MATH 149, FALL 2013

DISCRETE GEOMETRY

LECTURE NOTES

LENNY FUKSHANSKY

Contents

1. Introduction

Discrete Geometry is the study of arrangements of discrete sets of objects in space. The goal of this course is to give an introduction to some of the classical topics in the subject.

The origins of some of the topics we will discuss go back as far as early 17-th century. For instance, one of the big motivating problems in the subject is **Kepler's conjecture**, named after Johannes Kepler, who stated the conjecture in 1611. Suppose you want to pack balls of the same radius in a rectangular box with equal equal height, width, and length. One may ask for the densest possible arrangement of balls, i.e. how should we place balls into the box to maximize the number of balls that fit in? Here is one way to think of such problems. Let t be the height (= width = length) of the box, then its volume is t^3, and suppose that Λ is some arrangement of the balls inside of the box. Let us write $\operatorname{Vol}(\Lambda, t)$ to be the volume that the balls of this arrangement take up inside of the box with side t. Define the density $\delta(\Lambda)$ of this arrangement to be the limit of the ratio of the volume of the balls to the volume of the box as the side length of the box goes to infinity, i.e.

$$\delta(\Lambda) = \lim_{t \to \infty} \frac{\operatorname{Vol}(\Lambda, t)}{t^3}.$$

For which arrangement Λ is this density the largest it can be, and what is this largest value? Kepler conjectured that the maximum density, which is about 74%, is achieved by a so called **face-centered cubic** or **hexagonal close packing** arrangements. We will rigorously define these and other arrangements later in the course once we develop some necessary terminology. Kepler's conjecture has been open for nearly four hundred years, until its proof was announced by Thomas Hales in 1998 and published in 2005 ([16]). Of course analogous questions about optimal packing density of balls can be asked in spaces of different dimensions as well. The answer to such questions is only known in dimensions 1,2, and 3. However if one were to restrict to only a certain nice class of periodic arrangements associated with algebraic structures called **lattices**, then more is known. We will discuss some of these questions and results in more details later in the course.

Another example of a topic in discrete geometry that we will discuss in this course has to do with counting **integer lattice points** in various **convex sets**. Let us for instance consider a square $C(t)$ of side length $2t$, where t is an integer, centered at the origin in the plane. How many points with integer coordinates are contained in the interior or on the boundary of $C(t)$?

Exercise 1.1. *Prove that this number is equal to*

$$(2t+1)^2.$$

More generally, suppose that

$$C_N(t) = \{(x_1, \dots, x_N) \in \mathbb{R}^N : \max\{|x_1|, \dots, |x_N|\} \leq t\}$$

is a cube of side length $2t$, where t is an integer, centered at the origin in the Euclidean space $\mathbb{R}^N$. Prove that the number of points with integer coordinates in $C_N(t)$ is equal to

$$(2t+1)^N.$$

What if t is not an integer - can you generalize the formula to include such cases?

Of course we can ask the analogous question not just for a cube $C_N(t)$, but also for much more complicated sets S. It turns out that to give a precise answer to this question even in the plane is quite difficult. In special situations, however, a lot is known, and even in general something can be said as we will see. Let us give an example of the general principle. Consider the formula of Excercise 1.1, and apply Newton's binomial formula to it:

$$\begin{aligned} (2t+1)^N &= \sum_{i=0}^{N} \binom{N}{i} (2t)^i 1^{N-i} = (2t)^N + N(2t)^{N-1} + \cdots + 1 \\ &= \mathrm{Vol}(C_N(t)) + \text{ terms of smaller order.} \qquad (1) \end{aligned}$$

In other words, it seems that for large t the number of points with integer coordinates in $C_N(t)$ is approximately equal to the volume of this cube. This principle holds in more general situations as well, as we will see later in the course, however estimating the **remainder term** of this approximation in general (or, even better, getting exact formulas) is very hard. Moreover, to make sense of the formula (1) we first need to rigorously define what do we mean by volume.

Here is a slightly different related question. Suppose now that I have some **symmetric** set S centered at the origin in $\mathbb{R}^N$, so that the only integer lattice point (= point with integer coordinates) in S is the origin. Let us **homogeneously expand** S by a real parameter t, in other words consider sets of the form

$$tS = \{(tx_1, \dots, tx_N) : (x_1, \dots, x_N) \in S\},$$

where $t \in \mathbb{R}$. It is easy to see that if t is large enough, then tS will contain a non-zero integer lattice point. But how large does t have to be in order for this to happen? An answer to this question is given by a deep and influential theory of Minkowski, which has some fascinating connections to some topics in modern theoretical computer science.

We now have some basic idea of the flavor of questions asked in Discrete Geometry. In order to learn more we will need to develop some notation and machinery. Let us get to work!

2. Norms, sets, and volumes

Throughout these notes, unless explicitly stated otherwise, we will work in $\mathbb{R}^N$, where $N \geq 1$.

Definition 2.1. A function $\| \ \| : \mathbb{R}^N \to \mathbb{R}$ is called a **norm** if

(1) $\|\boldsymbol{x}\| \geq 0$ with equality if and only if $\boldsymbol{x} = \boldsymbol{0}$,
(2) $\|a\boldsymbol{x}\| = |a|\|\boldsymbol{x}\|$ for each $a \in \mathbb{R}$, $\boldsymbol{x} \in \mathbb{R}^N$,
(3) **Triangle inequality:** $\|\boldsymbol{x}+\boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ 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 $\boldsymbol{x} = (x_1, \dots, x_N) \in \mathbb{R}^N$. We also define the **sup-norm**, given by

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

Exercise 2.1. *Prove that* $\| \ \|_p$ *for each* $p \in \mathbb{Z}_{>0}$ *and* $| \ |$ *are indeed norms on* $\mathbb{R}^N$*.*

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 $\| \ \|_2$; recall that for every two points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$, Euclidean distance between them is given by

$$d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

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

Definition 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 $\mathbb{R}^N$

$$B_N = \{\boldsymbol{x} \in \mathbb{R}^N : \|\boldsymbol{x}\|_2 \leq 1\}$$

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

$$B_N^o = \{\boldsymbol{x} \in \mathbb{R}^N : \|\boldsymbol{x}\|_2 < 1\}$$

is not a compact set. If we now write

$$S_{N-1} = \{\boldsymbol{x} \in \mathbb{R}^N : \|\boldsymbol{x}\|_2 = 1\}$$

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

Exercise 2.2. *Let $\|\ \|$ be a norm on $\mathbb{R}^N$, and let $C \in \mathbb{R}$ be a positive number. Define*

$$A_N(C) = \{\boldsymbol{x} \in \mathbb{R}^N : \|\boldsymbol{x}\| \leq C\}.$$

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

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

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

It is easy to notice that whenever a convex set contains X, it must also contain $\operatorname{Co}(X)$. 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

$$\mathbb{H} = \left\{ \boldsymbol{x} \in \mathbb{R}^N : \sum_{i=1}^N a_i x_i = b \right\}, \tag{2}$$

for some $a_1, \dots, a_N, b \in \mathbb{R}$.

Exercise 2.3. *Prove that a hyperplane $\mathbb{H}$ as in (2) above 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 $1 \leq L \leq N$, to be $N-L$; thus co-dimension of $\mathbb{H}$ here is 1, as indicated above).*

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 $a_1, \dots, a_N, b \in \mathbb{R}$.

Exercise 2.4. *Prove that each convex polytope in $\mathbb{R}^N$ can be described as a bounded intersection of finitely many halfspaces, and vice versa.*

Remark 2.1. Exercise 2.4 is sometimes referred to as Minkowski-Weyl theorem.

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 here still work, as we will see later.

Definition 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 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 2.2. 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 neighbourhood $U \subseteq \partial X$ of $\boldsymbol{x}$ such that U is homeomorphic to $\mathbb{R}^{N-1}$. See Remark 2.4 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:

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

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 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 $f(X_1, \dots, X_N)$ 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.

Exercise 2.5. *Let $f(X_1, \dots, X_N)$ be a homogeneous polynomial of degree d with real coefficients. Prove that*

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

is a distance function.

As expected, distance functions are closely related to star bodies.

Exercise 2.6. *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.

In fact, a converse is also true.

Theorem 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 [6]. □

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 2.7. A subset $X \subseteq \mathbb{R}^N$ which contains $\mathbf{0}$ is called **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 Exercise 2.2, 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 $\boldsymbol{x}_1, \dots, \boldsymbol{x}_L \in \mathbb{R}^N$, i.e. $\operatorname{Co}(\mathbf{0}, \boldsymbol{x}_1, \dots, \boldsymbol{x}_L)$.

Exercise 2.7. *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.

Next we want to introduce the notion of volume for *bounded* sets in $\mathbb{R}^N$.

Definition 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 2.9. A bounded set X is said to have **Jordan volume** if its characteristic function is Riemann integrable, and then we define $\operatorname{Vol}(X)$ to be the value of this integral.

Remark 2.3. A set that has Jordan volume is also called **Jordan measurable**.

Definition 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.

Exercise 2.8. *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 $y \in Y$,*

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

Remark 2.4. 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**.

Exercise 2.9. *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?*

Theorem 2.2. *All convex sets and bounded ray sets have Jordan volume.*

Sketch of proof. We will only 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 S_{N-1} for the unit sphere centered at the origin in $\mathbb{R}^N$, i.e. $S_{N-1} = \partial B_N$. Define a map $\varphi : \partial X \to S_{N-1}$, given by

$$\varphi(\boldsymbol{x}) = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}.$$

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 $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{k(\varepsilon)} \in S_{N-1}$ such that if we let $\mathcal{C}_{\boldsymbol{x}_i}(\varepsilon)$ be an $(N-1)$-dimensional cap centered at $\boldsymbol{x}_i$ in S_{N-1} of radius ε, i.e.

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

then $S_{N-1} = \bigcup_{i=1}^{k(\varepsilon)} \mathcal{C}_{\boldsymbol{x}_i}(\varepsilon)$, 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

$$\|\boldsymbol{y}_i\|_2 = \max\{\|\boldsymbol{x}\|_2 : \boldsymbol{x} \in \varphi^{-1}\left(\mathcal{C}_{\boldsymbol{x}_i}(\varepsilon)\right)\},$$

and

$$\|\boldsymbol{z}_i\|_2 = \min\{\|\boldsymbol{x}\|_2 : \boldsymbol{x} \in \varphi^{-1}\left(\mathcal{C}_{\boldsymbol{x}_i}(\varepsilon)\right)\}.$$

Let $\delta_1(\varepsilon)$ and $\delta_2(\varepsilon)$ be minimal positive real numbers such that the spheres centered at the origin of radii $\|\boldsymbol{y}_i\|_2$ and $\|\boldsymbol{z}_i\|_2$ 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

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

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

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

Exercise 2.10. *Prove that cones like C_i^1 and C_i^2 have Jordan volume.*

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

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

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

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

□

This is Theorem 5 on p. 9 of [15], and the proof is also very similar.

3. Lattices

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

Exercise 3.1. *Prove that in this case $r \leq N$.*

Definition 3.1. A **lattice** Λ of **rank** r, $1 \leq r \leq N$, spanned by $\boldsymbol{a}_1, \dots, \boldsymbol{a}_r$ in $\mathbb{R}^N$ is the set of all possible linear combinations of the vectors $\boldsymbol{a}_1, \dots, \boldsymbol{a}_r$ with integer coefficients. In other words,

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

The set $\boldsymbol{a}_1, \dots, \boldsymbol{a}_r$ is called a **basis** for Λ. There are usually infinitely many different bases for a given lattice.

Exercise 3.2. *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 linear combinations with real coefficients of vectors from Λ).*

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 $\mathbb{R}^N$:

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

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

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

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 $\boldsymbol{a}_1, \dots, \boldsymbol{a}_r$ and $\boldsymbol{y} \in \Lambda$, then there exist $n_1, \dots, n_r \in \mathbb{Z}$ such that

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

where

$$\boldsymbol{n} = \begin{pmatrix} n_1 \\ \vdots \\ n_r \end{pmatrix} \in \mathbb{Z}^r,$$

and A is an $N \times r$ **basis matrix** for Λ of the form $A = (\boldsymbol{a}_1 \ \dots \ \boldsymbol{a}_r)$, 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 $N \times r$ 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 $U + \boldsymbol{x} = \{\boldsymbol{u} + \boldsymbol{x} : \boldsymbol{u} \in U\}$.

