

MATH 662 (SECTION 602), SPRING 2006

DIOPHANTINE APPROXIMATIONS AND GEOMETRY
OF NUMBERS

LECTURE NOTES

LENNY FUKSHANSKY

CONTENTS

Part 1. Geometry of Numbers	2
1. Norms, sets, and volumes	2
2. Lattices	8
3. Quadratic forms	17
4. Theorems of Blichfeldt and Minkowski	26
5. Successive minima	33
6. Inhomogeneous minimum	39
7. Sphere packings and coverings	43
8. Reduction theory	47
9. Lattice points in homogeneously expanding domains	51
10. Erhart polynomial	54
11. Siegel's lemma	60
Part 2. Diophantine Approximations	63
12. Dirichlet, Liouville, Roth	63
13. Absolute values	73
14. Heights	82
15. Lehmer's problem and Mahler's measure	89
16. Points of small height	94
References	100

Part 1. Geometry of Numbers

1. NORMS, SETS, AND VOLUMES

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

Definition 1.1. A function $\|\cdot\| : \mathbb{R}^N \rightarrow \mathbb{R}$ is called a **norm** if

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

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

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

for each $\mathbf{x} = (x_1, \dots, x_N) \in \mathbb{R}^N$. We also define the **sup-norm**, given by

$$\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq N} |x_i|.$$

Exercise 1.1. Prove that $\|\cdot\|_p$ for each $p \in \mathbb{Z}_{>0}$ and $\|\cdot\|_\infty$ are indeed norms on \mathbb{R}^N .

Unless stated otherwise, we will regard \mathbb{R}^N as a metric space, where the metric is given by the Euclidean norm $\|\cdot\|_2$; recall that for every two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$, Euclidean distance between them is given by

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

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

Definition 1.2. 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 \mathbf{x}, \mathbf{y} in this set $d(\mathbf{x}, \mathbf{y}) \leq M$.

For instance, the closed unit ball centered at the origin in \mathbb{R}^N

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

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

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

is not a compact set.

Definition 1.3. A compact subset $X \subseteq \mathbb{R}^N$ is called **convex** if whenever $\mathbf{x}, \mathbf{y} \in X$, then any point of the form

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

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

Exercise 1.2. Let $\|\cdot\|$ be a norm on \mathbb{R}^N , and let $C \in \mathbb{R}$ be a positive number. Define

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

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

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

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

It is easy to notice that whenever a convex set contains X , it must also contain $\text{Co}(X)$. Hence convex hull of a collection of points should be thought of as the *smallest* convex set containing all of them. Another name for the convex hull of a finite set of points is **convex polytope**.

Remark 1.1. Some sources allow convex sets to be infinite, and then define the notion of a convex hull for infinite sets as well, thus distinguishing between an arbitrary convex hull and a convex polytope. For us it will be the same thing.

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

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

for some $a_1, \dots, a_N, b \in \mathbb{R}$. Notice that each hyperplane divides \mathbb{R}^N into two halfspaces. More precisely, a closed **halfspace** \mathcal{H} in \mathbb{R}^N is a set of all $\mathbf{x} \in \mathbb{R}^N$ 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 1.3. Prove that each convex polytope in \mathbb{R}^N can be described as a bounded intersection of finitely many halfspaces, and vice versa.

Remark 1.2. Exercise 1.3 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 of Geometry of Numbers still work, as we will see later.

Definition 1.4. A set $X \subseteq \mathbb{R}^N$ is called a **ray set** if for every $\mathbf{x} \in X$, $t\mathbf{x} \in X$ for all $t \in [0, 1]$.

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

Definition 1.5. A set $X \subseteq \mathbb{R}^N$ is called a **star body** if for every $\mathbf{x} \in \mathbb{R}^N$ either $t\mathbf{x} \in X$ for all $t \in \mathbb{R}$, or there exists $t_0(\mathbf{x}) \in \mathbb{R}_{>0}$ such that $t\mathbf{x} \in X$ for all $t \in \mathbb{R}$ with $|t| \leq t_0(\mathbf{x})$, and $t\mathbf{x} \notin X$ for all $|t| > t_0(\mathbf{x})$.

Remark 1.3. We will also require all our star bodies to have boundary which is locally homeomorphic to \mathbb{R}^{N-1} . 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 1.6. A function $F : \mathbb{R}^N \rightarrow \mathbb{R}$ is called a **distance function** if

- (1) $F(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^N$,
- (2) F is continuous,
- (3) **Homogeneity:** $F(a\mathbf{x}) = aF(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^N$, $a \in \mathbb{R}_{\geq 0}$.

Notice that if $f(\mathbf{X}) \in \mathbb{R}[X_1, \dots, X_N]$ is a homogeneous polynomial of degree d with real coefficients, then

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

is a distance function.

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

Exercise 1.4. If F is a distance function on \mathbb{R}^N , prove that the set

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

is a bounded star body.

In fact, a converse is also true.

Theorem 1.1. Let X be a star body in \mathbb{R}^N . Then there exists a distance function F such that

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

Proof. Define F in the following way. For every $\mathbf{x} \in \mathbb{R}^N$ such that $t\mathbf{x} \in X$ for all $t \geq 0$, let $F(\mathbf{x}) = 0$. Suppose that $\mathbf{x} \in \mathbb{R}^N$ is such that there exists $t_0(\mathbf{x}) > 0$ with the property that $t\mathbf{x} \in X$ for all $t \leq t_0(\mathbf{x})$, and $t\mathbf{x} \notin X$ for all $t > t_0(\mathbf{x})$; for such \mathbf{x} define $F(\mathbf{x}) = \frac{1}{t_0(\mathbf{x})}$. 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]. \square

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

Definition 1.7. A subset $X \subseteq \mathbb{R}^N$ which contains $\mathbf{0}$ is called **0-symmetric** if whenever \mathbf{x} is in X , then so is $-\mathbf{x}$.

It is easy to see that every set $A_N(C)$ of Exercise 1.2, as well as every star body, is **0-symmetric**, although ray sets in general are not. In fact, star bodies are precisely the **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 $\mathbf{x}_1, \dots, \mathbf{x}_L \in \mathbb{R}^N$, i.e. $\text{Co}(\mathbf{0}, \mathbf{x}_1, \dots, \mathbf{x}_L)$.

Exercise 1.5. Let X be a star body, and let F be its distance function, i.e. $X = \{\mathbf{x} \in \mathbb{R}^N : F(\mathbf{x}) \leq 1\}$. Prove that

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

for all $\mathbf{x}, \mathbf{y} \in X$ 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 . In general of course volume of a set $X \subseteq \mathbb{R}^N$ is its Lebesgue measure. However, for all our purposes a smaller degree of generality will suffice.

Definition 1.8. **Characteristic function** of a set X is defined by

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

Definition 1.9. A bounded set X is said to have **Jordan volume** if its characteristic function is Riemann integrable, and then we define $\text{Vol}(X)$ to be the value of this integral.

Remark 1.4. A set that has Jordan volume is also called **Jordan measurable**.

Theorem 1.2. *All convex sets and bounded ray sets have Jordan volume.*

Proof. We will only prove this theorem for convex sets; for bounded ray sets the proof is similar. Let X be a convex set. We can assume that $\mathbf{0} \in X$; if not, we can just translate X so that it contains $\mathbf{0}$ - translation does not change measurability properties. Write ∂X for the boundary of X and 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 $\varphi : \partial X \rightarrow S_{N-1}$, given by

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

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

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

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

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

and

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

Let $\delta_1(\varepsilon)$ and $\delta_2(\varepsilon)$ be minimal positive real numbers such that the spheres centered at the origin of radii $\|\mathbf{y}_i\|_2$ and $\|\mathbf{z}_i\|_2$ are covered by caps of radii $\delta_1(\varepsilon)$ and $\delta_2(\varepsilon)$, $\mathcal{C}_{\mathbf{y}_i}(\varepsilon)$ and $\mathcal{C}_{\mathbf{z}_i}(\varepsilon)$, centered at \mathbf{y}_i and \mathbf{z}_i respectively. Define cones

$$C_i^1 = \text{Co}(\mathbf{0}, \mathcal{C}_{\mathbf{y}_i}(\varepsilon)), \quad C_i^2 = \text{Co}(\mathbf{0}, \mathcal{C}_{\mathbf{z}_i}(\varepsilon)),$$

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

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

and all the cones C_i^1, C_i^2 have Jordan volume, hence the same is true about their unions. Moreover,

$$\text{Vol} \left(\bigcup_{i=1}^{k(\varepsilon)} (C_i^1 \setminus C_i^2) \right) \rightarrow 0,$$

as $\varepsilon \rightarrow 0$. Hence X must have Jordan volume. \square

This is Theorem 5 on p. 9 of [17], and the proof is also very similar.

2. LATTICES

We start with an algebraic definition of lattices.

Definition 2.1. A **lattice** Λ of rank r , $1 \leq r \leq N$, in \mathbb{R}^N is a free \mathbb{Z} -module of rank r such that $\text{span}_{\mathbb{R}}(\Lambda)$ is a subspace of \mathbb{R}^N of dimension r .

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 \mathbb{Z}^N . In fact, every lattice Λ has a **basis**, i.e. a collection of linearly independent (over \mathbb{R}) vectors that span Λ over \mathbb{Z} . Then it is easy to see that a lattice Λ of rank $1 \leq r \leq N$ with a basis $\mathbf{a}_1, \dots, \mathbf{a}_r \in \mathbb{R}^N$ will always be of the form

$$\text{span}_{\mathbb{Z}}\{\mathbf{a}_1, \dots, \mathbf{a}_r\} = \bigoplus_{i=1}^r \mathbb{Z}\mathbf{a}_i.$$

Moreover, if $\mathbf{y} \in \Lambda$, then there exist $n_1, \dots, n_r \in \mathbb{Z}$ such that

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

where

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

and A is an $N \times r$ *basis matrix* for Λ of the form $A = (\mathbf{a}_1 \ \dots \ \mathbf{a}_r)$. 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 . Notice that bases are not unique; as we will see later, each lattice has bases with particularly nice properties.

There is an alternative description of lattices. Notice that \mathbb{R}^N is an abelian group under addition. Moreover, it is a **locally compact** group; this means that for every point $\mathbf{x} \in \mathbb{R}^N$ there exists an open set containing \mathbf{x} whose closure is compact, for instance take an open unit ball centered at \mathbf{x} . More generally, every subspace V of \mathbb{R}^N is also a locally compact abelian group. A subgroup Γ of V is called **discrete** if for each $\mathbf{x} \in \Gamma$ there exists an open set $S \subseteq V$ such that $S \cap \Gamma = \{\mathbf{x}\}$. For instance, \mathbb{Z}^N is a discrete subgroup of \mathbb{R}^N : for each point $\mathbf{x} \in \mathbb{Z}^N$ the open ball of radius $1/2$ centered at \mathbf{x} contains no other points of \mathbb{Z}^N . We say that a discrete subgroup Γ is **cocompact** in V if the quotient group V/Γ is compact. From now on we assume that all our discrete subgroups are cocompact.

Exercise 2.1. Let Λ be a lattice of rank r in \mathbb{R}^N . Then $V = \text{span}_{\mathbb{R}} \Lambda$ is an r -dimensional subspace of \mathbb{R}^N . Prove that Λ is a discrete subgroup of V .

In fact, the converse is also true; Exercise 2.1 and Theorem 2.1 are basic generalizations of Theorems 1 and 2 respectively on p. 18 of [17], the proofs are essentially the same; the idea behind this argument is quite important.

Theorem 2.1. Let V be an r -dimensional subspace of \mathbb{R}^N , and let Γ be a discrete 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 $\mathbf{a}_1, \dots, \mathbf{a}_r$ in Γ such that $\Gamma = \text{span}_{\mathbb{Z}}\{\mathbf{a}_1, \dots, \mathbf{a}_r\}$. We start by inductively constructing a collection of vectors $\mathbf{a}_1, \dots, \mathbf{a}_r$, and then show that it has the required properties.

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

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

and pick any $\mathbf{c} \in \Gamma \setminus H_{i-1}$. Let P_i be the closed parallelotope spanned by the vectors $\mathbf{a}_1, \dots, \mathbf{a}_{i-1}, \mathbf{c}$. Notice that since Γ is discrete in V , $\Gamma \cap P_i$ is a finite set. Moreover, since $\mathbf{c} \in P_i$, $\Gamma \cap P_i \not\subseteq H_{i-1}$. Then select \mathbf{a}_i such that

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

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

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

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

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

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

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

We will now show that

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

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

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

Notice that

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

and since Γ is a group, we must have

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

Then notice that

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

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

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

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

From now on, until further notice, our lattices will be of full rank in \mathbb{R}^N , that is of rank N .

Definition 2.2. An $N \times N$ non-singular matrix A is called a **basis matrix** for a lattice Λ in \mathbb{R}^N if its column vectors form a basis for Λ .

Theorem 2.2. Let Λ 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 \mathbf{b}_i of B can be expressed as

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

where $\mathbf{a}_1, \dots, \mathbf{a}_N$ 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 $1 \leq i \leq N$

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

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

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

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

Corollary 2.3. *If A and B are two basis matrices for the same lattice Λ , then*

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

Definition 2.3. The common determinant value of Corollary 2.3 is called the **determinant** of the lattice Λ , and is denoted by $\det(\Lambda)$.

Exercise 2.2. *Can you extend Theorem 2.2 and Corollaries 2.3 to lattices of rank $< N$ in \mathbb{R}^N ?*

We now talk about sublattices of a lattice. 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 . We will write $[\Lambda : \Omega]$ for the **index** of Ω in Λ as a subgroup. First we prove that a lattice always has a basis with “nice” properties; this is Theorem 1 on p. 11 of [6].

Then we clearly have

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

Therefore we must have $t_k - sv_{kk} \neq 0$ by minimality of k . But then (2) contradicts the minimality of $|v_{kk}|$: we could take $\mathbf{c} - s\mathbf{a}_k$ instead of \mathbf{a}_k , since it satisfies all the conditions that \mathbf{a}_k was chose to satisfy, and then $|v_{kk}|$ is replaced by the smaller nonzero number $|t_k - sv_{kk}|$. This proves that \mathbf{c} like this cannot exist, and so (1) is true, hence finishing one direction of the theorem.

Now suppose that we are given a basis $\mathbf{a}_1, \dots, \mathbf{a}_N$ for Ω . We want to prove that there exists a basis $\mathbf{b}_1, \dots, \mathbf{b}_N$ 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 2.1. Indeed, at i -th step of the basis construction in the proof of Theorem 2.1, we can choose i -th vector, call it \mathbf{b}_i , so that it lies in the span of the previous $i - 1$ vectors and the vector \mathbf{a}_i . Since $\mathbf{b}_1, \dots, \mathbf{b}_N$ constructed this way are linearly independent (in fact, they form a basis for Λ by the construction), we obtain that

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

for each $1 \leq i \leq N$. This proves the second half of our theorem. \square

Exercise 2.4. *Prove that it possible to select the coefficients v_{ij} in Theorem 2.4 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 2.1. Let the notation be as in Theorem 2.4. 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 2.4 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 metrics of the same lattice are always related by multiplication by a unimodular integral matrix (i.e. element of $GL_N(\mathbb{Z})$), Theorem 2.4 can be thought of as the construction of Hermite normal form. Exercise 2.4 places additional restrictions that make Hermite normal form unique.

Here is an important implication of Theorem 2.4; this is Lemma 1 on p. 14 of [6].

Theorem 2.5. *Let $\Omega \subseteq \Lambda$ be a sublattice. Then $\det(\Lambda) \mid \det(\Omega)$; moreover,*

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

Proof. Let $\mathbf{b}_1, \dots, \mathbf{b}_N$ be a basis for Λ , and $\mathbf{a}_1, \dots, \mathbf{a}_N$ be a basis for Ω , so that these two bases satisfy the conditions of Theorem 2.4, and write A and B for the corresponding basis matrices. Then notice that

$$B = VA,$$

where $V = (v_{ij})_{1 \leq i, j \leq N}$ is an $N \times N$ triangular matrix with entries as described in Theorem 2.4; 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 $\mathbf{c} \in \Lambda$ is contained in the same coset of Ω in Λ as precisely one of the vectors

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

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

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

Definition 2.4. 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

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

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

Remark 2.2. Notice that for every point $\mathbf{x} \in \mathbb{R}^N$ there exists uniquely a point $\mathbf{y} \in \mathcal{F}$ such that

$$\mathbf{x} \equiv \mathbf{y} \pmod{\Lambda},$$

i.e. \mathbf{x} lies in the coset $\mathbf{y} + \Lambda$ 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 2.5. Let Λ be a lattice, and $\mathbf{a}_1, \dots, \mathbf{a}_N$ be a basis for Λ . Then the set

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

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

Exercise 2.5. Prove that volume of a fundamental parallelootope 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 parallelootope, and call them *closed* fundamental domains. There is one more kind of closed fundamental domains which plays a central role in geometry of numbers.

Definition 2.6. The **Voronoi cell** of a lattice Λ is the set

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

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 $\mathbf{a}_1, \dots, \mathbf{a}_r$ be a basis for it. Write $A = (\mathbf{a}_1 \dots \mathbf{a}_r)$ for the corresponding $N \times r$ basis matrix of Λ , then A has rank r since its column vectors are linearly independent. For any $U \in GL_r(\mathbb{Z})$, UA is another basis matrix for Λ ; moreover, if B is any other basis matrix for Λ , there exists $U \in GL_r(\mathbb{Z})$ such that $B = AU$. In other words, $GL_r(\mathbb{Z})$ is the automorphism group of Λ . 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 $U \in GL_r(\mathbb{Z})$, 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 2.7. Let Λ be a lattice of rank $1 \leq r \leq N$ in \mathbb{R}^N , and let A be an $N \times r$ basis matrix for Λ . The **determinant** of Λ is defined to be

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

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

3. 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 their basic properties.

A **quadratic form** is a homogeneous polynomial of degree 2; for now, we will consider quadratic forms with real coefficients. More generally, we can talk about a **symmetric bilinear form**, that is a polynomial

$$B(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^N \sum_{j=1}^N b_{ij} X_i Y_j \in \mathbb{R}[X_1, \dots, X_N, Y_1, \dots, Y_N],$$

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

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

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

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix},$$

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

$$(3) \quad B(\mathbf{X}, \mathbf{Y}) = \mathbf{X}^t \mathcal{B} \mathbf{Y} = (\mathbf{X}^t \mathcal{B} \mathbf{Y})^t = \mathbf{Y}^t \mathcal{B}^t \mathbf{X} = \mathbf{Y}^t \mathcal{B} \mathbf{X} = B(\mathbf{Y}, \mathbf{X}),$$

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

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

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

Exercise 3.1. Let $Q(\mathbf{X})$ be a quadratic form in N variables. Prove that

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

is a symmetric bilinear form.

Definition 3.1. We define the **determinant** or **discriminant** of a symmetric bilinear form B and of its associated quadratic form Q to be the determinant of the coefficient matrix \mathcal{B} , and will denote it by $\det(B)$ or $\det(Q)$.

