Topics in Discrete Optimization

Lenny Fukshansky

Contents

Chapter 1. Preface i

1.1. What is discrete optimization? i

1.2. Asymptotic notation and computational complexity ii

Chapter 2. Knapsack and Frobenius problems 1

2.1. Complexity of knapsack problems 1

2.2. Approximating knapsack problem by relaxation 6

2.3. LP-polytope 9

2.4. Integer knapsack and counting integer lattice points in polytopes 12

2.5. Frobenius problem 18

2.6. Problems 25

Chapter 3. Geometry of Numbers 27

3.1. Lattices 27

3.2. Theorems of Blichfeldt and Minkowski 35

3.3. Successive minima 38

3.4. Inhomogeneous minimum 43

3.5. Problems 46

Chapter 4. Lattice Problems, Connections and Applications 48

4.1. Sphere packing, covering and kissing number problems 48

4.2. Lattice packings in dimension 2 54

4.3. Algorithmic problems on lattices 59

4.4. CVP is NP-hard 62

4.5. Geometry of the Frobenius problem 63

4.6. Lattice point counting 66

4.7. Applications of lattices in coding theory and cryptography 69

4.8. Euclidean frames 71

4.9. Problems 76

Bibliography 77

CHAPTER 1

Preface

1.1. What is discrete optimization?

Let $f(\boldsymbol{x})$ be a multi-variable function defined on some domain D. An *optimization problem* defined by f on D can then be formulated as:

Maximize / minimize $f(\boldsymbol{x})$ on D.

An optimization problem like this is called *discrete* if the domain D is a discrete set inside of some topological space, i.e. if every point of D is an isolated point (i.e., no open neighborhood of a point in D contains any other points of D). The condition of a point belonging to the domain D can often be formulated as a collection of certain constraints or inequalities.

In these notes, we will discuss several central discrete optimization problems, coming from different (but related) areas of discrete mathematics. Our goal will be to describe the problems with the necessary context and background they come from while focusing on their interpretation in the scope of discrete geometry. It is this geometric framework that naturally connects all the problems we will discuss and brings them together. Here are the specific problems we will be interested in:

- **The Knapsack Problems:** given a collection of objects with assigned weight and cost, maximize the objective function (i.e. total cost) while keeping the weight under the specified threshold (subject to possibly some additional constraints). In addition to its intrinsic mathematical significance, this problem often comes up in resource allocation.
- **The Frobenius Problem:** given a collection of relatively prime positive integers, find the largest positive integer that cannot be represented as their nonnegative integer linear combination. This problem appears in many areas of mathematics and is related to the knapsack problems.
- **The Main Problem of Coding Theory:** maximize the error-correcting capability of a linear code while keeping its codeword length bounded. This problem is central in the study of accurate data transmission over potentially noisy channels.
- **Optimization Problems on Lattices:** optimize packing density, covering thickness and kissing number of a Euclidean lattice in n dimensions. This is the main problem of lattice theory, a branch of mathematics at the intersection of number theory and discrete geometry. In addition to its theoretical value, it has numerous applications, for instance in digital and wireless communications.
- **Coherence Minimization on Euclidean Frames:** find frames (overdetermined spanning sets) in Euclidean vector spaces of large cardinality and small coherence. Such frames allow for sufficiently fast data transmission with efficient erasure-recovery capabilities.

1.2. Asymptotic notation and computational complexity

The main measure of "hardness" of different problems that we will discuss is given by their *computational complexity.* Here, we briefly and somewhat informally recall some basic notions of computational complexity. To start with, we need some asymptotic notation. Given two functions $f, g : \mathbb{R} \to \mathbb{R}$, we write $f(x) = O(g(x))$ if there exists a real constant C so that $f(x) \leq Cg(x)$ as $x \to \infty$. This is called *big-O notation.*

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

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

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

CHAPTER 2

Knapsack and Frobenius problems

2.1. Complexity of knapsack problems

The problems we will discuss here fit into the general *linear programming* or *linear optimization* paradigm. A *linear program (LP)* is a problem that can be stated in the following form:

Given vectors $\boldsymbol{c}$, $\boldsymbol{b}$ and a matrix A, find a vector $\boldsymbol{x} \geq \boldsymbol{0}$ that maximizes the objective function $\boldsymbol{c}^\top \boldsymbol{x}$ subject to the constraint $A\boldsymbol{x} \leq \boldsymbol{b}$.

A linear program is called an *integer linear program (ILP)* or simply an *integer program (IP)* if the solution vector $\boldsymbol{x}$ is required to have integer coordinates.

Suppose we have a knapsack that can hold weight no more than W. We want to pack it with objects of types 1 through n where an object of type i has weight w_i and price p_i. Our objective is to maximize the value of the knapsack. If we write x_i for the number of objects of type i that we take, we have the following optimization problem:

Maximize the objective function

$$\sum_{i=1}^{n} p_i x_i$$

under the constraint

$$\sum_{i=1}^{n} w_i x_i \leq W.$$

This is the basic prototype of a *knapsack problem.* Putting on additional constraints distinguishes different types of knapsack problems, for instance:

- **Binary knapsack problem (BKP):** the variables x_i can take values $0, 1$ only
- **Bounded knapsack problem (BndKP):** for each i, x_i is an integer with $0 \leq x_i \leq b_i$ for some upper bounds b_i.
- **Unbounded knapsack problem (UbndKP):** for each i, x_i is an integer.
- **Subset-sum problem (SSP):** for each i, $w_i = p_i$ and $x_i = 0, 1$.

All of these problems are NP-hard. To show this, first observe that SSP is the special case of BKP with $p_i = w_i$ for each i; also, BKP is the special case of BndKP with $b_i = 1$ for each i. On the other hand, UbndKP is clearly at least as hard as BndKP. Hence, NP-hardness of all of these problems follows from NP-completeness of SSP. We will show that SSP is NP-complete, more specifically we will deal with the the *decision version* of SSP:

Given a set of weights $S = \{w_1, \dots, w_n\}$ and a target value t, is there a subset $S' \subseteq S$ such that $\sum_{w_i \in S'} w_i = t$?

We first need some notation from Boolean logic. A *Boolean formula* is an expression built from Boolean variables (taking values TRUE = 1 or FALSE = 0) and operators AND ($\wedge$), OR ($\vee$), NOT ($\neg$) and parentheses separating different *clauses* of the formula. A Boolean formula is said to be *satisfiable* if it can be made TRUE by an appropriate assignment of variables.

EXAMPLE 2.1.1. *The formula*

$$x \wedge \neg y$$

is satisfiable: setting $x =$ TRUE, $y =$ FALSE makes this formula TRUE. On the other hand, the formula

$$x \wedge \neg x$$

is not satisfiable.

The (unrestricted) *Boolean satisfiability problem* SAT is the problem of determining if a given Boolean formula is satisfiable or not. It can be formally stated as the following decision problem:

INPUT: Boolean formula

OUTPUT: YES (satisfiable) or NO (not satisfiable)

THEOREM 2.1.1 (S. Cook (1971), L. Levin (1973)). *SAT is NP-complete.*

This was the first provable instance of an NP-complete problem – the notion did not properly exist before the work of Cook and Levin, who independently established this result. The fact that SAT is NP is not difficult to see: given any assignment of the variables, it can be verified in polynomial time whether they make the given formula TRUE or not. To show that it is NP-complete, one needs to prove that any NP problem can be reduced to an instance of SAT by a polynomial time algorithm. We do not prove this result here, however we will mention a (restricted) variation of the SAT problem, called *3-SAT* which is also known to be NP-complete: 3-SAT is the Boolean satisfiability problem, where every clause consists of no more than 3 *literals* (a literal is either a variable x, or negation of a variable $\neg x$).

In fact, every instance of a SAT formula can be transformed into a 3-SAT formula as follows. First notice that every formula can be rewritten in a way that the clauses are joined by $\wedge$ operator: for example, a formula like

$$(x_1 \wedge y_1) \vee (x_2 \wedge y_2)$$

can be transformed into

$$(x_1 \vee x_2) \wedge (y_1 \vee x_2) \wedge (x_1 \vee y_2) \wedge (y_1 \vee y_2).$$

Such a form is called a *generalized conjunctive normal form* for a given Boolean formula. Suppose now that there is a clause in an unrestricted SAT formula that looks like

$$\ell_1 \vee \dots \vee \ell_n,$$

where $\ell_1, \dots, \ell_n$ are literals. Introducing new variables $x_1, \dots, x_{n-2}$ we can rewrite this formula as

$$(\ell_1 \vee \ell_2 \vee x_1) \wedge (\neg x_1 \vee \ell_3 \vee x_2) \wedge \dots \wedge (\neg x_{n-3} \vee \ell_{n-2} \vee x_{n-2}) \wedge (\neg x_{n-2} \vee \ell_{n-1} \vee \ell_n).$$

This new formula is satisfiable if and only if the original is, and its length is at most 3 times longer than the original, which means that the reduction from SAT to 3-SAT implies only polynomial growth in the length of the formula. Hence SAT and 3-SAT have the same order of computational complexity, i.e. they are both NP-complete by the Cook-Levin theorem. We are now ready to show (somewhat informally) that the subset-sum problem is NP-complete by constructing a polynomial time reduction algorithm from 3-SAT to it.

Theorem 2.1.2. *The decision version of SSP is NP-complete.*

Sketch of proof. Recall that the running time is measured as a function of the input size. It is easy to see that our problem is NP. Indeed, let $S = \{w_1, \dots, w_m\}$ be the set of weights and T the target sum. Given a specific subset $S' \subseteq S$, it is simply a summation problem to verify whether S' sums to T – this summation algorithm runs in polynomial (in fact, linear) time.

Now we show that the decision version of SSP is NP-hard by constructing a polynomial-time reduction from 3-SAT to it. Let us write n for the number of variables and k for the number of clauses in our Boolean formula. We demonstrate this reduction on an example. Consider the Boolean formula

$$(x_1 \vee x_2 \vee x_3) \wedge (\neg x_1 \vee \neg x_2 \vee x_3) \wedge (x_1 \vee \neg x_2 \vee x_3) \wedge (x_1 \vee x_2 \vee \neg x_3).$$

There are four clauses which we label as c_1, c_2, c_3, c_4, so in this example $n = 3$, $k = 4$. This formula is satisfied if and only if each of the clauses is TRUE. Let us introduce two variables v_{i1}, v_{i2} for each of the Boolean variables x_i and two variables w_{i1}, w_{i2} for each of the clauses c_i along with $n - 1$ auxiliary variables $s_{i1}, \dots, s_{i(n-1)}$ for each c_i. With this notation, let us build a table consisting of four blocks as follows:

	x_1	x_2	x_3	c_1	c_2	c_3	c_4
v_{11}	1	0	0	1	0	1	1
v_{12}	1	0	0	0	1	0	0
v_{21}	0	1	0	1	0	0	1
v_{22}	0	1	0	0	1	1	0
v_{31}	0	0	1	1	1	1	0
v_{32}	0	0	1	0	0	0	1
s_{11}	0	0	0	1	0	0	0
s_{12}	0	0	0	1	0	0	0
s_{21}	0	0	0	0	1	0	0
s_{22}	0	0	0	0	1	0	0
s_{31}	0	0	0	0	0	1	0
s_{32}	0	0	0	0	0	1	0
s_{41}	0	0	0	0	0	0	1
s_{42}	0	0	0	0	0	0	1
T	1	1	1	3	3	3	3

The values in the cells of the table are assigned as follows.

- The variable v_{i1} stands for the TRUE value of x_i and v_{i2} stands for the FALSE value of x_i. Hence the cell corresponding to v_{i1}, x_j or v_{i2}, x_j gets a 1 if $i = j$ and 0 if $i \neq j$.

- The cell corresponding to v_{i1}, c_j gets a 1 if setting $x_i =$ TRUE makes c_j TRUE, and 0 otherwise. The cell corresponding to v_{i2}, c_j gets a 1 if setting $x_i =$ FALSE makes c_j TRUE, and 0 otherwise.
- The cell corresponding to s_{il}, x_j gets a 0 for all i, l, j.
- The cell corresponding to s_{il}, c_j gets a 1 if $i = j$ and 0 otherwise for each $1 \leq l \leq n-1$.
- The cells in the row labeled T (target sum) corresponding to a variable x_i get a 1 and those corresponding to a clause c_i get an $n =$ number of variables.

Now let S be the (multi-) set of numbers as written in rows except for the last one and the last row be the target sum T, so in our example $S =$

$$\{1001011, 1000100, 101001, 100110, 11110, 10001, 1000, 1000, 100, 100, 10, 10, 1, 1\}$$

and $T = 1113333$. We claim that the Boolean formula represented by this table is satisfiable if and only if there exists a subset S' of S which sums up to T. Indeed, notice that each row in the upper part of the table corresponds to a TRUE or FALSE value of the variable x_i (we would always pick precisely one of the rows v_{i1} and v_{i2}, since x_i must be assigned TRUE or FALSE but not both at the same time, and hence the corresponding digit of T would always be 1). Then we pick the rows in the bottom part of the table to compensate for those positions that are $< n$: the formula is not satisfiable if and only if there is a column corresponding to some c_i whose entries add up to a number $< n$ (this happens precisely when there is no choice of the variables making the clause c_i TRUE). In our example, the Boolean formula is satisfiable, since for instance the rows corresponding to v_{11}, v_{21}, v_{31} and s_{21}, s_{22}, s_{31}, s_{41} add up to T:

$$1001011 + 101001 + 11110 + 100 + 100 + 10 + 1 = 1113333.$$

Indeed, this choice of the subset S' corresponds to the assignment

$$x_1 = \text{TRUE},\ x_2 = \text{TRUE},\ x_3 = \text{TRUE},$$

which makes each of the clauses TRUE, hence making the formula TRUE.

Notice that each element of S has at most $n + k$ digits in it and there are at most $2n + (n-1)k$ elements in S. This ensures that this reduction procedure runs in polynomial time in the size of the input, which is itself a function of n and k. This completes the proof. □

Many of the problems mentioned in this section (e.g. ILP, 3-SAT and SSP) are among the original *Karp's 21 NP-complete problems.* In 1972, Richard Karp published a paper [**Kar72**] in which he showed the NP-completeness of 21 different natural combinatorial and graph theoretic problems. His main tool was Cook-Levin Theorem: he constructed polynomial time reductions from SAT to several problems and then used the Cook-Levin Theorem to establish NP-completeness of these problems. He then used these few original problems to show NP-completeness of the rest of his list of 21.

To close this section, let us also mention a simple but curious reformulation of the SAT problem in terms of polynomial vanishing. For a Boolean formula in conjunctive normal form (CNF), we construct a *CNF-polynomial* in the variables $x_1, \dots, x_n$ corresponding to this formula as follows:

(1) A literal x_i becomes a linear factor x_i and a literal $\neg x_i$ becomes a linear factor $1 - x_i$.
(2) Each clause becomes a product of linear factors corresponding to its literals, so disjunction becomes multiplication. We refer to such products as *literal-monomials.*
(3) Conjunction becomes addition.

Let us consider the following example:

$$(x_1 \vee x_2 \vee x_3) \wedge (\neg x_1 \vee \neg x_2 \vee x_3) \wedge (x_1 \vee \neg x_2 \vee x_3) \wedge (x_1 \vee x_2 \vee \neg x_3), \tag{2.1}$$

then the formula (2.1) correspond to the CNF-polynomial

$$f(x_1, x_2, x_3) = x_1x_2x_3 + (1 - x_1)(1 - x_2)x_3 + x_1(1 - x_2)x_3 + x_1x_2(1 - x_3). \tag{2.2}$$

Assign the values $0 =$ TRUE and $1 =$ FALSE. A CNF formula is satisfiable if and only if there exists a TRUE / FALSE assignment of the Boolean variables such that each clause is TRUE. This happens if and only if there exists a 0 / 1 assignment of the variables in the corresponding CNF-polynomial which makes every monomial 0. This property is equivalent to the CNF-polynomial vanishing at some vertex of the unit cube $[0, 1]^n$ in $\mathbb{R}^n$. For example, the polynomial in (2.2) vanishes at $(0, 1, 1)$; this corresponds to the assignment of the Boolean variables

$$x_1 = \text{TRUE},\ x_2 = \text{FALSE}, x_3 = \text{FALSE},$$

which indeed satisfies the formula (2.1). We can use this construction to prove the following observation.

Proposition 2.1.3. *The problem of determining whether a given multilinear polynomial of degree $k \geq 3$ in $n \geq 2$ variables vanishes at a vertex of the unit cube $[0, 1]^n$ in $\mathbb{R}^n$ is NP-complete.*

Proof. It is clear that this problem is NP: the procedure of evaluating a polynomial at a given point has polynomial complexity. The construction above shows a polynomial-time reduction from k-SAT, the Boolean k-satisfiability problem, to our problem for the corresponding CNF-polynomial. Since k-SAT is NP-hard for every $k \geq 3$, so must be our problem. □

2.2. Approximating knapsack problem by relaxation

The Binary knapsack problem (BKP) can be formulated as follows:

$$\begin{aligned}&\text{maximize } \sum_{i=1}^{n} p_i x_i\\ &\text{subject to } \sum_{i=1}^{n} w_i x_i \leq W,\\ &x_i \in \{0,1\}\ \forall\ 1 \leq i \leq n.\end{aligned}$$

Let us write $\boldsymbol{p} = (p_1, \dots, p_n)$ for the *profit vector* and $\boldsymbol{w} = (w_1, \dots, w_n)$ for the *weight vector*. This is arguably the most important of the knapsack problems. Indeed, SSP is a special case of BKP and BndKP can be reduced to BKP (in a larger number of variables) as we now show. BndKP can be formulated as follows:

$$\begin{aligned}&\text{maximize } \sum_{i=1}^{n} p_i x_i\\ &\text{subject to } \sum_{i=1}^{n} w_i x_i \leq W,\\ &x_i \in \mathbb{Z},\ 0 \leq x_i \leq b_i\ \forall\ 1 \leq i \leq n.\end{aligned}$$

For each $1 \leq i \leq n$, let $B_i := \{b_{i1}, \dots, b_{il_i}\}$ be a minimal (with respect to size) partition of b_i so that every integer between 0 and b_i is representable as a sum of some subcollection of B_i: such a partition always exists, since in the worst case scenario we can always take $B_i = \{1, \dots, 1\}$, but in general it will smaller. For example, if $b_i = 10$ we can take $B_i = \{1, 2, 3, 4\}$. Then we can introduce new variables y_{ik}, $1 \leq i \leq n$, $1 \leq k \leq l_i$, and rewrite the BndKP above as the following instance of BKP:

$$\begin{aligned}&\text{maximize } \sum_{i=1}^{n} p_i (b_{i1} y_{i1} + \dots + b_{il_i} y_{il_i})\\ &\text{subject to } \sum_{i=1}^{n} w_i (b_{i1} y_{i1} + \dots + b_{il_i} y_{il_i}) \leq W,\\ &y_{ik} \in \{0,1\}\ \forall\ 1 \leq i \leq n,\ 1 \leq k \leq l_i.\end{aligned}$$

These two problems are equivalent since there is bijection between all values of x_i between 0 and b_i and all values (written without repetition) of the sum $\sum_{k=1}^{l_i} b_{ik} y_{ik}$ as the variables y_{ik} assume values in the set $\{0, 1\}$.

From Section 2.1 we know that there is no known polynomial time algorithm to solve BKP. However, we can look for an approximate solution to BKP via certain relaxations. First such relaxation is the *continuous knapsack problem (CKP)*:

$$\begin{aligned}&\text{maximize } \sum_{i=1}^{n} p_i x_i\\ &\text{subject to } \sum_{i=1}^{n} w_i x_i \leq W,\\ &0 \leq x_i \leq 1\ \forall\ 1 \leq i \leq n.\end{aligned}$$

In other words, we no longer require x_i to take only integer values. Let us assume that the items are ordered so that

$$\frac{p_1}{w_1} \geq \frac{p_2}{w_2} \geq \cdots \geq \frac{p_n}{w_n}. \tag{2.3}$$

The following additional assumptions can be made for CKP:

(1) *Every weight $w_i \leq W$.* If this is not the case for some weight w_i, this weight (and its corresponding price p_i) can be eliminated, hence reducing the number of variables.

(2) $\sum_{i=1}^{n} w_i > W$. If this is not the case, then taking $x_i = 1$ for each i will maximize the objective function.

(3) *The inequalities in* (2.3) *are all strict.* Suppose not, then there exist some indices $1 < j < k < n$ such that

$$c := \frac{p_j}{w_j} = \cdots = \frac{p_k}{w_k},$$

which means that $\sum_{i=j}^{k} p_i x_i = c \sum_{i=j}^{k} w_i x_i$. Then set $t = \sum_{i=j}^{k} w_i$ and define a new variable $y = \frac{1}{t} \sum_{i=j}^{k} w_i x_i$; observe that $0 < y < 1$. In this case we can restate our instance of CKP as follows:

$$\begin{aligned}&\text{maximize } \sum_{i=1}^{j-1} p_i x_i + cty + \sum_{i=k+1}^{n} p_i x_i\\ &\text{subject to } \sum_{i=1}^{j-1} w_i x_i + ty + \sum_{i=k+1}^{n} w_i x_i \leq W,\\ &0 \leq x_i \leq 1 \ \forall\ i,\ 0 \leq y \leq 1.\end{aligned}$$

To solve CKP, we define the *critical index*

$$s = \min \left\{ j : \sum_{i=1}^{j} w_i > W \right\}.$$

Then $1 < s \leq n$ and we have the following result.

THEOREM 2.2.1. *[Dantzig, 1957] The optimal solution $\boldsymbol{x}^*$ to CKP is given by setting*

$$x_i^* = \begin{cases} 1 & \textit{if } 1 \leq i \leq s-1 \\ 0 & \textit{if } s+1 \leq i \leq n, \end{cases} \tag{2.4}$$

and $x_s^* = \frac{1}{w_s} \left(W - \sum_{j=1}^{s-1} w_j \right)$.