Exercise 3.3. *Let Λ be a lattice of rank r in $\mathbb{R}^N$. By Excercise 3.2, $V = \operatorname{span}_{\mathbb{R}} \Lambda$ is an r-dimensional subspace of $\mathbb{R}^N$. Prove that Λ is a discrete co-compact subset of V.*

We now need one more very important definition.

Definition 3.2. A subset G of $\mathbb{R}^N$ is called an **additive group** 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 $\boldsymbol{x} \in G$, $-\boldsymbol{x} \in G$.

If G and H are two additive groups in $\mathbb{R}^N$, and $H \subseteq G$, then we say that H is a **subgroup** of G.

Exercise 3.4. *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 Excercise 3.3 above. Prove that Λ and V are both additive groups, and Λ is a subgroup of V.*

Combining Excercises 3.3 and 3.4, we see that 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. Exercise 3.3 and Theorem 3.1 are basic generalizations of Theorems 1 and 2 respectively on p. 18 of [15], and the proofs are essentially the same; the idea behind this argument is quite important.

Theorem 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 $\boldsymbol{a}_1, \dots, \boldsymbol{a}_r$ in Γ such that $\Gamma = \operatorname{span}_{\mathbb{Z}}\{\boldsymbol{a}_1, \dots, \boldsymbol{a}_r\}$. We start by inductively constructing a collection of vectors $\boldsymbol{a}_1, \dots, \boldsymbol{a}_r$, and then show that it has the required properties.

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

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

and pick any $\boldsymbol{c} \in \Gamma \setminus H_{i-1}$: such $\boldsymbol{c}$ exists, since $\Gamma \nsubseteq H_{i-1}$ (otherwise Γ would not be co-compact in V). Let P_i be the closed parallelotope spanned by the vectors $\boldsymbol{a}_1, \dots, \boldsymbol{a}_{i-1}, \boldsymbol{c}$. Notice that since Γ is discrete in V, $\Gamma \cap P_i$ is a finite set. Moreover, since $\boldsymbol{c} \in P_i$, $\Gamma \cap P_i \nsubseteq H_{i-1}$. 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 $\boldsymbol{a}_1, \dots, \boldsymbol{a}_r$ be the collection of points chosen in this manner. Then we have

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

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

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

We will now show that

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

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

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

Notice that

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

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 $\boldsymbol{a}_1, \dots, \boldsymbol{a}_r$, 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 $\boldsymbol{x} \in \operatorname{span}_{\mathbb{Z}}\{\boldsymbol{a}_1, \dots, \boldsymbol{a}_r\}$. 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 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*

$$B = UA.$$

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 $\boldsymbol{a}_1, \dots, \boldsymbol{a}_N$ are column vectors of A, and u_{ij}'s are integers for all $1 \leq j \leq N$. This means that $B = UA$, 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 $A = WB$, where $W = (w_{ij})_{1 \leq i,j \leq N}$ is also an $N \times N$ matrix with integer entries. Then

$$B = UA = UWB,$$

which means that $UW = I_N$, the $N \times N$ identity matrix. Therefore

$$\det(UW) = \det(U)\det(W) = \det(I_N) = 1,$$

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

$$\Lambda \subseteq \operatorname{span}_{\mathbb{Z}}\{\boldsymbol{b}_1, \dots, \boldsymbol{b}_N\}.$$

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

Corollary 3.3. *If A and B are two basis matrices for the same lattice Λ, then*

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

Definition 3.3. The common determinant value of Corollary 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 3.4. 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.

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.

Definition 3.5. Suppose Λ is a lattice in $\mathbb{R}^N$ and $\Omega \subset \Lambda$ is a sublattice. For each $\boldsymbol{x} \in \Lambda$, the set

$$\boldsymbol{x} + \Omega = \{\boldsymbol{x} + \boldsymbol{y} : \boldsymbol{y} \in \Omega\}$$

is called a **coset** of Ω in Λ.

We now study some important properties of cosets.

Lemma 3.4. *Two cosets $\boldsymbol{x}+\Omega$ and $\boldsymbol{z}+\Omega$ of Ω in Λ are equal if and only if $\boldsymbol{x}-\boldsymbol{z} \in \Omega$.*

Proof. First assume $\boldsymbol{x}+\Omega = \boldsymbol{z}+\Omega$. Since $\boldsymbol{0} \in \Omega$,

$$\boldsymbol{x}+\boldsymbol{0} = \boldsymbol{x} \in \boldsymbol{x}+\Omega = \boldsymbol{z}+\Omega,$$

hence there exists some $\boldsymbol{y} \in \Omega$ such that $\boldsymbol{x} = \boldsymbol{z}+\boldsymbol{y}$, so

$$\boldsymbol{x}-\boldsymbol{z} = \boldsymbol{y} \in \Omega.$$

On the other hand, assume $\boldsymbol{x}-\boldsymbol{z} \in \Omega$, then there exists some $\boldsymbol{y} \in \Omega$ such that $\boldsymbol{x}-\boldsymbol{z} = \boldsymbol{y}$, therefore $\boldsymbol{x} = \boldsymbol{z}+\boldsymbol{y}$. This means that each element $\boldsymbol{x}+\boldsymbol{t} \in \boldsymbol{x}+\Omega$, where $\boldsymbol{t} \in \Omega$, can be expressed in the form

$$\boldsymbol{z}+\boldsymbol{y}+\boldsymbol{t} \in \boldsymbol{z}+\Omega,$$

since $\boldsymbol{y}+\boldsymbol{t} \in \Omega$ (this follows by Excercise 3.4, since Ω is a group, so it must be closed under addition). Thus we showed that $\boldsymbol{x}+\Omega \subseteq \boldsymbol{z}+\Omega$. In precisely the same way one can show that $\boldsymbol{z}+\Omega \subseteq \boldsymbol{x}+\Omega$, and this completes our proof. □

Corollary 3.5. *A coset $\boldsymbol{x}+\Omega$ is equal to Ω if and only if $\boldsymbol{x} \in \Omega$.*

Proof. Apply Lemma 3.4 with $\boldsymbol{z} = \boldsymbol{0} \in \Omega$. □

Exercise 3.5. *Given a lattice Λ and a real number μ, define*

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

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

Lemma 3.6. *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 $\boldsymbol{a}_1,\dots,\boldsymbol{a}_N$ be a basis for Ω and $\boldsymbol{b}_1,\dots,\boldsymbol{b}_N$ be a basis for Λ. Then

$$\operatorname{span}_{\mathbb{R}}\{\boldsymbol{a}_1,\dots,\boldsymbol{a}_N\} = \operatorname{span}_{\mathbb{R}}\{\boldsymbol{b}_1,\dots,\boldsymbol{b}_N\} = \mathbb{R}^N.$$

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

$$\begin{cases} \boldsymbol{a}_1 = u_{11}\boldsymbol{b}_1 + \dots + u_{1N}\boldsymbol{b}_N \\ \vdots \qquad \vdots \qquad\qquad \vdots \\ \boldsymbol{a}_N = u_{N1}\boldsymbol{b}_1 + \dots + u_{NN}\boldsymbol{b}_N. \end{cases}$$

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

$$\begin{cases} \boldsymbol{b}_1 = \frac{p_{11}}{q_{11}}\boldsymbol{a}_1 + \dots + \frac{p_{1N}}{q_{1N}}\boldsymbol{a}_N \\ \vdots \qquad \vdots \qquad\qquad \vdots \\ \boldsymbol{b}_N = \frac{p_{N1}}{q_{N1}}\boldsymbol{a}_1 + \dots + \frac{p_{NN}}{q_{NN}}\boldsymbol{a}_N. \end{cases}$$

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

$$\begin{cases} D\boldsymbol{b}_1 = \frac{Dp_{11}}{q_{11}}\boldsymbol{a}_1 + \cdots + \frac{Dp_{1N}}{q_{1N}}\boldsymbol{a}_N \\ \vdots \qquad\quad \vdots \qquad\qquad\quad \vdots \\ D\boldsymbol{b}_N = \frac{Dp_{N1}}{q_{N1}}\boldsymbol{a}_1 + \cdots + \frac{Dp_{NN}}{q_{NN}}\boldsymbol{a}_N \end{cases}$$

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

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

which completes the proof. □

We can now prove that a lattice always has a basis with "nice" properties with respect to a given sublattice; this is Theorem 1 on p. 11 of [6].

Theorem 3.7. *Let Λ be a lattice, and Ω a sublattice of Λ. For each basis $\boldsymbol{b}_1, \dots, \boldsymbol{b}_N$ of Λ, there exists a basis $\boldsymbol{a}_1, \dots, \boldsymbol{a}_N$ of Ω of the form*

$$\begin{cases} \boldsymbol{a}_1 = v_{11}\boldsymbol{b}_1 \\ \boldsymbol{a}_2 = v_{21}\boldsymbol{b}_1 + v_{22}\boldsymbol{b}_2 \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ \boldsymbol{a}_N = v_{N1}\boldsymbol{b}_1 + \cdots + v_{NN}\boldsymbol{b}_N, \end{cases}$$

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

Proof. Let $\boldsymbol{b}_1, \dots, \boldsymbol{b}_N$ be a basis for Λ. We will first prove the existence of a basis $\boldsymbol{a}_1, \dots, \boldsymbol{a}_N$ for Ω as claimed by the theorem. By Lemma 3.6, there exist integer multiples of $\boldsymbol{b}_1, \dots, \boldsymbol{b}_N$ in Ω, hence it is possible to choose a collection of vectors $\boldsymbol{a}_1, \dots, \boldsymbol{a}_N \in \Omega$ 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 $\boldsymbol{a}_1, \dots, \boldsymbol{a}_N$ is a basis for Ω. Clearly,

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

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

$$\Omega \subseteq \operatorname{span}_{\mathbb{Z}}\{\boldsymbol{a}_1, \dots, \boldsymbol{a}_N\}. \tag{3}$$

Suppose (3) is not true, then there exists $\boldsymbol{c} \in \Omega$ which is not in $\operatorname{span}_{\mathbb{Z}}\{\boldsymbol{a}_1, \dots, \boldsymbol{a}_N\}$. 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 $t_1, \dots, t_k$. 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

$$|t_k - s v_{kk}| < |v_{kk}|. \tag{4}$$

Then we clearly have

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

Therefore we must have $t_k - s v_{kk} \neq 0$ by minimality of k. But then (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 $|t_k - s v_{kk}|$. This proves that $\boldsymbol{c}$ like this cannot exist, and so (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 3.1. Indeed, at i-th step of the basis construction in the proof of Theorem 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. □

Exercise 3.6. *Prove that it is possible to select the coefficients v_{ij} in Theorem 3.7 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.*

Remark 3.1. Let the notation be as in Theorem 3.7. 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 $A = VB$. Then Theorem 3.7 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 3.7 can be thought of as the construction of **Hermite normal form** for an integral matrix. Exercise 3.6 places additional restrictions that make Hermite normal form unique.

Here is an important implication of Theorem 3.7; this is Lemma 1 on p. 14 of [6].

Theorem 3.8. *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 is equal to $\frac{\det(\Omega)}{\det(\Lambda)}$.*

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 3.7, and write A and B for the corresponding basis matrices. Then notice that

$$A = VB,$$

where $V = (v_{ij})_{1 \leq i,j \leq N}$ is an $N \times N$ triangular matix with entries as described in Theorem 3.7; in particular $\det(V) = \prod_{i=1}^N |v_{ii}|$. Hence

$$\det(\Omega) = |\det(A)| = |\det(V)||\det(B)| = \det(\Lambda) \prod_{i=1}^N |v_{ii}|,$$

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

$$q_1 \boldsymbol{b}_1 + \dots + q_N \boldsymbol{b}_N, \ 0 \leq q_i < v_{ii} \ \forall \ 1 \leq i \leq N,$$

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

Definition 3.6. The number of cosets of a sublattice Ω inside of a lattice Λ is called the **index** of Ω in Λ and is denoted by $[\Lambda : \Omega]$. Theorem 3.8 then guarantees that when Ω and Λ have the same rank,

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

in particular it is finite.