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

Lemma 3.1. 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 \mathbf{x} . Write $\bar{\lambda}$ for the complex conjugate of λ , and $\bar{\mathcal{B}}$ and $\bar{\mathbf{x}}$ for the matrix and vector correspondingly whose entries are complex conjugates of respective entries of \mathcal{B} and \mathbf{x} . Then $\mathcal{B}\mathbf{x} = \lambda\mathbf{x}$, and so

$$\mathcal{B}\bar{\mathbf{x}} = \bar{\mathcal{B}}\bar{\mathbf{x}} = \overline{\mathcal{B}\mathbf{x}} = \overline{\lambda\mathbf{x}} = \bar{\lambda}\bar{\mathbf{x}},$$

since \mathcal{B} is a real matrix, meaning that $\mathcal{B} = \bar{\mathcal{B}}$. Then, by (3)

$$\lambda(\mathbf{x}^t \bar{\mathbf{x}}) = (\lambda\mathbf{x})^t \bar{\mathbf{x}} = (\mathcal{B}\mathbf{x})^t \bar{\mathbf{x}} = \mathbf{x}^t \bar{\mathcal{B}}\bar{\mathbf{x}} = \mathbf{x}^t (\bar{\lambda}\bar{\mathbf{x}}) = \bar{\lambda}(\mathbf{x}^t \bar{\mathbf{x}}),$$

meaning that $\lambda = \bar{\lambda}$, since $\mathbf{x}^t \bar{\mathbf{x}} \neq 0$. Therefore $\lambda \in \mathbb{R}$. \square

Remark 3.1. Since eigenvectors corresponding to real eigenvalues of a matrix must be real, Lemma 3.1 implies that a real symmetric matrix has all real eigenvectors as well. In fact, even more is true.

Lemma 3.2. 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 \mathbf{x}_1 be an eigenvector of \mathcal{B} with the corresponding eigenvalue λ_1 . We can assume that $\|\mathbf{x}_1\|_2 = 1$. Use Gram-Schmidt orthogonalization process to extend \mathbf{x}_1 to an orthonormal basis for \mathbb{R}^N , and write U_1 for the corresponding basis matrix such that \mathbf{x}_1 is the first column. Then it is easy to notice that $U_1^{-1} = U_1^t$.

Exercise 3.2. Prove that the matrix $U^t \mathcal{B} U$ is of the form

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

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

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

is also symmetric.

Now we can apply induction hypothesis to the matrix A , thus obtaining an orthonormal basis for \mathbb{R}^{N-1} , consisting of eigenvectors of A , call them $\mathbf{y}_2, \dots, \mathbf{y}_N$. For each $2 \leq i \leq N$, define

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

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

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

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

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

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

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

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

Remark 3.2. An immediate implication of Lemma 3.2 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 3.2. A matrix $U \in GL_N(\mathbb{R})$ is called **orthogonal** if $U^{-1} = U^t$, and the subset of all such matrices in $GL_N(\mathbb{R})$

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

is easily seen to be a subgroup, and is called the **orthogonal group**.

Exercise 3.3. 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 .

Lemma 3.3. 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 3.2, we can pick an orthonormal basis $\mathbf{u}_1, \dots, \mathbf{u}_N$ for \mathbb{R}^N consisting of eigenvectors of \mathcal{B} . Then let

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

so by Exercise 3.3 the matrix U is orthogonal. Moreover, for each $1 \leq i \leq N$,

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

where λ_i is the corresponding eigenvalue, since

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

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

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

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

Remark 3.3. Lemma 3.3 is often referred to as the Principal Axis Theorem. The statements of Lemmas 3.1, 3.2, and 3.3 together are usually called the Spectral Theorem for symmetric matrices; it has many important applications in various areas of mathematics, especially in Functional Analysis, where it is usually interpreted as a statement about self-adjoint (or hermitian) linear operators. A more general version of Lemma 3.3, 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 3.3. Two real symmetric bilinear forms B_1 and B_2 in $2N$ variables are called **isometric** if there exists an isomorphism $\sigma : \mathbb{R}^N \rightarrow \mathbb{R}^N$ such that

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

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$. 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.

Remark 3.4. Isometry is easily seen to be an equivalence relation on real symmetric bilinear (respectively quadratic) forms, and the set of all isometries between two isometric real symmetric bilinear (respectively quadratic) forms is a group under function composition, which corresponds to respective matrix multiplication. We will slightly abuse notation by identifying isometries with their matrices; as long as we are operating on the whole of \mathbb{R}^N and not restricting to subspaces this causes no confusion.

Notice that it is possible to have an isometry from a bilinear form B to itself. This is the case when an isomorphism $\sigma : \mathbb{R}^N \rightarrow \mathbb{R}^N$ is such that $B(\sigma\mathbf{X}, \sigma\mathbf{Y}) = B(\mathbf{X}, \mathbf{Y})$, and so the same is true for the associated quadratic form Q . In this case it is easy to see that σ must come from $O_N(\mathbb{R})$; this means, in particular, that σ preserves $\det(B)$, which makes it well defined.

Definition 3.4. 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

$$B(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^N b_i X_i Y_i, \quad Q(\mathbf{X}) = \sum_{i=1}^N b_i X_i^2,$$

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 3.4. *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 3.3. □

Remark 3.5. 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 necessarily from $O_N(\mathbb{R})$). This procedure does however yield an invariant for nonsingular real quadratic forms, called signature.

Definition 3.5. A symmetric bilinear or quadratic form is called **nonsingular** (or **nondegenerate**, or **regular**) if its coefficient matrix is nonsingular.

Exercise 3.4. Let $B(\mathbf{X}, \mathbf{Y})$ be a symmetric bilinear form and $Q(\mathbf{X})$ its associated quadratic form. Prove that the following four conditions are equivalent:

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

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

- (4) Q is equivalent to a diagonal form with all coefficients nonzero.

Remark 3.6. Exercise 3.4 tells us, among other things, that the statement about a quadratic form being nonsingular is precisely equivalent to the statement about its variety being nonsingular. In other words, the notion of nonsingularity here coincides with the one in Algebraic Geometry.

We now deal with nonsingular quadratic forms until further notice.

Definition 3.6. A nonsingular diagonal quadratic form Q can be written as

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

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 [20]; the proof is essentially the same.

Theorem 3.5. *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 column vectors $\mathbf{u}_1, \dots, \mathbf{u}_N$ and $\mathbf{w}_1, \dots, \mathbf{w}_N$, respectively, arranged in such a way that

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

and

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

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

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

and

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

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

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

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. \square

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 3.5. *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 3.5 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 3.4 each of these classes contains a diagonal form. Some of these isometry classes are especially important for our purposes.

Definition 3.7. A quadratic form Q is called **positive** or **negative definite** if, respectively, $Q(\mathbf{x}) > 0$, or $Q(\mathbf{x}) < 0$ for each $\mathbf{0} \neq \mathbf{x} \in \mathbb{R}^N$; Q is called **positive** or **negative semi-definite** if, respectively, $Q(\mathbf{x}) \geq 0$, or $Q(\mathbf{x}) \leq 0$ for each $\mathbf{0} \neq \mathbf{x} \in \mathbb{R}^N$. Otherwise, Q is called **indefinite**.

Exercise 3.6. *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 $\mathbf{y} \in \Lambda$ if and only if $\mathbf{y} = A\mathbf{x}$ for some $\mathbf{x} \in \mathbb{Z}^N$. Notice that the Euclidean norm of \mathbf{y} in this case is

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

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 $\mathbf{x} \in \mathbb{R}^N$

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

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

Proposition 3.6. *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 $\Lambda = AZ^N$, 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 WAZ^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 3.7. *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 3.7, then

$$(4) \quad \det(\Lambda) = \sqrt{|\det(Q)|}.$$

4. THEOREMS OF BLICHFELDT AND MINKOWSKI

In this section we will discuss some of the famous theorems related to the following very classical problem in the geometry of numbers: given a set M and a lattice Λ in \mathbb{R}^N , how can we tell if M contains any points of Λ ? 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 [17], the proof is the same.

Theorem 4.1 (Blichfeldt, 1914). *Let M be a Jordan measurable set in \mathbb{R}^N . Suppose that $\text{Vol}(M) > 1$, or that M is closed, bounded, and $\text{Vol}(M) \geq 1$. Then there exist $\mathbf{x}, \mathbf{y} \in M$ such that $\mathbf{0} \neq \mathbf{x} - \mathbf{y} \in \mathbb{Z}^N$.*

Proof. First suppose that $\text{Vol}(M) > 1$. Let us assume that M is bounded: if not, then there must exist a bounded subset $M_1 \subseteq M$ such that $\text{Vol}(M_1) > 1$, so we can take M_1 instead of M . Let

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

and let

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

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

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

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 $M_r \subseteq P + \mathbf{u}_r$, $M_s \subseteq P + \mathbf{u}_s$, and $(P + \mathbf{u}_r) \cap (P + \mathbf{u}_s) = \emptyset$. This means that

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

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

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

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

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

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

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

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

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

By the argument above we know that for each r there exist

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

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

As a corollary of Theorem 4.1 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.

Theorem 4.2 (Minkowski). *Let $M \subset \mathbb{R}^N$ be a convex $\mathbf{0}$ -symmetric set with $\text{Vol}(M) \geq 2^N$. Then there exists $\mathbf{0} \neq \mathbf{x} \in M \cap \mathbb{Z}^N$.*

Proof. Notice that the set

$$\frac{1}{2}M = \left\{ \frac{1}{2}\mathbf{x} : \mathbf{x} \in M \right\}$$

is also convex, $\mathbf{0}$ -symmetric, and has volume $2^{-N} \text{Vol}(M) \geq 1$. Therefore, by Theorem 4.1, there exist $\frac{1}{2}\mathbf{x} \neq \frac{1}{2}\mathbf{y} \in \frac{1}{2}M$ such that

$$\frac{1}{2}\mathbf{x} - \frac{1}{2}\mathbf{y} \in \mathbb{Z}^N.$$

But, by symmetricity, since $\mathbf{y} \in M$, $-\mathbf{y} \in M$, and by convexity, since $\mathbf{x}, -\mathbf{y} \in M$,

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

This completes the proof. \square

Remark 4.1. This result is sharp: the open cube

$$C = \{\mathbf{x} \in \mathbb{R}^N : |\mathbf{x}| < 1\}$$

has volume equal to 2^N and contains no nonzero integer lattice points.

Next we consider 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 [17], and the proof is the same: this really is just a generalized Dirichlet's box principle.

Theorem 4.3. *Let $k \in \mathbb{Z}_{>0}$, and let $M \subseteq \mathbb{R}^N$ be a bounded Jordan measurable set with $\text{Vol}(M) > k$. Then there exist at least $k+1$ distinct points $\mathbf{u}_1, \dots, \mathbf{u}_{k+1} \in M$ such that*

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

Proof. For each $r \in \mathbb{Z}_{>0}$ let

$$N_r = \left| \left\{ \mathbf{x} \in M : \mathbf{x} = r^{-1}\mathbf{u} = \left(\frac{u_1}{r}, \dots, \frac{u_N}{r} \right), \mathbf{u} \in \mathbb{Z}^N \right\} \right|.$$

Then, as $r \rightarrow \infty$, $N_r \sim r^N \text{Vol}(M)$. In particular, for a sufficiently large r , $N_r > r^N k$. Notice that corresponding points \mathbf{u} with $r^{-1}\mathbf{u} \in M$ are distributed among at most r^N different residue classes mod r , hence there must exist a residue class mod r that contains at least $k+1$ different points $\mathbf{u}_1, \dots, \mathbf{u}_{k+1}$ with $r^{-1}\mathbf{u}_1, \dots, r^{-1}\mathbf{u}_{k+1} \in M$. Notice that for all $1 \leq i, j \leq k+1$

$$r^{-1}(\mathbf{u}_i - \mathbf{u}_j) \in \mathbb{Z}^N,$$

and this finishes the proof. \square

A generalized version of Minkowski's theorem follows as a corollary of Theorem 4.3, using the same type of argument as in the proof of Theorem 4.2, but now referring to Theorem 4.3 instead of Theorem 4.1; we skip the proof - it can be found for instance on p. 71 of [6].

Theorem 4.4. *Let $k \in \mathbb{Z}_{>0}$, and let $M \subset \mathbb{R}^N$ be a convex $\mathbf{0}$ -symmetric set with $\text{Vol}(M) > 2^N k$. Then there exists distinct nonzero points*

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

Exercise 4.1. *Prove versions of Theorems 4.1 - 4.4 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 $\mathbf{x} \in A^{-1}M \cap \mathbb{Z}^N$ if and only if $A\mathbf{x} \in M \cap \Lambda$. Now one can think of multiplication by A as a coordinate transformation with Jacobian equal to $\det(A)$ in order 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 4.1.

We will now discuss some applications of these results, following [17]. First we can prove **Minkowski's Linear Forms Theorem**; this is Theorem 3 on p. 43 of [17].

Theorem 4.5. Let $B = (b_{ij})_{1 \leq i, j \leq N} \in GL_N(\mathbb{R})$, and let

$$L_i(\mathbf{X}) = \sum_{j=1}^N b_{ij} X_j$$

be corresponding linear forms. Let $c_1, \dots, c_N \in \mathbb{R}_{>0}$ be such that

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

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

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

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

Proof. Consider parallelepiped

$$P = \{\mathbf{x} \in \mathbb{R}^N : |L_i(\mathbf{x})| \leq c_i \forall 1 \leq i \leq N\}.$$

Notice that

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

and so by Theorem 4.2 there exists $\mathbf{0} \neq \mathbf{x} \in P \cap \mathbb{Z}^N$. □

Next application is to positive definite quadratic forms; this is Theorem 4 on p. 44 of [17]. Let

$$(5) \quad \omega_N = \frac{\pi^{\frac{N}{2}}}{\Gamma\left(\frac{N+2}{2}\right)}$$

be the volume of a unit ball in \mathbb{R}^N .

Theorem 4.6. Let

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

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

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

Proof. Notice that for each $R \in \mathbb{R}_{>0}$ the set

$$E_R = \{\mathbf{x} \in \mathbb{R}^N : Q(\mathbf{x}) \leq R\}$$

is an ellipsoid centered at the origin with

$$\text{Vol}(E_R) = \omega_N \sqrt{\frac{R^N}{\det(B)}}.$$

Hence if

$$R = 4 \left(\frac{\det(B)}{\omega_N^2} \right)^{1/N},$$

then $\text{Vol}(E_R) = 2^N$, and so by Theorem 4.2 there exists $\mathbf{0} \neq \mathbf{x} \in E_R \cap \mathbb{Z}^N$. \square

Our next application is to produce a bound on the discriminant of a number field. For this we will briefly review some notation and basic facts. For a more in depth coverage of this material refer, for instance, to [25].

A **number field of degree** d is a finite field extension K of \mathbb{Q} with $[K : \mathbb{Q}] = d$. All such extensions are **primitive**, i.e. there exists $\beta \in K$ such that

$$K = \mathbb{Q}[\beta] = \mathbb{Q}(\beta).$$

In fact, dimension of $\mathbb{Q}[\beta]$ as a \mathbb{Q} -vector space is precisely d , and $1, \beta, \beta^2, \dots, \beta^{d-1}$ is a basis for K over \mathbb{Q} . All elements of K are **algebraic numbers**, i.e. for each $\alpha \in K$ there exists an irreducible polynomial $f_\alpha(x)$ with integer coefficients such that $f_\alpha(\alpha) = 0$. This polynomial is called the **minimal polynomial** of α . If $K = \mathbb{Q}(\beta)$, then $[K : \mathbb{Q}] = \deg(f_\beta)$. If $f_\alpha(x)$ is monic, i.e. has leading coefficient 1, then α is called an algebraic integer. Define \mathcal{O}_K to be the set of all algebraic integers in K , then \mathcal{O}_K is a ring. In fact, although \mathcal{O}_K is not necessarily a PID, it is always a **Dedekind domain**, i.e. prime factorization of elements in \mathcal{O}_K is not necessarily unique, but prime factorization of ideals is.

There are d embeddings of $K = \mathbb{Q}(\beta)$ into \mathbb{C} that fix \mathbb{Q} , sending β to one of its conjugates, i.e. to one of the roots of f_β . We will label these embeddings

$$\sigma_1, \dots, \sigma_r, \tau_1, \bar{\tau}_1, \dots, \tau_s, \bar{\tau}_s,$$

where σ_i for all $1 \leq i \leq r$ are real embeddings, and all $\tau_j, \bar{\tau}_j$ for all $1 \leq j \leq s$ are pairs of complex conjugate embeddings. In particular, $d = r + 2s$. Notice that for each $1 \leq i \leq r$, $\sigma_i(K) \subseteq \mathbb{R}$. Also, for each $1 \leq j \leq s$ and $\alpha \in K$, we can represent $\tau_j(\alpha)$ as a pair $(\tau_{j1}(\alpha), \tau_{j2}(\alpha)) \in \mathbb{R}^2$, where $\tau_{j1}(\alpha), \tau_{j2}(\alpha)$ are real and imaginary parts of $\tau_j(\alpha)$, respectively. Hence for each $1 \leq j \leq s$, $\tau_j(K) \subseteq \mathbb{R}^2$. Therefore

$$\Sigma = (\sigma_1, \dots, \sigma_r, \tau_{11}, \tau_{12}, \dots, \tau_{s1}, \tau_{s2}) : K \longrightarrow \mathbb{R}^d$$

is a vector space embedding of K into \mathbb{R}^d . In fact, $\Sigma(\mathcal{O}_K)$ is a lattice of full rank in \mathbb{R}^d . In particular, this means that we can pick a \mathbb{Z} -basis for \mathcal{O}_K , call it $\alpha_1, \dots, \alpha_d$. We define **discriminant** of K to be

$$\mathcal{D}_K = |\det(\alpha_i^{(j)})_{1 \leq i, j \leq d}|^2,$$

where $\alpha_i^{(j)} = \sigma_j(\alpha_i)$ for each $1 \leq j \leq r$, $\alpha_i^{(j)} = \tau_{j-r}(\alpha_i)$ for each $r+1 \leq j \leq r+s$, $\alpha_i^{(j)} = \bar{\tau}_{j-r-s}(\alpha_i)$ for each $r+s+1 \leq j \leq r+2s = d$; here each $\tau_j(\alpha_i)$ is thought of as a complex number, not as a pair of real numbers. \mathcal{D}_K is a very important **invariant** of K , i.e. does not depend on the choice of the basis $\alpha_1, \dots, \alpha_d$.

For each $a \in \mathcal{O}_K$, there exist integers u_1, \dots, u_d such that

$$a^{(j)} = \sum_{i=1}^d u_i \alpha_i^{(j)},$$

for all $1 \leq j \leq d$. In other words, these are particular sets of values of of the system of linear forms

$$L_j(\mathbf{x}) = \sum_{i=1}^d \alpha_i^{(j)} x_i, \quad 1 \leq j \leq d.$$

For a fixed number $r \in \mathbb{R}_{>0}$, consider a convex, $\mathbf{0}$ -symmetric set

$$C_r = \{\mathbf{x} \in \mathbb{R}^d : \sum_{j=1}^d |L_j(\mathbf{x})| \leq r\}.$$

The volume of this set turns out to be

$$\text{Vol}(C_r) = \frac{2^r \pi^s r^d}{d! \sqrt{|\mathcal{D}_K|}},$$

hence if $r = \{(4/\pi)^s d! \sqrt{|\mathcal{D}_K|}\}^{1/d}$, then $\text{Vol}(C_r) = 2^d$, and hence by Theorem 4.2 there exists $\mathbf{0} \neq \mathbf{u} \in C_r \cap \mathbb{Z}^d$. Then

$$L_1(\mathbf{u}), \dots, L_d(\mathbf{u})$$

are conjugates of some algebraic integer $\xi \in K$, and hence

$$|L_1(\mathbf{u}) \dots L_d(\mathbf{u})| \in \mathbb{Z}_{>0}.$$

Using the inequality of the arithmetic and geometric mean, we see that

$$1 \leq |L_1(\mathbf{u}) \dots L_d(\mathbf{u})| \leq \left(\frac{1}{d} \sum_{j=1}^d |L_j(\mathbf{u})| \right)^d,$$

and so

$$1 \leq \{(4/\pi)^s d! \sqrt{|\mathcal{D}_K|}\}^{1/d}.$$

We have proved the following result; this is Theorem 5 on p. 46 of [17].

