On Lattice Methods in Integer Optimization

Proefschrift

te verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op

vrijdag 27 september 2013 om 12:30

door

Frederik Jonas VON HEYMANN

Diplom-Mathematiker (Freie Universität Berlin)
geboren te Siegburg, Duitsland

Dit proefschrift is goedgekeurd door de promotor:

Prof. dr. ir. K.I. Aardal

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. ir. K.I. Aardal,	Technische Universiteit Delft, promotor
Prof. dr. C. Haase,	Goethe Universität Frankfurt am Main
Prof. dr. A. Lodi,	Università di Bologna
Prof. dr. F.H.J. Redig,	Technische Universiteit Delft
Prof. dr. A. Schürmann,	Universität Rostock
Prof. dr. F. Vallentin,	Universität zu Köln
Prof. dr. L.A. Wolsey,	Université catholique de Louvain
Prof. dr. J.M.A.M. van Neerven,	Technische Universiteit Delft, reservelid

Het onderzoek beschreven in dit proefschrift is mede gefinancierd door de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO), onder projectnummer 613.000.801

Screen Version
Print Version: ISBN 978-94-6191-874-1

Preface

Why are we doing anything we are doing? Why is it that I studied mathematics and ended up writing this thesis? There seems to be an unknown force, pushing us forward, driving us further and further to investigate the unknown. One might say we do so to make sure there are no lurking dangers, or to learn how to deal with them. Even if this sounds reasonable, the more convincing reason to me has always been that we want to feel the wonder of discovery.

For me personally, the main reason to study mathematics was to find out how things *really* work, interact, and are constructed. The hard lesson I had to learn, like anyone who has caught a glimpse into research, was that there will always be more things we do not know and understand, than there are things we do. Or, to quote Socrates: "The only true wisdom is to know that you know nothing."

The questions that drew me towards the topics of this thesis can maybe, very much simplified, be formulated as follows: Why can the discrete setting be more difficult than the continuous? Why is there ambiguity in the discrete counterparts to unique continuous objects? How can few dimensions be more complicated than many?

These questions will not be answered in this thesis, and maybe it is in their nature to remain unanswered, and instead guide us towards new discoveries. Some steps in this direction might be found on the pages to follow.

There are many people I would like to mention, people that made this thesis possible. First and foremost I want to thank my promoter Karen Aardal for her relentless support, and for being the best promoter anyone could ever ask for. I would be honored to call her a friend in the future.

I also want thank Andrea Lodi and Laurence Wolsey for answering my many questions and for making me feel welcome from the first meeting on. It was these meetings in Aussois, Brussels, and Bologna that showed me how much fun research can be when done with the right group of people.

My first year in Delft would have been very different without Achill Schürmann and Frank Vallentin, whose enthusiasm for mathematics and overall attitude towards science and life will always be an inspiration to me.

After my graduation in Berlin, Christian Haase gave me the opportunity to start my exploration of the discrete (and colorful) mathematics, and I owe him my deepest gratitude for this.

I thank Dion Gijswijt for always having an open door and for proof-reading parts of this thesis. The latter also applies to Anna Gundert, and I thank her for that, but even more for being my partner, keeping me sane, and so much more.

The time in Delft would have been much more dull without all the great people I had the pleasure to share the time and space with, and so I thank all (former) PhD students for the many inspiring and entertaining chats, specifically Sjoerd Dirksen, Matthijs Pronk, Pieter van den Berg, Sonja Cox, Guido Oud, Evan DeCorte, Shanfei Li, David de Laat, Jan Rozendaal, and Özlem Cavusoglu.

Finally, working on this thesis has certainly had an impact on my social life. I thank all my friends and family for putting up with me during this time, and in particular the ones who were hit the hardest: Anna, Peter, and my parents.

Remarks on the Screen Version

This version is almost verbatim identical with the version submitted and printed. What was changed is the layout: Here it is based on A4-paper, there are no left and right pages, and the footer and header are slightly altered for easier navigation on a screen. There is also a touch of color added, the page numbers correspond to the actual page-number in the document, and the index was removed since it is vastly inferior to any digital search-function.

The only exceptions to the above are this section and some very minor changes to fix layout-problems due to the changed text-width.

Contents

Introduction

Integer optimization is a powerful modeling tool both for problems of practical and more abstract origin. Since the 1970s we have seen huge progress in the size of problem instances that can be tackled. This progress is mostly due to the many results in polyhedral combinatorics and to algorithms and implementations related to the polyhedral results. In the theory of integer optimization we have also seen exciting results related to the algebraic structure of the set of integer points in polyhedra together with algorithms that exploit them.

One prominent such result is the integer programming algorithm of Lenstra [58] that finds, in polynomial time, an integer point in a polyhedron or concludes that no such point exists, if the dimension is fixed. The key ingredient in Lenstra's result is lattice basis reduction. In this thesis we will present results that make a step in the direction of merging the approach of polyhedral combinatorics with a reformulation technique built on lattice basis reduction.

In Chapters 2, 3, and 4, the leading question will be, generally speaking, how to solve the integer optimization program

$$\max \left\{ \boldsymbol{c}^{\mathsf{T}} \boldsymbol{x} \ : \ \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},\ \boldsymbol{x} \geq \boldsymbol{0}, \boldsymbol{x} \in \mathbb{Z}^n \right\}, \tag{1}$$

where $\boldsymbol{A}$ is an integer $m \times n$ matrix of full row rank, $\boldsymbol{b}$ an integer m-vector, and $\boldsymbol{c}$ an integer n-vector.

It is of particular interest here to observe that the presence of equality constraints implies that the feasible region will lie in an affine subspace of $\mathbb{R}^n$, and thus we can equivalently express the program in less variables than it is given to us, by reformulating it in terms of this subspace.

While the methods of branch-and-bound (or branch-and-cut) in connection with linear optimization and cutting plane algorithms provide successful all-purpose methods for integer optimization, they often do not perform satisfactorily if the feasible set of the linear relaxation is not full-dimensional.

Starting with the above-mentioned algorithm of Lenstra [58], several lattice-based approaches to reformulate the feasible region have been proposed, see, e.g., [2, 5, 25, 54, 59, 60, 63]. Here we will consider the reformulation as in [2]:

$$\boldsymbol{x} := \boldsymbol{x}^0 + \boldsymbol{Q}\boldsymbol{\mu}\,, \tag{2}$$

where $\boldsymbol{x}^0 \in \mathbb{Z}^n$ satisfies $\boldsymbol{A}\boldsymbol{x}^0 = \boldsymbol{b}$, $\boldsymbol{\mu} \in \mathbb{Z}^{n-m}$, and $\boldsymbol{Q}$ is a basis for the lattice $\ker_{\mathbb{Z}}(\boldsymbol{A}) = \{\boldsymbol{x} \in \mathbb{Z}^n \ : \ \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\}$. Then $\boldsymbol{Q}$ indeed captures the configuration of the integer points that lie in the affine subspace containing all feasible points.

Due to the nonnegativity requirements on the $\boldsymbol{x}$-variables, one now obtains an equivalent formulation of the integer program (1):

$$\max \left\{ \boldsymbol{c}^{\mathsf{T}}(\boldsymbol{x}^0 + \boldsymbol{Q}\boldsymbol{\mu}) \ : \ \boldsymbol{Q}\boldsymbol{\mu} \geq -\boldsymbol{x}^0 \right\} . \tag{3}$$

This reformulation has been shown to be of particular computational interest in the case where $\boldsymbol{Q}$ is reduced in the sense of Lovász [57].

Chapter 1 provides the basic notations and results we will make use of in the other chapters. A fair amount of details is provided, and where the technicalities seem to lead too far from the topic of the other chapters, further references are provided.

In Chapter 2, we will bring the reformulation given in (3) together with cutting plane algorithms, a type of algorithm that was first developed in the 1960s and 1970s (see, e.g., [10, 11, 20, 41, 42, 43, 69]) and is now an integral part of the most successful integer optimization programs: Given $P = \{\boldsymbol{x} \in \mathbb{R}^n \ : \ \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \ \boldsymbol{x} \geq \boldsymbol{0}\}$, we want to find inequalities that are satisfied by all $\boldsymbol{x} \in P \cap \mathbb{Z}^n$ but violated by some points in P, ideally leading us to the convex hull of $P \cap \mathbb{Z}^n$.

We will show that while all inequalities we obtain in the reformulation will theoretically be obtainable in the original space by similar methods, the reformulation still provides a good heuristic for inequalities that are more difficult to find without this technique.

This chapter is based on joint work with Karen Aardal, Andrea Lodi, and Laurence Wolsey.

In Chapter 3 we observe that one can enrich the reformulation given in (3) by not only taking the lattice $\ker_{\mathbb{Z}}(\boldsymbol{A})$ into consideration, but also the general shape of the polytope $\hat{P} = \left\{\boldsymbol{x} \in \mathbb{R}^{n-m} \ : \ \boldsymbol{Q}\boldsymbol{\mu} \geq -\boldsymbol{x}^0\right\}$ in this lattice.

Taking the shape into account when reducing the lattice has already been observed by Lovász in [44], as part of a proof of the theorem of Lenstra we mentioned before. However, this result is of theoretical nature based on the ellipsoid method for linear programming, and to date no implementation of the ellipsoid method has managed to convince the community of being more useful for the settings described in (1) and (3) than the classical approaches of branch-and-bound and cutting planes.

In [70], Nemirovski provided a constructive way of finding ellipsoids of large volume inside polytopes, even in the case where the polytope is not full-dimensional. We use this result to obtain an implementation of the idea to reduce the lattice $\ker_{\mathbb{Z}}(\boldsymbol{A})$ with respect to the shape of $\hat{P}$. We therefore obtain that the longer a reduced basis vector is, the smaller the number of lattice hyperplanes orthogonal to it we will need to cover all integer points in $\hat{P}$.

This chapter is based on joint work with Karen Aardal and Pim Otte.

Chapter 4 is more theoretical in nature. Some of the hard instances in the literature that have been successfully tackled by lattice-based techniques, such as market split and certain classes of knapsack instances, have randomly generated input $\boldsymbol{A}$. Since the considered instances are very hard even in low dimension, less experience is available for larger instances. In Chapter 4 we study such larger instances and observe that the LLL-reduced basis of $\ker_{\mathbb{Z}}(\boldsymbol{A})$ has a specific sparse structure.

In particular, this implies a map in which some of the original variables get a "rich" translation into a new variable space, whereas some variables are only substituted in the new space. If an original variable is important in the sense of branching or cutting planes, it is generally desirable to translate this variable in a non-trivial way.

We partially explain the obtained structure of the LLL-reduced basis in the case that the input matrix $\boldsymbol{A}$ consists of one row $\boldsymbol{a}$. Since the input is randomly generated, our analysis will be probabilistic. The key ingredient is a bound on the probability that the LLL algorithm will interchange two subsequent basis vectors.

This chapter is based on joint work with Karen Aardal, a conference-version of this work was published in the proceedings of IPCO 2013 [4], and a version similar to this chapter was accepted to *Mathematics of Operations Research*.

Chapter 5 marks a shift in focus and presents a topic from the more combinatorial side of optimization. Instead of optimizing a linear function in a lattice, with some given inequalities, we now want to optimize the shape of a set X of lattice points, such that the amount of lattice points in the vicinity of X is minimized.

This question is induced by the related continuous question of how to minimize the surface of a set of given volume. While for the continuous setting there is a unique solution in the form of the ball, in the discrete setting there is not always a unique optimal solution.

We study conditions for optimal solutions, in particular for dimension 2, and show that for a certain family of cardinalities the optimal solutions (in general dimension) are indeed unique.

This chapter is based on joint work with Aaron Dall and Birgit Vogtenhuber, and a version of it was published as a research report for the DocCourse Combinatorics and Geometry 2009 [31].

Chapter One
Lattices and Optimization Basics

In this thesis, we will consider integer optimization programs connected to lattices. The current chapter contains the basic definitions and some introductory examples. Readers well-versed in the theory of lattices, linear optimization, and polyhedral combinatorics will not discover any new theorems here and can safely skip to the next chapter.

The material and presentation of this chapter is largely inspired by the first parts of [1] by Aardal and Eisenbrand and [66] by Micciancio and Goldwasser. We assume that the reader is somewhat familiar with the basic notions from linear algebra as, e.g., the real numbers $\mathbb{R}$, vector spaces, and basic matrix-arithmetic. There is an abundance of books on these topics one could mention here as references. Without any claim that others are less suited, one in which all prerequisites are contained is the book [56] by Lang.

We begin by defining what we mean by a "lattice". Given $m \leq n$ linearly independent vectors $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m \in \mathbb{R}^n$, the set

$$L(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m) := \left\{ \sum_{i=1}^{m} x_i \boldsymbol{b}_i \ : \ x_i \in \mathbb{Z} \right\} \tag{1.1}$$

is called the *lattice* in $\mathbb{R}^n$ generated by (or associated to) $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$.

The integer n is called the *dimension* of the lattice, and m is the *rank*. If $n = m$, we call the lattice *full-dimensional*.

The vectors $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$ are called a *lattice basis* (or *basis of the lattice*), and we often represent them as a matrix $\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m]$. Accordingly, we often write $L(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m) = L(\boldsymbol{B})$, or even just L, if there is no danger of confusion.

Observe that (1.1) is also well-defined if $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$ are not linearly independent, and we will sometimes use $L(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m)$ to describe the smallest lattice (with respect to set-containment) containing $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$.

From Figure 1.1 we can already extract a very important observation: While every depicted basis is a basis of $\mathbb{R}^2$, the lattices are not all the same. We say that K is a *sublattice* of some given lattice L if $K \subseteq L$ is nonempty and again a lattice. This means that L and K might have the same rank, in contrast to the common definition of subspaces and their dimension. Let K be a sublattice of L, then K is a *pure sublattice* of L if $K = \operatorname{span}(K) \cap L$, i.e., if K is the restriction of L to some subspace. Note that in this case K and L are either identical, or K has smaller rank than L.

Equivalently to (1.1), one can also define a lattice L as a discrete additive subgroup of $\mathbb{R}^n$. Here *discrete* means that there is a real number $r > 0$ such that for any two distinct elements $\boldsymbol{x}, \boldsymbol{y} \in L$ we have $\|\boldsymbol{x} - \boldsymbol{y}\| > r$. An *additive subgroup* of $\mathbb{R}^n$ is a set L with $\boldsymbol{0} \in L$ and if $\boldsymbol{x}, \boldsymbol{y} \in L$, then also $\boldsymbol{x} \pm \boldsymbol{y} \in L$.

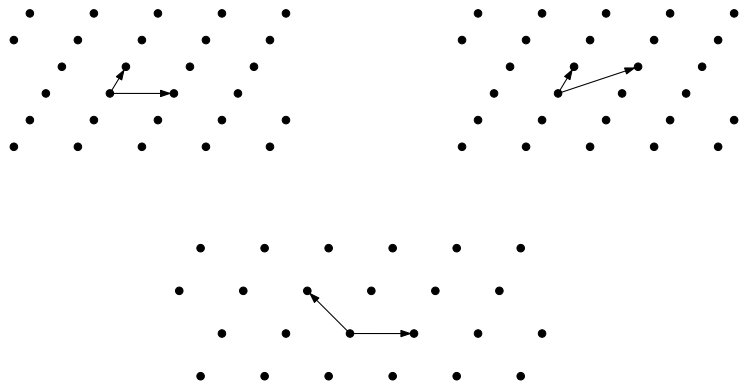

Figure 1.1: Some examples of 2-dimensional lattices with a basis given.

Although the latter is certainly the more elegant definition, we will primarily make use of the former one, as it is in many cases advantageous to have a concrete basis of the lattice at hand.

Given two sets of linearly independent vectors, we might wonder if they will generate the same lattice. To this end, recall that a square integer matrix is called *unimodular*, if it has a determinant of value ± 1. Also note that from this definition it follows that the inverse of a unimodular matrix is again unimodular.

The following operations on a matrix are called *elementary column operations*:

- exchanging two columns,
- multiplying a column by -1,
- adding an integer multiple of one column to another column.

It is well known that applying any finite number of operations of this kind to a matrix $\boldsymbol{M}$ can be expressed by multiplying $\boldsymbol{M}$ by a unimodular matrix.

With this it is easy to check that $\boldsymbol{B}$ and $\hat{\boldsymbol{B}}$ generate the same lattice if and only if we can find a unimodular matrix $\boldsymbol{U}$ with $\boldsymbol{B} = \hat{\boldsymbol{B}}\boldsymbol{U}$.

1.1 Lattice Invariants

The *determinant* $\det(L)$ of a lattice L is the volume of the *fundamental parallelepiped* spanned by a basis $\boldsymbol{B}$ of L. Here we compute the volume with respect to the rank of L, i.e., the determinant is always positive for a lattice with positive rank.

This determinant is invariant under the choice of the basis: Recall that the volume of a parallelepiped spanned by the columns of the matrix $\boldsymbol{B}$ is given by $\sqrt{\det(\boldsymbol{B}^\mathsf{T}\boldsymbol{B})}$, where $\boldsymbol{B}^\mathsf{T}$ denotes the transpose of $\boldsymbol{B}$. Then given two bases of the same lattice we can see that the unimodular matrix we need to transform one into the other will lead to a factor of 1 in the square root.

Another way to see that $\det(L)$ is an invariant under the choice of the basis is to observe that the volume of the parallelepiped is inverse to the density of the lattice: The sparser the lattice, the larger the determinant. More formally we have

$$\det(L) = \lim_{r\to\infty} \frac{\mathrm{vol}(B_r^m)}{|\{\boldsymbol{x} \in L \ : \ \|\boldsymbol{x}\| < r\}|},$$

where $\text{vol}(B_r^n)$ is the volume of the m-dimensional ball of radius r, and m is the rank of L.

Yet another way to compute the determinant comes from the Gram-Schmidt orthogonalization: Let $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$ be a basis of L, and let $\boldsymbol{x}^\mathsf{T}$ denote the transpose of vector $\boldsymbol{x}$. Then we define

$$\begin{aligned} \boldsymbol{b}_1^* &= \boldsymbol{b}_1, \\ \boldsymbol{b}_i^* &= \boldsymbol{b}_i - \sum_{j=1}^{i-1} \mu_{ij} \boldsymbol{b}_j^*, \quad 2 \le i \le m, \quad \text{where} \\ \mu_{ij} &= \frac{\boldsymbol{b}_i^\mathsf{T} \boldsymbol{b}_j^*}{\|\boldsymbol{b}_j^*\|^2}, \quad 1 \le j < i \le m. \end{aligned}$$

Geometrically, $\boldsymbol{b}_i^*$ is the component of $\boldsymbol{b}_i$ orthogonal to $\text{span}\{\boldsymbol{b}_1^*, \ldots, \boldsymbol{b}_{i-1}^*\}$ (and thus to $\text{span}\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{i-1}\}$), while μ_{ij} is the length, relative to the length of $\boldsymbol{b}_j^*$, of the component of $\boldsymbol{b}_i$ in the direction of $\boldsymbol{b}_j^*$.

Note that $\boldsymbol{b}_1^*, \ldots, \boldsymbol{b}_m^*$ are in general not in the lattice spanned by $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$, although they span the same Euclidean space. We also remark that we can express the relationship between a set of linearly independent vectors and their Gram-Schmidt orthogonalization by the equation $\boldsymbol{B} = \boldsymbol{B}^* \boldsymbol{R}$, where $\boldsymbol{B}^* = [\boldsymbol{b}_1^*, \ldots, \boldsymbol{b}_m^*]$ and

$$\boldsymbol{R} = \begin{bmatrix} 1 & \mu_{21} & \cdots & \mu_{m1} \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mu_{mm-1} \\ 0 & \cdots & 0 & 1 \end{bmatrix}.$$

In particular this implies that $\det(\boldsymbol{B}^\mathsf{T}\boldsymbol{B}) = \det(\boldsymbol{B}^{*\mathsf{T}}\boldsymbol{B}^*)$ and we compute

$$\det(L) = \sqrt{\det(\boldsymbol{B}^{*\mathsf{T}}\boldsymbol{B}^*)} = \|\boldsymbol{b}_1^*\| \cdot \ldots \cdot \|\boldsymbol{b}_m^*\|, \tag{1.2}$$

where the last equality holds, because the vectors $\boldsymbol{b}_i^*$, $i = 1, \ldots, m$, are pairwise orthogonal.

For an arbitrary matrix $\boldsymbol{A} \in \mathbb{R}^{n \times m}$ we get *Hadamard's inequality* instead of the above equality:

$$\sqrt{\det(\boldsymbol{A}^\mathsf{T}\boldsymbol{A})} \le \|\boldsymbol{a}_1\| \cdot \ldots \cdot \|\boldsymbol{a}_m\|. \tag{1.3}$$

To every lattice L we can also associate the *dual lattice*

$$L^\dagger = \{\boldsymbol{x} \in \text{span}(L) \mid \boldsymbol{x}^\mathsf{T}\boldsymbol{y} \in \mathbb{Z} \text{ for all } \boldsymbol{y} \in L\}.$$

Notice that $L^{\dagger\dagger} = L$, and that, if L is full-dimensional, the rows of $\boldsymbol{B}^{-1}$ form a basis of $L^\dagger$, where $\boldsymbol{B}$ is any basis of L.

Indeed, any row of $\boldsymbol{B}^{-1}$ is in $L^\dagger$, as $\boldsymbol{B}^{-1}\boldsymbol{B} = \boldsymbol{I}$. Conversely, if $\boldsymbol{x} \in L^\dagger$, then $\boldsymbol{x}^\mathsf{T}\boldsymbol{B}$ is integer and thus $\boldsymbol{x}^\mathsf{T} = (\boldsymbol{x}^\mathsf{T}\boldsymbol{B})\boldsymbol{B}^{-1}$ is an integer combination of the rows of $\boldsymbol{B}^{-1}$. In particular, this also implies that

$$\det(L^\dagger) = \frac{1}{\det(L)}. \tag{1.4}$$

It is not hard to see that Equation (1.4) also holds if L is not full-dimensional.

If K is a sublattice of L, then we define $K^\perp$ to be the sublattice of $L^\dagger$ orthogonal to K, i.e., $K^\perp = \{\boldsymbol{x} \in L^\dagger \mid \boldsymbol{x}^\mathsf{T}\boldsymbol{y} = 0 \text{ for all } \boldsymbol{y} \in K\}$. Let π denote the orthogonal projection of $\mathbb{R}^n$ onto $\operatorname{span}(K^\perp)$, then we define $L/K = \pi(L)$.

Then by construction $K^\perp$ is a pure sublattice of $L^\dagger$, and furthermore we get yet another way of computing $\det(L)$.

Proposition 1.1. *Let K be a sublattice of L, then*

$$K^\perp = (L/K)^\dagger, \tag{1.5}$$

and if K is a pure sublattice of L, then

$$\det(L) = \det(L/K) \cdot \det(K)\,. \tag{1.6}$$

Proof. If $\boldsymbol{x} \in (L/K)^\dagger$, then $\boldsymbol{x} \in \operatorname{span}(K^\perp)$ and thus $\boldsymbol{x}^\mathsf{T}\boldsymbol{y} = 0$ for all $\boldsymbol{y} \in \operatorname{span}(K)$ and in particular for $\boldsymbol{y} \in K$. Given $\boldsymbol{y}' \in L$, we can write $\boldsymbol{y}' = \boldsymbol{y}_1 + \boldsymbol{y}_2$ for some $\boldsymbol{y}_1 \in L/K$ and $\boldsymbol{y}_2 \in \operatorname{span}(K)$, and thus $\boldsymbol{x}^\mathsf{T}\boldsymbol{y}' = \boldsymbol{x}^\mathsf{T}\boldsymbol{y}_1 \in \mathbb{Z}$. Therefore, $\boldsymbol{x} \in L^\dagger$ and thus $\boldsymbol{x} \in K^\perp$.

Conversely, let $\boldsymbol{x} \in K^\perp$ and $\boldsymbol{y} \in L/K$. Then we know that there are vectors $\boldsymbol{y}_1 \in L$ and $\boldsymbol{y}_2$ orthogonal to $K^\perp$, such that $\boldsymbol{y} = \boldsymbol{y}_1 + \boldsymbol{y}_2$. Therefore, since $\boldsymbol{x} \in L^\dagger$, we get $\boldsymbol{x}^\mathsf{T}\boldsymbol{y} = \boldsymbol{x}^\mathsf{T}\boldsymbol{y}_1 \in \mathbb{Z}$, and together with the observation $\operatorname{span}(K^\perp) = \operatorname{span}(L/K)$ we get equation (1.5).

For equation (1.6) we first note that if K is a pure sublattice, then any basis $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k$ of K can be completed to a basis $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k, \boldsymbol{b}_{k+1}, \ldots, \boldsymbol{b}_m$ of L. Let P be the parallelepiped spanned by this basis.

Then the orthogonal projection of $\boldsymbol{b}_{k+1}, \ldots, \boldsymbol{b}_n$ onto the space orthogonal to K (and thus the first k basis vectors) does not change the volume of P.

Furthermore, the projected vectors are linearly independent and lie in L/K, and since the volume of P is the product of $\det(K)$ and the parallelepiped spanned by the projected vectors, they must form a basis of L/K. □

Lemma 1.2. *If K is a pure sublattice of $\mathbb{Z}^n$, then*

$$\det(K) = \det(K^\perp)\,. \tag{1.7}$$

Proof. By combining (1.6), (1.4), and (1.5), and letting $L = \mathbb{Z}^n$, we obtain

$$\det(K) = \frac{\det(L)}{\det(L/K)} = \frac{1}{\det(L/K)} = \det((L/K)^\dagger) = \det(K^\perp)\,.$$

□

Let $\boldsymbol{M} \in \mathbb{R}^{m\times n}$ be a matrix of full row rank. We say that $\boldsymbol{M}$ is in *Hermite Normal Form*, if it has the form $[\boldsymbol{H}, \boldsymbol{0}^{m\times(n-m)}]$, where $\boldsymbol{H}$ is a lower triangular non-negative $m \times m$ matrix in which the unique row maxima can be found along the diagonal.

Every rational $m \times n$ matrix $\boldsymbol{M}$ of full row rank has a unique Hermite normal form, $\operatorname{HNF}(\boldsymbol{M}) = [\boldsymbol{H}, \boldsymbol{0}^{m\times(n-m)}] = \boldsymbol{M}\boldsymbol{U}$, where $\boldsymbol{U}$ is unimodular. We can find $\operatorname{HNF}(\boldsymbol{M})$ by a series of elementary column operations, where constructing the lower triangular matrix $\boldsymbol{H}$ corresponds to the Euclidean algorithm (see Frumkin [37], based on work of von zur Gathen and Sieveking and Votjakov and Frumkin [40, 38], and references therein). Kannan and Bachem [9] also gave a direct polynomial-time method to bring a matrix into Hermite normal form.

To see the uniqueness, consider the lattice generated by the columns of $\boldsymbol{M}$ (and hence also $\boldsymbol{H}$). If there was another Hermite normal form of $\boldsymbol{M}$ with $\boldsymbol{H}' \neq \boldsymbol{H}$, then the first row where they differed (in, say, the j^{th} column) would give us a contradiction to the maximality of the diagonal elements, when looking at the difference of the j^{th} columns (as this difference is also in the lattice).

In this context, let us also define what we mean by the *size of a matrix or vector*. The size of an integer is the length of a reasonable encoding, e.g., its binary representation. The size of a rational number p/q (with p, q relatively prime) is $1 + \text{size}(p) + \text{size}(q)$. Then the size of a rational $n \times m$-matrix $\boldsymbol{A}$ is defined as

$$\text{size}(\boldsymbol{A}) = nm + \sum_{i,j} \text{size}(a_{ij}),$$

and the size of a vector of length n is given the same way, with $m = 1$.

When we said above that there is a polynomial-time algorithm to find the Hermite normal form of a matrix, then this means that the running time (and space-requirement) of the algorithm is polynomially bounded in the size of the input-matrix. In particular, the sizes of $\boldsymbol{H}$ and $\boldsymbol{U}$ are polynomial in size($\boldsymbol{M}$).

Lemma 1.3. *$L(\boldsymbol{B})$ is a pure sublattice of $\mathbb{Z}^n$ if and only if* $\text{HNF}(\boldsymbol{B}^\mathsf{T}) = [\boldsymbol{I}, \boldsymbol{0}]$.

Proof. First note that $L(\boldsymbol{B})$ is a pure sublattice of $\mathbb{Z}^n$ of rank m if and only if $\boldsymbol{B}$ has column-rank m and for every $\boldsymbol{z} \in \mathbb{R}^m$ it holds that

$$\boldsymbol{z}^\mathsf{T}\boldsymbol{B}^\mathsf{T} \text{ integer} \quad \Rightarrow \quad \boldsymbol{z} \text{ integer}.$$

Let $\text{HNF}(\boldsymbol{B}^\mathsf{T}) = [\boldsymbol{H}, \boldsymbol{0}]$. Then the above is equivalent to the condition that for every $\boldsymbol{z} \in \mathbb{R}^m$ it holds that

$$\boldsymbol{z}^\mathsf{T}\boldsymbol{H} \text{ integer} \quad \Rightarrow \quad \boldsymbol{z} \text{ integer}.$$

Clearly, this is satisfied for $\boldsymbol{H} = \boldsymbol{I}$. For $j = m, \ldots, 1$ define recursively $z_j = \frac{1}{h_{jj}} - \sum_{i=j+1}^{m} h_{ij} z_i$. Then it is not difficult to check that $\boldsymbol{z}^\mathsf{T}\boldsymbol{H} = \boldsymbol{1}$, and, by looking at the last column of $\boldsymbol{H}$ we see that h_{mm} must be 1 to ensure that $\boldsymbol{z}$ is integer. As $\boldsymbol{H}$ is a non-negative integer matrix, and its unique row maxima are on the diagonal, the other entries in the last row of $\boldsymbol{H}$ must be zero. We can now extend this argument inductively to the rest of $\boldsymbol{H}$. □

Lemma 1.4. *Let $\boldsymbol{A}$ be a rational $m \times n$-matrix of full row rank, and let* $\text{HNF}(\boldsymbol{A}) = [\boldsymbol{H}, \boldsymbol{0}]$. *Then for any rational vector $\boldsymbol{b}$ we have:*

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \text{ has an integer solution} \quad \Leftrightarrow \quad \boldsymbol{H}^{-1}\boldsymbol{b} \text{ is integer.}$$

Proof. Let $\boldsymbol{x}$ be given with $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, and let $[\boldsymbol{H}, \boldsymbol{0}] = \boldsymbol{A}\boldsymbol{U}$ where $\boldsymbol{U}$ is unimodular. Then we can compute

$$\boldsymbol{H}^{-1}\boldsymbol{b} = \boldsymbol{H}^{-1}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{H}^{-1}[\boldsymbol{H}, \boldsymbol{0}]\boldsymbol{U}^{-1}\boldsymbol{x} = [\boldsymbol{I}, \boldsymbol{0}]\boldsymbol{U}^{-1}\boldsymbol{x}.$$

Since $\boldsymbol{U}$ is unimodular, we know that $\boldsymbol{U}^{-1}$ is integer, and therefore if $\boldsymbol{x} \in \mathbb{Z}^n$, then $\boldsymbol{H}^{-1}\boldsymbol{b} = [\boldsymbol{I}, \boldsymbol{0}]\boldsymbol{U}^{-1}\boldsymbol{x} \in \mathbb{Z}^m$.

Conversely, let $\boldsymbol{H}^{-1}\boldsymbol{b}$ be integer. Then we know that

$$\boldsymbol{x} := \boldsymbol{U}\begin{pmatrix} \boldsymbol{H}^{-1}\boldsymbol{b} \\ \boldsymbol{0} \end{pmatrix}$$

is an integer vector, and furthermore that

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}\boldsymbol{U}\begin{pmatrix}\boldsymbol{H}^{-1}\boldsymbol{b}\\ \boldsymbol{0}\end{pmatrix} = [\boldsymbol{H}, \boldsymbol{0}]\begin{pmatrix}\boldsymbol{H}^{-1}\boldsymbol{b}\\ \boldsymbol{0}\end{pmatrix} = \boldsymbol{b}.$$

□

Note that the proof tells us even more:

Lemma 1.5. *Given $\boldsymbol{A}$ and $\boldsymbol{b}$, there is a polynomial time algorithm that gives us an integer vector $\boldsymbol{x}$ with $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, such that the size of $\boldsymbol{x}$ is polynomially bounded by the size of $\boldsymbol{A}$ and $\boldsymbol{b}$, or concludes that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ has no integer solution.*

Proof. The Hermite normal form of $\boldsymbol{A}$ can be computed in polynomial time and the size of $\boldsymbol{H}^{-1}$ is polynomially bounded in the size of $\boldsymbol{A}$. □

The *successive minima* $\lambda_1, \ldots, \lambda_m$ of a lattice L are the numbers defined as

$$\lambda_i = \inf\{r \in \mathbb{R} \ : \ \dim(\operatorname{span}(L \cap B(0, r))) \geq i\}.$$

Thus, λ_i is the radius of the smallest sphere around the origin containing i linearly independent vectors.

To find an upper bound on λ_1, we will use the following celebrated result of Minkowski.

Theorem 1.6 (Convex Body Theorem [67])**.** *Let K be a compact convex set in $\mathbb{R}^n$ of volume* $\operatorname{vol}(K)$ *that is symmetric about the origin. Let h be an integer and let L be a full-dimensional lattice with determinant* $\det(L)$*. Suppose that* $\operatorname{vol}(K) \geq h2^n \det(L)$*. Then K contains at least h pairs of points $\pm\boldsymbol{x}_j$, $1 \leq j \leq h$ that are distinct from each other and from the origin.*

Let L be a lattice and let S be the ball of radius $\sqrt{m}\det(L)^{1/m}$ in $\operatorname{span}(L)$, where m is the rank of L. Notice that S is indeed a compact convex set and is symmetric about the origin. Furthermore, S contains an m-dimensional hypercube with edges of length $2\det(L)^{1/m}$, and thus $\operatorname{vol}(S) > 2^m \det(L)$. We conclude that there is a ball S' of smaller radius, such that we still have $\operatorname{vol}(S') \geq 2^m \det(L)$. Then by Minkowski's theorem there is a nonzero lattice vector $\boldsymbol{v}$ with $\boldsymbol{v} \in S'$, which implies that

$$\lambda_1 < \sqrt{m}\det(L)^{1/m}. \tag{1.8}$$

A slightly stronger form of this inequality is known as *Minkowski's First Theorem*. Minkowski also proved a result involving all successive minima:

Theorem 1.7 (Minkowski's Second Theorem [67])**.** *For any lattice L of rank m, the successive minima $\lambda_1, \ldots, \lambda_m$ satisfy*

$$\left(\prod_{i=1}^{m} \lambda_i\right)^{1/m} < \sqrt{m}\det(L)^{1/m}.$$

Note that this implies the bound (1.8). We can also get a lower bound on λ_1 in the following way.

Lemma 1.8. *Let $\boldsymbol{B}$ be a basis of L, and let $\boldsymbol{B}^*$ be the corresponding Gram-Schmidt orthogonalization. Then*

$$\lambda_1 \geq \min_j \|\boldsymbol{b}_j^*\| > 0\,. \tag{1.9}$$

Proof. Let $\boldsymbol{Bx}$ be a nonzero lattice vector (i.e., $\boldsymbol{x} \in \mathbb{Z}^m$, $\boldsymbol{x} \neq \boldsymbol{0}$), and let i be the maximal index with $x_i \neq 0$. We will show that $|(\boldsymbol{Bx})^\mathsf{T}\boldsymbol{b}_i^*| \geq \|\boldsymbol{b}_i^*\|^2$. Using $|\boldsymbol{c}^\mathsf{T}\boldsymbol{d}| \leq \|\boldsymbol{c}\| \cdot \|\boldsymbol{d}\|$ for any two vectors $\boldsymbol{c}, \boldsymbol{d}$, we then get $\|\boldsymbol{Bx}\| \geq \|\boldsymbol{b}_i^*\| \geq \min_j \|\boldsymbol{b}_j^*\|$. As this holds for all lattice vectors, it must also be true for the shortest ones.

Now we compute

$$\begin{aligned}
(\boldsymbol{Bx})^\mathsf{T}\boldsymbol{b}_i^* &= \sum_{j=1}^{i} \boldsymbol{b}_j^\mathsf{T}\boldsymbol{b}_i^*\, x_j \\
&= \boldsymbol{b}_i^\mathsf{T}\boldsymbol{b}_i^*\, x_i \\
&= (\boldsymbol{b}_i^* + \sum_{j<i} \mu_{ij}\boldsymbol{b}_j^*)^\mathsf{T}\boldsymbol{b}_i^*\, x_i \\
&= (\boldsymbol{b}_i^*)^\mathsf{T}\boldsymbol{b}_i^*\, x_i + \sum_{j<i} \mu_{ij}(\boldsymbol{b}_j^*)^\mathsf{T}\boldsymbol{b}_i^*\, x_i \\
&= \|\boldsymbol{b}_i^*\|^2 x_i,
\end{aligned}$$

and since x_i is a nonzero integer, the claim follows. □

It is a classical problem for lattices to find λ_1, i.e., the length of a shortest nonzero vector in the lattice. This problem is commonly called the *Shortest Vector Problem*, or SVP. Very much related is the *Closest Vector Problem*, or CVP, in which we are given a point $\boldsymbol{r}$ (not necessarily in the lattice), and are asked to find the lattice vector closest to $\boldsymbol{r}$.

While the bound in (1.8) is asymptotically tight in the sense that there is a $c > 0$ such that for all $n \in \mathbb{N}$ we can find a lattice L of rank n with $\lambda_1(L) > c\sqrt{n}\det(L)^{1/n}$, in general this is not the case.

As an example consider the lattice L generated by $(\varepsilon, 0)^\mathsf{T}$ and $(0, 1/\varepsilon)^\mathsf{T}$, for small positive ε. Then $\det(L) = 1$ and the above bound gives us $\lambda_1(L) \leq \sqrt{2}$, while in fact we have $\lambda_1(L) = \varepsilon$.

Furthermore, the proof of Minkowski's Theorem is not constructive, so we are still lacking a computationally efficient method to even get a lattice vector of the above size. In fact, finding such a vector is a very challenging problem, as a look at the computational complexity of this problem will reveal. This also motivates why it is interesting to look at efficient algorithms for approximations of shortest vectors.

1.1.1 Complexity of some lattice problems

The way we defined SVP above is known as the *optimization version* of the problem. Finding a vector of minimal length is the *search version*, which is at least as difficult as the optimization version. Deciding whether there is a vector of length at most r, where $r > 0$ is given, is the *decision version*, which is at most as difficult as the previous two. Similar versions exist for CVP.

Without going too much into the technicalities (which are plentiful in this field), we will repeat some results on the complexity of these problems. (See the book of Micciancio and Goldwasser [66] for a more thorough treatment.)

In this spirit, we will in this section think of a *problem* as a question that, given some input, can be answered with *yes* or *no*. Given some valid input, a problem becomes an *instance* of the problem. Note that in most of this thesis, we will call problems *programs* when we are most interested in algorithms, and *problems* when we are interested in their complexity or other more structural properties.

A problem for which every instance can be solved by a deterministic Turing machine whose running-time is bounded by a polynomial in the size of the input, is said to belong to the class P. If instead for every yes-instance of the problem we can give a certificate of polynomial size that can be checked in polynomial time, then the problem belongs to the class NP. Clearly $\mathrm{P} \subseteq \mathrm{NP}$ since we can use the algorithm itself as certificate, and it is widely believed that the reverse is false. Even if it turns out that the classes P and NP are in fact equal, to date they provide a very useful tool to estimate how difficult it is for us to develop reasonably fast algorithms.

The most prominent method to show that a problem lies in one of these classes is by *reduction*: If A and B are two problems, then we say that A reduces to B if we can find a polynomial time algorithm which translates any instance of A into an instance of B of polynomial size (in terms of the size of the instance in A). Hence, if $B \in \mathrm{P}$ and A reduces to B, then we can also solve any instance of A in polynomial time. On the other hand, there are problems for which it is known that every problem in NP can be reduced to them. Such problems are called NP-*hard*, and if they lie in NP themselves, they are called NP-*complete*.

Theorem 1.9 ([34])**.** *The decision version of CVP is* NP*-complete.*

Theorem 1.10 ([64, 65])**.** *The decision version of SVP is in* NP*, and it is* NP*-complete under randomized reductions, or under the assumption that a conjecture on the density of square-free integers with bounded prime factors is true.*

A related problem comes from cryptography: A message is sent in the form of a $\{0, 1\}$-vector $\boldsymbol{x} = (x_1, \ldots, x_n)$. This message is encrypted using the public integer weights $a_1, \ldots, a_n$, and the encrypted message is $a_0 = \sum_{i=1}^{n} a_i x_i$. There is a hidden structure, called a trapdoor, between the weights, which is only known to the receiver. The idea is now that the following problem should be easy to solve if the trapdoor is known, but difficult without it:

$$\text{Determine a } \{0,1\}\text{-vector } \boldsymbol{x} \text{ such that } \sum_{i=1}^{n} a_i x_i = a_0. \tag{1.10}$$

Note that (1.10) is the search version of the *knapsack problem*, a problem class which is known to be NP-complete for the decision version and NP-hard for the other ones. Therefore one could expect that it will be difficult to solve. However, the additional information that (1.10) describes an encrypted message is enough to make it easier. To see this, define the *density* of the coefficients a_i, $i = 1, \ldots, n$, as

$$\delta(\boldsymbol{a}) = \frac{n}{\log_2(\max_{1 \le i \le n}\{a_i\})},$$

which is an approximation of the information rate at which bits are transmitted. The interesting case here is $\delta(\boldsymbol{a}) \le 1$, as for $\delta(\boldsymbol{a}) > 1$ the problem (1.10) has in general several solutions, which makes it unsuitable for transmitting a message. Earlier research had already shown that if the $\delta(\boldsymbol{a})$ is high, the trapdoor information is relatively hard to conceal.

Lagarias and Odlyzko [55] proposed a polynomial-time algorithm based on the Lenstra-Lenstra-Lovász basis reduction algorithm (see Chapter 1.2.2) that looks for short vectors in certain lattices. It is not guaranteed that this algorithm will find a solution $\boldsymbol{x}$ of (1.10), but nicely complementing the above results it almost always works for small densities. See also Aardal and Eisenbrand [1] for more details, and Coster, Joux, LaMaccia, Odlyzko, Schnorr, and Stern [30] for some improvements on the results in [55].

1.2 Lattice Bases

We will use the shortest vector problem to investigate which differences occur when we consider different bases of a lattice L.

As before, let $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m \in \mathbb{R}^n$ be a basis of L, and let $\boldsymbol{b}_1^*, \ldots, \boldsymbol{b}_m^*$ be the corresponding Gram-Schmidt vectors. Then

$$c = \frac{\|\boldsymbol{b}_1\| \cdot \ldots \cdot \|\boldsymbol{b}_m\|}{\det(L)}$$

is called the *orthogonality defect* of the basis. Note that by Hadamard's inequality we will always have $c \geq 1$ with equality if and only if the vectors $\boldsymbol{b}_i$ are pairwise orthogonal. We will see that it is desirable to find bases with small orthogonality defect. In the case of the shortest vector problem the reason is the following.

Proposition 1.11. *Let $\boldsymbol{v}$ be a lattice element, expressed in the basis as $\boldsymbol{v} = \sum_{i=1}^{m} \nu_i \boldsymbol{b}_i$. If $\boldsymbol{v}$ is a shortest vector, then $|\nu_i| \leq c$ for all i.*

Note that this implies that we can find the shortest vector by enumerating the $(2\lfloor c \rfloor + 1)^m$ vectors that, when expressed in the basis, have coefficients of magnitude at most c.