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

Definition 3.7. A **fundamental domain** of a lattice Λ of full rank in $\mathbb{R}^N$ is a Jordan measurable 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 $\boldsymbol{x} \neq \boldsymbol{y} \in \Lambda$, $(\mathcal{F} + \boldsymbol{x}) \cap (\mathcal{F} + \boldsymbol{y}) = \emptyset$.

Exercise 3.7. *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 $\mathbb{R}^N$.

Although each lattice has infinitely many different fundamental domains, they all have the same volume, which depends only on the lattice. This fact can be easily proved for a special class of fundamental domains.

Definition 3.8. Let Λ be a lattice, and $\boldsymbol{a}_1, \dots, \boldsymbol{a}_N$ be a basis for Λ. Then the set

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

is called a **fundamental parallelotope** of Λ with respect to the basis $\boldsymbol{a}_1, \dots, \boldsymbol{a}_N$. It is easy to see that this is an example of a fundamental domain for a lattice.

Exercise 3.8. *Prove that volume of a fundamental parallelotope is equal to the determinant of the 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. There is one more kind of closed fundamental domains which plays a central role in discrete geometry.

Definition 3.9. The **Voronoi cell** of a lattice Λ is the set

$$\mathcal{V} = \{\boldsymbol{x} \in \mathbb{R}^N : \|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x} - \boldsymbol{y}\|_2\ \forall\ \boldsymbol{y} \in \Lambda\}.$$

It is easy to see that $\mathcal{V}$ is a closed fundamental domain for Λ.

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 all the things we discussed here also have 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 $1 \leq r \leq N$, and let $\boldsymbol{a}_1, \dots, \boldsymbol{a}_r$ be a basis for it. Write $A = (\boldsymbol{a}_1 \ \dots \ \boldsymbol{a}_r)$ 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, UA 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 $M = AA^t$, so it is a square $r \times r$ non-singular matrix. Notice that if A and B are two basis matrices so that $B = UA$ for some U as above, then

$$\begin{aligned} \det(BB^t) &= \det((UA)(UA)^t) = \det(U(AA^t)U^t) \\ &= \det(U)^2 \det(AA^t) = \det(AA^t). \end{aligned}$$

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 3.10. 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

$$\det(\Lambda) = \sqrt{\det(AA^t)},$$

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 $\operatorname{span}_{\mathbb{R}}(\Lambda)$. The rest works essentially verbatim, keeping in mind that if $\Omega \subseteq \Lambda$ is a sublattice, then index $[\Lambda : \Omega]$ is only defined if $\operatorname{rk}(\Omega) = \operatorname{rk}(\Lambda)$.

4. Quadratic forms

In this section we outline the connection between lattices and positive definite quadratic forms. We start by defining quadratic forms and sketching some of their basic properties.

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 $X_1, \dots, X_N, Y_1, \dots, Y_N$ 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, $X_1, \dots, X_N$ and $Y_1, \dots, Y_N$. It is easy to see that a bilinear form $B(\boldsymbol{X}, \boldsymbol{Y})$ can also be written as

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

where

$$\mathcal{B} = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1N} \\ b_{12} & b_{22} & \dots & b_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ b_{1N} & b_{2N} & \dots & b_{NN} \end{pmatrix},$$

is the corresponding $N \times N$ symmetric coefficient matrix, 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

$$B(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{X}^t \mathcal{B} \boldsymbol{Y} = (\boldsymbol{X}^t \mathcal{B} \boldsymbol{Y})^t = \boldsymbol{Y}^t \mathcal{B}^t \boldsymbol{X} = \boldsymbol{Y}^t \mathcal{B} \boldsymbol{X} = B(\boldsymbol{Y}, \boldsymbol{X}), \tag{5}$$

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

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

Hence to each bilinear symmetric form in $2N$ variables there corresponds a quadratic form in N variables. The converse is also true.

Exercise 4.1. *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.

Definition 4.1. 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 4.1. *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 (5)

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

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

Remark 4.1. Since eigenvectors corresponding to real eigenvalues of a matrix must be real, Lemma 4.1 implies that a real symmetric matrix has all real eigenvectors as well. In fact, even more is true.

Lemma 4.2. *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 $\|\boldsymbol{x}_1\|_2 = 1$. 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 $U^{-1} = U^t$.

Exercise 4.2. *Prove that the matrix $U^t\mathcal{B}U$ 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.

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

$$U^t\mathcal{B}U\boldsymbol{y}'_i = \lambda_i \boldsymbol{y}'_i,$$

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

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

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

$$\|\boldsymbol{x}_i\|_2 = \left(U\begin{pmatrix} 0 \\ \boldsymbol{y}_i \end{pmatrix}\right)^t U\begin{pmatrix} 0 \\ \boldsymbol{y}_i \end{pmatrix} = (0, \boldsymbol{y}_i^t)U^tU\begin{pmatrix} 0 \\ \boldsymbol{y}_i \end{pmatrix} = \|\boldsymbol{y}_i\|_2 = 1,$$

meaning that $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N$ is precisely the basis we are looking for. □

Remark 4.2. An immediate implication of Lemma 4.2 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).

Definition 4.2. Let $GL_N(\mathbb{R})$ be the set of all invertible $N \times N$ matrices with real entries. We say that $GL_N(\mathbb{R})$ is a **matrix group under multiplication**, meaning that the following conditions hold:

(1) **Identity:** There exists the $N \times N$ identity matrix I_N in $GL_N(\mathbb{R})$, which has the property that $I_N A = AI_N = A$ for any $A \in GL_N(\mathbb{R})$,

(2) **Closure:** For every $A, B \in GL_N(\mathbb{R})$, $AB, BA \in GL_N(\mathbb{R})$,

(3) **Inverses:** For every $A \in GL_N(\mathbb{R})$, $A^{-1} \in GL_N(\mathbb{R})$.

$GL_N(\mathbb{R})$ is the called the $N \times N$ **real general linear group**. Any subset H of $GL_N(\mathbb{R})$ that satisfies the conditions (1)-(3) is called a **subgroup** of $GL_N(\mathbb{R})$.

Exercise 4.3. *Prove that conditions (1)-(3) in the definition above indeed hold for $GL_N(\mathbb{R})$.*

Definition 4.3. A matrix $U \in GL_N(\mathbb{R})$ is called **orthogonal** if $U^{-1} = U^t$, and the subset of all such matrices in $GL_N(\mathbb{R})$ is

$$O_N(\mathbb{R}) = \{U \in GL_N(\mathbb{R}) : U^{-1} = U^t\}.$$

Exercise 4.4. *Prove that $O_N(\mathbb{R})$ is a subgroup of $GL_N(\mathbb{R})$. It is called the $N \times N$* **real orthogonal group**.

Exercise 4.5. *Prove that a matrix U is in $O_N(\mathbb{R})$ if and only if its column vectors form an orthonormal basis for $\mathbb{R}^N$.*

Definition 4.4. Define $GL_N(\mathbb{Z})$ to be the set of all invertible $N \times N$ matrices with integer coordinates, whose inverses also have integral coordinates.

Exercise 4.6. *Prove that $GL_N(\mathbb{Z})$ is a subgroup of $GL_N(\mathbb{R})$, which consists of all matrices with integers coordinates whose determinant is equal to ± 1.*

Lemma 4.3. *Every real symmetric matrix $\mathcal{B}$ is diagonalizable by an orthogonal matrix, i.e. there exists a matrix $U \in O_N(\mathbb{R})$ such that $U^t \mathcal{B} U$ is a diagonal matrix.*

Proof. By Lemma 4.2, we can pick an orthonormal basis $\boldsymbol{u}_1, \dots, \boldsymbol{u}_N$ for $\mathbb{R}^N$ consisting of eigenvectors of $\mathcal{B}$. Then let

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

so by Exercise 4.5 the matrix U is orthogonal. Moreover, for each $1 \leq i \leq N$,

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

where λ_i is the corresponding eigenvalue, since

$$1 = \|\boldsymbol{u}_i\|_2^2 = \boldsymbol{u}_i^t \boldsymbol{u}_i.$$

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

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

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

Remark 4.3. Lemma 4.3 is often referred to as the Principal Axis Theorem. The statements of Lemmas 4.1, 4.2, and 4.3 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 4.3, asserting that any matrix is unitary-similar to an upper triangular matrix over an algebraically closed field, is usually called Schur's theorem.

What are the implications of these results for quadratic forms?

Definition 4.5. A nonsingular linear transformation $\sigma : \mathbb{R}^N \to \mathbb{R}^N$ is called an **isomorphism**. Notice that σ like this is always given by left-multiplication by an $N \times N$ non-singular matrix, and vice versa: left-multiplication by an $N \times N$ non-singular matrix with coefficients in $\mathbb{R}$ is always an isomorphism from $\mathbb{R}^N$ to $\mathbb{R}^N$. By abuse of notation, we will identify an isomorphism σ with its matrix, and hence we can say that the set of all possible isomorphisms of $\mathbb{R}^N$ with itself is precisely the group $GL_N(\mathbb{R})$.

Definition 4.6. Two real symmetric bilinear forms B_1 and B_2 in $2N$ variables are called **isometric** 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 isometric in this case, and the isomorphism σ is called an **isometry** of these bilinear (respectively, quadratic) forms.

Isometry is easily seen to be an equivalence relation on real symmetric bilinear (respectively quadratic) forms, so we can talk about **isometry classes** of real symmetric bilinear (respectively quadratic) forms.
Notice that it is possible to have an isometry from a bilinear form B to itself, which we will call an **autometry** of B. This is the case when an isomorphism $\sigma : \mathbb{R}^N \to \mathbb{R}^N$ is 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.

Exercise 4.7. *Prove that if σ is an autometry of a symmetric bilinear form B, then* $\det(\sigma) = \pm 1$. *Prove that the set of all autometries of a symmetric bilinear (respectively quadratic) is a group under matrix multiplication. Hence it must be a subgroup of $GL_N(\mathbb{R})$.*

Definition 4.7. 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 $b_1,\dots,b_N$ are precisely the diagonal entries of the matrix $\mathcal{B}$.

With this notation we readily obtain the following result.

Theorem 4.4. *Every real symmetric bilinear form, as well as its associated quadratic form, is isometric to a real diagonal form. In fact, there exists such an isometry whose matrix is in $O_N(\mathbb{R})$.*

Proof. This is an immediate consequence of Lemma 4.3. □

Remark 4.4. Notice that this diagonalization is not unique, i.e. it is possible for a bilinear or quadratic form to be isometric to more than one diagonal form (notice that an isometry can come from the whole group $GL_N(\mathbb{R})$, not necessarilly from $O_N(\mathbb{R})$). This procedure does however yield an invariant for nonsingular real quadratic forms, called signature.

Definition 4.8. A symmetric bilinear or quadratic form is called **nonsingular** (or **nondegenerate**, or **regular**) if its coefficient matrix is nonsingular.

Exercise 4.8. *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.

We now deal with nonsingular quadratic forms until further notice.

Definition 4.9. 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. Moreover, even 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 [18]; the proof is essentially the same.

Theorem 4.5. *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 $U, W \in O_N(\mathbb{R})$ be two different matrices that diagonalize $\mathcal{B}$ with colum vectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_N$ and $\boldsymbol{w}_1, \dots, \boldsymbol{w}_N$, respectively, arranged in such a way that

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

and

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

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

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

and

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

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. The following result, which we leave as an exercise, is due to Sylvester.

Exercise 4.9. *Prove that two nonsingular real quadratic forms in N variables are isometric if and only if they have the same signature.*

An immediate implication of Exercise 4.9 is that for each $N \geq 2$, there are precisely $N+1$ isometry classes of nonsingular real quadratic forms in N variables, and by Theorem 4.4 each of these classes contains a diagonal form. Some of these isometry classes are especially important for our purposes.

Definition 4.10. 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**.