PROOF. Our proof follows [**MT90**]. First observe that a vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ maximizing the objective function must satisfy the condition

$$\sum_{i=1}^{n} w_i x_i = W, \tag{2.5}$$

since otherwise some coordinates of $\boldsymbol{x}$ can be increased still under the weight restriction, which will increase the value of the objective function. Arguing towards a contradiction, suppose the optimal solution $\boldsymbol{x}$ is not of the form $\boldsymbol{x}^*$, say $x_i < 1$ for some $i < s$. Then there must exist some index $j \geq s$ such that $x_j > x_j^*$. Now, for a sufficiently small $\varepsilon > 0$, replace x_i by $x_i + \varepsilon$ and x_j by $x_j - \varepsilon w_i / w_j$, hence still preserving condition (2.5). However, this change will increase the objective function by

$$\varepsilon\left(p_i - \frac{p_j w_i}{w_j}\right),$$

which is positive, since $p_i/w_i > p_j/w_j$. This contradicts the optimality of the solution $\boldsymbol{x}$. The assumption $x_j > 0$ for some $j > s$ is handled analogously, also leading to a contradiction. Hence we must have the condition (2.4) satisfied for the optimal solution, and the formula for x_s^* follows from maximality. This completes the proof. □

The maximal value of the objective function in CKP is then easy to compute: it is

$$\sum_{i=1}^{s-1} p_i + \frac{p_s}{w_s}\left(W - \sum_{j=1}^{s-1} w_j\right).$$

This immediately implies an upper bound on the maximal value of the objective function for the associated instance of BKP:

$$U := \sum_{i=1}^{s-1} p_i + \left[\frac{p_s}{w_s}\left(W - \sum_{j=1}^{s-1} w_j\right)\right].$$

There are other known relaxations of BKP (such as the *Lagrangian relaxation*, stemming from an application of the method of Lagrange multipliers) leading to other upper bounds on the objective function. There is also a lot of literature on the algorithmic complexity of computing these bounds. Some of the known algorithms rely on the *greedy* approach (making the locally optimal choice at each stage of the algorithm) as well as the *branch-and-bound* method (recursively splitting the search space into smaller pieces and optimizing on each of those). We do not get into this material here, but mention a book by Martello and Toth [**MT90**] as a comprehensive source of algorithmic information about knapsack problems.

2.3. LP-polytope

Our goal in this section is to present a geometric interpretation of the knapsack problems. First we need some geometric notation. Recall that a *compact* (i.e. closed and bounded) subset $X \subset \mathbb{R}^n$ is called *convex* if for any pair $\boldsymbol{x}, \boldsymbol{y} \in X$, $t\boldsymbol{x} + (1-t)\boldsymbol{y} \in X$ for any $0 \leq t \leq 1$. An important special class of convex sets is convex hulls: the *convex hull* of a set $X \subset \mathbb{R}^n$ is

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

This is the smallest convex set (with respect to inclusion) containing X, so X is convex if and only if $X = \mathrm{Co}(X)$. A *convex polytope* is the convex hull of a finite collection of points. There is also a related notion of a convex polyhedron. A *halfspace* in $\mathbb{R}^n$ is a set

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

for some $a_1, \dots, a_n, b \in \mathbb{R}$, and the set

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

is called a *bounding hyperplane* of $\mathcal{H}$. A *convex polyhedron* is a compact intersection of a finite collection of halfspaces. Hence P is a convex polyhedron if and only if

$$P = \{ \boldsymbol{x} \in \mathbb{R}^n : A\boldsymbol{x} \leq \boldsymbol{b} \} \tag{2.6}$$

for an $m \times n$ real matrix A and a vector $\boldsymbol{b} \in \mathbb{R}^m$ such that this set is bounded.

Theorem 2.3.1 (Minkowski-Weyl). *A set $P \subset \mathbb{R}^n$ is a convex polytope if and only if it is a convex polyhedron.*

While we do not prove this theorem here, we point out one of its important consequence. A point $\boldsymbol{v}$ in a convex set $X \subset \mathbb{R}^n$ is called a *vertex* if there exists some $\boldsymbol{c} \in \mathbb{R}^n$ such that for all $\boldsymbol{c}^\top \boldsymbol{v} < \boldsymbol{c}^\top \boldsymbol{x}$ for all $\boldsymbol{x} \in X$. Then every convex polyhedron is the convex hull of its vertices, of which there are only finitely many. More generally, we can define a k-dimensional *face* of an n-dimensional polytope P, $1 \leq k < n$, to be a k-dimensional subset $F \subset P$ such that for some $\boldsymbol{c} \in \mathbb{R}^n$,

$$\boldsymbol{c}^\top \boldsymbol{v} = \boldsymbol{c}^\top \boldsymbol{u} \ \forall \ \boldsymbol{v}, \boldsymbol{u} \in F \text{ and } \boldsymbol{c}^\top \boldsymbol{v} < \boldsymbol{c}^\top \boldsymbol{x} \ \forall \ \boldsymbol{v} \in F, \boldsymbol{x} \in P.$$

Here, by dimension of a subset F we mean $\dim_{\mathbb{R}} (\operatorname{span}_{\mathbb{R}} F) - 1$, so vertices are 0-dimensional faces of P and every face of P contains at least one vertex. The polytope P can then be represented as the disjoint union of its *interior* P^o and its *boundary* ∂P, where ∂P is the union of all of the faces of P and $P^o = P \setminus \partial P$.

Given a linear program

$$\begin{aligned} &\text{maximize } \boldsymbol{p}(\boldsymbol{x}) = \sum_{i=1}^n p_i x_i \\ &\text{subject to } A\boldsymbol{x} \leq \boldsymbol{b} \end{aligned}$$

for an $m \times n$ matrix A and a vector $\boldsymbol{b} \in \mathbb{R}^m$ we can define the corresponding *LP-polytope* as in (2.6). Then the linear program can be reformulated as

$$\text{maximize } \boldsymbol{p}(\boldsymbol{x}) = \sum_{i=1}^{n} p_i x_i \text{ on } P.$$

In general, the polytope P can be unbounded, but we will focus specifically on the situations when it is compact – we refer to such linear programs as *bounded LPs*. Notice that our knapsack problems BKP, BndKP and SSP are all bounded LPs. With this notation, we can prove an important theorem.

Theorem 2.3.2. *The objective function $\boldsymbol{p}(\boldsymbol{x})$ is maximized at a vertex of P. In other words, there exists a vertex $\boldsymbol{v} \in P$ such that $\boldsymbol{p}(\boldsymbol{v}) \geq \boldsymbol{p}(\boldsymbol{x})$ for all $\boldsymbol{x} \in P$.*

Proof. Suppose that $\boldsymbol{v} \in P$ is a point such that $\boldsymbol{p}(\boldsymbol{v}) \geq \boldsymbol{p}(\boldsymbol{x})$ for all $\boldsymbol{x} \in P$, define b to be this maximal value, i.e. $b = \boldsymbol{p}(\boldsymbol{v})$. Define the hyperplane

$$\mathbb{H}_{\boldsymbol{p}}(b) = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{p}(\boldsymbol{x}) = b\}$$

and two halfspaces

$$\mathcal{H}_1 = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{p}(\boldsymbol{x}) \leq b\}, \ \mathcal{H}_2 = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{p}(\boldsymbol{x}) \geq b\}.$$

By our assumption, we must have $P \subset \mathcal{H}_1$ and hence $\mathbb{H}_{\boldsymbol{p}}(b)$ cannot intersect the interior of $\mathcal{P}$, i.e. $F = P \cap \mathbb{H}_{\boldsymbol{p}}(b) \neq \emptyset$ must be some (union of) face(s) of P. Then $\boldsymbol{p}(\boldsymbol{x})$ is constant on F and F contains a vertex $\boldsymbol{u}$ of P, so $\boldsymbol{p}(\boldsymbol{u}) = \boldsymbol{p}(\boldsymbol{v})$ is a maximal value of $\boldsymbol{p}(\boldsymbol{x})$ on P. □

Therefore Theorem 2.3.2 implies that to solve a given linear program we need to find all the vertices of the corresponding LP-polytope and identify an optimal one among them. This is done by George Dantzig's *simplex algorithm.* The main idea of the simplex algorithm is to start at a vertex of the LP-polytope and move along an *edge* (1-dimensional face) to a neighboring vertex corresponding to a larger value of the objective function. The algorithm terminates when no such vertex exists. While we will not get into the details of the algebraic implementation of this algorithm, we will demonstrate a geometric example with an instance of CKP.

Example 2.3.1. *Consider the following instance of CKP:*

$$\begin{aligned}&\textit{maximize } 3x_1 + 5x_2 + 2x_3\\ &\textit{subject to } 2x_1 + 5x_2 + 7x_3 \leq 11,\\ &x_1, x_2, x_3 \in [0,1].\end{aligned}$$

Define the corresponding LP-polytope to be $P = \left\{\boldsymbol{x} \in \mathbb{R}^3_{\geq 0} : A\boldsymbol{x} \leq \boldsymbol{b}\right\}$, where

$$A = \begin{pmatrix} 2 & 5 & 7 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \ \boldsymbol{b} = \begin{pmatrix} 12 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

We want to maximize the objective function $\boldsymbol{p}(\boldsymbol{x}) = 3x_1 + 5x_2 + 2x_3$ on P. Notice that P is the intersection of the unit cube $C = [0,1]^3$ with the halfspace

$$\mathcal{H} = \{\boldsymbol{x} \in \mathbb{R}^3 : 2x_1 + 5x_2 + 7x_3 \leq 12\},$$

and hence vertices of P lie on the edges of the unit cube C. These vertices are not hard to find in this case – they are the origin $\mathbf{0}$, the standard basis vectors $\boldsymbol{e}_1$, $\boldsymbol{e}_2$,

$\boldsymbol{e}_3$, their pairwise sums $\boldsymbol{e}_1+\boldsymbol{e}_2$, $\boldsymbol{e}_1+\boldsymbol{e}_3$, $\boldsymbol{e}_2+\boldsymbol{e}_3$ (this last one lying in the bounding hyperplane of $\mathcal{H}$), as well as the two more points in the bounding hyperplane of $\mathcal{H}$:

$$\boldsymbol{v}_1 = (1, 1, 5/7),\ \boldsymbol{v}_2 = (1, 3/5, 1).$$

We can now describe the geometric idea of the simplex algorithm in this example.

(1) *Start at the vertex $\mathbf{0}$ and pick the direction towards any vertex connected to it by an edge of P, since $\boldsymbol{p}(\mathbf{0}) = 0$. Say, we pick $\boldsymbol{e}_1$.*

(2) *Move to $\boldsymbol{e}_1$, where $\boldsymbol{p}(\boldsymbol{e}_1) = 3$. Pick a neighboring vertex with a larger value of $\boldsymbol{p}$, say $\boldsymbol{e}_1 + \boldsymbol{e}_2$.*

(3) *Move to $\boldsymbol{e}_1+\boldsymbol{e}_2$, where $\boldsymbol{p}(\boldsymbol{e}_1+\boldsymbol{e}_2) = 8 > \boldsymbol{p}(\boldsymbol{e}_1)$. Pick a neighboring vertex with a larger value of $\boldsymbol{p}$, which is $\boldsymbol{v}_1$.*

(4) *Move to $\boldsymbol{v}_1$, where $\boldsymbol{p}(\boldsymbol{v}_1) = 66/7 > \boldsymbol{p}(\boldsymbol{e}_1 + \boldsymbol{e}_2)$. No neighboring vertex of $\boldsymbol{v}_1$ gives a larger value of $\boldsymbol{p}$, thus we stop.*

(5) *Return the maximum value of $\boldsymbol{p}$ on P, which is $\boldsymbol{p}(\boldsymbol{v}_1) = 66/7$.*

Let us compare this result to the result yielded by Dantzig's Theorem 2.2.1 (the assumptions for this theorem are satisfied here). Ordering the items in our instance of CKP so that (2.3) *is satisfied, we have:*

$$\frac{p_1}{w_1} = \frac{3}{2} > \frac{p_2}{w_2} = \frac{5}{5} > \frac{p_3}{w_3} = \frac{2}{7}.$$

Then the critical index is $s = 3$. Hence the theorem guarantees that the optimal solution is

$$\left(1, 1, \frac{1}{7}(12 - (2 + 5))\right) = (1, 1, 5/7),$$

as expected.

This example demonstrates that in case the case our linear program is an instance of CKP the simplex algorithm essentially reduces to Theorem 2.2.1, however it applies far more generally than just CKP which is a big advantage. This being said, it still only applies to instance of LP, not ILP, and hence it does not directly help us with the knapsack problems. On the other hand, a knapsack problem can be formulated as maximization problem for an objective function on the set of *integer lattice points* (i.e. points of $\mathbb{Z}^n$) inside of the specific compact LP-polytope defined by the corresponding constraints. Such a set is finite, so if we could find all these integer lattice points, we could simply evaluate our objective function at all of them and pick the largest value. While not necessarily efficient, this would lead to a solution. The problem is that integer lattice points in polytope are difficult not only to find, but even to count. We will discuss such a counting problem in more details in the next section.

2.4. Integer knapsack and counting integer lattice points in polytopes

Let us start by defining a certain variation of the knapsack problems that is somewhat different from the previous versions we were discussing. This is a decision problem known as the *integer knapsack problem (IKP)*:

Given a set of weights $S = \{w_1, \dots, w_n\}$ and the target value t, do there exist $x_1, \dots, x_n \in \mathbb{Z}_{\geq 0}$ such that $\sum_{i=1}^n w_i x_i = t$?

Notice that this is a generalization of the decision version of SSP, where the variables $x_1, \dots, x_n$ were only allowed to take on values 0 or 1. Thus this problem is NP-hard. We can reformulate this problem geometrically by introducing the *knapsack polytope*

$$P(S,t) := \left\{ \boldsymbol{x} \in \mathbb{R}^n_{\geq 0} : \sum_{i=1}^n w_i x_i = t \right\}.$$

The problem then is to determine whether $P(S,t) \cap \mathbb{Z}^n = \emptyset$. In other words, we can ask whether the counting function $|P(S,t) \cap \mathbb{Z}^n| > 0$, which naturally leads to the question of counting integer lattice points in polytopes. This is the main focus of this section.

We discuss the following general question: given a compact convex polytope $P \subset \mathbb{R}^n$, what is the number of integer lattice points in it? In other words, we want to find the quantity $|P \cap \mathbb{Z}^n|$. Let us start with the two-dimensional situation, where we can prove a beautiful formula for even a somewhat more general situation. Let P be a *simple polygon* (with no holes or self-intersections) in $\mathbb{R}^2$, not necessarily convex, with integer vertices. Let us write $A(P)$ for the area of P, $I(P)$ for the number of integer lattice points in the interior of P and $B(P)$ for the number of integer lattice points on the boundary of P. The following famous theorem was proved by Georg Alexander Pick in 1899.

Theorem 2.4.1 (Pick's Theorem).

$$A(P) = I(P) + \frac{1}{2}B(P) - 1.$$

Sketch of Proof. Let $\boldsymbol{x} \in P \cap \mathbb{Z}^2$ and define $\alpha_P(\boldsymbol{x})$ to be the visibility angle of P from $\boldsymbol{x}$, i.e. it is the angle of the cone $C(\boldsymbol{x}) \cap P$ where $C(\boldsymbol{x})$ is a unit circle centered at $\boldsymbol{x}$. Notice that

$$\alpha_P(\boldsymbol{x}) = \begin{cases} 2\pi & \text{if } \boldsymbol{x} \text{ is an interior point of } P, \\ \pi & \text{if } \boldsymbol{x} \text{ is on the boundary of } P \text{ but not a vertex,} \end{cases}$$

and $\alpha_P(\boldsymbol{x})$ is the corresponding interior angle of P if $\boldsymbol{x}$ is a vertex. Define the weight enumerator

$$W(P) = \sum_{\boldsymbol{x} \in P \cap \mathbb{Z}^2} \frac{\alpha_P(\boldsymbol{x})}{2\pi}.$$

We will now sketch a proof of the formula $A(P) = W(P)$. First observe that $W(P)$ is additive, i.e. if $P = P_1 \cup P_2$ where P_1 and P_2 are polytopes sharing a piece of boundary then

$$W(P) = W(P_1) + W(P_2). \tag{2.7}$$

To see this, notice that interior points of P_1, P_2 remain interior points of P, boundary points on non-overlapping parts of the boundary remain boundary points, whereas a boundary point on the common part of the boundary of P_1 and P_2 is either a vertex or becomes an interior point. If such a point $\boldsymbol{x}$ is a vertex, then $\alpha_P(\boldsymbol{x}) = \alpha_{P_1}(\boldsymbol{x}) + \alpha_{P_2}(\boldsymbol{x})$; if $\boldsymbol{x}$ was on the joint boundary and became an interior point then it was counted with $\alpha_{P_1}(\boldsymbol{x}) = \pi$ in $W(P_1)$, $\alpha_{P_2}(\boldsymbol{x}) = \pi$ in $W(P_2)$ and will now be counted with $\alpha_P(\boldsymbol{x}) = 2\pi$ in $W(P)$.

The verification of the formula $A(P) = W(P)$ for rectangles and triangles is done in Problem 2.3. Now notice that any polygon P can be split into a union of triangles with non-overlapping interiors but possibly joint boundaries. This observation together with Problem 2.3 and (2.7) implies $A(P) = W(P)$ for all polygons.

Now let n = number of vertices of P, $m = B(P) - n$ = number of boundary integer lattice points that are not vertices, and $k = I(P)$ = number of internal integer lattice points. Let

$$\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\},\ \{\boldsymbol{y}_1, \dots, \boldsymbol{y}_m\},\ \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_k\}$$

be these sets of points, respectively. Recall that the sum of internal angles of P is $(n-2)\pi$ – this formula holds in general whether P is convex or not. Then

$$\begin{aligned} W(P) &= \frac{1}{2\pi}\left(\sum_{i=1}^{n} \alpha_P(\boldsymbol{x}_i) + \sum_{i=1}^{m} \alpha_P(\boldsymbol{y}_i) + \sum_{i=1}^{k} \alpha_P(\boldsymbol{z}_i)\right) \\ &= \frac{n-2}{2} + \frac{m}{2} + k = k + \frac{n+m}{2} - 1 = I(P) + \frac{1}{2}B(P) - 1. \end{aligned}$$

This completes the proof. □

Next we discuss the problem of counting integer lattice points in convex polytopes in dimensions ≥ 3. Specifically, we address the following question: how can we count the number of integer lattice points in homogeneous expansions of polytopes? An area of mathematics that aims to answer this question is called *Ehrhart theory*. Let $P \subseteq \mathbb{R}^n$ be a convex polytope such that $\mathrm{Vol}(P) > 0$, and vertices of P are points of $\mathbb{Z}^n$: such P is called a *lattice polytope*. Write

$$G_P(t) = |tP \cap \mathbb{Z}^n|.$$

We want to understand the behavior of $G_P(t)$ for all $t \in \mathbb{Z}_{>0}$; specifically, we will prove a famous theorem of Ehrhart, which states that $G_P(t)$ is a polynomial in t. Our presentation closely follows [**Ewa96**]. First we consider a special case of polytopes, namely simplices.

LEMMA 2.4.2. *Let $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n \in \mathbb{Z}^n$ be linearly independent, and define the simplex*

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

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

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

PROOF. Let A be the half-open parallelotope spanned by the vectors $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n$, i.e.

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

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

$$\boldsymbol{y} = \boldsymbol{x} + \sum_{i=1}^{n} \alpha_i \boldsymbol{a}_i, \tag{2.8}$$

where $\boldsymbol{x} \in A \cap \mathbb{Z}^n$ 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 $j\boldsymbol{a}_1, \dots, j\boldsymbol{a}_n$. We will determine the number of points of $\mathbb{Z}^n$ in $H_j \cap tS$, and the number of points of $\mathbb{Z}^n \cap tS$ in the strips of space bounded by H_{j-1} and H_j for each $1 \leq j \leq t$; notice that $H_0 = \{\mathbf{0}\}$.

First, let $\boldsymbol{x} = \mathbf{0}$ in (2.8). Then $\boldsymbol{y}$ as in (2.8) lies in H_j if and only if

$$\sum_{i=1}^{n} \alpha_i = j, \ 0 \leq \alpha_i \leq j \ \forall \ 1 \leq i \leq n. \tag{2.9}$$

We will prove now that there are precisely $\binom{n+j-1}{n-1}$ possibilities for $\alpha_1, \dots, \alpha_n$ satisfying (2.9) 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 (2.9) is

$$\sum_{\alpha_n=0}^{j} \binom{n+(j-\alpha_n)-2}{n-2} = \sum_{i=0}^{j} \binom{n+i-2}{n-2}. \tag{2.10}$$

Then our claim follows by combining (2.10) with the result of Problem 2.6:

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

Now to find the number of points $\boldsymbol{y}$ as in (2.8) with $\boldsymbol{x} = \mathbf{0}$ on $\bigcup_{j=0}^{t} H_j$, we sum over j, using the result of Excercise 2.6 once again:

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

If $\boldsymbol{x}$ in (2.8) lies properly between H_0 and H_1, then the number of possible $\boldsymbol{y}$ as given by (2.8) that lie in $\bigcup_{j=0}^{t} H_j$ reduces to $\binom{n+t-1}{n}$. Similarly, the number of possibilities for $\boldsymbol{y}$ as in (2.8) with $\boldsymbol{x}$ lying properly between H_{i-1} and H_i or on H_i is $\binom{n+t-i}{n}$ for each $1 \leq i \leq n$. Therefore, if β_i is the number of points $\boldsymbol{x} \in A \cap \mathbb{Z}^n$ which lie properly between H_{i-1} and H_i or on H_i, then the number of corresponding points $\boldsymbol{y}$ as in (2.8) is

$$\binom{n+t-i}{n}\beta_i.$$

Finally, in the case $t < n$, we let $\beta_i = 0$ for each $t+1 \leq i \leq n$. The statement of the lemma follows.