Theorem 4.7.

$$|\mathcal{D}_K| \geq \{(\pi/4)^s d^d / d!\}^2 \geq 1.$$

Notice that Theorem 4.7 in particular implies that $\mathcal{D}_K \rightarrow \infty$ as $d \rightarrow \infty$, i.e. a number field of large degree cannot have a small discriminant. A geometric significance of discriminant can be expressed by the following well-known identity:

$$\det(\Sigma(\mathcal{O}_K)) = 2^{-s} \sqrt{|\mathcal{D}_K|}.$$

Therefore Theorem 4.7 implies that when d is large, $\Sigma(\mathcal{O}_K)$ is a lattice not only of large rank, but also of large determinant. For a detailed discussion of further deep applications of Minkowski's theorem to algebraic number theory see chapter 5 of [22].

5. SUCCESSIVE MINIMA

Theorem 4.4 gives a criterion for a convex, $\mathbf{0}$ -symmetric set to contain a collection of lattice points. This collection 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 5.1. 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 **successive 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 5.1. Notice that the N linearly independent vectors $\mathbf{u}_1, \dots, \mathbf{u}_N$ 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

$$\Lambda = \left(\begin{array}{ccccc} 1 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 1 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{2} \end{array} \right) \mathbb{Z}^5,$$

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

$$\lambda_1 = \dots = \lambda_5 = 1,$$

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

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

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

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

An immediate observation is that

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

Moreover, Minkowski's convex body theorem implies that

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

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

Theorem 5.1. *With notation as above,*

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

Proof. We present the proof in case $\Lambda = \mathbb{Z}^N$, leaving generalization of the given argument to arbitrary lattices as an exercise. We start with a proof of the lower bound following [17], which is considerably easier than the upper bound. Let $\mathbf{u}_1, \dots, \mathbf{u}_N$ be the N linearly independent vectors corresponding to the respective successive minima $\lambda_1, \dots, \lambda_N$, and let

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

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

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

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

(6)

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

since $\det(U)$ is an integer. Since $P \subseteq M$, $\text{Vol}(M) \geq \text{Vol}(P)$. Combining this last observation with (6) yields the lower bound of the theorem.

Next we prove the upper bound. The argument we present is due to M. Henk [19], and is at least partially based on Minkowski's original geometric ideas. For each $1 \leq i \leq N$, let

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

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 $\mathbf{u}_1, \dots, \mathbf{u}_N$ to be the N linearly independent vectors corresponding to the respective successive minima $\lambda_1, \dots, \lambda_N$. In fact, notice that there exists a matrix $A \in GL_N(\mathbb{Z})$ such that

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

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

$$A\mathbf{u}_i \in \lambda'_i AM \cap E_i, \quad \forall 1 \leq i \leq N,$$

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

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

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

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

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

$$(7) \quad \text{Vol}(C_q^N + M_N) \leq (2q + \gamma)^N,$$

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

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

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

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

and so

$$(8) \quad \begin{aligned} (\mathbf{z} - \mathbf{x}) - (\mathbf{z} - \mathbf{y}) &= \mathbf{y} - \mathbf{x} \in \text{int}(M_1) - \text{int}(M_1) \\ &= \{\mathbf{z}_1 - \mathbf{z}_2 : \mathbf{z}_1, \mathbf{z}_2 \in M_1\} = \text{int}(\lambda_1 M), \end{aligned}$$

which would contradict minimality of λ_1 . Therefore

$$(9) \quad \text{Vol}(C_q^N + M_1) = (2q + 1)^N \text{Vol}(M_1) = (2q + 1)^N \left(\frac{\lambda_1}{2}\right)^N \text{Vol}(M).$$

To finish the proof, we need the following lemma.

Lemma 5.2. *For each $1 \leq i \leq N - 1$,*

$$(10) \quad \text{Vol}(C_q^N + M_{i+1}) \geq \left(\frac{\lambda_{i+1}}{\lambda_i}\right)^{N-i} \text{Vol}(C_q^N + M_i).$$

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

$$(\mathbf{x}_{i+1}, \dots, \mathbf{x}_N) \neq (\mathbf{y}_{i+1}, \dots, \mathbf{y}_N).$$

Then

$$(11) \quad (\mathbf{x} + \text{int}(M_{i+1})) \cap (\mathbf{y} + \text{int}(M_{i+1})) = \emptyset.$$

Indeed, suppose (11) is not true, i.e. there exists $\mathbf{z} \in (\mathbf{x} + \text{int}(M_{i+1})) \cap (\mathbf{y} + \text{int}(M_{i+1}))$. Then, as in (8) above, $\mathbf{x} - \mathbf{y} \in \text{int}(\lambda_{i+1}M)$. But we also have

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

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

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

On the other hand, at least one of

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

is not equal to 0. Hence $\mathbf{x} - \mathbf{y}, \mathbf{u}_1, \dots, \mathbf{u}_i$ are linearly independent, but this means that $\text{int}(\lambda_{i+1}M)$ contains $i + 1$ linearly independent points, contradicting minimality of λ_{i+1} . This proves (11). Notice that (11) implies

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

and

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

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

$$(12) \quad \text{Vol}(C_q^i + M_{i+1}) \geq \left(\frac{\lambda_{i+1}}{\lambda_i}\right)^{N-i} \text{Vol}(C_q^i + M_i).$$

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

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

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

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

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

and so to establish (12) it is sufficient to show that

$$(13) \quad \text{Vol}(C_q^i + f_1(M_i)) \geq \text{Vol}(C_q^i + M_i).$$

Let

$$E_i^\perp = \text{span}_{\mathbb{R}}\{\mathbf{e}_{i+1}, \dots, \mathbf{e}_N\},$$

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

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

in other words, although it is not necessarily true that $M_i \subseteq f_1(M_i)$, each section of M_i by a translate of E_i is contained in a translate of some such section of $f_1(M_i)$. Therefore

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

and hence

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

where Vol_i stands for the i -dimensional volume. This completes the proof of (13), and hence of the lemma. \square

Now, combining (7), (9), and (10), we obtain:

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

hence

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

as $q \rightarrow \infty$, 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.

6. INHOMOGENEOUS MINIMUM

Here we exhibit one important application of Minkowski's successive minima theorem. As before, let $\Lambda \subseteq \mathbb{R}^N$ be a lattice of full rank, and let $M \subseteq \mathbb{R}^N$ be a convex $\mathbf{0}$ -symmetric set of non-zero volume. Throughout this section, we let

$$\lambda_1 \leq \cdots \leq \lambda_N$$

to be the successive minima of M with respect to Λ . We define the **inhomogeneous minimum** of M with respect to Λ to be

$$\mu = \inf\{\lambda \in \mathbb{R}_{>0} : \lambda M + \Lambda = \mathbb{R}^N\}.$$

The main objective of this section is to obtain some basic bounds on μ . We start with the following result of Jarnik [21].

Lemma 6.1.

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

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

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

Recall from Theorem 1.1 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

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

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

Let $\mathbf{z} \in \mathbb{R}^N$ be an arbitrary point. We want to prove that there exists a point $\mathbf{v} \in \Lambda$ such that

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

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

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

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

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

We can also choose integer v_1, \dots, v_N such that

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

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

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

by the definition of a distance function and Exercise 1.5. This completes the proof. \square

Using Lemma 6.1 along with Minkowski's successive minima theorem, we can obtain some bounds on μ in terms of the determinant of Λ and volume of M . A nice bound can be easily obtained in an important special case.

Corollary 6.2. *If $\lambda_1 \geq 1$, then*

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

Proof. Since

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

Theorem 5.1 implies

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

and by Lemma 6.1,

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

The result follows by combining these two inequalities. \square

A general bound depending also on λ_1 was obtained by Scherk [34], once again using Minkowski's successive minima theorem (Theorem 5.1) and Jarnik's inequality (Lemma 6.1). He observed that if λ_1 is fixed and $\lambda_2, \dots, \lambda_N$ are subject to the conditions

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

then the maximum of the sum

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

is attained when

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

Hence we obtain Scherk's inequality for μ .

Corollary 6.3.

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

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

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

hence

$$(14) \quad \mu \geq \left(\frac{\det(\Lambda)}{\text{Vol}(M)} \right)^{1/N}.$$

In fact, by Theorem 5.1,

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

and combining this with (14), we obtain

$$(15) \quad \mu \geq \frac{\lambda_1}{2}.$$

Jarnik obtained a considerably better lower bound for μ in [21].

Lemma 6.4.

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

Proof. Let $\mathbf{u}_1, \dots, \mathbf{u}_N$ 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

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

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

$$(16) \quad F\left(\mathbf{x} - \frac{1}{2}\mathbf{u}_N\right) \geq \frac{1}{2}\lambda_N.$$

Suppose not, then there exists some $\mathbf{x} \in \Lambda$ such that $F(\mathbf{x} - \frac{1}{2}\mathbf{u}_N) < \frac{1}{2}\lambda_N$, and so, by Exercise 1.5

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

and similarly

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

Therefore, by definition of λ_N ,

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

and so $\mathbf{u}_N = \mathbf{x} + (\mathbf{u}_N - \mathbf{x}) \in \text{span}_{\mathbb{R}}\{\mathbf{u}_1, \dots, \mathbf{u}_{N-1}\}$, which is a contradiction. Hence we proved (16) for all $\mathbf{x} \in \Lambda$.

Exercise 6.1. *Prove that*

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

Then lemma follows by combining (16) with Exercise 6.1. □

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.

7. 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. Both of these would make nice topics for potential student presentations. An excellent comprehensive, although slightly outdated, reference on this subject is the celebrated book by Conway and Sloane [8]. 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

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

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

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

whenever B'_i and B'_j are spheres of radius larger than r such that $B_i \subsetneq B'_i$, $B_j \subsetneq B'_j$. 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**. Although clearly most packings are not lattices, it is not unreasonable to expect that best results may come from lattice packings; we will mostly be concerned with them.

Definition 7.1. Let $\Lambda \subseteq \mathbb{R}^N$ be a lattice of full rank. The **density** of corresponding sphere packing is defined to be

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

where ω_N is the volume of a unit ball in \mathbb{R}^N , given by (5), 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 [17] for the general version with proof).

Theorem 7.1. *In each dimension N there exist lattice packings with density*

$$(17) \quad \Delta \geq \frac{\zeta(N)}{2^{N-1}},$$

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

Ironically, all known proofs of Theorem 7.1 are non-constructive, so it is not generally known how to construct lattice packings with density as good as (17); in particular, in dimensions above 1000 the lattices whose existence is guaranteed by Theorem 7.1 are denser than all the presently known ones. Some of the known lattices with relatively high packing density come from rings of algebraic integers of number fields with large degree and small discriminant, and from irreducible smooth algebraic curves with many rational points over a finite field. These approaches are explained in details in the work of Michael Tsfasman.

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**

$$(18) \quad \begin{pmatrix} 1 & \frac{1}{2} \\ 0 & \frac{\sqrt{3}}{2} \end{pmatrix} \mathbb{Z}^2,$$

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 [18], and the lengthy computational part is still forthcoming. 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 7.2. Let $\Lambda \subseteq \mathbb{R}^N$ be a lattice of full rank. The **thickness** of corresponding sphere covering is defined to be

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

where ω_N is the volume of a unit ball in \mathbb{R}^N , given by (5), 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 6.1 and 6.4 we

obtain the following bounds on the covering radius in terms of successive minima of Λ :

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

The optimal sphere covering is only known in dimension $N = 2$, in which case it is given by the same hexagonal lattice (18), and is equal to ≈ 1.209199 . Best lattice coverings are not 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 **ESM lattices** (equal successive minima); another term **well rounded lattices** is also sometimes used. Notice that if Λ is ESM, then by Lemma 6.4 we have

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

although it is clearly impossible for equality to hold in this inequality.

Sphere packing and covering results have numerous engineering applications, among which there are applications to coding theory, telecommunications, and image processing. ESM lattices play an especially important role in these fields of study.

8. 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.

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

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

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

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

Definition 8.1. A basis $\{\mathbf{v}_1, \dots, \mathbf{v}_N\}$ of Λ is said to be **Minkowski reduced with respect to M** if for each $1 \leq i \leq N$, \mathbf{v}_i is such that

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

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 $\{\mathbf{v}_1, \dots, \mathbf{v}_N\}$ be a Minkowski reduced basis of Λ with respect to M . Then

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

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

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

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

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

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

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

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

$$(19) \quad Q(\mathbf{e}_i) \leq Q(\mathbf{x}) \quad \forall \mathbf{x} \in \mathbb{Z}^N, \quad \gcd(x_1, \dots, x_N) = 1.$$

If a positive definite quadratic form satisfies (19), we will say that it is **Minkowski reduced**.

Exercise 8.1. *Prove that every positive definite quadratic form is arithmetically equivalent to a Minkowski reduced form.*

Exercise 8.2. *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

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

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 [17], which is due to Minkowski, Mahler, and Weyl.

Theorem 8.1. *Let $\nu_1 = 1$, and $\nu_i = \left(\frac{3}{2}\right)^{i-2}$ for each $2 \leq i \leq N$. Then*

$$(20) \quad \lambda_i \leq F(\mathbf{v}_i) \leq \nu_i \lambda_i.$$

Moreover,

$$(21) \quad \prod_{i=1}^N F(\mathbf{v}_i) \leq 2^N \left(\frac{3}{2}\right)^{\frac{(N-1)(N-2)}{2}} \frac{\det(\Lambda)}{\text{Vol}(M)}.$$

Proof. It is easy to see that (21) follows immediately by combining (20) with Theorem 5.1, hence we only need to prove (20). We will only prove (20) 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 $F(\mathbf{v}_i) \geq \lambda_i$ for each $1 \leq i \leq N$, and that $F(\mathbf{v}_1) = \lambda_1$. Hence we only need to prove that for each $2 \leq i \leq N$

$$(22) \quad F(\mathbf{v}_i) \leq \nu_i \lambda_i.$$

Let $\mathbf{u}_1, \dots, \mathbf{u}_N$ be the linearly independent vectors corresponding to successive minima $\lambda_1, \dots, \lambda_N$, i.e.

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

Then, by linear independence, for each $2 \leq i \leq N$ at least one of $\mathbf{u}_1, \dots, \mathbf{u}_i$ does not belong to the subspace $\text{span}_{\mathbb{R}}\{\mathbf{v}_1, \dots, \mathbf{v}_{i-1}\}$, call

this vector \mathbf{u}_j . If the set $\mathbf{v}_1, \dots, \mathbf{v}_{i-1}, \mathbf{u}_j$ is extendable to a basis of \mathbb{Z}^N , then by construction of reduced basis we must have

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

and so it implies that $\lambda_i = F(\mathbf{v}_i)$, proving (22) in this case.

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

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

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

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

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

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

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

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

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

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

Combining this with the previous case, we conclude that

$$(23) \quad F(\mathbf{v}_i) \leq \max \left\{ \lambda_i, \frac{1}{2} \left(\sum_{k=1}^{i-1} F(\mathbf{v}_k) + \lambda_i \right) \right\}, \quad \forall 2 \leq i \leq N.$$

Hence we obtain

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

hence $F(\mathbf{v}_2) = \lambda_2$. More generally, one can easily deduce (22) from (23). This finishes the proof. \square

As a corollary of Theorem 8.1, we can easily deduce the following bound on the product of diagonal coefficients of reduced positive definite quadratic forms.

Exercise 8.3. *Let*

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

be a Minkowski reduced positive definite quadratic form. Then

$$(24) \quad \prod_{i=1}^N b_{ii} \leq \frac{4^N}{\omega_N^2} \left(\frac{3}{2} \right)^{\frac{(N-1)(N-2)}{2}} \det(Q),$$

where ω_N is the volume of a unit ball in \mathbb{R}^N , which is given by (5).

(Hint: let $\Lambda = \mathbb{Z}^N$, and let M be the convex body corresponding to the distance function $F = \sqrt{Q}$; apply Theorem 8.1.)

There are other reduction algorithms 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. Unfortunately, it is not known how to implement any of the reduction algorithms to work in polynomial time on a Turing machine, i.e. on a modern computer. In fact, the famous problem of finding the shortest non-zero vector of a given lattice is believed to be NP-complete. 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 $\mathbf{b}_1, \dots, \mathbf{b}_N$ for Λ such that

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

where $\mathbf{x} \in \Lambda$ is a shortest non-zero vector. Some good references on this subject are [23], [17], [1], and [29].

9. LATTICE POINTS IN HOMOGENEOUSLY EXPANDING DOMAINS

Let $M \subseteq \mathbb{R}^N$ be closed, bounded, and Jordan measurable with $\text{Vol}(M) > 0$, 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? If M is convex and $\mathbf{0}$ -symmetric, then a bound on this number can be derived from the generalized version of Minkowski's convex body theorem, namely Theorem 4.4. In this section, however, we consider a more general set M as above, and will be interested in asymptotic behaviour of the function

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

as $t \rightarrow \infty$. 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 9.1. Let S be a subset of some Euclidean space. A map

$$\varphi : S \rightarrow \mathbb{R}^N$$

is called a **Lipschitz map** if there exists $\mathcal{C} \in \mathbb{R}_{>0}$ such that for all $\mathbf{x}, \mathbf{y} \in S$

$$\|\varphi(\mathbf{x}) - \varphi(\mathbf{y})\|_2 \leq \mathcal{C}\|\mathbf{x} - \mathbf{y}\|_2.$$

We say that \mathcal{C} is the corresponding **Lipschitz constant**.

Let

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

be the closed unit cube.

Definition 9.2. We say that $S \subseteq \mathbb{R}^N$ is **Lipschitz parametrizable** if there exists a finite number of Lipschitz maps

$$\varphi_j : C^N \rightarrow S,$$

such that $S = \bigcup_j \varphi_j(C^N)$.

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 [22].

Theorem 9.1. *Let $t \in \mathbb{R}_{>0}$, then*

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

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

Proof. Let $\mathbf{x}_1, \dots, \mathbf{x}_N$ be a basis for Λ , and let \mathcal{F} be the corresponding fundamental parallelotope, i.e.

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

For each point $\mathbf{x} \in \Lambda$ we will write $\mathcal{F}_{\mathbf{x}}$ for the translate of \mathcal{F} by \mathbf{x} :

$$\mathcal{F}_{\mathbf{x}} = \mathcal{F} + \mathbf{x}.$$

Notice that if $\mathbf{x} \in tM \cap \Lambda$, then $\mathcal{F}_{\mathbf{x}} \cap tM \neq \emptyset$. Moreover, either

$$\mathcal{F}_{\mathbf{x}} \subseteq \text{int}(tM),$$

or

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

Let

$$\begin{aligned} m(t) &= |\{\mathbf{x} \in \Lambda : \mathcal{F}_{\mathbf{x}} \in \text{int}(tM)\}|, \\ b(t) &= |\{\mathbf{x} \in \Lambda : \mathcal{F}_{\mathbf{x}} \cap \partial(tM) \neq \emptyset\}|. \end{aligned}$$

Then clearly

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

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

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

hence

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

Therefore to conclude the proof we only need to estimate $b(t)$. Let

$$\varphi : C^{N-1} \rightarrow \partial M$$

be one of the Lipschitz parametrizing 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

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

for each $\mathbf{x}, \mathbf{y} \in C_i$, 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 $\mathbf{x} \in \Lambda$ such that the corresponding translate $\mathcal{F}_{\mathbf{x}}$ 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. \square

Theorem 9.1 provides an asymptotic formula for $G(t)$, demonstrating a very important general principle, namely that as $t \rightarrow \infty$, $G(t)$ grows like $\frac{\text{Vol}(M)}{\det(\Lambda)}t^N$, 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

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

A large amount of work has been done in this direction (see for instance pp. 140 - 147 of [17] for an overview of results and bibliography). This subject essentially originated in a paper of Davenport [9], who used a principle of Lipschitz [24]; also see [41] for a nice overview of Davenport's result and its generalizations. We present here without proof a result of P. G. Spain [39], which is a refinement of Davenport's bound, and can be thought of as a continuation of Theorem 9.1.