Exercise 4.10. *Prove that 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.*

Positive definite real quadratic forms are also sometimes called **norm forms**. 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

$$\|\boldsymbol{y}\|_2 = (A\boldsymbol{x})^t(A\boldsymbol{x}) = \boldsymbol{x}^t(A^tA)\boldsymbol{x} = Q_A(\boldsymbol{x}),$$

where Q_A is the quadratic form whose symmetric coefficient matrix is A^tA. By construction, Q_A must be a positive definite form. This quadratic form is called a **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 $U \in GL_N(\mathbb{Z})$ such that $C = AU$. Hence the matrix of the quadratic form Q_C is $(AU)^t(AU) = U^t(A^tA)U$; we call two such matrices $GL_N(\mathbb{Z})$-**congruent**. Notice in this case that for each $\boldsymbol{x} \in \mathbb{R}^N$

$$Q_C(\boldsymbol{x}) = \boldsymbol{x}^tU^t(A^tA)U\boldsymbol{x} = Q_A(U\boldsymbol{x}),$$

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

Proposition 4.6. *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 coefficient matrix $\mathcal{B}$, then there exists $U \in O_N(\mathbb{R})$ such that

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

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 $\mathcal{D} = \sqrt{\mathcal{D}}^t \sqrt{\mathcal{D}}$, and so

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

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 $W \in O_N(\mathbb{R})$

$$(WA)^t(WA) = A^t(W^tW)A = A^tA = \mathcal{B}.$$

Therefore Q is a norm form for every lattice $WA\mathbb{Z}^N$, where $W \in O_N(\mathbb{R})$. Let us call two lattices Λ_1 and Λ_2 **isometric** if there exists $W \in O_N(\mathbb{R})$ 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 4.7. *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 4.7, then

$$\det(\Lambda) = \sqrt{|\det(Q)|}. \tag{6}$$

5. 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? Although our discussion will be mostly limited to the $\mathbf{0}$-symmetric convex sets, we start with a fairly general result; this is Theorem 2 on p. 42 of [15], the proof is the same.

Theorem 5.1 (Blichfeldt, 1914)**.** *Let M be a Jordan measurable set in $\mathbb{R}^N$. Suppose that $\operatorname{Vol}(M) > 1$, or that M is closed, bounded, and $\operatorname{Vol}(M) \geq 1$. 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 $\operatorname{Vol}(M) > 1$. Let us assume that M is bounded: if not, then there must exist a bounded subset $M_1 \subseteq M$ such that $\operatorname{Vol}(M_1) > 1$, so we can take M_1 instead of M. 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 $S = \{\boldsymbol{u}_1, \dots, \boldsymbol{u}_{r_0}\}$. 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 $M_1', \dots, M_{r_0}' \subseteq P$. 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

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

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

$$\sum_{r=1}^{r_0} \operatorname{Vol}(M_r') > 1,$$

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

$$\operatorname{Vol}\left(\bigcup_{r=1}^{r_0} M_r'\right) \leq \operatorname{Vol}(P) = 1.$$

Hence the sets $M_1', \dots, M_{r_0}'$ 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 $\operatorname{Vol}(M) = 1$. 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 5.1 we can prove the following version of **Minkowski's Convex Body Theorem**; recall here that our convex sets are always compact, i.e. closed and bounded. For this proof, we will need one additional fact, that we state here as an exercise.

Exercise 5.1. *Let S and T be two Jordan measurable sets in $\mathbb{R}^N$ such that*

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

where $A \in GL^N(\mathbb{R})$. Prove that

$$\operatorname{Vol}(T) = |\det(A)| \operatorname{Vol}(S).$$

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.

Theorem 5.2 (Minkowski)**.** *Let $M \subset \mathbb{R}^N$ be a convex $\mathbf{0}$-symmetric set with $\operatorname{Vol}(M) \geq 2^N$. 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 Exercise 5.1 its volume is

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

Thererfore, by Theorem 5.1, 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 5.1. 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.

We also briefly mention a generalization of Blichfeldt's theorem which was proved by van der Corput in 1936, using a method of Mordell; this is Theorem 1 on p. 47 of [15], and the proof (which we do not include here) uses a generalized Dirichlet's box principle.

Theorem 5.3. *Let $k \in \mathbb{Z}_{>0}$, and let $M \subseteq \mathbb{R}^N$ be a bounded Jordan measurable set with* $\operatorname{Vol}(M) > k$. *Then there exist at least $k+1$ distinct points $\boldsymbol{u}_1, \dots, \boldsymbol{u}_{k+1} \in M$ such that*

$$\boldsymbol{u}_i - \boldsymbol{u}_j \in \mathbb{Z}^N \ \forall \ 1 \leq i, j \leq k+1.$$

A generalized version of Minkowski's theorem follows as a corollary of Theorem 5.3, using the same type of argument as in the proof of Theorem 5.2, but now referring to Theorem 5.3 instead of Theorem 5.1; we skip the proof - it can be found for instance on p. 71 of [6].

Theorem 5.4. *Let $k \in \mathbb{Z}_{>0}$, and let $M \subset \mathbb{R}^N$ be a convex $\mathbf{0}$-symmetric set with* $\operatorname{Vol}(M) > 2^N k$. *Then there exists distinct nonzero points*

$$\pm\boldsymbol{x}_1, \dots, \pm\boldsymbol{x}_k \in M \cap \mathbb{Z}^N.$$

Exercise 5.2. *Prove versions of Theorems 5.1 - 5.2 where $\mathbb{Z}^N$ is replaced by a general 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 $A \in GL_N(\mathbb{R})$. 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 Exercise 5.1 to relate the volume of $A^{-1}M$ to the volume of M.*

From now on we will assume the versions of Blichfeldt and Minkowski theorems for arbitrary lattices, as in Exercise 5.2.

We will now discuss a couple applications of these results, following [15]. First we can prove **Minkowski's Linear Forms Theorem**; this is Theorem 3 on p. 43 of [15].

Theorem 5.5. *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 $\boldsymbol{b}_1, \dots, \boldsymbol{b}_N$ 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 $2c_1, \dots, 2c_N$ centered at the origin in $\mathbb{R}^N$. Then by Exercise 5.1

$$\mathrm{Vol}(P) = |\det(B)|^{-1} \mathrm{Vol}(R) = |\det(B)|^{-1} 2^N c_1 \dots c_N = 2^N,$$

and so by Theorem 5.2 there exists $\mathbf{0} \neq \boldsymbol{x} \in P \cap \mathbb{Z}^N$. □

Next application is to positive definite quadratic forms; this is Theorem 4 on p. 44 of [15]. Let

$$\omega_N = \begin{cases} \frac{\pi^k}{k!} & \text{if } N = 2k \text{ for some } k \in \mathbb{Z} \\ \frac{2^{2k+1} k! \pi^k}{(2k+1)!} & \text{if } N = 2k+1 \text{ for some } k \in \mathbb{Z} \end{cases} \tag{7}$$

be the volume of a unit ball in $\mathbb{R}^N$. Hence the volume of a ball of radius r in $\mathbb{R}^N$ is $\omega_N r^N$.

Theorem 5.6. *Let*

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

be a positive definite quadratic form in N variables with symmetric coefficient matrix B. There exists $\boldsymbol{0} \neq \boldsymbol{x} \in \mathbb{Z}^N$ such that

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

Proof. As at the end of section 4 (proof of Theorem 4.7), we can decompose B as $B = A^t A$ 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

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

where $S_r = \{\boldsymbol{y} \in \mathbb{R}^N : \|\boldsymbol{y}\|_2^2 \leq r\}$ 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, and by Exercise 5.1

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

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 5.2 there exists $\boldsymbol{0} \neq \boldsymbol{x} \in E_r \cap \mathbb{Z}^N$. □

6. Successive minima

Theorem 5.4 gives a criterion for a convex, $\mathbf{0}$-symmetric set to contain a collection of lattice points. This collections however is not guaranteed to be linearly independent. A natural next question to ask is, given a convex, $\mathbf{0}$-symmetric set M and a lattice Λ, under which conditions does M contain i linearly independent points of Λ for each $1 \leq i \leq N$? To answer this question is the main objective of this section. We start with some terminology.

Definition 6.1. Let M be a convex, $\mathbf{0}$-symmetric set $M \subset \mathbb{R}^N$ of non-zero volume and $\Lambda \subseteq \mathbb{R}^N$ a lattice of full rank. For each $1 \leq i \leq N$ define the i-th **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 i linearly independent points of Λ.

Remark 6.1. Notice that the N linearly independent vectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_N$ corresponding to successive minima $\lambda_1, \dots, \lambda_N$, respectively, do not necessarily form a basis. It was already known to Minkowski that they do in dimensions $N = 1, \dots, 4$, but when $N = 5$ there is a well known counterexample. Let

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

and let $M = B_5$, the closed unit ball centered at $\mathbf{0}$ in $\mathbb{R}^N$. Then the successive minima of B_5 with respect to Λ is

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

since $\boldsymbol{e}_1, \dots, \boldsymbol{e}_5 \in B_5 \cap \Lambda$, and

$$\boldsymbol{x} = \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2} \right)^t \notin B_5.$$

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

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

An immediate observation is that

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N.$$

Moreover, Minkowski's convex body theorem implies that

$$\lambda_1 \leq 2 \left(\frac{\det(\Lambda)}{\operatorname{Vol}(M)} \right)^{1/N}.$$

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 6.1. *With notation as above,*

$$\frac{2^N \det(\Lambda)}{N! \operatorname{Vol}(M)} \leq \lambda_1 \dots \lambda_N \leq \frac{2^N \det(\Lambda)}{\operatorname{Vol}(M)}.$$

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 [15], which is considerably easier than the upper bound. Let $\boldsymbol{u}_1, \dots, \boldsymbol{u}_N$ be the N linearly independent vectors corresponding to the respective successive minima $\lambda_1, \dots, \lambda_N$, and let

$$U = (\boldsymbol{u}_1 \dots \boldsymbol{u}_N) = \begin{pmatrix} u_{11} & \dots & u_{N1} \\ \vdots & \ddots & \vdots \\ u_{1N} & \dots & u_{NN} \end{pmatrix}.$$

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

$$\pm \frac{\boldsymbol{u}_1}{\lambda_1}, \dots, \pm \frac{\boldsymbol{u}_N}{\lambda_N}$$

lie in M, hence M contains the convex hull P of these points, which is a generalized octahedron. In fact,

$$P = \begin{pmatrix} \lambda_1^{-1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_N^{-1} \end{pmatrix} U K_N, \tag{8}$$

where

$$K_N = \left\{ \boldsymbol{x} \in \mathbb{R}^N : \sum_{i=1}^N |x_i| \leq 1 \right\}$$

is the n-dimensional **cross-polytope** with vertices $\pm \boldsymbol{e}_1, \dots, \pm \boldsymbol{e}_N$.

Exercise 6.1. *Prove that the cross-polytope K_N has volume $2^N/N!$.*

Combining Exercise 6.1 with (8), we deduce that the volume of P is
(9)

$$\mathrm{Vol}(P) = \frac{2^N}{N!}\left|\det\begin{pmatrix} \frac{u_{11}}{\lambda_1} & \dots & \frac{u_{N1}}{\lambda_N} \\ \vdots & \ddots & \vdots \\ \frac{u_{1N}}{\lambda_1} & \dots & \frac{u_{NN}}{\lambda_N} \end{pmatrix}\right| = \frac{2^N|\det(U)|}{N!\,\lambda_1\dots\lambda_N} \geq \frac{2^N}{N!\,\lambda_1\dots\lambda_N},$$

since $\det(U)$ is an integer. Since $P \subseteq M$, $\mathrm{Vol}(M) \geq \mathrm{Vol}(P)$. Combining this last observation with (9) yields the lower bound of the theorem.

Next we prove the upper bound. The argument we present is due to M. Henk [17], and is at least partially based on Minkowski's original geometric ideas. For each $1 \leq i \leq N$, let

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

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

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

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

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

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

$$\mathrm{Vol}(C_q^N + M_N) \leq (2q+\gamma)^N, \tag{10}$$

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

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

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

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

and so

$$\begin{aligned}(\boldsymbol{z}-\boldsymbol{x}) - (\boldsymbol{z}-\boldsymbol{y}) &= \boldsymbol{y}-\boldsymbol{x} \in \mathrm{int}(M_1) - \mathrm{int}(M_1) \\ &= \{\boldsymbol{z}_1 - \boldsymbol{z}_2 : \boldsymbol{z}_1, \boldsymbol{z}_2 \in M_1\} = \mathrm{int}(\lambda_1 M), \end{aligned} \tag{11}$$

which would contradict minimality of λ_1. Therefore

$$\mathrm{Vol}(C_q^N + M_1) = (2q+1)^N \,\mathrm{Vol}(M_1) = (2q+1)^N \left(\frac{\lambda_1}{2}\right)^N \mathrm{Vol}(M). \tag{12}$$

To finish the proof, we need the following lemma.