Proof. Suppose there is a j with $|\nu_j| > c$. Since c is independent of the order in which the basis vectors are given, we may assume $j = m$. Note that, since $\boldsymbol{b}_i^*$ is an orthogonal projection of $\boldsymbol{b}_i$, we have $\|\boldsymbol{b}_i^*\| \leq \|\boldsymbol{b}_i\|$ for all i. Together with $\|\boldsymbol{b}_1\| \cdot \ldots \cdot \|\boldsymbol{b}_m\| = c \cdot \|\boldsymbol{b}_1^*\| \cdot \ldots \cdot \|\boldsymbol{b}_m^*\|$, we get $\|\boldsymbol{b}_m\| \leq c\|\boldsymbol{b}_m^*\|$. Furthermore,

$$\|\boldsymbol{v}\| = \|\nu_m \boldsymbol{b}_m + \sum_{i=1}^{m-1} \nu_i \boldsymbol{b}_i\| = \|\nu_m \boldsymbol{b}_m^* + \boldsymbol{u}\|,$$

where $\boldsymbol{u}$ is some vector in $\operatorname{span}\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{m-1}\}$. Since $\boldsymbol{b}_m$ is orthogonal to $\boldsymbol{u}$, in conclusion we get

$$\|\boldsymbol{v}\| = |\nu_m| \|\boldsymbol{b}_m^*\| + \|\boldsymbol{u}\| > c\|\boldsymbol{b}_m^*\| \geq \|\boldsymbol{b}_m\|.$$

But this implies that $\boldsymbol{v}$ cannot be a shortest vector. □

We will see in the next sections how to derive such a basis with bounded orthogonality defect in polynomial time.

1.2.1 Reduced Bases

Let L be a lattice of rank m. Given a basis of L, we will use their Gram-Schmidt vectors to check whether the basis has small orthogonality defect. Note that for reduced bases the indices of the basis vectors is important. Thus, if we change the order in which the basis vectors are given, this can also change whether or not a basis is reduced.

For fixed $y \in (\frac{1}{4}, 1)$ we call $\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m\}$ *y-reduced*, if

$$|\mu_{ij}| \leq \frac{1}{2}, \quad \text{for } 1 \leq j < i \leq m, \text{ and} \tag{1.11}$$

$$\|\boldsymbol{b}_i^* + \mu_{i\,i-1}\boldsymbol{b}_{i-1}^*\|^2 \geq y\,\|\boldsymbol{b}_{i-1}^*\|^2, \quad \text{for } 1 < i \leq m\,. \tag{1.12}$$

Condition (1.11) can be satisfied in two ways, depicted in Figure 1.2. Remember that we observed that $|\mu_{ij}|$ is the length of the projection of $\boldsymbol{b}_i$ onto $\boldsymbol{b}_j^*$, relative to the length of $\boldsymbol{b}_j^*$. If $\boldsymbol{b}_i$ is indeed almost orthogonal to $\boldsymbol{b}_j^*$, this projection will be short. However, it is also possible to satisfy (1.11) if $\boldsymbol{b}_i$ is much shorter than $\boldsymbol{b}_j^*$. Then $|\mu_{ij}|$ will be small, even if $\boldsymbol{b}_i$ and $\boldsymbol{b}_j^*$ are almost parallel.

Without Condition (1.12), we could therefore end up with a basis with $\|\boldsymbol{b}_1\| \gg \|\boldsymbol{b}_2\| \gg \cdots \gg \|\boldsymbol{b}_m\|$, and very large orthogonality defect.

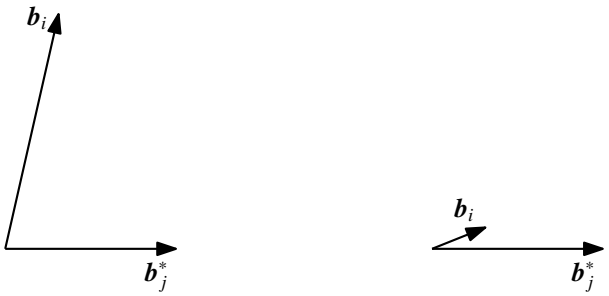

Figure 1.2: Two cases how Condition (1.11) can be satisfied.

Notice now that $\boldsymbol{b}_{i-1}^*$ is the projection of $\boldsymbol{b}_{i-1}$ onto the orthogonal complement of the subspace given by $\text{span}\{\boldsymbol{b}_1^*, \ldots, \boldsymbol{b}_{i-2}^*\}$, and $\boldsymbol{b}_i^* + \mu_{i\,i-1}\boldsymbol{b}_{i-1}^*$ is the projection of $\boldsymbol{b}_i$ onto the same space. If $\boldsymbol{b}_i$ is short compared to $\boldsymbol{b}_{i-1}^*$, then its projection will be even shorter, and (1.12) is violated.

Lemma 1.12. *If $\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m\}$ is y-reduced, then*

$$\|\boldsymbol{b}_1\| \leq \left(\frac{2}{\sqrt{4y-1}}\right)^{m-1} \lambda_1.$$

In particular, if $y = \frac{1}{4} + \left(\frac{3}{4}\right)^{m/(m-1)}$, then $\|\boldsymbol{b}_1\| \leq \left(\frac{2}{\sqrt{3}}\right)^m \lambda_1$.

Proof. As the basis y-reduced, we know that for all $i = 2, \ldots, m$ we have

$$\begin{aligned} y\|\boldsymbol{b}_{i-1}^*\|^2 &\leq \|\boldsymbol{b}_i^* + \mu_{i\,i-1}\boldsymbol{b}_{i-1}^*\|^2 \\ &= \|\boldsymbol{b}_i^*\|^2 + \mu_{i\,i-1}^2\|\boldsymbol{b}_{i-1}^*\|^2 \\ &\leq \|\boldsymbol{b}_i^*\|^2 + \frac{1}{4}\|\boldsymbol{b}_{i-1}^*\|^2. \end{aligned}$$

Rearranging the terms, we then get

$$(y - 1/4)\|\boldsymbol{b}_{i-1}^*\|^2 \leq \|\boldsymbol{b}_i^*\|^2.$$

Note that this implies a lower bound on how much the length of the vector $\boldsymbol{b}_j$ can be shorter than basis vectors of smaller index, for any $2 \leq j \leq m$. This is precisely the effect we expected from Condition (1.12).

Iterating this estimate, we get that for all $j \leq i$ we have

$$\left(y - \frac{1}{4}\right)^{i-j} \|\boldsymbol{b}_j^*\|^2 \leq \|\boldsymbol{b}_i^*\|^2,$$

and thus in particular, for all $i = 1, \ldots, m$,

$$\|\boldsymbol{b}_i^*\| \geq \left(y - \frac{1}{4}\right)^{\frac{i-1}{2}} \|\boldsymbol{b}_1^*\| \geq \left(y - \frac{1}{4}\right)^{\frac{m-1}{2}} \|\boldsymbol{b}_1\|,$$

where for the last inequality we recall that $\boldsymbol{b}_1^* = \boldsymbol{b}_1$ by the definition of the Gram-Schmidt vectors.

Using the lower bound $\lambda_1 \geq \min_i \|\boldsymbol{b}_i^*\|$ from Lemma 1.8, we are done. □

Lemma 1.13. *Let $\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m\}$ be $\frac{3}{4}$-reduced. Then the orthogonality defect is*

$$c \leq 2^{m(m-1)/4}.$$

Proof. Let $\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m\}$ be y-reduced. Then, as in the proof of Lemma 1.12, we estimate

$$\|\boldsymbol{b}_i^*\|^2 \leq \left(\frac{4}{4y-1}\right)^{j-i} \|\boldsymbol{b}_j^*\|^2,$$

for $j \geq i$. Thus, if we set $\mu_{ii} = 1$, we get by the definition of the Gram-Schmidt vectors and the previous considerations

$$\begin{aligned}\|\boldsymbol{b}_j\|^2 &= \sum_{i=1}^{j} \mu_{ij}^2 \|\boldsymbol{b}_i^*\|^2 \leq \|\boldsymbol{b}_j^*\|^2 + \frac{1}{4}\sum_{i=1}^{j-1} \|\boldsymbol{b}_i^*\|^2 \\ &\leq \|\boldsymbol{b}_j^*\|^2 + \frac{1}{4}\sum_{i=1}^{j-1} \left(\frac{4}{4y-1}\right)^{j-i} \|\boldsymbol{b}_j^*\|^2,\end{aligned}$$

and for $y = \frac{3}{4}$ we have $1 + \frac{1}{4}\sum_{i=1}^{j-1}\left(\frac{4}{4y-1}\right)^{j-i} = 1 + \frac{1}{4}\sum_{i=1}^{j-1} 2^{j-i} \leq 2^{j-1}$. The result then follows from the definition of the orthogonality defect. □

1.2.2 The LLL basis reduction algorithm

We are now ready to describe the lattice basis reduction algorithm of Lenstra, Lenstra, and Lovász [57]. We will not give a precise description of every step of the algorithm, but instead outline it to an extend such that the intrigued reader can complete it.

Let $L(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m) \subseteq \mathbb{Z}^n$ be a lattice of rank m. The two main operations, which will be applied repeatedly are

Length-reduction: Set $\boldsymbol{b}_i := \boldsymbol{b}_i - c_{ij}\boldsymbol{b}_j$, where $c_{ij} = \left\lceil \frac{\boldsymbol{b}_i^\mathsf{T}\boldsymbol{b}_j^*}{\|\boldsymbol{b}_j^*\|^2} \right\rfloor = \lceil \mu_{ij} \rfloor$;

and

Swap: Swap the indices of $\boldsymbol{b}_i$ and $\boldsymbol{b}_j$.

The reduction-step is designed to make sure that (1.11) is satisfied, while the swap-step will ensure that (1.12) holds. An important observation is that applying the reduction step with $j < i-1$ will not influence the the validity of (1.12), and no reduction step changes any of the Gram-Schmidt vectors. The LLL-algorithm now works as follows:

For $i = 2, \ldots, m$ and

While Conditions (1.11) and (1.12) are not satisfied

Do reduction-steps;

If index i violates (1.12), do a swap-step on i and $i-1$ and restart.

The reason we restart is that by swapping indices i and $i-1$, we influence all μ_{kj} with $k \in \{i, i-1\}$ and $j < k$. Note that for $j < i-1$ Condition (1.12) will now not be violated. However, it might be for $i-1$.

At this point it might not even be clear that this algorithm always terminates, let alone in polynomial time. The key ingredient to see that we do not have too many swap-steps is the *potential function*

$$\Phi(\boldsymbol{B}) = \|\boldsymbol{b}_1^*\|^{2m} \|\boldsymbol{b}_2^*\|^{2(m-1)} \cdot \ldots \cdot \|\boldsymbol{b}_m^*\|^2.$$

Notice that if we set $L_k = L(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k)$ for all $k = 1, \ldots, m$, then

$$\Phi(\boldsymbol{B}) = \prod_{k=1}^{m} \det(L_k)^2,$$

and thus $\Phi(\boldsymbol{B})$ is a positive integer, because $\det(K)^2$ also is, for any $K \subseteq \mathbb{Z}^n$.

Suppose now that we swap $\boldsymbol{b}_i$ and $\boldsymbol{b}_{i-1}$, because Condition (1.12) is violated. Let $\boldsymbol{B}'$ be the basis with the changed order. Note that L_k is unchanged for all $k \neq i-1$. Then let L'_{i-1} be the lattice corresponding to the new order and we compute

$$\begin{aligned} \frac{\Phi(\boldsymbol{B}')}{\Phi(\boldsymbol{B})} &= \frac{\det(L'_{i-1})^2}{\det(L_{i-1})^2} = \frac{(\prod_{j=1}^{i-2} \|\boldsymbol{b}_j^*\|^2) \cdot \|\boldsymbol{b}_i^* + \mu_{i\,i-1}\boldsymbol{b}_{i-1}^*\|^2}{\prod_{j=1}^{i-1} \|\boldsymbol{b}_j^*\|^2} \\ &= \frac{\|\boldsymbol{b}_i^* + \mu_{i\,i-1}\boldsymbol{b}_{i-1}^*\|^2}{\|\boldsymbol{b}_{i-1}^*\|^2} \\ &< y\,, \end{aligned}$$

where the second equality holds because, as we have observed before, $\boldsymbol{b}_i^* + \mu_{i\,i-1}\boldsymbol{b}_{i-1}^*$ is the projection of $\boldsymbol{b}_i$ onto the space which is orthogonal to $\text{span}\{\boldsymbol{b}_1^*, \ldots, \boldsymbol{b}_{i-2}^*\}$, which is precisely the definition of how to construct the $(i-1)^{st}$ Gram-Schmidt vector. The last inequality holds because Condition (1.12) was violated by $\boldsymbol{b}_i$ and $\boldsymbol{b}_{i-1}$.

Therefore, $\Phi(\boldsymbol{B})$ decreases at least by a factor of y at every swap step. Notice that we needed to assume $y < 1$ to make this argument work.

Using the bound $\Phi(\boldsymbol{B}) \leq (\|\boldsymbol{b}_1\| \cdot \ldots \cdot \|\boldsymbol{b}_m\|)^{2m}$, and taking care of the amount of operations needed for the reduction steps, one can now show that the number of iterations of the algorithm is bounded by $O(m \log(\|\boldsymbol{b}_1\| + \ldots + \|\boldsymbol{b}_m\|))$, and the running time is $O(m^6 (\log \beta)^3)$, where $\beta \geq 2$ is such that $\|\boldsymbol{b}_i\|^2 \leq \beta$ for $1 \leq i \leq m$.

For more detailed accounts of the running time see, e.g., [57, 22, 39]. Akhavi showed that the running time is linear even when $y = 1$ if we fix the dimension [7], but for general dimension no polynomial upper bound on the running time seems to be known.

1.3 Linear and Integer Optimization

Here and throughout this thesis, let $\boldsymbol{A}$ be a $m \times n$-matrix, and let $\boldsymbol{b}$ be a vector of length m. In most cases, we will assume that both are integer. While many statements will remain true also for rational or even irrational input, this is not always the case, and since we are particularly interested in algorithmic results, we will pass up on always achieving the most general statement. Also note that, given rational $\boldsymbol{A}$ and $\boldsymbol{b}$, multiplying with the least common multiple of all denominators will not change the input size, in the sense of polynomiality.

The classical *(Linear) Integer Optimization Program* is

$$\max\left\{\boldsymbol{c}^\mathsf{T}\boldsymbol{x} \;:\; \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \boldsymbol{x} \geq 0, \boldsymbol{x} \in \mathbb{Z}^n\right\}, \tag{IP}$$

where $\boldsymbol{c} \in \mathbb{Q}^n$.

If we allow some of the variables to be continuous, we get a *(Linear) Mixed Integer Optimization Program*

$$\max\left\{\boldsymbol{c}^\mathsf{T}\boldsymbol{x} \;:\; \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \boldsymbol{x} \geq 0, x_i \in \mathbb{Z} \text{ for } i \in I\right\}, \tag{MIP}$$

where again $\boldsymbol{c} \in \mathbb{Q}^n$, and $I \subseteq \{1, \ldots, n\}$.

Not surprisingly, if we drop the integrality requirements completely, we get a *(Linear) Optimization Program*

$$\max\left\{\boldsymbol{c}^\mathsf{T}\boldsymbol{x} \;:\; \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \boldsymbol{x} \geq 0\right\}. \tag{LP}$$

The way the above programs are given is called the *canonical form*. In particular for (IP) and (LP) we will frequently use the *standard form* instead:

$$\max\left\{\boldsymbol{c}^\mathsf{T}\boldsymbol{x} \;:\; \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \geq 0\right\}, \tag{s-OP}$$

where $\boldsymbol{b} \geq \boldsymbol{0}$, and for (IP) we of course additionally require $\boldsymbol{x} \in \mathbb{Z}^n$.

Note that also more general problems can be written in these forms. For example, if there is an unconstrained variable z in the program, we can replace it by $z = z^+ - z^-$ with $z^+, z^- \geq 0$.

Any equality-constraint $\sum_{j=1}^n a_{ij}x_j = b_i$ can be replaced by the two inequalities

$$\sum_{j=1}^n a_{ij}x_j \leq b_i \quad \text{and} \quad \sum_{j=1}^n (-a_{ij})x_j \leq -b_i,$$

and conversely, given an inequality $\sum_{j=1}^n a_{ij}x_j \leq b_i$, we can introduce an additional variable x_{n+i} and then write

$$\sum_{j=1}^n a_{ij}x_j + x_{n+i} = b_i, \quad x_{n+i} \geq 0.$$

Such a variable x_{n+i} is called a *slack variable*. If $b_i < 0$, additionally multiply both sides of the equation by -1. We will give a more formal definition of these correspondences at the end of the next subsection.

The vector $\boldsymbol{c}$ is often called the *objective function*, and the value $\boldsymbol{c}^\mathsf{T}\boldsymbol{x}$ for $\boldsymbol{x} \in \mathbb{R}^n$ is the *objective value*.

A vector $\boldsymbol{x}$ is called *feasible* for a program, if $\boldsymbol{x}$ satisfies all properties defining the set we want to optimize over. This set is also called the *feasible set* for the program.

1.3.1 Polytopes and Polyhedra

As we will frequently take a rather geometric viewpoint on lattice problems, let us review some of the basic concepts here. See, e.g., the book of Ziegler [85] for a very thorough treatment of this topic.

Given a system of inequalities $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$, we can view this as an intersection of finitely many closed halfspaces. We will denote this intersection as P, i.e.,

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

Such a set is a *polyhedron*, and if it is also bounded, we call it a *polytope*. It is easy to see that polyhedra, when defined this way, are always convex.

Given a set $K \subseteq \mathbb{R}^n$, the *convex hull* $\mathrm{conv}(K)$ of K is the intersection of all convex sets containing K.

It is a celebrated result due to Minkowski that any polytope can also be described as the convex hull of a finite set of points. However, the way a polytope is described can make a big difference algorithmically (see below).

A similar result can be formulated for polyhedra in general. To this end, define a *cone* as a subset of $\mathbb{R}^n$ that with any finite set of vectors also contains any non-negative linear combination of them. Then the *conical hull* of a set Y is the intersection of all cones that contain Y, and one can show that any polyhedron is the Minkowski sum of the convex hull of a finite set of points and the conical hull of a finite set of vectors.

Let $P \subseteq \mathbb{R}^n$ be a polyhedron. A linear inequality $\boldsymbol{c}^\mathsf{T}\boldsymbol{x} \leq c_0$ is *valid* for P if it is satisfied for all points $\boldsymbol{x} \in P$. A *face* of P is any set of the form

$$F = P \cap \left\{\boldsymbol{x} \in \mathbb{R}^n \ : \ \boldsymbol{c}^\mathsf{T}\boldsymbol{x} = c_0\right\},$$

where $\boldsymbol{c}^\mathsf{T}\boldsymbol{x} \leq c_0$ is a valid inequality for P. The *dimension* of a face F is the dimension of the smallest affine subspace it is contained in. Note that $\boldsymbol{0}^\mathsf{T}\boldsymbol{x} = 0$ is a valid inequality for any P and thus P itself is a face of P.

We will in particular be interested in faces of dimension 0, which are called *vertices*, and of dimension one less than the dimension of P, called *facets*. For linear optimization (see next subsection), we will also be interested in faces of dimension 1, called *edges*.

Note that if $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ are all of the vertices of P, then $P = \mathrm{conv}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$. On the other hand, if the inequalities $\boldsymbol{c}_1^\mathsf{T}\boldsymbol{x} \leq c_0^1, \ldots, \boldsymbol{c}_h^\mathsf{T}\boldsymbol{x} \leq c_0^h$ are valid for P and each facet of P is given by setting one of these inequalities to equality, then $P = \bigcap_{i=1}^h \boldsymbol{c}_i^\mathsf{T}\boldsymbol{x} \leq c_0^i$. Inequalities or this form are called *facet-defining*. Note that the second characterization can also be used to describe unbounded polyhedra.

Without any additional knowledge about a polytope P, we do not know whether describing it by its facets or by its vertices will be shorter. For example, the n-cube $C_n = \{\boldsymbol{x} \in \mathbb{R}^n \ : \ -1 \leq x_i \leq 1\}$ has 2^n vertices and $2n$ facets, while its dual (or polar), the crosspolytope $C_n^\Delta = \{\boldsymbol{x} \in \mathbb{R}^n \ : \ \sum_i |x_i| \leq 1\}$, has $2n$ vertices and 2^n facets.

Another observation we can derive from the above definitions is that if we have a point $\boldsymbol{x}$ that lies in a face of dimension k of a polytope P, then we can express $\boldsymbol{x}$ as convex combination of vertices of P, such that these vertices lie in an affine subspace of dimension at most k.

Let a polyhedron P be given, together with two inequalities $\boldsymbol{c}_1^\mathsf{T}\boldsymbol{x} \leq c_0^1$ and $\boldsymbol{c}_2^\mathsf{T}\boldsymbol{x} \leq c_0^2$. Then we say that $\boldsymbol{c}_1^\mathsf{T}\boldsymbol{x} \leq c_0^1$ *dominates* $\boldsymbol{c}_2^\mathsf{T}\boldsymbol{x} \leq c_0^2$ (with respect to P), if we can write $\boldsymbol{c}_2^\mathsf{T}\boldsymbol{x} \leq c_0^2$ as a *non-negative* linear combination of $\boldsymbol{c}_1^\mathsf{T}\boldsymbol{x} \leq c_0^1$ and the facet-defining inequalities of P. Geometrically, this means that $\boldsymbol{c}_2^\mathsf{T}\boldsymbol{x} \leq c_0^2$ is valid for $P \cap \left\{\boldsymbol{x} \in \mathbb{R}^n \;:\; \boldsymbol{c}_1^\mathsf{T}\boldsymbol{x} \leq c_0^1\right\}$.

Proposition 1.14 (See [76], Theorem 5.7)**.** *Let $P = \{\boldsymbol{x} \in \mathbb{R}^n \;:\; \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}\}$ be a polyhedron and $\boldsymbol{z} \in P$. Then $\boldsymbol{z}$ is a vertex of P if and only if there are n linearly independent rows $\boldsymbol{a}_i$ of $\boldsymbol{A}$ with $\boldsymbol{a}_i\boldsymbol{z} = b_i$.*

Proof. Let $\boldsymbol{z} \in P$ and let $\boldsymbol{A}_z$ be the matrix consisting of all rows $\boldsymbol{a}_i$ of $\boldsymbol{A}$ with $\boldsymbol{a}_i\boldsymbol{z} = b_i$. We will now show that $\boldsymbol{z}$ is a vertex of P if and only if $\operatorname{rank}(\boldsymbol{A}_z) = n$.

Suppose first that $\operatorname{rank}(\boldsymbol{A}_z) < n$. Then there exists a $\boldsymbol{c} \neq \boldsymbol{0}$ with $\boldsymbol{A}_z\boldsymbol{c} = \boldsymbol{0}$. Since $\boldsymbol{a}_j\boldsymbol{z} < b_j$ for any row of $\boldsymbol{A}$ not in $\boldsymbol{A}_z$, we find a $\delta > 0$ such that

$$\boldsymbol{a}_j(\boldsymbol{z} + \delta\boldsymbol{c}) \leq b_j \quad \text{and} \quad \boldsymbol{a}_j(\boldsymbol{z} - \delta\boldsymbol{c}) \leq b_j.$$

Since $\boldsymbol{A}_z\boldsymbol{c} = \boldsymbol{0}$ and $\boldsymbol{A}\boldsymbol{z} \leq \boldsymbol{b}$ it follows that

$$\boldsymbol{A}(\boldsymbol{z} + \delta\boldsymbol{c}) \leq \boldsymbol{b} \quad \text{and} \quad \boldsymbol{A}(\boldsymbol{z} - \delta\boldsymbol{c}) \leq \boldsymbol{b}.$$

Hence, $\boldsymbol{z} + \delta\boldsymbol{c}$ and $\boldsymbol{z} - \delta\boldsymbol{c}$ belong to P, and as $\boldsymbol{z}$ is a convex combination of them, $\boldsymbol{z}$ lies in a face of dimension ≥ 1 of P and is therefore not a vertex.

Conversely, suppose $\boldsymbol{z}$ is not a vertex. Then there exist points $\boldsymbol{x}, \boldsymbol{y} \in P$ such that $\boldsymbol{x} \neq \boldsymbol{z} \neq \boldsymbol{y}$ and $\boldsymbol{z} = \frac{1}{2}(\boldsymbol{x} + \boldsymbol{y})$. By construction, for every row $\boldsymbol{a}_i$ of $\boldsymbol{A}_z$ we have $\boldsymbol{a}_i\boldsymbol{z} = b_i$ and $\boldsymbol{a}_i\boldsymbol{v} \leq b_i$ for any $\boldsymbol{v} \in P$. Hence we get $\boldsymbol{a}_i(\boldsymbol{x} - \boldsymbol{z}) \leq 0$ and $\boldsymbol{a}_i(\boldsymbol{y} - \boldsymbol{z}) \leq 0$. As $\boldsymbol{y} - \boldsymbol{z} = -(\boldsymbol{x} - \boldsymbol{z})$, this implies $\boldsymbol{a}_i(\boldsymbol{x} - \boldsymbol{z}) = 0$ and thus $\boldsymbol{A}_z(\boldsymbol{x} - \boldsymbol{z}) = \boldsymbol{0}$. Since $\boldsymbol{x} - \boldsymbol{z} \neq \boldsymbol{0}$, we have shown that $\operatorname{rank}(\boldsymbol{A}_z) < n$. □

We will next give an explicit algebraic link between feasible points for the canonical form and feasible points for the standard form of a given optimization program.

Let $\boldsymbol{A}$ and $\boldsymbol{b}$ be given, and let $P = \{\boldsymbol{x} \in \mathbb{R}^n \;:\; \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \geq \boldsymbol{0}\}$. Assume for a moment that the last m columns of $\boldsymbol{A}$ are linearly independent. We will see below that this assumption on the rank of $\boldsymbol{A}$ is no restriction for our considerations. Also, let $N = \{1, \ldots, n-m\}$ and $B = \{n-m+1, \ldots, n\}$.

If we define $\boldsymbol{B}$ as the matrix consisting of the columns of $\boldsymbol{A}$ with indices in B, then we can reformulate $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ as $\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{B}^{-1}\boldsymbol{b}$, and define

$$\bar{\boldsymbol{A}} = \boldsymbol{B}^{-1}\boldsymbol{A} \quad \text{and} \quad \bar{\boldsymbol{b}} = \boldsymbol{B}^{-1}\boldsymbol{b}. \tag{1.13}$$

This leads to the equations

$$x_i = \bar{b}_i - \sum_{j \in N} \bar{a}_{ij} x_j, \quad i \in B, \tag{1.14}$$

and thus the fact that $\boldsymbol{x}$ is in P can equivalently be expressed by

$$\begin{aligned} \bar{b}_i - \sum_{j \in N} \bar{a}_{ij} x_j \geq 0 \qquad & i \in B \\ x_j \geq 0 \qquad & j \in N. \end{aligned} \tag{1.15}$$

Observe that (1.15) defines a polyhedron $\hat{P} \subseteq \mathbb{R}^{n-m}$, and if P is bounded, $\hat{P}$ is a polytope.

Conversely, let $\hat{P} \subseteq \mathbb{R}^{n-m}$ be a polytope given by n halfspaces, where the first $n-m$ are of the form

$$x_i \geq 0, \quad i = 1, \ldots, n-m,$$

and the remaining inequalities are given as

$$h_{i1}x_1 + \cdots + h_{i,n-m}x_{n-m} \leq g_i, \quad i = 1, \ldots, m.$$

Then we can introduce m slack variables $x_{n-m+1}, \ldots, x_n$ to obtain the equivalent system

$$\begin{aligned} \bar{\boldsymbol{A}}\boldsymbol{x} &= \bar{\boldsymbol{b}} \\ \boldsymbol{x} &\geq \boldsymbol{0} \end{aligned}$$

where $\bar{\boldsymbol{A}} = [\boldsymbol{H}, \boldsymbol{I}]$ and $\bar{\boldsymbol{b}} = (g_1, \ldots, g_m)^\mathsf{T}$.

Thus, any polytope $\hat{P} \subseteq \mathbb{R}^{n-m}$ in the positive orthant can be viewed as the feasible region P of a program (s-OP). Furthermore, any point $\hat{\boldsymbol{x}} \in \hat{P}$ can be transformed to $\boldsymbol{x} \in P$ by defining $x_j = \hat{x}_j$ for $j \in N$ and

$$x_i = g_i - \sum_{j \in N} h_{ij}\hat{x}_j, \quad i \in B,$$

and any $\boldsymbol{x} \in P$ can be transformed to $\hat{\boldsymbol{x}} \in \hat{P}$ by deleting the m coordinates with indices in B from $\boldsymbol{x}$.

We say that the points $\boldsymbol{x} \in P$ and $\bar{\boldsymbol{x}} \in \bar{P}$ are *corresponding* to each other, if we can obtain them from each other in the above manner.

1.3.2 Linear Optimization

We will describe how to solve a linear optimization program with a fair amount of details, since it will turn out that the method given below can be used in integer optimization programs as well, as we will explain later on.

Given (LP) in the standard form (s-OP), let P be the set of feasible points, where we assume that $\boldsymbol{A} \in \mathbb{Z}^{m \times n}$ with $\operatorname{rank}(\boldsymbol{A}) = m$.

We will now describe the *simplex method* (introduced by Dantzig [32]) in tableau form, which is one of the most influential algorithms in optimization. While in its most popular forms it was proven to not be a polynomial-time algorithm by Klee and Minty [53], in practice it is very fast and one can usually expect linear runtime (in the problem dimension). We will follow in large parts the exposition in the book of Papadimitriou and Steiglitz [71].

Generally speaking, the algorithm will turn out to work as follows: We start with some vertex $\boldsymbol{v}$ of P, and if it is not optimal (i.e., the maximum is not achieved there), we go to a vertex $\boldsymbol{v}'$ sharing an edge with $\boldsymbol{v}$, such that $\boldsymbol{c}^\mathsf{T}\boldsymbol{v}' > \boldsymbol{c}^\mathsf{T}\boldsymbol{v}$. This gives us a path along the edges of P that leads to an optimal vertex.

We will see in Theorem 1.17 how to determine whether or not a vertex is optimal. Furthermore, the way the algorithm is set up does not immediately reveal this view of things. However, it might be useful to keep it in mind, as the precise description can be quite technical at times.

A *basis* of $\boldsymbol{A}$ is a set of m linearly independent columns $\boldsymbol{a}_{i_1}, \ldots, \boldsymbol{a}_{i_m}$ of $\boldsymbol{A}$, often expressed as $\boldsymbol{B} = [\boldsymbol{a}_{i_1}, \ldots, \boldsymbol{a}_{i_m}]$. The *basic solution* corresponding to $\boldsymbol{B}$ is the vector $\boldsymbol{x} \in \mathbb{R}^n$ with

$$\begin{aligned} x_j &= 0 \quad \text{if } \boldsymbol{a}_j \notin \boldsymbol{B}, \\ x_{i_k} &= \text{the } k^{th} \text{ component of } \boldsymbol{B}^{-1}\boldsymbol{b}, \quad k = 1, \ldots, m. \end{aligned}$$

Observe that, given $\boldsymbol{B}$, we can find the basic solution by setting all coordinates corresponding to columns not in $\boldsymbol{B}$ to zero and solving the m remaining equations to determine the remaining coordinates of $\boldsymbol{x}$. The latter coordinates are called the *basic variables* corresponding to $\boldsymbol{B}$, and we set $B = \{i_1, \ldots, i_m\}$ and $N = \{1, \ldots, n\} \backslash B$.

If a basic solution $\boldsymbol{x}$ is in P, then $\boldsymbol{x}$ is a *basic feasible solution*.

Lemma 1.15. *Let $\boldsymbol{x}$ be a basic feasible solution corresponding to the basis $\boldsymbol{B}$. Then there exists a vector $\boldsymbol{c}$ such that $\boldsymbol{x}$ is the unique optimal solution of* (s-OP).

Proof. Consider the vector $\boldsymbol{c}$ given by

$$c_j = \begin{cases} 0 & \text{if } \boldsymbol{a}_j \in \boldsymbol{B} \\ -1 & \text{if } \boldsymbol{a}_j \notin \boldsymbol{B} \end{cases}$$

where $j = 1, \ldots, n$. Then $\boldsymbol{c}^\mathsf{T}\boldsymbol{x} = 0$ and clearly $\boldsymbol{x}$ is optimal, since any feasible solution of (s-OP) is non-negative.

Moreover, any $\hat{\boldsymbol{x}} \geq 0$ with $\boldsymbol{c}^\mathsf{T}\hat{\boldsymbol{x}} = 0$ must be zero in all non-basic components. Since $\boldsymbol{B}$ determines the feasible solution $\boldsymbol{x}$ uniquely, we must have $\hat{\boldsymbol{x}} = \boldsymbol{x}$. □

If we assume that P is nonempty, one can show that there is at least one basic feasible solution. Furthermore, if there is an upper bound for (s-OP), we can assume that P is bounded. (See [71], Theorem 2.1 and 2.2)

If we go back to the description of how to translate a linear program back and forth between standard and canonical form, observe that indeed we chose a specific basis to obtain $\hat{P} \subseteq \mathbb{R}^{n-m}$. Therefore, we can now be more specific and, given a basis $\boldsymbol{B}$, obtain the equivalent sets

$$P = \{\boldsymbol{x} \in \mathbb{R}^n \ : \ \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \geq \boldsymbol{0}\} \text{ and } P_{\boldsymbol{B}} = \left\{\boldsymbol{y} \in \mathbb{R}^d \ : \ \bar{\boldsymbol{A}}\boldsymbol{y} \leq \bar{\boldsymbol{b}}, \boldsymbol{y} \geq \boldsymbol{0}\right\}, \tag{1.16}$$

where $d = n - m$, and $\bar{\boldsymbol{A}}$ and $\bar{\boldsymbol{b}}$ are defined as in (1.13) (where we delete the basic columns). Note that we can reconstruct P from $P_{\boldsymbol{B}}$ as is described in (1.14), up to reordering of the indices. On the other hand, if $\boldsymbol{B}$ and $\boldsymbol{B}'$ are different bases, then $P_{\boldsymbol{B}}$ and $P_{\boldsymbol{B}'}$ are in general geometrically not the same polytope, but combinatorially they are.

Indeed, we can interpret going from P to $P_{\boldsymbol{B}}$ as setting all basic variables to zero, which is nothing else than projecting orthogonally onto the linear subspace defined by $x_i = 0$ for $i \in B$. Since P was bounded, and the only inequalities are the nonnegativity of the variables, it is not hard to see that the dimension of P and $P_{\boldsymbol{B}}$ must be the same. Similar arguments apply for all faces.

However, there are certain things that do change. For example, it could easily happen that P does not contain integer points, while $P_{\boldsymbol{B}}$ does. Thus, although this reformulation is very useful for linear programs, it is less so for integer programs.

Proposition 1.16 (See, e.g., [71], Theorem 2.4)**.** *Let $\boldsymbol{x} \in P$. Then $\boldsymbol{x}$ is a basic feasible solution of P corresponding to $\boldsymbol{B}$ if and only if the corresponding point $\hat{\boldsymbol{x}} \in P_{\boldsymbol{B}}$ is a vertex of $P_{\boldsymbol{B}}$.*

Let $\boldsymbol{x}$ be a basic feasible solution. If at least one basic variable has value zero, i.e., $\boldsymbol{x}$ contains more than $n - m$ zeros, then $\boldsymbol{x}$ is called *degenerate*.

Theorem 1.17 (See, e.g., [71])**.** *If two distinct bases $\boldsymbol{B}$ and $\boldsymbol{B}'$ correspond to the same basic feasible solution $\boldsymbol{x}$, then $\boldsymbol{x}$ is degenerate.*

Proof. Suppose $\boldsymbol{B}$ and $\boldsymbol{B}'$ both determine the basic feasible solution $\boldsymbol{x}$. Then $\boldsymbol{x}$ has zeros in the $n - m$ coordinates corresponding to columns not in $\boldsymbol{B}$. But there is at least one column in $\boldsymbol{B}$ that is not in $\boldsymbol{B}'$, and $\boldsymbol{x}$ must also be zero in that coordinate. □

Together with the observation that any linear function has its maximum over any polytope $\hat{P}$ in a vertex (using the representation of $\hat{P}$ as the convex hull of the vertices), we have thus shown that we can solve (s-OP) by finding a maximal basic feasible solution in P, and there are only finitely many of them. What is still missing is the rule how to get from one basis to a potentially better one.

Let $\boldsymbol{x}_0$ be a basic feasible solution of (s-OP) corresponding to the basis $\boldsymbol{B}$ given as $\boldsymbol{B} = [\boldsymbol{a}_{B(1)}, \dots, \boldsymbol{a}_{B(m)}]$. If the basic variables of $\boldsymbol{x}_0$ are given by $x_{1,0}, \dots, x_{m,0}$, then

$$\sum_{i=1}^{m} x_{i,0}\boldsymbol{a}_{B(i)} = \boldsymbol{b}, \quad \text{where} \quad x_{i,0} \geq 0. \tag{1.17}$$

As the columns of $\boldsymbol{B}$ are linearly independent, any nonbasic column $\boldsymbol{a}_j \notin \boldsymbol{B}$ can be written as a linear combination of the columns of $\boldsymbol{B}$:

$$\sum_{i=1}^{m} \bar{a}_{i,j}\boldsymbol{a}_{B(i)} = \boldsymbol{a}_j. \tag{1.18}$$

Now we can multiply (1.18) by a scalar $\theta > 0$ and subtract it from (1.17). This leads to

$$\sum_{i=1}^{m} (x_{i,0} - \theta\bar{a}_{i,j})\boldsymbol{a}_{B(i)} + \theta\boldsymbol{a}_j = \boldsymbol{b}. \tag{1.19}$$

If $\boldsymbol{x}_0$ is non-degenerate, let $\theta_0 = \min_{i \leq m} \left\{ \frac{x_{i,0}}{\bar{a}_{i,j}} \;:\; \bar{a}_{i,j} > 0 \right\}$. Then we see that for $\theta = \theta_0$, Equation (1.19) is of the form (1.17) for a different basis, and we have found a new basic feasible solution. If $\boldsymbol{x}_0$ is degenerate because some $x_{i,0} = 0$, but the corresponding $\bar{a}_{i,j}$ is positive, then we still replace $\boldsymbol{a}_{B(i)}$ by $\boldsymbol{a}_j$. We are then still at the same vertex, but we represent it by a different basis. Also note that if all $\bar{a}_{i,j}$, $i = 1, \dots, m$, were non-positive, P would be unbounded.

This method of moving from one basic feasible solution to another is called *pivoting*.

What we have yet to describe is how we choose the column we want to pivot on. To this end observe that the objective value $\boldsymbol{c}^\mathsf{T}\boldsymbol{x}_0$ of a basic feasible solution $\boldsymbol{x}_0$ is given by

$$z = \sum_{i=1}^{m} x_{i,0}c_{B(i)}.$$

If we now write

$$\boldsymbol{a}_j = \sum_{i=1}^{m} \bar{a}_{i,j} \boldsymbol{a}_{B(i)}$$

as in (1.18), then we can interpret this equality as meaning that if we want to increase the variable x_j by one unit, then, to keep the vector feasible, we need to reduce each of the variables $x_{B(i)}$ by the amount $\bar{a}_{i,j}$.

Thus, a unit increase in x_j results in a change in the objective of

$$\bar{c}_j = c_j - \sum_{i=1}^{m} \bar{a}_{i,j} c_{B(i)}, \tag{1.20}$$

the *relative cost* of column j. As we are maximizing $\boldsymbol{c}^\mathsf{T}\boldsymbol{x}$, it is profitable to pivot on column j exactly if $\bar{c}_j > 0$.

The *simplex tableau* now works as follows: Given the original equality constraints $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, we represent them in the form

	$x_1, \ldots, x_n$
z	$\bar{\boldsymbol{c}}$
$\boldsymbol{b}$	$\boldsymbol{A}$

where z and $\bar{\boldsymbol{c}}$ are defined as before.

Now we multiply rows in the lower part by nonzero constants and add these rows to other rows, until the columns $a_{B(i)}$ are transformed into the i^{th} unit vector. Note that these operations do not change the information in the equations. We also update the values of $\bar{\boldsymbol{c}}$ and z.

Observe that the lower part of the left-most column will now contain the values of the basic variables $x_{B(i)} = x_{i,0}$, $i = 1, \ldots, m$. Furthermore, if x_j is a nonbasic variable, then the j^{th} column contains precisely the numbers $\bar{a}_{i,j}$, $i = 1, \ldots, m$.

Therefore, if we want to pivot on column j, everything we described before can directly be read from the corresponding row in the tableau, and we can find the new tableau by the appropriate row operations again.

Also note that, since we carried out the row-operations to get unit vectors in the basic columns, the part of the tableau replacing the matrix $\boldsymbol{A}$ from the initial tableau is nothing else but $\boldsymbol{B}^{-1}\boldsymbol{A}$.

Theorem 1.18 (Optimality Criterion (see, e.g., [33]))**.** *Let $\boldsymbol{x}_0$ be a basic feasible solution, and let $\bar{\boldsymbol{c}}$ be defined as in* (1.20)*. If*

$$\bar{\boldsymbol{c}} \leq \boldsymbol{0}$$

then $\boldsymbol{x}_0$ is optimal.

Proof. Let $\boldsymbol{y} \geq \boldsymbol{0}$ be any feasible vector, not necessarily basic. Define $z_j = \sum_{i=1}^{m} \bar{a}_{i,j} c_{B(i)}$ for $j = 1, \ldots, n$, and $\boldsymbol{z} = (z_1, \ldots, z_n)^\mathsf{T}$. Let $\boldsymbol{c}_B = (c_{B(1)}, \ldots, c_{B(m)})^\mathsf{T}$ be the vector of components of $\boldsymbol{c}$ corresponding to basic variables. Then if $\bar{\boldsymbol{A}}$ denotes the current matrix replacing $\boldsymbol{A}$ in the tableau, we have by definition $\boldsymbol{z}^\mathsf{T} = \boldsymbol{c}_B\bar{\boldsymbol{A}} = \boldsymbol{c}_B\boldsymbol{B}^{-1}\boldsymbol{A}$ and thus

$$\boldsymbol{c}^\mathsf{T}\boldsymbol{y} = (\bar{\boldsymbol{c}} + \boldsymbol{z})^\mathsf{T}\boldsymbol{y} \leq \boldsymbol{z}^\mathsf{T}\boldsymbol{y} = \boldsymbol{c}_B\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{y} = \boldsymbol{c}_B\boldsymbol{B}^{-1}\boldsymbol{b} = \boldsymbol{c}^\mathsf{T}\boldsymbol{x}_0,$$

so $\boldsymbol{x}_0$ is optimal. □

Before we arrive at an optimal solution, there might be several columns we can pivot on. How to choose the next column can be done in several ways and this is known as *pivot selection* or *pivot rules*. As was mentioned in the beginning of this subsection, there are pivot rules that are proven to have non-polynomial running time for some instances, but are very fast in practice nonetheless.

As a final piece of the puzzle, we observe how to get an initial basic feasible solution, and why we can assume that $\boldsymbol{A}$ has full row rank.

Let us find a basic feasible solution first. If our program is given in the form $\boldsymbol{Ax} \leq \boldsymbol{b}$ with $\boldsymbol{b} \geq \boldsymbol{0}$, then the slack-variables we add to get equality-constraints form can be used as initial basic variables and we are done.

Otherwise, we bring the program into standard form, i.e., we find a matrix $\boldsymbol{A}$ and a vector $\boldsymbol{b} \geq \boldsymbol{0}$ such that the goal is to maximize over $\boldsymbol{c}^\mathsf{T}\boldsymbol{x}$ where $\boldsymbol{x}$ satisfies $\boldsymbol{Ax} = \boldsymbol{b}$ and $\boldsymbol{x} \geq \boldsymbol{0}$. Now we solve the program in two phases. The first one will get us a basic feasible solution for the program, and the second one is simply the one we described above.

For the first phase, we attach new artificial variables $x_1^a, \ldots, x_m^a \geq 0$ and replace $\boldsymbol{A}$ by $\hat{\boldsymbol{A}} = [\boldsymbol{A}, \boldsymbol{I}]$. Then with $x_i^a = b_i$, $i = 1, \ldots, m$, we have a basic feasible solution. Now we want to get one where all artificial variables are nonbasic.

