2. *Prove that*

$$\sum_{i=0}^{j} \binom{n+i-2}{n-2} = \binom{n+j-1}{n-1}.$$

PROBLEM 7.3. *Prove that $\sim$ as defined in Definition 7.4.1 is an equivalence relation on the set of all absolute values on a field K.*

PROBLEM 7.4. *Prove that the only absolute value equivalent to the trivial one is itself.*

PROBLEM 7.5. *Prove that two absolute values $|\ |_1$ and $|\ |_2$ on a field K are equivalent if and only if they induce the same topology.*

PROBLEM 7.6. *Prove that $|\ |_\infty$ is an archimedean absolute value on $\mathbb{Q}$.*

PROBLEM 7.7. *Prove that $|\ |_p$ is a non-archimedean absolute value on $\mathbb{Q}$ for each prime $p \in \mathbb{Z}$.*

PROBLEM 7.8. *Prove that*

$$\mathbb{Z} = \{a \in \mathbb{Q} : |a|_p \leq 1 \ \forall \text{ primes } p \in \mathbb{Z}\}.$$

PROBLEM 7.9. *Prove Lemma 7.4.2.*

PROBLEM 7.10. *Prove that $I = \{a \in \mathbb{Z} : |a| < 1\}$ is a prime ideal in $\mathbb{Z}$.*

PROBLEM 7.11. *Prove Theorem 7.4.3 (Artin - Whaples Product Formula over $\mathbb{Q}$): if $0 \neq a \in \mathbb{Q}$, then*

$$|a|_\infty \prod_{p \in \mathcal{P}} |a|_p = 1.$$

PROBLEM 7.12. *Prove that* (7.31) *defines an absolute value on a number field K, which restricts to the usual p-adic absolute value on $\mathbb{Q}$.*

PROBLEM 7.13. *Prove Lemma 7.4.5.*

PROBLEM 7.14. *Prove Lemma 7.4.6.*

PROBLEM 7.15. *Prove Lemma 7.4.7.*

PROBLEM 7.16. *Let K be a number field, $v \in M(K)$, $v|\infty$, and let P and Q be polynomials in one variable of degree $\leq n$ with coefficients in K. Use Cauchy's inequality to prove that*

$$H_v(PQ) \leq \sqrt{n+1}\; H_v(P)H_v(Q).$$

PROBLEM 7.17. *Prove that for every $\alpha \in \mathbb{A}$,*

$$h_1(\alpha) \leq h(\alpha) \leq \sqrt{2}\; h_1(\alpha).$$

PROBLEM 7.18. *Let $\alpha \in \mathbb{A}$ be of degree d, and let $f(x) \in \mathbb{Z}[x]$ be its minimal polynomial. Prove that*

$$h_1(\alpha) = M(f)^{\frac{1}{d}}.$$

PROBLEM 7.19. *Let A be an $n \times \ell$ integer matrix of rank $\ell < n$. Let $\Lambda = A\mathbb{Z}^\ell$ be a sublattice of $\mathbb{Z}^n$ of rank ℓ. Let*

$$V = \operatorname{span}_{\mathbb{R}} \Lambda = A\mathbb{R}^\ell$$

be the ℓ-dimensional subspace of $\mathbb{R}^n$ spanned by Λ, then $\Lambda = V \cap \mathbb{Z}^n$. The famous Cauchy-Binet formula then implies that the Schmidt height

$$H(V) = \sqrt{\det\left(A^\top A\right)} = \det \Lambda.$$

Use Cauchy-Binet formula along with Corollary 1.5.1 (stated verbatim for the full-rank lattice Λ in the Euclidean space V) to prove that there exists $\mathbf{0} \neq \boldsymbol{x} \in \Lambda$ such that

$$H(\boldsymbol{x}) \leq c_n H(V)^{1/\ell},$$

for some constant c_n depending only on n.

PROBLEM 7.20. *Let K be a number field. Prove that a point $\boldsymbol{x} = (x_0, x_1, \ldots, x_n) \in K^{n+1}$ with $x_0 \neq 0$ is a zero of a quadratic form $F(X_0, \ldots, X_n)$ if and only if the point $\boldsymbol{x}' = (x_1, \ldots, x_n) \in K^n$ is a zero of the quadratic polynomial*

$$F_1(X_1, \ldots, X_n) := F(1, X_1, \ldots, X_n).$$

Appendices

APPENDIX A

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 $\mathbf{0}$ denoting the identity element and $n\boldsymbol{x}$, for $n \in \mathbb{Z}$ and $\boldsymbol{x} \in G$, denoting the n-th power of the element $\boldsymbol{x}$. A collection of elements $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ in an abelian group G is called *linearly independent* if whenever

$$n_1\boldsymbol{x}_1 + \cdots + n_k\boldsymbol{x}_k = \mathbf{0}$$

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 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ be a basis for G, then

$$G = \left\{ \sum_{i=1}^{k} n_i\boldsymbol{x}_i : n_1, \ldots, n_k \in \mathbb{Z} \right\}.$$

Define a map $\varphi : G \to \mathbb{Z}^k$, given by

$$\varphi\left(\sum_{i=1}^{k} n_i\boldsymbol{x}_i \right) = \sum_{i=1}^{k} n_i\boldsymbol{e}_i.$$

We leave it to the reader to check that this is a group isomorphism. □

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 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ 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 $\mathbb{Z}^k \ncong \mathbb{Z}^m$ 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 $k = m$. □

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 1.3.