□

Let $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n \in \mathbb{Z}^n$ be linearly independent, and let

$$S = \mathrm{Co}(\mathbf{0}, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)$$

be the simplex as in Lemma 2.4.2. Define the *pseudo-simplex* associated with S

$$S_0 = S \setminus (\mathrm{Co}(\mathbf{0}, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_{n-1}) \ \cup\ \ldots\ \cup\ \mathrm{Co}(\mathbf{0}, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n))\,.$$

LEMMA 2.4.3. *$G(tS_0)$ is a polynomial in $t \in \mathbb{Z}_{\geq 0}$.*

PROOF. We argue by induction on dimension of S_0. If $\dim(S_0) = 0$, there is nothing to prove, so assume the lemma is true for pseudo-simplices of dimension $< n$. Let $F^{(1)}, \ldots, F^{(s)}$ be proper faces of S which contain $\mathbf{0}$ and satisfy

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

Then

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

is a disjoint union. By induction hypothesis,

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

is a polynomial in t. Hence, by Lemma 2.4.2,

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

is a polynomial in t. □

We are now ready to prove a theorem of Eugene Ehrhart's from the 1960s.

THEOREM 2.4.4 (Ehrhart). *Let P be a lattice polytope in $\mathbb{R}^n$. Then $G_P(t)$ is a polynomial in $t \in \mathbb{Z}_{\geq 0}$.*

PROOF. We can assume $\mathbf{0}$ to be a vertex of P, since such translation would not change the number of integer lattice points. Notice that each $(n-1)$-dimensional face of P which does not contain $\mathbf{0}$ can be given a decomposition as a simplicial complex whose 0-cells are the vertices of this face. We can then join each simplex, obtained in this manner, to $\mathbf{0}$ resulting in a decomposition of P into a simplicial complex whose 0-cells are precisely the vertices of P. Then P can be represented as a disjoint union

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

where $S_0^{(1)}, \ldots, S_0^{(r)}$ are precisely the cells of this simplicial complex which contain $\mathbf{0}$, but are not equal to $\{\mathbf{0}\}$. The theorem follows by Lemma 2.4.3. □

$G_P(t)$ as in Theorem 2.4.4 is called *Ehrhart polynomial* of P. An excellent reference on Ehrhart polynomials, their many fascinating properties, and connections to other important mathematical objects is [**BR06**]. For a general lattice polytope P very little is known about the coefficients of its Ehrhart polynomial $G_P(t)$. Let

$$G_P(t) = \sum_{i=0}^{n} c_i(P) t^i,$$

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

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

The rest of the coefficients of $G_P(t)$ are in general unknown, however there are known relations and identities that they satisfy; see [**BR06**] for further details.

Let us present the first simple example of Ehrhart polynomial. Consider the n-dimensional cube of sidelength 2 centered at the origin:

$$C_n = \{\boldsymbol{x} \in \mathbb{R}^n : |\boldsymbol{x}| \leq 1\}, \tag{2.11}$$

then for each $t \in \mathbb{Z}_{>0}$

$$|tC_n \cap \mathbb{Z}^n| = (2t+1)^n = \sum_{i=0}^{n} 2^i \binom{n}{k} t^i$$

is the corresponding Ehrhart polynomial. We will give two more explicit examples of Ehrhart polynomial. The first one is for an open simplex, which is precisely the interior of the simplex S of Lemma 2.4.2 with $\boldsymbol{a}_i = \boldsymbol{e}_i$ for each $1 \leq i \leq n$; the following observation along with the proof is due to S. I. Sobolev.

Proposition 2.4.5. *Define an open simplex*

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

Then $G_{S^\circ}(t) = 0$ *if* $t \leq n$, *and for every* $t \in \mathbb{Z}_{>n}$,

$$G_{S^\circ}(t) = \binom{t-1}{n}. \tag{2.12}$$

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

$$tS_1^\circ = \{\boldsymbol{x} \in \mathbb{R}^n : 0 < x_1 < \cdots < x_k < t\}.$$

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

Notice that (2.12) can be thought of as a geometric interpretation of binomial coefficients. The next example is closely related to the one in Proposition 2.4.5: it has been established in [**BCKV00**].

Proposition 2.4.6. *Let*

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

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

$$G_{\mathcal{S}_n}(t) = \sum_{i=0}^{\min\{t,n\}} 2^i \binom{n}{i} \binom{t}{i}. \tag{2.13}$$

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

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

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

Remark 2.4.1. A remarkable property of the polynomial in Proposition 2.4.6 is that the right hand side (2.13) is symmetric in t and n. This means that

$$|t\mathcal{S}_n \cap \mathbb{Z}^n| = |n\mathcal{S}_t \cap \mathbb{Z}^t|.$$

2.5. Frobenius problem

In this section we introduce a problem closely related to the IKP. For $n \geq 2$, consider integers

$$1 < a_1 < \cdots < a_n \text{ such that } \gcd(a_1, \ldots, a_n) = 1, \tag{2.14}$$

and define

$$S(a_1, \ldots, a_n) := \left\{ \sum_{i=1}^{n} a_i x_i : x_1, \ldots, x_n \in \mathbb{Z}_{\geq 0} \right\}.$$

This is an example of a *numerical semigroup*, i.e. a subset of $\mathbb{Z}_{\geq 0}$ containing 0 which is closed under addition. The set $\mathbb{N} \setminus S(a_1, \ldots, a_n)$ is called the set of *gaps* of $S(a_1, \ldots, a_n)$.

THEOREM 2.5.1. *The set of gaps of the numerical semigroup $S(a_1, \ldots, a_n)$ under the condition* (2.14) *is finite.*

PROOF. Since the n-tuple $a_1, \ldots, a_n$ is relatively prime, there exist integers $m_1, \ldots, m_n$ such that

$$a_1 m_1 + \cdots + a_n m_n = 1.$$

This sum has some positive and some negative terms, hence it can be written $A - B = 1$, where A and B are both nonnegative integer linear combinations of $a_1, \ldots, a_n$. Therefore $A, B \in S(a_1, \ldots, a_n)$. Now, let z be any positive integer, then Euclid's division lemma implies that

$$z = qa_1 + r, \ q, r \in \mathbb{Z}_{\geq 0}, \ 0 \leq r < a_1.$$

On the other hand, $r = r \times 1 = r(A - B)$. Notice that $qa_1 \in S(a_1, \ldots, a_n)$, and hence

$$z + (a_1 - 1)B = qa_1 + r(A - B) + (a_1 - 1)B = qa_1 + rA + (a_1 - r - 1)B \in S(a_1, \ldots, a_n),$$

since $a_1 - r - 1 \geq 0$. This implies that every integer $\geq (a_1 - 1)B$ is in $S(a_1, \ldots, a_n)$, and hence the number of gaps at most $(a_1 - 1)B - 1$. □

REMARK 2.5.1. The presentation of the above argument closely followed [**Ram05**].

The *Frobenius number* $g(a_1, \ldots, a_n)$ is defined to be the largest gap of $S(a_1, \ldots, a_n)$, i.e. the largest integer t that cannot be expressed in the form $t = \sum_{i=1}^{n} a_i x_i$ for some nonnegative integers $x_1, \ldots, x_n$. With this notation, we can formulate the *Frobenius Problem (FP)*:

Given an n-tuple $a_1, \ldots, a_n$ satisfying (2.14) find $g(a_1, \ldots, a_n)$.

Notice that FP can be stated in terms of the knapsack polytopes as follows: writing $\boldsymbol{a} = (a_1, \ldots, a_n)$, find the smallest positive integer g so that $P(\boldsymbol{a}, t) \cap \mathbb{Z}^n \neq \emptyset$ for every $t > g$. Alternatively, we can write it as follows:

$$\text{Find } \min\{g \in \mathbb{Z}_{>0} : |P(\boldsymbol{a}, t) \cap \mathbb{Z}^n| > 0 \ \forall \ b > t\}.$$

Notice also that

$$\sum_{i=1}^{n} a_i x_i = t \iff \sum_{i=1}^{n} a_i \left(\frac{x_i}{t} \right) = 1,$$

i.e. $\boldsymbol{x} \in P(\boldsymbol{a}, t)$ if and only if $\frac{1}{t}\boldsymbol{x} \in P(\boldsymbol{a}, 1)$, meaning that $P(\boldsymbol{a}, t) = tP(\boldsymbol{a}, 1)$ is a homogeneous expansion of a polytope that we discussed in the previous section. This, however, does not necessarily apply directly to the integer lattice points, since for $\boldsymbol{x} \in \mathbb{Z}^n$ the rescaled points $\frac{1}{t}\boldsymbol{x}$ may no longer be in $\mathbb{Z}^n$. Due to this knapsack connection, it is not surprising that FP is known to be NP-hard (specifically, there a polynomial-time algorithm that reduces IKP to FP). We begin our discussion of FP with the simplest case $n = 2$.

Let us start with a simple binary linear Diophantine equation of the form

$$ax + by = c, \tag{2.15}$$

in which a, b, c are nonzero integers. There are always rational solutions to (2.15). For which values of a, b, c does it have solutions in integers x, y? The greatest common divisor provides a criterion for the existence of solutions.

Lemma 2.5.2. *Let a, b, c be nonzero integers. Then* (2.15) *has a solution in integers x, y if and only if* $\gcd(a, b)|c$.

Proof. ($\Rightarrow$) Suppose that $ax + by = c$ for some $x, y \in \mathbb{Z}$. Since $\gcd(a, b)$ divides a and b, it divides $ax + by = c$.
($\Leftarrow$) If $\gcd(a, b)|c$, write $c = d\gcd(a, b)$, in which $d \in \mathbb{Z}$. By Euclid's Division Lemma, there exist $x', y' \in \mathbb{Z}$ such that $ax' + by' = \gcd(a, b)$. Thus, $a(dx') + b(dy') = d(ax' + by') = d\gcd(a, b) = c$ and hence (2.15) has integer solutions $x = dx'$ and $y = dy'$. □

In fact, we can classify all integer solutions to (2.15).

Theorem 2.5.3. *Let a, b, c be nonzero integers, and let $d = \gcd(a, b)$. Assume $d|c$. Then the equation $ax + by = c$ has infinitely many integer solutions. In fact, if x_0, y_0 is one such solution pair, then all solutions are given by*

$$x_t = x_0 - t\frac{b}{d}, \; y_t = y_0 + t\frac{a}{d} \tag{2.16}$$

as t ranges over all the integers.

Proof. First let $t \in \mathbb{Z}$ and x_t, y_t be as in (2.16). Then

$$ax_t + by_t = a\left(x_0 - t\frac{b}{d}\right) + b\left(y_0 + t\frac{a}{d}\right) \quad = (ax_0 + by_0) + t\left(\frac{ab}{d} - \frac{ab}{d}\right) = c,$$

hence our pair x, y is a solution to (2.15) for any $t \in \mathbb{Z}$.

We now show that any solution is of this form. Indeed, suppose x, y is a solution pair, then

$$ax_0 + by_0 = c = ax + by,$$

and so

$$a(x_0 - x) = b(y - y_0).$$

Let us divide both sides of the above equation by d and write $a' = a/d$, $b' = b/d$, then $\gcd(a', b') = 1$ and

$$a'(x_0 - x) = b'(y - y_0).$$

Then Euclid's Lemma implies that $a'|y - y_0$ and $b'|x_0 - x$, say $a' = \frac{y - y_0}{t}$ and $b' = \frac{x_0 - x}{s}$ for some integers t and s. Then we have

$$\frac{(y - y_0)(x_0 - x)}{t} = \frac{(x_0 - x)(y - y_0)}{s},$$

and so $s = t$. Therefore we obtain

$$y = y_0 + a't, \; x = x_0 - b't,$$

which is precisely what we wanted. □

Corollary 2.5.4. *If* $\gcd(a,b) = 1$*, then for any* c *the equation* $ax + by = c$ *has infinitely many solutions. Furthermore, if* x_0, y_0 *is one such solution pair, then all solutions are of the form*

$$x_t = x_0 - tb,\ y_t = y_0 + ta$$

for $t \in \mathbb{Z}$*.*

Example 2.5.1. *Let* $a = 4$*,* $b = 6$*,* $c = 9$*. Since* $\gcd(a,b) = 2 \nmid 9$*, the equations* $4x+6y = 9$ *has no integer solutions. On the other hand, if* $c = 10$*, then* $\gcd(a,b)|c$*, and so the equation* $4x+6y = 10$ *has infinitely many integer solutions. Since* $x = 1$*,* $y = 1$ *is one such solution, all solutions are of the form*

$$x_t = 1 - 3t, y_t = 1 + 2t$$

as t *ranges over all the integers.*

These observations also have a simple geometric interpretation. Notice that the set of integer solution pairs to (2.15)

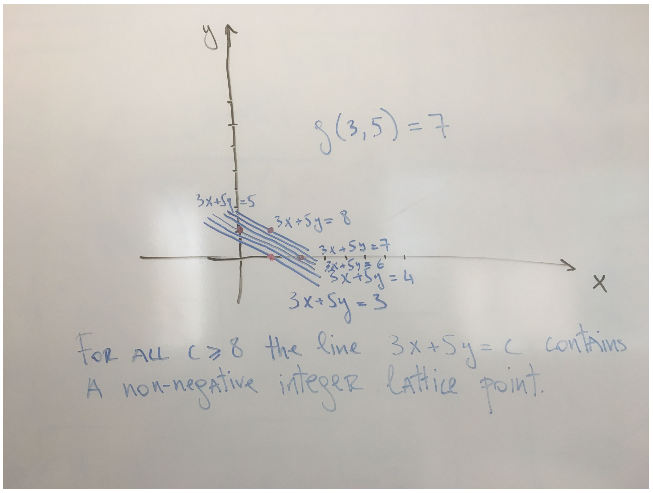
$$\big\{(x,y) \in \mathbb{Z}^2 : ax + by = c\big\}$$

is the set of all integer lattice points on the line given by the equation (2.15) in the Euclidean plane. For instance, the set of all such points in the case $a = 4, b = 6, c = 10$ of Example 2.5.1 is $\{(1 - 6t, 1 + 4t) : t \in \mathbb{Z}\}$.

Assume now that $c > 0$ and $\gcd(a,b)$ divides c, so the line $ax + by = c$ contains infinitely many integer lattice points, but does it necessarily contain any such points with nonnegative coordinates? Upon a quick inspection, we can see for instance that the line

$$3x + 5y = c \tag{2.17}$$

contains integer lattice points for any c, but no such points with $x, y \geq 0$ when $c = 1, 2, 4$. For which values of c is our line guaranteed to have nonnegative integer lattice points?



For all c ⩾ 8 the line 3x+5y = c contains A non-negative integer lattice point.

Here is an initial observation, which follows from Theorem 2.5.3 via a geometric argument.

COROLLARY 2.5.5. *Let a, b, c be positive integers with $d := \gcd(a, b)$ dividing c. If $c \geq ab/d$, then the equation* (2.15) *has integer solution pairs $x, y \geq 0$.*

PROOF. Let $t, s \in \mathbb{Z}$ and consider the solution pairs (x_t, y_t) and (x_s, y_s), as in (2.16), where (x_0, y_0) is some fixed solution pair. Notice that the Euclidean distance between the points (x_t, y_t) and (x_s, y_s) is

$$\sqrt{(x_t - x_s)^2 + (y_t - y_s)^2} = \sqrt{\frac{b^2}{d^2}(t-s)^2 + \frac{a^2}{d^2}(t-s)^2} = \frac{|t-s|\sqrt{a^2+b^2}}{d},$$

which is minimized when $|t - s| = 1$. Let $\ell_{a,b}(c)$ be the line $ax + by = c$ in the Euclidean plane, then the minimal distance between two integer lattice points on $\ell_{a,b}(c)$ is $\frac{\sqrt{a^2+b^2}}{d}$, which is assumed for any neighboring pair of integer lattice points (x_t, y_t) and (x_{t+1}, y_{t+1}). Notice that the intersection of the line $\ell_{a,b}(c)$ with the positive quadrant

$$\{(x, y) \in \mathbb{R}^2 : x, y \geq 0\}$$

is a line segment with endpoints $(c/a, 0)$ and $(0, c/b)$, so the length of this line segment is

$$\sqrt{\frac{c^2}{a^2} + \frac{c^2}{b^2}} = \frac{c\sqrt{a^2+b^2}}{ab}.$$

If the length of this line segment is no less than the distance between the neighboring integer lattice points, then the line segment must contain at least one integer lattice point. This means that when

$$\frac{c\sqrt{a^2+b^2}}{ab} \geq \frac{\sqrt{a^2+b^2}}{d},$$

the equation (2.15) has integer solution pairs $x, y \geq 0$. This happens when $c \geq ab/d$. □

Going back to the example of equation (2.17) and applying Corollary 2.5.5, we are guaranteed that there are nonnegative solutions at least for all $c \geq 15$. Checking by hand, we quickly see that in fact there are nonnegative solutions already for all $c \geq 8$, suggesting that the bound of Corollary 2.5.5 may not be very good. Indeed, we can obtain more precise results.

Let a, b be relatively prime positive integers, and suppose that we have unlimited supply of coins of denominations a and b. What is the maximal amount of change which we *cannot* give with such coins? This is precisely the Frobenius number $g(a, b)$ and we know from Corollary 2.5.5 that

$$g(a, b) < ab.$$

But is there an exact formula? This problem, although possibly in different terms was mentioned in the lectures of a famous German mathematician Ferdinand Georg Frobenius in the late 1800s, although Frobenius himself never published anything in these regards. Nonetheless, this problem became known as the (binary) Frobenius coin exchange problem with the maximal impossible amount of change $g(a, b)$ being the Frobenius number of a and b. Interestingly, closely related problems also appear in recreational mathematical literature under different names, such as postage stamp problem or the chicken McNugget problem. The origins of the latter name are curious: apparently, in the 1980s chicken McNuggets were sold by McDonalds in the UK in boxes of 3, 6 and 20 pieces, prompting a mathematician Henri Picciotto to ask what is the maximal number of nuggets that cannot be purchased (and then answering his own question – it is 43). Let us now derive a formula for the binary Frobenius number.

THEOREM 2.5.6. *Let* $\gcd(a,b)=1$*, then*

$$g(a,b)=(a-1)(b-1)-1.$$

In other words, this is the largest number that cannot be represented as $ax+by$ *with* x,y *nonnegative integers.*

PROOF. Since a and b are relatively prime, for every $c\in\mathbb{Z}$ there exist $x,y\in\mathbb{Z}$ such that

$$c=ax+by.$$

We will say that c is *representable* in terms of a and b if there exist such $x,y\geq 0$. Notice in fact that we can assume without loss of generality that $0\leq x<b$: if $x\geq b$, then $x=nb+x'$ for some $n,x'\in\mathbb{Z}$ with $0\leq x'<b$, and so

$$c=a(nb+x')+by=ax'+b(an+y),$$

meaning that we can replace x with x' by replacing y with $an+y$, if necessary.

Now, if $0\leq x<b$, then for every c there is a unique pair (x,y) such that $c=ax+by$, and so c is representable if and only if $y\geq 0$. Notice then that the largest non-representable c corresponds to the largest choice of x (namely, $x=b-1$) and the largest negative choice of y (namely, $y=-1$). This means that the largest non-representable integer is

$$g(a,b)=a(b-1)+b(-1)=ab-a-b=(a-1)(b-1)-1.$$

□

Theorem 2.5.6 therefore guarantees that for every $c>ab-a-b$ the line $ax+by=c$ contains a nonnegative integer lattice point, however for $c<ab-a-b$ such a point may or may not exist. Revisiting for instance our example (2.17), we see that while $g(3,5)=7$, the equation $3x+5y=c$ has nonnegative integer solutions for $c=3,5,6$, but does not for $c=1,2,4,7$, i.e. these are gaps of $S(3,5)$. Given a and b, we can ask how many gaps are there? This natural question was asked as a challenge problem in a journal called Educational Times by James Joseph Sylvester in 1884. Specifically, Sylvester, who has already obtained and published the answer himself in 1882, asked for a proof that this number is equal to $\frac{1}{2}(a-1)(b-1)$; in other words, out of $(a-1)(b-1)-1$ integers between 1 and the Frobenius number $g(a,b)$ about half are non-representable. A clever solution was produced by W. J. Curran Sharp. We prove this result here.

THEOREM 2.5.7. *The number of gaps with respect to a relatively prime pair of positive integers* a *and* b *is*

$$\frac{1}{2}(a-1)(b-1).$$

PROOF. Let $0\leq c\leq g(a,b)$, and define

$$c'=g(a,b)-c=ab-a-b-c.$$

By our argument in the proof of Theorem 2.5.6, there must exist the unique integers x,y with $0\leq x<b$ such that $c=ax+by$, then

$$c'=ab-a-b-c=ab-a-b-ax-by=ax'+by',$$

where $x' = b - x - 1$ and $y' = -y - 1$. Since $0 \leq x' < b$, we see that y' must also be unique.

Suppose that c is representable by a and b (including $c = 0$), then $y \geq 0$, and $y' < 0$, hence c' is not representable. On the other hand, assume that c is not representable, then $y < 0$, and so $y' \geq 0$, meaning that c' is representable. It is clear that c and c' are in a bijection with each other, and $c = c'$ if and only if

$$c = \frac{1}{2}(ab - a - b),$$

but this cannot be an integer, since a and b cannot both be even. Hence precisely a half of $g(a, b) + 1$ integers between 0 and $g(a, b)$ are representable and the rest are gaps, meaning that there are

$$\frac{1}{2}(g(a, b) + 1) = \frac{1}{2}(a - 1)(b - 1)$$

gaps. □

The Frobenius number has also been defined more generally. Let $n \geq 2$ be an integer and let

$$1 < a_1 < \cdots < a_n \tag{2.18}$$

be relatively prime integers. We say that a positive integer t is *representable* by the n-tuple $\boldsymbol{a} := (a_1, \ldots, a_n)$ if

$$t = a_1 x_1 + \cdots + a_n x_n \tag{2.19}$$

for some nonnegative integers $x_1, \ldots, x_n$, and we call each such solution $\boldsymbol{x} := (x_1, \ldots, x_n)$ of (2.19) a *representation for t in terms of $\boldsymbol{a}$*. Let $s \geq 0$ be an integer, then the *s-Frobenius number* of this n-tuple, $g_s(\boldsymbol{a})$, as defined by Beck and Robins in [**BR04**], is the largest positive integer that has at most s distinct representations in terms of $\boldsymbol{a}$. In the binary case ($n = 2$), Beck and Robins proved the following natural generalization of Theorem 2.5.6.