To this end we minimize $\sum_{i=1}^m x_i^a$ (or maximize the negative sum). If we can achieve zero as an objective value, then all artificial variables will indeed be zero. If the optimum is not zero, then $\boldsymbol{Ax}$ cannot be equal to $\boldsymbol{b}$ for any non-negative $\boldsymbol{x}$, and we can conclude that the system is infeasible.

The only detail remaining is that we have to make sure that the artificial variables are nonbasic in the optimal solution. If we stop with an artificial variable still in the basis, then we take a closer look at the corresponding row in the tableau. If all non-artificial variables have a zero as coefficient, then we have shown that this row in the original matrix $\boldsymbol{A}$ is not linearly independent from the others, since we arrived at the tableau by elementary row-operations. Thus we can just delete the row and continue. Thus, as a byproduct, we have shown how to make sure that $\boldsymbol{A}$ has full row-rank.

If some non-artificial coefficient is non-zero, we can pivot on that column, since it will neither change the objective nor lead to infeasibility. Strictly speaking this is not pivoting, since the relative cost might be negative. However, it is a valid way to remove one by one all artificial variables from the basis. We can then proceed with phase two, solving the original program.

1.3.3 Integer Optimization

When we want to solve integer linear programs, the most powerful tool to this date is using information from linear programs. We will briefly sketch two ways of doing so: Cutting plane algorithms and branch-and-bound methods.

The idea of cutting planes is based on the following observation: If we were to know that the rational polyhedron P we want to optimize over has only integer vertices, then any method solving the linear program will also solve the integer linear program. Thus, in the best of worlds, given P we would like to compute the convex hull of the integer points in P, called the *integer hull* P_I of P.

Let $H = \left\{\boldsymbol{x} \in \mathbb{R}^n \ : \ \boldsymbol{c}^\mathsf{T}\boldsymbol{x} \leq c_0\right\}$ be the halfspace defined by a valid inequality of P (see Subsection 1.3.1), where the components of $\boldsymbol{c}$ are relatively prime integers. Then we know

that $\boldsymbol{c}^\top \boldsymbol{x}$ will be integer for any $\boldsymbol{x} \in P \cap \mathbb{Z}^n$, and therefore $\boldsymbol{c}^\top \boldsymbol{x} \leq \lfloor c_0 \rfloor$ is valid for the integer hull of P. Geometrically, this corresponds to shifting the bounding hyperplane of H, until it contains an integer point.

Let $H_I = \left\{\boldsymbol{x} \in \mathbb{R}^n \ : \ \boldsymbol{c}^\top \boldsymbol{x} \leq \lfloor c_0 \rfloor\right\}$, then the bounding hyperplane of H_I is called a *cutting plane*, and we call the inequality $\boldsymbol{c}^\top \boldsymbol{x} \leq \lfloor c_0 \rfloor$ a *cut* for P. Note that by this definition a cut for P is not necessarily a valid inequality for P, but it is a valid inequality for P_I. More generally, any inequality we can derive by algebraic reasoning from a finite set of valid inequalities for P, such that the result is a valid inequality for P_I, will be called a cut. The cuts we derived above are equivalent to *Chvátal-Gomory rounding cuts*. We will introduce these cuts and several other families of cutting planes in greater detail in Chapter 2.

Cuttings planes can indeed be used to solve integer programs: If we define

$$P^{cl} = \bigcap_{H \supseteq P} H_I,$$

then it follows that $P \supseteq P^{cl} \supseteq P_I$, and, by repeating this procedure, $P \supseteq P^{cl} \supseteq (P^{cl})^{cl} \supseteq \cdots \supseteq P_I$. It turns out that in fact $P^{cl^t} = P_I$ for some integer t. This was proven for polytopes by Chvátal [20] (and indeed the smallest such t is called the *Chvátal rank* of P), and for rational polyhedra by Schrijver [74].

Moreover, P^{cl^k} is a polyhedron for any integer k [74]. Note that the latter implies that we only need to consider a finite set of halfspaces in every iteration. Since the cuts described above are, in some sense, the weakest type of cuts we will consider, all families of cutting planes in Chapter 2 will find the integer hull in finitely many iterations.

It should be noted, though, that this method is by far not guaranteed to terminate in polynomial time. To see this, consider the two-dimensional polytope P with the vertices $(0, 0)$, $(0, 1)$, and $(h, 1/2)$, for some integer $h > 0$ (see Figure 1.3; first mentioned in [20]). Then the integer hull of P is $\text{conv}\{(0, 0), (0, 1)\}$. However, it is not hard to show that P' will contain the point $(h - 1, 1/2)$, and therefore by induction we will need at least h iterations until we arrive at the integer hull.

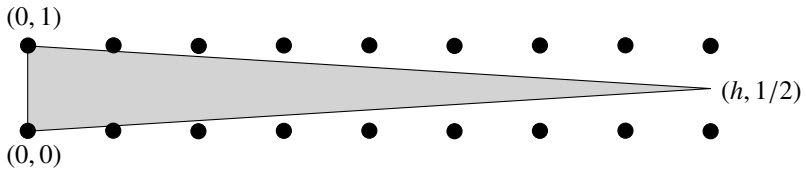

Figure 1.3: A polytope where we need at least h iterations of the cutting plane algorithm to find the integer hull.

As a positive result Cook, Coullard, and Turán [23] showed that if $P_I = \emptyset$, then there is a bound on the number of cutting planes in terms of the dimension. We will give more explicit descriptions of families of cutting planes in the next chapter.

The method known as *branch-and-bound* is also an iterative method, although based on a different principle: At stage 1 we have $\Pi_1 = \{P\}$. Suppose at stage k we have a collection $\Pi_k = \{P_1, \ldots, P_t\}$ such that

(i) $P_1, \ldots, P_t$ are pairwise disjoint polyhedra in $\mathbb{R}^n$;

(ii) all integer vectors in P are contained in $P_1 \cup \cdots \cup P_t$.

Then for $j = 1, \ldots, t$ we determine $\delta_j = \max\left\{\boldsymbol{c}^\mathsf{T}\boldsymbol{x} \ : \ \boldsymbol{x} \in P_j\right\}$, and let j^* be an index with $\delta_{j^*} \geq \delta_j$ for all j.

Let $\boldsymbol{x}^* \in P$ with $\boldsymbol{c}^\mathsf{T}\boldsymbol{x}^* = \delta_{j^*}$. If $\boldsymbol{x}^*$ is integral, we have found an optimal solution to the integer optimization program. If $\boldsymbol{x}^*$ is not integer in, say, component i, then define

$$Q_1 = \left\{\boldsymbol{x} \in P_{j^*} \ : \ x_i \leq \lfloor x_i^* \rfloor\right\} \quad \text{and} \quad Q_2 = \left\{\boldsymbol{x} \in P_{j^*} \ : \ x_i \geq \lceil x_i^* \rceil\right\}.$$

Let $\Pi_{k+1} = \{P_1, \ldots, P_{j^*-1}, Q_1, Q_2, P_{j^*+1}, \ldots, P_t\}$. It could happen that one or both of the Q_i are empty, in which case they are dropped from the collection. If the collection becomes empty, we conclude that the program is infeasibe.

Otherwise Π_{k+1} satisfies (i) and (ii), and thus we can start stage $k+1$. Note that we will only need to compute the δ_j for the two new polyhedra.

Again, the running time of this method is not polynomially bounded by the input size. Indeed, consider the integer program

$$\max\{x_1 \ : \ 2^m x_1 = (2^m + 1)x_2, 0 \leq x_1 \leq 2^m, x_1, x_2 \in \mathbb{Z}\}.$$

It is not hard to see that at stage k we have $\delta_{j^*} \geq 2^m - k$, while the only feasible solution is $(0, 0)$. Thus, the branch-and-bound method needs at least 2^m iterations, while the size of the program is linear in m.

Even if we assume that there is a family of collections Π_k which leads to an optimal solution with not too many stages, it is not obvious how to find it: at every stage we have up to n choices of components to branch on. Selecting good directions is a major challenge when we want to apply this method.

Note that the way we described the algorithm above, we do not necessarily stop when we find an integer solution, since it might have a smaller objective value than some δ_j for another P_j. However, we can at least use this to discard some of the polyhedra in our collection.

Suppose at some point the best completely integer solution we have found has objective value z_m. Then there is no hope of finding a better solution in a polyhedron P_i with $\delta_i \leq z_m$, and we can delete such polyhedra from the collection.

It should also be mentioned that the rule of branching on the maximal upper bound δ_{j^*} bares the risk of needing a large amount of storage-space, and a depth-first approach might be better in this respect. The other choice we make – which non-integer component to use for the branching – can also have an impact on the running time. One rule one can use, and that has proven to be useful in practice, is to choose the component that reduces the value δ_{j^*} the most and thus gives us the best chance to cut quickly.

Chapter Two
Reformulation-induced Cuts

In the following chapters, the leading question will be, generally speaking, how to solve the integer program

$$\max\left\{\boldsymbol{c}^{\mathsf{T}}\boldsymbol{x} \;:\; \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},\ \boldsymbol{x} \geq \boldsymbol{0}, \boldsymbol{x} \in \mathbb{Z}^n\right\}, \tag{eq-IP}$$

where $\boldsymbol{A}$ is an integer $m \times n$ matrix of full row rank and $\boldsymbol{b}$ an integer m-vector.

In this chapter, we will see how cutting plane algorithms can be combined with a reformulation of the program into a space of lower dimension (cf. Aardal, Hurkens, and Lenstra [2]).

In Chapter 1 we already described a method of how to reformulate the set of feasible points for the linear program. However, this method did not take into account the shape of the set relative to the integer lattice. Instead we therefore need a reformulation that preserves the structure of the sublattice we get when we intersect $\mathbb{Z}^n$ with the smallest affine subspace containing all feasible points of (eq-IP).

Recall that in Lemma 1.5 we were given a way to find, in polynomial time, a vector $\boldsymbol{x}_0 \in \mathbb{Z}^n$ with $\boldsymbol{A}\boldsymbol{x}_0 = \boldsymbol{b}$, or certify that no such vector exists. Clearly, if there is no integer solution to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, then (eq-IP) is infeasible. Further below we will describe a variation of this method to find such a vector $\boldsymbol{x}_0$, which arises naturally from other calculations. For the moment, however, let us assume we have $\boldsymbol{x}_0$ with the above properties.

Let $\boldsymbol{y} \in \mathbb{Z}^n$ with $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{b}$. Then $\boldsymbol{z} := \boldsymbol{x}_0 - \boldsymbol{y} \in \mathbb{Z}^n$ and $\boldsymbol{A}\boldsymbol{z} = \boldsymbol{0}$. Conversely, any $\boldsymbol{z} \in \mathbb{Z}^n$ with $\boldsymbol{A}\boldsymbol{z} = \boldsymbol{0}$ also defines a vector $\boldsymbol{y} := \boldsymbol{x}_0 - \boldsymbol{z} \in \mathbb{Z}^n$ with $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{b}$.

We have therefore found a way to translate all feasible points of (eq-IP) into a linear subspace. Observe that if $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{Z}^n$ with $\boldsymbol{A}\boldsymbol{x}_i = \boldsymbol{0}$, $i = 1, 2$, then also $\boldsymbol{A}(\boldsymbol{x}_1 \pm \boldsymbol{x}_2) = \boldsymbol{0}$ and therefore $\{\boldsymbol{x} \in \mathbb{Z}^n \;:\; \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\}$ is a lattice. Note that it is a pure sublattice of $\mathbb{Z}^n$. We will call this the *kernel lattice* of $\boldsymbol{A}$ and denote it by $\ker_{\mathbb{Z}}(\boldsymbol{A})$.

Hence, we have found a direct correspondence between the integer solutions of $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ and the lattice $\ker_{\mathbb{Z}}(\boldsymbol{A})$. More explicitly, let $\boldsymbol{Q}$ be a basis of $\ker_{\mathbb{Z}}(\boldsymbol{A})$. Then a vector $\boldsymbol{x} \in \mathbb{Z}^n$ satisfies $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ if and only if there is a vector $\boldsymbol{\mu} \in \mathbb{Z}^{n-m}$ with $\boldsymbol{x} = \boldsymbol{x}_0 + \boldsymbol{Q}\boldsymbol{\mu}$.

Due to the nonnegativity requirements on the $\boldsymbol{x}$-variables in the integer program (eq-IP), we now obtain an equivalent formulation

$$\max\left\{\boldsymbol{c}^{\mathsf{T}}(\boldsymbol{x}_0 + \boldsymbol{Q}\boldsymbol{\mu}) \;:\; \boldsymbol{Q}\boldsymbol{\mu} \geq -\boldsymbol{x}_0, \boldsymbol{\mu} \in \mathbb{Z}^{n-m}\right\}. \tag{2.1}$$

Note that this program is not necessarily of the form (IP) yet, as we do not require $\boldsymbol{\mu} \geq \boldsymbol{0}$. We will come back to this in a little while. However, we now have a program in $n - m$ variables, and with n inequalities.

To obtain the vector $\boldsymbol{x}_0$ and the basis $\boldsymbol{Q}$, we construct an auxiliary lattice $L \subseteq \mathbb{R}^{n+m+1}$. Let $\boldsymbol{a}_i$ denote the i^{th} row of $\boldsymbol{A}$, and we may assume $\gcd(a_{i1}, \ldots, a_{in}) = 1$, for $i = 1, \ldots, m$. The elements of L will now have the form

$$(x_1, \ldots, x_n, N_1 y, N_2(\boldsymbol{a}_1 \boldsymbol{x} - b_1 y), \ldots, N_2(\boldsymbol{a}_m \boldsymbol{x} - b_m y))^\mathsf{T},$$

where y is a new integer variable, and N_1 and N_2 are nonzero integer numbers. A basis of L is then given by the $(n + 1 + m) \times (n + 1)$-matrix

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & N_1 \\ N_2\boldsymbol{A} & -N_2\boldsymbol{b} \end{bmatrix}.$$

Proposition 2.1 ([2])**.** *Let $L \subseteq \mathbb{R}^{n+m+1}$ be defined as above. Then*

(i) an integer vector $\boldsymbol{x}$ satisfies $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ if and only if the vector $(\boldsymbol{x}^\mathsf{T}, N_1, \boldsymbol{0})^\mathsf{T}$ belongs to L,

(ii) an integer vector $\hat{\boldsymbol{x}}$ satisfies $\boldsymbol{A}\hat{\boldsymbol{x}} = \boldsymbol{0}$ if and only if the vector $(\hat{\boldsymbol{x}}^\mathsf{T}, 0, \boldsymbol{0})^\mathsf{T}$ belongs to L.

Proof. In (i) we set the variable $y = 1$. Observe that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ is satisfied if and only if the last m coordinates are zero. Similarly in (ii) we set $y = 0$. □

We will see in Theorem 2.3 that if there is an integer solution to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ and the numbers N_1, N_2 are chosen large enough, then the first $n - m$ columns of the basis we obtain by LLL-reducing $\boldsymbol{B}$ will be lattice vectors of the form given in (ii), and the $(n - m + 1)^{st}$ column is a vector of the form given in (i).

Lemma 2.2 ([40, 38])**.** *Let* $\mathrm{HNF}(\boldsymbol{A}) = [\boldsymbol{H}, \boldsymbol{0}] = \boldsymbol{A}\boldsymbol{U}$*, where $\boldsymbol{U}$ is a unimodular matrix of dimension $n \times n$. Define $n - m + 1$ vectors by*

$$(\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-m}) = \boldsymbol{U} \begin{bmatrix} \boldsymbol{H}^{-1}\boldsymbol{b} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}$$

Then $\boldsymbol{A}\boldsymbol{x}_0 = \boldsymbol{b}$ and $\boldsymbol{A}\boldsymbol{x}_j = \boldsymbol{0}$ for $1 \leq j \leq n - m$. Moreover, the vectors $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-m}$ are linearly independent and their sizes are polynomially bounded by the sizes of $\boldsymbol{A}$ and $\boldsymbol{b}$.

Proof. The equation $\boldsymbol{A}\boldsymbol{x}_0 = \boldsymbol{b}$ is the same we computed in the proof of Lemma 1.4. Let $\boldsymbol{e}_j$ be the j^{th} column of the identity matrix $\boldsymbol{I}$, then for each $\boldsymbol{x}_j$, $1 \leq j \leq n - m$, we have

$$\boldsymbol{A}\boldsymbol{x}_j = \boldsymbol{A}\boldsymbol{U}\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{e}_j \end{pmatrix} = [\boldsymbol{H}, \boldsymbol{0}]\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{e}_j \end{pmatrix} = \boldsymbol{0}.$$

The rest is clear from construction and Lemma 1.5. □

Theorem 2.3 ([2])**.** *Assume that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ has an integer solution, and let $\hat{\boldsymbol{B}} = [\hat{\boldsymbol{b}}_1, \ldots, \hat{\boldsymbol{b}}_{n+1}]$ be the basis of L we obtain by applying the LLL algorithm to the basis $\boldsymbol{B}$.*

Then there exist numbers N_{01} and N_{02}, such that for numbers N_1, N_2 with $N_1 > N_{01}$ and $N_2 > 2^{n+m}N_1^2 + N_{02}$ the following holds:

- *In $\hat{\boldsymbol{b}}_j$, all coordinates after the n^{th} are zero, for $1 \leq j \leq n - m$;*
- *In $\hat{\boldsymbol{b}}_{n-m+1}$, all coordinates after the $(n + 1)^{st}$ are zero;*

- *The* $(n+1)^{st}$ *coordinate of* $\hat{\boldsymbol{b}}_{n-m+1}$ *has cardinality* N_1.

In practice, the integers N_1 and N_2 can be chosen considerably smaller than the bounds given in the theorem. This is due to the fact that the bounds on the lengths of the vectors in a reduced basis are pessimistic for the huge majority of lattices one encounters.

Lemma 2.4. *Let* $\boldsymbol{x}_0 \in \mathbb{Z}^n$ *with* $\boldsymbol{A}\boldsymbol{x}_0 = \boldsymbol{b}$ *and let* $\boldsymbol{Q}$ *be a basis of* $\ker_{\mathbb{Z}}(\boldsymbol{A})$. *Then in polynomial time we can find* $\hat{\boldsymbol{x}}_0 \in \mathbb{Z}^n$ *such that* $\boldsymbol{A}\hat{\boldsymbol{x}}_0 = \boldsymbol{b}$ *and* $\boldsymbol{\mu} \geq 0$ *for all* $\boldsymbol{\mu} \in \mathbb{R}^d$ *with* $\boldsymbol{Q}\boldsymbol{\mu} \geq -\hat{\boldsymbol{x}}_0$, *or conclude that no such vector exists.*

Proof. For each $i = 1, \ldots, m$, we solve the linear program $\min\{\mu_i : \boldsymbol{Q}\boldsymbol{\mu} \geq -\boldsymbol{x}_0\}$. If one of them is unbounded, we are done. Otherwise let $\boldsymbol{\mu}^*$ be vector of the corresponding optimal values and define $\hat{\boldsymbol{x}}_0 = \boldsymbol{x}_0 + \boldsymbol{Q}\lfloor\boldsymbol{\mu}^*\rfloor$. Then $\hat{\boldsymbol{x}}_0$ is integer and satisfies $\boldsymbol{A}\hat{\boldsymbol{x}}_0 = \boldsymbol{b}$.

Furthermore, $\boldsymbol{\mu}$ satisfies $\boldsymbol{Q}\boldsymbol{\mu} \geq -\hat{\boldsymbol{x}}_0$ if and only if

$$\boldsymbol{Q}\boldsymbol{\mu} \geq -\boldsymbol{x}_0 - \boldsymbol{Q}\lfloor\boldsymbol{\mu}^*\rfloor,$$

which, in turn, is equivalent to

$$\boldsymbol{Q}(\boldsymbol{\mu} + \lfloor\boldsymbol{\mu}^*\rfloor) \geq -\boldsymbol{x}_0.$$

But then by the definition of $\boldsymbol{\mu}^*$ we have $\boldsymbol{\mu} + \lfloor\boldsymbol{\mu}^*\rfloor \geq \boldsymbol{\mu}^* \geq \lfloor\boldsymbol{\mu}^*\rfloor$ and thus $\boldsymbol{\mu} \geq \boldsymbol{0}$. □

Since we can check with linear programming whether an integer program is unbounded, we have now a complete description of a correspondence between the integer program (eq-IP) in dimension n in standard form, and the integer program (2.1) in dimension $d = n - m$ in canonical form.

Recall from Chapter 1.3.3 that given the set P of feasible points for the linear program, we can find the integer hull P_I by cutting planes, i.e., by taking valid inequalities for P and shift them in a way so they remain valid for P_I, and then iterating this procedure.

What we will do in this chapter is to combine cutting plane algorithms with the reformulation given above.

Note that the following section is quite long, because we will not simply state the cuts, but instead derive why they are valid for P_I. Subsequently, we will show that a cut derived for the reformulation by one of the (general) constructions below can also be found through the same method in the original formulation. However, as we show in the last section, if we restrict our attention to the constructions most commonly used in practice, cuts from the reformulation can be strictly stronger.

2.1 Families of Cutting Planes

The main reference this section is based on is the paper [27] by Cornuéjols and Li. While most of the cuts below are defined for mixed integer programs, we will formulate them for the pure integer case here. Furthermore, we will assume that the polyhedra are bounded, since these are the cases where integer programs differ from their linear relaxation.

Let $\boldsymbol{A} \in \mathbb{Z}^{m\times d}$ and $\boldsymbol{b} \in \mathbb{Z}^m$ be given, and as we have seen in (1.16), the sets

$$P = \left\{\boldsymbol{x} \in \mathbb{R}^d \ : \ \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \boldsymbol{x} \geq \boldsymbol{0}\right\} \text{ and } P' = \left\{\boldsymbol{z} \in \mathbb{R}^{d+m} \ : \ \boldsymbol{A}'\boldsymbol{z} = \boldsymbol{b}, \boldsymbol{z} \geq \boldsymbol{0}\right\}$$

are equivalent expressions, where $\boldsymbol{A}' = [\boldsymbol{A}, \boldsymbol{I}]$.

We already established that this kind of reformulation is not suited for integer programs with feasible set P'. However, it is not hard to see that an integer point in P' does translate to an integer point in P, so we may use this reformulation to go up in dimensions. Furthermore, we want to make use of the simplex method, which is constructed to take programs with equality-constraints as input.

We will from now on, for convenience, assume that $\boldsymbol{x} \geq \boldsymbol{0}$ is part of the constraints $\boldsymbol{Ax} \leq \boldsymbol{b}$. Recall that we defined the convex hull of the set of integer points in P as

$$P_I = \text{conv}(\left\{\boldsymbol{x} \in \mathbb{Z}^d \ : \ \boldsymbol{Ax} \leq \boldsymbol{b}\right\}),$$

and similarly P'_I for P'.

Finally, observe that for any $\boldsymbol{u} \in \mathbb{R}^m$ the equation $\boldsymbol{u}^\mathsf{T}\boldsymbol{A'z} = \boldsymbol{u}^\mathsf{T}\boldsymbol{b}$ is satisfied by all $\boldsymbol{z} \in P'$.

To be able to speak about how useful a family of cuts is in the description of P_I, we define the *elementary closure* associated with the family. Given a polytope P and a family F of cuts (see 1.3.3) for P, we define P_F as the convex set obtained as the intersection of all inequalities in F. While it is not obvious from this definition, it turns out that P_F is indeed a polyhedron again, for all families of cuts described in this section [20, 74, 24]. The concept of elementary closures was introduced by Chvátal [20].

Iterating this procedure will, in the pure integer case, stabilize at $\text{conv}(P_I)$ in finitely many steps (first proven by Gomory [41, 43]). If we allow for continuous variables as well, this finite convergence does not always occur, as was shown with a beautifully simple example by Cook, Kannan, and Schrijver in [24].

A function $g : \mathbb{R} \to \mathbb{R}$ is called *subadditive*, if $g(p) + g(q) \geq g(p + q)$ for all $p, q \in \mathbb{R}$. The following proposition forms the basis on which we will show the validity of some very useful inequalities for P_I.

Proposition 2.5 (See [69], Prop. II.1.4.1)**.** *If $\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{z} = \beta$ is satisfied for all $\boldsymbol{z} \in P'$ and g is a subadditive function with $g(0) = 0$, then*

$$\sum_{j=1}^{d+m} g(\alpha_j)z_j \geq g(\beta) \tag{2.2}$$

is a valid inequality for P'_I.

Proof. Note that $g(\alpha_j) \cdot 0 = 0 = g(0) = g(\alpha_j \cdot 0)$, and thus by induction

$$\begin{aligned} g(\alpha_j) \cdot k &= g(\alpha_j) + g(\alpha_j) \cdot (k-1) \geq g(\alpha_j) + g(\alpha_j \cdot (k-1)) \\ &\geq g(\alpha_j + \alpha_j \cdot (k-1)) = g(\alpha_j \cdot k) \end{aligned}$$

for all $k \in \mathbb{Z}$, $k \geq 1$. Note that the first inequality holds because of the induction hypothesis, while the second inequality holds because g is subadditive.

Since $\boldsymbol{z} \geq \boldsymbol{0}$, we can now compute for all $\boldsymbol{z} \in P'_I \cap \mathbb{Z}^{d+m}$

$$\sum_{j=1}^{d+m} g(\alpha_j)z_j \geq \sum_{j=1}^{d+m} g(\alpha_j z_j) \geq g\left(\sum_{j=1}^{d+m} \alpha_j z_j\right) = g(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{z}) = g(\beta),$$

where the last inequality again holds because g is subadditive. The inequality is then also valid for any point in the convex hull of $P'_I \cap \mathbb{Z}^{d+m}$. □

Note that here we forced the slack variables $(z_{d+1}, \dots, z_{d+m})$ to be integer as well. This could theoretically lead to unwanted effects, since there are no explicit integrality requirements for the slack variables and thus P'_I might be empty while P_I is not. To deal with this, we introduce another function h to take care of the slack variables separately.

The step in the proof where we used integrality was when we showed that $g(\alpha_j)z_j \geq g(\alpha_j z_j)$. This is certainly preserved if h is *positive homogeneous*, i.e., $h(p)\lambda = h(p\lambda)$ for all $\lambda \geq 0$ and $p \in \mathbb{R}$. Secondly, to keep the inequality valid, we want h to *dominate* g, i.e., $h(p) \geq g(p)$ for all $p \in \mathbb{R}$.

Thus, we get the following proposition.

Proposition 2.6. *If $\alpha^\top z = \beta$ is satisfied for all $z \in P'$, g is a subadditive function with $g(0) = 0$, and h is positive homogeneous and dominates g, then*

$$\sum_{i=1}^{d} g(\alpha_i)z_i + \sum_{j=d+1}^{d+m} h(\alpha_j)z_j \geq g(\beta) \tag{2.3}$$

is a valid inequality for $\operatorname{conv}(P' \cap (\mathbb{Z}^d \times \mathbb{R}^m)$. □

Chvátal-Gomory cuts and P_{CG}

Let $\boldsymbol{u} \in \mathbb{R}^m$ with $\boldsymbol{u} \geq \mathbf{0}$. Then the inequality $\boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{u}^\top \boldsymbol{b}$ is valid for P, and, since $\boldsymbol{x} \geq \mathbf{0}$, we only weaken the inequality by rounding down each coordinate of $\boldsymbol{u}^\top \boldsymbol{A}$. Therefore, the inequality $\left\lfloor \boldsymbol{u}^\top \boldsymbol{A} \right\rfloor \boldsymbol{x} \leq \boldsymbol{u}^\top \boldsymbol{b}$ is also valid for P, where $\lfloor \boldsymbol{v} \rfloor$ denotes the vector we get by rounding down every entry of the vector $\boldsymbol{v}$.

For all elements of $P_I \cap \mathbb{Z}^d$ we can also round down the right-hand side and obtain the inequality

$$\left\lfloor \boldsymbol{u}^\top \boldsymbol{A} \right\rfloor \boldsymbol{x} \leq \left\lfloor \boldsymbol{u}^\top \boldsymbol{b} \right\rfloor. \tag{2.4}$$

Once again, this inequality must then also be satisfied for convex combinations of points in $P_I \cap \mathbb{Z}^d$, and thus we have found a valid inequality for P_I. Cuts of this form were developed by Chvátal [20], and are generally referred to as *Chvátal cuts* [27]. we will see in Theorem 2.7 that we can concentrate on the case where $\mathbf{0} \leq \boldsymbol{u} < \mathbf{1}$.

There is, however, a different way to obtain them, developed earlier by Gomory [41, 42, 43]. We describe it here as well, as from this description we can easily obtain a very interesting subfamily of cuts from basic feasible solutions (see Subsection 1.3.2).

Note that $g(p) = p - \lfloor p \rfloor$ is subadditive, and since $(z_{d+1}, \dots, z_{d+m})^\top = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}$ implies that $\boldsymbol{z}$ is integer if $\boldsymbol{x}$ is integer (recall that we assumed $\boldsymbol{A}$ and $\boldsymbol{b}$ to be integer as well), we can use Proposition 2.5. Thus

$$\left(\boldsymbol{u}^\top \boldsymbol{A}' - \left\lfloor \boldsymbol{u}^\top \boldsymbol{A}' \right\rfloor\right) \boldsymbol{z} \geq (\boldsymbol{u}^\top \boldsymbol{b} - \left\lfloor \boldsymbol{u}^\top \boldsymbol{b} \right\rfloor)$$

is a valid inequality for P'_I. Now we substitute $(z_{d+1}, \dots, z_{d+m})^\top = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}$ and obtain $\boldsymbol{u}^\top \boldsymbol{A}'\boldsymbol{z} = (\boldsymbol{u}^\top \boldsymbol{A})\boldsymbol{x} + \boldsymbol{u}^\top(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})$. Hence

$$\left\lfloor \boldsymbol{u}^\top \boldsymbol{A} \right\rfloor \boldsymbol{x} - \lfloor \boldsymbol{u} \rfloor^\top \boldsymbol{A}\boldsymbol{x} \leq \left\lfloor \boldsymbol{u}^\top \boldsymbol{b} \right\rfloor - \lfloor \boldsymbol{u} \rfloor^\top \boldsymbol{b} \tag{2.5}$$

is a valid inequality for P_I. Cuts of this form are called *Gomory fractional cuts* [27].

Note that if $\mathbf{0} \leq \boldsymbol{u} < \mathbf{1}$, then (2.4) and (2.5) are the same. In fact, we get the following more general relationship.

Theorem 2.7 (see, e.g., [69])**.** *Any cut of the form* (2.5) *with the vector* $\boldsymbol{u}$ *can be derived as a cut of the form* (2.4) *with the vector* $\tilde{\boldsymbol{u}} = \boldsymbol{u} - \lfloor \boldsymbol{u} \rfloor$.

Conversely, any cut of the form (2.4) *is equal to or dominated by a cut of the form* (2.5).

Proof. First note that $\tilde{\boldsymbol{u}} = \boldsymbol{u} - \lfloor \boldsymbol{u} \rfloor \geq \boldsymbol{0}$ and $\left\lfloor \tilde{\boldsymbol{u}}^\mathsf{T} \boldsymbol{A} \right\rfloor = \left\lfloor \boldsymbol{u}^\mathsf{T} \boldsymbol{A} - \lfloor \boldsymbol{u} \rfloor^\mathsf{T} \boldsymbol{A} \right\rfloor = \left\lfloor \boldsymbol{u}^\mathsf{T} \boldsymbol{A} \right\rfloor - \lfloor \boldsymbol{u} \rfloor^\mathsf{T} \boldsymbol{A}$, and $\left\lfloor \tilde{\boldsymbol{u}}^\mathsf{T} \boldsymbol{b} \right\rfloor = \left\lfloor \boldsymbol{u}^\mathsf{T} \boldsymbol{b} - \lfloor \boldsymbol{u} \rfloor^\mathsf{T} \boldsymbol{b} \right\rfloor = \left\lfloor \boldsymbol{u}^\mathsf{T} \boldsymbol{b} \right\rfloor - \lfloor \boldsymbol{u} \rfloor^\mathsf{T} \boldsymbol{b}$, where we use that $\lfloor \boldsymbol{u} \rfloor^\mathsf{T} \boldsymbol{A}$ and $\lfloor \boldsymbol{u} \rfloor^\mathsf{T} \boldsymbol{b}$ are integer.

The first statement of the theorem then follows, as we have

$$\begin{aligned} & \left\lfloor \boldsymbol{u}^\mathsf{T} \bar{\boldsymbol{A}} \right\rfloor \bar{\boldsymbol{x}} - \lfloor \boldsymbol{u} \rfloor^\mathsf{T} \bar{\boldsymbol{A}} \bar{\boldsymbol{x}} \leq \left\lfloor \boldsymbol{u}^\mathsf{T} \boldsymbol{b} \right\rfloor - \lfloor \boldsymbol{u} \rfloor^\mathsf{T} \boldsymbol{b} \\ \Leftrightarrow \quad & \left\lfloor \tilde{\boldsymbol{u}}^\mathsf{T} \bar{\boldsymbol{A}} \right\rfloor \bar{\boldsymbol{x}} \leq \left\lfloor \tilde{\boldsymbol{u}}^\mathsf{T} \boldsymbol{b} \right\rfloor, \end{aligned}$$

For the second statement we first show that any cut of the form (2.4) with the vector $\boldsymbol{u} \geq \boldsymbol{0}$ can be written as a non-negative integer linear combination of a cut of the same form with the vector $\tilde{\boldsymbol{u}} = \boldsymbol{u} - \lfloor \boldsymbol{u} \rfloor$ and the inequalities $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$.

Indeed, by the above equations we have that $\left\lfloor \boldsymbol{u}^\mathsf{T} \boldsymbol{A} \right\rfloor \boldsymbol{x} \leq \left\lfloor \boldsymbol{u}^\mathsf{T} \boldsymbol{b} \right\rfloor$ can be obtained as the sum of $\left\lfloor \tilde{\boldsymbol{u}}^\mathsf{T} \boldsymbol{A} \right\rfloor \boldsymbol{x} \leq \left\lfloor \tilde{\boldsymbol{u}}^\mathsf{T} \boldsymbol{b} \right\rfloor$ and $\lfloor \boldsymbol{u} \rfloor^\mathsf{T} \boldsymbol{A}\boldsymbol{x} \leq \lfloor \boldsymbol{u} \rfloor^\mathsf{T} \boldsymbol{b}$.

As we have noted before, the cut $\left\lfloor \tilde{\boldsymbol{u}}^\mathsf{T} \boldsymbol{A} \right\rfloor \boldsymbol{x} \leq \left\lfloor \tilde{\boldsymbol{u}}^\mathsf{T} \boldsymbol{b} \right\rfloor$ is of the form (2.5). □

We also observe that had we used the seemingly weaker Proposition 2.6 instead of 2.5 (with $h = id$), we would have directly gotten Inequality (2.4).

As the inequalities in (2.4) and (2.5) form (basically) the same family of cuts, we will refer to them as *Chvátal-Gomory cuts*, which is also the name most commonly found in the literature. Let P_{CG} be the corresponding elementary closure.

Gomory mixed integer cuts and P_{GMI}

For the Chvátal-Gomory cuts we used the fractional parts of the reals as our subadditive function, now we will use a slightly more complicated function to derive potentially stronger cuts. For $0 \leq f_0 < 1$ and $p \in \mathbb{R}$ define $f(p) = p - \lfloor p \rfloor$ and

$$\gamma^{f_0}(p) = \begin{cases} f(p) & \text{if } f(p) \leq f_0 \\ \frac{f_0}{1-f_0}(1 - f(p)) & \text{otherwise.} \end{cases}$$

Proposition 2.8 (See [69], Prop. II.1.4.7)**.** *The function* γ^{f_0} *is subadditive for* $0 \leq f_0 < 1$.

Proof. Define $(q)^+ = \max\{0, q\}$ for $q \in \mathbb{R}$. Then it is not hard to see that $\gamma^{f_0}(p) = f(p) - \frac{(f(p)-f_0)^+}{1-f_0}$. Let $p, q \in \mathbb{R}$.

Case 1: $f(p) + f(q) < 1$. We compute

$$\begin{aligned} \gamma^{f_0}(p) + \gamma^{f_0}(q) &= f(p) - \frac{(f(p) - f_0)^+}{1 - f_0} + f(q) - \frac{(f(q) - f_0)^+}{1 - f_0} \\ &= f(p+q) - \frac{(f(p) - f_0)^+ + (f(q) - f_0)^+}{1 - f_0} \geq \gamma^{f_0}(p+q) \end{aligned}$$

Case 2: $f(p) + f(q) \geq 1$ and $f(q) \leq f_0$. Then

$$\begin{aligned} \gamma^{f_0}(p) + \gamma^{f_0}(q) &= f(p) - \frac{(f(p) - f_0)^+}{1 - f_0} + f(q) > f(p) + f(q) - 1 \\ &= f(p+q) \geq \gamma^{f_0}(p+q) \end{aligned}$$

Case 3: $f(p) + f(q) \geq 1$ and $f(p), f(q) > f_0$. then

$$\begin{aligned}
\gamma^{f_0}(p) + \gamma^{f_0}(q) &= f(p) - \frac{f(p) - f_0}{1 - f_0} + f(q) - \frac{f(q) - f_0}{1 - f_0} \\
&= f(p) + f(q) - \frac{f(p) + f(q) - f_0 - 1 + 1 - f_0}{1 - f_0} \\
&= f(p) + f(q) - 1 - \frac{f(p) + f(q) - 1 - f_0}{1 - f_0} \\
&= f(p + q) - \frac{f(p + q) - f_0}{1 - f_0} \\
&\geq \gamma^{f_0}(p + q)
\end{aligned}$$

□

Let $\boldsymbol{u} \in \mathbb{R}^m$, and define $\hat{\boldsymbol{a}}^\top = \boldsymbol{u}^\top \boldsymbol{A}'$ and $\hat{b} = \boldsymbol{u}^\top \boldsymbol{b}$. Then, as we have observed before, $\hat{\boldsymbol{a}}^\top \boldsymbol{z} = \hat{b}$ is satisfied by all $\boldsymbol{z} \in P'$. Furthermore, we define the fractional parts $f_i = f(\hat{a}_i) = \hat{a}_i - \lfloor \hat{a}_i \rfloor$, for $i = 1, \ldots, d + m$, and $f_0 = \hat{b} - \left\lfloor \hat{b} \right\rfloor$. Then, using again Proposition 2.5, we can conclude that the inequality

$$\sum_{f_i \leq f_0} f_i z_i + \frac{f_0}{1 - f_0} \sum_{f_i > f_0} (1 - f_i) z_i \geq f_0 \tag{2.6}$$

is valid for P'_I. Using the substitutions $(z_{d+1}, \ldots, z_{d+m})^\top = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}$, we get a valid inequality for P_I from this.

If we want to use Proposition 2.6 instead (as one might if $\boldsymbol{A}$ and $\boldsymbol{b}$ are not integer), then we can define $\delta^{f_0}(p) = p$ for $p \geq 0$ and $\delta^{f_0}(p) = -\frac{f_0}{1 - f_0} p$ for $p < 0$. It is not hard to check that δ^{f_0} is positive homogeneous and dominates γ^{f_0} and thus

$$\sum_{i \leq d; f_i \leq f_0} f_i x_i + \frac{f_0}{1 - f_0} \sum_{i \leq d; f_i > f_0} (1 - f_i) x_i + \sum_{u_j \geq 0} u_j z_{d+j} - \frac{f_0}{1 - f_0} \sum_{u_j < 0} u_j z_{d+j} \geq f_0 \tag{2.7}$$

is a valid inequality for $\operatorname{conv}(P' \cap (\mathbb{Z}^d \times \mathbb{R}^m)$. Again, we can derive a valid inequality for P_I from this by substituting the last m variables. The derivation of these cuts via subadditive functions is also nicely illustrated by Fischetti and Saturni in [36].

Cuts of the form (2.6) and (2.7) are called *Gomory mixed integer cuts*. They are of particular interest for *mixed* integer programs, but as we will see in Lemma 2.11 they are also a strengthening of Chvátal-Gomory cuts in the pure integer case. Let P_{GMI} be the corresponding elementary closure.

Split cuts and P_{S}

Let $(\boldsymbol{\pi}, \pi_0) \in \mathbb{Z}^{d+1}$, and let

$$\boldsymbol{c}\boldsymbol{x} - \alpha(\boldsymbol{\pi}\boldsymbol{x} - \pi_0) \leq c_0 \quad \text{and} \tag{2.8}$$

$$\boldsymbol{c}\boldsymbol{x} + \beta(\boldsymbol{\pi}\boldsymbol{x} - \pi_0 - 1) \leq c_0 \tag{2.9}$$

be valid inequalities for P with $\alpha, \beta \geq 0$, then

$$\boldsymbol{c}\boldsymbol{x} \leq c_0 \tag{2.10}$$

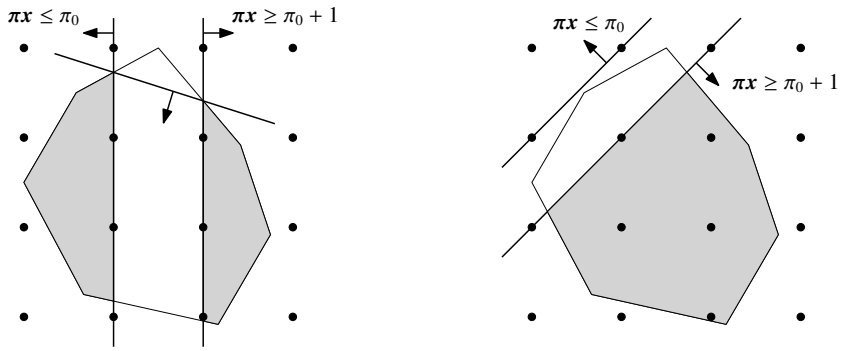

Figure 2.1: Geometric interpretation of split cuts.

is valid for $S_1 = P_I \cap \left\{x \in \mathbb{Z}^d \ : \ \pi x \leq \pi_0\right\}$ due to (2.8), and (2.10) is also valid for $S_2 = P_I \cap \left\{x \in \mathbb{Z}^d \ : \ \pi x \geq \pi_0 + 1\right\}$, due to (2.9). Since $P_I = \text{conv}(S_1 \cup S_2)$, Inequality (2.10) must be valid for P_I as well.

Cuts of this form are called *split cuts* or *disjunctive cuts* [11, 24] and were introduced by Balas [11]. A geometric interpretation of this argument can be found in Figure 2.1. Let P_S be the corresponding elementary closure.

Mixed integer rounding cuts and P_{MIR}

Assume the inequalities

$$c_1 x \leq c_0^1 \quad \text{and} \quad c_2 x \leq c_0^2$$

are valid for P, and $\pi = c_2 - c_1 \in \mathbb{Z}^d$. Define $\pi_0 = \left\lfloor c_0^2 - c_0^1 \right\rfloor$ and $\gamma = c_0^2 - c_0^1 - \pi_0$. We will show that this implies that

$$\pi x + (c_1 x - c_0^1)\frac{1}{1-\gamma} \leq \pi_0 \tag{2.11}$$

is valid for P_I.

Proposition 2.9. *Let $a \in \mathbb{R}^d$, $b, g \in \mathbb{R}$, $g \geq 0$, and consider the set*

$$T = \left\{(x, q) \in \mathbb{Z}^d \times \mathbb{R} \ : \ a^\mathsf{T} x - gq \leq b, x \geq \mathbf{0}, q \geq 0\right\}.$$

Then

$$\lfloor a \rfloor^\mathsf{T} x - \frac{1}{1-f_0} gq \leq \lfloor b \rfloor$$

is valid for T, where $f_0 = b - \lfloor b \rfloor$.

Proof. Consider first the case $-gq > f_0 - 1$. Then for any $(x, q) \in T$ we know that x is non-negative and therefore

$$\lfloor a \rfloor^\mathsf{T} x \leq a^\mathsf{T} x \leq b + gq < b - (f_0 - 1) = \lfloor b \rfloor + 1.$$

Since $\lfloor a \rfloor^\mathsf{T} x \in \mathbb{Z}$, this implies the inequality $\lfloor a \rfloor^\mathsf{T} x \leq \lfloor b \rfloor$, which we can add to $-\frac{1}{1-f_0} gq < 0$.