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 $\boldsymbol{x} + H$ where $\boldsymbol{x} \in G$. 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:

$$(\boldsymbol{x} + H) + (\boldsymbol{y} + H) = (\boldsymbol{x} + \boldsymbol{y}) + H.$$

G/H is called the *quotient group* of G modulo H. The identity element in this group is the trivial coset $\mathbf{0} + H = H = \boldsymbol{x} + H$ for every $\boldsymbol{x} \in H$, and inverse of $\boldsymbol{y} + H$ is $-\boldsymbol{y} + H$ for every $\boldsymbol{y} \in G$. 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 $\operatorname{Ker}(\varphi)$ is a subgroup of G and $\varphi(G)$ is a subgroup of E. The First Isomorphism Theorem states that

$$G/\operatorname{Ker}(\varphi) \cong \varphi(G). \tag{A.1}$$

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 addivtiely written. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ be a generating set for G, then every element $\boldsymbol{y} \in G$ is expressible as

$$\boldsymbol{y} = \sum_{i=1}^{k} n_i \boldsymbol{x}_i$$

for some $n_1, \ldots, n_k \in \mathbb{Z}$. Define a map $\varphi : \mathbb{Z}^k \to G$, given by

$$\varphi\left(\sum_{i=1}^{k} n_i \boldsymbol{e}_i\right) = \sum_{i=1}^{k} n_i \boldsymbol{x}_i.$$

We leave it to the reader to check that this is a group homomorphism. Let $K = \operatorname{Ker}(\varphi)$, 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 φ. Then M is also free abelian of rank $m \leq k$. Furthermore, M contains K: indeed, for every $\boldsymbol{x} \in K$, $\varphi(\boldsymbol{x}) = \mathbf{0} \in H$, hence $\boldsymbol{x} \in M$. 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 $\mathbb{Z}^m/\mathbb{Z}^\ell \cong \{\mathbf{0}\}$, the trivial group. Then assume that $m > \ell$. Considering the standard basis $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m$ for $\mathbb{Z}^m$, we can view $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_\ell$ 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 $\sum_{i=1}^{m} n_i \boldsymbol{e}_i + \mathbb{Z}^\ell$ in $\mathbb{Z}^m/\mathbb{Z}^\ell$ to $\sum_{i=m-\ell+1}^{m} n_i \boldsymbol{e}_i$

in $\mathbb{Z}^{m-\ell}$ (this is easily checked to be a group isomorphism). Now, $\mathbb{Z}^{m-\ell}$ is finitely generated, and hence we are done. □

Problems

PROBLEM A.1. *Suppose that G is a free abelian group. Prove that the following property holds: whenever $n\boldsymbol{x} = \mathbf{0}$ for some $n \in \mathbb{Z}$ and $\boldsymbol{x} \in G$, then either $n = 0$ or $\boldsymbol{x} = \mathbf{0}$.*

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

$$D_r^o = \{z \in \mathbb{C} : |z| < r\}.$$

Then $f(z)$ assumes its maximum value on D_r on its boundary

$$\partial D_r = \{z \in \mathbb{C} : |z| = r\} = D_r \setminus D_r^o.$$

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_r^o is a region in $\mathbb{C}$, by Theorem B.1 $f(z)$ cannot have a maximum on D_r^o. Thus it must be assumed on the boundary. □

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

$$1/|p(x)| \leq 1/|p(\beta)| \ \forall \ x \in D_r.$$

Now pick $t > r$ and D_t^o be the open disk of radius t. Then $D_r \subsetneq D_t^o$, and for all $x \in D_t^o \setminus D_r$,

$$1/|p(x)| < 1/|p(\alpha)| \leq 1/|p(\beta)|.$$

Hence $1/|p(x)|$ assumes its maximum on D_t^o at $x = \beta$. Since $1/p(x)$ is not a constant function (degree of $p(x)$ is > 0) and D_t^o is a region (it is open and connected), this violates the Maximum Modulus Principle. Hence $p(x)$ must have a zero in $\mathbb{C}$. $\square$

APPENDIX C

Brief remarks on exponential and logarithmic functions

We recall the basic properties of the exponential and logarithmic functions. We give only an abbreviated and restrictive definition here; for a detailed treatment of this important topic, the reader may want to consult a good book on complex analysis, such as [**Rud87**].