Lemma 6.2. *For each* $1 \leq i \leq N-1$,

$$\mathrm{Vol}(C_q^N + M_{i+1}) \geq \left(\frac{\lambda_{i+1}}{\lambda_i}\right)^{N-i} \mathrm{Vol}(C_q^N + M_i). \tag{13}$$

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

$$(\boldsymbol{x} + \mathrm{int}(M_{i+1})) \cap (\boldsymbol{y} + \mathrm{int}(M_{i+1})) = \emptyset. \tag{14}$$

Indeed, suppose (14) is not true, i.e. there exists $\boldsymbol{z} \in (\boldsymbol{x} + \mathrm{int}(M_{i+1})) \cap (\boldsymbol{y} + \mathrm{int}(M_{i+1}))$. Then, as in (11) above, $\boldsymbol{x} - \boldsymbol{y} \in \mathrm{int}(\lambda_{i+1} M)$. But we also have

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

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

$$u_{jk} = 0 \;\forall\; 1 \leq j \leq i,\; i+1 \leq k \leq N.$$

On the other hand, at least one of

$$x_k - y_k,\; i+1 \leq k \leq N,$$

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 (14). Notice that (14) implies

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

and

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

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

$$\mathrm{Vol}(C_q^i + M_{i+1}) \geq \left(\frac{\lambda_{i+1}}{\lambda_i}\right)^{N-i} \mathrm{Vol}(C_q^i + M_i). \tag{15}$$

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 (15) follows from the observation that

$$\mathrm{Vol}(C_q^i + f_1(M_i)) \geq \mathrm{Vol}(C_q^i + M_i). \tag{16}$$

This completes the proof of the lemma. □

Now, combining (10), (12), and (13), we obtain:

$$\begin{aligned}
(2q+\gamma)^N &\geq \mathrm{Vol}(C_q^N + M_N) \geq \left(\frac{\lambda_N}{\lambda_{N-1}}\right) \mathrm{Vol}(C_q^N + M_{N-1}) \geq \dots \\
&\geq \left(\frac{\lambda_N}{\lambda_{N-1}}\right)\left(\frac{\lambda_{N-1}}{\lambda_{N-2}}\right)^2 \dots \left(\frac{\lambda_2}{\lambda_1}\right)^{N-1} \mathrm{Vol}(C_q^N + M_1) \\
&= \lambda_N \dots \lambda_1 \frac{\mathrm{Vol}(M)}{2^N}(2q+1)^N,
\end{aligned}$$

hence

$$\lambda_1 \dots \lambda_N \leq \frac{2^N}{\mathrm{Vol}(M)}\left(\frac{2q+\gamma}{2q+1}\right)^N \to \frac{2^N}{\mathrm{Vol}(M)},$$

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

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 B_N in $\mathbb{R}^N$ centered at $\mathbf{0}$. We define the **successive minima of** Λ to be the successive minima of B_N with respect to Λ. Notice that successive minima are invariants of the lattice.

7. 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 non-zero volume. Throughout this section, we let

$$\lambda_1 \leq \dots \leq \lambda_N$$

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 [19].

Lemma 7.1.

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

Proof. Let F be the distance function corresponding to M, i.e. F is such that

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

Recall from Theorem 2.1 that such F exists, since M is a convex $\mathbf{0}$-symmetric set, hence a bounded star body. In fact, F can be defined by

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

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

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 $\boldsymbol{u}_1, \dots, \boldsymbol{u}_N$ be the linearly independent vectors corresponding to successive minima $\lambda_1, \dots, \lambda_N$, respectively. Then

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

Since $\boldsymbol{u}_1, \dots, \boldsymbol{u}_N$ form a basis for $\mathbb{R}^N$, there exist $a_1, \dots, a_N \in \mathbb{R}$ such that

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

We can also choose integer $v_1, \dots, v_N$ 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

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

by the definition of a distance function and Exercise 2.7. This completes the proof. □

Using Lemma 7.1 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 7.2. *If* $\lambda_1 \geq 1$, *then*

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

Proof. Since

$$1 \leq \lambda_1 \leq \dots \leq \lambda_N,$$

Theorem 6.1 implies

$$\lambda_N \leq \lambda_1 \dots \lambda_N \leq \frac{2^N \det(\Lambda)}{\operatorname{Vol}(M)},$$

and by Lemma 7.1,

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

The result follows by combining these two inequalities. □

A general bound depending also on λ_1 was obtained by Scherk [25], once again using Minkowski's successive minima theorem (Theorem 6.1) and Jarnik's inequality (Lemma 7.1) He observed that if λ_1 is fixed and $\lambda_2, \dots, \lambda_N$ are subject to the conditions

$$\lambda_1 \leq \dots \leq \lambda_N,\quad \lambda_1 \dots \lambda_N \leq \frac{2^N \det(\Lambda)}{\operatorname{Vol}(M)},$$

then the maximum of the sum

$$\lambda_1 + \dots + \lambda_N$$

is attained when

$$\lambda_1 = \lambda_2 = \cdots = \lambda_{N-1},\ \lambda_N = \frac{2^N \det(\Lambda)}{\lambda_1^{N-1} \operatorname{Vol}(M)}.$$

Hence we obtain Scherk's inequality for μ.

Corollary 7.3.

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

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

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

hence

$$\mu \geq \left(\frac{\det(\Lambda)}{\operatorname{Vol}(M)} \right)^{1/N}. \tag{17}$$

In fact, by Theorem 6.1,

$$\left(\frac{\det(\Lambda)}{\operatorname{Vol}(M)} \right)^{1/N} \geq \frac{(\lambda_1 \dots \lambda_N)^{1/N}}{2} \geq \frac{\lambda_1}{2},$$

and combining this with (17), we obtain

$$\mu \geq \frac{\lambda_1}{2}. \tag{18}$$

Jarnik obtained a considerably better lower bound for μ in [19].

Lemma 7.4.

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

Proof. Let $\boldsymbol{u}_1, \dots, \boldsymbol{u}_N$ be the linearly independent points of Λ corresponding to the successive minima $\lambda_1, \dots, \lambda_N$ of M with respect to Λ. Let F be the distance function of M, then

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

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

$$F\left(\boldsymbol{x} - \frac{1}{2}\boldsymbol{u}_N \right) \geq \frac{1}{2}\lambda_N. \tag{19}$$

Suppose not, then there exists some $\boldsymbol{x} \in \Lambda$ such that $F\left(\boldsymbol{x} - \frac{1}{2}\boldsymbol{u}_N\right) < \frac{1}{2}\lambda_N$, and so, by Exercise 2.7

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

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,

$$\boldsymbol{x}, \boldsymbol{u}_N - \boldsymbol{x} \in \operatorname{span}_{\mathbb{R}}\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_{N-1}\},$$

and so $\boldsymbol{u}_N = \boldsymbol{x} + (\boldsymbol{u}_N - \boldsymbol{x}) \in \operatorname{span}_{\mathbb{R}}\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_{N-1}\}$, which is a contradiction. Hence we proved (19) for all $\boldsymbol{x} \in \Lambda$.

Exercise 7.1. *Prove that*

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

Then lemma follows by combining (19) with Exercise 7.1. □

We define the **inhomogeneous minimum of** Λ to be the inhomogeneous minimum of the closed unit ball B_N with respect to Λ, since it will occur quite often. This is another invariant of the lattice.

8. Sphere packings and coverings

In this section we will very briefly discuss the two very old and famous problems that are closely related to the techniques in the geometry of numbers that we have so far developed, namely sphere packing and sphere covering. An excellent comprehensive, although slightly outdated, reference on this subject is the celebrated book by Conway and Sloane [7]. Throughout this section $N \geq 2$, since packing and covering problems in dimension $N = 1$ are clearly trivial.

Throughout this section by a sphere in $\mathbb{R}^N$ we will really 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 $B_i \subsetneq B_i'$, $B_j \subsetneq B_j'$. 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 8.1. 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 ω_N is the volume of a unit ball in $\mathbb{R}^N$, given by (7), and r is the **packing radius**, i.e. radius of each sphere in this lattice packing. It is easy to see that 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 non-zero 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**; we present a partial case of it without proof (see Theorem 1 on p. 200 of [15] for the general version with proof).

Theorem 8.1. *In each dimension N there exist lattice packings with density*

$$\Delta \geq \frac{\zeta(N)}{2^{N-1}}, \tag{20}$$

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

Ironically, all known proofs of Theorem 8.1 are non-constructive, so it is not generally known how to construct lattice packings with density as good as (20); in particular, in dimensions above 1000 the lattices whose existence is guaranteed by Theorem 8.1 are denser than all the presently known ones.

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 known are $N = 2, 3$. In case $N = 2$, Gauss has proved that the best possible lattice packing is given by the **hexagonal lattice**

$$\begin{pmatrix} 1 & \frac{1}{2} \\ 0 & \frac{\sqrt{3}}{2} \end{pmatrix} \mathbb{Z}^2, \tag{21}$$

and in 1940 L. Fejes Tóth proved that this indeed is the optimal packing. 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. It seems now that the famous Kepler's conjecture has been settled by Thomas Hales in 1998. Theoretical part of this proof is published only in 2005 [16], 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)). Best lattice packings are known in dimensions $N \leq 8$, however optimal packing is not known in any dimension $N > 3$. 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 8.2. 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 (7), 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 7.1 and 7.4 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 (21), 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

$$\lambda_1 = \dots = \lambda_N.$$

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 7.4 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.

9. Lattice packings in dimension 2

In this section we will prove that best lattice packing in $\mathbb{R}^2$ is achieved by the hexagonal lattice. First we show that our consideration can be reduced to well-rounded lattices.

Lemma 9.1. *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. Recall that in $\mathbb{R}^N$ for all $N < 5$ the vectors corresponding to successive minima in a lattice form a basis (we will call it a **minimal basis** for the lattice), hence $\boldsymbol{x}_1, \boldsymbol{x}_2$ and $\boldsymbol{y}_1, \boldsymbol{y}_2$ are bases for Λ and Ω, respectively. Notice that

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

Then:

$$\begin{aligned} \Delta(\Lambda) &= \frac{\lambda_1(\Lambda)^2 \omega_2}{4\det(\Lambda)} = \frac{\lambda_1(\Lambda)^2 \pi}{4\|\boldsymbol{x}_1\|_2 \|\boldsymbol{x}_2\|_2 \sin\theta} = \frac{\pi}{4\sin\theta} \\ &\geq \frac{\lambda_1(\Omega)^2 \pi}{4\|\boldsymbol{y}_1\|_2 \|\boldsymbol{y}_2\|_2 \sin\theta} = \frac{\lambda_1(\Omega)^2 \pi}{4\det(\Omega)} = \Delta(\Omega), \qquad (22) \end{aligned}$$

where $\omega_2 = \pi$ is the area of a unit circle in $\mathbb{R}^2$, as usual. This completes the proof. □

Notice that if $\{\boldsymbol{x}, \boldsymbol{y}\}$ is a minimal basis for a lattice Λ, then so are $\{-\boldsymbol{x}, \boldsymbol{y}\}$, $\{\boldsymbol{x}, -\boldsymbol{y}\}$, $\{-\boldsymbol{x}, -\boldsymbol{y}\}$. Out of these, let us agree to always pick the one with both vectors lying in the first quadrant, so that the angle θ between the vectors is in the interval $[0, \pi/2]$.