If we instead have $-gq \leq f_0 - 1$, then for any $(\boldsymbol{x}, q) \in T$ we get

$$\lfloor \boldsymbol{a} \rfloor^\mathsf{T} \boldsymbol{x} - \frac{1}{1-f_0} gq \leq \boldsymbol{a}^\mathsf{T} \boldsymbol{x} - \frac{1}{1-f_0} gq \leq b + gq - gq \frac{1}{1-f_0} = b - gq(\frac{1}{1-f_0} - 1)$$
$$= b + \frac{f_0}{1-f_0}(f_0 - 1) = b - f_0 = \lfloor b \rfloor .$$

□

Note that it is not hard to get a more general statement from this, with more real-valued variables and coefficients of different signs, see [69].

Now we observe that

$$\boldsymbol{c}_1 \boldsymbol{x} \leq c_0^1 \quad \Leftrightarrow \quad q := c_0^1 - \boldsymbol{c}_1 \boldsymbol{x} \geq 0$$

and

$$\boldsymbol{c}_2 \boldsymbol{x} \leq c_0^2 \Leftrightarrow \quad (\boldsymbol{c}_2 - \boldsymbol{c}_1)\boldsymbol{x} - (c_0^1 - \boldsymbol{c}_1 \boldsymbol{x}) \leq c_0^2 - c_0^1$$
$$\Leftrightarrow \quad (\boldsymbol{c}_2 - \boldsymbol{c}_1)\boldsymbol{x} - q \leq c_0^2 - c_0^1,$$

which together means that we can use Proposition 2.9 (with $g = 1$), and thus we have shown the validity of (2.11).

Cuts of the form (2.11) were introduced by Nemhauser and Wolsey [68] and are called *mixed integer rounding cuts*. Let P_{MIR} be the corresponding elementary closure.

There are several other families of cuts. In particular for the case $\boldsymbol{x} \in \{0,1\}^n$ (plus possibly some continuous variables) one could name *lift-and-project cuts* [12, 77, 61], *Sherali-Adams cuts* [77], and *Lovász-Schrijver cuts* [61]. A description of them, as well as a comparison of their strengths, can be found in [27].

All the above families have the well-studied sub-families which are derived from basic (feasible) solutions, and are in some sense easier to obtain: Note that the cuts described above start with either taking some valid inequalities for P, or some arbitrary (positive) vector $\boldsymbol{u}$. How these should be obtained in a way that will lead to a strong cut (i.e., a cut which is close to P_I in some meaningful way), is not part of the definition.

Cuts from basic feasible solutions

Recall that if $\hat{\boldsymbol{z}}$ is a basic feasible solution of P' with basis $\boldsymbol{B}$ and where B and N are the index sets of the basic and nonbasic variables, then, for $i \in B$,

$$\hat{z}_i + \sum_{j \in N} \bar{a}_{ij} \hat{z}_j = \bar{b}_i$$

is obtained as a linear combination of the equations in $\boldsymbol{A}'\boldsymbol{z} = \boldsymbol{b}$, with some multipliers $u_1, \ldots, u_m$. Thus, every basic variable in a basic feasible solution leads to a vector $\boldsymbol{u}$ of multipliers that we can use for Chvátal-Gomory or Gomory mixed integer cuts.

Let $\hat{\boldsymbol{x}}$ be the vertex of P corresponding to a basic feasible solution $\hat{\boldsymbol{z}}$ of P'. Then, as we have seen in the construction of the simplex algorithm, each of the d nonbasic variables in $\hat{\boldsymbol{z}}$ gives us a ray originating in $\hat{\boldsymbol{x}}$. Let C denote the cone defined by these d rays. Then inequalities that are valid for $C \cap \left\{ \boldsymbol{x} \in \mathbb{R}^d \ : \ x_k \leq \lfloor \hat{x}_k \rfloor \right\}$ and $C \cap \left\{ \boldsymbol{x} \in \mathbb{R}^d \ : \ x_k \geq \lfloor \hat{x}_k \rfloor + 1 \right\}$ are

also valid for P_I. Split cuts of this form are sometimes also called *intersection cuts from basic feasible solutions* [10] or *simple disjunctive cuts* [14].

Note that one can strengthen this last type of cuts by taking into account the integrality of the remaining variables, which is an alternative way of finding basic Gomory mixed integer cuts. This was shown by Balas and Jeroslow in [13].

Also note that we can get more (and possibly stronger) cuts if we replace basic feasible solutions by basic solutions that are not necessarily feasible.

We will come back to this approach of using basic feasible solutions for obtaining cuts in Section 2.3, where we give more details and illustrate the procedure with an example.

k-cuts

This family of cuts adds another twist to cuts from basic feasible solutions: Given a basic variable of such a solution, then, as we mentioned above, the corresponding row of the simplex-tableau is a linear combination of the original equations $\boldsymbol{A'z} = \boldsymbol{b}$. If we now multiply this row by a non-zero integer k, then the corresponding equation is still satisfied by all points in P', and we can obtain a Gomory mixed integer cut from it as we described before.

2.1.1 Comparing Elementary Closures

While it is certainly not obvious from the definitions, many of the above families define the same elementary closures.

Lemma 2.10 ([68])**.** $P_{\text{MIR}} = P_{\text{S}} = P_{\text{GMI}}$.

Proof. We will show the inclusions $\text{P}_{\text{MIR}} \subseteq P_{\text{S}} \subseteq P_{\text{GMI}} \subseteq P_{\text{MIR}}$.

$\text{P}_{\text{MIR}} \subseteq P_{\text{S}}$: Suppose $\boldsymbol{c}^\mathsf{T}\boldsymbol{x} \leq c_0$ is a split cut. By definition this means that there is an integer vector $(\boldsymbol{\pi}, \pi_0) \in \mathbb{Z}^{d+1}$ and $\alpha, \beta \geq 0$ such that

$$\begin{aligned} \boldsymbol{cx} - \alpha(\boldsymbol{\pi x} - \pi_0) &\leq c_0 \quad \text{and} \\ \boldsymbol{cx} + \beta(\boldsymbol{\pi x} - \pi_0 - 1) &\leq c_0 \end{aligned}$$

are valid for P. If $\alpha = \beta = 0$, then $\boldsymbol{c}^\mathsf{T}\boldsymbol{x} \leq c_0$ is valid for P and there is nothing to show. Otherwise we can multiply both inequalities by $\frac{1}{\alpha+\beta}$ and get

$$\begin{aligned} \frac{1}{\alpha+\beta}(\boldsymbol{c} - \alpha\boldsymbol{\pi})^\mathsf{T}\boldsymbol{x} &\leq \frac{c_0 - \alpha\pi_0}{\alpha+\beta} \quad \text{and} \\ \frac{1}{\alpha+\beta}(\boldsymbol{c} + \beta\boldsymbol{\pi})^\mathsf{T}\boldsymbol{x} &\leq \frac{c_0 + \beta\pi_0 + \beta}{\alpha+\beta}. \end{aligned} \tag{2.12}$$

Note that $\frac{1}{\alpha+\beta}(\boldsymbol{c} + \beta\boldsymbol{\pi} - (\boldsymbol{c} - \alpha\boldsymbol{\pi})) = \boldsymbol{\pi} \in \mathbb{Z}^d$ and the mixed integer rounding cut derived from the two above inequalities is again $\boldsymbol{c}^\mathsf{T}\boldsymbol{x} \leq c_0$.

$\text{P}_{\text{S}} \subseteq P_{\text{GMI}}$: Let $\hat{\boldsymbol{a}}$, $\hat{b}$, and f_i, $i = 0, \ldots, d+m$ be defined as before. We want to derive the cut

$$\sum_{f_i \leq f_0} f_i z_i + \frac{f_0}{1-f_0} \sum_{f_i > f_0} (1 - f_i) z_i \geq f_0$$

as a split cut. Define $\pi_0 = \left\lfloor \hat{b} \right\rfloor$ and for $i = 1, \ldots, d+m$

$$\boldsymbol{\pi}_i = \begin{cases} \lfloor \hat{\boldsymbol{a}}_i \rfloor & \text{if } f_i \leq f_0 \\ \lceil \hat{\boldsymbol{a}}_i \rceil & \text{if } f_i > f_0. \end{cases}$$

It is not hard to check that $\boldsymbol{\pi}\boldsymbol{z} \leq \pi_0$ together with $\hat{\boldsymbol{a}}^\mathsf{T}\boldsymbol{z} = \hat{b}$ implies the above inequality, and similarly for $\boldsymbol{\pi}\boldsymbol{z} \geq \pi_0 + 1$ together with $-\hat{\boldsymbol{a}}^\mathsf{T}\boldsymbol{z} = -\hat{b}$.

We can do the same for P instead of P'.

$\mathrm{P}_{\mathrm{GMI}} \subseteq P_{\mathrm{MIR}}$: Let $\boldsymbol{c}_1\boldsymbol{x} \leq c_0^1$ and $\boldsymbol{c}_2\boldsymbol{x} \leq c_0^2$ be two valid inequalities for P, with $\boldsymbol{\pi} := \boldsymbol{c}_2 - \boldsymbol{c}_1 \in \mathbb{Z}^d$.

Then we can find non-negative $\boldsymbol{\lambda}^1, \boldsymbol{\lambda}^2 \in \mathbb{R}^m$ such that $\boldsymbol{\lambda}^1\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{\lambda}^1\boldsymbol{b}$ dominates $\boldsymbol{c}_1\boldsymbol{x} \leq c_0^1$, and $\boldsymbol{\lambda}^2\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{\lambda}^2\boldsymbol{b}$ dominates $\boldsymbol{c}_2\boldsymbol{x} \leq c_0^2$.

Indeed, given a valid inequality (V) for P, there are rows of $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$ which dominate (V). Thus we can find a parallel hyperplane to the bounding hyperplane of (V) as a non-negative linear combination of the equalities $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, and then scale the coefficients such that they sum up to 1 (i.e., we find a convex combination). It is not hard to see that if the equalities with non-zero coefficients were tight for P (i.e., the inequalities are not strictly dominated by others of $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$), then the inequality given by this new hyperplane is also tight for P.

We will assume that the inequalities we are given are indeed equal to the ones we constructed as a convex combination of $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$.

Now define $\boldsymbol{u} = \boldsymbol{\lambda}^2 - \boldsymbol{\lambda}^1$ and note that $\boldsymbol{u}\boldsymbol{A}'\boldsymbol{z} = \boldsymbol{u}\boldsymbol{b}$ is satisfied by all points $\boldsymbol{z}$ in P'. Furthermore, let $\boldsymbol{s}$ be the vector of slack variables $(z_{d+1}, \ldots, z_{d+m})^\mathsf{T}$, and observe that

$$\boldsymbol{u}\boldsymbol{A}'\boldsymbol{z} = \boldsymbol{u}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{u}\boldsymbol{s} = (\boldsymbol{c}_2 - \boldsymbol{c}_1)\boldsymbol{x} + \boldsymbol{\lambda}^2\boldsymbol{s} - \boldsymbol{\lambda}^1\boldsymbol{s} \quad \text{and} \quad \boldsymbol{u}\boldsymbol{b} = c_0^2 - c_0^1.$$

Then the Gomory mixed integer cut we can derive from the equation $\boldsymbol{u}\boldsymbol{A}'\boldsymbol{z} = \boldsymbol{u}\boldsymbol{b}$ is, according to (2.7),

$$\boldsymbol{\lambda}^2\boldsymbol{s} + \frac{\gamma}{1-\gamma}\boldsymbol{\lambda}^1\boldsymbol{s} \geq \gamma,$$

where $\gamma = c_0^2 - c_0^1 - \left\lfloor c_0^2 - c_0^1 \right\rfloor$. Using $\boldsymbol{\lambda}^2\boldsymbol{s} = c_0^2 - c_0^1 + \boldsymbol{\lambda}^1\boldsymbol{s} - (\boldsymbol{c}_2 - \boldsymbol{c}_1)\boldsymbol{x}$ and $\boldsymbol{\lambda}^1\boldsymbol{s} = c_0^1 - \boldsymbol{c}_1\boldsymbol{x}$ (which follows from $\boldsymbol{s} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}$ and our assumptions on $\boldsymbol{\lambda}^1$), we transform this to

$$\begin{aligned} & (\boldsymbol{c}_2 - \boldsymbol{c}_1)\boldsymbol{x} - \boldsymbol{\lambda}^1\boldsymbol{s} - \frac{\gamma}{1-\gamma}\boldsymbol{\lambda}^1\boldsymbol{s} \leq \left\lfloor c_0^2 - c_0^1 \right\rfloor \\ \Leftrightarrow \quad & (\boldsymbol{c}_2 - \boldsymbol{c}_1)\boldsymbol{x} - \frac{1}{1-\gamma}(c_0^1 - \boldsymbol{c}_1\boldsymbol{x}) \leq \left\lfloor c_0^2 - c_0^1 \right\rfloor \\ \Leftrightarrow \quad & \boldsymbol{\pi}\boldsymbol{x} + (\boldsymbol{c}_1\boldsymbol{x} - c_0^1)\frac{1}{1-\gamma} \leq \left\lfloor c_0^2 - c_0^1 \right\rfloor. \end{aligned}$$

□

Lemma 2.11. *$P_{\mathrm{GMI}} \subseteq P_{\mathrm{CG}}$, and there are polytopes where the inclusion is strict.*

Proof. If $\boldsymbol{u} \in \mathbb{R}^m$ is nonnegative, then the Gomory mixed integer cut from (2.7) simplifies to

$$\sum_{i \le d; f_i \le f_0} f_i x_i + \frac{f_0}{1-f_0} \sum_{i \le d; f_i > f_0} (1-f_i) x_i + \sum_{j=1}^{m} u_j z_{d+j} \ge f_0.$$

If we now substitute $(z_{d+1}, \ldots, z_{d+m})^\mathsf{T} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}$ and multiply both sides of the inequality by $(1 - f_0)$, we get

$$(1-f_0) \sum_{i \le d; f_i \le f_0} f_i x_i + \sum_{i \le d; f_i > f_0} f_0(1-f_i) x_i + (1-f_0)(\boldsymbol{u}^\mathsf{T}\boldsymbol{b} - \boldsymbol{u}^\mathsf{T}\boldsymbol{A}\boldsymbol{x}) \ge f_0(1-f_0),$$

which in turn is equivalent to

$$\sum_{i \le d; f_i \le f_0} f_i x_i + \sum_{i \le d; f_i > f_0} f_0 x_i - \boldsymbol{u}^\mathsf{T}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{u}^\mathsf{T}\boldsymbol{b} - f_0\Big(\sum_{i \le d} f_i x_i - \boldsymbol{u}^\mathsf{T}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{u}^\mathsf{T}\boldsymbol{b}\Big) \ge f_0(1-f_0).$$

Then, since $\sum_{i \le d; f_i > f_0} f_0 x_i \le \sum_{i \le d; f_i > f_0} f_i x_i$, this implies the cut

$$\sum_{i \le d} f_i x_i - \boldsymbol{u}^\mathsf{T}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{u}^\mathsf{T}\boldsymbol{b} - f_0\Big(\sum_{i \le d} f_i x_i - \boldsymbol{u}^\mathsf{T}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{u}^\mathsf{T}\boldsymbol{b}\Big) \ge f_0(1-f_0),$$

and by dividing by $(1 - f_0)$ and by the definition of f_i for $i \ge 0$ we get indeed the Chvátal-Gomory cut $\left\lfloor \boldsymbol{u}^\mathsf{T}\boldsymbol{A} \right\rfloor \boldsymbol{x} \le \left\lfloor \boldsymbol{u}^\mathsf{T}\boldsymbol{b} \right\rfloor$.

To show that the inclusion is strict in some cases, define

$$\bar{P} = \{(x_1, x_2) \ : \ -2x_1 + x_2 \le 0,\ 2x_1 + x_2 \le 2,\ x_1, x_2 \ge 0\}.$$

A basic Gomory mixed integer cut we get from the optimal simplex tableau for $\max -x_1 + x_2$ is $z_3 + z_4 \ge 2$ or, in terms of the original variables, $x_2 \le 0$. In fact, it is not hard to see that therefore we have $\bar{P}_{GMI} = \bar{P}_I = \text{conv}\{(0,0),(1,0)\}$.

On the other hand, we will show that the point $(1/2, 1/2)$ satisfies any Chvátal-Gomory cut. Indeed, since $\bar{P}$ is a triangle with the vertices $(0,0)$, $(1,0)$, and $(1/2,1)$, any valid inequality for $\bar{P}$ must be valid for both $(0,0)$ and $(1,1)$, or both $(1,0)$ and $(0,1)$. Since the Chvátal-Gomory cut we derive from it will preserve validity for integer points, it will also be valid for the center point $(1/2, 1/2)$. □

2.1.2 A Non-Fulldimensional Example

Consider the integer program

$$\max\{x_1 \ : \ 5x_1 + 7x_2 = 23,\ x_1, x_2 \ge 0, x_1, x_2 \in \mathbb{Z}\},$$

see Figure 2.2.

If we now solve the linear relaxation of this program, the optimal solution is $(\frac{23}{5}, 0)$. Since x_2 is already integer, it is not unreasonable to try the disjunction $x_1 \le 4$ and $x_1 \ge 5$, which corresponds to the basic Chvátal-Gomory cut and also the basic Gomory mixed integer cut.

Thus, we add the inequality $x_1 \le 4$ to the program and solve again, which leads to the optimal point $(4, \frac{3}{7})$. Now x_2 is integer and we repeat the game with alternating roles, until we conclude infeasibility after adding 6 cuts.

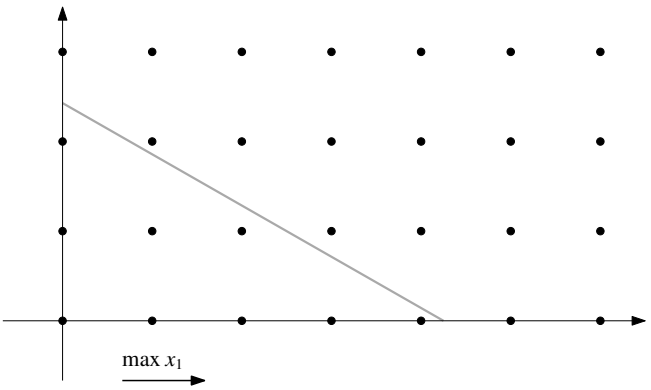

Figure 2.2: The linear relaxation of the program.

One can easily imagine an example that will require even more steps of this kind. What makes this program difficult to solve with cutting planes is the fact that we repeatedly find integer points that are close to the linear relaxation and thus prevent the more conservative methods from finding strong cuts.

However, since these integer points do not lie in the smallest affine space containing the linear relaxation of the program, we can exclude them from our consideration by applying the reformulation (2.1).

2.2 Cutting Planes in the Reformulation

Strictly speaking, the cuts we want to study in this chapter do not form a different family of cuts. Instead we describe an additional procedure, which we apply to the polytope before we use cuts from the above mentioned families. We first reformulate P in the lowest-dimensional affine subspace it is contained in, find cuts (by any method) in this reformulation, and then translate these cuts back to the original space.

The family of cuts we find by reformulating, finding all Chvátal-Gomory cuts in the new space, and translating them back to the original space, will be denoted by RCG, and the corresponding elementary closure by P_{RCG}. The elementary closure P_{RS} is defined analogously for split-cuts.

We restrict our attention to these two families of cuts, since we have seen in Lemma 2.10 that the elementary closures of Gomory mixed integer cuts and of mixed integer rounding cuts are the same as P_{S}.

We will now show that these elementary closures do not differ from the closures we obtain with these families without the reformulation. However, in Section 2.3 we will see that some of the most popular methods to derive cuts from these families do lead to cuts that will not be discovered using the same methods without the reformulation.

To show the equivalence of the elementary closures, we first observe the following.

Theorem 2.12 ([63])**.** *Let $\boldsymbol{B}$ be a basis of* $\ker_{\mathbb{Z}}(\boldsymbol{A})$*. Then there exists a matrix $W \in \mathbb{Z}^{(n-m)\times n}$ such that*

$$\boldsymbol{WB} = \boldsymbol{I}.$$

Proof. Let $\boldsymbol{U}$ be unimodular with $\boldsymbol{AU} = [\boldsymbol{H}, 0] = \mathrm{HNF}(\boldsymbol{A})$. Then the last $n - m$ columns of $\boldsymbol{U}$ form a basis of $\ker_{\mathbb{Z}}(\boldsymbol{A})$, call this basis $\hat{\boldsymbol{B}}$. We claim that for this basis the last $n - m$ rows of $\boldsymbol{U}^{-1}$ form the desired matrix $\hat{\boldsymbol{W}}$.

Indeed, $\boldsymbol{U}^{-1}$ and hence $\hat{\boldsymbol{W}}$ are integer matrices because $\boldsymbol{U}$ is unimodular, and $\hat{\boldsymbol{W}}\hat{\boldsymbol{B}} = \boldsymbol{I}$ by construction.

Now let $\boldsymbol{B}$ be any basis of $\ker_{\mathbb{Z}}(\boldsymbol{A})$. Then there is a unimodular matrix $\boldsymbol{V}$ with $\boldsymbol{B} = \hat{\boldsymbol{B}}\boldsymbol{V}$, and this directly implies that $\boldsymbol{W} = \boldsymbol{V}^{-1}\hat{\boldsymbol{W}}$ is a matrix with the desired properties. □

We will assume from now on that $\mathrm{HNF}(\boldsymbol{A}) = [\boldsymbol{I}, \boldsymbol{0}]$. If it is not, we can multiply both sides of the equation $\boldsymbol{Ax} = \boldsymbol{b}$ by $\boldsymbol{H}^{-1}$ to achieve it. Recall that this assumption is equivalent to the fact that the lattice generated by the rows of $\boldsymbol{A}$ is a pure sublattice of $\mathbb{Z}^n$.

Corollary 2.13. *Let* $P = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{Ax} = \boldsymbol{b}, \boldsymbol{x} \geq \boldsymbol{0}\}$ *with* $\boldsymbol{A} \in \mathbb{Z}^{m\times n}$ *and* $\boldsymbol{b} \in \mathbb{Z}^m$. *Then* $P_{\mathrm{CG}} = P_{\mathrm{RCG}}$.

Proof. Let $\boldsymbol{x} \in \mathbb{R}^n$ such that $\boldsymbol{Q\mu} = \boldsymbol{x} - \boldsymbol{x}^0$ for some $\boldsymbol{\mu} \in \mathbb{R}^{n-m}$, and $\boldsymbol{Q}$ and $\boldsymbol{x}^0$ as in (2.1). By Lemma 2.4 we may assume that $\boldsymbol{Q\mu} \geq -\boldsymbol{x}^0$ implies $\boldsymbol{\mu} \geq \boldsymbol{0}$, or conclude that (eq-IP) is unbounded.

We know from Theorem 2.12 that we can find an integer matrix $\boldsymbol{W}$ with $\boldsymbol{WQ} = \boldsymbol{I}$. Then

$$\boldsymbol{\mu} = \boldsymbol{I\mu} = \boldsymbol{WQ\mu} = \boldsymbol{W}(\boldsymbol{x} - \boldsymbol{x}^0),$$

where $\boldsymbol{W}$ and $\boldsymbol{Wx}^0$ are both integer.

Define $\hat{P} = \left\{\boldsymbol{\mu} \in \mathbb{R}^{n-m} : \boldsymbol{Q\mu} \geq -\boldsymbol{x}^0\right\}$ and let $\boldsymbol{dx} \leq d_0$ be valid for P, where $\boldsymbol{d}$ is integer. Note that this is the situation as described in Chapter 1.3.3. The cut $\boldsymbol{dx} \leq \lfloor d_0 \rfloor$ we derive from this is clearly equal to or dominated by a Chvátal-Gomory cut.

We now reformulate this inequality as

$$\boldsymbol{dQ\mu} \leq d_0 - \boldsymbol{dQx}^0,$$

where $\boldsymbol{dQ}$ is integer. Since $\boldsymbol{x}^0$ is also integer, we have

$$(d_0 - \boldsymbol{dQx}^0) - \left\lfloor d_0 - \boldsymbol{dQx}^0 \right\rfloor = d_0 - \lfloor d_0 \rfloor,$$

and thus the cut from the reformulated inequality is again dominated by a Chvátal-Gomory cut for $\hat{P}$.

Conversely, let $\boldsymbol{p\mu} \geq p_0$ be valid for $\hat{P}$ and such that $\boldsymbol{p}$ is integer. Then this is translated to

$$\boldsymbol{pWx} \geq p_0 + \boldsymbol{pWx}^0,$$

where again $\boldsymbol{pW}$ is integer and thus p_0 is the only possibly non-integer part on the right-hand side. □

Corollary 2.14. *Let* $P = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{Ax} = \boldsymbol{b}, \boldsymbol{x} \geq \boldsymbol{0}\}$ *with* $\boldsymbol{A} \in \mathbb{Z}^{m\times n}$ *and* $\boldsymbol{b} \in \mathbb{Z}^m$. *Then* $P_{\mathrm{S}} = P_{\mathrm{RS}}$.

Proof. Define again $\hat{P} = \left\{\boldsymbol{\mu} \in \mathbb{R}^{n-m} : \boldsymbol{Q\mu} \geq -\boldsymbol{x}^0\right\}$. As in the previous proof, we can use $\boldsymbol{Q}$ and $\boldsymbol{W}$ to translate the valid inequalities between the spaces. Thus, every split cut for P is a split cut for $\hat{P}$ and vice versa. □

2.3 Basic Cuts in the Reformulation

As was mentioned before, when we are asked to find a cut we are confronted with the challenge of how to choose the initial inequalities and possibly additional parameters like scaling-factors. The method of using basic feasible solutions has been applied quite successfully, and therefore we will use it to see how the reformulation-method described above can improve it.

In basic feasible solutions for P', the variables can have fractional coefficients which we can then use to derive, e.g., Chvátal-Gomory cuts or Gomory mixed integer cuts.

Note that, as we mentioned before, these cuts will always be split cuts. In particular, Chvátal-Gomory cuts can be expressed as split cuts, where one of the sets S_1 and S_2 is empty.

If we now do the reformulation first, and then find basic feasible solutions, we can find splits which, when translated back, are not cuts from basic feasible solutions as we described them before.

Interestingly, it turns out that they are k-cuts. Since there is no method known to determine useful values k from the original formulation, this observation provides a positive spin on the results of Cornuéjols, Li, and Vandenbussche [28], where it was, roughly speaking, shown that using k-cuts instead of just the standard Gomory mixed integer cuts will only lead to an improvement in half of the cases.

We will now describe the procedure of deriving cuts from basic cuts in the reformulation in more detail, where we illustrate each step with the integer program

$$\max\{x_1 \ : \ 16x_1 + 25x_2 + 47x_3 = 182, \boldsymbol{x} \geq \mathbf{0}\}.$$

Applying the reformulation first, we get $(1,\ 1,\ 3)^\mathsf{T}$ as our initial solution as described in the beginning of this chapter, and three inequalities corresponding to the nonnegativity:

$$\begin{aligned} -\bar{\mu}_1 + 9\bar{\mu}_2 &\geq -1 \\ -5\bar{\mu}_1 - 2\bar{\mu}_2 &\geq -1 \\ 3\bar{\mu}_1 - 2\bar{\mu}_2 &\geq -3 \end{aligned}$$

By chance this initial solution is feasible, which we did not require and will not make use of.

To make sure that we can apply all our cut-procedures, all points satisfying the new constraints must lie in the positive orthant. This can be achieved, for instance, by using the LP-relaxation and finding the minimum μ_i^* for each of the coordinates μ_i (see Lemma 2.4). If we now shift the program by $-(\lfloor\mu_1^*\rfloor, \ldots, \lfloor\mu_{n-m}^*\rfloor)^\mathsf{T}$, it will indeed lie in the positive orthant.

Doing this shifting in the example, we get $\mu_1 = \bar{\mu}_1 + 2$ and $\mu_2 = \bar{\mu}_2 + 1$, and our system becomes (see Figure 2.3)

$$\begin{aligned} -\mu_1 + 9\mu_2 &\geq 6 \\ -5\mu_1 - 2\mu_2 &\geq -13 \\ 3\mu_1 - 2\mu_2 &\geq 1 \end{aligned}$$

Next, we apply the simplex method to find a basic feasible solution. For the example the optimal tableau looks as follows:

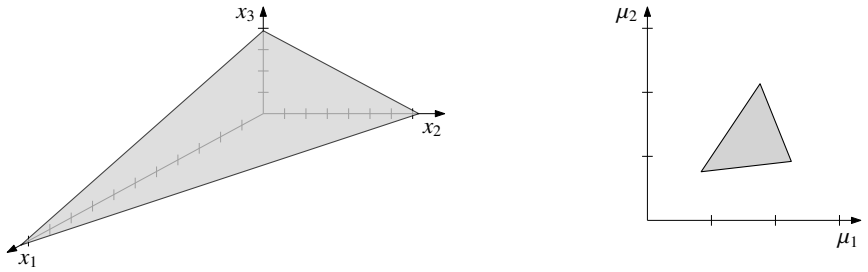

Figure 2.3: The example in the original space, and after reformulation and shift.

basis	$\bar{b}$	μ_1	μ_2	s_1	s_2	s_3
z	$\frac{139}{8}$				$-\frac{25}{16}$	$-\frac{47}{16}$
μ_1	$\frac{7}{4}$	1			$\frac{1}{8}$	$-\frac{1}{8}$
μ_2	$\frac{17}{8}$		1		$\frac{3}{16}$	$\frac{5}{16}$
s_1	$\frac{91}{8}$			1	$\frac{25}{16}$	$\frac{47}{16}$

Observe that there is a 1-to-1 correspondence between the variables $\boldsymbol{x}$ in the original space and the slack-variables $\boldsymbol{s}$, because both are defined as $\boldsymbol{Q\mu} + \bar{\boldsymbol{x}}^0$, where $\bar{\boldsymbol{x}}^0$ is the shifted initial solution. Moreover, because of the shift, all μ_i will be positive and thus basic in the optimal tableau. The remaining $n - m$ basic variables will be corresponding to basic variables in the original space. In our example, this is x_1 by our choice of objective function.

As a consequence, any cut generated from a row corresponding to a basic slack variable will be identical to the cut generated from the corresponding row in the tableau of the original space. In the example, the equation we get from the tableau is

$$s_1 + \frac{25}{16}s_2 + \frac{47}{16}s_3 = \frac{91}{8}.$$

To get the Gomory mixed integer cut, we compute the fractional parts $f_0 = \frac{3}{8}$, $f_1 = 0$, $f_2 = \frac{9}{16}$, and $f_3 = \frac{15}{16}$. Then if we enter this into Inequality (2.6), we get the cut $7s_2 + s_3 \geq 10$, or $7x_2 + x_3 \geq 10$.

Therefore, we are most interested in the Gomory mixed integer cuts we can get from the rows where the μ_i are basic. In our example, we get $s_2 + 3s_3 \geq 6$ from the μ_1-row, and $13s_2 + 11s_3 \geq 14$ from the μ_2-row. And again, we can just express the same cuts in x-variables: $x_2 + 3x_3 \geq 6$ and $13x_2 + 11x_3 \geq 14$.

Let us look at the cut from the μ_1-row more closely. If we translate it into the μ-variables, we get $-\mu_1 + 2\mu_2 \leq 1$. As a split cut, the inequalities are

$$\mu_1 \leq 1 \quad \text{and} \quad \mu_1 \geq 2,$$

see Figure 2.4.

Using the translation

$$\bar{\mu}_1 \;=\; -23 + 2x_1 + 3x_2 + 6x_3$$

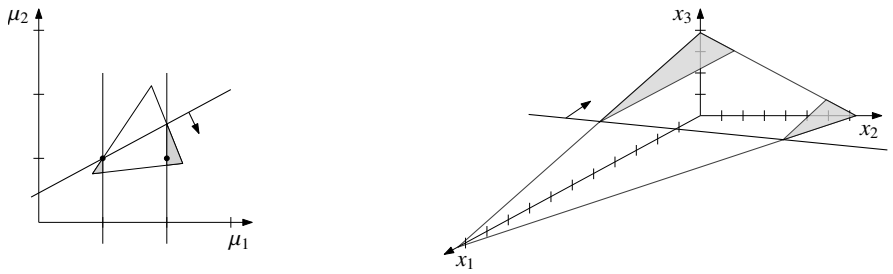

Figure 2.4: The cut $-\mu_1 + 2\mu_2 \leq 1$ and the corresponding split, and the cut and split translated to the original space. The dots mark the feasible integer solutions.

we get the corresponding split-inequalities

$$2x_1 + 3x_2 + 6x_3 \leq 22 \quad \text{and} \quad 2x_1 + 3x_2 + 6x_3 \geq 23$$

in the original variables.

We could also use this translation, together with $\bar{\mu}_2 = 58 - 5x_1 - 8x_2 - 15x_3$, to express the μ_1-cut in the form $12x_1 + 19x_2 + 36x_3 \geq 138$, but this is equivalent to the inequality $x_2 + 3x_3 \geq 6$, which we were able to read directly from the tableau of the reformulation.

We can recover the multiplier $\boldsymbol{u}$ that leads to this cut as a Gomory mixed integer cut, see Balas [10] and also, e.g., Fischetti, Lodi, and Tramontani [35]. In this example, we can certify that we have found a Gomory mixed integer cut by multiplying the original equation by $1/8$. Indeed, this leads to fractional parts $f_0 = 6/8$, $f_1 = 0$, $f_2 = 1/8$, and $f_3 = 7/8$.

While this shows that we can indeed find much stronger cuts by applying this reformulation-technique, there are also some signs for caution: Note, for example, that we only looked at one row of the simplex tableau in the reformulation. Repeating the same steps for the μ_2-row reveals that this is a Chvátal-Gomory cut and is dominated by the cut we get from the s_1-row, see Figure 2.5. Thus, a cut derived from the reformulation is not automatically stronger and might in fact actually be weaker than basic cuts generated from the original formulation.

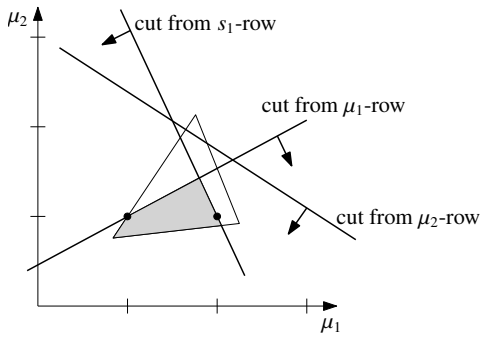

Figure 2.5: The cuts generated from the reformulation for $\max x_1$.

However, we are confident that this approach is worth pursuing further and can be developed into a method that can be applied as a heuristic to find strong cuts, in particular for the cases in which classical methods turn out to be slow.

This is based also on a (very) small number of experiments we performed: Using eight knapsack-instances, five with 5 and three with 3 variables, we compare the average of how much of the gap we close by using Chvátal-Gomory cuts. Since this comparison is somewhat unfair for the original formulation, where we have only one row and thus one cut, we run up to 5 rounds of cuts in the original space, and a single round in the reformulation.

In Table 2.1 we record the average of the fraction of the gap closed, and the sum of all cuts (over all 8 instances). In the reformulation, we close an average of 38% of the gap and have a total of 34 cuts.

Table 2.1: Performance of Chvátal-Gomory cuts in the original space.

# of rounds	1	2	3	4	5
% Gap closed	12	19	38	63	91
# of cuts	8	20	41	81	135

2.4 Notes

While in Theorem 2.12 shows the existence of an integer matrix that lets us translate back and forth between the coordinates given in $\mathbb{R}^n$ and the coordinates given in the basis $\boldsymbol{Q}$, it does not provide us with a direct way to find such a matrix. First, we observe that the matrix can be chosen such that it has polynomial size.

Proposition 2.15 ([63])**.** *We can compute a matrix $\boldsymbol{W}$ as in Theorem 2.12 in polynomial time and the size of its entries can be polynomially bounded in the number of bits required to store $\boldsymbol{A}$ and $\boldsymbol{B}$.*

Proof. The entries of $\boldsymbol{U}$ (and thus also of $\boldsymbol{U}^{-1}$ and $\hat{\boldsymbol{W}}$) are clearly polynomially bounded in the number of bits required to store $\boldsymbol{A}$.

Next we observe that the matrix $\boldsymbol{W}$ is not unique, since

$$(\boldsymbol{W} + [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_{n-m}]^\mathsf{T})\boldsymbol{B} = \boldsymbol{I}$$

for any $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{n-m}$ in the lattice $\{\boldsymbol{x} \in \mathbb{Z}^n \ : \ \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$.

Let $\boldsymbol{w}_i$ be the i^{th} row of the matrix $\boldsymbol{W}$ we constructed in the proof of Theorem 2.12, and let $\boldsymbol{\pi}_i = (\boldsymbol{A}^\mathsf{T}\boldsymbol{A})^{-1}\boldsymbol{A}^\mathsf{T}\boldsymbol{w}_i^\mathsf{T}$. Note that

$$\boldsymbol{w}_i^\mathsf{T} = [\boldsymbol{I} - \boldsymbol{A}(\boldsymbol{A}^\mathsf{T}\boldsymbol{A})^{-1}\boldsymbol{A}^\mathsf{T}]\boldsymbol{w}_i + \boldsymbol{A}\boldsymbol{\pi}_i = (\boldsymbol{B}(\boldsymbol{B}^\mathsf{T}\boldsymbol{B})^{-1}\boldsymbol{B}^\mathsf{T})\boldsymbol{w}_i^\mathsf{T} + \boldsymbol{A}\boldsymbol{\pi}_i.$$

Set $\hat{\boldsymbol{\pi}}_i = \boldsymbol{\pi}_i - \lceil \boldsymbol{\pi}_i \rfloor$ and $\tilde{\boldsymbol{w}}_i^\mathsf{T} = \boldsymbol{w}_i^\mathsf{T} - \boldsymbol{A}\lceil \boldsymbol{\pi}_i \rfloor = (\boldsymbol{B}(\boldsymbol{B}^\mathsf{T}\boldsymbol{B})^{-1}\boldsymbol{B}^\mathsf{T})\boldsymbol{w}_i^\mathsf{T} + \boldsymbol{A}\hat{\boldsymbol{\pi}}_i$. Then

$$\begin{aligned} \|\tilde{\boldsymbol{w}}_i\|^2 &= |(\boldsymbol{B}(\boldsymbol{B}^\mathsf{T}\boldsymbol{B})^{-1}\boldsymbol{e}_i + \boldsymbol{A}\hat{\boldsymbol{\pi}}_i)^\mathsf{T}(\boldsymbol{B}(\boldsymbol{B}^\mathsf{T}\boldsymbol{B})^{-1}\boldsymbol{e}_i + \boldsymbol{A}\hat{\boldsymbol{\pi}}_i)| \\ &= |\boldsymbol{e}_i^\mathsf{T}(\boldsymbol{B}^\mathsf{T}\boldsymbol{B})^{-1}\boldsymbol{B}^\mathsf{T}\boldsymbol{B}(\boldsymbol{B}^\mathsf{T}\boldsymbol{B})^{-1}\boldsymbol{e}_i + \boldsymbol{B}^\mathsf{T}\boldsymbol{A} + \boldsymbol{A}^\mathsf{T}\boldsymbol{B} + \hat{\boldsymbol{\pi}}_i^\mathsf{T}\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\hat{\boldsymbol{\pi}}_i| \\ &= |\boldsymbol{e}_i^\mathsf{T}(\boldsymbol{B}^\mathsf{T}\boldsymbol{B})^{-1}\boldsymbol{e}_i + \hat{\boldsymbol{\pi}}_i^\mathsf{T}\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\hat{\boldsymbol{\pi}}_i| \end{aligned}$$

and the cardinality of the elements in $\hat{\boldsymbol{\pi}}_i$ are bounded by $1/2$.

The polynomial runtime for finding such a $\tilde{\boldsymbol{W}}$ is then clear from the proof of Theorem 2.12. □

However, there is also another way to aquire $\boldsymbol{W}$, which may be considered a bit more direct: Since $\ker_{\mathbb{Z}}(\boldsymbol{A})$ is a pure sublattice of $\mathbb{Z}^n$ and $\boldsymbol{Q}$ is a lattice basis of $\ker_{\mathbb{Z}}(\boldsymbol{A})$ we know that $\mathrm{HNF}(\boldsymbol{Q}^\mathsf{T}) = [\boldsymbol{I}, \boldsymbol{0}]$ (see Lemma 1.3), and therefore we find, in polynomial time, a unimodular matrix $\boldsymbol{U}$ such that

$$\boldsymbol{U}^\mathsf{T}\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{I} \\ \boldsymbol{0} \end{bmatrix}.$$

The first $n - m$ rows of $\boldsymbol{U}^\mathsf{T}$ form the desired matrix $\boldsymbol{W}$.

Chapter Three
Ellipsoidal Basis Reduction

Given the integer program

$$\max\left\{\boldsymbol{c}^{\mathsf{T}}\boldsymbol{x} \;:\; \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},\ \boldsymbol{x} \geq \boldsymbol{0}, \boldsymbol{x} \in \mathbb{Z}^n\right\}, \tag{eq-IP}$$

and performing the reformulation

$$\max\left\{\boldsymbol{c}^{\mathsf{T}}(\boldsymbol{x}_0 + \boldsymbol{Q}\boldsymbol{\mu}) \;:\; \boldsymbol{Q}\boldsymbol{\mu} \geq -\boldsymbol{x}_0, \boldsymbol{\mu} \in \mathbb{Z}^{n-m}\right\}. \tag{3.1}$$

as we described it in Chapter 2, we notice a potential difficulty:

When reducing the basis of $\ker_{\mathbb{Z}}(\boldsymbol{A})$ we do not take into account the shape of the feasible region of the relaxation $P = \{\boldsymbol{x} \in \mathbb{R}^n \;:\; \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \geq 0\}$.

As long as the vertices of P have more or less the same distance to each other, i.e., P is more or less regular, not looking at the shape of P does not pose any problems. Potentially, however, we could get an irregular shape of P, as shown in Figure 3.1. When we now reduce the basis to identify directions for cuts or for branching, this can lead to undesirable effects: While the basis has a long vector in some direction, the polytope is flat in a different one. Thus, the effect one hopes to obtain by identifying long basis vectors will not occur.

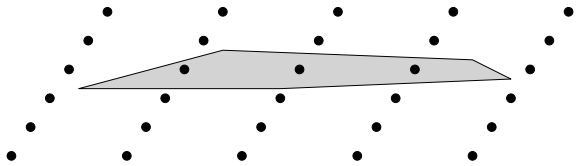

Figure 3.1: A polytope with properties reversing the effect of the reformulation.

Instead of looking at the lattice alone, we will now also take the general shape of P into account, as was already suggested by Lenstra in [58].

Before we treat the general case, we briefly look at an idea suggested by Aardal and Wolsey [6] for the case where $\boldsymbol{A}$ has only one row $\boldsymbol{a}$ and all elements a_i are positive. Aardal and Wolsey applied a map $\boldsymbol{D}_1$ to the standard lattice and to the polytope

$$P = \{\boldsymbol{x} \in \mathbb{R}^n \;:\; \boldsymbol{a}\boldsymbol{x} = b,\ \boldsymbol{x} \geq \boldsymbol{0}\},$$

which is a simplex, such that the mapped simplex

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

intersects the coordinate axes at the point $b\boldsymbol{e}_j$, $j = 1, \ldots, n$, yielding a perfectly regular polytope. The map $\boldsymbol{D}_1$ is given by the diagonal matrix

$$\boldsymbol{D}_1 = \begin{pmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_n \end{pmatrix}. \tag{3.2}$$

Under this map, the lattice $\ker_{\mathbb{Z}}(\boldsymbol{a})$ becomes

$$\ker_{\boldsymbol{D}_1\mathbb{Z}}(\mathbf{1}) = \boldsymbol{D}_1 \ker_{\mathbb{Z}}(\boldsymbol{a}) = \left\{ \boldsymbol{x} \in \boldsymbol{D}_1\mathbb{Z}^n \ : \ \sum_{i=1}^{n} x_i = 0 \right\}.$$

Reducing a basis for the lattice $\ker_{\boldsymbol{D}_1\mathbb{Z}}(\mathbf{1})$ with respect to the Euclidean norm is equivalent to reducing a basis for $\ker_{\mathbb{Z}}(\boldsymbol{a})$ with respect to the norm $\|\boldsymbol{x}\|^2_{\boldsymbol{D}_1} = \sum_{i=1}^n a_i^2 x_i^2$. Computations indicate that reducing a basis for $\ker_{\mathbb{Z}}(\boldsymbol{a})$ with respect to the norm $\|\boldsymbol{x}\|^2_{\boldsymbol{D}_1} = \sum_{i=1}^n a_i^2 x_i^2$ rather than the Euclidean norm yields a significant reduction in the number of branch-and-bound nodes needed to solve randomly generated instances. Selected results for instances with $n = 100,\ 200,\ 300$ are given in Table 3.1. The elements a_i are generated uniformly at random from the integers in the interval $[15,000,\ 150,000]$. The first column in the table gives the instance number. In columns 2–4 we report on the number of branch-and-bound nodes needed to solve the instance for the original formulation (eq-IP), the reformulation (3.1) in which $\boldsymbol{Q}$ is reduced with respect to the Euclidean norm, and the reformulation in which $\boldsymbol{Q}$ is reduced with respect to the norm $\|\boldsymbol{x}\|^2_{\boldsymbol{D}_1}$.