We first define the *exponential function* $f_a : \mathbb{C} \to \mathbb{C}$ given by $f_a(x) = a^x$ for each base $a \in \mathbb{C}$ and outline some of its basic properties. We will do this in multiples steps. First assume that $0 \neq a \in \mathbb{C}$ and $b \in \mathbb{N}$, then

$$a^b := a \cdot \cdots \cdot a \text{ taken } b \text{ times}, \ a^0 := 1, \ 0^b := 0, \ a^{-b} := (a^{-1})^b.$$

If $b = \frac{m}{n} \in \mathbb{Q}_{>0}$ with $\gcd(m, n) = 1$ (a fraction can always be reduced) and a is a positive real number, then a^b is defined as the unique positive real root (Problem C.1) of the polynomial

$$x^n - a^m \in \mathbb{R}[x].$$

We also define $a^{-b} := (a^{-1})^b$, same as for integer exponents. Now, if $b \in \mathbb{R}$, then there exists a rational Cauchy sequence $\{c_n\}_{n=1}^{\infty}$ converging to b, and so we define

$$a^b := \lim_{n \to \infty} a^{c_n}. \tag{C.1}$$

Equation (C.1) above needs some clarification. Consider the sequence $(a_n) = \{a^{c_n}\}_{n=1}^{\infty}$. Each element a_n of this sequence is a real number, and it can be shown that this sequence is a *Cauchy sequence of real number*, meaning that for every positive *real* number ϵ there exists a positive integer N such that for all integers $m, n > N$,

$$|a_m - a_n| < \epsilon.$$

A Cauchy sequence of real numbers always converges to a real number. This means that every Cauchy sequence of real numbers is equivalent to some Cauchy sequence of rational numbers. A field with this property is called *complete*, and $\mathbb{R}$ is the most common example of a complete field. Hence a^b in (C.1) is precisely the equivalence class of the Cauchy sequence $\{a^{c_n}\}_{n=1}^{\infty}$.

LEMMA C.1. *Let $a, b \in \mathbb{R}_{>0}$, $c, d \in \mathbb{R}$. Then*

$$a^{c+d} = a^c a^d, \ (ab)^c = a^c b^c, a^{cd} = (a^c)^d. \tag{C.2}$$

This means that for each $a \in \mathbb{R}_{>0}$, $a \neq 1$, the exponential map $x \mapsto a^x$ is an injective group homomorphism from $\mathbb{Z}$, $\mathbb{Q}$, or $\mathbb{R}$ (viewed as additive groups) to $\mathbb{R}^+ = \mathbb{R}_{>0}$ (viewed as a multiplicative group).

PROOF. Problem C.2. □

In fact, the exponential map is an isomorphism of abelian groups $(\mathbb{R}, +)$ and $(\mathbb{R}^+, \cdot)$ (we do not prove it here, but a proof can be found in many standard algebra and analysis books). The inverse of this isomorphism is called the *logarithmic function* with base a, denoted $\log_a x$.

Next, let us recall a definition of e. Let $a \in \mathbb{R}_{>0}$ and consider the exponential function with base a, $f_a(x) = a^x$ for $x \in \mathbb{R}$. Notice that the derivative of this function at x is

$$f_a'(x) = \lim_{h\to 0} \frac{a^{x+h} - a^x}{h} = a^x \lim_{h\to 0} \frac{a^h - 1}{h},$$

and so

$$f_a'(0) = \lim_{h\to 0} \frac{a^h - 1}{h}.$$

This limit depends only on a, and there exists a unique value of a for which this limit is equal to 1. This value is called e. Hence e is the unique value of the base a for which the graph of $f_a(x)$ has slope $= 1$ at $x = 0$, as well as $f_a(x) = f_a'(x)$ for all x. It is also possible to define e in terms of its well-known properties:

$$e = \sum_{n=0}^{\infty} \frac{1}{n!} = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n = 2.71828\ldots,$$

as well as

$$\int_1^e \frac{1}{x}\, dx = 1.$$

This number is denoted by e in honor of Leonard Euler, who was first to prove its irrationality in 1737, although the number itself was first introduced by Jacob Bernoulli in 1683.

Recall from calculus the following power series expansions:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!},\ \cos x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!},\ \sin x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!}. \tag{C.3}$$

These power series converge for all $x \in \mathbb{C}$ (Problem C.3), and hence one can treat these power series expansions as definitions of e^x, $\cos x$, and $\sin x$ for any complex number x. In case of e^x, one can also derive an easier to use Euler's formula (Problem C.4):

$$e^{ix} = \cos x + i \sin x \tag{C.4}$$

for all $x \in \mathbb{C}$. Notice that the argument of a complex number is not uniquely defined: it is easy to see from Euler's formula that if θ is equal to $\arg(a)$ then so is $\theta + 2\pi n$ for any $n \in \mathbb{Z}$. This problem leads to the general logarithmic function not actually being a function in the usual meaning of this word, but a *multivalued function* instead. We avoid this complication by restricting the argument: from here on, we will assume that

$$-\pi \le \arg(a) < \pi \ \forall\ a \in \mathbb{C}$$

whenever it matters. Placing this restriction is usually called selecting the *principal branch*.