Lemma 9.2. *Let $\Lambda \subset \mathbb{R}^2$ be a lattice of full rank with successive minima $\lambda_1 \leq \lambda_2$, and let $\boldsymbol{x}, \boldsymbol{y}$ be the basis vectors corresponding to λ_1, λ_2, respectively. Let $\theta \in [0, \pi/2]$ be the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$. Then*

$$\pi/3 \leq \theta \leq \pi/2.$$

Proof. Notice that $\boldsymbol{x}^t\boldsymbol{y} > 0$, since both vectors are in the first quadrant. Assume that $\theta < \pi/3$, then

$$\frac{1}{2} < \cos\theta = \frac{\boldsymbol{x}^t\boldsymbol{y}}{\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2} = \frac{\boldsymbol{x}^t\boldsymbol{y}}{\lambda_1^2\lambda_2^2},$$

and hence

$$\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 = (\boldsymbol{x}-\boldsymbol{y})^t(\boldsymbol{x}-\boldsymbol{y}) = \|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2 - 2\boldsymbol{x}^t\boldsymbol{y} < \lambda_2^2,$$

meaning that $\|\boldsymbol{x}-\boldsymbol{y}\|_2 < \lambda_2$, where $\boldsymbol{x}-\boldsymbol{y} \neq 0$, and $\boldsymbol{x}, \boldsymbol{x}-\boldsymbol{y}$ are linearly independent. But this contradicts the fact that λ_2 is the second successive minimum of Λ, hence we must have $\pi/3 \leq \theta \leq \pi/2$. This completes the proof. □

Lemma 9.3. *Let $\Lambda \subset \mathbb{R}^2$ be a lattice of full rank, and let $\boldsymbol{x}, \boldsymbol{y}$ be a basis for Λ such that*

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

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

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

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

If $ab \geq 0$, then clearly $\|\boldsymbol{z}\|_2^2 \geq \|\boldsymbol{x}\|_2^2$. Now suppose $ab < 0$, then again

$$\|\boldsymbol{z}\|_2^2 \geq (a^2+b^2-|ab|)\|\boldsymbol{x}\|_2^2 \geq \|\boldsymbol{x}\|_2^2,$$

since $\cos\theta \leq 1/2$. Therefore $\boldsymbol{x}, \boldsymbol{y}$ are shortest non-zero 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 9.4. *Let Ω 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*

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

is WR with successive minima equal to λ_1.

Proof. By Lemma 9.2, 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 9.3, Ω_{WR} is WR with successive minima equal to λ_1. □

Now combining Lemma 9.1 with Lemma 9.4 implies that the packing density of the WR lattice Ω_{WR} is no smaller than that of Ω. Therefore the maximum packing density among lattices in $\mathbb{R}^2$ must occur on 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 $\mathbb{R}^2$, (22) 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, this angle θ is in the first quadrant, it is also an invariant of the lattice, and we call it the **angle of** Λ, denoted by $\theta(\Lambda)$.

Theorem 9.5. *The largest lattice packing density in* $\mathbb{R}^2$ *is achieved by the hexagonal lattice, and this density is equal to* $\frac{\pi}{2\sqrt{3}} = 0.906899\dots$

Proof. Lemma 9.1 says that the largest lattice packing density in $\mathbb{R}^2$ is attained by some WR lattice Λ, and (22) implies that

$$\Delta(\Lambda) = \frac{\pi}{4\sin\theta(\Lambda)}, \tag{23}$$

meaning that the smaller is $\sin\theta(\Lambda)$ the larger is $\Delta(\Lambda)$. Lemma 9.2 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 9.3 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 (23). This completes the proof. □

Remark 9.1. In fact, the density of Theorem 9.5 is attained by *any* lattice Λ in $\mathbb{R}^2$ with $\theta(\Lambda) = \pi/3$. There are infinitely many such lattices, but all of them are **similar** to Λ_h in the sense that they can be obtained by rotation and dilation of Λ_h (i.e. they are all of the form $\alpha A \Lambda_h$, where $0 \neq \alpha \in \mathbb{R}$ and $A \in O_2(\mathbb{R})$).

10. Reduction theory

Throughout this section we let $M \subseteq \mathbb{R}^N$ be a $\mathbf{0}$-symmetric convex set of non-zero volume, and let $\Lambda \subseteq \mathbb{R}^N$ be a lattice of full rank, as before. In section 5 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.

As before, let us write F for the distance function which corresponds to M, i.e.

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

Recall that since M is a convex $\mathbf{0}$-symmetric set

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

Also write $\lambda_1, \dots, \lambda_N$ for the successive minima of M with respect to Λ.

Definition 10.1. A basis $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_N\}$ 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

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

In the frequently occurring case when M is the closed unit ball 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 non-zero vector in Λ.

From here on let $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_N\}$ 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 $M = B_N$, then $F = \| \ \|_2$. Write A for the corresponding basis matrix of Λ, i.e. $A = (\boldsymbol{v}_1 \dots \boldsymbol{v}_N)$, and so $\Lambda = A\mathbb{Z}^N$. Let Q be the corresponding positive definite quadratic form, i.e. for each $\boldsymbol{x} \in \mathbb{R}^N$

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

Then, as we noted before, $Q(\boldsymbol{x}) = \|A\boldsymbol{x}\|_2^2$. In particular, for each $1 \leq i \leq N$,

$$Q(\boldsymbol{e}_i) = \|\boldsymbol{v}_i\|_2^2.$$

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

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

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

$$Q(\boldsymbol{e}_i) \leq Q(\boldsymbol{x}) \ \forall \ \boldsymbol{x} \in \mathbb{Z}^N, \ \gcd(x_i, \dots, x_N) = 1. \tag{24}$$

If a positive definite quadratic form satisfies (24), we will say that it is **Minkowski reduced**.

Exercise 10.1. *Prove that every positive definite quadratic form is arithmetically equivalent to a Minkowski reduced form.*

Exercise 10.2. *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*

$$0 < b_{11} \leq b_{22} \leq \dots \leq b_{NN},$$

and

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

Now let us drop the assumption that $M = B_N$, but preserve the rest of notation as above. We can prove the following analogue of Minkowski's successive minima theorem; this is essentially Theorem 2 on p. 66 of [15], which is due to Minkowski, Mahler, and Weyl.

Theorem 10.1. *Let $\nu_1 = 1$, and $\nu_i = \left(\frac{3}{2}\right)^{i-2}$ for each $2 \leq i \leq N$. Then*

$$\lambda_i \leq F(\boldsymbol{v}_i) \leq \nu_i \lambda_i. \tag{25}$$

Moreover,

$$\prod_{i=1}^N F(\boldsymbol{v}_i) \leq 2^N \left(\frac{3}{2}\right)^{\frac{(N-1)(N-2)}{2}} \frac{\det(\Lambda)}{\operatorname{Vol}(M)}. \tag{26}$$

Proof. It is easy to see that (26) follows immediately by combining (25) with Theorem 6.1, hence we only need to prove (25). We will only prove (25) in case $\Lambda = \mathbb{Z}^N$, leaving the general case as an exercise for the reader.

It is obvious by definition of 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 $2 \leq i \leq N$

$$F(\boldsymbol{v}_i) \leq \nu_i \lambda_i. \tag{27}$$

Let $\boldsymbol{u}_1, \dots, \boldsymbol{u}_N$ be the linearly independent vectors corresponding to successive minima $\lambda_1, \dots, \lambda_N$, 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 $\boldsymbol{u}_1, \dots, \boldsymbol{u}_i$ does not belong to the subspace $\operatorname{span}_{\mathbb{R}}\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_{i-1}\}$, call this vector $\boldsymbol{u}_j$. If the set $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{i-1}, \boldsymbol{u}_j$ is extendable to a basis of $\mathbb{Z}^N$, 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 (27) in this case.

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

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

where $k_1, \dots, k_{i-1}, m \in \mathbb{Z}$, and $m \geq 2$. Indeed, $m \neq 0$ since $\boldsymbol{u}_j \notin \operatorname{span}_{\mathbb{R}}\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_{i-1}\}$; on the other hand, if $m = 1$ then

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

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

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

where $\alpha_1, \dots, \alpha_{i-1} \in \mathbb{R}$. 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 $\boldsymbol{v} - \boldsymbol{v}' \in \operatorname{span}_{\mathbb{Z}}\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_{i-1}\}$, it must be that $\boldsymbol{v}' \in \mathbb{Z}^N$, and the set $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{i-1}, \boldsymbol{v}'$ is extendable to a basis of $\mathbb{Z}^N$. 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

$$F(\boldsymbol{v}_i) \leq \max\left\{\lambda_i, \frac{1}{2}\left(\sum_{k=1}^{i-1} F(\boldsymbol{v}_k) + \lambda_i\right)\right\},\ \forall\ 2 \leq i \leq N. \tag{28}$$

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 (27) from (28). This finishes the proof. □

As a corollary of Theorem 10.1, we can easily deduce the following bound on the product of diagonal coefficients of reduced positive definite quadratic forms.

Exercise 10.3. *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. Then

$$\prod_{i=1}^{N} b_{ii} \leq \frac{4^N}{\omega_N^2}\left(\frac{3}{2}\right)^{\frac{(N-1)(N-2)}{2}} \det(Q), \tag{29}$$

where ω_N is the volume of a unit ball in $\mathbb{R}^N$, which is given by (7).

(Hint: let $\Lambda = \mathbb{Z}^N$, and let M be the convex body corresponding to the distance function $F = \sqrt{Q}$; apply Theorem 10.1.)

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 non-zero 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 a famous problem in theoretical computer science, called the *shortest vector problem.* We briefly discuss this problem in the next section.

11. Shortest vector problem and computational complexity

Let $\Lambda \subset \mathbb{R}^N$ be a lattice of full rank, and let

$$B_N = \left\{ \boldsymbol{x} \in \mathbb{R}^N : \|\boldsymbol{x}\|_2 \leq 1 \right\}$$

be a closed unit ball in $\mathbb{R}^N$ centered at the origin, as usual. Let λ_1 be the first successive minimum of Λ with respect to B_N. Then

$$\lambda_1 = \inf \left\{ \lambda \in \mathbb{R}_{>0} : \lambda B_N \cap \Lambda \neq \{\boldsymbol{0}\} \right\},$$

and so there exists a vector $\boldsymbol{0} \neq \boldsymbol{w} \in \lambda_1 B_N \cap \Lambda$, meaning that

$$\|\boldsymbol{w}\|_2 = \lambda_1 = \min \left\{ \|\boldsymbol{x}\|_2 : \boldsymbol{x} \in \Lambda \setminus \{\boldsymbol{0}\} \right\}.$$

Such a vector $\boldsymbol{w}$ is called a **shortest vector** in Λ. The famous **shortest vector problem (SVP)** asks for an algorithm that allows to find a shortest vector in a given lattice Λ. This problem has been studied by Gauss, Dirichlet, Hermite, Minkowski, and many other mathematicians. As we discussed above, if $\boldsymbol{v}_1, \dots, \boldsymbol{v}_N$ is a Minkowski reduced basis for Λ with respect to B_N, then $\boldsymbol{v}_1$ is a shortest vector in Λ. But the question is how do you actually find it? The problem with Minkowski's reduction algorithm is that it is hard to implement. Let us explain what we mean by this. For this we will need to briefly introduce the notion of computational complexity, an important concept in theoretical computer science.

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. Elementary operations on a Turing include reading a symbol and writing a symbol, along with fast-forward and rewind, and correspond to elementary operations on a computer. We will say that a given problem can be solved in **polynomial time** on a Turing machine if the number of elementary operations required to solve the problem on a computer is bounded from above by a fixed polynomial function in the size of the input. The class of all polynomial-time problems is denoted by P. This is our first example of a **computational complexity class**.

For some problems we may not know whether it is possible to solve them on a computer 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 ≠ 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 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.

The shortest vector problem is known to be NP-complete. In particular, this means that it is not known how to implement Minkowski reduction to work in polynomial time on a Turing machine, i.e. on a modern computer. However, for practical applications, it is often sufficient to produce a close enough approximation to such shortest vector. The most famous such approximation algorithm is LLL, which stands for Lenstra, Lenstra, Lovasz. LLL is a polynomial time reduction algorithm that, given a lattice Λ, produces a basis $\boldsymbol{b}_1, \dots, \boldsymbol{b}_N$ for Λ such that

$$\min_{1 \leq i \leq N} \|\boldsymbol{b}_i\|_2 \leq 2^{N-1} \|\boldsymbol{w}\|,$$

where $\boldsymbol{w} \in \Lambda$ is a shortest non-zero vector. Some good references on this subject are [21], [15], [2], and [23].

There are many known examples of NP-complete problems (over 3000, it seems). Another famous example of an NP-complete problem with discrete geometry interpretation to it is the Coin Exchange Problem of Frobenius (see [1] for a detailed account, and [3], [14] for a more geometric interpretation).