Theorem 9.2. *Let the notation be as in Theorem 9.1, and let \mathcal{C} be the maximal Lipschitz constant corresponding to parametrization of ∂M . Then for each $t \in \mathbb{R}_{>0}$,*

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

10. ERHART POLYNOMIAL

As in section 9, let $M \subseteq \mathbb{R}^N$ be closed, bounded, Jordan measurable with $\text{Vol}(M) > 0$, 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 9.1 and 9.2, we can conclude that

$$(25) \quad G(t, M, \Lambda) = |tM \cap \Lambda| \leq \frac{\text{Vol}(M)}{\det(\Lambda)} t^N + \sum_{i=0}^{N-1} 2^N \mathcal{C}^i \binom{N-1}{i} t^i,$$

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

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

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

$$(26) \quad |tM \cap \Lambda| = (2t+1)^N = \sum_{i=0}^N 2^i \binom{N}{i} t^i,$$

which is similar to the upper bound of (25) in this case.

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

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

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 [11]. First we consider a special case of polytopes, namely simplices.

Lemma 10.1. *Let $\mathbf{a}_1, \dots, \mathbf{a}_N \in \mathbb{Z}^N$ be linearly independent, and define the simplex*

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

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

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

Proof. Let A be the half-open parallelotope spanned by the vectors $\mathbf{a}_1, \dots, \mathbf{a}_N$, i.e.

$$A = \left\{ \sum_{i=1}^N t_i \mathbf{a}_i : 0 \leq t_i < 1 \forall 1 \leq i \leq N \right\}.$$

For every $\mathbf{y} \in tS \cap \mathbb{Z}^N$ there exists a unique representation of \mathbf{y} of the form

$$(27) \quad \mathbf{y} = \mathbf{x} + \sum_{i=1}^N \alpha_i \mathbf{a}_i,$$

where $\mathbf{x} \in A \cap \mathbb{Z}^N$ and $\alpha_1, \dots, \alpha_N \in \mathbb{Z}_{\geq 0}$. For each $0 \leq j \leq t$, let H_j be the hyperplane which passes through the points $j\mathbf{a}_1, \dots, j\mathbf{a}_N$. We will determine the number of points of \mathbb{Z}^N in $H_j \cap tS$, and the number of points of $\mathbb{Z}^N \cap tS$ in the strips of space bounded by H_{j-1} and H_j for each $1 \leq j \leq t$; notice that $H_0 = \{\mathbf{0}\}$.

First, let $\mathbf{x} = \mathbf{0}$ in (27). Then \mathbf{y} as in (27) lies in H_j if and only if

$$(28) \quad \sum_{i=1}^N \alpha_i = j, \quad 0 \leq \alpha_i \leq j \quad \forall 1 \leq i \leq N.$$

We will prove now that there are precisely $\binom{N+j-1}{N-1}$ possibilities for $\alpha_1, \dots, \alpha_N$ satisfying (28) 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 $\alpha_1, \dots, \alpha_{N-1}$ 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 $\alpha_1, \dots, \alpha_N$ satisfying (28) is

$$(29) \quad \sum_{\alpha_N=0}^j \binom{N+(j-\alpha_N)-2}{N-2} = \sum_{i=0}^j \binom{N+i-2}{N-2}.$$

Then our claim follows by combining (29) with the result of the following exercise.

Exercise 10.1. *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 \mathbf{y} as in (27) with $\mathbf{x} = \mathbf{0}$ on $\bigcup_{j=0}^t H_j$, we sum over j , using the result of Exercise 10.1 once again:

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

If \mathbf{x} in (27) lies properly between H_0 and H_1 , then the number of possible \mathbf{y} as given by (27) that lie in $\bigcup_{j=0}^t H_j$ reduces to $\binom{N+t-1}{N}$. Similarly, the number of possibilities for \mathbf{y} as in (27) with \mathbf{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 $\mathbf{x} \in A \cap \mathbb{Z}^N$ which lie properly between H_{i-1} and H_i or on H_i , then the number of corresponding points \mathbf{y} as in (27) 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. \square

Let $\mathbf{a}_1, \dots, \mathbf{a}_N \in \mathbb{Z}^N$ be linearly independent, and let S be the simplex $\text{Co}(\mathbf{0}, \mathbf{a}_1, \dots, \mathbf{a}_N)$, as in Lemma 10.1. Define the **pseudo-simplex** associated with S

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

Lemma 10.2. $G(tS_0)$ 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 $\mathbf{0}$ and satisfy

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

Then

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

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 10.1,

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

is a polynomial in t . \square

We are now ready to prove Erhart's theorem.

Theorem 10.3 (Ehrhart). *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

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

where $S_0^{(1)}, \dots, S_0^{(r)}$ are precisely the cells of this simplicial complex which contain $\mathbf{0}$, but are not equal to $\{\mathbf{0}\}$. The theorem follows by Lemma 10.2. \square

$G(t\mathcal{P})$ as in Theorem 10.3 is called **Ehrhart polynomial** of \mathcal{P} . An excellent reference on Ehrhart polynomials, their many fascinating properties, and connections to other important mathematical objects is [2]. For a general lattice polytope \mathcal{P} very little is known about the coefficients of its Ehrhart 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 $\text{Vol}(\mathcal{P})$, 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 [2] for further details.

Notice that (26) provides an explicit example of Ehrhart polynomial in the simple case of a cube. To conclude this section, we will give two more explicit examples of Ehrhart polynomial. The first one is for an open simplex, which is precisely the interior of the simplex S of Lemma

10.1 with $\mathbf{a}_i = \mathbf{e}_i$ for each $1 \leq i \leq N$; the following observation along with the proof is due to S. I. Sobolev.

Proposition 10.4. *Define an open simplex*

$$S^\circ = \left\{ \mathbf{x} \in \mathbb{R}^N : x_i > 0 \forall 1 \leq i \leq N, \sum_{i=1}^N x_i < 1 \right\}.$$

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

$$(30) \quad G(tS^\circ) = \binom{t-1}{N}.$$

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

$$tS_1^\circ = \{ \mathbf{x} \in \mathbb{R}^N : 0 < x_1 < \cdots < x_N < t \}.$$

This transformation is volume-preserving and maps \mathbb{Z}^N to itself. Integral points of tS_1° correspond to increasing sequences of integers $0 < y_1 < \cdots < y_N < t$. The number of such sequences is precisely $\binom{t-1}{N}$, which is the number of all possible N -element subsets of the set $\{1, \dots, t-1\}$. \square

Notice that (30) can be thought of as a geometric interpretation of binomial coefficients. The next example is related to the one in Proposition 10.4, but is more general.

Proposition 10.5 ([4]). *Let*

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

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

$$(31) \quad G(t\mathcal{S}_N) = \sum_{i=0}^{\min\{t, N\}} 2^i \binom{N}{i} \binom{t}{i}.$$

Proof. Notice that for each $0 \leq i \leq \min\{t, N\}$ the number of points in $t\mathcal{S}_N \cap \mathbb{Z}^N$ with precisely i nonzero coordinates is

$$2^i \binom{N}{i} \binom{t}{i}.$$

Indeed, the number of choices of which coordinates are nonzero is $\binom{N}{i}$; for each such choice there are 2^i choices of \pm signs, and $\binom{t}{i}$ choices of absolute values. Summing over all $0 \leq i \leq \min\{t, N\}$ completes the proof. \square

Remark 10.1. A remarkable property of the polynomial in Proposition 10.5 is that the right hand side (31) is symmetric in t and N . This means that

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

11. SIEGEL'S LEMMA

In this section we conclude the Geometry of Numbers part of these notes by discussing a topic which has many important connections to Diophantine Approximations, thus providing a reasonable transition. Recall that 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 similar problem for certain lattices of not full rank. More specifically, for the rest of this section $\Lambda \subseteq \mathbb{R}^N$ will be a lattice of rank $N - M$, $1 \leq M < N$. More specifically, let

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

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

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

then $\Lambda \subseteq \mathbb{R}^N$ 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 $\mathbf{x} \in \Lambda$. Here is one way to do it. Suppose that we can prove that there must exist a nonzero vector $\mathbf{x} \in \Lambda$ with

$$(32) \quad \|\mathbf{x}\|_2 \leq N|\mathbf{x}| \leq f(A),$$

where $f(A) = f(a_{11}, \dots, a_{MN})$ is some explicit function of the entries of A . Then for each vector $\mathbf{x} \in \mathbb{Z}^N$ with $\|\mathbf{x}\|_2 \leq f(A)$ we can check whether $\mathbf{x} \in \Lambda$, 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 $A\mathbf{x} = \mathbf{0}$. Therefore we are interested in proving the existence of a nonzero vector $\mathbf{x} \in \Lambda$ with explicitly bounded norm, as suggested by (32). An idea of this sort was first used by A. Thue in 1909 [40], but formally stated only in 1929 by C. L. Siegel [38]. Our presentation partially follows [36].

Theorem 11.1 (Siegel's Lemma). *With notation as above, there exists $\mathbf{0} \neq \mathbf{x} \in \Lambda$ with*

$$(33) \quad |\mathbf{x}| < 2 + (N|A|)^{\frac{M}{N-M}},$$

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

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

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

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 $T_A : \mathbb{R}^N \rightarrow \mathbb{R}^M$ be a linear map, given by $T_A(\mathbf{x}) = A\mathbf{x}$ for each $\mathbf{x} \in \mathbb{R}^N$. Notice that for every $\mathbf{x} \in C_H^N$,

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

i.e. T_A maps C_H^N onto $C_{N|A|H}^M \subseteq \mathbb{R}^M$, since $\text{rk}(A) = M$. 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 $C_H^N \cap \mathbb{Z}^N$ onto $C_{N|A|H}^M \cap \mathbb{Z}^M$ in a one-to-one manner. Hence, there must exist $\mathbf{x} \neq \mathbf{y} \in C_H^N \cap \mathbb{Z}^N$ such that $T_A(\mathbf{x}) = T_A(\mathbf{y})$, i.e.

$$T_A(\mathbf{x} - \mathbf{y}) = \mathbf{0},$$

and so $\mathbf{x} - \mathbf{y} \in \Lambda$. On the other hand,

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

and this finishes the proof. \square

Notice that the main underlying idea in the proof of Siegel's Lemma was Dirichlet's box principle, which we have already seen before. It will make further important appearances in the theory of Diophantine Approximations. It is remarkable that the exponent $\frac{M}{N-M}$ in the upper bound of (33) 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 $A\mathbf{x} = \mathbf{0}$ must have $x_N = R^{N-1}x_1$. Therefore, if

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

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

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

Siegel's Lemma type of results have been proved in a considerably more sophisticated forms with the use of height functions and connections to Diophantine Geometry by a number of authors. Most notably, see the celebrated papers of Bombieri and Vaaler [3] and of Roy and Thunder [32], as well as a very nice overview of this subject in [36]. The original motivation came from Diophantine Approximations and Transcendental Number Theory, where this principle was used to construct polynomials of bounded height and high order of vanishing at a prescribed set of points. Siegel's Lemma, being a result in the Geometry of Numbers, is also often thought of as a result in Diophantine Approximations, since by exhibiting explicit search bounds we are *approximating* solutions of a system of linear Diophantine equations.

Part 2. Diophantine Approximations

12. DIRICHLET, LIOUVILLE, ROTH

Recall a classical theorem from real analysis (see, for instance, Theorem 1.20 on p.9 of [33]).

Theorem 12.1. *The set of rational number \mathbb{Q} is **dense** inside of the set of real number \mathbb{R} , i.e. if $x < y \in \mathbb{R}$, then there exists $z \in \mathbb{Q}$ such that*

$$x < z < y.$$

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

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

so $nx + 1 < ny$. Let $m \in \mathbb{Z}_{>0}$ be such that

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

then we have

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

and hence

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

Let $z = \frac{m}{n} \in \mathbb{Q}$, and this finishes the proof. \square

Theorem 12.1 implies that given a real number we can approximate it arbitrarily well by rational numbers. For many purposes we may want to control how “complicated” the rational numbers we use for such approximations are, i.e. we may want to bound the size of their denominators. This is the starting point of the theory of Diophantine Approximations. The first result in this direction dates back to Dirichlet, and is proved with the use of Dirichlet’s box principle; in fact, this is most likely the theorem to which this principle owes its name. For the rest of this section we follow [36].

Theorem 12.2 (Dirichlet, (1842)). *Let $\alpha \in \mathbb{R}$, and let $Q \in \mathbb{Z}_{>0}$. There exist relatively prime integers p, q with $1 \leq q \leq Q$ such that*

$$(34) \quad \left| \alpha - \frac{p}{q} \right| \leq \frac{1}{q(Q+1)}.$$

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

$$(35) \quad \left| \alpha - \frac{p}{q} \right| \leq \frac{1}{q^2}.$$

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

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

Consider the numbers $\{l\alpha\}$, $1 \leq l \leq Q+1$, where $\{ \}$ denotes the fractional part function, i.e. $\{x\} = x - [x]$. Clearly, all of these numbers are distinct and lie in the interval $[0, 1)$.

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

$$(36) \quad |l\alpha - [l\alpha]| \leq \frac{1}{Q+1},$$

or

$$(37) \quad |l\alpha - [l\alpha] - 1| \leq \frac{1}{Q+1}.$$

This means that there exists an integer $1 \leq l \leq Q$ and an integer m equal to either $[l\alpha]$ or $[l\alpha] - 1$, depending on whether (36) or (37) holds, such that

$$|l\alpha - m| \leq \frac{1}{Q+1}.$$

Let $d = \gcd(l, m)$, and let $p = \frac{m}{d}$ and $q = \frac{l}{d}$, then

$$|qd\alpha - pd| \leq \frac{1}{Q+1},$$

meaning that

$$\left| \alpha - \frac{p}{q} \right| \leq \frac{1}{qd(Q+1)} \leq \frac{1}{q(Q+1)},$$

proving (34) in this case.

Case 2. Now assume that one of the subintervals $\left[\frac{i-1}{Q+1}, \frac{i}{Q+1} \right)$ for some $1 \leq i \leq Q+1$ does not contain any of the numbers $\{l\alpha\}$, $1 \leq l \leq Q+1$. Since there are $Q+1$ such numbers and $Q+1$ subintervals, one of the subintervals must contain two such numbers, say $\left[\frac{j-1}{Q+1}, \frac{j}{Q+1} \right)$ for

some $1 \leq j \leq Q + 1$ contains $\{l\alpha\}$ and $\{m\alpha\}$ for some $1 \leq l < m \leq Q + 1$. Therefore

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

Once again, let $d = \gcd((m - l), ([m\alpha] - [l\alpha]))$, and let $p = \frac{[m\alpha] - [l\alpha]}{d}$ and $q = \frac{m - l}{d}$, and so in the same way as above we obtain (34).

Exercise 12.1. Prove that if $\alpha = \frac{a}{Q+1}$ for some integer a with

$$\gcd(a, Q + 1) = 1,$$

then there is equality in (34).

We can now derive (35) follows from (34): since $q \leq Q$,

$$(38) \quad \left| \alpha - \frac{p}{q} \right| \leq \frac{1}{q(Q + 1)} < \frac{1}{q^2}.$$

Now suppose that there are only finitely many rationals that satisfy (35), call them

$$\frac{p_1}{q_1}, \dots, \frac{p_k}{q_k}.$$

Let

$$\delta = \min_{1 \leq i \leq k} \left| \alpha - \frac{p_i}{q_i} \right|,$$

then $\delta > 0$, since α is irrational. Let $Q \in \mathbb{Z}_{>0}$ be such that

$$\frac{1}{Q} < \delta.$$

By (38), there must exist $\frac{p}{q}$ with $1 \leq q \leq Q$ such that

$$\left| \alpha - \frac{p}{q} \right| \leq \frac{1}{q(Q + 1)} < \delta,$$

hence $\frac{p}{q} \notin \left\{ \frac{p_1}{q_1}, \dots, \frac{p_k}{q_k} \right\}$, which is a contradiction. Thus there must be infinitely many such rationals. \square

Remark 12.1. Notice that the argument that derives (35) from (34) is very similar to Euclid's proof of the infinitude of primes.

Hurwitz (1891) improved Dirichlet's bound (35) slightly by showing that for any irrational $\alpha \in \mathbb{R}$ there exist infinitely many distinct rational numbers $\frac{p}{q}$ such that

$$(39) \quad \left| \alpha - \frac{p}{q} \right| \leq \frac{1}{\sqrt{5} q^2}.$$

We will now show that in a certain sense (39) is best possible.

Lemma 12.3. *Let $\alpha \in \mathbb{R}$ be a quadratic irrational satisfying $f(\alpha) = 0$, where*

$$f(x) = ax^2 + bx + c$$

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

$$(40) \quad \left| \alpha - \frac{p}{q} \right| < \frac{1}{Aq^2}.$$

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

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

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

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

Now suppose that for some $\frac{p}{q} \in \mathbb{Q}$ (40) holds. Notice that since $f(x)$ is a quadratic polynomial with irrational roots, then

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

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

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

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

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

The left hand side of this inequality is not 0 since $A > \sqrt{D}$, and hence

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

This implies that there are only finitely many possibilities for the denominator q , but for each such q there can be only finitely many p so that (40) holds. This completes the proof. \square

Remark 12.2. Let $\alpha = \frac{1+\sqrt{5}}{2}$, then the corresponding polynomial

$$f(x) = x^2 - x - 1,$$

and its discriminant is $D = 5$. By Lemma 12.3, if $A > \sqrt{5}$ then there are only finitely many $\frac{p}{q} \in \mathbb{Q}$ such that

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

which proves that Hurwitz's bound (39) is best possible.

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

$$(41) \quad \left| \alpha - \frac{p}{q} \right| \geq \frac{C(\alpha)}{q^2}.$$

In other words, quadratic irrationals are *badly approximable*.

Definition 12.1. An irrational number α is called **badly approximable** if there exists a positive real constant $C(\alpha)$ such that (41) holds for any $\frac{p}{q} \in \mathbb{Q}$.

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

Theorem 12.4 (Liouville). *Let $\alpha \in \mathbb{R}$ be an algebraic number of degree $d = \deg(f) \geq 2$, where $f(x) \in \mathbb{Z}[x]$ is the minimal polynomial of α over \mathbb{Q} . Then there exists a positive real constant $C(\alpha)$ such that for any $\frac{p}{q} \in \mathbb{Q}$*

$$(42) \quad \left| \alpha - \frac{p}{q} \right| \geq \frac{C(\alpha)}{q^d}.$$

Proof. Let

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

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

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

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

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

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

Liouville used his theorem to construct the first known example of a transcendental number (recall that a complex number is called **transcendental** if it is not a root of any polynomial with integer coefficients, i.e. if it is not algebraic).

Corollary 12.5 (Liouville). *The number*

$$\alpha = \sum_{n=1}^{\infty} \frac{1}{a^{n!}}$$

is transcendental for any $a \in \mathbb{Z}_{>0}$.