Now let $a \in \mathbb{C}$ and $b \in \mathbb{R}$. We can define

$$a^b = \left(|a| e^{i \arg(a)}\right)^b := |a|^b e^{ib \arg(a)} = |a|^b \left(\cos(b \arg(a)) + i \sin(b \arg(a))\right),$$

by Euler's formula. In other words, a^b is the complex number with modulus $|a|^b$ and argument $b\arg(a)$. It is easy to notice that the properties (C.2) of Exercise C.1 apply to this situation as well.

We can now define the exponential for any base and exponent. We want our general definition to be consistent with all previous cases, which in particular means that me must have

$$(ab)^c = a^c b^c,\ a^{b+c} = a^b a^c,\ a^{bc} = (a^b)^c$$

for all $a, b, c \in \mathbb{C}$. Let $a, b \in \mathbb{C}$. It will be convenient to write $a = |a|e^{i\arg(a)}$, $b = b_1 + b_2 i$. Then

$$a^b = a^{b_1} a^{ib_2}.$$

We already know what a^{b_1} is, so it only remains to define

$$a^{ib_2} = |a|^{ib_2}\left(e^{i\arg(a)}\right)^{ib_2} = |a|^{ib_2} e^{i^2 b_2 \arg(a)} = |a|^{ib_2} e^{-b_2\arg(a)}.$$

Now, $|a| \in \mathbb{R}^\times$, and hence by our discussion above

$$|a| = e^{\ln|a|},$$

where $\ln := \log_e$. Then

$$|a|^{ib_2} = \left(e^{\ln|a|}\right)^{ib_2} = e^{ib_2\ln|a|}.$$

Thus we have

$$a^b := |a|^{b_1} e^{-b_2\arg(a)} e^{i(b_1\arg(a)+b_2\ln|a|))} = |a|^{b_1 - \frac{b_2\arg(a)}{\ln|a|}} e^{i(b_1\arg(a)+b_2\ln|a|))} \tag{C.5}$$

for any $a = |a|e^{i\arg(a)}$ and $b = b_1 + ib_2$ in $\mathbb{C}$.

Now for every $a \in \mathbb{C}$, we have the exponential function with base a, $f_a : \mathbb{C} \to \mathbb{C}$ given by $f_a(x) = a^x$. It is a homomorphism of groups $(\mathbb{C}, +)$ and $(\mathbb{C}^\times, \cdot)$, which is surjective whenever $a \neq 0, \pm 1$, however is not injective: its kernel is equal to $\{2n\pi i : n \in \mathbb{Z}\}$ as can be seen from Euler's formula. Restricting the argument of x to the interval $[-\pi, \pi)$, as discussed above, we can define the inverse of f_a, the *logarithmic function*, denoted by $\log_a$: since f_a is surjective, for each $y \in \mathbb{C}$ there exists the unique $x \in \mathbb{C}$ with $\arg(x) \in [-\pi, \pi)$ such that $a^x = y$; define $\log_a(y)$ to be this x. Unfortunately, our restriction of argument causes the logarithmic function not to be continuous. This difficulty can be overcome by introduction of a *Riemann surface* for the logarithm function, which is usually done in complex analysis (see, for instance, [**Rud87**]).

Problems

PROBLEM C.1. *Let $p(x) = x^n - a^m$ for $n, m \in \mathbb{N}$ with $\gcd(m, n) = 1$ and $a \in \mathbb{R}_{>0}$, as above. Prove that $p(x)$ has precisely one positive real root.*

PROBLEM C.2. *Prove Lemma C.1.*

PROBLEM C.3. *Prove that the power series in* (C.3) *converge for all $x \in \mathbb{C}$.*

PROBLEM C.4. *Use expansions* (C.3) *to prove Euler's formula, established by him in 1740:*

$$e^{ix} = \cos x + i \sin x$$

for all $x \in \mathbb{C}$. *Furthermore, using Euler's formula, prove that any complex number* $a + bi$ *can be written as*

$$a + bi = |a + bi| e^{i\theta} = \sqrt{a^2 + b^2}\ e^{i\theta}$$

for some $\theta \in \mathbb{R}$. *Here* $\sqrt{a^2 + b^2}$ *is called the modulus and* θ *the argument of* $a + bi$, *denoted* $\arg(a + bi)$. *It is not hard to notice that modulus and argument identify the complex number uniquely.*

PROBLEM C.5. *Derive a power series expansion for the exponential function* $f_a(x) = a^x$ *with base* $a \in \mathbb{C}$, $a \neq 0, \pm 1$, *which converges for all* $x \in \mathbb{C}$.