THEOREM 2.5.8. *Let* $\gcd(a, b) = 1$ *and* $s \geq 0$*, then*

$$g_s(a, b) = (s + 1)ab - (a + b).$$

In the case $s = 0$, the formula of Theorem 2.5.6 is recovered.

This is a generalization of the classical Frobenius number $g_0(\boldsymbol{a})$, i.e., the largest positive integer that has no such representations. The Frobenius number has been studied extensively by a variety of authors, starting as early as late 19th century; see [**Ram05**] for a detailed account and bibliography. Generalizing Theorem 2.5.1, the condition

$$\gcd(a_1, \ldots, a_n) = 1 \tag{2.20}$$

implies that $g_s(\boldsymbol{a})$ exists for every s, but the NP-hardness of FP (and the fact that P vs NP is an open problem) in particular implies that no general closed form formulas for the Frobenius numbers is known, sparking interest in upper and lower bounds. Frobenius numbers and their various generalizations tend to play an important role in several areas of mathematics, including theory of numerical semigroups, commutative algebra, algebraic geometry, number theory, combinatorics, operations research, and theoretical computer science, to name a few. The literature on this subject is vast with a large number of relevant references available in the bibliography to the book [**Ram05**]. We will talk more about the Frobenius number and its beautiful geometric connections in Section 4.5.

2.6. Problems

Problem 2.1. *Consider the Boolean formula*

$$B = (x \wedge y) \vee (\neg x \wedge \neg y) \vee (\neg x \wedge z) \vee (y \wedge \neg z).$$

Part a. *Rewrite B in generalized conjunctive normal form.*
Part b. *Rewrite the formula you obtained in part a in the form with clauses consisting of three literals each.*
Part c. *Construct the table to reduce your formula from part b to an instance of SSP, as in the proof of Theorem 2.1.2.*
Part d. *Decide if the instance of SSP you obtained is solvable or not. If so, what values of the Boolean variables make the formula from part b satisfiable? How about the original formula B?*

Problem 2.2. *Consider the following instance of BKP:*

$$\begin{aligned} &\textit{maximize } 4x_1 + 5x_2 + 7x_3 \\ &\textit{subject to } 6x_1 + 3x_2 + 5x_3 \leq 12, \\ &x_1, x_2, x_3 \in \{0, 1\}. \end{aligned}$$

Part a. *Use Theorem 2.2.1 to solve the CKP relaxation of this problem.*
Part b. *Define the corresponding LP-polytope for the CKP problem in part a and find its vertices.*
Part c. *Use the geometric description of the simplex method as in Example 2.3.1 to find the solution. Make sure it is consistent with part a.*
Part d. *List all the integer lattice points in the LP-polytope from part b and solve BKP by evaluating the objective function at each one of them and comparing. Was the solution to CKP that you found a good approximation to the BKP solution?*

Problem 2.3. *Prove the formula $A(P) = W(P)$ as in the proof of Pick's Theorem (Theorem 2.4.1) for rectangles and triangles.*

Hint: *First prove it for a rectangle, then for a right triangle – splitting rectangle into two of them and applying additivity, and then for an arbitrary triangle by embedding it into a rectangle and applying additivity.*

Problem 2.4. *Suppose P is a polygon with integer vertices and h simple polygonal holes, each also with integer vertices. Can you generalize Pick's theorem to P?*

Problem 2.5. *Let n be a positive integer. The* **Farey series** *F_n of order n is the set of all reduced nonnegative rationals in the interval $[0, 1]$ with denominators no bigger than n written in increasing order, e.g.*

$$F_5 = \left\{ \frac{0}{1}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{2}{5}, \frac{1}{2}, \frac{3}{5}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \frac{1}{1} \right\}.$$

Let $\frac{a}{b}, \frac{c}{d} \in F_n$, where $n = \max\{b, d\}$. Let P be the parallelogram with vertices

$$(0, 0), (a, b), (c, d), (a + c, b + d).$$

Use Pick's theorem to prove that $\frac{a}{b}, \frac{c}{d}$ are neighbors in F_n if and only if the area of P is equal to 1.

Problem 2.6. *Prove that*

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

Problem 2.7. *Compute Ehrhart polynomials of a rectangle, a right triangle, and a right trapezoid.*

Problem 2.8. *Let a, b be positive relatively prime integers. Express the Frobenius number $g(a,b)$ in terms of areas of parallelograms with a side of length $\sqrt{a^2+b^2}$.*

Problem 2.9. *Let $a_1, \ldots, a_n$ be positive relatively prime integers.*

Part a. Express the s-Frobenius number $g_s(a_1, \ldots, a_n)$ in terms of the restricted partition function

$$p_{a_1,\ldots,a_n}(t) = \#\left\{(x_1, \ldots, x_n) \in \mathbb{Z}^n_{\geq 0} : \sum_{i=1} a_i x_i = t\right\}.$$

Part b. Prove the recursive formula

$$p_{a_1,\ldots,a_n}(t) = \sum_{m \geq 0} p_{a_1,\ldots,a_{n-1}}(t - m a_n).$$

Part c. Interpret the recursive formula from part b in terms of the s-Frobenius numbers.

Problem 2.10. *Let $a_1, \ldots, a_n$ be positive relatively prime integers. For each $s \geq 0$, define $S_s(a_1, \ldots, a_n)$ to be the set of all integers that have more than s representations in the form $\sum_{i=1}^n a_i x_i$ with n-tuple of nonnegative integers $x_1, \ldots, x_n \in \mathbb{Z}_{\geq 0}$. Prove that*

$$\cdots \subseteq S_s(a_1, \ldots, a_n) \subseteq S_{s-1}(a_1, \ldots, a_n) \subseteq \cdots \subseteq S_0(a_1, \ldots, a_n) = S(a_1, \ldots, a_n)$$

is a sequence of numerical semigroups.

Problem 2.11. *Let a, b and c, d be two pairs of positive relatively prime integers. Suppose*

$$g(a,b) \geq g(c,d).$$

Does this mean that $g_1(a,b) \geq g_1(c,d)$? Prove or give a counterexample.

CHAPTER 3

Geometry of Numbers

3.1. Lattices

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

Definition 3.1.1. A *lattice* Λ of *rank* r, $1 \leq r \leq n$, spanned by $\boldsymbol{a}_1, \dots, \boldsymbol{a}_r$ in $\mathbb{R}^n$ is the set of all possible linear combinations of the vectors $\boldsymbol{a}_1, \dots, \boldsymbol{a}_r$ with integer coefficients. In other words,

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

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

Notice that in general a lattice in $\mathbb{R}^n$ can have any rank $1 \leq r \leq n$. We will often however talk specifically about lattices of rank n, that is of full rank. The most obvious example of a lattice is the set of all points with integer coordinates in $\mathbb{R}^n$:

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

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

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

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

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

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

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

where

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

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

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

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

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

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

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

By Problems 3.3 and 3.4 a lattice Λ of rank r in $\mathbb{R}^n$ is a discrete co-compact subgroup of $V = \operatorname{span}_{\mathbb{R}} \Lambda$. In fact, the converse is also true.

THEOREM 3.1.1. *Let V be an r-dimensional subspace of $\mathbb{R}^n$, and let Γ be a discrete co-compact subgroup of V. Then Γ is a lattice of rank r in $\mathbb{R}^n$.*

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

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

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

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

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

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

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

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

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

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

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

We will now show that

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

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

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

Notice that

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

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

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

Then notice that

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

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

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

since $\boldsymbol{z}$ lies in the parallelotope spanned by $\boldsymbol{a}_1, \dots, \boldsymbol{a}_r$, and hence in P_r as in our construction above. Therefore $c_r = [c_r]$. We proceed in the same manner to conclude that $c_i = [c_i]$ for each $1 \leq i \leq r$, and hence $\boldsymbol{x} \in \operatorname{span}_{\mathbb{Z}}\{\boldsymbol{a}_1, \dots, \boldsymbol{a}_r\}$. Since this is true for every $\boldsymbol{x} \in \Gamma$, we are done. □

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

THEOREM 3.1.2. *Let Λ be a lattice of rank n in $\mathbb{R}^n$, and let A be a basis matrix for Λ. Then B is another basis matrix for Λ if and only if there exists an $n \times n$ integral matrix U with determinant ± 1 such that*

$$B = AU.$$

PROOF. First suppose that B is a basis matrix. Notice that, since A is a basis matrix, for every $1 \leq i \leq n$ the i-th column vector $\boldsymbol{b}_i$ of B can be expressed as

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

where $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n$ are column vectors of A, and u_{ij}'s are integers for all $1 \leq j \leq n$. This means that $B = AU$, where $U = (u_{ij})_{1 \leq i,j \leq n}$ is an $n \times n$ matrix with integer entries. On the other hand, since B is also a basis matrix, we also have for every $1 \leq i \leq n$

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

where w_{ij}'s are also integers for all $1 \leq j \leq N$. Hence $A = BW$, where $W = (w_{ij})_{1 \leq i,j \leq n}$ is also an $n \times n$ matrix with integer entries. Then

$$B = AU = BWU,$$

which means that $WU = I_n$, the $n \times n$ identity matrix. Therefore

$$\det(WU) = \det(W)\det(U) = \det(I_n) = 1,$$

but $\det(U), \det(W) \in \mathbb{Z}$ since U and W are integral matrices. This means that

$$\det(U) = \det(W) = \pm 1.$$

Next assume that $B = UA$ for some integral $n \times n$ matrix U with $\det(U) = \pm 1$. This means that $\det(B) = \pm \det(A) \neq 0$, hence column vectors of B are linearly independent. Also, U is invertible over $\mathbb{Z}$, meaning that $U^{-1} = (w_{ij})_{1 \leq i,j \leq n}$ is also an integral matrix, hence $A = U^{-1}B$. This means that column vectors of A are in the span of the column vectors of B, and so

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

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

Corollary 3.1.3. *If A and B are two basis matrices for the same lattice Λ, then*

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

Definition 3.1.2. The common determinant value of Corollary 3.1.3 is called the *determinant* of the lattice Λ, and is denoted by $\det(\Lambda)$.

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

Definition 3.1.3. If Λ and Ω are both lattices in $\mathbb{R}^n$, and $\Omega \subseteq \Lambda$, then we say that Ω is a *sublattice* of Λ.

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

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

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

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

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

Lemma 3.1.4. *Let Ω be a subattice of Λ. There exists a positive integer D such that $D\Lambda \subseteq \Omega$.*

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

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

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

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

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

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

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

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

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

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

which completes the proof. □

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

Theorem 3.1.5. *Let Λ be a lattice, and Ω a sublattice of Λ. For each basis $\boldsymbol{b}_1, \dots, \boldsymbol{b}_n$ of Λ, there exists a basis $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n$ of Ω of the form*

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

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

Proof. Let $\boldsymbol{b}_1, \dots, \boldsymbol{b}_n$ be a basis for Λ. We will first prove the existence of a basis $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n$ for Ω as claimed by the theorem. By Lemma 3.1.4, there exist integer multiples of $\boldsymbol{b}_1, \dots, \boldsymbol{b}_n$ in Ω, hence it is possible to choose a collection of vectors $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n \in \Omega$ of the form

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

for each $1 \leq i \leq n$ with $v_{ii} \neq 0$. Clearly, by construction, such a collection of vectors will be linearly independent. In fact, let us pick each $\boldsymbol{a}_i$ so that $|v_{ii}|$ is as small as possible, but not 0. We will now show that $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n$ is a basis for Ω. Clearly,

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

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

$$\Omega \subseteq \operatorname{span}_{\mathbb{Z}}\{\boldsymbol{a}_1, \dots, \boldsymbol{a}_n\}. \tag{3.1}$$

Suppose (3.1) is not true, then there exists $\boldsymbol{c} \in \Omega$ which is not in $\operatorname{span}_{\mathbb{Z}}\{\boldsymbol{a}_1, \dots, \boldsymbol{a}_n\}$. Since $\boldsymbol{c} \in \Lambda$, we can write

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

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

$$|t_k - s v_{kk}| < |v_{kk}|. \tag{3.2}$$

Then we clearly have

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

Therefore we must have $t_k - s v_{kk} \neq 0$ by minimality of k. But then (3.2) contradicts the minimality of $|v_{kk}|$: we could take $\boldsymbol{c} - s\boldsymbol{a}_k$ instead of $\boldsymbol{a}_k$, since it satisfies all the conditions that $\boldsymbol{a}_k$ was chosen to satisfy, and then $|v_{kk}|$ is replaced by the smaller nonzero number $|t_k - s v_{kk}|$. This proves that $\boldsymbol{c}$ like this cannot exist, and so (3.1) is true, hence finishing one direction of the theorem.

Now suppose that we are given a basis $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n$ for Ω. We want to prove that there exists a basis $\boldsymbol{b}_1, \dots, \boldsymbol{b}_n$ for Λ such that relations in the statement of the theorem hold. This is a direct consequence of the argument in the proof of Theorem 3.1.1. Indeed, at i-th step of the basis construction in the proof of Theorem 3.1.1, we can choose i-th vector, call it $\boldsymbol{b}_i$, so that it lies in the span of the previous $i-1$ vectors and the vector $\boldsymbol{a}_i$. Since $\boldsymbol{b}_1, \dots, \boldsymbol{b}_n$ constructed this way are linearly independent (in fact, they form a basis for Λ by the construction), we obtain that

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

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

In fact, it is possible to select the coefficients v_{ij} in Theorem 3.1.5 so that the matrix $(v_{ij})_{1 \leq i,j \leq n}$ is upper (or lower) triangular with non-negative entries, and the largest entry of each row (or column) is on the diagonal: we leave the proof of this to Problem 3.9.

Remark 3.1.1. Let the notation be as in Theorem 3.1.5. Notice that if A is any basis matrix for Ω and B is any basis for Λ, then there exists an integral matrix V such that $A = BV$. Then Theorem 3.1.5 implies that for a given B there exists an A such that V is lower triangular, and for for a given A exists a B such that V is lower triangular. Since two different basis matrices of the same lattice are always related by multiplication by an integral matrix with determinant equal to ±1, Theorem 3.1.5 can be thought of as the construction of *Hermite normal form* for an integral matrix. Problem 3.9 places additional restrictions that make Hermite normal form unique.

Here is an important implication of Theorem 3.1.5.

THEOREM 3.1.6. *Let $\Omega \subseteq \Lambda$ be a sublattice. Then $\frac{\det(\Omega)}{\det(\Lambda)}$ is an integer; moreover, the number of cosets of Ω in Λ, i.e. the index of Ω as a subgroup of Λ is*

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

PROOF. Let $\boldsymbol{b}_1, \dots, \boldsymbol{b}_n$ be a basis for Λ, and $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n$ be a basis for Ω, so that these two bases satisfy the conditions of Theorem 3.1.5, and write A and B for the corresponding basis matrices. Then notice that

$$B = AV,$$

where $V = (v_{ij})_{1 \leq i,j \leq n}$ is an $n \times n$ triangular matix with entries as described in Theorem 3.1.5; in particular $\det(V) = \prod_{i=1}^n |v_{ii}|$. Hence

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

which proves the first part of the theorem.

Moreover, notice that each vector $\boldsymbol{c} \in \Lambda$ is contained in the same coset of Ω in Λ as precisely one of the vectors

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

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

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

DEFINITION 3.1.4. A *fundamental domain* of a lattice Λ of full rank in $\mathbb{R}^n$ is a convex set $\mathcal{F} \subseteq \mathbb{R}^n$ containing $\boldsymbol{0}$, so that

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

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

In other words, a fundamental domain of a lattice $\Lambda \subset \mathbb{R}^n$ is a *full set of coset representatives of* Λ *in* $\mathbb{R}^n$ (see Problem 3.10). Although each lattice has infinitely many different fundamental domains, they all have the same volume, which is equal to the determinant of the lattice. This fact can be easily proved for a special class of fundamental domains (see Problem 3.11).

DEFINITION 3.1.5. Let Λ be a lattice, and $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n$ be a basis for Λ. Then the set

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

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

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

Definition 3.1.6. The *Voronoi cell* of a lattice Λ is the set

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

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

Notice that everything we discussed so far also has analogues for lattices of not necessarily full rank. We mention this here briefly without proofs. Let Λ be a lattice in $\mathbb{R}^n$ of rank $1 \leq r \leq n$, and let $\boldsymbol{a}_1, \dots, \boldsymbol{a}_r$ be a basis for it. Write $A = (\boldsymbol{a}_1 \ \dots \ \boldsymbol{a}_r)$ for the corresponding $n \times r$ basis matrix of Λ, then A has rank r since its column vectors are linearly independent. For any $r \times r$ integral matrix U with determinant ± 1, AU is another basis matrix for Λ; moreover, if B is any other basis matrix for Λ, there exists such a U so that $B = AU$. For each basis matrix A of Λ, we define the corresponding *Gram matrix* to be $M = A^\top A$, so it is a square $r \times r$ nonsingular matrix. Notice that if A and B are two basis matrices so that $B = UA$ for some U as above, then

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

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

Definition 3.1.7. Let Λ be a lattice of rank $1 \leq r \leq n$ in $\mathbb{R}^n$, and let A be an $n \times r$ basis matrix for Λ. The *determinant* of Λ is defined to be

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

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

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

3.2. Theorems of Blichfeldt and Minkowski

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

THEOREM 3.2.1 (Blichfeldt, 1914). *Let M be a compact convex set in $\mathbb{R}^n$. Suppose that $\operatorname{Vol}(M) \geq 1$. Then there exist $\boldsymbol{x}, \boldsymbol{y} \in M$ such that $\boldsymbol{0} \neq \boldsymbol{x} - \boldsymbol{y} \in \mathbb{Z}^n$.*

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

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

and let

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

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

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

so that $M'_1, \dots, M'_{r_0} \subseteq P$. On the other hand, $\bigcup_{r=1}^{r_0} M_r = M$, and $M_r \cap M_s = \emptyset$ for all $1 \leq r \neq s \leq r_0$, since $M_r \subseteq P + \boldsymbol{u}_r$, $M_s \subseteq P + \boldsymbol{u}_s$, and $(P + \boldsymbol{u}_r) \cap (P + \boldsymbol{u}_s) = \emptyset$. This means that

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

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

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

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

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

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

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

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

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

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

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

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

As a corollary of Theorem 3.2.1 we can prove the following version of *Minkowski Convex Body Theorem.*

Theorem 3.2.2 (Minkowski). *Let $M \subset \mathbb{R}^n$ be a compact convex $\mathbf{0}$-symmetric set with* $\operatorname{Vol}(M) \geq 2^n$. *Then there exists $\mathbf{0} \neq \boldsymbol{x} \in M \cap \mathbb{Z}^n$.*

Proof. Notice that the set

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

is also convex, $\mathbf{0}$-symmetric, and by Problem 3.12 its volume is

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

Thererfore, by Theorem 3.2.1, there exist $\frac{1}{2}\boldsymbol{x} \neq \frac{1}{2}\boldsymbol{y} \in \frac{1}{2}M$ such that

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

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

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

This completes the proof. □

Remark 3.2.1. This result is sharp: for any $\varepsilon > 0$, the cube

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

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

Problem 3.13 extends Blichfeldt and Minkowski theorems to arbitrary lattices as follows:

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

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

Theorem 3.2.3. *Let $B = (b_{ij})_{1 \leq i,j \leq n} \in \mathrm{GL}_n(\mathbb{R})$, and for each $1 \leq i \leq n$ define a linear form with coefficients $b_{i1}, \dots, b_{in}$ by*

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

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

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

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

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

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

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

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

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

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

where $R = \{\boldsymbol{x} \in \mathbb{R}^n : |x_i| \leq c_i \ \forall \ 1 \leq i \leq n\}$ is the rectangular box with sides of length $2c_1, \dots, 2c_n$ centered at the origin in $\mathbb{R}^n$. Then by Problem 3.12,

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

and so by Theorem 3.2.2 there exists $\mathbf{0} \neq \boldsymbol{x} \in P \cap \mathbb{Z}^n$. □

3.3. Successive minima

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

Corollary 3.3.1. *Let $M \subset \mathbb{R}^n$ be a compact convex $\mathbf{0}$-symmetric and $\Lambda \subset \mathbb{R}^n$ a lattice of full rank. Define the first successive minimum of M with respect to Λ to be*

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

Then

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

Proof. The fact that λ_1 has to be positive readily follows from Λ being a discrete set. Hence we only have to prove the upper bound. By Theorem 3.2.2 for a general lattice Λ (Problem 3.13), if

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

then λM contains a nonzero point of Λ. On the other hand, by Problem 3.12,

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

Hence as long as

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

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

The above corollary thus provides an estimate as to how much should the set M be expanded to contain a nonzero point of the lattice Λ: this is the meaning of λ_1, it is precisely this expansion factor. A natural next question to ask is how much should we expand M to contain 2 linearly independent points of Λ, 3 linearly independent points of Λ, etc. To answer this question is the main objective of this section. We start with a definition.