Table 3.1: The number of branch-and-bound nodes needed to solve the various reformulations.

n	eq-IP	(3.1) classical	(3.1) $\boldsymbol{D}_1$-norm
100_1	191,968	2,077	130
100_2	104,367	1,880	103
200_1	94,761	597	47
300_1	48,146	1,230	710

If the matrix $\boldsymbol{A}$ consists of more than one row we do not know the extreme points of the corresponding polytope P explicitly. The result we will treat further in this chapter can roughly be described as follows:

> *One can use an ellipsoid of maximum volume inside P to transform the original set of inequalities to a set of inequalities that yields a regular description of P.*

We will of course need to define what we mean precisely by these notions. First, let $\boldsymbol{M}$ be a $d \times d$ non-singular matrix, and $\boldsymbol{t} \in \mathbb{R}^d$. Then the mapping $T(\boldsymbol{x}) = t + \boldsymbol{M} \cdot \boldsymbol{x}$ is called an *affine transformation*. Note that, since $\boldsymbol{M}$ is non-singular, T^{-1} is well-defined and again an affine transformation. Let $S = B(0,1)$ be the unit ball, then if T is an affine transformation, $T(S)$ is called an *ellipsoid*. Define $\boldsymbol{D} = \boldsymbol{M}\boldsymbol{M}^{\mathsf{T}}$. We observe that we can describe $T(S)$ alternatively

the other way around: instead of mapping the elements of S, we collect all points that have an inverse under T with norm at most 1, i.e.,

$$T(S) = \text{ell}(\boldsymbol{t}, \boldsymbol{D}) = \left\{\boldsymbol{x} \in \mathbb{R}^d \ : \ (\boldsymbol{x} - \boldsymbol{t})^\mathsf{T} \boldsymbol{D}^{-1} (\boldsymbol{x} - \boldsymbol{t})^\mathsf{T} \leq 1\right\}.$$

Note that from this expression we can easily derive $\text{vol}(T(S)) = \sqrt{\det(\boldsymbol{D})} \cdot \text{vol}(S)$, and, since $\boldsymbol{M}$ is non-singular, $\boldsymbol{D}$ is a positive definite matrix.

The above observations on the shape of a compact convex set K, relative to a lattice L, are very much connected to the *lattice-width* of K, which is defined as

$$\min_{\boldsymbol{c} \in L^\dagger \setminus \{\boldsymbol{0}\}} \left\{\max\left\{\left\lfloor \boldsymbol{c}^\mathsf{T} \boldsymbol{x} \right\rfloor \ : \ \boldsymbol{x} \in K\right\} - \min\left\{\left\lceil \boldsymbol{c}^\mathsf{T} \boldsymbol{x} \right\rceil \ : \ \boldsymbol{x} \in K\right\}\right\},$$

where $L^\dagger$ is the dual lattice to L (see Chapter 1). This tells us the smallest number of lattice hyperplanes we need in order to cover $K \cap L$: One more than its width. Note that in contrast to the definition given in much of the literature, we here define the lattice width to be integer for any compact convex set.

It should be mentioned that we will make no direct use of the lattice width. It is given here simply because it is a guiding concept one should keep in mind in the following considerations.

As we will see, finding a direction in which we only need a small amount of lattice slices to cover the lattice points in a polytope P is of key importance in the proof below for the seminal theorem of H.W. Lenstra, Jr. [58], which demonstrates that integer programs can be solved in polynomial time if the dimension is fixed.

In the original proof, Lenstra constructs a simplex of positive volume inside of P which contains integer points if and only if P contains integer points. The advantage over the complete description of P is, that a simplex has the minimal amount of facets and vertices for its dimension, while still being full-dimensional.

However, Lovász noticed a different way of proving this (see [44, 75]), combining basis reduction with the ellipsoid method. Thus, before we have a closer look at this proof, we briefly review the ellipsoid method.

3.1 The Ellipsoid Method

This method was developed by Shor [78, 79] and Yudin and Nemirovski [83, 84], and gives us an alternative way to solve linear programs. It was refined and, most importantly, shown to have polynomial runtime by Khachiyan [52, 51]. The ellipsoid method solves the *(linear) feasibility program*: Given $\boldsymbol{A}$ and $\boldsymbol{b}$, find a vector $\boldsymbol{y}$ with $\boldsymbol{A}\boldsymbol{y} \leq \boldsymbol{b}$, or conclude that no such vector exists. However, we will see in Lemma 3.3 that this method can be adapted to solve the optimization program (LP) in polynomial time as well.

Recall that, as we mentioned before, the simplex method for (LP) the way it is commonly used is not a polynomial time algorithm. Surprisingly, however, there is no implementation known so far for the ellipsoid method with an average running-time anywhere near competitive to the simplex method.

Let $P = \left\{\boldsymbol{x} \in \mathbb{R}^d \;:\; \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}\right\}$ be a polyhedron and let us assume, for the moment, that P is bounded and full-dimensional. We will determine a sequence of ellipsoids $E_0, E_1, E_2, \ldots$ with decreasing volume, such that $P \subseteq E_i$ for every i. We stop when we find that the center $\boldsymbol{z}_i$ of E_i lies in P.

Let $v = 4d^2\phi$, where ϕ is the maximal row size of $[\boldsymbol{A}, \boldsymbol{b}]$. It can be shown that each vertex of P has size at most v (see [75], Theorem 10.2). Then, if we set $R = 2^v$, we get $P \subseteq \left\{\boldsymbol{x} \in \mathbb{R}^d \;:\; \|\boldsymbol{x}\| \leq R\right\} =: E_0$.

Suppose now $E_i \supseteq P$ has been determined. If its center $\boldsymbol{z}_i$ satisfies all inequalities in $\boldsymbol{A}\boldsymbol{z}_i \leq \boldsymbol{b}$, we stop. Otherwise, let $\boldsymbol{a}_k\boldsymbol{x} \leq b_k$ be an inequality violated by $\boldsymbol{z}_i$. Then we construct E_{i+1} as the ellipsoid of smallest volume containing the half-ellipsoid $E_i \cap \left\{\boldsymbol{x} \in \mathbb{R}^d \;:\; \boldsymbol{a}_k\boldsymbol{x} \leq \boldsymbol{a}_k\boldsymbol{z}_i\right\}$. Note that again we have $P \subseteq E_{i+1}$, since $P \subseteq E_i$ and

$$P \subseteq \left\{\boldsymbol{x} \in \mathbb{R}^d \;:\; \boldsymbol{a}_k\boldsymbol{x} \leq b_k\right\} \subseteq \left\{\boldsymbol{x} \in \mathbb{R}^d \;:\; \boldsymbol{a}_k\boldsymbol{x} \leq \boldsymbol{a}_k\boldsymbol{z}_i\right\}.$$

We will show that the volume of E_{i+1} is smaller than the volume of E_i, by at least a factor depending only on the dimension d. Since P was assumed to be full-dimensional, it has positive volume. Therefore we can already conclude that there is a finite N such that the center $\boldsymbol{z}_N$ of E_N lies in P. We will show below how to bound it polynomially in the size of $\boldsymbol{A}$ and $\boldsymbol{b}$.

First, however, we describe how to find an ellipsoid of smallest volume containing a given half of an ellipsoid, and also how to bound its volume.

Theorem 3.1 (see, e.g., [75])**.** *Let $E = \mathrm{ell}(\boldsymbol{z}, \boldsymbol{D}) \subseteq \mathbb{R}^d$, and $\boldsymbol{a} \in \mathbb{R}^d$. Then there is a unique ellipsoid $E' = \mathrm{ell}(\boldsymbol{z}', \boldsymbol{D}')$ containing $E \cap \left\{\boldsymbol{x} \in \mathbb{R}^d \;:\; \boldsymbol{a}\boldsymbol{x} \leq \boldsymbol{a}\boldsymbol{z}\right\}$ such that E' has the smallest possible volume. More specifically,*

$$\begin{aligned} \boldsymbol{z}' &= \boldsymbol{z} - \frac{1}{d+1} \cdot \frac{\boldsymbol{D}\boldsymbol{a}}{\sqrt{\boldsymbol{a}^\mathsf{T}\boldsymbol{D}\boldsymbol{a}}}, \\ \boldsymbol{D}' &= \frac{d^2}{d^2-1}\left(\boldsymbol{D} - \frac{2}{d+1} \cdot \frac{\boldsymbol{D}\boldsymbol{a}\boldsymbol{a}^\mathsf{T}\boldsymbol{D}}{\boldsymbol{a}^\mathsf{T}\boldsymbol{D}\boldsymbol{a}}\right). \end{aligned} \tag{3.3}$$

Moreover,

$$\frac{\mathrm{vol}\, E'}{\mathrm{vol}\, E} < e^{-\frac{1}{2d+2}}. \tag{3.4}$$

Proof. It is not hard to check that for $E = \mathrm{ell}(\boldsymbol{0}, \boldsymbol{I})$ the ellipsoid E' as given in (3.3) is indeed the unique ellipsoid containing half the unit ball split along $\boldsymbol{a}\boldsymbol{x} = 0$, with minimal volume.

The uniqueness and construction for general ellipsoids follows from this, since any ellipsoid is an affine transformation of the unit ball and such transformations preserve set-inclusions.

What is left to show is the bound on the volume. Note that $\mathrm{vol}(\mathrm{ell}(\hat{\boldsymbol{z}}, \hat{\boldsymbol{D}})) = \sqrt{\det(\hat{\boldsymbol{D}})} \cdot \mathrm{vol}(\mathrm{ell}(\boldsymbol{0}, \boldsymbol{I}))$ implies

$$\frac{\mathrm{vol}\, E'}{\mathrm{vol}\, E} = \sqrt{\frac{\det(\boldsymbol{D}')}{\det(\boldsymbol{D})}}.$$

We can again assume that $E = \mathrm{ell}(\boldsymbol{0}, \boldsymbol{I})$, and therefore

$$\sqrt{\frac{\det(\boldsymbol{D}')}{\det(\boldsymbol{D})}} = \sqrt{\det(\boldsymbol{D}')} = \left(\frac{d^2}{d^2-1}\right)^{d/2} \cdot \left(\det\left(\boldsymbol{I} - \frac{2}{d+1} \cdot \frac{\boldsymbol{a}\boldsymbol{a}^\mathsf{T}}{\boldsymbol{a}^t\boldsymbol{a}}\right)\right)^{1/2},$$

by the definition of $\boldsymbol{D}'$. Note that the matrix $\frac{\boldsymbol{a}\boldsymbol{a}^\mathsf{T}}{\boldsymbol{a}^\mathsf{T}\boldsymbol{a}}$ has 1 as its unique nonzero eigenvalue, and thus

$$\sqrt{\det(\boldsymbol{D}')} = \left(\frac{d^2}{d^2-1}\right)^{d/2}\left(1-\frac{2}{d+1}\right)^{1/2} = \left(\frac{d^2}{d^2-1}\right)^{(d-1)/2}\left(\frac{d}{d+1}\right).$$

Finally, we use that $1+q < e^q$ for all $q \neq 0$, and therefore

$$\frac{d^2}{d^2-1} = 1+\frac{1}{d^2-1} < e^{1/(d^2-1)} \quad \text{and} \quad \frac{d}{d+1} = 1-\frac{1}{d+1} < e^{-1/(d+1)}.$$

Thus,

$$\frac{\operatorname{vol} E'}{\operatorname{vol} E} < e^{(d-1)/(2d^2-2)}e^{-1/(d+1)} = e^{-\frac{1}{2d+2}}.$$

□

As we started the algorithm with E_0 of volume at most $(2R)^d$, we can conclude by induction that

$$\operatorname{vol} E_i \leq e^{-\frac{i}{2(d+1)}} \cdot (2R)^d.$$

On the other hand, if $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_d$ are affinely independent vertices of P, then

$$\operatorname{vol}(P) \geq \operatorname{vol}(\operatorname{conv}(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_d)) = \frac{1}{d!}\left|\det\begin{pmatrix} 1 & \cdots & 1 \\ \boldsymbol{x}_0 & & \boldsymbol{x}_d \end{pmatrix}\right|.$$

Since for each $i = 0, \ldots, d$ the vertex $\boldsymbol{x}_i$ has size at most v, we know that the determinant has denominator at most 2^{dv}.Therefore, we can conclude that

$$\operatorname{vol}(P) \geq d^{-d}2^{-dv} \geq 2^{-2dv}.$$

Suppose we have not found a center $\boldsymbol{z}_i \in P$ for some $i \leq N$ and $N = 16d^2v$. If P is nonempty, this implies

$$2^{-2dv} \leq \operatorname{vol}(P) \leq \operatorname{vol}(E_N) < e^{-N/(2d+2)} \cdot (2R)^d \leq 2^{-2dv},$$

which is a contradiction.

Therefore, if P is nonempty, we find a feasible $\boldsymbol{z}_i$ before reaching E_N. As N is polynomially bounded in the sizes of $\boldsymbol{A}$ and $\boldsymbol{b}$, we will find a feasible point in polynomial time, if it exists.

There are some technicalities we glossed over here. The first one is that we assumed that we can calculate with infinite precision for the construction of E' in Theorem 3.1. As this included calculations with square-roots, it is not a realistic assumption for actual computations. There are ways to approximate with sufficient precision, see, e.g., [71] (Section 8.7.4). Going into these details here, however, would lead us too far away from the main topic.

Second, we assumed that P is full-dimensional and bounded. *Detecting* whether P is full-dimensional is easy to achieve: If we reach E_N without finding a center $\boldsymbol{z}_i$ in P, then P is not full-dimensional.

However, this does not give us a direct indication of how to proceed, thus we need more technicalities. As is, e.g., described in [75] (Section 13.4), we can "inflate" P by a tiny amount to make it full-dimensional. Given a solution for the inflated version of P, we can then find one for the original version. To ensure boundedness, one can show that P is nonempty if and only if we find a feasible point inside a "very large" sphere (whose radius is polynomially bounded by the sizes of $\boldsymbol{A}$ and $\boldsymbol{b}$). See, e.g., [75] Chapters 14 and 15 for more details.

In particular, we get the following refinement of the idea of finding a sequence of ellipsoids of decreasing volume, all containing P:

Theorem 3.2 (See [75] p.206)**.** *If $P = \{\boldsymbol{x} \in \mathbb{R}^n \ : \ \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}\}$ is full-dimensional and bounded, and a rational γ with $0 < \gamma < 1/n$ is given, then there is an algorithm that finds an ellipsoid* $\text{ell}(z, \boldsymbol{D})$ *such that $\boldsymbol{z}$ is rational and* $\text{ell}(z, \gamma^2\boldsymbol{D}) \subseteq P \subseteq \text{ell}(z, \boldsymbol{D})$*, in time polynomially bounded by n, $\frac{1}{1-\gamma n}$, and the sizes of $\boldsymbol{A}$, $\boldsymbol{b}$, and γ.*

Lemma 3.3 (see [75], Corollary 14.1a)**.** *If we can solve the linear feasibility program in polynomial time, then we can also solve the optimization program* (LP) *in polynomial time.*

Proof. Let $P \subseteq \mathbb{R}^n$ be a rational polyhedron of size at most φ, where $n, \varphi \in \mathbb{N}$. We have seen in Chapter 1 that if $\max\left\{\boldsymbol{c}^\mathsf{T}\boldsymbol{x} \ : \ \boldsymbol{x} \in P\right\}$ is finite, then it is attained by a vertex of P. Since vertices of P have size at most $4n^2\varphi$ (see again Theorem 10.2 in [75]), the size of any finite maximum will be bounded by $2(\text{size}(c) + 4n^2\varphi)$.

Next, define $\tau = 3\,\text{size}(c) + 12n^2\varphi$ and set $m_0 = -2^\tau$ and $M_0 = 2^\tau$. We will now repeatedly update these values and use the algorithm for the feasibility program to test whether

$$P \cap \left\{\boldsymbol{x} \in \mathbb{R}^n \ : \ \boldsymbol{c}^\mathsf{T}\boldsymbol{x} \geq \frac{1}{2}(m_i + M_i)\right\}$$

is empty.

If it is, we set $m_{i+1} = m_i$ and $M_{i+1} = \frac{1}{2}(m_i + M_i)$. If the set is not empty, then instead we set $m_{i+1} = \frac{1}{2}(m_i + M_i)$ and $M_{i+1} = M_i$. We stop when we reach $i = 3\tau + 2 =: K$. We thus performed polynomially many iterations and there are now three possibilities.

Case 1. $m_K \neq m_0$ and $M_K \neq M_0$.

Then $M := \max\left\{\boldsymbol{c}^\mathsf{T}\boldsymbol{x} \ : \ \boldsymbol{x} \in P\right\}$ must be finite, with $m_K \leq M < M_K$. Note that $(M_K - m_K) = 2^{-2\tau-1}$. Since $\text{size}(M) < \tau$, the denominator of M is at most 2^τ, and using continued fractions it is not hard to see that this determines M uniquely (see also Corollary 6.3a in [75]).

We can find a vector $\boldsymbol{x} \in P$ with $\boldsymbol{c}^\mathsf{T}\boldsymbol{x} = M$ by using the feasibility algorithm on $P \cap \left\{\boldsymbol{x} \in \mathbb{R}^n \ : \ \boldsymbol{c}^\mathsf{T}\boldsymbol{x} = M\right\}$.

Case 2. $M_K = M_0$.

Then $\max\left\{\boldsymbol{c}^\mathsf{T}\boldsymbol{x} \ : \ \boldsymbol{x} \in P\right\}$ is unbounded, since there is a vector $\boldsymbol{x} \in P$ with $\boldsymbol{c}^\mathsf{T}\boldsymbol{x} \geq m_K = 2^\tau - 2^{-2\tau-1} > 2^{2(\text{size}(c)+4n^2\varphi)}$.

Case 3. $m_K = m_0$.

Then P is empty, since otherwise there would be an $\boldsymbol{x} \in P$ with $\text{size}(\boldsymbol{x}) \leq 4n^2\varphi$. But then $\boldsymbol{c}^\mathsf{T}\boldsymbol{x} \geq -2^{\text{size}(c)+4n^2\varphi} \geq -2^\tau + 2^{-2\tau-1} = M_K$, and since we did not change m_i in the last step, this is a contradiction to the definition of M_K.

□

Lemma 3.4. *If we can solve the linear feasibility program in polynomial time, then we can also find the smallest affine subspace containing P in polynomial time.*

Proof. We first show the statement for the case that $P = L$ is a linear subspace of $\mathbb{R}^n$. To this end, we want to find linearly independent vectors $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_t, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_d$, such that

$$L = \left\{\boldsymbol{x} \in \mathbb{R}^n \ : \ \boldsymbol{c}_1^\mathsf{T}\boldsymbol{x} = \cdots = \boldsymbol{c}_t^\mathsf{T}\boldsymbol{x} = 0\right\} = \text{span}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d\}.$$

We will do this iteratively, starting with $d = 0$.

Suppose we have found linearly independent vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d$, then, possibly after permuting coordinates, we get that

$$[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d] = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix},$$

such that $\boldsymbol{X}_1$ is a non-singular matrix of rank d.

Let L_d consist of all vectors in L with zeros in the first d coordinates, i.e.,

$$L_d = L \cap \left\{ \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{y} \end{pmatrix} \in \mathbb{R}^n \ : \ \boldsymbol{y} \in \mathbb{R}^{n-d} \right\}.$$

To find $\boldsymbol{x}_{d+1}$ (if it exists), we just need to find a non-zero vector in L_d. We can do so by solving the feasibility program for the polyhedra

$$P_{d,i} := L_d \cap \{\boldsymbol{x} \in \mathbb{R}^n \ : \ x_i \geq 1\}$$

for $i = d + 1, \ldots, n$.

If we find a vector in one of these polyhedra, we can permute coordinates again such that it is in $P_{d,d+1}$. We then call this vector $\boldsymbol{x}_{d+1}$ and restart the procedure for $d + 1$.

If all of the $P_{d,i}$ are empty, then $L_d = \{\boldsymbol{0}\}$ and hence $L = \operatorname{span}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d\}$. It is also not difficult to check that furthermore

$$L = \left\{ \boldsymbol{x} \in \mathbb{R}^n \ : \ [\boldsymbol{X}_2 \cdot \boldsymbol{X}_1^{-1}, -\boldsymbol{I}]\boldsymbol{x} = \boldsymbol{0} \right\},$$

which gives us the vectors $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_t$ as required; we just need to extract a maximal system of linearly independent rows.

Now let P be any polyhedron. With the feasibility program we determine some $\boldsymbol{x}_0 \in P$ or conclude that P is empty (in which case the alline hull is also empty).

If P is not empty, let L be the linear space we get by shifting the affine hull of P by $-\boldsymbol{x}_0$. Note that we ca also solve the feasibility program for L in polynomial time. Indeed, to test whether $\boldsymbol{z} \neq \boldsymbol{0}$ is in L, we can invoke the optimization program of Lemma 3.3 for P with objective $\boldsymbol{z}$ and $-\boldsymbol{z}$. The optimal vectors in P are identical if and only if $\boldsymbol{z} \notin L$.

Now we can, as we did above, find linearly independent vectors $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_t$ such that $L = \left\{ \boldsymbol{x} \in \mathbb{R}^n \ : \ \boldsymbol{c}_1^\mathsf{T}\boldsymbol{x} = \cdots = \boldsymbol{c}_t^\mathsf{T}\boldsymbol{x} = 0 \right\}$, and thus the affine hull of P is given by

$$\left\{ \boldsymbol{x} \in \mathbb{R}^n \ : \ \boldsymbol{c}_1^\mathsf{T}\boldsymbol{x} = \boldsymbol{c}_1^\mathsf{T}\boldsymbol{x}_0, \cdots \boldsymbol{c}_t^\mathsf{T}\boldsymbol{x} = \boldsymbol{c}_t^\mathsf{T}\boldsymbol{x}_0 \right\}.$$

□

3.2 Lenstra's Algorithm

We reproduce Lenstra's algorithm here in the version given by Schrijver in [75]. The main tool is the following theorem, which gives us a constructive upper bound on the lattice-width. The key to the proof is applying lattice reduction to the lattice we get from $\mathbb{Z}^n$ after the affine transformation that makes the polytope as round as possible. In other words, we reduce with respect to the shape of the polytope.

Theorem 3.5 ([44]). *There exists a polynomial algorithm which finds, for any system $\boldsymbol{Ax} \leq \boldsymbol{b}$ of rational linear inequalities, either an integer vector $\boldsymbol{y}$ satisfying $\boldsymbol{Ay} \leq \boldsymbol{b}$, or a nonzero integer vector $\boldsymbol{c}$ such that*

$$\max\{\boldsymbol{cx} \ : \ \boldsymbol{Ax} \leq \boldsymbol{b}\} - \min\{\boldsymbol{cx} \ : \ \boldsymbol{Ax} \leq \boldsymbol{b}\} \leq 2n(n+1)2^{n(n-1)/4}, \tag{3.5}$$

where n is the number of columns of A.

Proof. Define $P = \{\boldsymbol{x} \in \mathbb{R}^n \ : \ \boldsymbol{Ax} \leq \boldsymbol{b}\}$. With the ellipsoid method, following Lemma 3.4, we determine the smallest affine subspace containing P. If P is not full-dimensional, this will give us a nonzero vector $\boldsymbol{c}$ and a rational δ such that $P \subseteq \left\{\boldsymbol{x} \in \mathbb{R}^n \ : \ \boldsymbol{c}^\mathsf{T}\boldsymbol{x} = \delta\right\}$, and we can stop.

Thus we may assume that P is full-dimensional and, for the moment, additionally we assume that P is bounded.

Then by Theorem 3.2 we find, in polynomial time, a rational vector z and a positive definite matrix $\boldsymbol{D}$ such that

$$\operatorname{ell}\left(z, \frac{1}{(n+1)^2}\boldsymbol{D}\right) \subseteq P \subseteq \operatorname{ell}(z, \boldsymbol{D}).$$

Let $\boldsymbol{D}^{-1} = \boldsymbol{M}\boldsymbol{M}^\mathsf{T}$ where $\boldsymbol{M}$ is non-singular. Note that we can find $\boldsymbol{M}$ by using the eigendecomposition of $\boldsymbol{D}^{-1}$. If we now consider the lattice L with basis $\boldsymbol{M}^\mathsf{T}$, then this is the image of $\mathbb{Z}^n$ under the affine transformation we need to apply to P to ensure that the ellipsoid we get from Theorem 3.2 is the unit ball.

Note that, given a basis of L, the inverse under this transformation will always give us a basis of $\mathbb{Z}^n$. Therefore, instead of applying the LLL-reduction algorithm to $\boldsymbol{M}^\mathsf{T}$, we apply it to the standard unit basis of $\mathbb{Z}^n$, where we use the norm $/\!/\boldsymbol{x}/\!/ := \|\boldsymbol{M}^\mathsf{T}\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^\mathsf{T}\boldsymbol{D}^{-1}\boldsymbol{x}}$ instead of the Euclidean norm.

Let $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ be the basis of $\mathbb{Z}^n$ we find this way. Note that the reduced basis of L is given by $\boldsymbol{M}^\mathsf{T}\boldsymbol{b}_1, \ldots, \boldsymbol{M}^\mathsf{T}\boldsymbol{b}_n$.

We know by Lemma 1.13 that the orthogonality defect of L is at most $2^{n(n-1)/4}$ and thus

$$/\!/\boldsymbol{b}_1/\!/ \cdot \ldots \cdot /\!/\boldsymbol{b}_n/\!/ \leq 2^{n(n-1)/4}\sqrt{\det(\boldsymbol{D}^{-1})}.$$

Now we can write the center of the ellipsoids z in terms of this new basis, $z = \lambda_1\boldsymbol{b}_1 + \cdots + \lambda_n\boldsymbol{b}_n$ and define

$$\boldsymbol{y} = \lfloor\lambda_1\rfloor\,\boldsymbol{b}_1 + \cdots + \lfloor\lambda_n\rfloor\,\boldsymbol{b}_n.$$

If $\boldsymbol{y}$ satisfies $\boldsymbol{Ay} \leq \boldsymbol{b}$, we are done. Otherwise, $\boldsymbol{y}$ will also not be in the smaller ellipsoid $\operatorname{ell}\left(z, \frac{1}{(n+1)^2}\boldsymbol{D}\right)$, which means

$$\begin{aligned}\frac{1}{(n+1)^2} &< (\boldsymbol{y}-z)^\mathsf{T}\boldsymbol{D}^{-1}(\boldsymbol{y}-z) = /\!/\boldsymbol{y}-z/\!/^2 \\ &= /\!/\lambda_1 - \lfloor\lambda_1\rfloor\,\boldsymbol{b}_1 + \cdots + \lambda_n - \lfloor\lambda_n\rfloor\,\boldsymbol{b}_n/\!/^2.\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{1}{n+1} &< (\lambda_1 - \lfloor\lambda_1\rfloor)/\!/\boldsymbol{b}_1/\!/ + \cdots + (\lambda_n - \lfloor\lambda_n\rfloor)/\!/\boldsymbol{b}_n/\!/ \\ &\leq /\!/\boldsymbol{b}_1/\!/ + \cdots + /\!/\boldsymbol{b}_n/\!/.\end{aligned}$$

Without loss of generality we can assume that $\boldsymbol{b}_n$ has the maximal norm among the basis vectors. Then we know that $\|\boldsymbol{b}_n\| \geq \frac{1}{n(n+1)}$.

Let $\boldsymbol{B}_i$ be the matrix with columns $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_i$, for $i = 1, \ldots, n$. If $\boldsymbol{c}$ is a nonzero vector that is orthogonal to $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{n-1}$, then $\boldsymbol{M}^{-1}\boldsymbol{c}$ is a nonzero vector that is orthogonal to $\boldsymbol{M}^\mathsf{T}\boldsymbol{b}_1, \ldots, \boldsymbol{M}^\mathsf{T}\boldsymbol{b}_{n-1}$. Furthermore, observe that $\det(\boldsymbol{B}_n) = 1$ by construction.

Thus we can compute

$$\begin{aligned}\det L &= \sqrt{\det(\boldsymbol{D}^{-1})} = \sqrt{\det(\boldsymbol{B}_n^\mathsf{T}\boldsymbol{M}\boldsymbol{M}^\mathsf{T}\boldsymbol{B}_n)} \\ &= \frac{|(\boldsymbol{M}^{-1}\boldsymbol{c})^\mathsf{T}\boldsymbol{M}^\mathsf{T}\boldsymbol{b}_n|}{\|\boldsymbol{M}^{-1}\boldsymbol{c}\|}\sqrt{\det(\boldsymbol{B}_{n-1}^\mathsf{T}\boldsymbol{M}\boldsymbol{M}^\mathsf{T}\boldsymbol{B}_{n-1})} \\ &= \frac{|\boldsymbol{c}^\mathsf{T}\boldsymbol{b}_n|}{\sqrt{\boldsymbol{c}^\mathsf{T}\boldsymbol{D}\boldsymbol{c}}}\sqrt{\det(\boldsymbol{B}_{n-1}^\mathsf{T}\boldsymbol{D}^{-1}\boldsymbol{B}_{n-1})}\,,\end{aligned}$$

and therefore

$$(\boldsymbol{c}^\mathsf{T}\boldsymbol{D}\boldsymbol{c})\det(\boldsymbol{D}^{-1}) = (\boldsymbol{c}^\mathsf{T}\boldsymbol{b}_n)^2\det(\boldsymbol{B}_{n-1}^\mathsf{T}\boldsymbol{D}^{-1}\boldsymbol{B}_{n-1}).$$

We may further assume that $\boldsymbol{c}$ is integer and that the components are relatively prime. Since $\boldsymbol{B}_n$ is a basis of $\mathbb{Z}^n$ and $\boldsymbol{c}$ is orthogonal to all $\boldsymbol{b}_i$ with $i < n$, we conclude that in fact we must have $\boldsymbol{c} = \pm\boldsymbol{b}_n$ and therefore in particular $(\boldsymbol{c}^\mathsf{T}\boldsymbol{b}_n)^2 = 1$.

By Hadamard's inequality (see (1.3)), we have furthermore

$$\begin{aligned}\det(\boldsymbol{B}_{n-1}^\mathsf{T}\boldsymbol{D}^{-1}\boldsymbol{B}_{n-1}) &\leq (\boldsymbol{b}_1^\mathsf{T}\boldsymbol{D}^{-1}\boldsymbol{b}_1)\cdot\ldots\cdot(\boldsymbol{b}_{n-1}^\mathsf{T}\boldsymbol{D}^{-1}\boldsymbol{b}_{n-1}) \\ &= \|\boldsymbol{b}_1\|^2\cdot\ldots\cdot\|\boldsymbol{b}_{n-1}\|^2.\end{aligned}$$

Thus, taking all this together, if $\boldsymbol{x}^\mathsf{T}\boldsymbol{D}^{-1}\boldsymbol{x} \leq 1$, then we have

$$\begin{aligned}|\boldsymbol{c}^\mathsf{T}\boldsymbol{x}| &\leq \|\boldsymbol{c}^\mathsf{T}\boldsymbol{M}\|\cdot\|\boldsymbol{M}^{-1}\boldsymbol{x}\| = \sqrt{\boldsymbol{c}^\mathsf{T}\boldsymbol{D}\boldsymbol{c}}\cdot\sqrt{\boldsymbol{x}^\mathsf{T}\boldsymbol{D}^{-1}\boldsymbol{x}} \leq \sqrt{\boldsymbol{c}^\mathsf{T}\boldsymbol{D}\boldsymbol{c}} \\ &= \sqrt{\det\boldsymbol{D}}\cdot\sqrt{\det(\boldsymbol{B}_{n-1}^\mathsf{T}\boldsymbol{D}^{-1}\boldsymbol{B}_{n-1})} \leq \sqrt{\det\boldsymbol{D}}\cdot\|\boldsymbol{b}_1\|\cdot\ldots\cdot\|\boldsymbol{b}_{n-1}\| \\ &\leq 2^{n(n-1)/4}\|\boldsymbol{b}_n\|^{-1} \leq n(n+1)2^{n(n-1)/4},\end{aligned}$$

where we use the Cauchy-Schwarz inequality for the first step.

In conclusion, if $\boldsymbol{y}_1, \boldsymbol{y}_2$ are in ell$(\boldsymbol{z}, \boldsymbol{D})$, then

$$|\boldsymbol{c}^\mathsf{T}\boldsymbol{y}_1 - \boldsymbol{c}^\mathsf{T}\boldsymbol{y}_2| \leq |\boldsymbol{c}^\mathsf{T}(\boldsymbol{y}_1 - \boldsymbol{z})| + |\boldsymbol{c}^\mathsf{T}(\boldsymbol{y}_2 - \boldsymbol{z})| \leq 2n(n+1)2^{n(n-1)/4},$$

and thus $\boldsymbol{c}$ satisfies (3.5).

What remains now is to show what to do if P is unbounded. Using Lemma 1.5, we observe that if we intersect P with a large enough ball (that still has a radius of polynomial size), all vertices of P and a good part more lies in the intersection (see [75] p.258 for details). Then it is just a matter of choosing the radius correctly to see that if the vector $\boldsymbol{y}$ we construct above does not lie in the intersection, then the $\boldsymbol{c}$ this leads to satisfies (3.5) for the whole polyhedron P. □

Corollary 3.6 (Lenstra's algorithm)**.** *For each fixed natural number n, there exists a polynomial algorithm which finds an integer solution for a given rational system $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$, in n variables, or decides that no such solution exists.*

Proof. We use induction on n, where the case $n = 1$ is easy. Now let $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$ be given, where $\boldsymbol{A}$ has n columns, and apply the algorithm of Theorem 3.5 to this system of inequalities. If we get an integer solution from this, we are done. Suppose we get a nonzero integer vector $\boldsymbol{c}$ that satisfies (3.5). We may assume that the components of $\boldsymbol{c}$ are relatively prime.

We define $P = \{\boldsymbol{x} \in \mathbb{R}^n \ : \ \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}\}$, and determine $\mu := \min\left\{\boldsymbol{c}^\mathsf{T}\boldsymbol{x} \ : \ \boldsymbol{x} \in P\right\}$. Then for $t = \lceil\mu\rceil, \ldots, \left\lfloor\mu + 2n(n+1)2^{n(n-1)/4}\right\rfloor$, consider the polyhedra

$$P_t = \left\{\boldsymbol{x} \in P \ : \ \boldsymbol{c}^\mathsf{T}\boldsymbol{x} = t\right\}.$$

Each integer solution of $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$ is in one of these P_t.

Let $\boldsymbol{U}$ be such that $\begin{bmatrix}\boldsymbol{c}^\mathsf{T}\\ \boldsymbol{U}\end{bmatrix}$ is an $n \times n$ unimodular integer matrix. Then $\boldsymbol{U}$ maps any $\boldsymbol{x} \in P_t$ into

$$Q_t = \left\{\boldsymbol{y} \in \mathbb{R}^{n-1} \ : \ \boldsymbol{A}\begin{bmatrix}\boldsymbol{c}^\mathsf{T}\\ \boldsymbol{U}\end{bmatrix}^{-1}\begin{pmatrix}t\\ \boldsymbol{y}\end{pmatrix} \leq \boldsymbol{b}\right\},$$

and moreover, if $\boldsymbol{x} \in P_t \cap \mathbb{Z}^n$ then $\boldsymbol{U}\boldsymbol{x} \in Q_t \cap \mathbb{Z}^{n-1}$, and if $\boldsymbol{y} \in Q_t \cap \mathbb{Z}^{n-1}$ then $\begin{bmatrix}\boldsymbol{c}^\mathsf{T}\\ \boldsymbol{U}\end{bmatrix}^{-1}\begin{pmatrix}t\\ \boldsymbol{y}\end{pmatrix} \in P_t \cap \mathbb{Z}^n$.

Note that we can indeed always find such a matrix $\boldsymbol{U}$: Perform elementary column operations on $\boldsymbol{c}^\mathsf{T}$ until we have a unit vector, say $\boldsymbol{e}_i$. This must occur at some point because the components of $\boldsymbol{c}$ are relatively prime. Then we get $\boldsymbol{U}$ by inserting $\boldsymbol{0}$ as i^{th} column into the unit matrix and reversing the column operations one by one.

Thus, we have reformulated the program to at most $2n(n+1)2^{n(n-1)/4} + 1$ programs of dimension $n - 1$. □

3.3 Ellipsoids and Reformulations

We are now ready to return to our original question: Let

$$P' = \{\boldsymbol{\mu} \in \mathbb{R}^{n-m} \ : \ \boldsymbol{Q}\boldsymbol{\mu} \geq -\boldsymbol{x}_0\}$$

be the polytope we get from reformulating $P = \{\boldsymbol{x} \in \mathbb{R}^n \ : \ \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \geq \boldsymbol{0}\}$ as in Chapter 2. How can we incorporate the general shape of P' when we reduce the lattice $\ker_{\mathbb{Z}}(\boldsymbol{A})$?

Note that, by construction, P' will be full-dimensional and we may assume bounded, thus we can describe its shape directly as we did in the proof of Theorem 3.5. We want to find a rational vector $\boldsymbol{z}$ and a positive definite matrix $\boldsymbol{D}$, such that

$$\operatorname{ell}\left(\boldsymbol{z}, \frac{1}{(n+1)^2}\boldsymbol{D}\right) \subseteq P' \subseteq \operatorname{ell}(\boldsymbol{z}, \boldsymbol{D}).$$

If we then take $\boldsymbol{M}$ to be the non-singular matrix with $\boldsymbol{D}^{-1} = \boldsymbol{M}\boldsymbol{M}^\mathsf{T}$ and consider the lattice $L(\boldsymbol{M}^\mathsf{T}\boldsymbol{Q})$, then a reduced basis will have the desired property that lattice hyperplanes generated by all basis vectors but one long one will give us a decent approximation of the lattice width, or in other words, the length of a basis vector is now related to the width of P' in the direction of the vector.

At this point, the question arises why this is not used in practice. As we mentioned before, the idea has been around for some time, and was explicitly described already at least 30 years ago. If we observe the steps once more, there are two large building blocks: finding a small

ellipsoid around P', and reducing lattice bases (once to get a basis of the kernel lattice of $\boldsymbol{A}$, once to reduce it with respect to the ellipsoid).

While lattice reduction is known to slow down considerably when we go from dozens of variables to hundreds, our experiments indicate that the larger issue is the ellipsoid. Indeed, the time spent on finding an improved basis based on the ellipsoid (in comparison to the reformulation from Chapter 2) always significantly exceeded the amount of time saved on branching afterwards. Note that because of the discouraging nature of this, we only did a very limited number of experiments in this setting.

Instead of trying to find an ellipsoid containing P' with minimum volume (also called *Loewner ellipsoid*), we can try to find an ellipsoid contained in P' of maximum volume (also called *John ellipsoid*). See, e.g., the short survey of Henk [47] for some historical remarks.

While the Loewner ellipsoid gives us a lower bound on the volume of the John ellipsoid by Theorem 3.2, there is no exact relationship known between these ellipsoids, ecxept for some special cases, where the polytope has particular symmetries.

The task to find such an ellipsoid of maximum volume is of interest for Control Theory [18], and in particular for the *Inscribed Ellipsoid method*, see [80].

In [70], Nemirovski describes a path-following interior point method for approximating saddle points of a certain kind of functions, which can be used for finding approximations of the ellipsoid with largest possible volume inside P'. To find an ellipsoid with volume at least $(1 - \varepsilon)$ times the maximal one, the running time of this method is $O(d^{3.5} \ln(d\varepsilon^{-1}R))$, where d is the dimension of P' and R is the ratio of two radii such that, for a fixed center, the Euclidean ball of the smaller radius is contained in P', and the Euclidean ball of the larger radius contains P'.

A restriction of this method is, that we need to know a bound on the size of the elements of P', which is passed to the algorithm. By Lemma 1.5 this bound can be chosen of polynomial size, but choosing a large bound here will slow down the algorithm significantly.

However, this algorithm does not require the polytope to be full-dimensional, and we can thus apply it also to the polytope P in the original space. This has the additional advantage that we can use the positive definite matrix $\boldsymbol{D}$ describing the ellipsoid precisely the way we used the diagonal matrix for the knapsack-program at the beginning of the chapter.

Since prgrams where the components are required to be binary are particularly suitable for the algorithm of Nemirovski (as they come with a small bound R), we used the notorious market split problems (see Cornuéjols and Dawande [26]) to test the practicability of this procedure. We get indeed a correct procedure. However, there was no consistent improvement in the running time, when compared to using the reduction described in Chapter 2, and in fact for small dimensions the additional time we spend on the computation of the ellipsoid is most of the time not recovered, see Table 3.2.

We tested five instances with 40 variables and four instances with 50 variables. The particular method we used for solving the instances (after reformulating) did not permit larger instances. We have done some tests with different solving methods and slightly more variables, with similarly inconsistent outcome. Note that the time spent on computing the ellipsoid is not part of the table, but lies consistently by about 2-5 seconds.

These results come as no surprise if we take a closer look at the ellipsoids we get, because as it turns out, they are on average very close to a Euclidean sphere, in particular for a larger

Table 3.2: Running time (in seconds) for market split problems with and without ellipsoidal norm.

instance	5x40_1	5x40_2	5x40_3	5x40_4	5x40_5
LLL	0.3	0.12	0.26	0.15	0.15
LLL+E	0.34	0.22	0.3	0.24	0.29
instance	6x50_1	6x50_2	6x50_3	6x50_4	
LLL	97.7	61.6	40.6	286.9	
LLL+E	36	35.4	140.2	147.9	

number of variables. Thus, the improvement we might get from obtaining a different reduced basis is often smaller than the additional time we spent on investigating the shape of P.

This observation also raises the question of how to capture the difficulty of market split instances mathematically.

Thus, an exciting challenge for further research will be to investigate possibilities of detecting whether the feasible region of a program might have a more accentuated shape and will therefore enable the method of reducing with respect to the ellipsoidal norm to have a more consistently positive effect on the running time.