Bibliography

[AFH12] I. Aliev, L. Fukshansky, and M. Henk. Generalized Frobenius numbers: bounds and average behavior. *Acta Arithm.*, 155:53–63, 2012.

[AG07] I. Aliev and P. M. Gruber. An optimal lower bound for the Frobenius problem. *J. Number Theory*, 123(1):71–79, 2007.

[AH09] I. Aliev and M. Henk. Integer knapsacks: Average behavior of the Frobenius numbers. *Math. Oper. Res.*, 34(3):698–705, 2009.

[AH10] I. Aliev and M. Henk. On feasibility of integer knapsacks. *SIAM J. Optim.*, 20(6):2978–2993, 2010.

[AHH11] I. Aliev, M. Henk, and A. Hinrichs. Expected Frobenius numbers. *J. Comb. Theory A*, 118:525–531, 2011.

[AHL13] I. Aliev, M. Henk, and E. Linke. Integer points in knapsack polytopes and s-covering radius. *Electron. J. Combin.*, 20(2, Paper 42):17 pp., 2013.

[ALL16] I. Aliev, J. De Loera, and Q. Louveaux. Parametric polyhedra with at least k lattice points: their semigroup structure and the k-Frobenius problem. In *Recent trends in combinatorics, IMA Vol. Math. Appl., 159*, pages 753–778. Springer, 2016.

[Arn99] V. I. Arnold. Weak asymptotics for the numbers of solutions of diophantine problems. *Funct. Anal. Appl.*, 33(4):292–293, 1999.

[Arn06] V. I. Arnold. Geometry and growth rate of Frobenius numbers of additive semigroups. *Math. Phys. Anal. Geom.*, 9(2):95–108, 2006.

[Bac18] R. Bachter. On the number of perfect lattices. *J. Théor. Nombres Bordeaux*, 30(3):917–945, 2018.

[BCKV00] D. Bump, K. K. Choi, P. Kurlberg, and J. Vaaler. A local Riemann hypothesis, I. *Math. Z.*, 233(1):1–19, 2000.

[BDE12] T. D. Browning, R. Dietmann, and P. D. T. A. Elliott. Least zero of a cubic form. *Math. Ann.*, 352(3):745–778, 2012.

[BF17] A. Böttcher and L. Fukshansky. Addendum to "Lattices from equiangular tight frames". *Linear Algebra Appl.*, 531:592–601, 2017.

[BFE+19] A. Böttcher, L. Fukshansky, S. Eisenbarth, S. R. Garcia, and H. Maharaj. Spherical 2-designs and lattices from abelian groups. *Discrete Comput. Geom.*, 61(1):123–135, 2019.

[BFG+16] A. Böttcher, L. Fukshansky, S. R. Garcia, H. Maharaj, and D. Needell. Lattices from tight equiangular frames. *Linear Algebra Appl.*, 510:395–420, 2016.

[BG06] E. Bombieri and W. Gubler. *Heights in Diophantine geometry*. New Mathematical Monographs, 4. Cambridge University Press, Cambridge, 2006.

[Bor02] P. Borwein. *Computational Excursions in Analysis and Number Theory*. Canadian Mathematical Society, 2002.

[BR04] M. Beck and S. Robins. A formula related to the Frobenius problem in two dimensions. *Number theory (New York, 2003), Springer, New York*, pages 17–23, 2004.

[BR06] M. Beck and S. Robins. *Computing the Continuous Discretely. Integer-Point Enumeration in Polyhedra*. Springer-Verlag, 2006.

[BS07] J. Bourgain and Y. G. Sinai. Limit behaviour of large Frobenius numbers. *Russ. Math. Surv.*, 62(4):713–725, 2007.

[BV83] E. Bombieri and J. D. Vaaler. On Siegel's lemma. *Invent. Math.*, 73(1):11–32, 1983.

[Cas53] J. W. S. Cassels. A short proof of the Minkowski-Hlawka theorem. *Proc. Cambridge Philos. Soc.*, 49:165–166, 1953.

[Cas55] J. W. S. Cassels. Bounds for the least solutions of homogeneous quadratic equations. *Proc. Cambridge Philos. Soc.*, 51:262–264, 1955.

[Cas56] J. W. S. Cassels. Addendum to the paper: Bounds for the least solutions of homogeneous quadratic equations. *Proc. Cambridge Philos. Soc.*, 52:602, 1956.

[Cas57] J. W. S. Cassels. *An introduction to Diophantine approximation*. Cambridge Tracts in Mathematics and Mathematical Physics, No. 45. Cambridge University Press, New York, 1957.

[Cas59] J. W. S. Cassels. *An Introduction to the Geometry of Numbers*. Springer-Verlag, 1959.

[Cas78] J. W. S. Cassels. *Rational quadratic forms*. London Mathematical Society Monographs, 13. Academic Press, Inc., 1978.