Definition 3.3.1. Let M be a convex, $\mathbf{0}$-symmetric set $M \subset \mathbb{R}^n$ of non-zero volume and $\Lambda \subseteq \mathbb{R}^n$ a lattice of full rank. For each $1 \leq i \leq n$ define the i-th *succesive minimum* of M with respect to Λ, λ_i, to be the infimum of all positive real numbers λ such that the set λM contains at least i linearly independent points of Λ. In other words,

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

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

Remark 3.3.1. Notice that the n linearly independent vectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_n$ corresponding to successive minima $\lambda_1, \dots, \lambda_n$, respectively, do not necessarily form a basis. It was already known to Minkowski that they do in dimensions $n = 1, \dots, 4$, but when $n = 5$ there is a well known counterexample. Let

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

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

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

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

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

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

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

An immediate observation is that

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

and Corollary 3.3.1 gives an upper bound on λ_1. Can we produce bounds on all the successive minima in terms of $\operatorname{Vol}(M)$ and $\det(\Lambda)$? This question is answered by *Minkowski's Successive Minima Theorem.*

THEOREM 3.3.2. *With notation as above,*

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

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

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

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

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

lie in M, hence M contains the convex hull P of these points, which is a generalized octahedron. Any polyhedron in $\mathbb{R}^n$ can be decomposed as a union of simplices that pairwise intersect only in the boundary. A *standard simplex* in $\mathbb{R}^n$ is the convex hull of n points, so that no 3 of them are co-linear, no 4 of them are co-planar, etc., no k of them lie in a $(k-1)$-dimensional subspace of $\mathbb{R}^n$, and so that their convex hull does not contain any integer lattice points in its interior. The volume of a standard simplex in $\mathbb{R}^n$ is $1/n!$ (Problem 3.14).

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

$$\begin{pmatrix} \frac{u_{11}}{\lambda_1} & \dots & \frac{u_{n1}}{\lambda_n} \\ \vdots & \ddots & \vdots \\ \frac{u_{1n}}{\lambda_1} & \dots & \frac{u_{nn}}{\lambda_n} \end{pmatrix},$$

therefore its volume is

$$\text{(3.3)} \qquad \operatorname{Vol}(P) = \frac{2^n}{n!} \left| \det \begin{pmatrix} \frac{u_{11}}{\lambda_1} & \dots & \frac{u_{n1}}{\lambda_n} \\ \vdots & \ddots & \vdots \\ \frac{u_{1n}}{\lambda_1} & \dots & \frac{u_{nn}}{\lambda_n} \end{pmatrix} \right| = \frac{2^n |\det(U)|}{n!\ \lambda_1 \dots \lambda_N} \geq \frac{2^n}{n!\ \lambda_1 \dots \lambda_n},$$

since $\det(U)$ is an integer. Since $P \subseteq M$, $\operatorname{Vol}(M) \geq \operatorname{Vol}(P)$. Combining this last observation with (3.3) yields the lower bound of the theorem.

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

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

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

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

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

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

for each $1 \leq i \leq n$, i.e. we can rotate each $\operatorname{span}_{\mathbb{R}}\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_i\}$ so that it is contained in E_i. Moreover, volume of AM is the same as volume of M, since $\det(A) = 1$ (i.e. rotation does not change volumes), and

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

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

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

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

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

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

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

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

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

$$\text{(3.4)} \qquad \operatorname{Vol}(C_q^n + M_n) \leq (2q + \gamma)^n,$$

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

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

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

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

and so

$$\begin{aligned}(\boldsymbol{z}-\boldsymbol{x})-(\boldsymbol{z}-\boldsymbol{y}) &= \boldsymbol{y}-\boldsymbol{x} \in \operatorname{int}(M_1)-\operatorname{int}(M_1)\\ &= \{\boldsymbol{z}_1-\boldsymbol{z}_2 : \boldsymbol{z}_1, \boldsymbol{z}_2 \in M_1\} = \operatorname{int}(\lambda_1 M),\end{aligned} \tag{3.5}$$

which would contradict minimality of λ_1. Therefore

$$\operatorname{Vol}(C_q^n + M_1) = (2q+1)^n \operatorname{Vol}(M_1) = (2q+1)^n \left(\frac{\lambda_1}{2}\right)^n \operatorname{Vol}(M). \tag{3.6}$$

To finish the proof, we need the following lemma.

Lemma 3.3.3. *For each* $1 \leq i \leq n-1$,

$$\operatorname{Vol}(C_q^n + M_{i+1}) \geq \left(\frac{\lambda_{i+1}}{\lambda_i}\right)^{n-i} \operatorname{Vol}(C_q^n + M_i). \tag{3.7}$$

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

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

Then

$$(\boldsymbol{x} + \operatorname{int}(M_{i+1})) \cap (\boldsymbol{y} + \operatorname{int}(M_{i+1})) = \emptyset. \tag{3.8}$$

Indeed, suppose (3.8) is not true, i.e. there exists $\boldsymbol{z} \in (\boldsymbol{x} + \operatorname{int}(M_{i+1})) \cap (\boldsymbol{y} + \operatorname{int}(M_{i+1}))$. Then, as in (3.5) above, $\boldsymbol{x} - \boldsymbol{y} \in \operatorname{int}(\lambda_{i+1} M)$. But we also have

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

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

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

On the other hand, at least one of

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

is not equal to 0. Hence $\boldsymbol{x} - \boldsymbol{y}, \boldsymbol{u}_1, \dots, \boldsymbol{u}_i$ are linearly independent, but this means that $\operatorname{int}(\lambda_{i+1} M)$ contains $i+1$ linearly independent points, contradicting minimality of λ_{i+1}. This proves (3.8). Notice that (3.8) implies

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

and

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

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

$$\operatorname{Vol}(C_q^i + M_{i+1}) \geq \left(\frac{\lambda_{i+1}}{\lambda_i}\right)^{n-i} \operatorname{Vol}(C_q^i + M_i). \tag{3.9}$$

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

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

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

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

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

This implies that

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

and so to establish (3.9) it is sufficient to show that

$$\mathrm{Vol}(C_q^i + f_1(M_i)) \geq \mathrm{Vol}(C_q^i + M_i). \tag{3.10}$$

Let

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

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

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

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

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

and hence

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

where Vol_i stands for the i-dimensional volume. This completes the proof of (3.10), and hence of the lemma. □

Now, combining (3.4), (3.6), and (3.7), we obtain:

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

hence

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

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

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

3.4. Inhomogeneous minimum

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

$$\lambda_1 \leq \cdots \leq \lambda_n$$

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

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

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

Lemma 3.4.1.

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

Proof. Let us define a function

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

for every $\boldsymbol{x} \in \mathbb{R}^n$. This function is a norm (Problem 3.15). Then

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

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

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

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

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

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

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

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

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

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

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

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

Using Lemma 3.4.1 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 3.4.2. *If $\lambda_1 \geq 1$, then*

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

Proof. Since

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

Theorem 3.3.2 implies

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

and by Lemma 3.4.1,

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

The result follows by combining these two inequalities. □

A general bound depending also on λ_1 was obtained by Scherk [**Sch50**], once again using Minkowski's successive minima theorem (Theorem 3.3.2) and Jarnik's inequality (Lemma 3.4.1) He observed that if λ_1 is fixed and $\lambda_2, \dots, \lambda_n$ are subject to the conditions

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

then the maximum of the sum

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

is attained when

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

Hence we obtain Scherk's inequality for μ.

Corollary 3.4.3.

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

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

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

hence

$$\mu \geq \left(\frac{\det(\Lambda)}{\operatorname{Vol}(M)} \right)^{1/n}. \tag{3.11}$$

In fact, by Theorem 3.3.2,

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

and combining this with (3.11), we obtain

$$\mu \geq \frac{\lambda_1}{2}. \tag{3.12}$$

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

Lemma 3.4.4.

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

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

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

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

$$F\left(\boldsymbol{x} - \frac{1}{2}\boldsymbol{u}_n\right) \geq \frac{1}{2}\lambda_n. \tag{3.13}$$

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

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

and similarly

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

Therefore, by definition of λ_n,

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

and so $\boldsymbol{u}_n = \boldsymbol{x} + (\boldsymbol{u}_n - \boldsymbol{x}) \in \operatorname{span}_{\mathbb{R}}\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_{n-1}\}$, which is a contradiction. Hence we proved (3.13) for all $\boldsymbol{x} \in \Lambda$. Further, by Problem 3.16,

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

Then lemma follows by combining this observation with (3.13). $\square$

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

3.5. Problems

PROBLEM 3.1. *Let $\boldsymbol{a}_1, \dots, \boldsymbol{a}_r \in \mathbb{R}^n$ be linearly independent points. Prove that $r \leq n$.*

PROBLEM 3.2. *Prove that if Λ is a lattice of rank r in $\mathbb{R}^n$, $1 \leq r \leq n$, then $\operatorname{span}_{\mathbb{R}} \Lambda$ is a subspace of $\mathbb{R}^n$ of dimension r (by $\operatorname{span}_{\mathbb{R}} \Lambda$ we mean the set of all finite real linear combinations of vectors from Λ).*

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

PROBLEM 3.4. *Let Λ be a lattice of rank r in $\mathbb{R}^n$, and let $V = \operatorname{span}_{\mathbb{R}} \Lambda$ be an r-dimensional subspace of $\mathbb{R}^n$, as in Problem 3.3 above. Prove that Λ and V are both additive groups, and Λ is a subgroup of V.*

PROBLEM 3.5. *Let Λ be a lattice and Ω a subset of Λ. Prove that Ω is a sublattice of Λ if and only if it is a subgroup of the abelian group Λ.*

PROBLEM 3.6. *Let Λ be a lattice and Ω a sublattice of Λ of the same rank. Prove that two cosets $\boldsymbol{x} + \Omega$ and $\boldsymbol{y} + \Omega$ of Ω in Λ are equal if and only if $\boldsymbol{x} - \boldsymbol{y} \in \Omega$. Conclude that a coset $\boldsymbol{x} + \Omega$ is equal to Ω if and only if $\boldsymbol{x} \in \Omega$.*

PROBLEM 3.7. *Let Λ be a lattice and $\Omega \subseteq \Lambda$ a sublattice. Suppose that the quotient group Λ/Ω is finite. Prove that rank of Ω is the same as rank of Λ.*

PROBLEM 3.8. *Given a lattice Λ and a real number μ, define*

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

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

PROBLEM 3.9. *Prove that it is possible to select the coefficients v_{ij} in Theorem 3.1.5 so that the matrix $(v_{ij})_{1 \leq i,j \leq n}$ is upper (or lower) triangular with non-negative entries, and the largest entry of each row (or column) is on the diagonal.*

PROBLEM 3.10. *Prove that for every point $\boldsymbol{x} \in \mathbb{R}^n$ there exists uniquely a point $\boldsymbol{y} \in \mathcal{F}$ such that*

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

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

PROBLEM 3.11. *Prove that volume of a fundamental parallelotope is equal to the determinant of the lattice.*

Problem 3.12. *Let S be a compact convex set in $\mathbb{R}^n$, $A \in \mathrm{GL}_n(\mathbb{R})$, and define*

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

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

Problem 3.13. *Prove versions of Theorems 3.2.1 - 3.2.2 where $\mathbb{Z}^n$ is replaced by an arbitrary lattice $\Lambda \subseteq \mathbb{R}^n$ or rank n and the lower bounds on volume of M are multiplied by* $\det(\Lambda)$.
Hint: Let $\Lambda = A\mathbb{Z}^n$ for some $A \in \mathrm{GL}_n(\mathbb{R})$. Then a point $\boldsymbol{x} \in A^{-1}M \cap \mathbb{Z}^n$ if and only if $A\boldsymbol{x} \in M \cap \Lambda$. Now use Problem 3.12 to relate the volume of $A^{-1}M$ to the volume of M.

Problem 3.14. *Prove that a standard simplex in $\mathbb{R}^n$ has volume* $1/n!$.

Problem 3.15. *Let $M \subset \mathbb{R}^n$ be a compact convex $\mathbf{0}$-symmetric set. Define a function $F : \mathbb{R}^n \to \mathbb{R}$, given by*

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

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

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

Problem 3.16. *Let F be a norm like in Problem 3.15. Prove that the inhomogeneous minimum of the corresponding set M with respect to the full-rank lattice $\Lambda \subset \mathbb{R}^n$ satisfies*

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

CHAPTER 4

Lattice Problems, Connections and Applications

4.1. Sphere packing, covering and kissing number problems

Lattices play an important role in discrete optimization from classical problems to the modern day applications, such as theoretical computer science, digital communications, coding theory and cryptography, to name a few. We start with an overview of three old and celebrated problems that are closely related to the techniques in the geometry of numbers that we have so far developed, namely sphere packing, sphere covering and kissing number problems. An excellent comprehensive, although slightly outdated, reference on this subject is the well-known book by Conway and Sloane [**CS99**].

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

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

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

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

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

Definition 4.1.1. Let $\Lambda \subseteq \mathbb{R}^n$ be a lattice of full rank. The *density* of corresponding sphere packing is defined to be

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

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

$$\omega_n = \begin{cases} \frac{\pi^k}{k!} & \text{if } n = 2k \text{ for some } k \in \mathbb{Z} \\ \frac{2^{2k+1} k! \pi^k}{(2k+1)!} & \text{if } n = 2k+1 \text{ for some } k \in \mathbb{Z}. \end{cases} \tag{4.1}$$

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

The first observation we can make is that the packing radius r must depend on the lattice. In fact, it is easy to see that r is precisely one half of the length of the shortest 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.*

Theorem 4.1.1. *In each dimension n there exist lattice packings with density*

$$\Delta \geq \frac{\zeta(n)}{2^{n-1}}, \tag{4.2}$$

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

All known proofs of Theorem 4.1.1 are nonconstructive, so it is not generally known how to construct lattice packings with density as good as (4.2); in particular, in dimensions above 1000 the lattices whose existence is guaranteed by Theorem 4.1.1 are denser than all the presently known ones. We refer to [**GL87**] and [**Cas59**] for many further details on this famous theorem. Here we present a very brief outline of its proof, following [**Cas53**]. The first observation is that this theorem readily follows from the following result.

Theorem 4.1.2. *Let M be a convex bounded $\mathbf{0}$-symmetric set in $\mathbb{R}^n$ with volume $< 2\zeta(n)$. Then there exists a lattice Λ in $\mathbb{R}^n$ of determinant 1 such that M contains no points of Λ except for $\mathbf{0}$.*

Now, to prove Theorem 4.1.2, we can argue as follows. Let χ_M be the characteristic function of the set M, i.e.

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

for every $\boldsymbol{x} \in \mathbb{R}^n$. For parameters T, $\xi_1, \dots, \xi_{n-1}$ to be specified, let us define a lattice $\Lambda = \Lambda_T(\xi_1, \dots, x_{n-1}) :=$

$$\left\{ \left(T(a_1 + \xi_1 b), \dots, T(a_{n-1} + \xi_{n-1} b), T^{-(n-1)} b \right) : a_1, \dots, a_{n-1}, b \in \mathbb{Z} \right\},$$

in other words

$$\Lambda = \begin{pmatrix} T & 0 & \dots & 0 & \xi_1 \\ 0 & T & \dots & 0 & \xi_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & T & \xi_{n-1} \\ 0 & 0 & \dots & 0 & T^{-(n-1)} \end{pmatrix} \mathbb{Z}^n. \tag{4.3}$$

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

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

and so taking T to be sufficiently large we can ensure that none of them are in M, since M is bounded. Thus assume that T is large enough so that the only points of Λ in M have $b \neq 0$. Notice that M contains a nonzero point of Λ if and only if it contains a primitive point of Λ, where we say that $\boldsymbol{x} \in \Lambda$ is primitive if it is not a scalar multiple of another point in Λ. The number of symmetric pairs of primitive points of Λ in M is given by the counting function $\eta_T(\xi_1, \dots, \xi_{n-1}) =$

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

The argument of [**Cas53**] then proceeds to integrate this expression over all $0 \leq \xi_i \leq 1$, $1 \leq i \leq n-1$, obtaining an expression in terms of the volume of M. Taking a limit as $T \to \infty$, it is then concluded that since this volume is $< 2\zeta(n)$, the average of the counting function $\eta_T(\xi_1, \dots, \xi_{n-1})$ is less than 1. Hence there must exist some lattice of the form (4.3) which contains no nonzero points in M.

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

$$\Lambda_h := \begin{pmatrix} 1 & \frac{1}{2} \\ 0 & \frac{\sqrt{3}}{2} \end{pmatrix} \mathbb{Z}^2, \tag{4.4}$$

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

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

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

The density of this packing is ≈ 0.74048. Once again, it has been shown by Gauss in 1831 that this is the densest lattice packing, however until recently it was still not proved that this is the optimal packing. The famous Kepler's conjecture has been settled by Thomas Hales in 1998. Theoretical part of this proof is published only in 2005 [**Hal05**], and the lengthy computational part was published in a series of papers in the journal of Discrete and Computational Geometry (vol. 36, no. 1 (2006)).

Dimensions $n = 8$ and $n = 24$ were settled in 2016, a week apart from each other. Maryna Viazovska [**Via17**], building on previous work of Cohn and Elkies [**CE03**], discovered a “magic” function that implied optimality of the exceptional root lattice E_8 for packing density in $\mathbb{R}^8$. Working jointly with Cohn, Kumar, Miller and Radchenko [**CKM$^+$17**], she then immediately extended her method to dimension 24, where the optimal packing density is given by the famous Leech lattice. Detailed constructions of these remarkable lattices can be found in Conway and Sloane’s book [**CS99**]. This outlines the currently known results for optimal sphere packing configurations in general. On the other hand, best lattice packings are known in dimensions $n \leq 8$, as well as $n = 24$. There are dimensions in which the best known packings are not lattice packings, for instance $n = 11$.

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

Definition 4.1.2. Let $\Lambda \subseteq \mathbb{R}^n$ be a lattice of full rank. The *thickness* Θ of corresponding sphere covering is defined to be

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

where ω_n is the volume of a unit ball in $\mathbb{R}^n$, given by (4.1), and R is the *covering radius*, i.e. radius of each sphere in this lattice covering. It is easy to see that R is precisely the radius of the smallest ball circumscribed around the Voronoi cell $\mathcal{V}$ of Λ, i.e. the *circumradius* of $\mathcal{V}$. Clearly $\Theta \geq 1$.

Notice that the covering radius R is precisely μ, the inhomogeneous minimum of the lattice Λ. Hence combining Lemmas 3.4.1 and 3.4.4 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 (4.4), and is equal to ≈ 1.209199. Best possible lattice coverings are currently known only in dimensions $n \leq 5$, and it is not known in general whether optimal coverings in each dimension are necessarily given by lattices. Once again, there are dimensions in which the best known coverings are not lattice coverings.

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

$$\lambda_1 = \dots = \lambda_n.$$

Lattices with these property are called *well-rounded lattices*, abbreviated *WR*; another term *ESM lattices* (equal successive minima) is also sometimes used. Notice that if Λ is WR, then by Lemma 3.4.4 we have

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

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

Another closely related classical question is known as the *kissing number problem*: given a sphere in $\mathbb{R}^n$ how many other non-overlapping spheres of the same radius can touch it? In other words, if we take the ball centered at the origin in a sphere packing, how many other balls are adjacent to it? Unlike the packing and covering problems, the answer here is easy to obtain in dimension 2: it is 6, and we leave it as an exercise for the reader (Problem 4.2). Although the term "kissing number" is contemporary (with an allusion to billiards, where the balls are said to kiss when they bounce), the 3-dimensional version of this problem was the subject of a famous dispute between Isaac Newton and David Gregory in 1694. It was known at that time how to place 12 unit balls around a central unit ball, however the gaps between the neighboring balls in this arrangement were large enough for Gregory to conjecture that perhaps a 13-th ball can some how be fit in. Newton thought that it was not possible. The problem was finally solved by Schütte and van der Waerden in 1953 [**SvdW53**] (see also [**Lee56**] by J. Leech, 1956), confirming that the kissing number in $\mathbb{R}^3$ is equal to 12. The only other dimensions where the maximal kissing number is known are $n = 4, 8, 24$. More specifically, if we write $\tau(n)$ for the maximal possible kissing number in dimension n, then it is known that

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

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

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

are due to Wyner, Kabatianski and Levenshtein; see [**CS99**] for detailed references and many further details.

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

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

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

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

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

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

A matrix $U \in \mathrm{GL}_n(\mathbb{R})$ is called *orthogonal* if $U^{-1} = U^\top$, and the subset of all such matrices in $\mathrm{GL}_n(\mathbb{R})$ is

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

This is a subgroup of $\mathrm{GL}_n(\mathbb{R})$ (Problem 4.5). Discrete optimization problems on the space of lattices in a given dimension, as those discussed above, are usually considered up to the equivalence relation of *similarity*: two lattices L and M of full rank in $\mathbb{R}^n$ are called *similar*, denoted $L \sim M$, if there exists $\alpha \in \mathbb{R}$ and an orthogonal matrix $U \in \mathcal{O}_n(\mathbb{R})$ such that $L = \alpha UM$. This is an equivalence relation on the space of all full-rank lattices in $\mathbb{R}^n$ (Problem 4.3), and we refer to the equivalence classes under this relation as *similarity classes*. If lattices L and M are similar, then they have the same packing density, covering thickness, and kissing number (Problem 4.4). We use the perspective of similarity classes in the next section when considering lattice packing density in the plane.

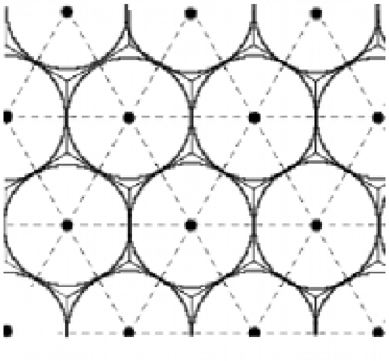


FIGURE 1. Hexagonal lattice with Voronoi cell translates and associated circle packing

4.2. Lattice packings in dimension 2

Our goal here is to prove that the best lattice packing in $\mathbb{R}^2$ is achieved by the hexagonal lattice Λ_h as defined in (4.4) above (see Figure 1). Specifically, we will prove the following theorem.

THEOREM 4.2.1. *Let L be a lattice of rank 2 in $\mathbb{R}^2$. Then*

$$\Delta(L) \leq \Delta(\Lambda_h) = \frac{\pi}{2\sqrt{3}} = 0.906899\ldots,$$

and the equality holds if any only if $L \sim \Lambda_h$.