Chapter Four

On the Structure of Kernel Lattice Bases

We again look at the integer program

$$\max\left\{\boldsymbol{c}^{\mathsf{T}}\boldsymbol{x} \;:\; \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},\ \boldsymbol{x} \geq \boldsymbol{0}\right\}, \tag{eq-IP}$$

where $\boldsymbol{A}$ is an integer $m \times n$ matrix of full row rank and $\boldsymbol{b}$ an integer m-vector. As before, we are interested in the reformulation

$$\boldsymbol{x} := \boldsymbol{x}^0 + \boldsymbol{Q}\boldsymbol{\lambda}\,, \tag{4.1}$$

where $\boldsymbol{x}^0 \in \mathbb{Z}^n$ satisfies $\boldsymbol{A}\boldsymbol{x}^0 = \boldsymbol{b}$, $\boldsymbol{\lambda} \in \mathbb{Z}^{n-m}$, and $\boldsymbol{Q}$ is a basis for the lattice $\ker_{\mathbb{Z}}(\boldsymbol{A}) = \{\boldsymbol{x} \in \mathbb{Z}^n \;:\; \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\}$. The equivalent formulation of the integer program (eq-IP) is then

$$\max\left\{\boldsymbol{c}^{\mathsf{T}}(\boldsymbol{x}^0 + \boldsymbol{Q}\boldsymbol{\lambda}) \;:\; \boldsymbol{Q}\boldsymbol{\lambda} \geq -\boldsymbol{x}^0\right\}. \tag{4.2}$$

Several authors have studied knapsack instances that have a certain structure that makes them particularly difficult to solve by "standard" methods such as branch-and-bound. Examples of such instances can be found in [3, 29, 54]. Common for these instances is that the input is generated in such a way that the resulting lattice $\ker_{\mathbb{Z}}(\boldsymbol{A})$ has a very particular structure that makes the reformulated instances almost trivial to solve. Other instances that are randomly generated without any particular structure of the $\boldsymbol{A}$-matrix, such as the market split instances [26] and knapsack instances studied in [3, 5], have no particular lattice structure. Yet they are practically unsolvable by branch-and-bound in the original $\boldsymbol{x}$-variable space, whereas their lattice reformulation solves rather easily, at least up to a certain dimension. It is still to be understood why the lattice reformulation for these instances is computationally more effective.

If we consider randomly generated instances without any particular lattice structure and solve small instances, such as $n - m \leq 50$, one typically observes that the number of zeros in the LLL-reduced basis $\boldsymbol{Q}$ is small. In higher dimension, and here "high" is depending on the input, a certain sparser structure will start to appear.

More specifically, we observe computationally that $\boldsymbol{Q}$ has a certain number of rows with rich interaction between the variables $\boldsymbol{x}$ and $\boldsymbol{\lambda}$, but from some point on this interaction breaks down almost instantly and we get one '1' per row, i.e., $\boldsymbol{Q}$ yields variable substitutions. To be able to better understand the relative effectiveness of the lattice reformulation, and in order to be able to apply the lattice reformulation in a (more) useful way in higher dimension, it is important to identify the variables that have a nontrivial translation into the new $\boldsymbol{\lambda}$-variable space.

In this chapter we partially explain the phenomenon described above for the case that $m = 1$, that is, $\boldsymbol{A}$ consists of a single row $\boldsymbol{a} = (a_1, \ldots, a_n)$. As the structure of $\boldsymbol{Q}$ depends

on the choice of $\boldsymbol{a}$, our analysis will be probabilistic. To this end, we assume that the entries of our input vector $\boldsymbol{a}$ are drawn independently and uniformly at random from an interval $[l, \ldots, u] := [l, u] \cap \mathbb{Z}$, where $0 < l < u$. We notice that explaining the phenomenon is related to the analysis of the probability that the LLL-algorithm (see Chapter 1.2.2) performs a basis vector interchange after a basis vector with a certain index k has been considered by the algorithm.

Let $\boldsymbol{Q} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{n-1}]$ be an LLL y-reduced basis of $\ker_{\mathbb{Z}}(\boldsymbol{a})$, and let $\boldsymbol{b}_1^*, \ldots, \boldsymbol{b}_{n-1}^*$ be the Gram-Schmidt vectors corresponding to $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{n-1}$. If $\|\boldsymbol{b}_{i+1}^*\|^2 \geq y\|\boldsymbol{b}_i^*\|^2$, then basis vectors $i+1$ and i will not be interchanged. We will show that, starting with a basis $\bar{\boldsymbol{Q}}$ of $\ker_{\mathbb{Z}}(\boldsymbol{a})$ of a certain structure, the probability that the LLL-algorithm performs basis vector interchanges becomes increasingly small the higher the index of the basis vector. In particular, for given l, u, and reduction factor y, we derive a constant c and a k_0, such that for $k \geq k_0$ we have

$$\Pr\left(\|\boldsymbol{b}_{k+1}^*\|^2 < y\|\boldsymbol{b}_k^*\|^2\right) \quad \leq \quad e^{-c(k+1)} + 2^{-(k+1)/2}.$$

Note that, stated in this form, it is an asymptotic result, but we will see that the values of k_0 are very similar to the ones observed in the experiments.

To derive a bound on $\Pr\left(\|\boldsymbol{b}_{k+1}^*\|^2 < y\|\boldsymbol{b}_k^*\|^2\right)$ we first need to be able to express the length of the Gram-Schmidt vectors $\boldsymbol{b}_j^*$ in terms of the input vector $\boldsymbol{a}$. This is done in Section 4.1 and results in Expression (4.9). The bound on $\Pr\left(\|\boldsymbol{b}_{k+1}^*\|^2 < y\|\boldsymbol{b}_k^*\|^2\right)$ is derived through several steps in Section 4.2. In this derivation, the challenge is that $\|\boldsymbol{b}_{k+1}^*\|^2$ and $\|\boldsymbol{b}_k^*\|^2$ are not independent. To estimate the mean of the ratio $\|\boldsymbol{b}_{k+1}^*\|^2/\|\boldsymbol{b}_k^*\|^2$, we use a result by Pittenger [72], and to estimate how much this ratio deviates from the mean we use the Azuma-Hoeffding inequality [8, 48]. Some further discussion and computational indications are provided in Sections 4.3 and 4.4. We notice that the theoretical results correspond well to the observed practical behavior of the LLL algorithm on the considered class of input.

4.1 More on lattices and reduced bases

We first repeat some facts about lattices and reduced bases of lattices from Chapter 1, and prove the Key-Lemma 4.2.

Let L be a lattice in $\mathbb{R}^n$, and let $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$, $m \leq n$, be a basis of L. Recall from Chapter 1 that the Gram-Schmidt vectors are defined as follows:

$$\begin{aligned}
\boldsymbol{b}_1^* &= \boldsymbol{b}_1, \\
\boldsymbol{b}_i^* &= \boldsymbol{b}_i - \sum_{j=1}^{i-1} \mu_{ij}\boldsymbol{b}_j^*, \quad 2 \leq i \leq m, \quad \text{where} \\
\mu_{ij} &= \frac{\boldsymbol{b}_i^{\mathsf{T}}\boldsymbol{b}_j^*}{\|\boldsymbol{b}_j^*\|^2}, \quad 1 \leq j < i \leq m.
\end{aligned}$$

We also recall that a basis $\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m\}$ is called *y-reduced*, if

$$|\mu_{ij}| \leq \frac{1}{2}, \quad \text{for } 1 \leq j < i \leq m-1, \text{ and} \tag{4.3}$$

$$\|\boldsymbol{b}_i^* + \mu_{i,i-1}\boldsymbol{b}_{i-1}^*\|^2 \geq y\,\|\boldsymbol{b}_{i-1}^*\|^2, \quad \text{for } 1 < i \leq m-1\,. \tag{4.4}$$

Notice that, as $\boldsymbol{b}_1^*, \ldots, \boldsymbol{b}_m^*$ are pairwise orthogonal, Inequality (4.4) is satisfied if

$$\|\boldsymbol{b}_i^*\|^2 \geq y\|\boldsymbol{b}_{i-1}^*\|^2, \quad \text{for } 1 < i \leq m-1\,. \tag{4.5}$$

If Condition (4.3) is violated, i.e., $|\mu_{kj}| > 1/2$ for some $j < k$, then the LLL-algorithm will perform a *size reduction* by setting $\boldsymbol{b}_k := \boldsymbol{b}_k - \lfloor \mu_{kj} \rceil \boldsymbol{b}_j$. Notice that this operation will not change the Gram-Schmidt vector $\boldsymbol{b}_k^*$. If Condition (4.4) is violated for $i = j$, then vectors $\boldsymbol{b}_{j-1}$ and $\boldsymbol{b}_j$ are *interchanged*. This operation does affect several of the μ-values. Moreover, the new vector $\boldsymbol{b}_{j-1}^*$ will be the old vector $\boldsymbol{b}_j^* + \mu_{j,j-1}\boldsymbol{b}_{j-1}^*$. See Section 1.2 for more details.

We also recall from Chapter 1 that $\det(L) = \prod_{i=1}^{m} \|\boldsymbol{b}_i^*\|$ (see (1.2)), and that if K is a pure sublattice of $\mathbb{Z}^n$, then $\det(K) = \det(K^\perp)$ (see (1.7)).

Let us now consider a vector $\boldsymbol{a} \in \mathbb{Z}^n$ such that $\gcd(a_1, \ldots, a_n) = 1$. We again define the kernel lattice of $\boldsymbol{a}$ as the set $\ker_{\mathbb{Z}}(\boldsymbol{a}) := \{\boldsymbol{x} \in \mathbb{Z} \,:\, \boldsymbol{a}\boldsymbol{x} = 0\}$. As we have seen in Chapter 1, the lattice $\ker_{\mathbb{Z}}(\boldsymbol{a})$ is a pure sublattice of $\mathbb{Z}^n$.

With all this in mind again, we now show that the lattice $\ker_{\mathbb{Z}}(\boldsymbol{a})$ has a basis of the following form:

$$\boldsymbol{Q} = \begin{pmatrix} x & x & \cdots & x \\ x & x & \cdots & x \\ 0 & x & \cdots & x \\ \vdots & 0 & \ddots & x \\ 0 & \cdots & 0 & x \end{pmatrix}, \tag{4.6}$$

where each 'x' denotes some integer number that may be different from zero. This is a corollary of a classical result on lattice bases (see, e.g., [19] Theorem 1 and corollaries), but it is also not difficult to prove directly.

Lemma 4.1. *The lattice* $\ker_{\mathbb{Z}}(\boldsymbol{a})$ *has a basis* $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{n-1}$ *of the following form:*

$$\mathbb{Z}\boldsymbol{b}_1 + \ldots + \mathbb{Z}\boldsymbol{b}_k = \ker_{\mathbb{Z}}(\boldsymbol{a}) \cap (\mathbb{Z}^{k+1} \times 0^{n-k-1})$$

for any $1 \leq k \leq n-1$.

Proof. Write $c_i = \min\left\{|y_i| > 0 \,:\, \boldsymbol{y} \in \ker_{\mathbb{Z}}(\boldsymbol{a}), y_j = 0 \text{ for } j > i\right\}$, where $2 \leq i \leq n$. Note that the set we minimize over is not empty, because at least vectors of the form

$$(-a_i, 0, \ldots, 0, a_1, 0, \ldots, 0)^\mathsf{T},$$

where a_1 appears in the i^{th} position, are in $\ker_{\mathbb{Z}}(\boldsymbol{a})$ for any $i \in \{2, \ldots, n\}$. Now choose

$$\boldsymbol{b}_i \in \left\{\boldsymbol{x} \in \ker_{\mathbb{Z}}(\boldsymbol{a}) \,:\, x_{i+1} = c_{i+1}, x_j = 0 \text{ for } j > i+1\right\}. \tag{4.7}$$

To see that this is indeed a lattice-basis, let $\boldsymbol{z} \in \ker_{\mathbb{Z}}(\boldsymbol{a})$ and let k be the highest index of a non-zero coordinate of $\boldsymbol{z}$. Let $\boldsymbol{Q} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{n-1}]$, where $\boldsymbol{b}_i$ satisfies (4.7).

We want to find $\boldsymbol{\lambda} \in \mathbb{Z}^{n-1}$ such that $\boldsymbol{z} = \boldsymbol{Q}\boldsymbol{\lambda}$. Observe that $\frac{z_k}{c_k}$ must be integer, because otherwise there is a $c' \in \mathbb{Z}$ such that $0 < |z_k - c'c_k| < c_k$, which contradicts the minimality of c_k. Therefore we may define $\lambda_{k-1} := \frac{z_k}{c_k}$.

Setting $\boldsymbol{z} = \boldsymbol{z} - \lambda_{k-1}\boldsymbol{b}_{k-1}$, this gives us a recursive construction for the integer coefficients $\lambda_1, \ldots, \lambda_{n-1}$ to express $\boldsymbol{z}$ in terms of our basis. □

One can additionally observe that if $\gcd(a_1, \ldots, a_i) = 1$ for some $1 \leq i \leq n$ then the last non-zero element of the basis vectors $\boldsymbol{b}_i, \ldots, \boldsymbol{b}_{n-1}$ is equal to ± 1. We will follow up on this idea in Section 4.4.

Let L_k be the sublattice given by the basis $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k$ as described in Lemma 4.1, for $1 \leq k \leq m$. Then we have $L_1 \subseteq L_2 \subseteq \cdots \subseteq L_{n-1} = \ker_{\mathbb{Z}}(\boldsymbol{a})$ and $\det(L_k) = \prod_{i=1}^{k} \|\boldsymbol{b}_i^*\|$. Also, because of the specific structure of the basis, we can express L_k as

$$L_k = \left\{ \boldsymbol{x} \in \mathbb{Z}^n \ : \ (a_1, \ldots, a_{k+1}, 0, \ldots, 0)\boldsymbol{x} = 0, x_j = 0, k+2 \leq j \leq n \right\}.$$

We can extend the above observations to conclude the following:

Lemma 4.2. *Let $L_1, \ldots, L_{n-1}$ be given as above and let $1 \leq k \leq n-1$. If $\gcd(a_1, \ldots, a_{k+1}) = 1$, then*

$$\det(L_k) = \sqrt{\sum_{i=1}^{k+1} a_i^2}, \tag{4.8}$$

and thus we get in particular

$$\|\boldsymbol{b}_k^*\|^2 = \frac{\sum_{i=1}^{k+1} a_i^2}{\sum_{i=1}^{k} a_i^2}. \tag{4.9}$$

Proof. Observe that $(a_1, \ldots, a_{k+1}, 0, \ldots, 0)^{\mathsf{T}}$ and the unit vectors $\boldsymbol{e}_j$, with $k+2 \leq j \leq n$, are an orthogonal basis of $L_k^{\perp}$. Using (1.2) and the fact that $\det(K) = \det(K^{\perp})$ for pure sublattices of $\mathbb{Z}^n$ (see (1.7)), we get (4.8).

Equation (4.9) follows from (1.2) in combination with (4.8) for L_k and L_{k-1}. □

4.2 Probabilistic analysis

Here we present the main result of the chapter, namely a bound on the probability that the LLL-algorithm will perform a basis vector interchange after basis vector $\boldsymbol{b}_k$ is considered. We assume that the elements a_i of the vector $\boldsymbol{a}$ are drawn independently and uniformly at random from an interval $[l, \ldots, u] := [l, u] \cap \mathbb{Z}$, where $0 < l < u$, and that the starting basis of $\ker_{\mathbb{Z}}(\boldsymbol{a})$ is a basis of the structure given in Lemma 4.1. Recall from Section 4.1 that if, for given reduction factor $y \in (\frac{1}{4}, 1)$,

$$\|\boldsymbol{b}_{i+1}^*\|^2 \geq y\|\boldsymbol{b}_i^*\|^2, \quad \text{for } 1 \leq i < n-1\,,$$

then the LLL-algorithm will not interchange basis vectors $\boldsymbol{b}_i$ and $\boldsymbol{b}_{i+1}$.

We will prove the following result:

Theorem 4.3. *Let $y \in (\frac{1}{4}, 1)$ be fixed. Then, for k large enough, we get*

$$\Pr\left(\frac{\|\boldsymbol{b}_{k+1}^*\|^2}{\|\boldsymbol{b}_k^*\|^2} < y \right) \quad \leq \quad e^{-c(k+1)} + 2^{-(k+1)/2}\,,$$

where $c > 0$ depends on u, l, and y.

We can provide explicit bounds on c when k is large enough. To increase accessibility to the proof, we build our result from several lemmas. We start by noticing that for any $1 \leq k < n-1$

$$\begin{aligned} \Pr\left(\|\boldsymbol{b}_{k+1}^*\|^2 < y\|\boldsymbol{b}_k^*\|^2 \right) \leq \ & \Pr\left(\|\boldsymbol{b}_{k+1}^*\|^2 < y\|\boldsymbol{b}_k^*\|^2 \mid \gcd(a_1, \ldots, a_{k+1}) = 1 \right) \\ & + \Pr(\gcd(a_1, \ldots, a_{k+1}) > 1), \end{aligned} \tag{4.10}$$

and hence we can bound the two terms separately. The last one can be bounded in the following way:

Lemma 4.4. *Let $a_1, \ldots, a_n$ be chosen independently and uniformly at random from the set $[l, \ldots, u]$ for some integers $0 < l < u$, and let l and u be fixed. Then*

$$\Pr(\gcd(a_1, \ldots, a_{k+1}) > 1) \leq \left(\frac{1}{2}\right)^{(k+1)/2}$$

for any $k \geq \frac{\log_2\left(\left\lfloor \frac{u}{2} \right\rfloor + 1\right)}{\log_2\left(\frac{u-l+1}{u-l+2}\right) + \frac{1}{2}}$.

Proof. First observe that if $u - l < 2$ then the statement is true as consecutive integers have no common divisor except 1. Thus we may assume $u - l \geq 2$.

For any prime p let $A_{l,u}(p) = \{x \in \mathbb{Z} \,:\, l \leq x \leq u \text{ and } p \text{ divides } x\}$. As l and u are fixed for the moment, we write $A_{l,u}(p) = A(p)$.

We observe that in fact $A(p) = \left\{\left\lceil \frac{l}{p} \right\rceil p, \left(\left\lceil \frac{l}{p} \right\rceil + 1\right) p, \ldots, \left\lfloor \frac{u}{p} \right\rfloor p\right\}$, and thus $|A(p)| = \left\lfloor \frac{u}{p} \right\rfloor - \left\lceil \frac{l}{p} \right\rceil + 1$. Set $s := u - l + 1$, then for any prime $p \leq u$ we have

$$\begin{aligned}
\Pr(a_i \in A(p)) &= \frac{\left\lfloor \frac{u}{p} \right\rfloor - \left\lceil \frac{l}{p} \right\rceil + 1}{s} \leq \frac{\frac{u}{p} - \frac{l}{p} + 1}{s} = \frac{\frac{1}{p}(u-l) + 1}{s} \\
&= \frac{1}{s} + \frac{1}{p}\left(\frac{u-l}{s}\right) \leq \frac{1}{s} + \frac{1}{2}\left(\frac{u-l}{s}\right) \\
&= \frac{s+1}{2s}
\end{aligned}$$

Next, let $Y(p, (k+1))$ denote the event that $a_1, \ldots, a_{k+1}$ are all divisible by p, i.e.,

$$Y(p, (k+1)) = \bigcap_{1 \leq i \leq (k+1)} \{a_i \in A(p)\},$$

and as the a_i are chosen independently

$$\Pr(Y(p, (k+1))) = \prod_{i=1}^{k+1} \Pr(a_i \in A(p)) \leq \left(\frac{s+1}{2s}\right)^{k+1}.$$

Let P_u be the set of prime numbers less than or equal to u. Then

$$\Pr(\gcd(a_1, \ldots, a_{k+1}) > 1) = \Pr\left(\bigcup_{p \in P_u} Y(p, (k+1))\right) \leq \sum_{p \in P_u} \Pr(Y(p, (k+1)))$$

by the union-bound. Then by using the above estimates for $\Pr(Y(p, (k+1)))$, and observing that all primes larger than 2 are odd, we get

$$\Pr(\gcd(a_1, \ldots, a_{k+1}) > 1) \leq (\left\lfloor \frac{u}{2} \right\rfloor + 1)\left(\frac{s+1}{2s}\right)^{k+1}. \tag{4.11}$$

Comparing the right-hand-side of (4.11) with $\left(\frac{1}{2}\right)^{(k+1)/2}$ and solving for $(k+1)$ yields the bound given in the Lemma. Note that this is where we need $s \geq 3$. □

Remark 4.5. *More precisely, we have shown that*

$$\Pr(\gcd(a_1, \ldots, a_{k+1}) > 1) \leq \left(\frac{1}{2}\right)^{(k+1)\delta}$$

for $\delta < \log_2(\frac{s}{s+1}) + 1$ *and* $k \geq \frac{\log_2(\lfloor \frac{u}{2} \rfloor + 1)}{\log_2(\frac{s}{s+1}) + (1-\delta)}$.

We only fixed $\delta = 1/2$ *so as not to have yet another variable in Lemma 4.4.*

Next, for given reduction factor y, we want to derive a bound on the first term of Expression (4.10), i.e.:

$$\Pr\left(\frac{\|\boldsymbol{b}_{k+1}^*\|^2}{\|\boldsymbol{b}_k^*\|^2} < y \mid \gcd(a_1, \ldots, a_{k+1}) = 1\right).$$

Showing that the ratio between $\|\boldsymbol{b}_{k+1}^*\|^2$ and $\|\boldsymbol{b}_k^*\|^2$ behaves the way we suspect is not straightforward as the two quantities are not independent. To estimate the mean of this ratio we use a result by Pittenger [72], which we state below in a form that is adapted to our situation.

Theorem 4.6 ([72], adapted)**.** *Let* X *be a random variable on some positive domain. Choose* $c > 0$ *such that* $X - c \geq 0$ *and define* $\mu = \mathbb{E}[X]$ *and* $\sigma^2 = Var(X)$*. Then*

$$\begin{aligned} \frac{1}{\mu} &\leq \mathbb{E}\left[\frac{1}{X}\right] \\ &\leq \frac{\mu^3 c - 3\mu^2 c^2 + 3\mu c^3 - c^4 + \sigma^2\mu^2 - \sigma^2\mu c + \sigma^4}{\mu^4 c - 3\mu^3 c^2 + 3\mu^2 c^3 - \mu c^4 + 2\sigma^2\mu^2 c - 3\sigma^2\mu c^2 + \sigma^2 c^3 + \sigma^4 c}. \end{aligned} \tag{4.12}$$

For convenience of notation we define $X_k := \sum_{i=1}^{k} a_i^2$. We first estimate the following mean.

Lemma 4.7. *Let* $a_1, \ldots, a_n$ *be chosen independently and uniformly at random from the set* $[l, \ldots, u]$ *for some integers* $0 < l < u$*, let* $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{n-1}$ *be given as in Lemma 4.1, and let* $1 < k < n$*.*

If $\gcd(a_1, \ldots, a_{k+1}) = 1$*, there exists a function* $f(k) \in \Theta(\frac{1}{k^2})$ *such that*

$$1 + \frac{1}{k} \leq \mathbb{E}\left[\|\boldsymbol{b}_k^*\|^2\right] \leq 1 + \frac{1}{k} + f(k), \tag{4.13}$$

and we can give one such $f(k)$ *explicitly.*

Proof. First note that we can use equation (4.9) which yields

$$\|\boldsymbol{b}_k^*\|^2 = \frac{X_{k+1}}{X_k} = 1 + \frac{a_{k+1}^2}{X_k}.$$

Then, by the linearity of the mean and the independence of the a_i, we get

$$\mathbb{E}\left[\|\boldsymbol{b}_k^*\|^2\right] = 1 + \mathbb{E}[a_{k+1}^2]\mathbb{E}\left[\frac{1}{X_k}\right]. \tag{4.14}$$

We can compute

$$\hat{\mu} := \mathbb{E}[a_i^2] = \frac{1}{u-l}\sum_{j=l}^{u} j^2 \quad \text{and} \quad \hat{\mu}_2 := \mathbb{E}[a_i^4] = \frac{1}{u-l}\sum_{j=l}^{u} j^4$$

for any $i = 1, \ldots, n$. Notice that $\hat{\mu}$ and $\hat{\mu}_2$ are independent of k. The last mean in (4.14), $\mathbb{E}\left[\frac{1}{X_k}\right]$, will be approximated by using Theorem 4.6, where we choose $c = kl^2$.

Observe that $\mathbb{E}[X_k] = k\hat{\mu}$, $\mathbb{E}[X_k - c] = \mathbb{E}[X_k] - c$, and $Var(X_k - c) = Var(X_k)$, and it is easy to check that $Var(X_k) = k\hat{\sigma}^2$, where $\hat{\sigma}^2 := Var(a_i^2) = \hat{\mu}_2 - \hat{\mu}^2$.

If we input this in the inequalities of Theorem 4.6, we obtain:

$$\frac{1}{k\hat{\mu}} \leq \mathbb{E}\left[\frac{1}{X_k}\right] \leq \frac{1}{k\hat{\mu}} + \Theta(\frac{1}{k^2}) \,.$$

The precise calculations are given in Section 4.5. By combining the previous expression with $\hat{\mu} = \mathbb{E}[a_{k+1}^2]$ we obtain the desired bounds for $\mathbb{E}\left[\|\boldsymbol{b}_k^*\|^2\right]$ (see Expression (4.14)):

$$1 + \frac{1}{k} \leq 1 + \mathbb{E}[a_{k+1}^2]\mathbb{E}\left[\frac{1}{X_k}\right] \leq 1 + \frac{1}{k} + \Theta(\frac{1}{k^2}).$$

□

Note that by using Theorem 4.6, we can compute an explicit upper bound in (4.13). We present this upper bound in Section 4.5.

Lemma 4.8. *Let $a_1, \ldots, a_n$ be chosen independently and uniformly at random from the set $[l, \ldots, u]$ for some integers $0 < l < u$. Then for any $1 \leq k < n - 1$ with $\gcd(a_1, \ldots, a_{k+1}) = 1$ we get*

$$\left|1 - \mathbb{E}\left[\|\boldsymbol{b}_{k+1}^*\|^2 / \|\boldsymbol{b}_k^*\|^2\right]\right| = O\left(\frac{1}{k}\right).$$

Proof. Note that under the given condition on k we have

$$\begin{aligned}
\frac{\|\boldsymbol{b}_{k+1}^*\|^2}{\|\boldsymbol{b}_k^*\|^2} = \frac{X_{k+2}}{X_{k+1}} \cdot \frac{X_k}{X_{k+1}} &= \frac{(X_{k+1} + a_{k+2}^2)(X_{k+1} - a_{k+1}^2)}{X_{k+1}^2} \\
&= \frac{X_{k+1}^2 + X_{k+1}(a_{k+2}^2 - a_{k+1}^2) - a_{k+1}^2 a_{k+2}^2}{X_{k+1}^2} \\
&= 1 + \frac{a_{k+2}^2}{X_{k+1}} - \frac{a_{k+1}^2}{X_{k+1}} - \frac{a_{k+1}^2 a_{k+2}^2}{X_{k+1}^2} \\
&= \|\boldsymbol{b}_{k+1}^*\|^2 - a_{k+1}^2 \left(\frac{1}{X_{k+1}} + \frac{a_{k+2}^2}{X_{k+1}^2}\right),
\end{aligned} \tag{4.15}$$

and thus

$$\|\boldsymbol{b}_{k+1}^*\|^2 - u^2 \left(\frac{1}{X_{k+1}} + \frac{a_{k+2}^2}{X_{k+1}^2}\right) \leq \frac{\|\boldsymbol{b}_{k+1}^*\|^2}{\|\boldsymbol{b}_k^*\|^2} \leq \|\boldsymbol{b}_{k+1}^*\|^2 - l^2 \left(\frac{1}{X_{k+1}} + \frac{a_{k+2}^2}{X_{k+1}^2}\right).$$

Observe that a_{k+2}^2 and X_{k+1}^2 are independent. Furthermore, we already derived bounds for $\mathbb{E}[\|\boldsymbol{b}_{k+1}^*\|^2]$ and $\mathbb{E}[1/X_{k+1}]$ in Lemma 4.7 and its proof.

To bound $\mathbb{E}[1/X_{k+1}^2]$ we use Theorem 4.6 in a similar fashion as for $\mathbb{E}[1/X_{k+1}]$. Details of this calculation are given in Section 4.5. The result is that

$$\frac{1}{k^2\hat{\mu}^2 + k(\hat{\mu}_2 - \hat{\mu}^2)} \leq \mathbb{E}\left[\frac{1}{X_k^2}\right] \leq \frac{1}{k^2\hat{\mu}^2} + \Theta(1/k^3).$$

Fitting the parts back together, we get

$$\mathbb{E}\left[\frac{\|\boldsymbol{b}_{k+1}^*\|^2}{\|\boldsymbol{b}_k^*\|^2}\right] \geq 1 + \frac{1}{k+1} - u^2\left(\frac{1}{(k+1)\hat{\mu}} + \frac{\hat{\mu}}{(k+1)^2\hat{\mu}^2 + (k+1)(\hat{\mu}_2 - \hat{\mu}^2)}\right)$$

and

$$\mathbb{E}\left[\frac{\|\boldsymbol{b}_{k+1}^*\|^2}{\|\boldsymbol{b}_k^*\|^2}\right] \leq 1 + \frac{1}{k+1} + \Theta(\frac{1}{k^2}) - l^2\left(\frac{1}{(k+1)\hat{\mu}} + \Theta(\frac{1}{k^2}) + \frac{\hat{\mu}}{(k+1)^2\hat{\mu}^2} + \Theta(\frac{1}{k^3})\right).$$

Therefore, we conclude that

$$1 - \frac{u^2 - \hat{\mu}}{(k+1)\hat{\mu}} - \Theta\left(\frac{1}{k^2}\right) \leq \mathbb{E}\left[\frac{\|\boldsymbol{b}_{k+1}^*\|^2}{\|\boldsymbol{b}_k^*\|^2}\right] \leq 1 + \frac{\hat{\mu} - l^2}{(k+1)\hat{\mu}} + \Theta\left(\frac{1}{k^2}\right). \tag{4.16}$$

□

As with Lemma 4.7, we give explicit upper and lower bounds in Section 4.5.

Returning to Inequality (4.10), we will in fact only need the lower bound for $\mathbb{E}\left[\frac{\|\boldsymbol{b}_{k+1}^*\|^2}{\|\boldsymbol{b}_k^*\|^2}\right]$ from (4.16), to see that for any given reduction factor y we can find a $k(y)$ such that the mean is larger than y for any $k \geq k(y)$. More precisely:

Corollary 4.9. *Let $a_1, \ldots, a_n$ be chosen independently and uniformly at random from the set $[l, \ldots, u]$ for some integers $0 < l < u$, and let $y \in (1/4, 1)$ be fixed. Define $\hat{\mu} := \mathbb{E}[a_i^2]$ and $\hat{\sigma}^2 := Var(a_i^2)$.*

Suppose $k \leq n$ is given, and $\gcd(a_1, \ldots, a_{k+1}) = 1$. If k satisfies

$$1 - \frac{u^2 - \hat{\mu}}{(k+1)\hat{\mu}} - \frac{u^2\hat{\mu}}{(k+1)^2\hat{\mu}^2 + (k+1)\hat{\sigma}^2} > y, \tag{4.17}$$

then $\mathbb{E}\left[\frac{\|\boldsymbol{b}_{k+1}^*\|^2}{\|\boldsymbol{b}_k^*\|^2}\right] > y$.

Note that (4.17) can be solved explicitly for $k+1$, giving us a lower bound on k. We omit this calculation here as the solution is long and does not seem illuminating as to what size is sufficient for k. We will give some examples for given l, u, and y in Section 4.4.

If we can now also control the probability of $\|\boldsymbol{b}_{k+1}^*\|/\|\boldsymbol{b}_k^*\|$ deviating by more than a small amount from its mean for given $\boldsymbol{a}$, we have found a bound on the first term in the right-hand side of Inequality (4.10). For this we apply the inequality of Azuma-Hoeffding (cf. [8, 48]):

Let $Z_1, \ldots, Z_N$ be independent random variables, where Z_i takes values in the space Λ_i, and let $f : \prod_{i=1}^N \Lambda_i \to \mathbb{R}$. Define the following Lipschitz condition for the numbers $c_1, \ldots, c_N$:

(L) If the vectors $z, z' \in \prod_{i=1}^N \Lambda_i$ differ only in the j^{th} coordinate, then $|f(z) - f(z')| \leq c_j$, for $j = 1, \ldots, N$.

Theorem 4.10 (see [49])**.** *If f is measurable and satisfies (L), then, for any $t \geq 0$, the random variable $X = f(Z_1, \ldots, Z_N)$ satisfies*

$$\begin{aligned} \Pr(X \geq \mathbb{E}[X] + t) &\leq e^{\frac{-2t^2}{\sum_{i=1}^N c_i^2}} \quad \textit{and} \\ \Pr(X \leq \mathbb{E}[X] - t) &\leq e^{\frac{-2t^2}{\sum_{i=1}^N c_i^2}}. \end{aligned} \tag{4.18}$$

Thus, we indeed have a bound on the probability that a random variable satisfying (L) will deviate more than a little bit from its mean. Note that the bound gets stronger if we find small c_i and choose t large. As with Lemma 4.8, we will ultimately just need one of the bounds, in this case (4.18). Applied to our situation, we obtain the following result.

Theorem 4.11. *Let $a_1, \ldots, a_n$ be chosen independently and uniformly at random from the set $[l, \ldots, u]$ for some integers $0 < l < u$, and let $y \in (1/4, 1)$ be fixed.*

Suppose $k < n$ is given, and $\gcd(a_1, \ldots, a_{k+1}) = 1$. If k satisfies (4.17), then

$$\Pr\left(\frac{\|\boldsymbol{b}_{k+1}^*\|^2}{\|\boldsymbol{b}_k^*\|^2} \leq y\right) \quad \leq \quad e^{-t^2 k\hat{c}},$$

where $\hat{c} > 0$ depends on u and l, and $t > 0$ depends on u, l, and y.

Sketch. This proof contains the main ideas and we refer to Section 4.5 for all technical details.

Set $N = k + 2$, $\Lambda = \Lambda_i = \{l^2, (l+1)^2 \ldots, u^2\}$, let Z_i be uniformly distributed, and set

$$f(\boldsymbol{z}) = 1 + \frac{z_N}{\sum_{i=1}^{N-1} z_i} - \frac{z_{N-1}}{\sum_{i=1}^{N-1} z_i} - \frac{z_{N-1}z_N}{(\sum_{i=1}^{N-1} z_i)^2}$$

for any $\boldsymbol{z} \in \Lambda^N$. Recall from (4.15) that under the given conditions this represents $\frac{\|\boldsymbol{b}_{k+1}^*\|^2}{\|\boldsymbol{b}_k^*\|^2}$.

For $j \in \{1, \ldots, N\}$, let $\boldsymbol{z}, \boldsymbol{z}' \in \Lambda^N$ with $z_i = z_i'$ for $i \neq j$ and $z_j \neq z_j'$. For $j = N$ it is not hard to see that

$$|f(\boldsymbol{z}) - f(\boldsymbol{z}')| \quad \leq \quad \frac{(u^2 - l^2)(N-2)}{(N-1)^2 l^2}$$

for $N > 3$. Similarly, using some standard calculus machinery, one can show that for $j = N-1$ we get

$$|f(\boldsymbol{z}) - f(\boldsymbol{z}')| \quad \leq \quad \frac{u^2 - l^2}{(N-1)l^2} - \frac{u^2}{(N-1)^2 l^2} + \frac{u^4}{((N-2)l^2 + u^2)^2},$$

again assuming only that $N > 3$.

If $j \leq N-2$, it is straightforward that $|f(\boldsymbol{z}) - f(\boldsymbol{z}')| \in O(\frac{1}{N})$, but when trying to find a good upper bound, we run into the trouble that this difference is not a monotone function in all z_i. Therefore, the maximum might not be obtained at the boundary, and the outcome differs depending on the relative distance between l and u, as well as on their cardinality.

However, for "typical" values of l and u (i.e., where l and u are not both large but close together), the above difference is monotone, and we can compute the maximum. See Section 4.4 for examples.

We conclude that f satisfies (L) for constants c_i with

$$\sum_{i=1}^{N} c_i^2 \in \left((N-2)O(\frac{1}{N})^2 + O(\frac{1}{N})^2 + O(\frac{1}{N})^2\right) = O(1/N),$$

which means that $\frac{1}{\sum_{i=1}^{N} c_i^2} \in \Omega(k+1)$.

To finish the proof, let $T(k)$ be the left hand side in (4.17). Note that $T(k) \leq T(k')$ for $k < k'$ and under the given conditions on k we can set $t := T(k) - y > 0$. As

$$\begin{aligned}\Pr\left(\frac{\|\boldsymbol{b}_{k+1}^*\|^2}{\|\boldsymbol{b}_k^*\|^2} \leq y\right) &= \Pr\left(\frac{\|\boldsymbol{b}_{k+1}^*\|^2}{\|\boldsymbol{b}_k^*\|^2} \leq T(k) - t\right) \\ &\leq \Pr\left(\frac{\|\boldsymbol{b}_{k+1}^*\|^2}{\|\boldsymbol{b}_k^*\|^2} \leq \mathbb{E}\left[\frac{\|\boldsymbol{b}_{k+1}^*\|^2}{\|\boldsymbol{b}_k^*\|^2}\right] - t\right),\end{aligned}$$

we are done by using (4.18) from Theorem 4.10. Note that the last inequality holds by Corollary 4.9. □

To summarize, we proved in Lemma 4.4 and in Theorem 4.11 that for fixed reduction factor $y \in (1/4,\ 1)$, and for fixed l, u the following holds:

$$\Pr(\gcd(a_1, \ldots, a_{k+1}) > 1) \leq \left(\frac{1}{2}\right)^{(k+1)/2} \text{ for any } k \geq \frac{\log_2\left(\left\lfloor \frac{u}{2} \right\rfloor + 1\right)}{\log_2\left(\frac{u-l+1}{u-l+2}\right) + \frac{1}{2}} \tag{4.19}$$

and,

$$\Pr\left(\|\boldsymbol{b}^*_{k+1}\|^2 < y\|\boldsymbol{b}^*_k\|^2 \mid \gcd(a_1, \ldots, a_{k+1}) = 1\right) \ \leq \ e^{-t^2(k+1)\hat{c}}, \tag{4.20}$$

where $\hat{c} > 0$ depends on u and l, and $t > 0$ depends on u, l, and y. Adding the right-hand sides of Inequalities (4.19) and (4.20) yields the upper bound on $\Pr\left(\|\boldsymbol{b}^*_{k+1}\|^2/\|\boldsymbol{b}^*_k\|^2 \leq y\right)$ as stated in Theorem 4.3.

4.3 Discussion

If we again look at a basis $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k$ that is obtained by applying the LLL reduction algorithm to an input basis of the format described in Lemma 4.1 in Section 4.1, we showed that for not too small k it will most likely have the following structure:

$$\left(\begin{array}{c|c} X_1 & X_2 \\ \hline \mathbf{0} & X_3 \end{array}\right).$$

The dimension of the submatrices X_1, X_2 and X_3 are $(k+1) \times k$, $(k+1) \times (n-(k+1))$, and $(n-(k+1)) \times (n-(k+1))$ respectively. All the elements of X_1 and X_2 may be non-zero, and X_3 is upper triangular.

In our computations, however, we see even more structure in the reduced basis, as discussed in the introduction. More precisely, we observe a reduced basis of the following form:

$$\left(\begin{array}{c|c} X_1 & \bar{X}_2 \\ \hline \mathbf{0} & I \end{array}\right), \tag{4.21}$$

that is, $X_3 = I$. So, a remaining question to address is why this is the case. We pointed out in Section 4.1 that if $\gcd(a_1, \ldots, a_{k+1}) = 1$, then it follows from the proof of Lemma 4.1 that the last nonzero element in each of the columns $\boldsymbol{b}_{k+1}, \ldots, \boldsymbol{b}_{n-1}$ must be ± 1. Therefore we know that the first column of X_3 is $(1, 0, \ldots, 0)^\mathsf{T}$. The second column of X_3 is $(x, 1, 0, \ldots, 0)^\mathsf{T}$, and so on. Here, again, x just denotes that the element may be non-zero. So, by subtracting x times vector $\boldsymbol{b}_{k+1}$ from vector $\boldsymbol{b}_{k+2}$ yields a unit column $(0, 1, 0, \ldots, 0)^\mathsf{T}$ as the second column of X_3. This procedure can now be repeated for the remaining basis vectors to produce $X_3 = I$. Notice that these operations are elementary column operations.

Observation 4.12. *If we apply column operations as described above to the basis given in Lemma 4.1, then every part of the analysis where we assumed the basis to be given as in Lemma 4.1 also works for this new lattice basis.*

So, indeed, $\ker_{\mathbb{Z}}(\boldsymbol{a})$ has a basis of the structure given in (4.21), and we observe in our computational experiments that such a basis is y-reduced if the input vector $\boldsymbol{a}$ satisfies the assumptions given in the beginning of Section 4.2. Here we give qualitative arguments for why this is the case.

Suppose that the elementary column operations performed to obtain $X_3 = I$ yield a basis that is not size reduced. Then we can add any linear integer combination of the first k basis vectors to any of the last $n - (k + 1)$ vectors without destroying the identity matrix structure of submatrix X_3, since the first k vectors have zeros as the last $n - (k + 1)$ elements. These elementary column operations can be viewed as size reductions. If we consider the first k basis vectors we empirically observe that the absolute values of the non-zero elements (i.e., elements in submatrix X_1) are small, and that the vectors are almost orthogonal since they are reduced. Since all a_i-elements are positive, each basis vector has a mixture of positive, negative and zero elements. Apparently, once these size reductions are done, the basis is reduced, i.e., no further swaps are needed. This is in line with the results presented in Subsection 4.2 that the expected length of the Gram-Schmidt vectors $\boldsymbol{b}_k^*$ becomes arbitrarily close to one with increasing values of k, see also reduction Condition (4.5).

4.4 Computations

4.4.1 Single-row instances

We now present some computational indications and start with the case of $m = 1$, that is, the matrix $\boldsymbol{A}$ consists of one row. This is the case to which our analysis has been applied. We consider two classes of input vectors $\boldsymbol{a}$, namely a_i drawn from intervals $[l, \ldots, u] = [100, \ldots, 1,000]$ and $[l, \ldots, u] = [15,000, \ldots, 150,000]$ and three different instance sizes: $n = 50,\ 100,\ 200$. For each size and input interval we have generated ten instances. In the two sets of columns we report on the average number of "dense" rows of a y-reduced basis $\boldsymbol{Q}$, and the minimum and maximum number of dense rows for the given instance size. The number of "non-dense" rows is computed as follows. Starting from the last row of the reduced basis $\boldsymbol{Q}$, going in the order of decreasing row indices, we count the number of *subsequent rows* that have just one element equal to one in it and all other elements equal to zero. The rest of the rows are counted as "dense". So, a row with just one '1' and the rest zeros is counted as dense if there is a row with higher index that contains 2 or more non-zeros. A row with a single element '1' and the rest zeros correspond to an x-variable just being substituted by a λ-variable in the Reformulation (4.1).

In Table 4.2 we give an upper bound on $\Pr(\gcd(a_1, \ldots, a_{k+1}) > 1)$ for k greater than or equal to the value given in the table. This probability is computed according to Lemma 4.4 for the intervals $[l, \ldots, u] = [100, \ldots, 1,000]$ and $[l, \ldots, u] = [15,000, \ldots, 150,000]$. That is, for the interval $[l, \ldots, u] = [100, \ldots, 1,000]$, the probability that $\gcd(a_1, \ldots, a_{k+1}) > 1$ is less than or equal to 0.0014 for $k \geq 19$. Notice that this value of k is only depending on l and u, and not on n.

In Table 4.2 we also give the value of $k(y)$ for reduction factor $y = 95/100$ such that $\mathbb{E}\left[\|\boldsymbol{b}_{k+1}^*\|^2 / \|\boldsymbol{b}_k^*\|^2\right] > y$ for all $k \geq k(y)$. The values given in the table are very close to the values we observe empirically. The values of k for which the global probability of interchanging basis vectors is small is not equally close to the outcome of our experiments. This is not so surprising given the generality of the Azuma-Hoeffding inequality.

Table 4.1: Results for input vectors $\boldsymbol{a}$ with a_i drawn independently and uniformly at random from the interval $[l, \ldots, u] = [100, \ldots, 1,000]$ and from the interval $[l, \ldots, u] = [15,000, \ldots, 150,000]$.