[CE03] H. Cohn and N. Elkies. New upper bounds on sphere packings. I. *Ann. of Math. (2)*, 157(2):689–714, 2003.

[CKM+17] H. Cohn, A. Kumar, S. D. Miller, D. Radchenko, and M. S. Viazovska. The sphere packing problem in dimension 24. *Ann. of Math. (2)*, 185(3):1017–1033, 2017.

[Cla] P. Clark. *Geometry of Numbers with Applications to Number Theory*.

[Coh00] H. Cohen. *A Course in Computational Algebraic Number Theory*. GTM. 138. Springer, 2000.

[CS99] J. H. Conway and N. J. A. Sloane. *Sphere Packings, Lattices, and Groups. 3rd edition*. Springer-Verlag, 1999.

[CZ98] P. B. Cohen and U. Zannier. Multiplicative dependence and bounded height, an example. *Algebraic number theory and Diophantine analysis*, pages 93–101, 1998.

[Dav51] H. Davenport. On a principle of Lipschitz. *J. London Math. Soc.*, 26:179–183, 1951.

[DF03] D. S. Dummit and R. M. Foote. *Abstract Algebra*. Wiley, 3rd edition, 2003.

[Dob79] E. Dobrowolski. On a question of lehmer and the number of irreducible factors of a polynomial. *Acta Arith.*, 34(4):391–401, 1979.

[Dys48] F. J. Dyson. On the product of four non-homogeneous linear forms. *Ann. of Math.*., 49:82–109, 1948.

[EE93] B. Edixhoven and J.-H. Evertse (Eds.). *Diophantine Approximation and Abelian Varieties*. Springer-Verlag, 1993.

[EW99] G. Everest and T. Ward. *Height of Polynomials and Entropy in Algebraic Dynamics*. Universitext, Springer-Verlag, 1999.

[Ewa96] G. Ewald. *Combinatorial convexity and algebraic geometry*. Springer-Verlag, 1996.

[FH13] L. Fukshansky and G. Henshaw. Lattice point counting and height bounds over number fields and quaternion algebras. *Online J. Anal. Comb.*, 8:20 pp., 2013.

[FM18] L. Fukshansky and N. Moshchevitin. On an effective variation of Kronecker's approximation theorem avoiding algebraic sets. *Proc. Amer. Math. Soc.*, 146:4151–4163, 2018.

[FNPX19] L. Fukshansky, D. Needell, J. Park, and Y. Xin. Lattices from tight frames and vertex transitive graphs. *Electron. J. Combin.*, 26(3):Paper No. 3.49, 30 pp., 2019.

[FR07] L. Fukshansky and S. Robins. Frobenius problem and the covering radius of a lattice. *Discrete Comput. Geom.*, 37(3):471–483, 2007.

[FS11] L. Fukshansky and A. Schürmann. Bounds on generalized Frobenius numbers. *European J. Combin.*, 32(3):361–368, 2011.

[FS20] L. Fukshansky and Y. Shi. Positive semigroups and generalized Frobenius numbers over totally real number fields. *Mosc. J. Comb. Number Theory*, 9(1):29–41, 2020.

[Fuk04] L. Fukshansky. Small zeros of quadratic forms with linear conditions. *J. Number Theory*, 108(1):29–43, 2004.

[Fuk06a] L. Fukshansky. Integral points of small height outside of a hypersurface. *Monatsh. Math.*, 147(1):25–41, 2006.

[Fuk06b] L. Fukshansky. Siegel's lemma with additional conditions. *J. Number Theory*, 120(1):13–25, 2006.

[Fuk09a] L. Fukshansky. On similarity classes of well-rounded sublattices of Z^2. *J. Number Theory*, 129(10):2530–2556, 2009.

[Fuk09b] L. Fukshansky. Search bounds for zeros of polynomials over $\overline{\mathbf{Q}}$. *Rocky Mountain J. Math.*, 39(3):789–804, 2009.

[Fuk11] L. Fukshansky. Revisiting the hexagonal lattice: on optimal lattice circle packing. *Elem. Math.*, 66(1):1–9, 2011.

[Fuk13] L. Fukshansky. Heights and quadratic forms: on Cassels' theorem and its generalizations. In W. K. Chan, L. Fukshansky, R. Schulze-Pillot, and J. D. Vaaler, editors,

Diophantine methods, lattices, and arithmetic theory of quadratic forms, Contemp. Math., 587, pages 77–94. Amer. Math. Soc., Providence, RI, 2013.

[GL87] P. M. Gruber and C. G. Lekkerkerker. *Geometry of Numbers*. North-Holland Publishing Co., 1987.

[GM16] S. M. Gonek and H. L. Montgomery. Kronecker's approximation theorem. *Indag. Math. (N.S.)*, 27:506–523, 2016.

[Hal05] T. Hales. A proof of the Kepler conjecture. *Ann. of Math. (2)*, 162(3):1065–1185, 2005.