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

LEMMA 4.2.2. *Let Λ be a lattice in $\mathbb{R}^2$ with successive minima $\lambda_1 \leq \lambda_2$ and let $\boldsymbol{x}_1, \boldsymbol{x}_2$ be the vectors in Λ corresponding to λ_1, λ_2, respectively. Then $\boldsymbol{x}_1, \boldsymbol{x}_2$ form a basis for Λ.*

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

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

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

$$(\boldsymbol{x}_1 \; \boldsymbol{x}_2) = (\boldsymbol{y}_1 \; \boldsymbol{y}_2) \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix}. \tag{4.5}$$

Let θ_x be the angle between $\boldsymbol{x}_1, \boldsymbol{x}_2$, and θ_y be the angle between $\boldsymbol{y}_1, \boldsymbol{y}_2$, then $\pi/3 \leq \theta_x \leq \pi/2$ by Problem 4.7. Moreover, $\pi/3 \leq \theta_y \leq \pi/2$: indeed, suppose

$\theta_y < \pi/3$, then by Problem 4.6,

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

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

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

then $\mathcal{D}$ is a positive integer, and taking determinants of both sides of (4.5), we obtain

$$\|\boldsymbol{x}_1\| \|\boldsymbol{x}_2\| \sin\theta_x = \mathcal{D} \|\boldsymbol{y}_1\| \|\boldsymbol{y}_2\| \sin\theta_y. \tag{4.6}$$

Notice that by definition of successive minima, $\|\boldsymbol{x}_1\| \|\boldsymbol{x}_2\| \leq \|\boldsymbol{y}_1\| \|\boldsymbol{y}_2\|$, and hence (4.6) implies that

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

meaning that $\mathcal{D} = 1$. Combining this observation with (4.5), we see that

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

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

As we know from Remark 3.3.1 in Section 3.3, the statement of Lemma 4.2.2 does not generally hold for $d \geq 5$. We will call a basis for a lattice as in Lemma 4.2.2 a *minimal basis*. The goal of the next three lemmas is to show that the lattice packing density function Δ attains its maximum in $\mathbb{R}^2$ on the set of well-rounded lattices.

Lemma 4.2.3. *Let Λ and Ω be lattices of full rank in $\mathbb{R}^2$ with successive minima $\lambda_1(\Lambda), \lambda_2(\Lambda)$ and $\lambda_1(\Omega), \lambda_2(\Omega)$ respectively. Let $\boldsymbol{x}_1, \boldsymbol{x}_2$ and $\boldsymbol{y}_1, \boldsymbol{y}_2$ be vectors in Λ and Ω, respectively, corresponding to successive minima. Suppose that $\boldsymbol{x}_1 = \boldsymbol{y}_1$, and angles between the vectors $\boldsymbol{x}_1, \boldsymbol{x}_2$ and $\boldsymbol{y}_1, \boldsymbol{y}_2$ are equal, call this common value θ. Suppose also that*

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

Then

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

Proof. By Lemma 4.2.2, $\boldsymbol{x}_1, \boldsymbol{x}_2$ and $\boldsymbol{y}_1, \boldsymbol{y}_2$ are minimal bases for Λ and Ω, respectively. Notice that

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

Then

$$\begin{aligned} \Delta(\Lambda) &= \frac{\pi\lambda_1(\Lambda)^2}{4\det(\Lambda)} = \frac{\lambda_1(\Lambda)^2\pi}{4\|\boldsymbol{x}_1\|\|\boldsymbol{x}_2\|\sin\theta} = \frac{\pi}{4\sin\theta} \\ &\geq \frac{\lambda_1(\Omega)^2\pi}{4\|\boldsymbol{y}_1\|\|\boldsymbol{y}_2\|\sin\theta} = \frac{\lambda_1(\Omega)^2\pi}{4\det(\Omega)} = \Delta(\Omega). \end{aligned} \tag{4.7}$$

□

The following lemma is a converse to Problem 4.7.

LEMMA 4.2.4. *Let $\Lambda \subset \mathbb{R}^2$ be a lattice of full rank, and let $\boldsymbol{x}_1, \boldsymbol{x}_2$ be a basis for Λ such that*

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

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

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

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

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

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

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

LEMMA 4.2.5. *Let Λ be a lattice in $\mathbb{R}^2$ with successive minima λ_1, λ_2 and corresponding basis vectors $\boldsymbol{x}_1, \boldsymbol{x}_2$, respectively. Then the lattice*

$$\Lambda_{\mathrm{WR}} = \begin{pmatrix} \boldsymbol{x}_1 & \frac{\lambda_1}{\lambda_2}\boldsymbol{x}_2 \end{pmatrix} \mathbb{Z}^2$$

is WR with successive minima equal to λ_1.

PROOF. By Problem 4.7, the angle θ between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is in the interval $[\pi/3, \pi/2]$, and clearly this is the same as the angle between the vectors $\boldsymbol{x}_1$ and $\frac{\lambda_1}{\lambda_2}\boldsymbol{x}_2$. Then by Lemma 4.2.4, Λ_{WR} is WR with successive minima equal to λ_1. □

Now combining Lemma 4.2.3 with Lemma 4.2.5 implies that

$$\Delta(\Lambda_{\mathrm{WR}}) \geq \Delta(\Lambda) \tag{4.8}$$

for any lattice $\Lambda \subset \mathbb{R}^2$, and (4.7) readily implies that the equality in (4.8) occurs if and only if $\Lambda = \Lambda_{\mathrm{WR}}$, which happens if and only if Λ is well-rounded. Therefore the maximum packing density among lattices in $\mathbb{R}^2$ must occur at a WR lattice, and so for the rest of this section we talk about WR lattices only. Next observation is that for any WR lattice Λ in $\mathbb{R}^2$, (4.7) implies:

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

meaning that $\sin\theta$ is an invariant of Λ, and does not depend on the specific choice of the minimal basis. Since by our conventional choice of the minimal basis and Problem 4.7, this angle θ is in the interval $[\pi/3, \pi/2]$, it is also an invariant of the lattice, and we call it the *angle of* Λ, denoted by $\theta(\Lambda)$.

LEMMA 4.2.6. *Let Λ be a WR lattice in $\mathbb{R}^2$. A lattice $\Omega \subset \mathbb{R}^2$ is similar to Λ if and only if Ω is also WR and $\theta(\Lambda) = \theta(\Omega)$.*

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

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

Then $\|\boldsymbol{y}_1\| = \|\boldsymbol{y}_2\|$, and the angle between $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ is $\theta(\Lambda) \in [\pi/3, \pi/2]$. By Lemma 4.2.4 it follows that $\boldsymbol{y}_1, \boldsymbol{y}_2$ is a minimal basis for Ω, and so Ω is WR and $\theta(\Omega) = \theta(\Lambda)$.

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

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

Then $\boldsymbol{x}_1, \boldsymbol{x}_2$ and $\boldsymbol{z}_1, \boldsymbol{z}_2$ are pairs of points on the circle of radius $\lambda(\Lambda)$ centered at the origin in $\mathbb{R}^2$ with equal angles between them. Therefore, there exists a 2×2 real orthogonal matrix U such that

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

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

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

Proof of Theorem 4.2.1. The density inequality (4.8) says that the largest lattice packing density in $\mathbb{R}^2$ is achieved by some WR lattice Λ, and (4.7) implies that

$$\Delta(\Lambda) = \frac{\pi}{4\sin\theta(\Lambda)}, \tag{4.9}$$

meaning that a smaller $\sin\theta(\Lambda)$ corresponds to a larger $\Delta(\Lambda)$. Problem 4.7 implies that $\theta(\Lambda) \geq \pi/3$, meaning that $\sin\theta(\Lambda) \geq \sqrt{3}/2$. Notice that if Λ is the hexagonal lattice

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

then $\sin\theta(\Lambda) = \sqrt{3}/2$, meaning that the angle between the basis vectors $(1, 0)$ and $(1/2, \sqrt{3}/2)$ is $\theta = \pi/3$, and so by Lemma 4.2.4 this is a minimal basis and $\theta(\Lambda) = \pi/3$. Hence the largest lattice packing density in $\mathbb{R}^2$ is achieved by the hexagonal lattice. This value now follows from (4.9).

Now suppose that for some lattice Λ, $\Delta(\Lambda) = \Delta(\Lambda_h)$, then by (4.8) and a short argument after it Λ must be WR, and so

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

Then $\theta(\Lambda) = \pi/3$, and so Λ is similar to Λ_h by Lemma 4.2.6. This completes the proof. □

While we have only settled the question of best lattice packing in dimension two, we saw that well-roundedness is an essential property for a lattice to be a good contender for optimal packing density. There are, however, infinitely many WR lattices in the plane, even up to similarity, and only one of them worked well. One can then ask what properties must a lattice have to maximize packing density?

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

$$\|\boldsymbol{v}\|^2 = \sum_{i=1}^{m} c_i (\boldsymbol{v}^\top \boldsymbol{x}_i)^2 \tag{4.10}$$

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

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

spans the real vector space of $n \times n$ symmetric matrices. These properties are preserved on similarity classes (Problem 4.8), and up to similarity there are only finitely many perfect or eutactic lattices in every dimension. For instance, up to similarity, the hexagonal lattice is the only one in the plane that is both, perfect and eutactic (Problem 4.9).

Suppose that $\Lambda = A\mathbb{Z}^n$ is a lattice with basis matrix A, then, as we know, B is another basis matrix for Λ if and only if $B = AU$ for some $U \in \mathrm{GL}_n(\mathbb{Z})$. In this way, the space of full-rank lattices in $\mathbb{R}^n$ can be identified with the set of orbits of $\mathrm{GL}_n(\mathbb{R})$ under the action by $\mathrm{GL}_n(\mathbb{Z})$ by right multiplication. The packing density Δ is a continuous function on this space, and hence we can talk about its local extremum points. A famous theorem of Georgy Voronoi (1908) states that a lattice is a local maximum of the packing density function in its dimension if and only if it is perfect and eutactic. Hence, combining Problem 4.9 with Voronoi's theorem gives another proof of unique optimality of the hexagonal lattice for lattice packing in the plane. Further, Voronoi's theorem suggests a way of looking for the maximizer of the lattice packing density in every dimension: identify the finite set of perfect and eutactic lattices, compute their packing density and choose the largest. Unfortunately, this approach is not very practical, since already in dimension 9 the number of perfect lattices is over 9 million.

4.3. Algorithmic problems on lattices

There is a class of algorithmic problems studied in computational number theory, discrete geometry and theoretical computer science, which are commonly referred to as the *lattice problems*. One of their distinguishing features is that they are provably known to be very hard to solve in the sense of computational complexity of algorithms involved. As usual, we write $\Lambda \subset \mathbb{R}^n$ for a lattice of full rank and

$$0 < \lambda_1 \leq \dots \leq \lambda_n$$

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

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

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

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

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

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

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

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

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

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

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

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

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

On the other hand, γ-approximate versions of these problems for γ exponential in n are known to be polynomial. The most famous such approximation algorithm is LLL, which was discovered by A. Lenstra, H. Lenstra and L. Lovasz in 1982 [**LLL82**]. LLL is a polynomial time reduction algorithm that, given a lattice Λ, produces a basis $\boldsymbol{b}_1, \dots, \boldsymbol{b}_n$ for Λ such that

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

and

$$\prod_{i=1}^{n} \|\boldsymbol{b}_i\| \leq 2^{\frac{n(n-1)}{4}} \det(\Lambda). \tag{4.11}$$

We can compare this to the upper bound given by Minkowski's Successive Minima Theorem (Theorem 3.3.2):

$$\prod_{i=1}^{n} \lambda_i \leq \frac{2^n}{\omega_n} \det(\Lambda). \tag{4.12}$$

For instance, when $n = 2k$ the bound (4.11) gives

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

while (4.12) gives

$$\prod_{i=1}^{n} \lambda_i \leq \frac{4^k k!}{\pi^k} \det(\Lambda).$$

Let us briefly describe the main idea behind LLL. The first observation is that an orthogonal basis, if one exists in a lattice, is always the shortest one. Indeed, suppose $\boldsymbol{u}_1, \dots, \boldsymbol{u}_n$ is such a basis, then for any $a_1, \dots, a_n \in \mathbb{Z}$,

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

which implies that the shortest basis vectors can only be obtained by taking one of the coefficients $a_i = \pm 1$ and the rest 0. Of course, most lattices do not have orthogonal bases, in which case finding a short basis is much harder. Still, the basic principle of constructing a short basis is based on looking for vectors that would be "close to orthogonal".

We observed in Section 4.2 (in particular, see Problems 4.6, 4.7, Lemma 4.2.4) that the angle between a pair of shortest vectors must be between $[\pi/3, 2\pi/3]$, i.e. these vectors are "near-orthogonal": in fact, these vectors have to be as close to orthogonal as possible within the lattice. This is the underlying idea behind the classical *Lagrange-Gauss Algorithm* for finding a shortest basis for a lattice in $\mathbb{R}^2$. Specifically, an ordered basis $\boldsymbol{b}_1, \boldsymbol{b}_2$ for a planar lattice Λ consists of vectors corresponding to successive minima λ_1, λ_2 of Λ, respectively, if and only if

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

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

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

where $\lfloor \mu \rceil$ stands for the nearest integer to μ, produces a shorter second basis vector. We leave the proof of this as an exercise (Problem 4.10). Hence we can formulate the Gauss-Lagrange Algorithm:

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

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

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

This algorithm terminates in a finite number of steps (Problem 4.11).

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

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

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

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

The nearest integer to μ is 1, so we set

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

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

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

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

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

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

(1) For all $1 \leq j < i \leq n$, $|\mu_{ij}| \leq 1/2$

(2) For some parameter $\delta \in [1/4, 1)$, for all $1 \leq k \leq n$,

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

Traditionally, δ is taken to be 3/4. While we will not go into further details about the LLL, some good more detailed references on this subject include the original paper [**LLL82**], as well as more recent books [**Coh00**], [**Bor02**], and [**HPS08**].

4.4. CVP is NP-hard

In this section we discuss the complexity of the decision version of the CVP for sublattices of the integer lattice $\mathbb{Z}^n$. Specifically, here is the problem we are considering:

Given an $n \times m$ integer basis matrix B, $m \leq n$, a target vector $\boldsymbol{t} \in \mathbb{Z}^n$ and a (usually rational) number $r > 0$, does there exist a vector $\boldsymbol{x} \in B\mathbb{Z}^m$ such that $\|\boldsymbol{x} - \boldsymbol{t}\| \leq r$?

We will now explicitly show that this problem is NP-hard.

Theorem 4.4.1. *The decision version of CVP is NP-complete.*

Proof. First notice that, given a vector $\boldsymbol{x} \in B\mathbb{Z}^m$, checking whether $\|\boldsymbol{x}-\boldsymbol{t}\| \leq r$ is a polynomial problem: it comes down to computing the difference vector, evaluating its Euclidean norm, and comparing it to r. Hence our problem is NP.

To show that it is NP-hard, we will construct a polynomial-time reduction algorithm from SSP (the subset sum problem) to decision CVP. Since we know that SSP is NP-hard (Theorem 2.1.2), the result will follow. Let

$$\boldsymbol{a} = (a_1, \dots, a_n), s \tag{4.13}$$

be an instance of SSP, i.e. $\boldsymbol{a}$ is the n-tuple of weights and s is the target sum. Define the $(n+1) \times n$ basis matrix B for a lattice $B\mathbb{Z}^n \subset \mathbb{Z}^{n+1}$ by

$$B = \begin{pmatrix} \boldsymbol{a} \\ 2I_n \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & \dots & a_n \\ 2 & 0 & \dots & 0 \\ 0 & 2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 2 \end{pmatrix},$$

and let $\boldsymbol{t} = (s, 1, \dots, 1)^\top \in \mathbb{Z}^{n+1}$ be the target vector. The we can consider the instance of decision CVP with B, $\boldsymbol{t}$ and $r = \sqrt{n}$, i.e.

$$\text{does there exist a vector } \boldsymbol{x} \in B\mathbb{Z}^n \text{ so that } \|\boldsymbol{x} - \boldsymbol{t}\| \leq \sqrt{n}? \tag{4.14}$$

Assume (4.13) is a YES instance of SSP, i.e. $\sum_{i=1}^n a_i x_i = s$ for some choice of coefficients $x_1, \dots, x_n \in \{0, 1\}$, then

$$B\boldsymbol{x} - \boldsymbol{t} = \begin{pmatrix} \sum_{i=1}^n a_i x_i - s \\ 2x_1 - 1 \\ \vdots \\ 2x_n - 1 \end{pmatrix},$$

and so $\|B\boldsymbol{x} - \boldsymbol{t}\|^2 = \sum_{i=1}^n |2x_i - 1|^2 = n$, since $2x_i - 1 = \pm 1$ for every i, and thus the answer to (4.14) is also YES. Conversely, a YES instance of (4.14) with B and $\boldsymbol{t}$ as above gives a vector $\boldsymbol{y} = B\boldsymbol{x} \in B\mathbb{Z}^n$ such that

$$\|B\boldsymbol{x} - \boldsymbol{t}\|^2 = \left|\sum_{i=1}^n a_i x_i - s\right|^2 + \sum_{i=1}^n |2x_i - 1|^2 \leq n,$$

which can only be true if $\sum_{i=1}^n a_i x_i - s = 0$, since again $2x_i - 1 = \pm 1$ for every i. Thus we obtain a YES instance of (4.13).

Hence we have a reduction from SSP to decision CVP, which is polynomial-time by construction. □

4.5. Geometry of the Frobenius problem

In this section we will apply the newly-acquired knowledge of lattices and their geometric invariants to the Frobenius problem. A geometric approach to the classical Frobenius problem has been pioneered in the influential paper of R. Kannan [**Kan92**], leading to a polynomial-time algorithm to find the Frobenius number for each fixed n. Bounds on the classical Frobenius number stemming from further geometry of numbers applications have been obtained in [**FR07**] and [**AG07**]. These ideas have also been extended to the more general s-Frobenius problem in [**FS11**] and [**AFH12**]. A higher-dimensional analogue of the Frobenius problem has also been considered in the recent years by several authors, notably in [**AH10**], [**AHL13**], and [**ALL16**]. A generalization of this problem to certain number fields has been studied in [**FS20**].

Let us briefly describe Kannan's approach to the Frobenius problem. Let

$$L_{\boldsymbol{a}} = \left\{ \boldsymbol{x} \in \mathbb{Z}^{n-1} : \sum_{i=1}^{n-1} a_i x_i \equiv 0 \pmod{a_n} \right\},$$

then $L_{\boldsymbol{a}}$ is a sublattice of $\mathbb{Z}^{n-1}$ of full rank. Define also a simplex

$$S_{\boldsymbol{a}} = \left\{ \boldsymbol{x} \in \mathbb{R}^{n-1}_{\geq 0} : \sum_{i=1}^{n-1} a_i x_i \leq 1 \right\}.$$

With this notation, Kannan proved the following remarkable identity.

Theorem 4.5.1.

$$g_0(\boldsymbol{a}) = \mu(S_{\boldsymbol{a}}, L_{\boldsymbol{a}}) - \sum_{i=1}^{n} a_i. \tag{4.15}$$

where $\mu(S_{\boldsymbol{a}}, L_{\boldsymbol{a}})$ is the inhomogeneous minimum (also known as the covering radius) of $S_{\boldsymbol{a}}$ with respect to L, namely

$$\mu(S_{\boldsymbol{a}}, L_{\boldsymbol{a}}) = \inf\left\{ t \in \mathbb{R}_{>0} : tS_{\boldsymbol{a}} + L_{\boldsymbol{a}} = \mathbb{R}^{n-1} \right\}. \tag{4.16}$$

Proof. Kannan's argument consists of an upper and a lower bound on the inhomogeneous minimum. First we show that

$$\mu(S_{\boldsymbol{a}}, L_{\boldsymbol{a}}) \leq g_0(\boldsymbol{a}) + \sum_{i=1}^{n} a_i. \tag{4.17}$$

Assume that $\boldsymbol{y} \in \mathbb{Z}^{n-1}$ is such that $\sum_{i=1}^{n-1} a_i y_i \equiv m \,(\mathrm{mod}\, a_n)$. Let t_m be the smallest positive integer congruent to m modulo a_n that is representable as a nonnegative integer linear combination of $a_1, \dots, a_{n-1}$. Then there exist coefficients $x_1, \dots, x_n \in \mathbb{Z}_{\geq 0}$ such that

$$t_m = a_1 x_1 + \cdots + a_{n-1} x_{n-1} + a_n x_n = m + a_n x_n.$$

Let $\boldsymbol{x}' = (x_1, \dots, x_{n-1})$ for this choice of the coefficients, and observe that $\boldsymbol{y} - \boldsymbol{x}' \in L_{\boldsymbol{a}}$. Further, $\boldsymbol{x}' \in mS_{\boldsymbol{a}} \subseteq t_m S_{\boldsymbol{a}}$ and so $\boldsymbol{y} = (\boldsymbol{y} - \boldsymbol{x}') + \boldsymbol{x}' \in L_{\boldsymbol{a}} + t_m S_{\boldsymbol{a}}$. Since the choice of $\boldsymbol{y} \in \mathbb{Z}^{n-1}$ was arbitrary, we conclude that $\mathbb{Z}^{n-1} \subseteq L_{\boldsymbol{a}} + t_m S_{\boldsymbol{a}}$. Additionally, $t_m \leq g_0(\boldsymbol{a}) + a_n$, and thus

$$\mathbb{Z}^{n-1} \subseteq L_{\boldsymbol{a}} + (g_0(\boldsymbol{a}) + a_n) S_{\boldsymbol{a}}.$$

Also notice that for any point $\boldsymbol{z} \in \mathbb{R}^{n-1}$, the integer part $[\boldsymbol{z}] = ([z_1], \dots, [z_{n-1}]) \in \mathbb{Z}^{n-1}$ and the point $\boldsymbol{z}' = \boldsymbol{z} - [\boldsymbol{z}]$ has each coordinate ≤ 1, and so satisfies the inequality

$$\sum_{i=1}^{n-1} a_i z_i' \leq \sum_{i=1}^{n-1} a_i.$$

This means that

$$\boldsymbol{z} = [\boldsymbol{z}] + \boldsymbol{z}' \in \mathbb{Z}^{n-1} + \left(\sum_{i=1}^{n-1} a_i \right) S_{\boldsymbol{a}}.$$

Thus

$$\mathbb{R}^{n-1} \subseteq \mathbb{Z}^{n-1} + \left(\sum_{i=1}^{n-1} a_i \right) S_{\boldsymbol{a}} \subseteq L_{\boldsymbol{a}} + \left(g_0(\boldsymbol{a}) + \sum_{i=1}^{n} a_i \right) S_{\boldsymbol{a}},$$

which implies (4.17).