12. Siegel's lemma

In the discussion of the shortest vector problem we were concerned with a polynomial-time algorithm that would allow us to find the shortest nonzero vector in a lattice of full rank in $\mathbb{R}^N$. Such an algorithm is not currently known, and is not necessarily believed to exist. Here we discuss a different approach to a similar problem for certain lattices of not full rank. For the rest of this section $\Lambda \subset \mathbb{R}^N$ will be a lattice of rank $N-M$, $1 \leq M < N$. More specifically, let

$$A = \begin{pmatrix} a_{11} & \dots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{M1} & \dots & a_{MN} \end{pmatrix}$$

be an $M \times N$ matrix with integer entries and rank equal to M. Define

$$\Lambda = \{\boldsymbol{x} \in \mathbb{Z}^N : A\boldsymbol{x} = \boldsymbol{0}\}.$$

Exercise 12.1. *Prove that Λ is a lattice of rank $N-M$.*

We will say that Λ is the **null-lattice** of the matrix A. Suppose we want to find a shortest nonzero vector $\boldsymbol{x} \in \Lambda$. Here is one way to do it. Suppose that we can prove that there must exist a nonzero vector $\boldsymbol{x} \in \Lambda$ with

$$\|\boldsymbol{x}\|_2 \leq N|\boldsymbol{x}| \leq f(A), \tag{30}$$

where $|\boldsymbol{x}| = \max_{1 \leq i \leq N} |x_i|$ is the usual sup-norm of $\boldsymbol{x}$, and $f(A) = f(a_{11}, \dots, a_{MN})$ is some explicit function of the entries of A. Then for each vector $\boldsymbol{x} \in \mathbb{Z}^N$ with $\|\boldsymbol{x}\|_2 \leq f(A)$ we can check whether $\boldsymbol{x} \in \Lambda$, ordering them in the order of ascending norm, and hence finding a shortest nonzero vector in Λ; $f(A)$ like this is often called a **search bound** for solutions of the linear system $A\boldsymbol{x} = \boldsymbol{0}$. Therefore we are interested in proving the existence of a nonzero vector $\boldsymbol{x} \in \Lambda$ with explicitly bounded norm, as suggested by (30). An idea of this sort was first used by A. Thue in 1909 [29], but formally stated only in 1929 by C. L. Siegel [27]. Our presentation partially follows [26].

Theorem 12.1 (Siegel's Lemma)**.** *With notation as above, there exists $\boldsymbol{0} \neq \boldsymbol{x} \in \Lambda$ with*

$$|\boldsymbol{x}| < 2 + (N|A|)^{\frac{M}{N-M}}, \tag{31}$$

where $|A| = \max\{|a_{mn}| : 1 \leq m \leq M,\ 1 \leq n \leq N\}$.

Proof. Let $H \in \mathbb{Z}_{>0}$, and let

$$C_H^N = \{\boldsymbol{x} \in \mathbb{R}^N : |\boldsymbol{x}| \leq H\}$$

be the cube centered at the origin in $\mathbb{R}^N$ with sidelength $2H$. Then

$$|C_H^N \cap \mathbb{Z}^N| = (2H+1)^N.$$

Let $T_A : \mathbb{R}^N \to \mathbb{R}^M$ 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 $\boldsymbol{x} \in C_H^N$,

$$|T_A(\boldsymbol{x})| \leq N|A|H,$$

i.e. T_A maps C_H^N into $C_{N|A|H}^M \subseteq \mathbb{R}^M$, since $\mathrm{rk}(A) = M$. Now

$$|C_{N|A|H}^M \cap \mathbb{Z}^M| = (2N|A|H+1)^M.$$

Now let us choose H to be a positive integer satisfying

$$(N|A|)^{\frac{M}{N-M}} \leq 2H < (N|A|)^{\frac{M}{N-M}} + 2.$$

Then

$$\begin{aligned} |C_H^N \cap \mathbb{Z}^N| &= (2H+1)^N = (2H+1)^M (2H+1)^{N-M} \\ &\geq (2H+1)^M (N|A|)^M > (2N|A|H+1)^M \\ &= |C_{N|A|H}^M \cap \mathbb{Z}^M|. \end{aligned}$$

This means that T_A cannot be mapping $C_H^N \cap \mathbb{Z}^N$ into $C_{N|A|H}^M \cap \mathbb{Z}^M$ in a one-to-one manner. Hence, there must exist $\boldsymbol{x} \neq \boldsymbol{y} \in C_H^N \cap \mathbb{Z}^N$ 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,

$$|\boldsymbol{x} - \boldsymbol{y}| \leq |\boldsymbol{x}| + |\boldsymbol{y}| \leq 2H < (N|A|)^{\frac{M}{N-M}} + 1,$$

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 $\frac{M}{N-M}$ in the upper bound of (31) cannot be improved. To see this, let for instance $M = N-1$ 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 $A\boldsymbol{x} = \boldsymbol{0}$ must have $x_N = R^{N-1}x_1$. Therefore, if

$$\Lambda = \{\boldsymbol{x} \in \mathbb{Z}^N : A\boldsymbol{x} = \boldsymbol{0}\},$$

and $\mathbf{0} \neq \boldsymbol{x} \in \Lambda$, then

$$|\boldsymbol{x}| \geq R^{N-1} = |A|^{\frac{M}{N-M}}.$$

Siegel's Lemma-type results have been proved in a considerably more general settings by a number of authors, employing quite sophisticated machinery from number theory and arithmetic geometry. Most notably, see the celebrated papers of Bombieri and Vaaler [4] and of Roy and Thunder [24], as well as a very nice overview of this subject in [26]. For some of the more recent related results also see [10], [11], [12], [13]. The original motivation for Siegel's Lemma came from Diophantine approximation and transcendental number theory.

13. Lattice points in homogeneously expanding domains

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? In this section we will at least partially answer this question. We will be interested in the *asymptotic behavior* of the function

$$G(t) = G(t, M, \Lambda) = |tM \cap \Lambda|$$

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 $G(t)$.

Definition 13.1. 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

$$C^N = \{\boldsymbol{x} \in \mathbb{R}^N : 0 \leq x_i \leq 1 \ \forall\ 1 \leq i \leq N\}$$

be the closed unit cube.

Definition 13.2. We say that $S \subseteq \mathbb{R}^N$ 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 13.3. 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 [20].

Theorem 13.1. *Let $t \in \mathbb{R}_{>0}$, then*

$$G(t) = \frac{\operatorname{Vol}(M)}{\det(\Lambda)} t^N + O(t^{N-1}),$$

where the constant in O-notation depends on Λ, N, and Lipschitz constants.

Proof. Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ 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

$$\mathcal{F}_{\boldsymbol{x}} \subseteq \operatorname{int}(tM),$$

or

$$\mathcal{F}_{\boldsymbol{x}} \cap \partial(tM) \neq \emptyset.$$

Let

$$m(t) = \left| \{ \boldsymbol{x} \in \Lambda : \mathcal{F}_{\boldsymbol{x}} \subseteq \operatorname{int}(tM) \} \right|,$$
$$b(t) = \left| \{ \boldsymbol{x} \in \Lambda : \mathcal{F}_{\boldsymbol{x}} \cap \partial(tM) \neq \emptyset \} \right|.$$

Then clearly

$$m(t) \leq G(t) \leq m(t) + b(t).$$

Moreover, since $\operatorname{Vol}(\mathcal{F}) = \det(\Lambda)$

$$m(t) \det(\Lambda) \leq \operatorname{Vol}(tM) = t^N \operatorname{Vol}(M) \leq (m(t) + b(t)) \det(\Lambda),$$

hence

$$m(t) \leq \frac{\operatorname{Vol}(M)}{\det(\Lambda)} t^N \leq m(t) + b(t).$$

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 $1/[t]$, then we can represent C^{N-1} as a union of $[t]^{N-1}$ small cubes with sidelength $1/[t]$ each, call them $C_1, \dots, C_{[t]^{N-1}}$. For each such C_i, we have

$$\|\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y})\|_2 \leq \mathcal{C} \|\boldsymbol{x} - \boldsymbol{y}\|_2 \leq \frac{\mathcal{C}\sqrt{N-1}}{[t]},$$

for each $\boldsymbol{x}, \boldsymbol{y} \in C_i$, i.e. the image of each such C_i under φ has diameter at most $\frac{\mathcal{C}\sqrt{N-1}}{[t]}$. Hence image of each such C_i under the map $t\varphi$ has diameter at most

$$\mathcal{C}\sqrt{N-1}\,\frac{t}{[t]} \leq 2\,\mathcal{C}\sqrt{N-1}.$$

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 13.1 provides an asymptotic formula for $G(t)$, demonstrating a very important general principle, namely that as $t \to \infty$, $G(T)$ 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

$$\left| G(t) - \frac{\mathrm{Vol}(M)}{\det(\Lambda)}t^N \right| \;?$$

A large amount of work has been done in this direction (see for instance pp. 140 - 147 of [15] for an overview of results and bibliography). This subject essentially originated in a paper of Davenport [8], who used a principle of Lipschitz [22]; also see [30] for a nice overview of Davenport's result and its generalizations. We present here without proof a result of P. G. Spain [28], which is a refinement of Davenport's bound, and can be thought of as a continuation of Theorem 13.1.

Theorem 13.2. *Let the notation be as in Theorem 13.1, and let $\mathcal{C}$ be the maximal Lipschitz constant corresponding to parametrization of ∂M. Then for each $t \in \mathbb{R}_{>0}$,*

$$\left| G(t) - \frac{\mathrm{Vol}(M)}{\det(\Lambda)}t^N \right| \leq 2^N(\mathcal{C}t+1)^{N-1}.$$

14. Erhart polynomial

As in section 9, let $M \subseteq \mathbb{R}^N$ be closed, bounded, Jordan measurable with $\operatorname{Vol}(M) > 0$, and suppose that ∂M is Lipschitz parametrizable with maximal Lipschitz constant $\mathcal{C}$. Let $\Lambda \subseteq \mathbb{R}^N$ be a lattice of full rank, then from Theorems 13.1 and 13.2, we can conclude that

$$G(t, M, \Lambda) = |tM \cap \Lambda| \leq \frac{\operatorname{Vol}(M)}{\det(\Lambda)} t^N + \sum_{i=0}^{N-1} 2^N \mathcal{C}^i \binom{N-1}{i} t^i, \tag{32}$$

i.e. there is a polynomial bound on $G(t, M, \Lambda)$ with coefficients dependent on $\mathcal{C}$. Under which conditions is $G(t, M, \Lambda)$ equal to a polynomial? This is known to happen for a more special class of sets. Here is the simplest example of such a situation. Let $\Lambda = \mathbb{Z}^N$, and

$$M = \{\boldsymbol{x} \in \mathbb{R}^N : |\boldsymbol{x}| \leq 1\},$$

then ∂M is Lipschitz parametrizable by linear maps, so maximal Lipschitz constant is equal to 1. Clearly for each $t \in \mathbb{Z}_{>0}$

$$|tM \cap \Lambda| = (2t+1)^N = \sum_{i=0}^{N} 2^i \binom{N}{i} t^i, \tag{33}$$

which is similar to the upper bound of (32) in this case.

For the rest of this section, let $\mathcal{P} \subseteq \mathbb{R}^N$ be a convex polytope such that $\operatorname{Vol}(\mathcal{P}) > 0$, and vertices of $\mathcal{P}$ are points of $\mathbb{Z}^N$; we will say that $\mathcal{P}$ is a **lattice polytope**. Write

$$G(t\mathcal{P}) = \left| t\mathcal{P} \cap \mathbb{Z}^N \right|.$$

We want to understand the behaviour of $G(t\mathcal{P})$ for all $t \in \mathbb{Z}_{>0}$; specifically, we will prove a famous theorem of Erhart, which states that $G(t\mathcal{P})$ is a polynomial in t. Our presentation closely follows [9]. First we consider a special case of polytopes, namely simplices.