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

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

Then

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

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

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

and then we have

$$(43) \quad \left| \alpha - \frac{p_k}{q_k} \right| < \frac{\mathcal{C}}{a^{(k+1)!}} = \frac{\mathcal{C}}{q_k^{(k+1)}} < \frac{\mathcal{C}}{q_k^k}.$$

Now suppose that α is algebraic of degree d . Then, by Theorem 12.4, there exists a constant $C(\alpha)$ such that

$$\left| \alpha - \frac{p_k}{q_k} \right| \geq \frac{C(\alpha)}{q_k^d},$$

for every $k \in \mathbb{Z}_{>0}$. However, if we take k large enough so that

$$\frac{\mathcal{C}}{q_k^k} < \frac{C(\alpha)}{q_k^d},$$

then (43) implies a contradiction; more specifically, we just need to take k large enough so that

$$k!(k-d) > \frac{\ln \mathcal{C} - \ln C(\alpha)}{\ln a}.$$

This completes the proof. \square

Remark 12.3. Numbers that can be proved to be transcendental using Liouville's theorem are called **Liouville numbers**; they form a set of measure zero. In particular, e and π are not Liouville numbers; moreover, most transcendental numbers are not Liouville. In general, almost all numbers are transcendental, since the set of all algebraic numbers is countable, and hence of measure zero.

Theorem 12.4 implies that if α is an algebraic number of degree $d \geq 2$ and $\mu > d$, then there are only finitely many $\frac{p}{q} \in \mathbb{Q}$ with $\gcd(p, q) = 1$ such that

$$(44) \quad \left| \alpha - \frac{p}{q} \right| < \frac{1}{q^\mu}.$$

Indeed, suppose there were infinitely many rational numbers for which (44) holds. Let $C(\alpha)$ be the constant guaranteed by Theorem 12.4. Let Q be an integer so that $C(\alpha) > \frac{1}{Q^{\mu-d}}$. Clearly there can be only finitely many $\frac{p}{q}$ with $\gcd(p, q) = 1$ for which (44) holds with $q \leq Q$, hence there must be infinitely many such rationals with $q > Q$. Suppose $\frac{p}{q}$ is one of them, then

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

which contradicts (42). This proves finiteness of the number of solutions for (44).

For an algebraic number α of degree $d \geq 2$, what is the smallest possible μ for which (44) will have only finitely many solutions? Combining the discussion above with Dirichlet's theorem (Theorem 12.2), we see that

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

for any $\delta > 0$. In 1908 Thue proved that $\mu \leq \frac{d+2}{2} + \delta$; in 1921 Siegel proved that $\mu \leq \sqrt{2} d + \delta$. Dyson (1947) and Gelfond (1952) proved that $\mu \leq \sqrt{2d} + \delta$. The major breakthrough came with the famous

theorem of Roth (1955) [31], for which he received a Fields medal in 1958.

Theorem 12.6 (Roth). *Let α be an algebraic number. For any $\delta > 0$, there are only finitely many rationals $\frac{p}{q}$ with $\gcd(p, q) = 1$ such that*

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

Remark 12.4. Dirichlet’s theorem shows that Roth’s theorem is best possible, i.e. the exponent on q in the upper bound cannot be improved. Notice also that in case α has degree 2, Lemma 12.3 gives a better result. An outline of the proof of Roth’s theorem can be found in [36]; complete versions of the proof can be found in [35], [10], and [31].

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

Definition 12.2. A set $S \subseteq \mathbb{R}$ is called a *C-set* for a real number $C > 1$ if for any two numbers m, n in S , $m \leq Cn$ and $n \leq Cm$.

Notice for instance that a *C-set* consisting of integers must be finite, although unless we know at least one of its elements, we cannot say anything about its cardinality.

Definition 12.3. A set $S \subseteq \mathbb{R}$ is called a *γ -set* for a real number $\gamma > 1$ if whenever $m, n \in S$ and $m < n$, then $\gamma m \leq n$.

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

Lemma 12.7. *Let $C > 1$ and $\gamma > 1$, and suppose that $S \subseteq \mathbb{R}$ is a (C, γ) -set. Then*

$$(45) \quad |S| \leq 1 + \frac{\log C}{\log \gamma}.$$

Proof. Clearly S is a finite set, so assume

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

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

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

and

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

Hence

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

and (45) follows. □

Definition 12.4. Given $C > 1$, a **window of exponential width C** is an interval of real numbers x of type

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

for some $w > 1$.

We can now use Lemma 12.7 to prove a bound on the number of good rational approximations to a real number α in a window of exponential width C for any $C > 1$. We will say that a rational number $\frac{p}{q}$ is *reduced* if $\gcd(p, q) = 1$.

Lemma 12.8. Let $\alpha \in \mathbb{R}$, $\delta > 0$, and $C > 1$. Let $N_C(\alpha)$ be the number of reduced rational numbers $\frac{p}{q}$ such that

$$(46) \quad \left| \alpha - \frac{p}{q} \right| < \frac{1}{2q^{2+\delta}}$$

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

$$(47) \quad N_C(\alpha) \leq 1 + \frac{\log C}{\log(1 + \delta)}.$$

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

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

for some $w > 1$, hence $x \leq y^C$ and $y \leq x^C$. Now suppose that $\frac{p_1}{q_1} \neq \frac{p_2}{q_2}$ are reduced fractions that satisfy (46) with $q_1 \leq q_2$ in a window of exponential width C . Then

$$\begin{aligned} \frac{1}{q_1 q_2} &\leq \left| \frac{p_1}{q_1} - \frac{p_2}{q_2} \right| = \left| \left(\frac{p_1}{q_1} - \alpha \right) + \left(\alpha - \frac{p_2}{q_2} \right) \right| \\ &\leq \left| \alpha - \frac{p_1}{q_1} \right| + \left| \alpha - \frac{p_2}{q_2} \right| < \frac{1}{2q_1^{2+\delta}} + \frac{1}{2q_2^{2+\delta}} \leq \frac{1}{q_1^{2+\delta}}, \end{aligned}$$

and so

$$q_2 > q_1^{1+\delta}.$$

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

$$\gamma \log q_1 < \log q_2,$$

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

$$\log q_1 \leq C \log q_2, \quad \log q_2 \leq C \log q_1,$$

that is these logarithms also form a C -set, hence they form a (C, γ) -set, and by Lemma 12.7 the cardinality of this set is

$$\leq 1 + \frac{\log C}{\log \gamma} = 1 + \frac{\log C}{\log(1 + \delta)},$$

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

Remark 12.5. Suppose that $1 < A < B$ are given, and suppose that we want to know the number of reduced rational approximations $\frac{p}{q}$ to the real number α with

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

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

$$A = e^{\log A} \leq q \leq B = (e^{\log A})^{\frac{\log B}{\log A}},$$

and so by Lemma 12.8, the number of such approximations is

$$\leq 1 + \frac{\log \left(\frac{\log B}{\log A} \right)}{\log(1 + \delta)}.$$

Definition 12.5. Let $\alpha \in \mathbb{R}$ and let $\delta > 0$. We will call $\frac{p}{q} \in \mathbb{Q}$ a δ -**approximation** to α if $q > 0$, $\gcd(p, q) = 1$, and

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

A method similar to the proof of Lemma 12.8 yields the following result; a proof of this can be found on p. 59 of [36].

Lemma 12.9. *Let $\alpha \in \mathbb{R}$, $\delta > 0$. The number of δ -approximations $\frac{p}{q}$ to α in a window $w \leq q \leq w^C$, where $w \geq 4^{1/\delta}$ is*

$$\leq 1 + \frac{\log 2C}{\log(1 + \delta)}.$$

13. ABSOLUTE VALUES

We now start introducing the basic machinery of absolute values and heights, which we will then use to investigate further questions in Diophantine Approximations.

Definition 13.1. Let K be a field. An **absolute value** on K is a function $|\cdot| : K \rightarrow \mathbb{R}_{\geq 0}$ such that for all $x, y \in K$ we have:

- (1) $|x| \geq 0$ with equality if and only if $x = 0$,
- (2) $|xy| = |x||y|$,
- (3) **Triangle inequality:** $|x + y| \leq |x| + |y|$.

Sometimes (3) can be replaced by the stronger property:

- (4) **Ultrametric inequality:** $|x + y| \leq \max\{|x|, |y|\}$.

If $|\cdot|$ satisfies (1), (2), (3), but fails (4), we say that it is **archimedean** absolute value; if it also satisfies (4), it is called **non-archimedean**.

Here is the most basic example of an absolute value on K : it is called the **trivial** absolute value, and is defined by

$$|x| = \begin{cases} 0 & \text{if } x = 0, \\ 1 & \text{if } x \neq 0. \end{cases}$$

This is the only possible absolute value on a finite field.

We will say that two absolute values $|\cdot|_1$ and $|\cdot|_2$ on K are **equivalent** if there exists $\theta \in \mathbb{R}_{>0}$ such that

$$|x|_1 = |x|_2^\theta$$

for all $x \in K$. In this case we will write $|\cdot|_1 \sim |\cdot|_2$. It is easy to see that an archimedean absolute value cannot be equivalent to a non-archimedean one.

Exercise 13.1. Prove that \sim as defined above is an equivalence relation on the set of all absolute values on K .

Exercise 13.2. Prove that the only absolute value equivalent to the trivial one is itself.

Equivalence classes of nontrivial absolute values on K are called **places**. The set of all places of K will be denoted by $M(K)$. Notice that an absolute value $|\cdot|$ defines a metric on K :

$$(x, y) \rightarrow |x - y|$$

for every $x, y \in K$. Therefore $|\cdot|$ induces a metric topology on K . Moreover, we can talk about the *completion* of K with respect to this topology. K equipped with the metric induced by $|\cdot|$ is a metric space,

we will write $(K, | \cdot |)$ to mean that we are thinking of K as a metric space with respect to this metric. Recall that a metric space $(K, | \cdot |)$ is called **complete** if every Cauchy sequence in K converges to a point in K . The **completion** of $(K, | \cdot |)$ is the set of all equivalence classes of Cauchy sequences on $(K, | \cdot |)$, where two Cauchy sequences $\{a_n\}$ and $\{b_n\}$ are equivalent if

$$\lim_{n \rightarrow \infty} |a_n - b_n| = 0.$$

Notice that $| \cdot |$ is also defined on the completion of $(K, | \cdot |)$, and so this completion also has a metric topology induced by $| \cdot |$. Then $(K, | \cdot |)$ is complete if and only if it is equal to its completion; by “equal” here we mean isometrically isomorphic as fields: it is a well known fact that completion of a field is also a field, where addition and multiplication on Cauchy sequences are defined component-wise.

Exercise 13.3. *Prove that two absolute values $| \cdot |_1$ and $| \cdot |_2$ on K are equivalent if and only if they induce the same topology.*

Notice that for an absolute value $| \cdot |$ on K , $x \rightarrow |x|$ is a homomorphism from the multiplicative group $K^\times = \{x \in K : x \neq 0\}$ to multiplicative group $\mathbb{R}_{>0}$. Therefore:

- (1) $|1| = 1$,
- (2) $|\zeta| = 1$ for every root of unity $\zeta \in K$, i.e. for every $\zeta \in K$ such that $\zeta^n = 1$ for some $n \in \mathbb{Z}_{>0}$,
- (3) $|-x| = |x|$, for all $x \in K^\times$,
- (4) $|x^{-1}| = |x|^{-1}$, for all $x \in K^\times$.

If L/K is an extension of fields and $| \cdot |$ is an absolute value on L , then its restriction to K is an absolute value on K . It is in general possible that $| \cdot |$ is non-trivial on L , but is trivial on K .

We will now demonstrate some standard absolute values on \mathbb{Q} . The first one is the usual absolute value, which we will denote by $| \cdot |_\infty$:

$$|x|_\infty = \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x < 0. \end{cases}$$

Exercise 13.4. *Prove that $| \cdot |_\infty$ is an archimedean absolute value on \mathbb{Q} .*

Notice that $| \cdot |_\infty$ induces the usual metric topology on \mathbb{Q} ; the completion of \mathbb{Q} with respect to this topology is \mathbb{R} . Sometimes we will write \mathbb{Q}_∞ instead of \mathbb{R} to stress this fact.

Now let $p \in \mathbb{Z}$ be a prime, and define the **p -adic** absolute value $| \cdot |_p$ on \mathbb{Q} as follows. For each $n \in \mathbb{Z}$, let

$$|n|_p = p^{-\mu(n)},$$

where $p^{\mu(n)}$ is the largest power of p dividing n , hence $|n|_p \leq 1$ for each $n \in \mathbb{Z}$. Now for each $\frac{m}{n} \in \mathbb{Q}$, let

$$\left| \frac{m}{n} \right|_p = \frac{|m|_p}{|n|_p}.$$

Exercise 13.5. Prove that $|\cdot|_p$ is a non-archimedean absolute value on \mathbb{Q} for each prime $p \in \mathbb{Z}$.

Exercise 13.6. Prove that

$$\mathbb{Z} = \{a \in \mathbb{Q} : |a|_p \leq 1 \ \forall \text{ primes } p \in \mathbb{Z}\}.$$

The topology induced by $|\cdot|_p$ on \mathbb{Q} is called **p -adic topology**; the completion of \mathbb{Q} with respect to this is called the field of **p -adic numbers**, and is denoted by \mathbb{Q}_p . The set

$$\mathbb{Z}_p = \{a \in \mathbb{Q}_p : |a|_p \leq 1\}$$

is a ring, and is called the ring of **p -adic integers**. Exercise 13.6 implies that $\mathbb{Z} \subseteq \mathbb{Z}_p$ for every prime $p \in \mathbb{Z}$. Moreover, if we write \mathcal{P} for the set of all primes in \mathbb{Z} , then

$$\mathbb{Z} = \bigcap_{p \in \mathcal{P}} \mathbb{Z}_p.$$

Theorem 13.1 (Ostrowski, 1935). Any non-trivial absolute value on \mathbb{Q} is equivalent to either $|\cdot|_\infty$ or $|\cdot|_p$ for some $p \in \mathcal{P}$.

Proof. We start with the following easy to prove fact.

Exercise 13.7. An absolute value $|\cdot|$ on \mathbb{Q} is non-archimedean if and only if $|n| \leq 1$ for every $n \in \mathbb{Z}$. Moreover, for any absolute value $|\cdot|$ on \mathbb{Q} there exists $\rho \in \mathbb{R}_{>0}$ such that

$$(48) \quad |n| \leq |n|_\infty^\rho.$$

Now suppose $|\cdot|$ is an absolute value on \mathbb{Q} . We will use Exercise 13.7 throughout this proof, assuming without loss of generality that $\rho = 1$ in (48); indeed, $|\cdot|^{\frac{1}{\rho}}$ is equivalent to $|\cdot|$, so it is not important whether we prove that $|\cdot|^{\frac{1}{\rho}}$ or $|\cdot|$ is equivalent to $|\cdot|_\infty$ or $|\cdot|_p$ for some $p \in \mathcal{P}$.

Let $a, b \in \mathbb{Z}_{>0}$, $a > 1, b > 1$. For any $\nu \in \mathbb{Z}_{>0}$, there exists integers c_0, \dots, c_n with $0 \leq c_i < a$ and $c_n \neq 0$ such that

$$b^\nu = c_0 + c_1 a + \dots + c_n a^n.$$

Notice that by Exercise 13.7 for each $0 \leq i \leq n$,

$$|c_i| \leq |c_i|_\infty \leq |a|_\infty = a.$$

Also notice that

$$a^n \leq c_n a^n \leq b^\nu,$$

and so $n \leq \frac{\nu \log b}{\log a}$. Then

$$\begin{aligned} |b|^\nu = |b^\nu| &\leq \sum_{i=0}^n |c_i| |a|^i \leq (n+1) a \max\{1, |a|\}^n \\ &\leq \left(1 + \frac{\nu \log b}{\log a}\right) a \max\{1, |a|\}^n. \end{aligned}$$

Therefore

$$|b| \leq \left(1 + \frac{\nu \log b}{\log a}\right)^{1/\nu} a^{1/\nu} \max\{1, |a|\}^{\frac{\log b}{\log a}} \rightarrow \max\left\{1, |a|^{\frac{\log b}{\log a}}\right\},$$

as $\nu \rightarrow \infty$, in other words

$$(49) \quad |b| \leq \max\left\{1, |a|^{\frac{\log b}{\log a}}\right\}.$$

Case 1. Assume $|\cdot|$ is archimedean. Then by Exercise 13.7, there exists $b \in \mathbb{Z}$ such that $|b| > 1$. Then by (49), $|a| > 1$ for every $a \in \mathbb{Z}$ except for $-1, 0, 1$. Therefore if $a, b \in \mathbb{Z}$, $a, b > 1$, then

$$|b|^{\frac{1}{\log b}} \leq |a|^{\frac{1}{\log a}} \leq |b|^{\frac{1}{\log b}},$$

and so

$$|b|^{\frac{1}{\log b}} = |a|^{\frac{1}{\log a}}.$$

We have

$$1 < |b| \leq |b|_\infty = b,$$

so $|b| = |b|_\infty^\rho = b^\rho$ for some $0 < \rho \leq 1$, and hence

$$|a| = |b|^{\frac{\log a}{\log b}} = b^{\rho \frac{\log a}{\log b}} = a^\rho = |a|_\infty^\rho.$$

Same way therefore $|\alpha| = |\alpha|_\infty^\rho$ for every $\alpha \in \mathbb{Q}$.

Case 2. Assume $|\cdot|$ is non-archimedean. Then by Exercise 13.7, $|n| \leq 1$ for every $n \in \mathbb{Z}$, and since $|\cdot|$ is non-trivial, there exists $a \in \mathbb{Z}$ such that $|a| < 1$. Let

$$I = \{a \in \mathbb{Z} : |a| < 1\}.$$

Exercise 13.8. Prove that I is a prime ideal in \mathbb{Z} .

Therefore there exists a prime $p \in \mathbb{Z}$ such that $I = p\mathbb{Z}$. Let $0 \neq \alpha \in \mathbb{Q}$. Write

$$\alpha = p^r \frac{x}{y}$$

with $x, y \in \mathbb{Z}$ such that $p \nmid xy$. Then $x, y \notin I$, hence

$$|x| = |y| = 1,$$

and so

$$|\alpha| = |p^r| = |p|^r.$$

Since $p \in I$, $|p| < 1$, so $|p| = p^{-s}$ for some $s > 0$. Then

$$|\alpha| = p^{-rs} = |r|_p^s.$$

We have shown that $|\cdot|$ must be equivalent to either $|\cdot|_\infty$ or $|\cdot|_p$ for some prime p . This completes the proof. \square

Therefore we can write

$$M(\mathbb{Q}) = \{\infty\} \cup \mathcal{P},$$

this way indexing the archimedean place by ∞ , and non-archimedean places by p for each $p \in \mathcal{P}$.

Theorem 13.2 (Artin - Whaples Product Formula). *If $0 \neq a \in \mathbb{Q}$, then*

$$|a|_\infty \prod_{p \in \mathcal{P}} |a|_p = 1.$$

Proof. Exercise. \square

Next we discuss absolute values on a number field K . If $|\cdot|$ is an absolute value on K , its restriction to \mathbb{Q} is an absolute value on \mathbb{Q} , and so must belong to either ∞ or one of the p -adic places on \mathbb{Q} . Hence absolute values on K are precisely extensions of those on \mathbb{Q} . If $v \in M(K)$, we will write $|\cdot|_v$ for an absolute value that represents it. We know that $|\cdot|_v$ extends either $|\cdot|_\infty$ or $|\cdot|_p$ for some $p \in \mathcal{P}$, and we say that v **lies over** ∞ or p respectively; we denote it by writing $v|\infty$ or $v|p$. The place $v \in M(K)$ is archimedean if and only if $v|\infty$. Sometimes we will write $v \nmid \infty$ to mean that v is non-archimedean, i.e. lies over some p -adic place of \mathbb{Q} . For each place $u \in M(\mathbb{Q})$ there may be more than one place $v \in M(K)$ such that $v|u$, however each place $v \in M(K)$ lies over precisely one place $u \in M(\mathbb{Q})$.