	$l = 100, \quad u = 1,000$			$l = 15,000, \quad u = 150,000$		
n	average # dense rows	min # dense rows	max # dense rows	average # dense rows	min # dense rows	max # dense rows
50	22.4	18	28	28.6	26	32
100	24.1	19	33	30.2	26	36
200	26.7	20	40	31.1	27	44

Table 4.2: Column two gives an upper bound on $\Pr(\gcd(a_1, \ldots, a_{k+1}) > 1)$ for k greater than or equal to the value given in column 3, cf. Lemma 4.4. In the fourth column we give the value of $k(y)$ for reduction factor $y = 95/100$, such that $\mathbb{E}\left[\|\boldsymbol{b}^*_{k+1}\|^2/\|\boldsymbol{b}^*_k\|^2\right] > y$ for all $k \geq k(y)$. The last two columns give the smallest k for which the bounds of Lemma 4.4 and Theorem 4.11 guarantee $\Pr\left(\|\boldsymbol{b}^*_{k+1}\|^2/\|\boldsymbol{b}^*_k\|^2 \leq y\right) \leq \varepsilon$ for $y = 95/100$.

Interval	Probability $\leq$	$k \geq$	$k(y)$	$k(\varepsilon = 0.05)$	$k(\varepsilon = 0.01)$
$[100, \ldots, 1,000]$	0.0014	19	36	4864	5788
$[15,000, \ldots, 150,000]$	0.000008	34	36	4864	5788

The empirical indication we observe from Table 4.1 is that for larger instances, only relatively few of the $\boldsymbol{x}$-variables have a non-trivial translation into $\boldsymbol{\lambda}$-variables. This is well in line with the theoretical result reported in Table 4.2 that the expected value of $\|\boldsymbol{b}^*_{k+1}\|^2/\|\boldsymbol{b}^*_k\|^2$ is greater than the reduction factor for all $k \geq 36$ for both of the considered intervals. Yet, we observe that if we solve the instances using Reformulation (4.2) rather than the original formulation (eq-IP), the number of branch-and-bound nodes needed in $\boldsymbol{\lambda}$-space could be one to two orders of magnitude smaller than in the original space. Thus, there is a computationally important structure in the $\boldsymbol{\lambda}$-space, but this structure is not arbitrarily "spread", but contained in a limited subset of the variables.

Suppose now that a row $\boldsymbol{a}\boldsymbol{x} = b$ is part of a larger problem formulation, and that we expect this row to be important in the formulation in the sense of obtaining a good branching direction or a useful cut. If we wish to obtain this information through the lattice reformulation, then we need to be careful in indexing the $\boldsymbol{x}$-variables appropriately.

4.4.2 Multi-row instances

We also present computational results for multi-row instances. Our analysis does not extend to this case, but the computations indicate that a similar situation as for the single-row case applies. In Table 4.3 we report on the number of dense rows for instances with $m = 5$ and $n = 50,\ 100,\ 200$. For each size we generated ten instances. The elements a_{ij} of the $\boldsymbol{A}$-matrix are generated from the interval $[l, \ldots, u] = [100, \ldots, 1,000]$. As for the single-row instances in Table 4.1 the results in Table 4.3 indicate that the number of dense rows is not much affected by the increase in the number of variables.

Table 4.3: Results for matrices $\boldsymbol{A} = [a_{ij}]$ with $m = 5$, $n = 50,\ 100,\ 200$, and a_{ij} drawn independently and uniformly at random from the interval $[l, \ldots, u] = [100, \ldots, 1,000]$.

$m \times n$	average # dense rows	min # dense rows	max # dense rows
5×50	48.7	46	50
5×100	53.8	46	66
5×200	52.5	49	65

We do, however, observe that for a given interval $[l, \ldots, u]$ and number of variables, the number of dense rows becomes larger if the number m of rows of $\boldsymbol{A}$ increases. We have therefore also considered instances of sizes $2 \times 50,\ 3 \times 50$, and 4×50 in Table 4.4 in order to compare to the results for instances of size 1×50 in Table 4.1 and of size 5×50 in Table 4.3. For each of the sizes in Table 4.4 we have generated five instances.

Table 4.4: Results for matrices $\boldsymbol{A} = [a_{ij}]$ with $n = 50$ and $m = 2,\ 3,\ 4$, and a_{ij} drawn independently and uniformly at random from the interval $[l, \ldots, u] = [100, \ldots, 1,000]$.

$m \times n$	average # dense rows	min # dense rows	max # dense rows
2×50	36.8	31	44
3×50	40.2	33	48
4×50	45.2	41	48

Finally, we also generated market split instances of various sizes since they are recognized for their difficulty. These instances were proposed by Cornuéjols and Dawande [26] and are generated as follows. For a given number m of rows, the number of variables n is equal to $10(m - 1)$. The elements a_{ij} of the matrix $\boldsymbol{A}$ are generated from the interval $[l, \ldots, u] = [0, \ldots, 99]$. To get instances comparable with the ones reported on in Table 4.3 in terms of the number of variables, we let $m = 6,\ 11,\ 21$, which yields instances of sizes $6 \times 50,\ 11 \times 100,\ 21 \times 200$. For each instance size we generated five instances. The results are given in Table 4.5.

Table 4.5: Results for market split instances of sizes $6 \times 50,\ 11 \times 100,\ 21 \times 200$.

$m \times n$	average # dense rows	min # dense rows	max # dense rows
6×50	46.6	44	49
11×100	68.6	65	78
21×200	**200**	**200**	**200**

Here we in particular observe that the instances of size 21×200 give different results than expected. If we look at the output we notice that for the first 90-100 rows of the reduced basis

$\boldsymbol{Q}$ the number of non-zeros is in the order of hundred per row. Then the number of non-zeros drops quite sharply to the order of 10-20 for the remaining rows, but there are basically no rows with just one nonzero. This seems related to the fact that the number of variables (200) is much larger than the number of integers in the interval (100). It remains to be investigated how difficult these instances are to solve both in the original space and in the reformulated space.

4.4.3 Solving Instances

So far we have just been concerned with the structure of the reduced basis $\boldsymbol{Q}$, without addressing the question of the influence of this structure on the running time when actually solving the instances. To try to give an answer to this question we used instances with up to 50 variables, to make sure that on the one hand the effect of the sparse part occured, and on the other hand the instances were still solvable with standard methods within an acceptable time-limit.

For a fixed instance, we permuted the order of the coefficients $a_1, \ldots, a_n$, such that they were in increasing, decreasing, or random order, and compared how quickly the reformulated instance with reduced $\boldsymbol{Q}$ can be solved by CPLEX in the standard setting. There was no consistent improvement for one order over the others, and also the number of dense rows did not change much when given the same instance with re-ordered coefficients.

Another interesting observation is that if we reduce less strictly, i.e., with smaller reduction factor y, this has little to no influence on the running time, but we get slightly more dense rows. A possible explanation for this behavior could be that, since the coefficients are chosen uniformly at random, already a relatively small amount can act as representatives of the instance.

One behavior that was consistent is that all instances were solved much faster when reformulated with a reduced $\boldsymbol{Q}$, in comparison to the original formulation in n variables.

Acknowledgement

We wish to acknowledge the fruitful discussion with Andrea Lodi and Laurence Wolsey that lead to the question addressed in this contribution. We also want to thank Hendrik Lenstra for his helpful suggestions, and Andrea Lodi for inspiring the section on solving the instances.

4.5 Notes

Details on Lemma 4.7. Recall $\hat{\mu} := \mathbb{E}[a_i^2]$, $X_k := \sum_{i=1}^{k} a_i^2$, and $\hat{\sigma} := Var(a_i^2)$.

Then $\mathbb{E}[X_k] = k\hat{\mu}$ and $Var(X_k) = k\hat{\sigma}$. If we set $X = X_k$ and $c = kl^2$ and input all this into (4.12), then simplifying for k results in

$$\begin{aligned}\frac{1}{k\hat{\mu}} &\leq \mathbb{E}\left[\frac{1}{X_k}\right] \\ &\leq \frac{k^2(\hat{\mu}^3 l^2 - 3\hat{\mu}^2 l^4 + 3\hat{\mu} l^6 - l^8) + k(\hat{\sigma}^2\hat{\mu}^2 - \hat{\sigma}^2\hat{\mu} l^2) + \hat{\sigma}^4}{k^3(\hat{\mu}^4 l^2 - 3\hat{\mu}^3 l^4 + 3\hat{\mu}^2 l^6 - \hat{\mu} l^8) + k^2(2\hat{\sigma}^2\hat{\mu}^2 l^2 - 3\hat{\sigma}^2\hat{\mu} l^4 + \hat{\sigma}^2 l^6) + k\hat{\sigma}^4 l^2}\,.\end{aligned}$$

Details on Lemma 4.8. We want to set $X = X_k^2$ in (4.12), hence we need to know $\mathbb{E}[X_k^2]$ and $Var(X_k^2)$. To this end we state the following useful equation.

Proposition 4.13. *Let $x_i \in \mathbb{R}$ for $i \in [1, \dots, m]$. Then*

$$\left(\sum_{i=1}^{m} x_i\right)^4 = \sum_{i=1}^{m} x_i^4 + 4\sum_{i\neq j} x_i^3 x_j + 6\sum_{i<j} x_i^2 x_j^2 + 12\sum_{\substack{j<k \\ j,k\neq i}} x_i^2 x_j x_k + 24\sum_{i<j<k<l} x_i x_j x_k x_l\,.$$

Proof. This is an easy exercise in induction, where one first shows that

$$\left(\sum_{i=1}^{m} x_i\right)^3 = \sum_{i=1}^{m} x_i^3 + 3\sum_{i\neq j} x_i^2 x_j + 6\sum_{i<j<k} x_i x_j x_k\,.$$

□

Now compute

$$\mathbb{E}[X_k^2] = \sum_{i=1}^{k} \mathbb{E}[a_i^4] + 2\sum_{i<j} \mathbb{E}[a_i]^2 = k\mathbb{E}[a_i^4] + k(k-1)\hat{\mu}^2 = k^2\hat{\mu}^2 + k\hat{\sigma}^2\,,$$

and for $Var(X_k^2)$ we compute

$$\begin{aligned}
\mathbb{E}[X_k^4] &= \mathbb{E}[\sum_{i=1}^{k}(a_i^2)^4 + 4\sum_{i\neq j}(a_i^2)^3 a_j^2 + 6\sum_{i<j}(a_i^2)^2(a_j^2)^2 \\
&\qquad + 12\sum_{\substack{j<k \\ j,k\neq i}}(a_i^2)^2 a_j^2 a_k^2 + 24\sum_{i<j<k<l} a_i^2 a_j^2 a_k^2 a_l^2] \\
&= k\mathbb{E}[a_i^8] + 4k(k-1)\mathbb{E}[a_i^6]\hat{\mu} - 6\frac{k(k-1)}{2}\hat{\mu}_2^2 \\
&\quad + 12\frac{k(k-1)(k-2)}{2}\hat{\mu}_2\hat{\mu}^2 + 24\frac{k(k-1)(k-2)(k-3)}{24}\hat{\mu}^4 \\
&= k^4\hat{\mu}^4 + k^3 6\hat{\mu}^2(\hat{\mu}_2 - \hat{\mu}^2) + k^2(4\mathbb{E}[a_i^6]\hat{\mu} + 3\hat{\mu}_2^2 - 18\hat{\mu}_2\hat{\mu}^2 + 11\hat{\mu}^4) \\
&\quad + k(\mathbb{E}[a_i^8] - 4\mathbb{E}[a_i^6]\hat{\mu} - 3\hat{\mu}_2^2 + 12\hat{\mu}_2\hat{\mu}^2 - 6\hat{\mu}^4)\,.
\end{aligned}$$

Therefore,

$$\begin{aligned}
Var(X_k^2) &= \mathbb{E}[X_k^4] - \mathbb{E}[X_k^2]^2 \\
&= k^4\hat{\mu}^4 + k^3 6\hat{\mu}^2(\hat{\mu}_2 - \hat{\mu}^2) + k^2(4\mathbb{E}[a_i^6]\hat{\mu} + 3\hat{\mu}_2^2 - 18\hat{\mu}_2\hat{\mu}^2 + 11\hat{\mu}^4) \\
&\quad + k(\mathbb{E}[a_i^8] - 4\mathbb{E}[a_i^6]\hat{\mu} - 3\hat{\mu}_2^2 + 12\hat{\mu}_2\hat{\mu}^2 - 6\hat{\mu}^4) \\
&\quad - (k^4\hat{\mu}^4 + k^3 2\hat{\mu}^2(\hat{\mu}_2 - \hat{\mu}^2) + k^2(\hat{\mu}_2 - \hat{\mu}^2)^2) \\
&= k^3 4\hat{\mu}^2(\hat{\mu}_2 - \hat{\mu}^2) + k^2(4\mathbb{E}[a_i^6]\hat{\mu} + 2\hat{\mu}_2^2 - 16\hat{\mu}_2\hat{\mu}^2 + 10\hat{\mu}^4) \\
&\quad + k(\mathbb{E}[a_i^8] - 4\mathbb{E}[a_i^6]\hat{\mu} - 3\hat{\mu}_2^2 + 12\hat{\mu}_2\hat{\mu}^2 - 6\hat{\mu}^4) \\
&= k^3 4\hat{\sigma}^2\hat{\mu}^2 + k^2(2\hat{\sigma}^4 - 8\hat{\sigma}^2\hat{\mu}^2 - 4\hat{\mu}_2\hat{\mu}^2 + 4\mathbb{E}[a_i^6]\hat{\mu}) \\
&\quad + k(\mathbb{E}[a_i^8] - 4\mathbb{E}[a_i^6]\hat{\mu} - 3\hat{\sigma}^4 + 3\hat{\sigma}^2\hat{\mu}^2 + 3\hat{\mu}_2\hat{\mu}) \\
&=: k^3 p_3 + k^2 p_2 + k p_1\,,
\end{aligned}$$

and finally, by using this expression in (4.12), we get $\frac{1}{k^2\hat{\mu}^2+k\hat{\sigma}^2} \leq \mathbb{E}\left[\frac{1}{X_k^2}\right]$ and

$$\mathbb{E}\left[\frac{1}{X_k^2}\right] \leq \frac{k^6\hat{p}_6 + k^5\hat{p}_5 + k^4\hat{p}_4 + k^3\hat{p}_3 + k^2\hat{p}_2 + k\hat{p}_1 + \hat{p}_0}{k^8\hat{q}_8 + k^7\hat{q}_7 + k^6\hat{q}_6 + k^5\hat{q}_5 + k^4\hat{q}_4 + k^3\hat{q}_3 + k^2\hat{q}_2},$$

where

$$\begin{aligned}
\hat{p}_6 &= l^4\hat{\mu}^6 - 3l^8\hat{\mu}^4 + 3l^{12}\hat{\mu}^2 - l^{16},\\
\hat{p}_5 &= 3l^4\hat{\mu}^4\hat{\sigma}^2 - 6l^8\hat{\mu}^2\hat{\sigma}^2 + 3l^{12}\hat{\sigma}^2 + p_3\hat{\mu}^4 - p_3l^4\hat{\mu}^2,\\
\hat{p}_4 &= 3l^4\hat{\mu}^2\hat{\sigma}^4 - 3l^8\hat{\sigma}^4 + 2p_3\hat{\mu}^2\hat{\sigma}^2 + p_2\hat{\mu}^4 - p_3l^4\hat{\sigma}^2 - p_2l^4\hat{\mu}^2 + p_3^2,\\
\hat{p}_3 &= l^4\hat{\sigma}^6 + p_3\hat{\sigma}^4 + 2p_2\hat{\mu}^2\hat{\sigma}^2 + p_1\hat{\mu}^4 + p_2l^4\hat{\sigma}^2 - p_1l^4\hat{\mu}^2 + 2p_2p_3,\\
\hat{p}_2 &= p_2\hat{\sigma}^4 + 2p_1\hat{\mu}^2\hat{\sigma}^2 - p_1l^4\hat{\sigma}^2 + 2p_1p_3 + p_2^2,\\
\hat{p}_1 &= p_1\hat{\sigma}^4 + 2p_1p_2,\\
\hat{p}_0 &= p_1^2,\\
\hat{q}_8 &= l^4\hat{\mu}^8 - 3l^8\hat{\mu}^6 + 3l^{12}\hat{\mu}^4 - l^{16}\hat{\mu}^2,\\
\hat{q}_7 &= 4l^4\hat{\mu}^6\hat{\sigma}^2 - 9l^8\hat{\mu}^4\hat{\sigma}^2 + 6l^{12}\hat{\mu}^2\hat{\sigma}^2 - l^{16}\hat{\sigma}^2 + 2p_3l^4\hat{\mu}^4 - 3p_3l^8\hat{\mu}^2 + p_3l^{12},\\
\hat{q}_6 &= 6l^4\hat{\mu}^4\hat{\sigma}^4 - 9l^8\hat{\mu}^2\hat{\sigma}^4 + 3l^{12}\hat{\sigma}^4 + 4p_3l^4\hat{\mu}^2\hat{\sigma}^2 + 2p_2l^4\hat{\mu}^4 - 3p_3l^8\hat{\sigma}^2\\
&\quad - 3p_2l^8\hat{\mu}^2 + p_2l^{12} + p_3^2l^4\\
\hat{q}_5 &= 4l^4\hat{\mu}^2\hat{\sigma}^6 - 3l^8\hat{\sigma}^6 + 2p_3l^4\hat{\sigma}^4 + 4p_2l^4\hat{\mu}^2\hat{\sigma}^2 + 2p_1l^4\hat{\mu}^4 - 3p_2l^8\hat{\sigma}^2\\
&\quad - 3p_1l^8\hat{\mu}^2 + p_1l^{12} + 2p_2p_3l^4\\
\hat{q}_4 &= l^4\hat{\sigma}^8 + 2p_2l^4\hat{\sigma}^4 + 4p_1l^4\hat{\mu}^2\hat{\sigma}^2 - 3p_1l^8\hat{\sigma}^2 + 2p_1p_3l^4 + p_2^2l^4,\\
\hat{q}_3 &= 2p_1l^4\hat{\sigma}^4 + 2p_1p_2l^4,\\
\hat{q}_2 &= p_1^2l^4.
\end{aligned}$$

Note that $\hat{q}_8 = \hat{\mu}^2\hat{p}_6$. It might also be of interest that $\hat{\sigma}^2$ neither appears in the leading term of the numerator nor of the denominator.

Proof of Theorem 4.11

We want to compute good constants for which $\frac{\|\boldsymbol{b}_{k+1}^*\|^2}{\|\boldsymbol{b}_k^*\|^2}$ satisfies the condition (L) under the conditions of Theorem 4.11. As in the proof, we set

$$f(z) = 1 + \frac{z_N - z_{N-1}}{\sum^{N-1} z_i} - \frac{z_N z_{N-1}}{(\sum^{N-1} z_i)^2},$$

where $z_j \in \{l^2, \dots, u^2\}$.

The goal is to get upper bounds on $|f(z) - f(z')|$, with $z_j = z'_j$ for all but one $j \in \{1, \dots, N\}$.

We will distinguish between three cases: z and z' differ in the last, the second-to-last, or some other coordinate $j < N-1$.

For better readability, set $x_1 := z_N$, $x_2 := z_{N-1}$, $x_3 := \sum_{i\neq j}^{N-2} z_i$, and $x_4 := z_j$ (and define x'_i similarly). Thus, we are looking at the function

$$f(x) = 1 + \frac{x_1 - x_2}{x_2 + x_3 + x_4} - \frac{x_1 x_2}{(x_2 + x_3 + x_4)^2}.$$

Whenever we talk about this function, we mean the real-valued function on the domain $[l^2, u^2] \times [l^2, u^2] \times [(N-3)l^2, (N-3)u^2] \times [l^2, u^2]$.

Lemma 4.14. *$f(x) > 0$ for $0 < l < u$.*

Proof. Observe that $f(x) > 0$ is equivalent to

$$\begin{aligned} & (x_2 + x_3 + x_4)^2 + (x_1 - x_2)(x_2 + x_3 + x_4) - x_1x_2 > 0 \\ \Leftrightarrow \quad & x_2^2 + 2x_2(x_3 + x_4) + (x_3 + x_4)^2 + x_1x_2 - x_2^2 + (x_3 + x_4)(x_1 - x_2) - x_1x_2 > 0 \\ \Leftrightarrow \quad & x_2(x_3 + x_4) + (x_3 + x_4)^2 + x_1(x_3 + x_4) > 0\,. \end{aligned}$$

□

Lemma 4.15. *Let $f(x)$ be given as above, and $0 < l < u$, then*

(a) f is increasing in x_1;

(b) f is decreasing in x_2;

(c)
- *If $x_1 \leq x_2$, then f is increasing in x_4;*
- *If $x_1 > x_2$ and $x_3 > x_2\frac{x_1+x_2}{x_1-x_2} - l^2$, then f is decreasing in x_4;*
- *If $x_1 > x_2$ and $x_3 \leq x_2\frac{x_1+x_2}{x_1-x_2} - l^2$, then f has a (for fixed values x_1, x_2, x_3) unique maximum at $x_4 = x_2\frac{x_1+x_2}{x_1-x_2} - x_3$.*

Proof. We compute

$$f_{x_1} = \frac{1}{x_2 + x_3 + x_4} - \frac{x_2}{(x_2 + x_3 + x_4)^2} = \frac{x_3}{(x_2 + x_3 + x_4)^2} > 0\,.$$

This proves part (a).

Also, we can directly compute

$$\begin{aligned} f_{x_2} &= \frac{-(x_2 + x_3 + x_4) - (x_1 - x_2)}{(x_2 + x_3 + x_4)^2} - \frac{x_1(x_2 + x_3 + x_4) - 2x_1x_2}{(x_2 + x_3 + x_4)^3} \\ &= \frac{-(x_1 + x_3 + x_4)(x_2 + x_3 + x_4) - x_1x_2 - x_1(x_3 + x_4) + 2x_1x_2}{(x_2 + x_3 + x_4)^3} \\ &= \frac{-2x_1(x_3 + x_4) - x_2(x_3 + x_4) - (x_3 + x_4)^2}{(x_2 + x_3 + x_4)^3} < 0\,. \end{aligned}$$

This proves part (b).

For part (c), we first compute

$$\begin{aligned} f_{x_4} &= -\frac{x_1 - x_2}{(x_2 + x_3 + x_4)^2} + \frac{2x_1x_2}{(x_2 + x_3 + x_4)^3} \\ &= \frac{-x_1x_2 - x_1x_3 - x_1x_4 + x_2^2 + x_2x_3 + x_2x_4 + 2x_1x_2}{(x_2 + x_3 + x_4)^3} \\ &= \frac{x_1x_2 + x_2^2 - (x_3 + x_4)(x_1 - x_2)}{(x_2 + x_3 + x_4)^3}\,. \end{aligned}$$

Thus, if $x_1 \leq x_2$, then $f_{x_4} > 0$. Assume now that $x_1 > x_2$. Then

$$\begin{aligned} f_{x_4} = 0 &\Leftrightarrow x_2(x_1 + x_2) = (x_3 + x_4)(x_1 - x_2) \\ &\Leftrightarrow x_4 = \frac{x_2(x_1 + x_2)}{x_1 - x_2} - x_3\,, \end{aligned}$$

where the last equivalence holds because $x_1 \neq x_2$. We know that by definition we have $x_4 \geq l^2$, thus if $x_2\frac{x_1+x_2}{x_1-x_2} - x_3 < l^2$, then f_{x_4} is non-zero for all values of x_4 in this part of the domain of f. It is a straightforward computation to see that in fact this extremum is a maximum. □

Case 1: $x_1 \neq x_1'$**.** We may assume $x_1 > x_1'$. Then by Lemma 4.14 and 4.15(a), we know that $|f(x) - f(x')| = f(x) - f(x')$ and the difference is maximized for $x_1 = u^2$ and $x_1' = l^2$.

Observe that

$$f(x) - f(x') = \frac{x_1 - x_2 - (x_1' - x_2)}{x_2 + x_3 + x_4} - \frac{x_2(x_1 - x_1')}{(x_2 + x_3 + x_4)^2} = \frac{(x_1 - x_1')(x_3 + x_4)}{(x_2 + x_3 + x_4)^2}\,. \tag{4.22}$$

Now set $x_5 := x_3 + x_4$ (as these variables always appear paired), and consider $g(x_2, x_5) := \frac{x_5}{(x_2+x_5)^2}$. Then we have $g_{x_2} = \frac{-2x_5}{(x_2+x_5)^3} < 0$, and thus the difference (4.22) is maximal for $x_2 = l^2$.

Furthermore, we have $g_{x_5}(l^2, x_5) = \frac{(l^2+x_5)-2x_5}{(l^2+x_5)^3} = \frac{l^2-x_5}{(l^2+x_5)^3}$, and this is negative if $N > 3$. Thus, for $N > 3$, the difference (4.22) is maximal if $x_3 = (N-3)l^2$ and $x_4 = l^2$.

We conclude that in Case 1 we have (for $N > 3$)

$$|f(x) - f(x')| \leq \frac{(u^2 - l^2)(N-2)l^2}{((N-1)l^2)^2} = \frac{(N-2)(u^2 - l^2)}{(N-1)^2 l^2} =: c_1\,.$$

Case 2: $x_2 \neq x_2'$**.** We may assume $x_2 < x_2'$, then by Lemma 4.14 and 4.15(b), we know that

$$\begin{aligned} |f(x)-f(x')| &= \frac{(x_1 - x_2)(x_2 + x_3 + x_4) - x_1x_2}{(x_2 + x_3 + x_4)^2} - \frac{(x_1 - x_2')(x_2' + x_3 + x_4) - x_1x_2'}{(x_2' + x_3 + x_4)^2} \\ &= \frac{x_1(x_3 + x_4) - x_2(x_2 + x_3 + x_4)}{(x_2 + x_3 + x_4)^2} - \frac{x_1(x_3 + x_4) - x_2'(x_2' + x_3 + x_4)}{(x_2' + x_3 + x_4)^2}\,, \end{aligned} \tag{4.23}$$

and (4.23) is maximal for $x_2 = l^2$ and $x_2' = u^2$.

We again set $x_5 = x_3 + x_4$ and consider the RHS of (4.23) as the function $g(x_1, x_2, x_2', x_5)$. Then

$$g_{x_1} = x_5\left(\frac{1}{(x_2 + x_5)^2} - \frac{1}{(x_2' + x_5)^2}\right) > 0,$$

and thus (4.23) is maximal for $x_1 = u^2$.

To determine x_5, we are interested in

$$\begin{aligned} g_{x_5} =& \frac{(x_1 - x_2)(x_2 + x_5) - 2(x_1x_5 - x_2(x_2 + x_5))}{(x_2 + x_5)^3} \\ &- \frac{(x_1 - x_2')(x_2' + x_5) - 2(x_1x_5 - x_2'(x_2' + x_5))}{(x_2' + x_5)^3} \\ =& \frac{x_5(x_2 - x_1) + x_2(x_1 + x_2)}{(x_2 + x_5)^3} - \frac{x_5(x_2' - x_1) + x_2'(x_1 + x_2')}{(x_2' + x_5)^3}\,. \end{aligned}$$

As we have all the values except the one for x_5, we only need to look at

$$h(x_5) := g_{x_5}(u^2, l^2, u^2, x_5) = \frac{l^2(u^2 + l^2) - x_5(u^2 - l^2)}{(l^2 + x_5)^3} - \frac{u^2(2u^2)}{(u^2 + x_5)^3}.$$

Claim 4.16. *For $x_5 \geq 2l^2$ we have $h(x_5) < 0$.*

Equivalently, we will show

$$\tilde{h}(x_5) := (u^2l^2 + l^4)(u^2 + x_5)^3 - x_5(u^2 - l^2)(u^2 + x_5)^3 - 2u^4(l^2 + x_5)^3 < 0$$

for $x_5 \geq 2l^2$. To this end, observe that

$$\begin{aligned}\tilde{h}(2l^2) =&(u^2l^2 + l^4)(u^2 + 2l^2)^3 - 2(l^2u^2 - l^4)(u^2 + 2l^2)^3 - 2u^4(3l^2)^3\\ =&24l^{10} + 28u^2l^8 - 48u^4l^6 - 3u^6l^4 - u^8l^2 < 52u^2l^8 - 52u^4l^6 < 0\,,\end{aligned}$$

and furthermore that

$$\begin{aligned}\tilde{h}'(x_5) =&(3(u^2l^2 + l^4) - (u^2 - l^2)((u^2 + x_5) - 3x_5(u^2 - l^2))(u^2 + x_5)^2\\ &- 6u^4(l^2 + x_5)^2\\ =&-4(u^2 - l^2)x_5^3 - (15u^4 - 12u^2l^2 - 3l^4)x_5^2 - (6u^6 - 6u^2l^4)x_5\\ &+ 4u^6l^2 - 3u^4l^4 - u^8\,.\end{aligned}$$

As all coefficients except the last one are negative, we know that if $\tilde{h}'(y_0) < 0$ for some $y_0 > 0$, then $\tilde{h}'(y) < 0$ for all $y \geq y_0$. It is straightforward to check that indeed $\tilde{h}'(2l^2) < 0$. Therefore, $\tilde{h}$ is decreasing for $x_5 \geq 2l^2$ and $\tilde{h}(2l^2) < 0$, and Claim 4.16 is proven.

Thus, if $N > 3$ we have to set $x_5 = (N-2)l^2$ to maximize (4.23).

In summary, for Case 2 we get (for $N > 3$)

$$|f(x) - f(x')| \leq \frac{(u^2 - l^2)}{(N-1)l^2} - \frac{u^2}{(N-1)^2l^2} + \frac{u^4}{((N-2)l^2 + u^2)^2} =: c_2\,.$$

Case 3.1: $x_4 \neq x_4'$ and $x_1 \leq x_2$. We can assume $x_4 > x_4'$, then by Lemma 4.14 and 4.15(c), we know that

$$\begin{aligned}|f(x) - f(x')| =&\frac{x_1 - x_2}{x_2 + x_3 + x_4} - \frac{x_1x_2}{(x_2 + x_3 + x_4)^2}\\ &- \frac{x_1 - x_2}{x_2 + x_3 + x_4'} + \frac{x_1x_2}{(x_2 + x_3 + x_4')^2}\,,\end{aligned} \tag{4.24}$$

and this is maximal if $x_4 = u^2$ and $x_4' = l^2$.

We again interpret (4.24) as the function $g(x_1, x_2, x_3, x_4, x_4')$ and compute

$$\begin{aligned}g_{x_1} =&\frac{1}{x_2 + x_3 + x_4} - \frac{x_2}{(x_2 + x_3 + x_4)^2} - \frac{1}{x_2 + x_3 + x_4'} + \frac{x_2}{(x_2 + x_3 + x_4')^2}\\ =&\frac{x_3 + x_4}{(x_2 + x_3 + x_4)^2} - \frac{x_3 + x_4'}{(x_2 + x_3 + x_4')^2}\,.\end{aligned}$$

As we already determined x_4 and x_4', we restrict our attention to

$$g_{x_1}(x_1, x_2, x_3, u^2, l^2) = \frac{x_3 + u^2}{(x_2 + x_3 + u^2)^2} - \frac{x_3 + l^2}{(x_2 + x_3 + l^2)^2} =: h(x_2, x_3)\,.$$

Claim 4.17. *If $N - 3 > \frac{\sqrt{(u^2-l^2)^2+4u^4}-u^2-l^2}{2l^2}$, then $h(x_2, x_3) < 0$.*

Equivalently, we show that

$$\begin{aligned}&(x_3 + u^2)(x_2 + x_3 + l^2)^2 - (x_3 + l^2)(x_2 + x_3 + u^2)^2 < 0\\ \Leftrightarrow\quad &x_3^2 + x_3(u^2 + l^2) + u^2l^2 - x_2^2 > 0.\end{aligned} \tag{4.25}$$

Define $k = N - 3$. Because $x_2 \leq u^2$ and $x_3 \geq (N-3)l^2 = kl^2$, we will get (4.25), if we can show that

$$\begin{aligned} & k^2l^4 + kl^2(u^2 + l^2) + u^2(l^2 - u^2) > 0 \\ \Leftrightarrow \quad & (k + \frac{u^2 + l^2}{2l^2})^2 > \frac{4u^2(u^2 - l^2) + (u^2 + l^2)^2}{4l^4} \\ \Leftrightarrow \quad & k > \frac{\sqrt{4u^4 + (u^2 - l^2)^2} - u^2 - l^2}{2l^2}\,. \end{aligned}$$

Thus we have Claim 4.17. In particular, we conclude that for N large enough, (4.24) is maximal for $x_1 = l^2$.

Next, we investigate

$$\begin{aligned} g_{x_2} =& \frac{-(x_2 + x_3 + x_4) - (x_1 - x_2)}{(x_2 + x_3 + x_4)^2} - \frac{x_1(x_2 + x_3 + x_4) - 2x_1x_2}{(x_2 + x_3 + x_4)^3} \\ & - \frac{-(x_2 + x_3 + x_4') - (x_1 - x_2)}{(x_2 + x_3 + x_4')^2} + \frac{x_1(x_2 + x_3 + x_4') - 2x_1x_2}{(x_2 + x_3 + x_4')^3} \\ =& \frac{-(x_3 + x_4)(2x_1 + x_2 + x_3 + x_4)}{(x_2 + x_3 + x_4)^3} - \frac{-(x_3 + x_4')(2x_1 + x_2 + x_3 + x_4')}{(x_2 + x_3 + x_4')^3}\,. \end{aligned}$$

Claim 4.18. *If* $N > \sqrt{4 + 2\frac{u^2}{l^2} + \frac{u^4}{l^4}}$, *then* $g_{x_2}(l^2, x_2, x_3, u^2, l^2)$ *is positive.*

We show this by showing that the function $h(y) = \frac{(x_3+y)(2x_1+x_2+x_3+y)}{(x_2+x_3+y)^3}$ is decreasing for $y \geq l^2$ (and x_i in the usual interval). For this we compute

$$\begin{aligned} h'(y) =& \frac{(2y + 2x_1 + x_2 + 2x_3)(y + x_2 + x_3)}{(x_2 + x_3 + y)^4} \\ & - \frac{3y^2 + (6x_1 + 3x_2 + 6x_3)y + 3x_3(2x_1 + x_2 + x_3)}{(x_2 + x_3 + y)^4}\,, \end{aligned}$$

which is negative if we can show that

$$y^2 + (4x_1 + 2x_3)y - 2x_1x_2 - x_2^2 + 4x_1x_3 + x_3^2 > 0\,. \tag{4.26}$$

We already know that we will set $x_1 = l^2$. We also know that $y \geq l^2$, $x_2 \leq u^2$, and $x_3 \leq (N-3)l^2$. Thus (4.26) is satisfied in our case, if we have

$$\begin{aligned} & l^4 + 4l^4 + 2(N-3)l^4 - 2u^2l^2 - u^4 + 4(N-3)l^4 + (N-3)^2l^4 > 0 \\ \Leftrightarrow \quad & N^2 > 4 + \frac{2u^2l^2 + u^4}{l^4}, \end{aligned}$$

which is what we claimed.

Therefore, for N as in Claim 4.18, we maximize (4.24) if we set $x_2 = u^2$.

Now that we have all values except x_3, let us consider

$$\hat{g}(x_3) = \frac{l^2 - u^2}{2u^2 + x_3} - \frac{u^2l^2}{(2u^2 + x_3)^2} - \frac{l^2 - u^2}{u^2 + l^2 + x_3} + \frac{u^2l^2}{(u^2 + l^2 + x_3)^2}\,.$$

It is a straightforward computation to see that $\hat{g}'(x_3)$ is equal to

$$-(u^2-l^2)\Big[\frac{2(u^2-l^2)x_3^3+3(3u^4-l^4)x_3^2}{(u^2+l^2+x_3)^3(2u^2+x_3)^3} + \frac{(13u^6+15u^4l^2-3u^2l^4-l^6)x_3+2u^4(3u^4+l^4+8u^2l^2)}{(u^2+l^2+x_3)^3(2u^2+x_3)^3}\Big],$$

and as this is negative (for any $x_3 > 0$ in fact), we set $x_3 = (N-3)l^2$.

Therefore, in Case 3.1 we get (for N large enough)

$$\begin{aligned}|f(x)-f(x')| \leq &(u^2-l^2)\left(\frac{1}{u^2+(N-2)l^2}-\frac{1}{2u^2+(N-3)l^2}\right)\\ &+u^2l^2\left(\frac{1}{(u^2+(N-2)l^2)^2}-\frac{1}{(2u^2+(N-3)l^2)^2}\right) =: c_{3.1}\,.\end{aligned}$$

Case 3.2: $x_4 \neq x_4'$**,** $x_1 > x_2$**, and** $x_3 > x_2\frac{x_1+x_2}{x_1-x_2} - l^2$**.** Using Lemma 4.14 and 4.15(c), and assuming $x_4 < x_4'$, we first note that we get (4.24) again and this is maximal if $x_4 = l^2$ and $x_4' = u^2$.

As before, we use these values and restrict our attention to

$$g_{x_1}(x_1, x_2, x_3, l^2, u^2) = -h(x_2, x_3)\,.$$

Thus, by Claim 4.17, we maximize (4.24) by setting $x_1 = u^2$.

Using this now in a very similar fashion as in Claim 4.18 (with the difference of setting $x_1 = u^2$, and using that x_4 is now smaller than x_4'), we conclude that for $N > \sqrt{8-4\frac{u^2}{l^2}+3\frac{u^4}{l^4}}$, (4.24) is maximized for $x_2 = l^2$.

To get x_3, we again put in all the other values and get that (4.24) is at most equal to

$$\hat{g}_1(x_3) = \frac{u^2-l^2}{2l^2+x_3} - \frac{u^2l^2}{(2l^2+x_3)^2} - \frac{u^2-l^2}{u^2+l^2+x_3} + \frac{u^2l^2}{(u^2+l^2+x_3)^2}\,.$$

It is another straightforward computation to see that $\hat{g}_1'(x_3)$ is equal to

$$-(u^2-l^2)\Big[\frac{2(u^2-l^2)x_3^3+3(u^4-3l^4)x_3^2+(u^6+3u^4l^2-15u^2l^4-13l^6)x_3}{(u^2+l^2+x_3)^3(2l^2+x_3)^3} - \frac{2l^4(u^4+8u^2l^2+3l^4)}{(u^2+l^2+x_3)^3(2l^2+x_3)^3}\Big]. \tag{4.27}$$

The sign of this expression depends on the ratios between l, u, and N.

To go on from here, we will restrict ourselves to the values of l and u we used in the computations. Similar computations are possible for different values, but we chose these because they seem reasonable in their order of magnitude and ratio to get a relevant statement.

Therefore, from here on, we assume that

$$[l,\dots,u] = \begin{cases}[100,\dots,1000] & \text{or}\\ [15000,\dots,150000]\,.\end{cases} \tag{4.28}$$

Observation 4.19. *If l, u are given as in* (4.28)*, then N satisfies all inequalities we came across in Cases 3.1 and 3.2, if $N \geq 173$.*

Observation 4.20. *If l, u are given as in* (4.28)*, and $N \geq 1$, then* (4.27) *is negative.*

Thus, we set $x_3 = (N-3)l^2$. Then, for N large enough, we obtain

$$\begin{aligned}|f(x) - f(x')| \leq &(u^2 - l^2)\left(\frac{1}{(N-1)l^2} - \frac{1}{u^2 + (N-2)l^2}\right) \\ &- u^2l^2\left(\frac{1}{((N-1)l^2)^2} - \frac{1}{(u^2 + (N-2)l^2)^2}\right) =: c_{3.2}\,.\end{aligned}$$

Observation 4.21. *If l, u are given as in* (4.28) *and x_i chosen as above, then $x_3 > x_2\frac{x_1+x_2}{x_1-x_2} - l^2$ for $N \geq 1$.*

Claim 4.22. *If $u \geq 3l$, then $c_{3.1} < c_{3.2}$.*

Solving $c_{3.2} - c_{3.1} > 0$ for N, this reduces to showing that

$$\begin{aligned}l^4(u^2 - l^2)(2u^4 - 8u^2l^2 + 5l^4)N^2 + 2l^2(2u^8 - 14u^6l^2 + 33u^4l^4 - 31u^2l^6 + 10l^8)N \\ +(4u^8l^2 + 38u^6l^4 - 68u^4l^6 + 53u^2l^8 - 15l^{10}) > 0\,.\end{aligned}$$

Notice that the leading coefficient is positive if $2u^4 - 8u^2l^2 + 5l^4 > 0$. Indeed, we observe that

$$2u^4 - 8u^2l^2 + 5l^4 > u^4 - 8u^2l^2 + l^4 = (u^2 - 4l^2)^2 - 15l^4 \geq 0$$

for $u \geq 3l$. One can easily check that this restriction on l, u will also imply that the other coefficients are positive.

Case 3.3: $x_4 \neq x_4'$, $x_1 > x_2$, and $x_3 \leq x_2\frac{x_1+x_2}{x_1-x_2} - l^2$. Observe that

$$x_2\frac{x_1 + x_2}{x_1 - x_2} - l^2 < u^2\frac{2u^2}{u^2 - l^2} - l^2,$$

and for l, u as in (4.28) this is smaller than $202l^2$. Therefore, certainly for $N \geq 205$, this case will not occur.

Chapter Five
Discrete Isoperimetric Sets

The question we want to study here is inspired by a well-known problem from analysis: Given a positive volume, which shape has minimal boundary-size? In a Euclidean space, the answer is of course the ball.

There are many interesting questions one can derive from this for discrete settings. We will mention some of them after we fix some notation, but in particular we refer to the paper of Bezrukov [16] for a more thorough survey. In this chapter, we want to consider the following question: Given the lattice $\mathbb{Z}^n$ and a cardinality M, we want to find lattice point configurations of this size such that the number of lattice points with distance 1 to the set is small. Here and throughout this chapter, distance is measured in the L^1 norm. Let us call sets of cardinality M with the smallest number of such 'neighbors' *optimal*. Wang and Wang proved in [81] that sets of points with coordinate sum less than or equal to some integer k are optimal. We will additionally show that they are unique for their cardinalities.

We will also discuss the question of how to characterize optimal sets in general, and if by adding the points of distance 1 to an optimal set we will always get an optimal set. For the plane we prove the latter and give a non-trivial necessary condition for optimality.

But first let us fix some notation, where most of it is adopted from topology. Although it might seem a little odd in the discrete context at first, we hope that it turns out to be helpful for the intuition. Afterwards, we briefly discuss the history of the questions addressed in this chapter and give precise formulations of the problems.

Consider the lattice $\mathbb{Z}^n$ with the L_1-distance $d(\boldsymbol{x},\boldsymbol{y}) = \sum_i |x_i - y_i|$, where $\boldsymbol{x} = (x_1,\dots,x_n)$ and $\boldsymbol{y} = (y_1,\dots,y_n)$. Given a finite set $X \subseteq \mathbb{Z}^n$, the (n-dimensional) *neighborhood* of X is

$$\mathsf{Nbhd}^n(X) := \{\boldsymbol{y} \in \mathbb{Z}^n \backslash X \ : \ d(\boldsymbol{y}, X) = 1\},$$

and the the size of this neighborhood is $n(X) := |\mathsf{Nbhd}^n(X)|$. If there is no danger of confusion, we will write $\mathsf{Nbhd}^n(X) =: \mathsf{Nbhd}(X)$. Note that here $n(\cdot)$ is a function and should not be confused with the dimension.

Further, the *boundary* ∂X of X are the points of X that are next to some point of $\mathsf{Nbhd}(X)$, and the *interior* $\operatorname{int} X$ of X are the points of X for which all neighbors are in X, i.e.,

$$\partial X := \{\boldsymbol{x} \in X \ : \ d(\boldsymbol{x}, \mathsf{Nbhd}(X)) = 1\} = \mathsf{Nbhd}(\mathsf{Nbhd}(X)) \cap X,$$

$$\operatorname{int} X := \{\boldsymbol{x} \in X \ : \ d(\boldsymbol{x}, \mathsf{Nbhd}(X)) > 1\} = X \backslash \partial X.$$

We also define $\operatorname{int}_j(X)$ as the set of points in X such that all neighbors in the directions normal to the j-axis are in X, that is

$$\operatorname{int}_j(X) = \{p \in X \ : \ p \pm e_i \in X \text{ for all } i \neq j\}.$$

For any $r \in \mathbb{N}$, the set $B_r^n := \{\boldsymbol{x} \in \mathbb{Z}^n : \sum_i x_i \leq r\}$ is called the *ball* (of radius r). Note that the neighborhood $\mathsf{Nbhd}(B_r^n)$ of a ball B_r^n contains exactly the points of $\mathbb{Z}^n$ with coordinate sum $r + 1$.