Next we establish that

$$\mu(S_{\boldsymbol{a}}, L_{\boldsymbol{a}}) \geq g_0(\boldsymbol{a}) + \sum_{i=1}^{n} a_i. \tag{4.18}$$

For this, we first need an auxiliary lemma.

Lemma 4.5.2. $g_0(\boldsymbol{a}) = \max_{1 \leq m \leq a_n - 1} t_m - a_n$.

Proof. If a positive integer T is congruent to 0 modulo a_n, then T is just a multiple of a_n. Otherwise, $T \equiv m \pmod{a_n}$ for some $1 \leq m \leq a_n - 1$ and thus it is representable as a nonnegative linear combination of $a_1, \dots, a_n$ if and only if it is $\geq t_m$. □

Back to the proof of (4.18), let us consider the set $(g_0(\boldsymbol{a}) + a_n) S_{\boldsymbol{a}} + L_{\boldsymbol{a}}$. We will show that $g_0(\boldsymbol{a}) + a_n$ is the smallest positive real value of t so that $t S_{\boldsymbol{a}} + L_{\boldsymbol{a}}$ contains $\mathbb{Z}^{n-1}$. Suppose not, then there exists some $t' < g_0(\boldsymbol{a}) + a_n$ so that $\mathbb{Z}^{n-1} \subset t' S_{\boldsymbol{a}} + L_{\boldsymbol{a}}$. Pick any $1 \leq m \leq a_n - 1$ and take $\boldsymbol{y} \in \mathbb{Z}^{n-1}$ be such that

$$\sum_{i=1}^{n-1} a_i y_i \equiv m \pmod{a_n}.$$

Since $\boldsymbol{y} \in t' S_{ba} + \boldsymbol{x}$ for some $\boldsymbol{x} \in L_{\boldsymbol{a}}$, we must have $\boldsymbol{y} - \boldsymbol{x} \in t' S_{\boldsymbol{a}}$. However,

$$\sum_{i=1}^{n-1} a_i (y_i - x_i) \equiv m \pmod{a_n} \text{ and } y_i - z_i \geq 0 \ \forall \ 1 \leq i \leq n$$

implies that $t_m \leq t'$. This is true for any choice of m, Lemma 4.5.2 implies

$$g_0(\boldsymbol{a}) = \max_{1 \leq m \leq a_n - 1} t_m - a_n \leq t' - a_n < g_0(\boldsymbol{a})$$

by our assumption on t'. This is a contradiction, hence

$$g_0(\boldsymbol{a}) + a_n = \min \left\{ t > 0 : \mathbb{Z}^{n-1} \subset t S_{\boldsymbol{a}} + L_{\boldsymbol{a}} \right\}.$$

Therefore there must exist $\boldsymbol{y} \in \mathbb{Z}^{n-1}$ such that for any $\boldsymbol{x} \in L_{\boldsymbol{a}}$ with $y_i - x_i \geq 0$ for all i, we have

$$\sum_{i=1}^{n-1} a_i (y_i - x_i) \geq g_0(\boldsymbol{a}) + a_n. \tag{4.19}$$

Let $\varepsilon \in (0,1)$ and define the point $\boldsymbol{p} \in \mathbb{R}^{n-1}$ by $p_i = y_i + (1-\varepsilon)$ for every i. Suppose $\boldsymbol{x} \in L_{\boldsymbol{a}}$ is such that $x_i \geq p_i$ for every i. Since all x_i's are integers, we must have $x_i \leq y_i$ for every i, and so

$$\sum_{i=1}^{n-1} a_i(p_i - x_i) = (1-\varepsilon)\sum_{i=1}^{n-1} a_i + \sum_{i=1}^{n-1} a_i(y_i - x_i) \geq (1-\varepsilon)\sum_{i=1}^{n-1} a_i + (g_0(\boldsymbol{a}) + a_n),$$

by (4.19). Now, $\mu(S_{\boldsymbol{a}}, L_{ba})$ is $\geq$ than the left hand side of this inequality, which holds for any $\varepsilon \in (0,1)$. Thus we must have

$$\mu(S_{\boldsymbol{a}}, L_{\boldsymbol{a}}) \geq \sum_{i=1}^{n-1} a_i + (g_0(\boldsymbol{a}) + a_n) = g_0(\boldsymbol{a}) + \sum_{i=1}^{n} a_i.$$

This completes the proof. □

On the other hand, Kannan showed that in every fixed dimension n there is a polynomial-time algorithm to find the covering radius, given $S_{\boldsymbol{a}}$ and $L_{\boldsymbol{a}}$ (which is to say, given $\boldsymbol{a}$). This result, along with his identity (4.15) implies a polynomial-time algorithm for the Frobenius number in fixed dimension. Kannan's Theorem 4.5.1 has been extended to the s-Frobenius numbers in [**AFH12**]. For integer $s \geq 1$, define

$$\mu_s(S_{\boldsymbol{a}}, L_{\boldsymbol{a}}) = \min\{t > 0 : \forall\ \boldsymbol{x} \in \mathbb{R}^n\ \exists\ \boldsymbol{b}_1, \dots, \boldsymbol{b}_s \in L_{\boldsymbol{a}} \text{ s.t. } \boldsymbol{x} \in \boldsymbol{b}_i + tS_{\boldsymbol{a}}\} \tag{4.20}$$

be the smallest positive number t such that any $\boldsymbol{x} \in \mathbb{R}^n$ is covered by at least s lattice translates of $tS_{\boldsymbol{a}}$: this $\mu_s(S_{\boldsymbol{a}}, L_{\boldsymbol{a}})$ is called the *s-covering radius* of $S_{\boldsymbol{a}}$ with respect to $L_{\boldsymbol{a}}$. If $s = 1$, this is precisely the classical covering radius as in (4.16). With this notation, the following theorem is established in [**AFH12**].

THEOREM 4.5.3.

$$g_s(\boldsymbol{a}) = \mu_{s+1}(S_{\boldsymbol{a}}, L_{\boldsymbol{a}}) - \sum_{i=1}^{n} a_i.$$

Such geometric ideas have also been used by different authors to give expected values of Frobenius numbers with respect to the uniform probability distribution on ensembles of vectors in $\mathbb{Z}^n$ defined with respect to different norms; see [**Arn99**], [**Arn06**], [**AH09**], [**AHH11**], [**BS07**], [**Li15**], [**Mar10**], [**Str12**], [**SSU09**], [**Ust10**], and [**AFH12**] for results on average behavior of Frobenius numbers.

4.6. Lattice point counting

All of the lattice point counting results in the previous chapters were specifically for integer lattice points in polytopes, which is a rather special class of convex bodies in $\mathbb{R}^n$ and only one lattice. What can be said for more general convex bodies and lattices? Let $M \subseteq \mathbb{R}^n$ be closed, bounded, and Jordan measurable with $\operatorname{Vol}(M) > 0$, and let $\Lambda \subseteq \mathbb{R}^n$ be a lattice of full rank. Suppose we homogeneously expand M by a positive real parameter t, i.e. for each positive real value of t we will consider the set tM. How many points of Λ are there in tM as t grows? To partially answer this question, we will be interested in the *asymptotic behavior* of the function

$$G_{M,\Lambda}(t) = |tM \cap \Lambda|$$

as $t \to \infty$. In general, this is a very difficult question. We will need to make some additional assumptions on M in order to study $G_{M,\Lambda}(t)$.

Definition 4.6.1. Let S be a subset of some Eucildean space. A map

$$\varphi : S \to \mathbb{R}^n$$

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

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

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

Let C_n be the cube as in (2.11). We say that a set $S \subseteq \mathbb{R}^n$ is *Lipschitz parametrizable* if there exists a finite number of Lipschitz maps

$$\varphi_j : C_n \to S,$$

such that $S = \bigcup_j \varphi_j(C_n)$.

Definition 4.6.2. Let $f(t)$ and $g(t)$ be two functions defined on $\mathbb{R}$. We will say that

$$f(t) = O(g(t)) \text{ as } t \to \infty$$

if there exists a positive real number $\mathcal{B}$ and a real number t_0 such that for all $t \geq t_0$,

$$|f(t)| \leq \mathcal{B}|g(t)|.$$

We usually use the O-notation to emphasize the fact that $f(t)$ behaves similar to $g(t)$ when t is large. This is quite useful if $g(t)$ is a simpler function than $f(t)$; in this case, such a statement helps us to understand the *asymptotic behavior* of $f(t)$, namely its behavior as $t \to \infty$.

Let ∂M be the boundary of M, and assume that ∂M is $(n-1)$-Lipschitz parametrizable. Notice that for $t \in \mathbb{R}_{>0}$, $\partial(tM) = t\partial M$. The following result is Theorem 2 on p. 128 of [**Lan94**].

Theorem 4.6.1. *Let $t \in \mathbb{R}_{>0}$, then*

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

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

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

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

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

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

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

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

or

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

Let

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

Then clearly

$$m(t) \leq G_{M,\Lambda}(t) \leq m(t) + b(t).$$

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

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

hence

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

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

$$\varphi : C_{n-1} \to \partial M$$

be one of the Lipschitz paramterizing maps for a piece of the boundary of M, and let $\mathcal{C}$ be the maximum of all Lipschitz constants corresponding to these maps. Then $t\varphi$ parametrizes a corresponding piece of $\partial(tM) = t\partial M$. Cut up each side of C_{n-1} into segments of length $1/[2t]$, then we can represent C_{n-1} as a union of $[t]^{n-1}$ small cubes with sidelength $1/[2t]$ each, call them $C^1, \dots, C^{[t]^{n-1}}$. For each such C_i, we have

$$\|\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y})\|_2 \leq \mathcal{C}\|\boldsymbol{x} - \boldsymbol{y}\|_2 \leq \frac{\mathcal{C}\sqrt{n-1}}{[2t]},$$

for each $\boldsymbol{x}, \boldsymbol{y} \in C_i$, i.e. the image of each such C_i under φ has diameter at most $\frac{\mathcal{C}\sqrt{n-1}}{[2t]}$. Hence image of each such C_i under the map $t\varphi$ has diameter at most

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

Clearly therefore the number of $\boldsymbol{x} \in \Lambda$ such that the corresponding translate $\mathcal{F}_{\boldsymbol{x}}$ has nonempty intersection with $t\varphi(C_i)$, for each $1 \leq i \leq [t]^{n-1}$, is bounded by some constant $\mathcal{C}'$ that depends only on Λ, $\mathcal{C}$, and n. Hence

$$b(t) \leq \mathcal{C}'[t]^{n-1}.$$

This completes the proof. □

Theorem 4.6.1 provides an asymptotic formula for $G_{M,\Lambda}(t)$, demonstrating an important general principle, namely that as $t \to \infty$, $G_{M,\Lambda}(t)$ grows like $\frac{\mathrm{Vol}(M)}{\det(\Lambda)} t^n$, which is what one would expect. However, it does not give any explicit information about the constant in the error term $O(t^{n-1})$. Can this constant be somehow bounded, i.e. what can be said about the quantity

$$\left| G_{M,\Lambda}(t) - \frac{\mathrm{Vol}(M)}{\det(\Lambda)} t^n \right| \ ?$$

A large amount of work has been done in this direction (see for instance pp. 140 - 147 of [**GL87**] for an overview of results and bibliography). This subject essentially originated in a paper of Davenport [**Dav51**], who used a principle of Lipschitz [**Lip65**]; also see [**Thu93**] for a nice overview of Davenport's result and its generalizations and [**Wid12**] for further recent results. We present here without proof a result of P. G. Spain [**Spa95**], which is a refinement of Davenport's bound, and can be thought of as a continuation of Theorem 4.6.1.

Theorem 4.6.2. *Let the notation be as in Theorem 4.6.1, and let $\mathcal{C}$ be the maximal Lipschitz constant corresponding to parametrization of ∂M. Then for each $t \in \mathbb{R}_{>0}$,*

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

Finally, for very explicit inequalities in the case of counting lattice points in rectangular boxes see [**Fuk06a**], [**Fuk06b**] and [**FH13**].

4.7. Applications of lattices in coding theory and cryptography

Here we very briefly mention two applications of lattices in digital communications. First of this is to *coding theory.* The theory of error correcting codes assumes transmission of a signal from transmitter to receiver over a potentially noisy channel. There is a possibility of two types of errors in the channel:

1. *Erasure:* a character in the signal codeword was erased in transmission.
2. *Alteration:* a character in the signal codeword was alterated in transmission.

We will briefly talk about erasures in the next section. Here, we will say a few words about how lattices can be used to deal with alterations. The main idea of constructing good error correcting codes is to ensure large distance between the codewords (here distance can be defined in different ways, most commonly the Hamming distance, which we do not discuss here). Imagine, for instance, that we use points of a full-rank lattice $L \subset \mathbb{R}^n$ as our codewords. Specifically, let r be a sufficiently large integer and let

$$L_r = \{ \boldsymbol{x} \in L : \|\boldsymbol{x}\| \leq r \}.$$

We can use L_r as our code-space for transmission of information, i.e. signals are converted to points of L_r for transmission. If a codeword $\boldsymbol{x} \in L_r$ is transmitted, an error due to alteration in a noisy channel may result in the introduction of a sufficiently small error vector $\boldsymbol{\varepsilon}$ so that the received codeword is

$$\boldsymbol{x} + \boldsymbol{\varepsilon}.$$

The correction mechanism then needs to strip-off the error and return $\boldsymbol{x}$. Write λ_1 for the first successive minimum of L. Assuming that $\|\boldsymbol{\varepsilon}\| < \lambda_1/2$, we see that $\boldsymbol{x}$ is the solution of CVP on L with the input point $\boldsymbol{x} + \boldsymbol{\varepsilon}$, i.e. the error correction comes down to solving an instance of CVP. While CVP is hard in general, it can be made much easier provided we know a shortest basis for our lattice L: in that case, it can be solved by Babai's nearest hyperplane algorithm (see [**MG02**]) for details).

Another use of lattices in coding theory comes from design of transmitter networks. Given a lattice $\Lambda \in \mathbb{R}^n$, we can regard its nonzero points as transmitters which interfere with the transmitter at the origin, and then a standard measure of the *total interference* of Λ is given by $E_\Lambda(2)$, where

$$E_\Lambda(s) = \sum_{\boldsymbol{x} \in \Lambda \setminus \{\boldsymbol{0}\}} \frac{1}{\|\boldsymbol{x}\|^{2s}} \tag{4.21}$$

is the Epstein zeta-function of Λ, and the signal-to-noise ratio of Λ is defined by

$$\mathrm{SNR}(\Lambda) = 10 \log_{10} \frac{1}{9 E_\Lambda(2)}, \tag{4.22}$$

as in [**BSW97**]. Suppose that we have a network of transmitters positioned at the points of a planar lattice Λ. The plane is tiled with translates of the Voronoi cell of Λ, which are the cells serviced by the corresponding transmitters at their centers. The packing density of Λ is precisely the proportion of the plane covered by the transmitter network. WR lattices allow for transmitters of the same power. To maximize $\mathrm{SNR}(\Lambda)$ on the set of all planar WR lattices of a fixed determinant Δ is the same as to minimize $E_\Lambda(2)$ on this set. This optimization problem is discussed in [**BSW97**] and [**FHL$^+$12**].

Another application of lattices comes from cryptography, resulting in a subarea knows as *lattice cryptography.* The book [**MG02**] is an excellent introduction to this exciting and active area of research. Here we only mention a basic connection. Cryptography assumes transmission of information over an unsecured channel, which allows for for intruders to intercept the message. The goal is to encode a message in a way that allows the intended receiver to easily decode it, but makes decoding very hard for intruders. Asymmetric cryptography then recognizes that the transmitter does not need to be able to decode the message, only the receiver needs to be able to do this. Encoding is done with use of a *public key*, i.e. a piece of publicly available information, while the decoding requires a *private key* known only to the receiver. The security of such a scheme is based on the assumption that decoding without the private key is a computationally hard problem. We describe one example of a lattice crypto-system based on CVP: this is the GGH encryption scheme, named after its creators O. Goldreich, S. Goldwasser and S. Halevi. Let L be a lattice in $\mathbb{R}^n$, and define:

- *Private key:* a shortest basis B for L and a matrix $U \in \mathrm{GL}_n(\mathbb{Z})$,
- *Public key:* a basis $B' = BU$ for L.

Let $\boldsymbol{m} \in \mathbb{Z}^n$ be message text and let $\boldsymbol{\varepsilon}$ be a small error vector. To encrypt $\boldsymbol{m}$, take

$$\boldsymbol{m}' = B'\boldsymbol{m} + \boldsymbol{\varepsilon}.$$

If we know the private key B and U, we can compute B^{-1} and U^{-1} and decrypt as follows:

$$B^{-1}\boldsymbol{m}' = B^{-1}BU\boldsymbol{m} + B^{-1}\boldsymbol{\varepsilon} = U\boldsymbol{m} + B^{-1}\boldsymbol{\varepsilon}$$

then use Babai's nearest hyperplane algorithm to solve the corresponding instance of CVP retrieving $U\boldsymbol{m}$ and compute $\boldsymbol{m}$ multiplying by U^{-1}. On the other hand, the intruders possessing only the public key would attempt to do the same and receive:

$$(B')^{-1}\boldsymbol{m}' = \boldsymbol{m} + (B')^{-1}\boldsymbol{\varepsilon} = \boldsymbol{m} + (B')^{-1}\boldsymbol{\varepsilon},$$

where $\|(B')^{-1}\boldsymbol{\varepsilon}\|$ can be sufficiently big to produce an incorrect CVP solution, i.e. the resulting lattice vector would be different from the message text $\boldsymbol{m}$. Unfortunately, this algorithm does have some security issues as demonstrated by P. Nguyen in 1999. This being said, it still serves as a good illustration of the lattice encryption idea. There is also a good number of other more secure encryption schemes based on lattices and high complexity of lattice problems.

4.8. Euclidean frames

An (n,k)-*frame* in a Euclidean space $\mathbb{R}^k$ is a set $\{\boldsymbol{x}_1,\dots,\boldsymbol{x}_n\}$, $n \geq k$, that satisfies the following property: there exist positive real constants γ_1, γ_2 such that for every vector $\boldsymbol{v} \in \mathbb{R}^k$,

$$\gamma_1 \|\boldsymbol{v}\|^2 \leq \sum_{i=1}^n \langle \boldsymbol{v}, \boldsymbol{x}_i \rangle^2 \leq \gamma_2 \|\boldsymbol{v}\|^2,$$

where $\langle\ ,\ \rangle$ denotes the usual Euclidean inner product. It is not difficult to see (Problem 4.12) that a frame is a spanning set for $\mathbb{R}^k$. In this section we will discuss some particular types of frames, their properties and their connections to lattices. In particular, we will be interested in *uniform* frames, meaning that $\|\boldsymbol{x}_1\| = \dots = \|\boldsymbol{x}_n\|$ (if this common value is 1, we call such a frame *unit*). Further, we will say that a uniform (n,k)-frame $\{\boldsymbol{x}_1,\dots,\boldsymbol{x}_n\}$ is *tight* if $\gamma_1 = \gamma_2$, i.e. if there exists a positive real constant c such that for every vector $\boldsymbol{v} \in \mathbb{R}^k$,

$$\|\boldsymbol{v}\|^2 = c \sum_{i=1}^n \langle \boldsymbol{v}, \boldsymbol{x}_i \rangle^2 . \tag{4.23}$$

Compare (4.23) to (4.10) and observe the following fact: a full-rank lattice $\Lambda \subset \mathbb{R}^k$ with the set of minimal vectors

$$S(\Lambda) = \{\boldsymbol{x}_1,\dots,\boldsymbol{x}_n\}$$

is strongly eutactic if and only if $S(\Lambda)$ forms a uniform tight (n,k)-frame in $\mathbb{R}^k$. Problem 4.12 implies that such a lattice must also be WR (a fact that we mentioned before, as it is true for eutactic and for perfect lattices too).

We will be especially interested in a more specialized class of tight frames. A uniform tight (n,k)-frame $\{\boldsymbol{x}_1,\dots,\boldsymbol{x}_n\}$ is called *equiangular* (abbreviated ETF = equiangular tight frame) if there exists a real constant α such that

$$\alpha = \frac{|\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle|}{\|\boldsymbol{x}_i\|\|\boldsymbol{x}_j\|}\ \forall\ i \neq j,$$

in other words if absolute value of the cosine of the angle between any distinct pair of frame vectors $\boldsymbol{x}_i, \boldsymbol{x}_j$ is the same. By a certain abuse of notation, we will refer to α as the *angle* of this ETF. ETF's are a generalization of an orthonormal basis: indeed, an orthonormal basis for $\mathbb{R}^k$ is a unit (k,k)-ETF of angle 0. The first nontrivial example of an ETF is the *Mercedes-Benz frame* of three equiangular unit vectors in $\mathbb{R}^2$:

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1/2 \\ \sqrt{3}/2 \end{pmatrix}, \begin{pmatrix} -1/2 \\ \sqrt{3}/2 \end{pmatrix}.$$

The angle of this $(3,2)$-ETF is $1/2$, and it is equal to $S'(\Lambda_h)$ where Λ_h is the hexagonal lattice and $S'(\Lambda_h)$ is the subset of the set of minimal vectors $S(\Lambda_h)$ obtained by choosing one vector from each ± pair. More generally, we can ask how large can an ETF in $\mathbb{R}^k$ be? To this end, there is the following bound.