Lemma 14.1. *Let $\boldsymbol{a}_1, \dots, \boldsymbol{a}_N \in \mathbb{Z}^N$ be linearly independent, and define the simplex*

$$S = \operatorname{Co}(\boldsymbol{0}, \boldsymbol{a}_1, \dots, \boldsymbol{a}_N) = \left\{ \sum_{i=1}^N t_i \boldsymbol{a}_i : t_i \geq 0 \ \forall\ 1 \leq i \leq N,\ \sum_{i=1}^N t_i \leq 1 \right\}.$$

Then there exist $\beta_1, \dots, \beta_N \in \mathbb{Z}_{\geq 0}$ such that for every $t \in \mathbb{Z}_{>0}$, we have

$$G(tS) = \left| tS \cap \mathbb{Z}^N \right| = \binom{N+t}{N} + \sum_{i=1}^N \binom{N+t-i}{N} \beta_i.$$

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

$$\boldsymbol{y} = \boldsymbol{x} + \sum_{i=1}^{N} \alpha_i \boldsymbol{a}_i, \tag{34}$$

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 $H_0 = \{\boldsymbol{0}\}$.

First, let $\boldsymbol{x} = \boldsymbol{0}$ in (34). Then $\boldsymbol{y}$ as in (34) lies in H_j if and only if

$$\sum_{i=1}^{N} \alpha_i = j, \; 0 \leq \alpha_i \leq j \; \forall \; 1 \leq i \leq N. \tag{35}$$

We will prove now that there are precisely $\binom{N+j-1}{N-1}$ possibilities for $\alpha_1, \ldots, \alpha_N$ satisfying (35) 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 (35) is

$$\sum_{\alpha_N=0}^{j} \binom{N + (j - \alpha_N) - 2}{N - 2} = \sum_{i=0}^{j} \binom{N + i - 2}{N - 2}. \tag{36}$$

Then our claim follows by combining (36) with the result of the following excercise.

Exercise 14.1. *Prove that*

$$\sum_{i=0}^{j} \binom{N + i - 2}{N - 2} = \binom{N + j - 1}{N - 1}.$$

Now to find the number of points $\boldsymbol{y}$ as in (34) with $\boldsymbol{x} = \boldsymbol{0}$ on $\bigcup_{j=0}^t H_j$, we sum over j, using the result of Excercise 14.1 once again:

$$\sum_{j=0}^t \binom{N+j-1}{N-1} = \binom{N+t}{N}.$$

If $\boldsymbol{x}$ in (34) lies properly between H_0 and H_1, then the number of possible $\boldsymbol{y}$ as given by (34) 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 (34) 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 (34) 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 $\boldsymbol{a}_1, \dots, \boldsymbol{a}_N \in \mathbb{Z}^N$ be linearly independent, and let S be the simplex $\operatorname{Co}(\boldsymbol{0}, \boldsymbol{a}_1, \dots, \boldsymbol{a}_N)$, as in Lemma 14.1. Define the **pseudo-simplex** associated with S

$$S_0 = S \setminus (\operatorname{Co}(\boldsymbol{0}, \boldsymbol{a}_1, \dots, \boldsymbol{a}_{N-1}) \ \cup \ \dots \ \cup \ \operatorname{Co}(\boldsymbol{0}, \boldsymbol{a}_2, \dots, \boldsymbol{a}_N)).$$

Lemma 14.2. *$G(tS_0)$ 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 $F^{(1)}, \dots, F^{(s)}$ be proper faces of S which contain $\boldsymbol{0}$ and satisfy

$$0 < \dim(F^{(i)}) < N, \ \forall\ 1 \leq i \leq s.$$

Then

$$S \setminus S_0 = \{\boldsymbol{0}\} \ \cup F_0^{(1)} \ \cup \ \dots \ \cup \ F_0^{(s)}$$

is a disjoint union. By induction hypothesis,

$$G(t(S \setminus S_0)) = 1 + G(tF_0^{(1)}) + \dots + G(tF_0^{(s)})$$

is a polynomial in t. Hence, by Lemma 14.1,

$$G(tS_0) = G(tS) - G(t(S \setminus S_0)) = G(tS) - 1 - G(tF_0^{(1)}) - \dots - G(tF_0^{(s)})$$

is a polynomial in t. □

We are now ready to prove Erhart's theorem.

Theorem 14.3 (Erhart)**.** *Let $\mathcal{P}$ be a lattice polytope in $\mathbb{R}^N$. Then $G(t\mathcal{P})$ is a polynomial in $t \in \mathbb{Z}_{\geq 0}$.*

Proof. We can assume $\mathbf{0}$ to be a vertex of $\mathcal{P}$, since such translation would not change the number of integer lattice points. Notice that each $(N-1)$-dimensional face of $\mathcal{P}$ 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 $\mathcal{P}$ into a simplicial complex whose 0-cells are precisely the vertices of $\mathcal{P}$. Then $\mathcal{P}$ can be represented as a disjoint union

$$\mathcal{P} = \{\mathbf{0}\} \, \cup S_0^{(1)} \, \cup \, \ldots \, \cup \, S_0^{(r)},$$

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 14.2. □

$G(t\mathcal{P})$ as in Theorem 14.3 is called **Erhart polynomial** of $\mathcal{P}$. An excellent reference on Erhart polynomials, their many fascinating properties, and connections to other important mathematical objects is [3]. For a general lattice polytope $\mathcal{P}$ very little is known about the coefficients of its Erhart polynomial $G(t\mathcal{P})$. Let

$$G(t\mathcal{P}) = \sum_{i=0}^{N} c_i(\mathcal{P}) t^i,$$

then it is known that the leading coefficient $c_N(\mathcal{P})$ is equal to $\mathrm{Vol}(\mathcal{P})$, and $c_{N-1}(\mathcal{P})$ is $(N-1)$-dimensional volume of the boundary $\partial\mathcal{P}$, which is normalized by the determinants of the sublattices induced by the corresponding faces of $\mathcal{P}$. Also, $c_0(\mathcal{P})$ is the combinatorial **Euler characteristic** $\chi(\mathcal{P})$:

$$\chi(\mathcal{P}) = \sum_{i=0}^{N} (-1)^i (\text{number of } i - \text{dimensional faces of } \mathcal{P}).$$

The rest of the coefficients of $G(t\mathcal{P})$ are in general unknown, however there are known relations and identities that they satisfy; see [3] for further details.

Notice that (33) provides an explicit example of Erhart polynomial in the simple case of a cube. To conclude this section, we will give two more explicit examples of Erhart polynomial. The first one is for an open simplex, which is precisely the interior of the simplex S of Lemma

14.1 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 14.4. *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 $G(tS^\circ) = 0$ if $t \leq N$, and for every $t \in \mathbb{Z}_{>N}$,

$$G(tS^\circ) = \binom{t-1}{N}. \tag{37}$$

Proof. Let $t > N$, and notice that the simplex tS° can be mapped by an affine transformation to the simplex

$$tS_1^\circ = \left\{ \boldsymbol{x} \in \mathbb{R}^N : 0 < x_1 < \dots < x_k < t \right\}.$$

This transformation is volume-preserving and maps $\mathbb{Z}^N$ to itself. Integral points of tS_1° correspond to increasing sequences of integers $0 < y_1 < \dots < y_N < t$. 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 (37) can be thought of as a geometric interpretation of binomial coeddicients. The next example is related to the one in Proposition 14.4, but is more general.

Proposition 14.5 ([5]). *Let*

$$\mathcal{S}_N = \left\{ \boldsymbol{x} \in \mathbb{R}^N : \sum_{i=1}^N |x_i| \leq 1 \right\}.$$

Then for every $t \in \mathbb{Z}_{>0}$

$$G(t\mathcal{S}_N) = \sum_{i=0}^{\min\{t,N\}} 2^i \binom{N}{i} \binom{t}{i}. \tag{38}$$

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 ± signs, and $\binom{t}{i}$ choices of absolute values. Summing over all $0 \leq i \leq \min\{t, N\}$ completes the proof. □

Remark 14.1. A remarkable property of the polynomial in Proposition 14.5 is that the right hand side (38) is symmetric in t and N. This means that

$$|t\mathcal{S}_N \cap \mathbb{Z}^N| = |N\mathcal{S}_t \cap \mathbb{Z}^t|.$$

References

[1] J. L. Ramirez Alfonsin. *The Diophantine Frobenius Problem.* Oxford University Press, 2005.

[2] A. H. Banihashemi and A. K. Khandani. On the complexity of decoding lattices using the Korkin-Zolotarev reduced basis. *IEEE Trans. Inform. Theory*, 44(1):162–171, 1998.

[3] M. Beck and S. Robins. *Computing the Continuous Discretely. Integer-Point Enumeration in Polyhedra.* Springer-Verlag, 2006.

[4] E. Bombieri and J. D. Vaaler. On Siegel's lemma. *Invent. Math.*, 73(1):11–32, 1983.

[5] D. Bump, K. K. Choi, P. Kurlberg, and J. Vaaler. A local Riemann hypothesis, I. *Math. Z.*, 233(1):1–19, 2000.

[6] J. W. S. Cassels. *An Introduction to the Geometry of Numbers.* Springer-Verlag, 1959.

[7] J. H. Conway and N. J. A. Sloane. *Sphere Packings, Lattices, and Groups.* Springer-Verlag, 1988.

[8] H. Davenport. On a principle of Lipschitz. *J. London Math. Soc.*, 26:179–183, 1951.

[9] G. Ewald. *Combinatorial convexity and algebraic geometry.* Springer-Verlag, 1996.

[10] L. Fukshansky. Integral points of small height outside of a hypersurface. *Monatsh. Math.*, 147(1):25–41, 2006.

[11] L. Fukshansky. Siegel's lemma with additional conditions. *J. Number Theory*, 120(1):13–25, 2006.

[12] L. Fukshansky. Effective structure theorems for symplectic spaces via height. In *Quadratic forms – algebra, arithmetic, and geometry, Contemp. Math.*, volume 493, pages 117–129, Amer. Math. Soc., Providence, RI, 2009.

[13] L. Fukshansky. Algebraic points of small height missing a union of varieties. *J. Number Theory*, 130(10):2099–2118, 2010.

[14] L. Fukshansky and S. Robins. Frobenius problem and the covering radius of a lattice. *Discrete Comput. Geom.*, 37(3):471–483, 2007.

[15] P. M. Gruber and C. G. Lekkerkerker. *Geometry of Numbers.* North-Holland Publishing Co., 1987.

[16] T. Hales. A proof of the Kepler conjecture. *Ann. of Math. (2)*, 162(3):1065–1185, 2005.

[17] 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.

[18] B. Jacob. *Linear Algebra.* W.H. Freeman and Company, 1990.

[19] V. Jarnik. Zwei Bemerkungen zur Geometrie de Zahlen. *Věstnik Krălovské České Společnosit Nauk*, 1941.

[20] S. Lang. *Algebraic Number Theory.* Springer-Verlag, 1994.

[21] A. K. Lenstra, H. W. Lenstra, and L. Lovasz. Factoring polynomials with rational coefficients. *Math. Ann.*, 261:515–534, 1982.

[22] R. Lipschitz. Über die asymptotischen Gesetze von gewissen Gattungen zahlentheoretischer Funktionen. *Monatsber. der Berliner Academie*, pages 174–185, 1865.

[23] M. Pohst. On the computation of lattice vectors of minimal length, successive minima, and reduced bases with applications. *ACM SIGSAM Bulletin*, 15(1):37–44, 1981.

[24] D. Roy and J. L. Thunder. An absolute Siegel's lemma. *J. Reine Angew. Math.*, 476:1–26, 1996.

[25] P. Scherk. Convex bodies off center. *Archiv Math.*, 3:303, 1950.

[26] W. M. Schmidt. *Diophantine Approximations and Diophantine Equations.* Springer-Verlag, 1991.

[27] C. L. Siegel. Zur theorie der quadratischen formen. *Nachr. Akad. Wiss. Gttingen Math.-Phys. Kl. II*, pages 21–46, 1972.

[28] P. G. Spain. Lipschitz: a new version of old principle. *Bull. London Math. Soc.*, 27:565–566, 1995.

[29] A. Thue. Uber Annaherungswerte algebraischer Zahlen. *J. Reine Angew. Math.*, 135:284–305, 1909.

[30] J. L. Thunder. The number of solutions of bounded height to a system of linear equations. *J. Number Theory*, 43:228–250, 1993.

Department of Mathematics, Claremont McKenna College, 850 Columbia Avenue, Claremont, CA 91711

E-mail address: `lenny@cmc.edu`