[Hen02] M. Henk. Successive minima and lattice points. *IV International Conference in Stochastic Geometry, Convex Bodies, Empirical Measures and Applications to Engineering Science, Vol. I (Tropea, 2001). Rend. Circ. Mat. Palermo (2) Suppl. No. 70, part I*, pages 377–384, 2002.

[HGRS09] R. J. Hans-Gill, M. Raka, and R. Sehmi. On conjectures of Minkowski and Woods for $n = 7$. *J. Number Theory*, 129(5):1011–1033, 2009.

[HGRS11] R. J. Hans-Gill, M. Raka, and R. Sehmi. On conjectures of Minkowski and Woods for $n = 8$. *Acta Arith.*, 147(4):337–385, 2011.

[Hla80] E. Hlawka. Approximation von irrationalzahlen und pythagoraische tripel. In *Lectures from the Colloquium on the Occasion of Ernst Peschl's 70th birthday*, volume 121, pages 1–32. Bonner Math. Schriften, Univ. Bonn, Bonn, 1980.

[HPS08] J. Hoffstein, J. Pipher, and J. H. Silverman. *An Introduction to Mathematical Cryptography*. Springer, 2008.

[HW08] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers*. Oxford University Press; 6 edition, 2008.

[Jac90] B. Jacob. *Linear Algebra*. W.H. Freeman and Company, 1990.

[Jar41] V. Jarnik. Zwei Bemerkungen zur Geometrie de Zahlen. *Věstnik Krălovské České Společnosit Nauk*, 1941.

[Kan92] R. Kannan. Lattice translates of a polytope and the Frobenius problem. *Combinatorica*, 12(2):161–177, 1992.

[Kar13] O. Karpenkov. *Geometry of Continued Fractions*. Algorithms and Computation in Mathematics, 26. Springer, Heidelberg, 2013.

[KM15] D. Kleinbock and K. Merrill. Rational approximation on spheres. *Israel J. Math.*, 209(1):293–322, 2015.

[Kop80] H. G. Kopetzky. Rationale approximationen am einheitskreis. *Monatsh. Math.*, 89(4):293–300, 1980.

[Kop81] H. G. Kopetzky. Diophantische approximationen auf kreisen und zyklische minima von quadratischen formen. Technical Report 179, Forschungszentrum Graz, Mathematisch-Statistische Sektion, Graz, 1981.

[KR16] L. Kathuria and M. Raka. On conjectures of Minkowski and Woods for $n = 9$. *Proc. Indian Acad. Sci. Math. Sci.*, 126(4):501–548, 2016.

[Kuc09] M. Kuczma. *An introduction to the theory of functional equations and inequalities. Cauchy's equation and Jensen's inequality*. Birkhäuser, Basel, 2nd edition, 2009.

[Lad19] F. Ladisch. Lattices of finite abelian groups. *Discrete Comput. Geom.*, 2019.

[Lan94] S. Lang. *Algebraic Number Theory*. Springer-Verlag, 1994.

[Lee56] J. Leech. The problem of thirteen spheres. *Math. Gaz.*, 40:22–23, 1956.

[Leh33] D. H. Lehmer. Factorization of certain cyclotomic functions. *Ann. of Math. (2)*, 34(3):461–479, 1933.

[Li15] H. Li. Effective limit distribution of the Frobenius numbers. *Compos. Math.*, 151(5):898–916, 2015.

[Lip65] R. Lipschitz. *Monatsber. der Berliner Academie*, pages 174–185, 1865.

[LLL82] A. K. Lenstra, H. W. Lenstra, and L. Lovasz. Factoring polynomials with rational coefficients. *Math. Ann.*, 261:515–534, 1982.

[Mar03] J. Martinet. *Perfect Lattices in Euclidean Spaces*. Springer-Verlag, 2003.

[Mar10] J. Marklof. The asymptotic distribution of Frobenius numbers. *Invent. Math*, 181:179–207, 2010.

[Mas98] D. W. Masser. How to solve a quadratic equation in rationals. *Bull. London Math. Soc.*, 30(1):24–28, 1998.

[Mas02] D. W. Masser. Search bounds for Diophantine equations. *A panorama of number theory or the view from Baker's garden (Zurich, 1999)*, pages 247–259, 2002.

[Mat70] Yu. V. Matijasevich. The diophantineness of enumerable sets. *Dokl. Akad. Nauk SSSR*, 191:279–282, 1970.

[McM05] C. T. McMullen. Minkowski's conjecture, well-rounded lattices and topological dimension. *J. Amer. Math. Soc.*, 18:711–734, 2005.

[Min00] H. Minkowski. Über die annäherung an eine reelle Grösse durch rationale Zahlen. *Math. Ann.*, 54(1–2):91–124, 1900.

[Mos16] N. Moshchevitin. Über die rationalen Punkte auf der Sphäre. *Monatsh. Math.*, 179(1):105–112, 2016.