First we describe all archimedean places of K . As in section 4 above, let $\sigma_1, \dots, \sigma_r$ be real embeddings of K , and $\tau_1, \bar{\tau}_1, \dots, \tau_s, \bar{\tau}_s$ conjugate pairs of complex embeddings, then

$$r + 2s = d = [K : \mathbb{Q}].$$

Notice that since $\mathbb{Q}_\infty = \mathbb{R} \subset \mathbb{C}$, the absolute value $|\cdot|_\infty$ is defined on \mathbb{R} and on \mathbb{C} . Also, for each $a \in K$

$$\sigma_i(a) \in \mathbb{R}, \tau_j(a), \bar{\tau}_j(a) \in \mathbb{C}$$

for each $1 \leq i \leq r$ and $1 \leq j \leq s$. If ρ is one of these embeddings, then we define an absolute value $|\cdot|_\rho$ on K by

$$|a|_\rho = |\rho(a)|_\infty.$$

It is easy to notice that if $|\cdot|_{\tau_j} = |\cdot|_{\bar{\tau}_j}$ for each $1 \leq j \leq s$. However, the absolute values

$$|\cdot|_{\sigma_1}, \dots, |\cdot|_{\sigma_r}, |\cdot|_{\tau_1}, \dots, |\cdot|_{\tau_s}$$

are not equivalent to each other. These represent all the archimedean places of K . For each $v \in M(K)$, we will write K_v for the completion of K at v . If $v|u$ for some $u \in M(\mathbb{Q})$, then K_v/\mathbb{Q}_u is an extension of fields, and we will define the **local degree** of K at v to be the degree of this extension, and denote it by

$$d_v = [K_v : \mathbb{Q}_u].$$

We will also write sometimes \mathbb{Q}_v where $v \in M(K)$ to mean \mathbb{Q}_u , where $u \in M(\mathbb{Q})$ is the unique place over which v lies. Notice that if $v \in M(K)$ is archimedean, then K_v is either \mathbb{R} or \mathbb{C} , depending on whether v is real or complex, i.e. corresponds to a real or to a complex embedding. Therefore, for each $v|\infty$

$$d_v = [K_v : \mathbb{Q}_\infty] = [K_v : \mathbb{R}] = \begin{cases} 1 & \text{if } v \text{ is real} \\ 2 & \text{if } v \text{ is complex.} \end{cases}$$

Therefore

$$\sum_{v|\infty} d_v = r + 2s = d.$$

Next we describe non-archimedean places of K . Let p be a prime in \mathbb{Z} , so that $(p) = p\mathbb{Z}$ is a prime ideal in \mathbb{Z} . Recall that \mathcal{O}_K , the ring of algebraic integers of K , is a Dedekind domain, which means that there is unique factorization into prime ideals in \mathcal{O}_K . Notice that $\mathbb{Z} \in \mathcal{O}_K$, and so $p\mathcal{O}_K$ is an ideal in \mathcal{O}_K , although it may no longer be prime. Then there exist prime ideals P_1, \dots, P_k and positive integers e_1, \dots, e_k such that

$$p\mathcal{O}_K = P_1^{e_1} \dots P_k^{e_k},$$

and $\sum_{i=1}^k e_i = d$; each such e_i is called the **ramification degree** of P_i over p . First we define $|0|_{P_i} = 0$. Now let $0 \neq a \in \mathcal{O}_K$, then for each P_i , $1 \leq i \leq k$, define

$$\text{ord}_{P_i} a = \max\{j \in \mathbb{Z} : a \in P_i^j\},$$

and let

$$|a|_{P_i} = p^{-\frac{\text{ord}_{P_i} a}{e_i}}.$$

The number $\text{ord}_{P_i} a$ is well-defined due to unique factorization of ideals into powers of prime ideals: it is precisely the power to which P_i divides $a\mathcal{O}_K$. Notice that K is the field of fractions of \mathcal{O}_K , i.e.

$$K = \left\{ \frac{a}{b} : a, b \in \mathcal{O}_K \right\}.$$

Then for each $\alpha = \frac{a}{b} \in K$ with $a, b \in \mathcal{O}_K$, define

$$(50) \quad |\alpha|_{P_i} = \frac{|a|_{P_i}}{|b|_{P_i}}.$$

Exercise 13.9. *Prove that (50) defines an absolute value on K , which restricts to the usual p -adic absolute value on \mathbb{Q} .*

Hence for each prime p in \mathbb{Z} , we defined absolute values lying over it; these are all the non-archimedean places of K . Suppose $v \in M(K)$ lies over p , and P_i is the corresponding prime ideal of \mathcal{O}_K with ramification degree e_i over p . In a Dedekind domain every non-zero prime ideal is maximal, hence P_i is a maximal ideal, and so \mathcal{O}_K/P_i is a field; in fact, it is a finite field of characteristic p , meaning that

$$|\mathcal{O}_K/P_i| = p^{f_i},$$

for some $f_i \in \mathbb{Z}_{>0}$. This f_i is called the **inertia degree** of P_i over p . Its significance for our purposes is that the local degree $d_v = [K_v : \mathbb{Q}_p]$ is equal to $e_i f_i$. A result from algebraic number theory implies that if P_1, \dots, P_k are prime ideals in \mathcal{O}_K lying over a rational prime p with respective ramification degrees e_1, \dots, e_k and ramification degrees f_1, \dots, f_k , then

$$\sum_{i=1}^k e_i f_i = d.$$

In particular this means that

$$\sum_{v|u} d_v = d$$

is true for any $u \in M(\mathbb{Q})$. The Artin - Whaples product formula works over a number field in the same way as over \mathbb{Q} .

Theorem 13.3. *If $0 \neq a \in K$, then*

$$\prod_{v \in M(K)} |a|_v^{d_v} = 1.$$

Example. Let $K = \mathbb{Q}(\sqrt{2})$, then $d = 2$. Since K is totally real, there are no complex embeddings. Hence if $v \in M(K)$ is archimedean, then $K_v = \mathbb{R}$, and so $d_v = 1$. Since

$$\sum_{v|\infty} d_v = 2,$$

K must have two archimedean places. These are precisely the places corresponding to embeddings $\sigma_1, \sigma_2 : K \rightarrow \mathbb{R}$, given by

$$\sigma_1(\sqrt{2}) = \sqrt{2}, \quad \sigma_2(\sqrt{2}) = -\sqrt{2},$$

and fixing \mathbb{Q} , hence σ_1 is the identity. Let v_1, v_2 be the archimedean places corresponding to embeddings σ_1, σ_2 respectively. Notice that for every $\alpha \in K$, there exist $a, b \in \mathbb{Q}$ such that $\alpha = a + b\sqrt{2}$, hence

$$|\alpha|_{v_1} = |\sigma_1(a + b\sqrt{2})|_\infty = |a + b\sqrt{2}|_\infty,$$

and

$$|\alpha|_{v_2} = |\sigma_2(a + b\sqrt{2})|_\infty = |a - b\sqrt{2}|_\infty.$$

Now let us look at non-archimedean places of K . Consider for instance all places $v \in M(K)$ lying over 7. Notice that

$$7 = (3 + \sqrt{2})(3 - \sqrt{2}),$$

therefore the ideal $7\mathcal{O}_K$ no longer prime in \mathcal{O}_K splits as the product of these two prime ideals:

$$7\mathcal{O}_K = P_1P_2,$$

where $P_1 = (3 + \sqrt{2})\mathcal{O}_K$ and $P_2 = (3 - \sqrt{2})\mathcal{O}_K$. This means that there are two places lying over 7, corresponding to P_1 and P_2 , call them u_1 and u_2 respectively. Then $d_{u_1} = d_{u_2} = 1$. Notice for instance that

$$3 + \sqrt{2} \in P_1, \quad 3 + \sqrt{2} \notin P_1^2, \quad 3 + \sqrt{2} \notin P_2,$$

$$3 - \sqrt{2} \in P_2, \quad 3 - \sqrt{2} \notin P_2^2, \quad 3 - \sqrt{2} \notin P_1,$$

hence

$$|3 + \sqrt{2}|_{u_1} = 7^{-1}, \quad |3 - \sqrt{2}|_{u_1} = 7^0,$$

$$|3 + \sqrt{2}|_{u_2} = 7^0, \quad |3 - \sqrt{2}|_{u_2} = 7^{-1}.$$

Recall that prime ideals in \mathcal{O}_K are maximal. This implies that $3 \pm \sqrt{2}$ are not contained in any other prime ideal of \mathcal{O}_K , hence for every place $v \in M(K)$ which is not equal to v_1, v_2, u_1 , or u_2 , $|3 \pm \sqrt{2}|_v = 1$. Hence

$$\prod_{v \in M(K)} |3 \pm \sqrt{2}|_v = |3 + \sqrt{2}|_\infty |3 - \sqrt{2}|_\infty 7^{-1} = 1.$$

This is a demonstration of the product formula at work.

Remark 13.1. The same construction of absolute values as described in this section can be carried out for any field extension of number fields L/K . In this case, we would replace the ground field \mathbb{Q} with K , and talk about places of L lying over places of K in the same precise manner. We will assume this more general construction in the next section.

14. HEIGHTS

In this section we introduce *height functions*, which serve as the main tool used to measure arithmetic complexity. We use the notation of the previous section, in particular let K be a number field of degree d over \mathbb{Q} , and let $M(K)$ be its set of places. Let $N \geq 2$ be an integer. For each place v of K we define a **local height** H_v for each vector $\mathbf{x} \in K_v^N$ by

$$H_v(\mathbf{x}) = \begin{cases} \left(\sum_{i=1}^N |x_i|_v^2 \right)^{\frac{1}{2}} & \text{if } v|\infty, \\ \max_{1 \leq i \leq N} |x_i|_v & \text{if } v \nmid \infty. \end{cases}$$

Then for each $\mathbf{0} \neq \mathbf{x} \in K^N$, define the **global height** H_K by

$$(51) \quad H_K(\mathbf{x}) = \prod_{v \in M(K)} H_v(\mathbf{x})^{d_v}.$$

Notice that for each $\mathbf{0} \neq \mathbf{x} \in K^N$, $H_v(\mathbf{x}) = 1$ for all but finitely many places v of K , hence the product in (51) is actually finite, therefore convergent, meaning that H_K is well-defined. Also notice that if $0 \neq \alpha \in K$ and $\mathbf{0} \neq \mathbf{x} \in K^N$, then

$$(52) \quad \begin{aligned} H_K(\alpha \mathbf{x}) &= \prod_{v \in M(K)} |\alpha|_v^{d_v} H_v(\mathbf{x})^{d_v} \\ &= \left(\prod_{v \in M(K)} |\alpha|_v^{d_v} \right) \prod_{v \in M(K)} H_v(\mathbf{x})^{d_v} = H_K(\mathbf{x}) \end{aligned}$$

by the product formula. This means that H_K is a *homogeneous* function, and so is *projectively defined*. Indeed, define an equivalence relation on $K^N \setminus \{\mathbf{0}\}$ by writing $\mathbf{x} \sim \mathbf{y}$ whenever $\mathbf{x} = \alpha \mathbf{y}$ for some $0 \neq \alpha \in K$. It is easy to check that this indeed is an equivalence relation, and we write $[x_1 : \cdots : x_N]$ for the equivalence class of the vector $\mathbf{x} = (x_1, \dots, x_N) \in K^N$, which is called the **projective point** corresponding to \mathbf{x} . The space of all projective points on K^N is called the $(N-1)$ -dimensional **projective space** over K , i.e.

$$\mathbb{P}^{N-1}(K) = \{[x_1 : \cdots : x_N] : (x_1, \dots, x_N) \in K^N \setminus \{\mathbf{0}\}\}.$$

Notice that this is precisely the space of all lines through the origin in K^N , i.e. the space of 1-dimensional subspaces of K^N . This is the simplest example of the more general construction of Grassmannian that we will encounter later. Then (52) implies that H_K is well-defined on $\mathbb{P}^{N-1}(K)$, i.e. it can be viewed as a function $H_K : \mathbb{P}^{N-1}(K) \rightarrow \mathbb{R}_{>0}$.

Notice that the definition of H_K depends on K . Let L be an extension of K of degree e , hence $[L : \mathbb{Q}] = de$. For each place $v \in M(L)$, we will write $e_v = [L_v : K_v]$, hence $[L_v : \mathbb{Q}_v] = d_v e_v$. Also notice that

$$\sum_{v \in M(L), v|u} e_v = e$$

for each place $u \in M(K)$. Suppose that $\mathbf{0} \neq \mathbf{x} \in K^N$, then

$$H_L(\mathbf{x}) = \prod_{v \in M(L)} H_v(\mathbf{x})^{d_v e_v} = \prod_{u \in M(K)} \prod_{v \in M(L), v|u} H_v(\mathbf{x})^{d_u e_v},$$

but since $\mathbf{x} \in K^N$, $H_v(\mathbf{x}) = H_{v'}(\mathbf{x})$ whenever $v, v' \in M(L)$ lie over the same place $u \in M(K)$. Hence:

$$H_L(\mathbf{x}) = \prod_{u \in M(K)} H_u(\mathbf{x})^{d_u \sum_{v \in M(L), v|u} e_v} = \prod_{u \in M(K)} H_u(\mathbf{x})^{d_u e} = H_K(\mathbf{x})^e.$$

This suggests that if we want a height function that does not depend on the field of definition, we may want to introduce the normalizing exponent $\frac{1}{[K:\mathbb{Q}]}$.

Definition 14.1. Let $\overline{\mathbb{Q}}$ be the algebraic closure of \mathbb{Q} , i.e. the field of all algebraic numbers. Define the **absolute height** $H : \overline{\mathbb{Q}}^N \setminus \{\mathbf{0}\} \rightarrow \mathbb{R}_{>0}$ by

$$H(\mathbf{x}) = H_K(\mathbf{x})^{\frac{1}{[K:\mathbb{Q}]}}$$

for every $\mathbf{0} \neq \mathbf{x} \in \overline{\mathbb{Q}}^N$, where K is any number field containing the coordinates of \mathbf{x} . By the discussion above, H does not depend on the choice of this number field. Once again, notice that H is projectively defined. We will also adopt a convention that $H(\mathbf{0}) = 1$.

We also define the **inhomogeneous height** $h_K : K^N \rightarrow \mathbb{R}_{>0}$ by

$$h_K(\mathbf{x}) = H_K(1, \mathbf{x}),$$

for every $\mathbf{x} \in K^N$, and the **absolute inhomogeneous height** $h : \overline{\mathbb{Q}}^N \rightarrow \mathbb{R}_{>0}$ by

$$h(\mathbf{x}) = h_K(\mathbf{x})^{\frac{1}{[K:\mathbb{Q}]}}$$

for every $\mathbf{x} \in \overline{\mathbb{Q}}^N$, where K is any number field containing the coordinates of \mathbf{x} . Notice that h_K and h are no longer projectively defined, i.e. if $\alpha \in \overline{\mathbb{Q}}$, then $h(\alpha \mathbf{x})$ is not necessarily equal to $h(\mathbf{x})$. Also notice that for every $\mathbf{x} \in \overline{\mathbb{Q}}^N$,

$$H(\mathbf{x}) \leq h(\mathbf{x}).$$

For any algebraic number $\alpha \in \overline{\mathbb{Q}}$, we define its **Weil height** to be

$$h(\alpha) = H(1, \alpha).$$

We now briefly outline the basic properties of heights.

Exercise 14.1. First suppose that $\mathbf{x} \in \mathbb{Z}$ is such that

$$\gcd(x_1, \dots, x_N) = 1.$$

Prove that

$$H(\mathbf{x}) = \|\mathbf{x}\|_2 = (x_1^2 + \dots + x_N^2)^{\frac{1}{2}},$$

i.e. height on \mathbb{Z}^N is just the Euclidean norm.

Next assume that $0 \neq x_0 \in \mathbb{Z}$, and

$$\mathbf{x} = \left(\frac{x_1}{x_0}, \dots, \frac{x_N}{x_0} \right) \in \mathbb{Q}^N,$$

is such that $\gcd(x_0, x_1, \dots, x_N) = 1$. Then prove that

$$h(\mathbf{x}) = (x_0^2 + x_1^2 + \dots + x_N^2)^{\frac{1}{2}},$$

i.e. the inhomogeneous height on \mathbb{Q}^N is just the Euclidean norm of the integral vector (x_0, x_1, \dots, x_N) .

Exercise 14.2. Prove that if $m_1, \dots, m_k \in \mathbb{Z}$, and $\mathbf{x}_1, \dots, \mathbf{x}_k \in \overline{\mathbb{Q}}^N$, then

$$h\left(\sum_{i=1}^k m_i \mathbf{x}_i\right) \leq \left(\sum_{i=1}^k m_i^2\right)^{\frac{1}{2}} \prod_{i=1}^k h(\mathbf{x}_i).$$

In particular, this means that if $\alpha_1, \dots, \alpha_k \in \overline{\mathbb{Q}}$, then

$$h\left(\sum_{i=1}^k m_i \alpha_i\right) \leq \left(\sum_{i=1}^k m_i^2\right)^{\frac{1}{2}} \prod_{i=1}^k h(\alpha_i).$$

Prove also that for any $\alpha, \beta \in \overline{\mathbb{Q}}$,

$$h(\alpha\beta) \leq h(\alpha)h(\beta).$$

Exercise 14.3. Suppose that K and L are isomorphic number fields with $\sigma : K \rightarrow L$ an isomorphism. We also write σ for the isomorphism it induces from K^N to L^N for each integer $N \geq 1$. Prove that

$$H(\sigma(\mathbf{x})) = H(\mathbf{x})$$

for each $\mathbf{x} \in K$. Hence conjugate vectors have the same height. Notice in particular that this implies that conjugate algebraic numbers have the same height.

The notion of height also extends to polynomials. In particular, if F is a polynomial with coefficients $a_1, \dots, a_N \in \overline{\mathbb{Q}}$, then we define

$$H(F) = H(a_1, \dots, a_N).$$

Lemma 14.1. *Let $P(X), Q(X) \in \overline{\mathbb{Q}}[X]$ be polynomials in one variable with coefficients in $\overline{\mathbb{Q}}$ of degrees n_1, n_2 respectively, and let $n = \min\{n_1, n_2\}$. Then*

$$H(PQ) \leq \sqrt{n+1} H(P)H(Q).$$

Proof. Let K be a number field containing coefficients of P and Q , and suppose it has degree d over \mathbb{Q} . It is easy to observe that for every $v \in M(K)$ such that $v \nmid \infty$,

$$H_v(PQ) = H_v(P)H_v(Q),$$

where these are precisely the local heights of corresponding coefficient vectors.