Theorem 4.8.1 (Gerzon). *If $X = \{\boldsymbol{x}_1,\dots,\boldsymbol{x}_n\}$ is an ETF in $\mathbb{R}^k$, then*

$$n \leq \frac{k(k+1)}{2}.$$

PROOF. Normalizing, if necessary, we can assume that X is a unit frame of angle α, then

$$\alpha = |\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle| \ \forall\ i \neq j,\ \ |\langle \boldsymbol{x}_i, \boldsymbol{x}_i \rangle| = 1 \ \forall\ 1 \leq i \leq n.$$

For each $1 \leq i \leq n$, consider the $k \times k$ symmetric matrix $\boldsymbol{x}_i \boldsymbol{x}_i^\top$ as a vector in $\mathbb{R}^{k^2}$, then we can compute the inner products

$$\left\langle \boldsymbol{x}_i \boldsymbol{x}_i^\top, \boldsymbol{x}_j \boldsymbol{x}_j^\top \right\rangle = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle^2 = \begin{cases} \alpha^2 & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases}$$

Consider a symmetric $k \times k$ matrix

$$A = \sum_{i=1}^n a_i \boldsymbol{x}_i \boldsymbol{x}_i^\top,$$

then the squared norm of A is

$$\begin{aligned} \langle A, A \rangle &= \sum_{i=1}^n \sum_{j=1}^n a_i a_j \left\langle \boldsymbol{x}_i \boldsymbol{x}_i^\top, \boldsymbol{x}_j \boldsymbol{x}_j^\top \right\rangle = \sum_{i=1}^n \sum_{j=1}^n \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle^2 a_i a_j \\ &= \sum_{i=1}^n a_i^2 + 2\alpha^2 \sum_{i=1}^n \sum_{j=i+1}^n a_i a_j, \end{aligned}$$

which is a positive definite quadratic form in the variables $a_1, \dots, a_n$. Hence $A = \boldsymbol{0}$ if and only if $\langle A, A \rangle = 0$, which happens if and only if $a_1 = \dots = a_n = 0$. Thus the set

$$X^* = \left\{ \boldsymbol{x}_i \boldsymbol{x}_i^\top : 1 \leq i \leq n \right\}$$

is linearly independent, meaning that its cardinality n cannot be larger than the dimension of the space of all real symmetric $k \times k$ matrices, which is $\frac{k(k+1)}{2}$. □

Notice that Gerzon's bound is sharp, as demonstrated by the Mercedes-Benz example: $\frac{2(2+1)}{2} = 3$. ETF's achieving Gerzon's bound are called *maximal.* One can ask in which dimensions do maximal ETFs occur? In fact, only four examples are known: $(3,2)$, $(6,3)$, $(28,7)$ and $(276,23)$. Out of these examples, besides the hexagonal lattice, the $(28,7)$-ETF is $S'(\Lambda)$ of a certain perfect strongly eutactic lattice in $\mathbb{R}^7$ and the $(276,23)$-ETF appears among the set of minimal vectors of the famous 24-dimensional Leech lattice. There are many other examples of ETFs (although not maximal) appearing as sets of minimal vectors of strongly eutactic lattices, however these lattices are not perfect: perfection would require $|S'(\Lambda)| \geq \frac{k(k+1)}{2}$, which implies maximal ETF; see [**BFG$^+$16**], [**BF17**] for details.

Frames have applications in information transmission, for instance in recovering erasures (as defined in Section 4.7 above) in signal transmission. If $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\} \subset \mathbb{R}^k$ is an (n,k)-frame, then an arbitrary vector $\boldsymbol{v} \in \mathbb{R}^k$ can be written in these frame coordinates as

$$\boldsymbol{v}(X) = (v_1(X), \dots, v_n(X)),$$

where $v_i(X) = \langle \boldsymbol{v}, \boldsymbol{x}_i \rangle$ for each $1 \leq i \leq n$. This generalizes the notion of coordinates of a vector with respect to a basis. We can now transmit the message encoded by $\boldsymbol{v}$ by instead transmitting its vector of coordinates $\boldsymbol{v}(X)$. The advantage of using an overdetermined frame is that the vector $\boldsymbol{v}$ can be reconstructed with a certain degree of accuracy even if some of the coordinates were lost in transmission, i.e. if erasures occurred (see [**HP04**] for more information on this). Hence, ideally we want the cardinality n of the frame to be large as compared to the dimension k: this gives more coordinates, an so a better potential chance at accurate reconstruction of a signal in the presence of erasures. On the other hand, the accuracy of reconstruction is increased if the frame vectors are not "aligned" with each other, i.e. if the angles between them are large. To this end, let us define the frame coherence.

Given an arbitrary (n,k)-frame $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\} \subset \mathbb{R}^k$, we define its *coherence* to be

$$C(X) = \max_{1 \leq i \neq j \leq n} \frac{|\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle|}{\|\boldsymbol{x}_i\| \|\boldsymbol{x}_j\|},$$

i.e. maximal absolute value of cosines of the angles between pairs of frame vectors. This is a measure of how aligned with each other frame vectors are. For instance, coherence of an orthogonal basis is 0 and coherence of an ETF is its angle. Thinking of frame vectors as frequencies encoding a signal, coherence represents the measure of interference between different frequencies used in transmission: the lower this interference is the better it is for accurate transmission. Notice, however, that the larger is the cardinality of a frame the more aligned its vectors would have to be, and so the higher would be its coherence. Hence we have the following optimization problem.

PROBLEM 4.1. *Construct (n,k)-frame $X \subset \mathbb{R}^k$ with large cardinality n as compared to the dimension k and low coherence $C(X)$.*

The first question to ask then is how small can coherence be? To this end, we have the following inequality.

THEOREM 4.8.2. *(Welch) Given an (n,k)-frame $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\} \subset \mathbb{R}^k$,*

$$C(X) \geq \sqrt{\frac{n-k}{k(n-1)}}.$$

The equality is achieved if and only if X is an ETF.

PROOF. Rescaling the vectors, if necessary, we can assume that X is a unit frame: rescaling does not change coherence. Let us write

$$A = (\boldsymbol{x}_1 \ \dots \ \boldsymbol{x}_n)$$

for the $k \times n$ matrix whose columns are the frame vectors. Define the corresponding $n \times n$ *Gram matrix* to be

$$G = A^\top A,$$

then the ij-entry of G is equal to $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$. In particular, all the diagonal entries of G are equal to 1, and so its trace $\mathrm{Tr}(G) = n$. On the other hand, trace of G is equal to the sum of its eigenvalues $\lambda_1, \dots, \lambda_n$. Notice that G is a positive semi-definite matrix of rank k, so we can assume that

$$\lambda_1, \dots, \lambda_k > 0, \ \lambda_{k+1} = \dots = \lambda_n = 0.$$

Then

$$n^2 = \mathrm{Tr}(G)^2 = \left(\sum_{i=1}^k \lambda_i \right)^2 \leq k \sum_{i=1}^k \lambda_i^2 = k \sum_{i=1}^n \lambda_i^2. \tag{4.24}$$

by Cauchy-Schwartz inequality. Recall that the *Frobenius norm* of this Gram matrix is simply its Euclidean norms viewed as a vector in $\mathbb{R}^{n^2}$, i.e.

$$\|G\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} | \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle |^2 .$$

The symmetric matrix G is diagonalizable by some orthogonal matrix U, i.e.

$$UGU^\top = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n \end{pmatrix},$$

and the Frobenius norm is invariant under such diagonalization, i.e.

$$\|G\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} | \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle |^2 = \sum_{i=1}^{n} \lambda_i^2 . \tag{4.25}$$

Combining (4.24) and (4.25), we obtain

$$\sum_{i=1}^{n} \sum_{j=1}^{n} | \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle |^2 \geq \frac{n^2}{k} . \tag{4.26}$$

On the other hand, notice that $| \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle | \leq C(X)$ (with equality if an only if X is an ETF) for $i \neq j$ and $| \langle \boldsymbol{x}_i, \boldsymbol{x}_i \rangle | = 1$, hence

$$\frac{n^2}{k} \leq n + \sum_{i \neq j} | \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle |^2 \leq n + (n^2 - n) C(X)^2,$$

with the last inequality being equality if and only if X is an ETF, therefore

$$C(X)^2 \geq \frac{1}{n^2 - n} \left(\frac{n^2}{k} - n \right) = \frac{n-k}{k(n-1)},$$

which establishes the Welch bound. Further, if X is an ETF all nonzero eigenvalues of G are the same, and so

$$n = \operatorname{Tr}(G) = k\lambda_1,$$

meaning that

$$\sum_{i=1}^{n} \lambda_i^2 = k\lambda_1^2 = k \left(\frac{n^2}{k^2} \right) = \frac{n^2}{k},$$

in other words there is an equality in (4.26). This establishes equality in the Welch bound if and only if X is an ETF. □

We then have an immediate consequence of Welch's bound for ETFs.

Corollary 4.8.3. *If X is an (n, k) ETF of angle α, then*

$$\alpha = \sqrt{\frac{n-k}{k(n-1)}} .$$

In fact, there are some additional remarkable properties an (n, k) ETF of angle α must possess.

- *(Neumann)* If $n > 2k$, then $1/\alpha$ is an odd integer.

- *(Sustik, Tropp, Dhillon, Heath)* Let $1 < k < n-1$. Suppose $n \neq 2k$, then $1/\alpha$ is an odd integer and the quantity
$$\sqrt{\frac{(n-k)(n-1)}{k}}$$
is also an odd integer. If $n = 2k$, then k must be an odd integer and $n-1$ the sum of two squares.

These and related properties are used to eliminate the pairs (n, k) for which ETFs cannot exist. One of the main goals here is to find more maximal ETFs beyond the four mentioned above.

4.9. Problems

PROBLEM 4.2. *Prove that the optimal kissing number in* $\mathbb{R}^2$ *is equal to* 6.

PROBLEM 4.3. *Prove that similarity is an equivalence relation on the set of all lattices of full rank in* $\mathbb{R}^n$.

PROBLEM 4.4. *Assume two full-rank lattices* L *and* M *in* $\mathbb{R}^n$ *are similar. Prove that they have the same packing density, covering thickness and kissing number.*

PROBLEM 4.5. *Prove that the set of all real orthogonal* $n \times n$ *matrices* $\mathcal{O}_n(\mathbb{R})$ *is a subgroup of* $\mathrm{GL}_n(\mathbb{R})$.

PROBLEM 4.6. *Let* $\boldsymbol{x}_1$ *and* $\boldsymbol{x}_2$ *be nonzero vectors in* $\mathbb{R}^2$ *so that the angle* θ *between them satisfies* $0 < \theta < \frac{\pi}{3}$. *Prove that*

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

PROBLEM 4.7. *Let* $\Lambda \subset \mathbb{R}^2$ *be a lattice of full rank with successive minima* $\lambda_1 \leq \lambda_2$, *and let* $\boldsymbol{x}_1, \boldsymbol{x}_2$ *be the vectors in* Λ *corresponding to* λ_1, λ_2, *respectively. Let* $\theta \in [0, \pi/2]$ *be the angle between* $\boldsymbol{x}_1$ *and* $\boldsymbol{x}_2$. *Prove that*

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

PROBLEM 4.8. *Let* L *and* M *be two similar lattices. Prove that if* L *is eutactic (respectively, strongly eutactic, perfect), then so is* M.

PROBLEM 4.9. *Prove that the hexagonal lattice* Λ_h *is both, perfect and eutactic. Further, prove that if* L *is a perfect lattice in* $\mathbb{R}^2$, *then* $L \sim \Lambda_h$.

PROBLEM 4.10. *Prove that an ordered basis* $\boldsymbol{b}_1, \boldsymbol{b}_2$ *for a planar lattice* Λ *consists of vectors corresponding to successive minima* λ_1, λ_2, *respectively, if and only if*

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

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

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

where $\lfloor \mu \rceil$ *stands for the nearest integer to* μ, *produces a shorter second basis vector.*

PROBLEM 4.11. *Prove that the Gauss-Lagrange Algorithm as discussed in Section 4.3 terminates in a finite number of steps.*

PROBLEM 4.12. *Let* $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ *be an* (n, k)*-frame in* $\mathbb{R}^k$, $n \geq k$. *Prove that* $\mathbb{R}^k = \mathrm{span}_{\mathbb{R}}\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$.

Bibliography

[AFH12] I. Aliev, L. Fukshansky, and M. Henk. Generalized Frobenius numbers: bounds and average behavior. *Acta Arithm.*, 155:53–63, 2012.

[AG07] I. Aliev and P. M. Gruber. An optimal lower bound for the Frobenius problem. *J. Number Theory*, 123(1):71–79, 2007.

[AH09] I. Aliev and M. Henk. Integer knapsacks: Average behavior of the Frobenius numbers. *Math. Oper. Res.*, 34(3):698–705, 2009.

[AH10] I. Aliev and M. Henk. On feasibility of integer knapsacks. *SIAM J. Optim.*, 20(6):2978–2993, 2010.

[AHH11] I. Aliev, M. Henk, and A. Hinrichs. Expected Frobenius numbers. *J. Comb. Theory A*, 118:525–531, 2011.

[AHL13] I. Aliev, M. Henk, and E. Linke. Integer points in knapsack polytopes and s-covering radius. *Electron. J. Combin.*, 20(2, Paper 42):17 pp., 2013.

[ALL16] I. Aliev, J. De Loera, and Q. Louveaux. Parametric polyhedra with at least k lattice points: their semigroup structure and the k-Frobenius problem. In *Recent trends in combinatorics, IMA Vol. Math. Appl., 159*, pages 753–778. Springer, 2016.

[Arn99] V. I. Arnold. Weak asymptotics for the numbers of solutions of diophantine problems. *Funct. Anal. Appl.*, 33(4):292–293, 1999.

[Arn06] V. I. Arnold. Geometry and growth rate of Frobenius numbers of additive semigroups. *Math. Phys. Anal. Geom.*, 9(2):95–108, 2006.

[BCKV00] D. Bump, K. K. Choi, P. Kurlberg, and J. Vaaler. A local Riemann hypothesis, I. *Math. Z.*, 233(1):1–19, 2000.

[BF17] A. Böttcher and L. Fukshansky. Addendum to "Lattices from equiangular tight frames". *Linear Algebra Appl.*, 531:592–601, 2017.

[BFG$^+$16] A. Böttcher, L. Fukshansky, S. R. Garcia, H. Maharaj, and D. Needell. Lattices from tight equiangular frames. *Linear Algebra Appl.*, 510:395–420, 2016.

[Bor02] P. Borwein. *Computational Excursions in Analysis and Number Theory*. Canadian Mathematical Society, 2002.

[BR04] M. Beck and S. Robins. A formula related to the Frobenius problem in two dimensions. *Number theory (New York, 2003), Springer, New York*, pages 17–23, 2004.

[BR06] M. Beck and S. Robins. *Computing the Continuous Discretely. Integer-Point Enumeration in Polyhedra*. Springer-Verlag, 2006.

[BS07] J. Bourgain and Y. G. Sinai. Limit behaviour of large Frobenius numbers. *Russ. Math. Surv.*, 62(4):713–725, 2007.

[BSW97] M. Bernstein, N. J. A. Sloane, and P. E. Wright. On sublattices of the hexagonal lattice. *Discrete Math.*, 170(1-3):29–39, 1997.

[Cas53] J. W. S. Cassels. A short proof of the Minkowski-Hlawka theorem. *Proc. Cambridge Philos. Soc.*, 49:165–166, 1953.

[Cas59] J. W. S. Cassels. *An Introduction to the Geometry of Numbers*. Springer-Verlag, 1959.

[CE03] H. Cohn and N. Elkies. New upper bounds on sphere packings. I. *Ann. of Math. (2)*, 157(2):689–714, 2003.

[CKM$^+$17] H. Cohn, A. Kumar, S. D. Miller, D. Radchenko, and M. S. Viazovska. The sphere packing problem in dimension 24. *Ann. of Math. (2)*, 185(3):1017–1033, 2017.

[Coh00] H. Cohen. *A Course in Computational Algebraic Number Theory*. GTM. 138. Springer, 2000.

[CS99] J. H. Conway and N. J. A. Sloane. *Sphere Packings, Lattices, and Groups. 3rd edition*. Springer-Verlag, 1999.

[Dav51] H. Davenport. On a principle of Lipschitz. *J. London Math. Soc.*, 26:179–183, 1951.

[Ewa96] G. Ewald. *Combinatorial convexity and algebraic geometry.* Springer-Verlag, 1996.

[FH13] L. Fukshansky and G. Henshaw. Lattice point counting and height bounds over number fields and quaternion algebras. *Online J. Anal. Comb.*, 8:20 pp., 2013.

[FHL$^+$12] L. Fukshansky, G. Henshaw, P. Liao, M. Prince, X. Sun, and S. Whitehead. On integral well-rounded lattices in the plane. *Discrete Comput. Geom.*, 48(3):735–748, 2012.

[FR07] L. Fukshansky and S. Robins. Frobenius problem and the covering radius of a lattice. *Discrete Comput. Geom.*, 37(3):471–483, 2007.

[FS11] L. Fukshansky and A. Schürmann. Bounds on generalized Frobenius numbers. *European J. Combin.*, 32(3):361–368, 2011.

[FS20] L. Fukshansky and Y. Shi. Positive semigroups and generalized Frobenius numbers over totally real number fields. *Mosc. J. Comb. Number Theory*, 9(1):29–41, 2020.

[Fuk06a] L. Fukshansky. Integral points of small height outside of a hypersurface. *Monatsh. Math.*, 147(1):25–41, 2006.

[Fuk06b] L. Fukshansky. Siegel's lemma with additional conditions. *J. Number Theory*, 120(1):13–25, 2006.

[Fuk11] L. Fukshansky. Revisiting the hexagonal lattice: on optimal lattice circle packing. *Elem. Math.*, 66(1):1–9, 2011.

[GL87] P. M. Gruber and C. G. Lekkerkerker. *Geometry of Numbers.* North-Holland Publishing Co., 1987.

[Hal05] T. Hales. A proof of the Kepler conjecture. *Ann. of Math. (2)*, 162(3):1065–1185, 2005.

[Hen02] M. Henk. Successive minima and lattice points. *IV International Conference in Stochastic Geometry, Convex Bodies, Empirical Measures and Applications to Engineering Science, Vol. I (Tropea, 2001). Rend. Circ. Mat. Palermo (2) Suppl. No. 70, part I*, pages 377–384, 2002.

[HP04] R.B. Holmes and V.I. Paulsen. Optimal frames for erasures. *Linear Algebra Appl.*, 377:31–51, 2004.

[HPS08] J. Hoffstein, J. Pipher, and J. H. Silverman. *An Introduction to Mathematical Cryptography.* Springer, 2008.

[Jar41] V. Jarnik. Zwei Bemerkungen zur Geometrie de Zahlen. *Věstnik Krălovské České Společnosit Nauk*, 1941.

[Kan92] R. Kannan. Lattice translates of a polytope and the Frobenius problem. *Combinatorica*, 12(2):161–177, 1992.

[Kar72] R. Karp. Reducibility among combinatorial problems. In *Complexity of computer computations (Proc. Sympos., IBM Thomas J. Watson Res. Center, Yorktown Heights, N.Y., 1972)*, pages 85–103. Plenum, New York, 1972.

[Lan94] S. Lang. *Algebraic Number Theory.* Springer-Verlag, 1994.

[Lee56] J. Leech. The problem of thirteen spheres. *Math. Gaz.*, 40:22–23, 1956.

[Li15] H. Li. Effective limit distribution of the Frobenius numbers. *Compos. Math.*, 151(5):898–916, 2015.

[Lip65] R. Lipschitz. *Monatsber. der Berliner Academie*, pages 174–185, 1865.

[LLL82] A. K. Lenstra, H. W. Lenstra, and L. Lovasz. Factoring polynomials with rational coefficients. *Math. Ann.*, 261:515–534, 1982.

[Mar10] J. Marklof. The asymptotic distribution of Frobenius numbers. *Invent. Math*, 181:179–207, 2010.

[MG02] D. Micciancio and S. Goldwasser. *Complexity of lattice problems. A cryptographic perspective.* The Kluwer International Series in Engineering and Computer Science, 671. Kluwer Academic Publishers, Boston, MA, 2002.

[MT90] S. Martello and P. Toth. *Knapsack problems. Algorithms and computer implementations.* Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Ltd., Chichester, 1990.

[Ram05] J. L. Ramírez Alfonsín. *The Diophantine Frobenius problem.* Oxford University Press, 2005.

[Sch50] P. Scherk. Convex bodies off center. *Archiv Math.*, 3:303, 1950.

[Spa95] P. G. Spain. Lipschitz: a new version of old principle. *Bull. London Math. Soc.*, 27:565–566, 1995.

[SSU09] V. Shchur, Ya. Sinai, and A. Ustinov. Limiting distribution of Frobenius numbers for $n = 3$. *Journal of Number Theory*, 129:2778–2789, 2009.

[Str12] A. Strömbergsson. On the limit distribution of Frobenius numbers. *Acta Arith.*, 152.(1):81–107, 2012.

[SvdW53] K. Schütte and B. L. van der Waerden. Das Problem der dreizehn Kugeln. *Math. Ann.*, 125:325–334, 1953.

[Thu93] J. L. Thunder. The number of solutions of bounded height to a system of linear equations. *J. Number Theory*, 43:228–250, 1993.

[Ust10] A. Ustinov. On the distribution of Frobenius numbers with three arguments. *Izv. Math.*, 74:1023–1049, 2010.

[Via17] M. S. Viazovska. The sphere packing problem in dimension 8. *Ann. of Math. (2)*, 185(3):991–1015, 2017.

[Wid12] M. Widmer. Lipschitz class, narrow class, and counting lattice points. *Proc. Amer. Math. Soc.*, 140(2):677–689, 2012.