[MR14] M. R. Murty and P. Rath. *Transcendental Numbers*. Springer, New York, 2014.

[MS17] J. Marklof and A. Strömbergsson. The three gap theorem and the space of lattices. *Amer. Math. Monthly*, 124(8):741–745, 2017.

[MSSW06] C. J. Moreno and Jr. S. S. Wagstaff. *Sums of Squares of Integers*. Chapman & Hall, 2006.

[MTB06] S. J. Miller and R. Takloo-Bighash. *An invitation to modern number theory. With a foreword by Peter Sarnak*. Princeton University Press, Princeton, NJ, 2006.

[Nor49] D. G. Northcott. An inequality in the theory of arithmetic on algebraic varieties. *Proc. Camb. Phil. Soc.*, 45:502–509, 510–518, 1949.

[Rag75] S. Raghavan. Bounds of minimal solutions of diophantine equations. *Nachr. Akad. Wiss. Gottingen, Math. Phys. Kl.*, 9:109–114, 1975.

[Ram05] J. L. Ramírez Alfonsín. *The Diophantine Frobenius problem*. Oxford University Press, 2005.

[Rem23] R. Remak. Verallgemeinerung eines Minkowskischen Satzes. *Math. Z.*, 18(1):173–200, 1923.

[Rot55] K.F. Roth. Rational approximations to algebraic numbers. *Mathematika*, 2:1–20, 1955.

[RSW17] O. Regev, U. Shapira, and B. Weiss. Counterexamples to a conjecture of Woods. *Duke Math. J.*, 166(13):2443–2446, 2017.

[RT96] D. Roy and J. L. Thunder. An absolute Siegel's lemma. *J. Reine Angew. Math.*, 476:1–26, 1996.

[Rud87] W. Rudin. *Real and Complex Analysis*. McGraw-Hill Book Co., New York, 3rd edition, 1987.

[Sam70] P. Samuel. *Algebraic theory of numbers. Translated from the French by Allan J. Silberger*. Houghton Mifflin Co., Boston, Mass., 109 pp., 1970.

[Sch50] P. Scherk. Convex bodies off center. *Archiv Math.*, 3:303, 1950.

[Sch67] W. M. Schmidt. On heights of algebraic subspaces and Diophantine approximations. *Ann. of Math.*, 85(2):430–472, 1967.

[Sch80] W. M. Schmidt. *Diophantine Approximation*. Springer-Verlag, 1980.

[Sch91] W. M. Schmidt. *Diophantine Approximations and Diophantine Equations*. Springer-Verlag, 1991.

[Sku72] B. F. Skubenko. On Minkowski's conjecture for $n = 5$. *Soviet Math. Dokl.*, 13:1136–1138, 1972.

[Smy71] C. Smyth. On the product of the conjugates outside the unit circle of an algebraic integer. *Bull. London Math. Soc.*, 3:169–175, 1971.

[Spa95] P. G. Spain. Lipschitz: a new version of old principle. *Bull. London Math. Soc.*, 27:565–566, 1995.

[SSU09] V. Shchur, Ya. Sinai, and A. Ustinov. Limiting distribution of Frobenius numbers for $n = 3$. *Journal of Number Theory*, 129:2778–2789, 2009.

[ST02] I. Stewart and D. Tall. *Algebraic number theory and Fermat's last theorem*. Third edition. A K Peters, Ltd., Natick, MA, 2002.

[Str12] A. Strömbergsson. On the limit distribution of Frobenius numbers. *Acta Arith.*, 152.(1):81–107, 2012.

[SvdW53] K. Schütte and B. L. van der Waerden. Das Problem der dreizehn Kugeln. *Math. Ann.*, 125:325–334, 1953.

[Thu93] J. L. Thunder. The number of solutions of bounded height to a system of linear equations. *J. Number Theory*, 43:228–250, 1993.

[TV91] M. A. Tsfasman and S. G. Vlãdut. *Algebraic-geometric codes*. Mathematics and its Applications (Soviet Series), 58. Kluwer Academic Publishers Group, Dordrecht, 1991.

[Ust10] A. Ustinov. On the distribution of Frobenius numbers with three arguments. *Izv. Math.*, 74:1023–1049, 2010.

[Via17] M. S. Viazovska. The sphere packing problem in dimension 8. *Ann. of Math. (2)*, 185(3):991–1015, 2017.

[Wid09] M. Widmer. Counting points of fixed degree and bounded height. *Acta Arith.*, 140(2):145–168, 2009.

[Wid10] M. Widmer. Counting points of fixed degree and bounded height on linear varieties. *J. Number Theory*, 130(8):1763–1784, 2010.

[Wid12] M. Widmer. Lipschitz class, narrow class, and counting lattice points. *Proc. Amer. Math. Soc.*, 140(2):677–689, 2012.

[Woo72] A. C. Woods. Covering six space with spheres. *J. Number Theory*, 4:157–180, 1972.