Exercise 14.4. *Let $v \in M(K)$, $v \mid \infty$. Use Cauchy's inequality to prove that*

$$H_v(PQ) \leq \sqrt{n+1} H_v(P)H_v(Q).$$

Then

$$\begin{aligned} H(PQ) &= \prod_{v \in M(K)} H_v(PQ)^{\frac{d_v}{d}} \\ &\leq \prod_{v \nmid \infty} (H_v(P)H_v(Q))^{\frac{d_v}{d}} \prod_{v \mid \infty} \left(|n+1|_v^{\frac{1}{2}} H_v(P)H_v(Q) \right)^{\frac{d_v}{d}} \\ &= H(P)H(Q) \prod_{v \mid \infty} |n+1|_v^{\frac{d_v}{2d}} \\ &= \left(\sqrt{n+1} \right)^{\frac{\sum_{v \mid \infty} d_v}{d}} H(P)H(Q) = \sqrt{n+1} H(P)H(Q). \end{aligned}$$

This completes the proof. □

Corollary 14.2. *Suppose that*

$$P(X) = a_d(X - \alpha_1) \dots (X - \alpha_d),$$

where $a_d, \alpha_1, \dots, \alpha_d \in \overline{\mathbb{Q}}$. Then

$$(53) \quad H(P) \leq 2^{\frac{d-1}{2}} h(\alpha_1) \dots h(\alpha_d).$$

Proof. Notice that here we can view $P(X)$ as a product of d linear polynomials in one variable, hence applying Lemma 14.1 $d-1$ times yields (53). □

For a vector $\mathbf{x} \in \overline{\mathbb{Q}}^N$, we define its **degree** to be

$$\deg(\mathbf{x}) = [\mathbb{Q}(x_1, \dots, x_N) : \mathbb{Q}].$$

Also, for a projective point $[\mathbf{x}]$ we write $\deg([\mathbf{x}])$ to mean the minimum of $\deg(\mathbf{x})$ taken over all representatives of $[\mathbf{x}]$. We are now ready to prove the fundamental property of heights.

Theorem 14.3 (Northcott, 1949). *Let N, d, B be positive integers. Then the set*

$$S_N(B, d) = \{[\mathbf{x}] \in \mathbb{P}^{N-1}(\overline{\mathbb{Q}}) : \deg([\mathbf{x}]) \leq d, H(\mathbf{x}) \leq B\}$$

is finite.

Proof. If $\mathbf{x} = (x_1, x_2, \dots, x_N) \in \overline{\mathbb{Q}}^N$ with $x_i \neq 0$ for some $1 \leq i \leq N$, then $H(\mathbf{x}) = H\left(\frac{\mathbf{x}}{x_i}\right)$. Therefore we can always choose a representative \mathbf{x} of $[\mathbf{x}] \in \mathbb{P}^{N-1}(\overline{\mathbb{Q}})$ with one coordinate equal to 1. Without loss of generality assume $\mathbf{x} = (1, x_2, \dots, x_N) \in \overline{\mathbb{Q}}^N$, then

$$H(\mathbf{x}) \geq H(1, x_i) = h(x_i), \quad \forall 2 \leq i \leq N.$$

Therefore it suffices to prove that the set

$$S(B, d) = \{\alpha \in \overline{\mathbb{Q}} : \deg(\alpha) \leq d, h(\alpha) \leq B\}$$

is finite. Notice that if $\alpha \in S(B, d)$, then it must be a root of a monic polynomial with rational coefficients of degree at most d

$$P(X) = (X - \alpha_1)(X - \alpha_2) \dots (X - \alpha_d),$$

where $\alpha = \alpha_1, \alpha_2, \dots, \alpha_d$ are conjugates of α . By Exercise 14.3, $h(\alpha) = h(\alpha_i)$ for every $1 \leq i \leq d$, and so $h(\alpha_i) \leq B$ for all $1 \leq i \leq d$. By Corollary 14.2,

$$(54) \quad H(P) \leq 2^{\frac{d-1}{2}} h(\alpha_1) \dots h(\alpha_d) \leq 2^{\frac{d-1}{2}} B^d.$$

Since $P(x)$ is monic, let $(\frac{m_0}{m}, \dots, \frac{m_{d-1}}{m}, 1) \in \mathbb{Q}$ be the coefficient vector of P , written in such a way that $\gcd(m, m_0, \dots, m_{d-1}) = 1$. Then by Exercise 14.1,

$$H(P) = \sqrt{m^2 + m_0^2 + \dots + m_{d-1}^2} = \|\mathbf{m}\|_2,$$

where $\mathbf{m} = (m, m_0, \dots, m_{d-1}) \in \mathbb{Z}^{d+2}$, and $\|\cdot\|_2$ stands for the Euclidean norm, as usual. It is now easy to see that there are only finitely many integral vectors \mathbf{m} with $\|\mathbf{m}\|_2 \leq 2^{\frac{d-1}{2}} B^d$, and so there are only finitely many polynomials P satisfying (54). This means that $S(B, d)$ must be finite, and so completes the proof. \square

Remark 14.1. The cardinality of $S_N(B, d)$ has been investigated by various authors, starting with a result of Schanuel in 1979. More recently there were upper and lower bounds produced by Schmidt, Thunder,

Masser, and Vaaler, among others, however there still is no known asymptotic formula for $|S_N(B, d)|$.

Next we will show how the notion of height can be extended to subspaces of K^N . Let $V \subseteq K^N$ be an L -dimensional subspace, $1 \leq L \leq N$. Let $\mathbf{x}_1, \dots, \mathbf{x}_L$ be a basis for V , and write $X = (\mathbf{x}_1 \ \dots \ \mathbf{x}_L)$ for the corresponding $N \times L$ basis matrix. Let \mathcal{I} be the set of subsets of $\{1, \dots, N\}$ of cardinality L , then

$$|\mathcal{I}| = \binom{N}{L}.$$

For each $I \in \mathcal{I}$, let X_I be the $L \times L$ submatrix of X whose rows are indexed by elements of I . We introduce lexicographic ordering on elements of \mathcal{I} , and write

$$\mathcal{I} = \left\{ I_1, \dots, I_{\binom{N}{L}} \right\}$$

with respect to that order. Then define a vector of **Grassmann coordinates** of V with respect to the basis $\mathbf{x}_1, \dots, \mathbf{x}_L$ to be

$$g(X) = \left(\det(X_{I_1}), \dots, \det\left(X_{I_{\binom{N}{L}}}\right) \right) \in K^{\binom{N}{L}}.$$

Suppose $\mathbf{y}_1, \dots, \mathbf{y}_L$ is a different basis for V , and write Y for the corresponding basis matrix. Then there exists a matrix $U \in GL_L(K)$ such that

$$Y = XU,$$

and so it is easy to see that

$$g(Y) = \det(U)g(X).$$

As before, we write $[g(X)]$ for the projective point in $\mathbb{P}^{\binom{N}{L}-1}$ represented by the vector $g(X)$, hence $[g(X)] = [g(Y)]$, and so we denote this projective point $[g(V)]$ to indicate that it does not depend on the choice of the basis. Define

$$\mathbb{G}_N^L(K) = \{[g(V)] : V \subseteq K^N, \dim_K(V) = L\}.$$

$\mathbb{G}_N^L(K)$ is called the $\binom{N}{L}$ -**Grassmann component** of K^N , and this is the projective space whose points correspond to L -dimensional subspaces of K^N . Notice that this is a generalization of the projective space $\mathbb{P}^{N-1}(K)$, which can be thought of as the space of one-dimensional subspaces of K^N . This is perhaps the simplest example of a parameter space, i.e. of a general type of objects in algebraic geometry which are called *moduli spaces*.

Using this notation, we can now define height of an L -dimensional subspace V of K^N by

$$H(V) = H(g(V)).$$

Of course, this works in precisely the same manner for subspaces of $\overline{\mathbb{Q}}^N$. Height can also be defined for more general objects, such as algebraic varieties and intersection cycles; this is done in a manner similar in spirit to the simplest case of linear varieties (namely vector subspaces) that we considered here, namely by parametrizing these objects in an appropriate manner. This, however, is more in the realm of arithmetic geometry, and out of the scope of these notes.

For the rest of these notes we will be concerned with various applications of height function machinery to problems in Diophantine approximations. We start here by very briefly going back to Roth's theorem. Recall that for a positive real number δ , a δ -approximation to an algebraic number α is a rational number $\frac{p}{q}$ with $q > 0$, $\gcd(p, q) = 1$, and

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

Then Roth's theorem (Theorem 12.6) can be restated by saying that any algebraic number α has only finitely many δ -approximations for each $\delta > 0$. A natural question to ask is how many δ -approximations are there for a fixed algebraic α and a fixed $\delta > 0$? Recall that in section 12 we were counting the number of δ - and related approximations to α in fixed intervals and windows of exponential width. We can now state a result on the overall number of δ -approximations. There are various bounds produced by Davenport, Roth, Luckhardt, Mueller, and Schmidt, among others. A version of the following theorem is due to Bombieri and Van der Poorten.

Theorem 14.4. *Let α be an algebraic number of degree $d \geq 3$, and let $0 < \delta < 1$. Then the number of δ -approximations to α is less than*

$$(55) \quad \frac{\log_+ \log h(\alpha)}{\log(1 + \delta)} + c(d, \delta),$$

where $\log_+ a = \max\{0, \log a\}$, and

$$c(d, \delta) = \frac{10^8}{\delta^5} (\log 2d)^2 \log \left(\left(\frac{50}{\delta} \right)^2 \log 2d \right).$$

The first term in (55) is best possible, however $c(d, \delta)$ can likely be improved; see [36] for many further details on this subject.

15. LEHMER'S PROBLEM AND MAHLER'S MEASURE

Notice that the bound on the number of δ -approximations to an algebraic number α in Theorem 14.4 depends on the height of α ; in particular, the smaller $h(\alpha)$ is, the fewer such approximations are there. This underlines the meaning of $h(\alpha)$ in the following sense. Height measures arithmetic complexity of an algebraic number, incorporating together its size (i.e. archimedean norms) with its divisibility properties (i.e. non-archimedean norms). Hence, roughly speaking, the larger height is, the more “complicated” the algebraic number is, the “closer” it is to transcendental numbers. We know, on the other hand, that transcendental numbers are better approximable than algebraic numbers, so the dependence of the number of δ -approximations on the height makes sense. But how small can $h(\alpha)$ be? In order to investigate this question, it will be convenient to use a different height h_1 on algebraic numbers, which is closely related to h . For each $\alpha \in \overline{\mathbb{Q}}$, define

$$h_1(\alpha) = \prod_{v \in M(K)} \max\{1, |\alpha|_v\}^{\frac{d_v}{d}},$$

where $K = \mathbb{Q}[\alpha]$, $d = [K : \mathbb{Q}]$, and $d_v = [K_v : \mathbb{Q}_v]$.

Exercise 15.1. *Prove that for every $\alpha \in \overline{\mathbb{Q}}$,*

$$h_1(\alpha) \leq h(\alpha) \leq \sqrt{2} h_1(\alpha).$$

Then instead of investigating lower bounds for $h(\alpha)$ we will talk about lower bounds for $h_1(\alpha)$. It is easy to see that

$$h_1(\alpha) \geq 1$$

with equality if and only if α is either 0 or a root of unity. So suppose that $\alpha \in \overline{\mathbb{Q}}$ is of fixed degree d , and $h_1(\alpha) > 1$. Then how small can $h_1(\alpha)$ be? In other words, is there a gap in values of $h_1(\alpha)$, or is it continuous? In this direction we state a famous conjecture of D. H. Lehmer, dating back to 1932.

Conjecture 15.1 (Lehmer). *There exists an absolute constant $C \in \mathbb{R}_{>1}$ such that for any algebraic number α of degree d which is not 0 or a root of unity, we have*

$$h_1(\alpha) \geq C^{\frac{1}{d}}.$$

In this section we will review some of the material required to understand this outstanding conjecture and its significance. For more detailed information on the material of this section see [37], which is an excellent account of this fascinating subject.

Let $\alpha_1, \dots, \alpha_d$ be conjugate algebraic numbers of degree d , and let

$$f(x) = a_d \prod_{i=1}^d (x - \alpha_i) = \sum_{i=0}^d a_i x^i \in \mathbb{Z}[x]$$

be their minimal polynomial. We define the **Mahler's measure** of $f(x)$ to be

$$(56) \quad M(f) = |a_d|_\infty \prod_{i=1}^d \max\{1, |\alpha_i|_\infty\}.$$

Exercise 15.2. Let $\alpha \in \overline{\mathbb{Q}}$ be of degree d , and let $f(x) \in \mathbb{Z}[x]$ be its minimal polynomial. Prove that

$$h_1(\alpha) = M(f)^{\frac{1}{d}}.$$

Using Exercise 15.2, we can restate Lehmer's conjecture in terms of Mahler's measure, which incidentally is how it was originally stated.

Conjecture 15.2 (Lehmer). *There exists an absolute constant $C \in \mathbb{R}_{>1}$ such that for any polynomial $f(x) \in \mathbb{Z}[x]$ which is not a product of cyclotomics and a power of x , we have*

$$M(f) \geq C.$$

Moreover,

$$C = M(g) = 1.1762808\dots,$$

where

$$g(x) = x^{10} + x^9 - x^7 - x^6 - x^5 - x^4 - x^3 + x + 1$$

is the so-called **Lehmer's polynomial**.

We can think of Mahler's measure as another height function defined on polynomials in $\mathbb{Z}[x]$; in particular it satisfies Northcott's finiteness property. More specifically, let $B, d \in \mathbb{R}_{\geq 1}$, and consider the set

$$S(B, d) = \{f(x) \in \mathbb{Z}[x] : \deg(f) \leq d, M(f) \leq B\}.$$

Northcott's theorem (Theorem 14.3) implies that it is finite. Indeed, if $f(x) \in S(B, d)$, then its roots $\alpha_1, \dots, \alpha_d$ must be in the set

$$S'(B, d) = \{\alpha \in \overline{\mathbb{Q}} : \deg(\alpha) \leq d, h(\alpha) \leq B^{\frac{1}{d}} \leq B\},$$

which is finite by Theorem 14.3.

The definition of Mahler's measure immediately implies a few nice properties that it satisfies. First of all, $M(f) = 1$ if and only if $f(x)$ is a product of cyclotomic polynomials and x^n for some $n \in \mathbb{Z}_{>0}$. Also

notice that unlike the regular height H on polynomials, which only satisfies Lemma 14.1, M is a multiplicative function, i.e.

$$M(fq) = M(f)M(q),$$

for any two polynomials $f(x), q(x)$. On the other hand, Mahler proved that $M(f)$ is comparable to $H(f)$, specifically

$$(57) \quad H(f) \ll M(f) \ll H(f).$$

Notice that the definition of Mahler's measure automatically extends to polynomials in $\mathbb{C}[x]$. Although we can no longer think of it as a height function, since the roots of a polynomial in $\mathbb{C}[x]$ are not necessarily algebraic, the multiplicative property remains true. Mahler's measure can also be represented by an integral formula by an application of the classical Jensen's formula from complex analysis.

Theorem 15.1 (Jensen's formula). *For any $\alpha \in \mathbb{C}$,*

$$e^{\int_0^1 \log|\alpha - e^{2\pi i\theta}| d\theta} = \max\{1, |\alpha|\},$$

where $|\cdot|$ stands for the regular $|\cdot|_\infty$ absolute value on \mathbb{C} .

This is a standard theorem, a proof of which can be found for instance in [37], Lemma 1.9. An immediate application of this is the following result, which was originally proved by Mahler; in fact, this along with (57) is the reason why $M(f)$ is called Mahler's measure.

Corollary 15.2. *For any non-zero $f(x) \in \mathbb{C}[x]$,*

$$M(f) = e^{\int_0^1 \log|f(e^{2\pi i\theta})| d\theta}.$$

Proof. Let $d = \deg(f)$, then there exist $a_d, \alpha_1, \dots, \alpha_d \in \mathbb{C}$ such that

$$f(x) = a_d \prod_{j=1}^d (x - \alpha_j),$$

and so

$$\log|f(e^{2\pi i\theta})| = \log|a_d| + \sum_{j=1}^d \log|\alpha_j - e^{2\pi i\theta}|.$$

Therefore, by Theorem 15.1

$$\begin{aligned} e^{\int_0^1 \log|f(e^{2\pi i\theta})| d\theta} &= |a_d| \prod_{j=1}^d e^{\int_0^1 \log|\alpha_j - e^{2\pi i\theta}| d\theta} \\ &= |a_d| \prod_{j=1}^d \max\{1, |\alpha_j|\} = M(f). \end{aligned}$$

This completes the proof. \square

Corollary 15.2 therefore implies that Mahler's measure is a continuous function on $\mathbb{C}[x]$.

Now let us review some of the results in the direction of Lehmer's conjecture. For a polynomial $f(x) \in \mathbb{C}[x]$ of degree d , define its **reciprocal polynomial**

$$f^*(x) = x^d f(x^{-1}).$$

We will say that $f(x)$ is **reciprocal** if $f(x) = f^*(x)$, and that it is **non-reciprocal** otherwise. The reciprocal condition on $f(x)$ is equivalent to the condition that its coefficients read forward same as backward.

Theorem 15.3 (Smyth, 1971). *If $f(x) \in \mathbb{Z}[x]$ is non-reciprocal such that $f(0)f(1) \neq 0$, then*

$$M(f) \geq M(x^3 - x - 1) = 1.324\dots$$

Smyth's theorem implies that we can restrict our attention to reciprocal polynomials. However, in this case no absolute lower bound is known: the best unconditional bound known depends on the degree of $f(x)$, and looks as follows.

Theorem 15.4 (Dobrowolski, 1979). *Let $\varepsilon > 0$. Then there exists $d_0(\varepsilon) \in \mathbb{Z}$ such that for every $f(x) \in \mathbb{Z}[x]$ of degree $d > d_0(\varepsilon)$,*

$$(58) \quad M(f) > 1 + (1 - \varepsilon) \left(\frac{\log \log d}{\log d} \right)^3.$$

Actually, Cantor, Straus, Loubotin, and Rausch slightly improved Dobrowolski's method in several papers published between 1982 and 1985, which allowed to replace $(1 - \varepsilon)$ in the lower bound of (58) with $(9/4 - \varepsilon)$, however getting rid of the dependence on d seems to be a major obstacle. Dobrowolski's theorem stated in terms of height of algebraic numbers implies that there exists a positive constant C_1 such that for every $\alpha \in \overline{\mathbb{Q}}$ of degree d , we have

$$h_1(\alpha) > \left(1 + C_1 \left(\frac{\log \log d}{\log d} \right)^3 \right)^{\frac{1}{d}},$$

which is weaker than the bound of the form $C^{\frac{1}{d}}$ proposed by Conjecture 15.1 where the constant C is independent of d . However, if we consider pairs of algebraic numbers α and $1 - \alpha$, it turns out to be possible to produce a lower bound on the product of their heights which is completely independent of d . The following Lehmer-type result with an implicit constant was first obtained by Zhang in 1992 with the use

of Arakelov theory; in 1995 Zagier obtained an elementary proof of it with an explicit constant.

Theorem 15.5 (Zhang, Zagier). *Let $\alpha \in \overline{\mathbb{Q}}$ be not equal to 0, 1, or $\frac{1 \pm \sqrt{-3}}{2}$, then*

$$h_1(\alpha)h_1(1 - \alpha) \geq \sqrt{\frac{1 + \sqrt{5}}{2}},$$

with equality if and only if α or $1 - \alpha$ is a primitive 10-th root of unity.

Notice that Theorem 15.5 can be viewed as a uniform lower bound on the height of algebraic points on the curve

$$x + y = 1.$$

There is also a uniform upper bound on all such points that satisfy one additional condition.

Definition 15.1. Two numbers $\alpha, \beta \in \overline{\mathbb{Q}}$ are called **multiplicatively dependent** if there exist $t \in \overline{\mathbb{Q}}$, $a, b \in \mathbb{Z}$, and $\varepsilon_1, \varepsilon_2$ roots of unity such that

$$\alpha = \varepsilon_1 t^a, \beta = \varepsilon_2 t^b.$$

The following bound was obtained in 1998 by P. Cohen and U. Zannier in [7].

Theorem 15.6 (Cohen, Zannier). *Let $0, 1 \neq \alpha \in \overline{\mathbb{Q}}$ be such that α and $1 - \alpha$ are multiplicatively dependent, then*

$$h_1(\alpha) \leq 2,$$

with equality if and only if $\alpha = 2$ or $1/2$.

Hence by combining the results of Theorems 15.5 and 15.6, we conclude that if (x, y) is an algebraic point on the curve

$$x + y = 1$$

with multiplicatively dependent coordinates $\neq 0, 1, \frac{1 \pm \sqrt{-3}}{2}$, then

$$\sqrt{\frac{1 + \sqrt{5}}{2}} \leq h_1(x)h_1(y) \leq 4.$$

This is a rare example of a uniform bound for the height of a set of algebraic points on an algebraic variety.

16. POINTS OF SMALL HEIGHT

In the previous section we saw an example of uniform bounds on height of points on a simple curve. In a more general situation we cannot hope to obtain such nice results. In fact, on higher dimensional varieties it is usually very difficult to prove existence of even a single point of relatively small height. If however we were to prove existence with an explicit bound on height, it would reduce the search for such points to a finite set due to Northcott's theorem. In other words, we can think of a bound on the height of a point on variety over a fixed number field K as a **search bound** in the same sense as in section 11. We start discussing this topic with the case of a linear variety, i.e. by revisiting Siegel's lemma, but this time in its much more powerful form over a number field. The following version of Siegel's lemma was proved by Bombieri and Vaaler in 1983, see [3].

Theorem 16.1 (Bombieri, Vaaler). *Let V be an L -dimensional subspace of K^N , $L < N$. Then there exists a basis $\mathbf{x}_1, \dots, \mathbf{x}_L \in \mathcal{O}_K^N$ for V such that*

$$(59) \quad \prod_{i=1}^L H(\mathbf{x}_i) \leq \{N|\mathcal{D}_K|^{1/d}\}^{L/2} H(V),$$

where \mathcal{D}_K is the discriminant of K , and $d = [K : \mathbb{Q}]$ as usual.

In other words, Theorem 16.1 states that a subspace V of K^N has a basis of relatively small height with coordinates in \mathcal{O}_K , where the bound on height is explicit and depends on height of V . In particular, it implies the existence of a non-zero point of small height in V , bounded as follows.

Corollary 16.2. *Let V be an L -dimensional subspace of K^N , $L < N$. Then there exists $\mathbf{0} \neq \mathbf{x} \in \mathcal{O}_K^N \cap V$ such that*

$$(60) \quad H(\mathbf{x}) \leq \{N|\mathcal{D}_K|^{1/d}\}^{1/2} H(V)^{1/L}.$$

The dependence on $H(V)$ in (59) and (60) is sharp. An analogous bound has been proved for a small-height basis of a subspace V of $\overline{\mathbb{Q}}^N$ by Roy and Thunder, see [32], where the constant in the upper bound does not depend on any number field; this is often desired, since \mathcal{D}_K can be quite large.

Next we can force additional conditions on the point in question, for instance require that it lies in V outside of the union of a collection of subspaces of K^N . In this direction we have the following extension of Siegel's lemma, see [16], [14].

Theorem 16.3 (Fukshansky, 2004). *Let V be an L -dimensional subspace of K^N , $1 \leq L \leq N$. Let $l = \lfloor \frac{N}{2} \rfloor$, and let $1 \leq s < L$ be an integer. Let W_1, \dots, W_M be nonzero subspaces of K^N with*

$$\max_{1 \leq i \leq M} \{\dim_K(W_i)\} \leq s.$$

There exists a point $\mathbf{x} \in V \setminus \bigcup_{i=1}^M W_i$ such that

$$(61) \quad H(\mathbf{x}) \ll_{K,N,L,s} H(V)^d \left\{ \left(\sum_{i=1}^M \frac{1}{H(W_i)^d} \right)^{\frac{1}{(L-s)d}} + M^{\frac{1}{(L-s)d+1}} \right\},$$

where the constant in the upper bound is explicit, and depends in particular on \mathcal{D}_K .

The dependence on $H(V)$ in the upper bound of Theorem 16.3 is sharp.

Next we consider the case of a quadratic hypersurface. Namely, let

$$F(\mathbf{X}) = \sum_{i=1}^N \sum_{j=1}^N f_{ij} X_i X_j \in K[X_1, \dots, X_N]$$

be a quadratic form in N variables with coefficients in the number field K of degree d over \mathbb{Q} . We say that F is **isotropic** over K if there exists $\mathbf{0} \neq \mathbf{x} \in K^N$ such that $F(\mathbf{x}) = 0$. Provided that F is isotropic over K , we are interested in proving the existence of a non-zero point of bounded height in the quadratic variety

$$\mathcal{V}_K(F) = \{\mathbf{x} \in K^N : F(\mathbf{x}) = 0\}$$

with an explicit bound on height. The following theorem was originally proved by Cassels in 1955 for the case $K = \mathbb{Q}$, and then extended to arbitrary number fields by Raghavan in 1975; see [5] and [30].

Theorem 16.4 (Cassels, Raghavan). *Let F be a quadratic form, which is isotropic over K as above, then there exists $\mathbf{0} \neq \mathbf{x} \in \mathcal{V}_K(F)$ such that*

$$H(\mathbf{x}) \ll_{K,N} H(F)^{\frac{N-1}{2}},$$

where the constant in the upper bound is explicit, and depends in particular on \mathcal{D}_K .

The dependence of $H(F)$ in the upper bound of Theorem 16.4 is best possible. What is instead of being a quadratic form, F is an inhomogeneous quadratic polynomial over K ? In other words, let

$$P(\mathbf{X}) = \sum_{i=1}^N \sum_{j=1}^N f_{ij} X_i X_j + \sum_{i=1}^N f_{0i} X_i + f_{00} \in K[X_1, \dots, X_N],$$

and suppose that

$$\mathcal{V}_K(P) = \{\mathbf{x} \in K^N : P(\mathbf{x}) = 0\}$$

is not empty. We want to prove the existence of a point $\mathbf{x} \in \mathcal{V}_K(F)$ of bounded height. Notice that we can “homogenize” F by adding one more variable X_0 , i.e. consider the quadratic form in $N + 1$ variables

$$F(\mathbf{X}) = \sum_{i=0}^N \sum_{j=1}^N f_{ij} X_i X_j \in K[X_0, \dots, X_N].$$

Exercise 16.1. *A point $\mathbf{x} = (x_0, x_1, \dots, x_N) \in K^{N+1}$ with $x_0 \neq 0$ is a zero of F if and only if the point $\mathbf{x}' = (x_1, \dots, x_N) \in K^N$ is a zero of P .*

Hence we want to look for small-height zeros of F with additional condition $x_0 \neq 0$. The following theorem was originally proved for the case $K = \mathbb{Q}$ by Masser in 1998 and extended over an arbitrary number field by Fukshansky in 2003; see [26], [15].

Theorem 16.5 (Masser, Fukshansky). *Let F be a quadratic form in $N + 1 \geq 2$ variables with coefficients in K . Suppose that there exists $\mathbf{x} = (x_0, \dots, x_N) \in K^{N+1}$ such that $F(\mathbf{x}) = 0$ and $x_0 \neq 0$, then there exists such \mathbf{x} with*

$$(62) \quad H(\mathbf{x}) \ll_{K,N} H(F)^{\frac{N+1}{2}},$$

where the constant in the upper bound is explicit, and depends in particular on \mathcal{D}_K .

This implies that if an inhomogeneous quadratic polynomial in N variables with coefficients in K has a zero over K , then it has such a zero of height bounded as in (62). The exponent in the upper bound of (62) is best possible as demonstrated by an example of Masser.

Next consider a certain generalization of Theorem 16.5. Namely, notice that the condition $x_0 \neq 0$ is a non-vanishing condition on a simple linear form. Then we can state the following more general result, see [15].

Theorem 16.6 (Fukshansky, 2003). *Let F be a quadratic form in $N+1$ variables with coefficients in K , as above. Let*

$$L_1(\mathbf{X}), \dots, L_M(\mathbf{X}) \in K[X_0, X_1, \dots, X_N]$$

be linear forms in $N+1$ variables with coefficients in K . Suppose that there exists a point $\mathbf{x} \in K^N$ such that $F(\mathbf{x}) = 0$, and $L_i(\mathbf{x}) \neq 0$ for each $1 \leq i \leq M$. Then there exists such a point \mathbf{x} with

$$(63) \quad H(\mathbf{x}) \ll_{K,N,M} H(F)^{\frac{N+1}{2} + (M-1)(N+2)} \prod_{i=1}^M H(L_i)^{\frac{(2M-1)N}{M}},$$

where the constant in the upper bound is explicit, and depends in particular on \mathcal{D}_K .

Notice that the additional arithmetic condition on non-vanishing of linear forms in Theorem 16.6 is similar to the additional condition on missing a collection of subspaces in Theorem 16.3: a linear form does not vanish at \mathbf{x} if and only if \mathbf{x} is outside of its nullspace. As a corollary of Theorem 16.6, we can state the following Cassels-type result on singular and non-singular points in a quadratic variety, see [15] and [12].

Corollary 16.7 (Fukshansky, 2005). *Let F be a quadratic form in N variables with coefficients in K . Let*

$$\mathcal{V}_K(F) = \{\mathbf{x} \in K^N : F(\mathbf{x}) = 0\},$$

as above. If there exists a non-singular point $\mathbf{x} \in \mathcal{V}_K(F)$, then there exists such a point with

$$(64) \quad H(\mathbf{x}) \ll_{K,N} H(F)^{\frac{N-1}{2}}.$$

Also, if there exists a singular point $\mathbf{0} \neq \mathbf{y} \in \mathcal{V}_K(F)$, then there exists such a point with

$$(65) \quad H(\mathbf{y}) \ll_{K,N} H(F)^{\frac{r}{2(N-r)}},$$

where r is the rank of F on K^N , $1 \leq r < N$, since $\mathcal{V}_K(F)$ contains non-zero singular points. The constants in the upper bounds of (64) and (65) are explicit, and depend in particular on \mathcal{D}_K .

What can be said about bounds on height of solutions of polynomials of degree higher than 2 in an arbitrary number of variables over a fixed number field K ? This problem seems to be completely out of reach at the present time. In fact, if $K = \mathbb{Q}$ and F is a homogeneous polynomial, such a bound would provide an algorithm to decide whether a homogeneous Diophantine equation has an integral solution, and so

would imply a positive answer to Hilbert's 10th problem in this case, i.e. this would mean that there exists an algorithm to decide whether such an equation has non-trivial integral solutions. However, by the famous theorem of Matijasevich [28] Hilbert's 10th problem is undecidable. This means that in general such bounds do not exist over \mathbb{Q} ; in fact, they are unlikely to exist over any fixed number field. Moreover, it is known they do not exist over \mathbb{Q} even for a quartic polynomial (see [27] for details). This problem becomes considerably easier if we were to allow for solutions to lie over some extension of K of bounded degree over K . In fact, the following basic bound is quite easy to prove (see [13]).

Lemma 16.8. *Let $M \geq 1$, $N \geq 2$, and $F(X_1, \dots, X_N)$ be a homogeneous polynomial in N variables of degree M with coefficients in a number field K . There exists $\mathbf{0} \neq \mathbf{z} \in \overline{\mathbb{Q}}^N$ with $\deg_K(\mathbf{z}) \leq M$ such that $F(\mathbf{z}) = 0$ and*

$$(66) \quad H(\mathbf{z}) \leq \sqrt{2} H(F)^{1/M}.$$

In fact, it is possible to prove a stronger statement by requiring the point \mathbf{x} in question to satisfy some additional arithmetic conditions. Write \mathbb{G}_m^N for the multiplicative torus $(\overline{\mathbb{Q}}^\times)^N$. The following theorem is proved in [13].

Theorem 16.9 (Fukshansky, 2004). *Let $F(X_1, \dots, X_N)$ be a homogeneous polynomial in $N \geq 2$ variables of degree $M \geq 1$ over a number field K , and let $A \in GL_N(K)$. Then either there exists $\mathbf{0} \neq \mathbf{x} \in K^N$ such that $F(\mathbf{x}) = 0$ and*

$$(67) \quad H(\mathbf{x}) \leq H(A),$$

or there exists $\mathbf{x} \in A\mathbb{G}_m^N$ with $\deg_K(\mathbf{x}) \leq M$ such that $F(\mathbf{x}) = 0$, and

$$(68) \quad H(\mathbf{x}) \ll_{M,N} H(A)^2 H(F)^{1/M},$$

where the constant in the upper bound is explicit, and depends only on M and N .

Theorem 16.9 asserts that for each element A of $GL_N(K)$ either there exists a zero of F over K whose height is bounded by $H(A)$, or there exists a small-height zero of F over $\overline{\mathbb{Q}}$ which lies outside of the union of nullspaces of row vectors of A^{-1} ; for instance, if $A = I_N$ this means that there exists a small-height zero of F with all coordinates nonzero. As a corollary of this result it is also possible to derive a bound on the height of a zero of an *inhomogeneous* polynomial with additional arithmetic conditions; this has also been done in [13].

To conclude these notes, it is useful to mention some of the further topics. These include further connections with arithmetic geometry via the study of points of bounded height on various more specialized varieties. This requires a more extensive theory of heights, including heights of subvarieties of an algebraic variety and of projective intersection cycles; one way to define such heights is via Chow forms. Another interesting direction is the study of multi-dimensional Mahler's measure, which is defined by an integral formula analogous to the one in section 15; there is no analogue of the product formula definition of Mahler's measure in the higher dimensional cases. Values of Mahler's measure are connected with special values of L -functions, volumes of hyperbolic manifolds, and K -theory, among other things. It is also possible to study various connections and applications of Lehmer's conjecture, which include dynamical systems, ergodic theory, and hyperbolic topology. On the geometry of numbers side of things, one can investigate an analogue of Minkowski's beautiful theory over so-called *adelic* spaces with number fields playing the role of lattices. These are just a few of the vast array of topics related to Diophantine approximations and geometry of numbers.

REFERENCES

- [1] 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.
- [2] M. Beck and S. Robins. *Computing the Continuous Discretely. Integer-Point Enumeration in Polyhedra*. Springer-Verlag, to appear.
- [3] E. Bombieri and J. D. Vaaler. On Siegel’s lemma. *Invent. Math.*, 73(1):11–32, 1983.
- [4] D. Bump, K. K. Choi, P. Kurlberg, and J. Vaaler. A local Riemann hypothesis, I. *Math. Z.*, 233(1):1–19, 2000.
- [5] J. W. S. Cassels. Bounds for the least solutions of homogeneous quadratic equations. *Proc. Cambridge Philos. Soc.*, 51:262–264, 1955.
- [6] J. W. S. Cassels. *An Introduction to the Geometry of Numbers*. Springer-Verlag, 1959.
- [7] P. B. Cohen and U. Zannier. Multiplicative dependence and bounded height, an example. *Algebraic number theory and Diophantine analysis*, pages 93–101, 1998.
- [8] J. H. Conway and N. J. A. Sloane. *Sphere Packings, Lattices, and Groups*. Springer-Verlag, 1988.
- [9] H. Davenport. On a principle of Lipschitz. *J. London Math. Soc.*, 26:179–183, 1951.
- [10] B. Edixhoven and J.-H. Evertse (Eds.). *Diophantine Approximation and Abelian Varieties*. Springer-Verlag, 1993.
- [11] G. Ewald. *Combinatorial convexity and algebraic geometry*. Springer-Verlag, 1996.
- [12] L. Fukshansky. On effective Witt decomposition and Cartan-Dieudonné theorem. *to appear in Canad. J. Math.*
- [13] L. Fukshansky. Search bounds for zeros of polynomials over $\overline{\mathbf{Q}}$. *preprint*.
- [14] L. Fukshansky. Siegel’s lemma with additional conditions. *to appear in J. Number Theory*.
- [15] L. Fukshansky. Small zeros of quadratic forms with linear conditions. *J. Number Theory*, 108(1):29–43, 2004.
- [16] L. Fukshansky. Integral points of small height outside of a hypersurface. *Monatsh. Math.*, 147(1):25–41, 2006.
- [17] P. M. Gruber and C. G. Lekkerkerker. *Geometry of Numbers*. North-Holland Publishing Co., 1987.
- [18] T. Hales. A proof of the Kepler conjecture. *Ann. of Math. (2)*, 162(3):1065–1185, 2005.
- [19] 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.
- [20] B. Jacob. *Linear Algebra*. W.H. Freeman and Company, 1990.
- [21] V. Jarník. Zwei Bemerkungen zur Geometrie de Zahlen. *Věstník Královské České Společnosti Nauk*, 1941.
- [22] S. Lang. *Algebraic Number Theory*. Springer-Verlag, 1994.
- [23] A. K. Lenstra, H. W. Lenstra, and L. Lovasz. Factoring polynomials with rational coefficients. *Math. Ann.*, 261:515–534, 1982.

- [24] R. Lipschitz. *Monatsber. der Berliner Academie*, pages 174–185, 1865.
- [25] D. Marcus. *Number Fields*. Springer-Verlag, 1977.
- [26] D. W. Masser. How to solve a quadratic equation in rationals. *Bull. London Math. Soc.*, 30(1):24–28, 1998.
- [27] D. W. Masser. Search bounds for Diophantine equations. *A panorama of number theory or the view from Baker's garden (Zurich, 1999)*, pages 247–259, 2002.
- [28] Yu. V. Matijasevich. The diophantineness of enumerable sets. *Dokl. Akad. Nauk SSSR*, 191:279–282, 1970.
- [29] M. Pohst. On the computation of lattice vectors of minimal length, successive minima, and reduced bases with applications. *technical report*.
- [30] S. Raghavan. Bounds of minimal solutions of diophantine equations. *Nachr. Akad. Wiss. Göttingen, Math. Phys. Kl.*, 9:109–114, 1975.
- [31] K.F. Roth. Rational approximations to algebraic numbers. *Mathematika*, 2:1–20, 1955.
- [32] D. Roy and J. L. Thunder. An absolute Siegel's lemma. *J. Reine Angew. Math.*, 476:1–26, 1996.
- [33] W. Rudin. *Principles of Mathematical Analysis*. McGraw Hill, Inc., 1976.
- [34] P. Scherk. Convex bodies off center. *Archiv Math.*, 3:303, 1950.
- [35] W. M. Schmidt. *Diophantine Approximation*. Springer-Verlag, 1980.
- [36] W. M. Schmidt. *Diophantine Approximations and Diophantine Equations*. Springer-Verlag, 1991.
- [37] W. M. Schmidt. *Height of Polynomials and Entropy in Algebraic Dynamics*. Springer-Verlag, 1999.
- [38] C. L. Siegel. Zur theorie der quadratischen formen. *Nachr. Akad. Wiss. Göttingen Math.-Phys. Kl. II*, pages 21–46, 1972.
- [39] P. G. Spain. Lipschitz: a new version of old principle. *Bull. London Math. Soc.*, 27:565–566, 1995.
- [40] A. Thue. Über Annäherungswerte algebraischer Zahlen. *J. Reine Angew. Math.*, 135:284–305, 1909.
- [41] 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, TEXAS A&M UNIVERSITY, TAMU 3368,
COLLEGE STATION, TEXAS 77843-3368

E-mail address: lenny@math.tamu.edu