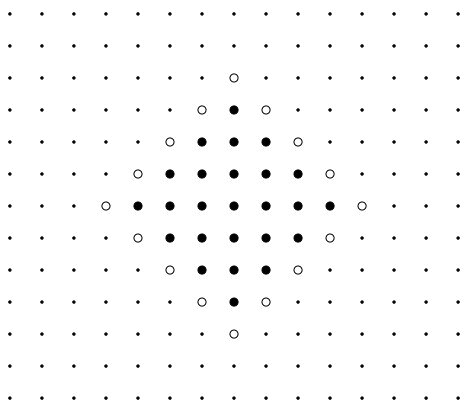

Figure 5.1: The ball B_3^2 and its neighborhood (empty dots).

A finite set $X \subseteq \mathbb{Z}^n$ is called *optimal*, if the size of its neighborhood $n(X)$ is minimal among all sets $Y \subseteq \mathbb{Z}^n$ with $|Y| = |X|$.

5.1 Background and Formulation of the Problems

Isoperimetric problems of this kind have arisen in a number of different contexts, with several definitions of neighborhood, and several different underlying finite and infinite lattices. One approach to a solution is to provide an ordering of the lattice points such that the first j of them form an optimal set of their cardinality for every j. For the Boolean lattice (chains of length two) this is the celebrated theorem of Harper [46]. Katona [50], and Clements and Lindström [21] solved this for chains of arbitrary length l. Bezrukov and Serra [17] considered the problem for Cartesian powers of graphs.

Quite some time before that, Macaulay [62] presented an ordering of the non-negative n-dimensional integer points $\mathbb{Z}_+^n$ having coordinate sum $\leq k$, such that the first j of them have a minimum number of neighbors with coordinate sum $k + 1$ among all sets of k-sum points.

As a similar (and equivalent) result, Wang and Wang [81] constructed sets in $\mathbb{Z}_+^n$ that minimize the number of neighbors in $\mathbb{Z}_+^n$, and extended this to an ordering of the points of $\mathbb{Z}^n$ such that the first j of them minimize the number of neighbors in $\mathbb{Z}^n$. They called these optimal sets *standard spheres*. To avoid confusion with the other notation borrowed from topology, we will call them *standard minimizers*. Basically, a standard minimizer is a ball plus possibly some points of the neighborhood of the ball (see Section 5.7 for a precise definition of the order and the sets).

As it happens, the sequence of standard minimizers enjoys the property that it is closed under the operation of adding the neighborhood. In particular, balls B_r^n are optimal sets in $\mathbb{Z}^n$ (conjectured for $\mathbb{Z}_+^n$ by Hack [45] and for $\mathbb{Z}^n$ by Rivest [73]), which is the analogous result to the classical isoperimetric problem in Euclidean space. In contrast with its Euclidean counterpart, standard minimizers are not, in general, the only optimal sets in lattices.

In the case of the Boolean lattice, all optimal sets were characterized by Bezrukov [15], while for general lattices the complete characterization is still an open problem.

For a survey about different types of isoperimetric problems, as well as a thorough list of references, see [16].

In this chapter we address the following questions.

Problem 5.1. *Is it true that balls are the only optimal sets of their cardinality in the n-dimensional lattice?*

Problem 5.2. *Let $X \subseteq \mathbb{Z}^n$ be an optimal set. Is it true that $X \cup \mathsf{Nbhd}(X)$ is also an optimal set?*

Problem 5.3. *What are necessary and sufficient conditions for sets $X \subseteq \mathbb{Z}^n$ to be optimal?*

The first two problems were answered positively for the Boolean lattice (see [16] and references therein for Problem 5.1, respectively [82] for Problem 5.2).

In Section 5.3 and 5.4 we show that Problem 5.1 has a positive answer also in $\mathbb{Z}^n$.

We conclude in Section 5.5 with some necessary conditions for optimality in $\mathbb{Z}^2$, from which we can deduce that the answer to Problem 5.2 is yes in the two-dimensional case.

5.2 Basic Observations

We call a set $X \subseteq \mathbb{Z}^n$ *connected* if there is no lattice hyperplane orthogonal to a standard basis vector e_i, such that there are elements of X on both sides of the hyperplane, but no points of X on it.

Note that this is a very weak notion of connectedness, and one can easily come up with examples that are connected in this sense, but which do not correspond to what we commonly consider a connected set. We will, however, rarely use this, and a stronger definition would require a lot more care in the following.

Proposition 5.4. *A necessary condition for a finite set $X \subseteq \mathbb{Z}^n$ to be optimal is that X is connected.*

Proof. Let X be a finite set and suppose we find a lattice hyperplane H orthogonal to e_i, such that there are no points of X in H, but there are points of $X_1 \subseteq X$ in direction e_i and $X_2 \subseteq X$ in direction $-e_i$ from H.

Find a pair of points $x_1 \in X_1$ and $x_2 \in X_2$ that minimizes the distance in direction e_i between X_1 and X_2. By construction, $x_1 - e_i$ is not in X, and neither is $x_2 + e_i$. If we now shift X_1 such that x_1 is at position $x_2 + e_i$, then this reduces the size of the neighborhood by at least one.

□

This proposition implies that it is no restriction that from now on all sets $X \subseteq \mathbb{Z}^n$ will be considered to be finite and connected.

Given a set $X \subseteq \mathbb{Z}^n$, we want to consider slices of X with respect to some coordinate direction. To this end we denote by

$$L_{k,l}(X) = \{\boldsymbol{x} \in X : x_k = l\}$$

the l^{th} *k-level* of X. Every k-level lies in an $(n-1)$-dimensional hyperplane, and we denote by $L^{n-1}_{k,l}(X) \subseteq \mathbb{Z}^{n-1}$ the $(n-1)$-dimensional set that results from $L_{k,l}(X)$ by omitting the k^{th} coordinate:

$$L^{n-1}_{k,l}(X) = \left\{(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n) \in \mathbb{Z}^{n-1} \ : \ (x_1, \ldots, x_n) \in L_{k,l}(X)\right\}.$$

The 0^{th} k-level is also called *k-base*.

In the other direction, given some set $Y \subseteq \mathbb{Z}^{n-1}$ then we define the l^{th} *k-extension* of Y as the set $Y^n_{k,l} \subseteq \mathbb{Z}^n$ that is obtained by adding a (new) k^{th} coordinate with value l:

$$Y^n_{k,l} = \{(x_1, \ldots, x_n) \in \mathbb{Z}^n \ : \ x_k = l, (x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n) \in Y\}.$$

Lemma 5.5. *The size of the neighborhood of some k-level, $n(L_{k,l}(X))$, equals the size of the neighborhood of $L^{n-1}_{k,l}(X)$ (in $n-1$ dimensions) plus two times the size of $L_{k,l}(X)$.*

Proof. Every point $x \in L_{k,l}(X)$ has exactly two neighbors y^1, y^2 with $y^i_k \neq l$, namely $y^i_j = x_j$ for $j \neq k$ and $y^i_k = l \pm 1$. □

Lemma 5.6. *Let $X \subseteq \mathbb{Z}^n$ and let $L_{k,l}(X)$ be some level (in any coordinate-direction) with nonempty interior $I = \operatorname{int} L^{n-1}_{k,l}(X)$. Then adding (any subset of) the k-extensions $I^n_{k,l-1}$ and $I^n_{k,l+1}$ to X does not increase the size of the neighborhood $n(X)$.*

By adding we here mean taking the union of the point sets.

Proof. Every point in the mentioned k-extension is in the neighborhood of $L_{k,l}(X)$, and so are any neighbors with k^{th} coordinate equal to $l \pm 1$. Thus there is only one neighbor that is not yet accounted for, and as the point itself reduced the size of the neighborhood by one, the total count stays the same. □

Lemma 5.7. *If $X \subseteq \mathbb{Z}^n$ contains some level $L_{k,l}(X)$ such that the following two statements hold:*

1. $L^{n-1}_{k,j+1}(X) \subseteq \operatorname{int} L^{n-1}_{k,j}(X)$ *for all* $j \geq l$;
2. $L^{n-1}_{k,j-1}(X) \subseteq \operatorname{int} L^{n-1}_{k,j}(X)$ *for all* $j \leq l$,

then the size of the neighborhood of X is $n(X) = n(L_{k,l}(X))$.

Proof. Use Lemma 5.5 and Lemma 5.6 (shifted). □

Lemma 5.8. *The cardinality of any ball $|B^n_r|$ is odd.*

Proof. Since B^n_r is centrally symmetric, it consists of pairs of points, plus the origin. □

Lemma 5.9. *Let $X \subseteq \mathbb{Z}^n$, then*

$$n(X) \geq |L_{k,l_{\max}}(X)| + |L_{k,l_{\min}}(X)| + \sum_{l=l_{\min}}^{l_{\max}} n(L^{n-1}_{k,l}(X)),$$

where

$$\begin{aligned} l_{\min} &:= \min\{l \in \mathbb{Z} : L_{k,l}(X) \neq \emptyset\}, \text{ and} \\ l_{\max} &:= \max\{l \in \mathbb{Z} : L_{k,l}(X) \neq \emptyset\}. \end{aligned}$$

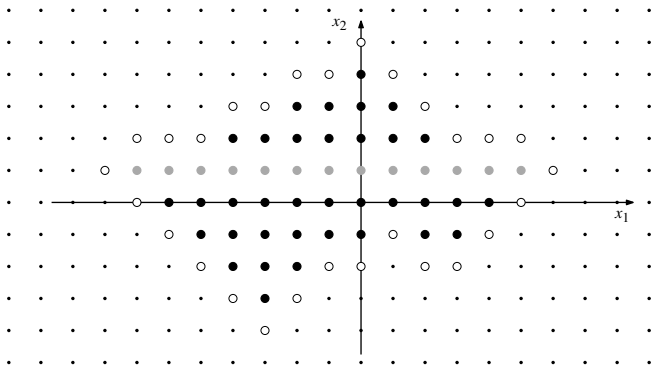

Figure 5.2: The highlighted level, $L_{2,1}(X)$, satisfies the conditions in Lemma 5.7.

Proof. Observe that we can write $\mathsf{Nbhd}(X)$ as a disjoint union of sets, distinguished by the value of the k^{th} coordinate:

$$\mathsf{Nbhd}(X) = (\mathsf{Nbhd}(X) \cap L_{k,l_{\min}-1}(\mathbb{Z}^n)) \cup \cdots \cup (\mathsf{Nbhd}(X) \cap L_{k,l_{\max}+1}(\mathbb{Z}^n)).$$

We also note that $|\mathsf{Nbhd}(X) \cap L_{k,l_{\min}-1}(\mathbb{Z}^n)| = |L_{k,l_{\min}}(X)|$ by arguments similar to Lemma 5.5, and for the same reason $|\mathsf{Nbhd}(X) \cap L_{k,l_{\max}+1}(\mathbb{Z}^n)| = |L_{k,l_{\max}}(X)|$. Finally, we have

$$|\mathsf{Nbhd}(X) \cap L_{k,l}(\mathbb{Z}^n)| \geq n(L_{k,l}^{n-1}(X))$$

for all $l_{\max} \leq l \leq l_{\min}$, which gives us the desired inequality. □

5.3 Uniqueness for Balls in Dimension 2

For better readability we defer the proof that standard minimizers are optimal to Section 5.7. While this proof is directly taken from the paper of Wang and Wang [81], we reproduce it here as well, as the presentation and notation is slightly different.

Proposition 5.10. *The optimal neighborhood size is increasing with the cardinality of the point set: if X and Y are optimal sets in $\mathbb{Z}^2$ with $|X| < |Y|$, then $n(X) \leq n(Y)$.*

As this holds for standard minimizers (of any dimension), it has to be true for all optimal sets.

Proposition 5.11. *Consider the ball B_r^2, with $|B_r^2| = s$. Then the neighborhood of any set $X \subseteq \mathbb{Z}^2$ with $|X| = s + 1$ has size $n(X) \geq n(B_r^2) + 1$.*

Proof. This follows again directly from the results about standard minimizers. The standard minimizer with $s + 1$ points is B_r^2 plus a point $x \in \mathsf{Nbhd}(B_r^2)$ in the positive orthant (see Section 5.7). The grid point x has coordinate sum $r + 1$, and thus it has in each coordinate direction one neighbor with coordinate sum $r + 2$.

Both these neighbors lie in $\mathsf{Nbhd}(B_r^2 \cup \{x\}) \backslash \mathsf{Nbhd}(B_r^2)$. Thus, the neighborhood of $B_r^2 \cup \{x\}$ has at least size $n(B_r^2) + 1$.

As standard minimizers are optimal, any set of cardinality $s + 1$ needs to have at least this neighborhood size. □

Remark 5.12. *Given that standard minimizers are optimal sets in any finite dimension, we can reconstruct Propositions 5.10 and 5.11 for general dimension, where in the latter statement we get* $n(X) \geq n(B_r^2) + n - 1$.

Given a set $X \subseteq \mathbb{Z}^2$, assume w.l.o.g. that the origin $(0, 0)$ is part of X, and consider the following four *diagonal tangents* of X (defined by equalities):

$$\begin{aligned} t_{++} &: \quad +x_1 + x_2 = c_{++} \geq 0 \\ t_{+-} &: \quad +x_1 - x_2 = c_{+-} \geq 0 \\ t_{-+} &: \quad -x_1 + x_2 = c_{-+} \geq 0 \\ t_{--} &: \quad -x_1 - x_2 = c_{--} \geq 0 \end{aligned}$$

where $c_{\pm,\pm}$ are chosen so that for each equality there is some point in X that satisfies it and every point in X satisfies the inequalities obtained by replacing $=$ with $\leq$. The *diagonal hull* $DH(X)$ of X is the set of all integer points in the region bounded by these tangents. Further, X is called *diagonal convex* if $X = DH(X)$.

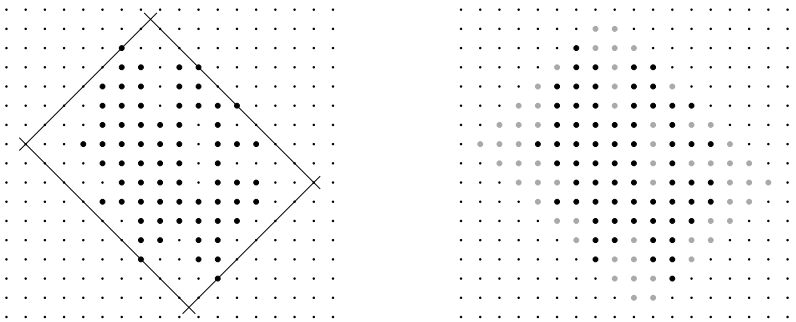

Figure 5.3: A set with its diagonals (left) and the diagonal hull of the set (right, new points highlighted).

If a set $X \subseteq \mathbb{Z}^2$ is diagonal convex, then each point on the boundary lies on at most two of the diagonal tangents, and X resembles a rectangle. However, if two diagonal tangents intersect in a non-integer point, then there is a (connected) portion of ∂X of cardinality 2 that is parallel to some coordinate direction. We will refer to such parts as *axis-aligned* components (of ∂X). Depending on the axes they are aligned to, two such parts may be *orthogonal* or *parallel* to each other.

Accordingly, for each of the diagonal tangents t we call $\partial X \cap t$ a *diagonal* (of ∂X).

Note that ∂X might have none, two, or four axis-aligned components. For example, the diagonal hull of the set X in Figure 5.3 has two axis-aligned components, both of which are parallel to the horizontal axis.

Proposition 5.13. *Let* $X \subseteq \mathbb{Z}^2$*, then* $n(DH(X)) \leq n(X)$.

Proof. $DH(X)$ is obtained from X by repeatedly applying the operation from Lemma 5.6 until there are no points in any direction that can be added. Note that here we use that the sets are connected, and thus in any coordinate-direction X will have at least two neighbors, from the first to the last level. □

Theorem 5.14. *Balls* B_r^2 *are unique optimal sets of their cardinality.*

Proof. Assume that $X \subseteq \mathbb{Z}^2$ is a set with $|X| = |B_r^2|$ and $n(X) = n(B_r^2)$ for some $r \in \mathbb{N} \setminus \{0\}$. Then X has to be diagonal convex. Otherwise $|DH(X)| > |B_r^2|$ and $n(DH(X)) \leq n(X) = n(B_r^2)$, which contradicts Proposition 5.11.

There are four cases for the possible number and relative positions of the axis-aligned components: ∂X can have four, two parallel (opposite), none, or two orthogonal axis-aligned components of size 2 (see Figure 5.4).

Our main strategy will be to transfer parts of ∂X to some part of $\mathsf{Nbhd}(X)$ without increasing the size of the neighborhood. For these new sets it is then easy to see that they cannot simultaneously be optimal and have the cardinality of a ball.

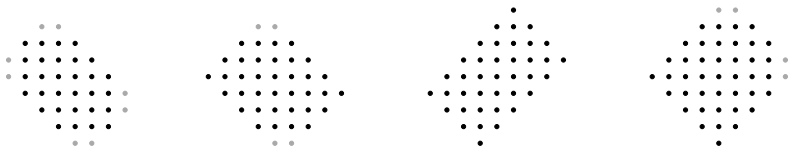

Figure 5.4: The four basic shapes of diagonal convex sets in $\mathbb{Z}^2$ (axis-aligned components are highlighted).

Case 1. ∂X has four axis-aligned components.

Consider two opposite diagonals. They both have the same length k, and are adjacent to $k+1$ points of the neighborhood. Remove all points from one of these diagonals, and add them along the other diagonal (from bottom to top, see Figure 5.5). This results in a set X' with $n(X') = n(X)$ that is not diagonal convex (as it contains an axis-aligned component of size 3), which is a contradiction to the assumption that X is optimal and has the cardinality of a ball.

Figure 5.5: Case 1: Arranging all points from a diagonal along the opposite diagonal gives the same neighborhood-size and an axis-aligned component of size 3 (where the neighborhood is indicated by the empty dots).

Case 2. ∂X has exactly two parallel axis-aligned components.

Consider the levels in the coordinate direction k in which the axis-aligned components both constitute a level. Then every k-level has even cardinality, and thus $|X|$ is even. By Lemma 5.8, this is a contradiction to X having the cardinality of a ball.

Case 3. ∂X has no axis-aligned components.

Let k and l the lengths of the diagonals (each pair of opposite diagonals has the same length).

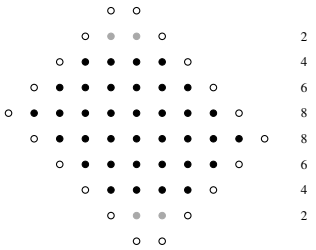

Figure 5.6: Case 2: Even parity.

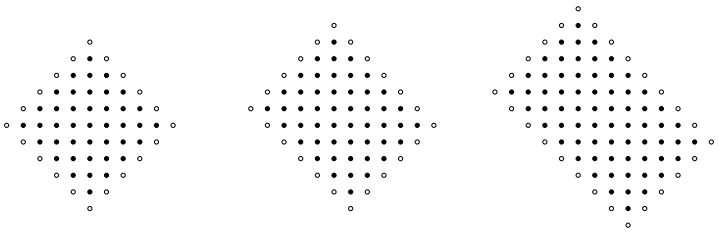

Figure 5.7: Case 3: $k = l$ (left), $k = l + 1$ (middle), and $k > l + 1$ (right).

Case 3.1 If $k = l$, then X is a ball.

Case 3.2 If $k > l$, then remove all points from one of the shorter diagonals and add them along one of the (at least before) longer diagonals.
The resulting set X' has $n(X') = n(X)$. Further, for $k = l + 1$ it is a diagonal convex set with two parallel axis-aligned components of size 2, while for $k > l + 1$, it is not diagonal convex. In either case, we get a contradiction to the assumption that X is optimal and has the cardinality of a ball (remember Proposition 5.11).

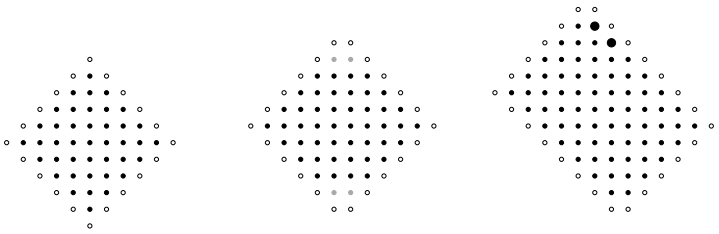

Figure 5.8: Case 3: changed point sets. Fat points are additional ones.

Case 4. ∂X has two orthogonal axis-aligned components.
Consider the diagonal that connects the two axis-aligned components, and say it has length k. Then the opposite diagonal has length $k + 1$ and the two other diagonals both have length l.

Case 4.1. If $k = l$, then this is exactly a ball minus one diagonal. These are optimal sets, as one easily checks that this defines a standard minimizer (with the definition given in the next section). But obviously these cannot have the cardinality of a ball.

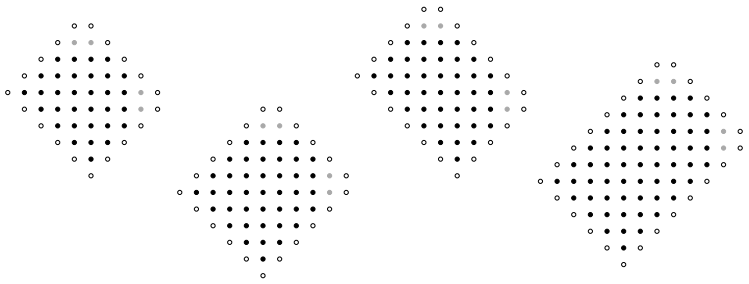

Figure 5.9: Case 4: left to right: $k = l$, $k = l - 1$, $k > l$, and $k < l - 1$.

Case 4.2. If $k = l - 1$, then this is exactly a ball plus one diagonal. Again, these are standard minimizers, but cannot have the cardinality of a ball.

Case 4.3. If $k > l$, then we remove a diagonal of length l and add it again along the diagonal of length k. In the resulting point set there is at least one point missing in this new diagonal, and adding it does not increase the size of the neighborhood. By Proposition 5.11 this contradicts our assumptions.

Case 4.4. If $k < l - 1$, then we remove the diagonal of length k and add it along a diagonal of length l. Again, there is at least one point missing in the new diagonal and adding it does not increase the size of the neighborhood.

□

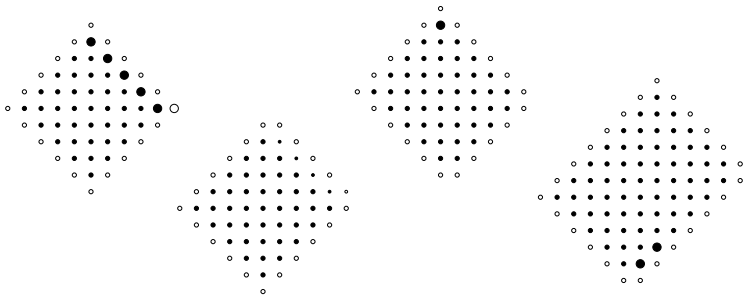

Figure 5.10: Case 4: changed point sets. Fat points are additional ones.

5.4 Uniqueness for Balls in General Dimension

We are now going to use the proof of optimality for standard minimizers to show the uniqueness of balls as optimal sets.

Theorem 5.15. *Balls B_r^n are unique optimal sets of their cardinality.*

Proof. The proof is by induction on the dimension n, the induction base $n = 2$ being Theorem 5.14 from the last section.

A major tool in the proof of optimality for standard minimizers is the *k-normalization* $N_k(X)$ of a set $X \subseteq \mathbb{Z}^n$:

- replace all (nonempty) k-levels of X by $(n-1)$-dimensional standard minimizers of the same cardinality, and then
- change the order of the levels, such that the largest one is the k-base and the remaining ones are arranged around the k-base in decreasing order and alternating sign in the k^{th} component, that is, $|L_{k,i}| \geq |L_{k,i+1}|$ for $i \geq 0$ and $|L_{k,j}| \geq |L_{k,j-1}|$ for $j \leq 0$.

Roughly, to show that standard minimizers are optimal sets, take an arbitrary set and repeatedly apply 1-normalization and n-normalization to it. This series of transformations terminates with a standard minimizer after a finite number of steps, and this procedure does not increase the size of the neighborhood (see Section 5.7 for details).

For the induction step consider some optimal set $X \subseteq \mathbb{Z}^n$ with $|X| = |B_r^n|$ for some r. If we repeatedly apply 1-normalization and n-normalization to X, then X is transformed to B_r^n, as this is the standard minimizer of this cardinality. Consider the set Y that occurs in this transformation process exactly before the normalization step that yield a stable set. We assume the last step is in direction 1 to reduce the number of variables in the following.

Then there is a one-to-one correspondence between the 1-levels of Y and the 1-levels of B_r^n, such that all cardinalities match. Note that the levels of B_r^n are $(n-1)$-dimensional balls. Define

$$\begin{aligned} l_{\min} &:= \min\{l \in \mathbb{Z} : L_{1,l}(Y) \neq \emptyset\}, \text{ and} \\ l_{\max} &:= \max\{l \in \mathbb{Z} : L_{1,l}(Y) \neq \emptyset\}. \end{aligned}$$

Then as a lower bound for the size of the neighborhood of Y we have

$$n(Y) \geq |L_{1,l_{\max}}(Y)| + |L_{1,l_{\min}}(Y)| + \sum_{l=l_{\min}}^{l_{\max}} n(L_{1,l}^{n-1}(Y))$$

by Lemma 5.9.

Now as every level $L_{1,l}(Y)$ has the cardinality of a ball $B_{r_l}^{n-1}$ for some r_l, it follows by the induction hypothesis that they all really have to be these balls, as otherwise $n(L_{1,l}^{n-1}(Y)) > n(B_{r_l}^{n-1})$ and thus $n(X) \geq n(Y) > n(B_r^n)$, which would be a contradiction to the assumption that X is optimal.

Finally consider the order of the 1-levels in Y. If they are not ordered in the same way as for the corresponding ball B_r^n, then there exist two adjacent levels $L_{1,i}(Y) = B_{r_i}^{n-1}$ and $L_{1,j}(Y) = B_{r_j}^{n-1}$ such that $r_j < r_i - 1$. But then

$$I = \operatorname{int}\left(L_{1,i}^{n-1}(Y)\right) \setminus L_{1,j}^{n-1}(Y) \neq \emptyset$$

and thus by Lemma 5.6 we can add points to Y without increasing the size of the neighborhood $n(Y)$. This is, once again, a contradiction to Proposition 5.11 since Y was assumed to be optimal and of the same cardinality as some B_r^n.

Thus $X = Y = B_r^n$, which completes the proof. □

5.5 Necessary Conditions for Optimal Sets

Back to Dimension 2

In Proposition 5.4 we showed that connectedness is a necessary condition for a set $X \subseteq \mathbb{Z}^2$ to be optimal. As a first step towards further conditions we consider the shapes of X and $Y = X \cup \mathsf{Nbhd}(X)$ for diagonal convex sets, and the relation between $n(X)$ and $n(Y)$.

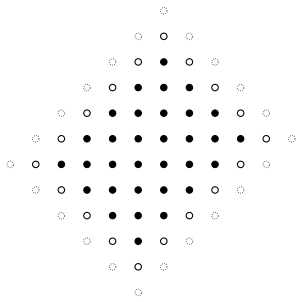

Figure 5.11: A diagonal convex set X, together with $\mathsf{Nbhd}(X)$ and $\mathsf{Nbhd}(X \cup \mathsf{Nbhd}(X))$.

Proposition 5.16. *If a set $X \subseteq \mathbb{Z}^2$ is diagonal convex, then $Y = X \cup \mathsf{Nbhd}(X)$ is again diagonal convex and $n(Y) = n(X) + 4$.*

Proof. The lengths of the axis-aligned parts of $\mathsf{Nbhd}(Y)$ and $\mathsf{Nbhd}(X)$ are identical, while the lengths of the diagonal parts of $\mathsf{Nbhd}(Y)$ are the lengths of the diagonal parts of $\mathsf{Nbhd}(X)$ plus one. □

This proposition tells us that diagonal convex sets have a shape that is 'stable' under the operation of adding the neighborhood, and that the size of the neighborhood behaves in a nice way. But there is a far larger class of sets that behaves in essentially the same way:

Consider some set $X \subseteq \mathbb{Z}^2$, such that the neighborhood $\mathsf{Nbhd}(X)$ is a simple cycle. By this we mean that for any two points in $\mathsf{Nbhd}(X)$ there are exactly two disjoint paths between them.

A path between x and y in X is a sequence $(v_j)_{1 \le j \le k}$ such that $v_1 = x$, $v_k = y$, $v_i \in X$ for all i, and any two consecutive points in the path differ by at most one in each coordinate, i.e., $v_j - v_{j+1} = \sum_{i \le n} c_i e_i$ with $c_i \in \{-1, 0, 1\}$.

Note that if we regard the cycle $\mathsf{Nbhd}(X)$ as a (not necessarily convex) polygon, the grid points in the interior of the polygon are all points of X. Observe that the inside and outside angles at any $v \in \mathsf{Nbhd}(X)$ are at least 90 degrees, and at least 135 degrees if one of the polygon edges adjacent to v is in a coordinate direction. An interior angle smaller than this would give $d(v, X) > 1$, and an exterior angle would lead to a sub-cycle of length 3 in $\mathsf{Nbhd}(X)$.

We will proceed through the cycle $\mathsf{Nbhd}(X)$ in the counter-clockwise direction and consider the occurring direction changes with respect to the oriented coordinate directions (e_i, σ) with $i \in \{1, 2\}$ and $\sigma \in \{+, -\}$:

- Choose as starting (and ending) point the topmost point of the tangent $t_{++} : x_1 + x_2 = c > 0$.
- Choose as starting (and ending) direction $(e_1, -)$.
- Proceeding through the cycle, remember the current coordinate direction (e_i, σ), as well as the number of occurred direction changes. In every step x_k to x_{k+1}
  1. keep the direction (e_i, σ) if $\sigma = \operatorname{sign}(x_{k+1} - x_k)_i$,
  2. otherwise change the direction to (e_j, σ') with $j \neq i$ and $\sigma' = \operatorname{sign}(x_{k+1} - x_k)_j$,

  where for every $y \in \mathbb{R} \backslash \mathbf{0}$ we define $\operatorname{sign}(y) = +$ if $y > 0$ and $\operatorname{sign}(y) = -$ if $y < 0$.

- as the starting point is reached with a direction different from $(e_1, -)$, the turn to this direction (according to the turning rule above) counts an as occurring turn.

Every considered (and counted) direction change is a (left or right) turn by 90 degrees. As we require the starting direction to be identical to the ending direction, we turn 360 degrees in total. Thus we count an even number $k \geq 4$ of direction changes (each of the oriented directions at least once).

Also, by the above observations about the occurring angles, we notice that for every turn there is a diagonal part of $\mathsf{Nbhd}(X)$ of size at least 2, that could be seen as being in either of the two directions before and after the turn. We will call such a diagonal part of $\mathsf{Nbhd}(X)$ a *connecting diagonal*.

We call a set X *close-to-convex* if $\mathsf{Nbhd}(X)$ is a cycle and if in the process described above there are four changes in direction (i.e., every oriented coordinate direction appears exactly once).

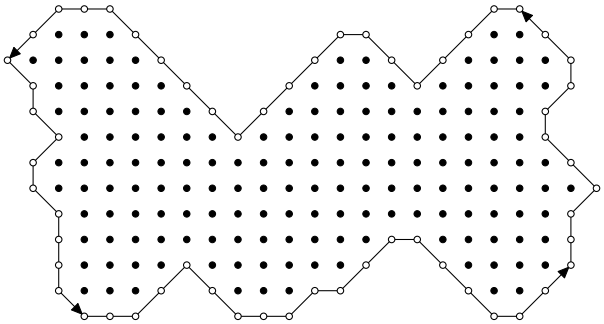

Figure 5.12: Proceeding through the neighborhood of a close-to-convex set. Direction changes occur after the arrows.

Observation 5.17. *If a set $X \subseteq \mathbb{Z}^2$ is close-to-convex, then $Y = X \cup \mathsf{Nbhd}(X)$ is again close-to-convex, and $n(Y) = n(X) + 4$. Moreover, Y has the same shape as X except for the four connecting diagonals which each get longer by one. See Figure 5.13 for an example.*

Note that each connecting diagonal is identical to the intersection of $X \cup \mathsf{Nbhd}(X)$ with one of its diagonal tangents.

Observation 5.18. *The standard minimizers that are presented in [81] are optimal sets that are close-to-convex, and for any standard minimizer S, $S \cup \mathsf{Nbhd}(S)$ is again a standard minimizer (and thus optimal). This implies that for any optimal set $X \subseteq \mathbb{Z}^2$ the inequality $n(Y) \geq n(X) + 4$ holds for $Y = X \cup \mathsf{Nbhd}(X)$.*

Let us go back to general connected sets $X \subseteq \mathbb{Z}^2$. We denote the cycle $C(X) \subseteq \mathsf{Nbhd}(X)$ for which X lies in the interior of the polygon defined by this cycle as the *cycle that surrounds* X. Further we call the (finite) set $\mathsf{cl}(X)$ of grid points enclosed by $C(X)$ the *closure* of X.

An ordered subset $\{x_1, x_2, \ldots, x_n\} \subseteq \mathbb{Z}^2$ is a *lattice path* if the elements are distinct and $d(x_i, x_{i+1}) = 1$ for all i. We note that every lattice path is a path, as defined in section 1, but the converse is not true.

A *hole* in X is a subset $H \subseteq \mathsf{cl}(X) \backslash X$ such that

1. H is connected, and

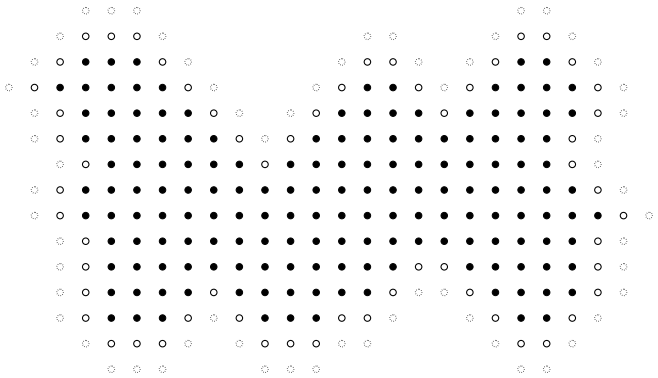

Figure 5.13: A close-to-convex set and two layers of neighborhood.

2. for every $h \in H$, every lattice path from h to $C(X)$ contains some element of X.

Proposition 5.19. *If $X \subseteq \mathbb{Z}^2$ is connected, then for $Y = X \cup \mathsf{Nbhd}(X)$ we have $n(Y) \leq n(X)+4$.*

Proof. For any close-to-convex set this is obviously true.

Now if X is not close-to-convex, then either its neighborhood $\mathsf{Nbhd}(X)$ does not form a cycle or we will get more than four turns when proceeding through the cycle like described above.

In the latter case, assume we have counted $2k + 4$ turns. Then we have counted exactly $k + 4$ left turns and k right turns. For every left turn the connecting diagonal gets longer by (at most) one, while for each right turn the connecting diagonal gets shorter by (at least) one (see Figure 5.14). All parts in between are just translated by 1 and are thus (at most) as long as they were.

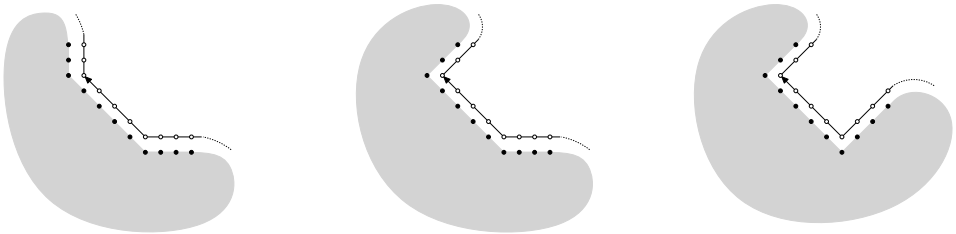

Figure 5.14: All possible right turns (up to symmetry).

Note that $\mathsf{Nbhd}(Y)$ need not be a cycle, and that the *at most* and *at least* statements from above stem from the fact that points of $\mathsf{Nbhd}(Y)$ might be created by duplication of more than one part of $\mathsf{Nbhd}(X)$, see Figures 5.15 and 5.16 for examples.

Now what is left is the situation when $\mathsf{Nbhd}(X)$ is not a cycle. Here, consider the cycle $C(X) \subseteq \mathsf{Nbhd}(X)$ that surrounds X.

For the part of the neighborhood of $\mathsf{Nbhd}(Y)$ that lies outside of $C(X)$ the same arguments as above can be applied.

For the inside part observe that every component of $\mathsf{Nbhd}(Y)$ corresponds to a hole of $X \cup \mathsf{Nbhd}(X)$, see Figure 5.17, and thus the size of the neighborhood inside $C(X)$ is decreasing. □

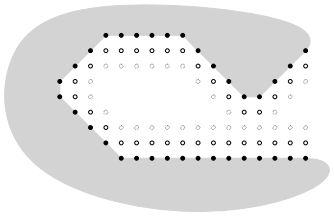

Figure 5.15: $\mathsf{Nbhd}(Y)$ does not form a cycle: creating a hole.

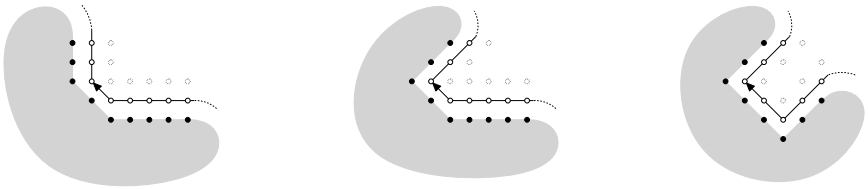

Figure 5.16: $\mathsf{Nbhd}(Y)$ does not form a cycle: Right turns with short connecting diagonal.

Proposition 5.20. *If $X \subseteq \mathbb{Z}^2$ is not close-to-convex, then X cannot be optimal.*

Proof. Assume that X is connected but not close-to-convex, and consider again the cycle $C(X) \subseteq \mathsf{Nbhd}(X)$ that surrounds X.

If $C(X) \subsetneq \mathsf{Nbhd}(X)$, then the set Y containing all points inside this cycle $C(X)$ has $|Y| > |X|$ and $n(Y) < n(X)$. Thus X cannot be optimal.

If $C(X) = \mathsf{Nbhd}(X)$ then consider $Z = X \cup \mathsf{Nbhd}(X)$. From the proof of Proposition 5.19 we know that if $\mathsf{Nbhd}(Z)$ is a cycle again, then the length of every right turn connecting diagonal is shorter than the corresponding one in $\mathsf{Nbhd}(X)$. Thus, after finitely many steps of adding the neighborhood we obtain a set that is not optimal (as its neighborhood is not a cycle). But as Proposition 5.19 holds for every step of adding the neighborhood, X cannot be optimal either. □

Next we show that optimality is not affected by addition of the neighborhood. More precisely, we have:

Proposition 5.21. *Consider any connected set $X \subseteq \mathbb{Z}^2$ and its union with its neighborhood $X' = X \cup \mathsf{Nbhd}(X)$. Let Y and Y' be the standard minimizer of cardinality $|Y| = |X|$ and $|Y'| = |X'|$, respectively. If $n(X) = n(Y) + k$ for some $k \geq 0$, then $n(X') \leq n(Y') + k$.*

Proof. $|Y'| = |X'| = |X| + n(X) = |Y| + n(Y) + k \geq |Y \cup n(Y)|$. Thus from Proposition 5.10, it follows that

$$n(Y') \geq n(Y \cup n(Y)) = n(Y) + 4 = n(X) - k + 4 = n(X') - k.$$

□

Now we can state the answer to Problem 5.2, i.e., we have shown that, at least in dimension 2, the set of optimal sets is closed under the operation of adding the neighborhood. We record this solution in the following corollary.

Corollary 5.22. *If a set $X \subseteq \mathbb{Z}^2$ is optimal, then $n(X \cup \mathsf{Nbhd}(X)) = n(X) + 4$ and $X \cup \mathsf{Nbhd}(X)$ is also optimal.*

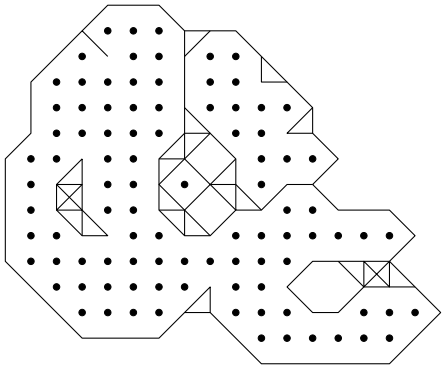

Figure 5.17: A connected set X visualizing some possible shapes. The lines indicate all paths in $\mathsf{Nbhd}(X)$.

5.6 Outlook

One could try to find a good definition of 'close-to-convex' in higher dimensions. There are, however, some problems associated with this. For example, we made excessive use of the concepts of paths and directions, and even when looking in two coordinate-directions separately at a time, the interaction between them has to be addressed.

Also, we would be interested in good sufficient conditions for a set to be optimal, starting already in dimension 2, but of course also for the general case.

5.7 Notes

5.7.1 Definitions and Properties of standard minimizers

Here and in Section 5.7.2 we follow largely [81].

A *standard minimizer* in n dimensions is the set of points in the n-ball of some radius $m \in \mathbb{N}$, together with those points of the n-sphere of radius $m+1$ (which is the set $\{\boldsymbol{x} \in \mathbb{Z}^n \ : \ \sum x_i = m+1\}$), whose projection normal to the first coordinate is a standard minimizer in $n-1$ dimensions. The name is chosen in anticipation of their property to be optimal sets, i.e., to minimize the size of the neighborhood.

First we will discuss the results for $\mathbb{Z}^n_+$, then we will see how we can get the results for $\mathbb{Z}^n$. We adapt all definitions we have made for $\mathbb{Z}^n$ also for the restriction to the set $\mathbb{Z}^n_+$. To avoid confusion, we will also write *positive* standard minimizer, if we mean a standard minimizer in $\mathbb{Z}^n_+$.

A standard minimizer in $\mathbb{Z}^n_+$ can also be described as follows: The points $\boldsymbol{y} \in \mathbb{Z}^n_+$ with $\sum_{i=1}^n y_i \leq q$ can be viewed as representations of $P := \sum y_i q^i$, i.e., as q-adic numbers. In this light, a positive standard minimizer consists of:

(1) $\boldsymbol{y} \in \mathbb{Z}^n_+$ with $\sum y_i \leq m$ (i.e., the ball B^n_m);

(2) $\boldsymbol{y} \in \mathbb{Z}^n_+$ with $\sum y_i = m+1$ and $(y_1, \ldots, y_n) \succ (r_1, \ldots, r_n)$ for some r and some canonical order (e.g., $P \succ R$ in the $(m+1)$-adic numbers).

The set described in (1) is also called the *core* of the positive standard minimizer, the set in (2) is called its *outer shell*. As an order, we will use the above mentioned for q-adic numbers.

We will now gather some more or less obvious relations between these objects in $\mathbb{Z}_+^n$.

Evidently, for a positive standard minimizer X, the interior of X is contained in its core. Conversely, a core point will not be in $\mathrm{int}(X)$, if some neighbor is not in X. But by the description of the shell in (2), this means that $\boldsymbol{y} \in core(X)$ satisfies

$$\boldsymbol{y} \notin \mathrm{int}(X) \Leftrightarrow \boldsymbol{y} + e_n \notin X.$$

Note that we thus only have to look at the 'upper' neighbor, and also only at the last coordinate.

These facts might also be expressed as follows: Let X be a positive standard minimizer and $\boldsymbol{x} \in X$. Then

$$\boldsymbol{x} + e_n \in X \Rightarrow \boldsymbol{x} \in \mathrm{int}_n(X) \Rightarrow \boldsymbol{x} \in \mathrm{int}(X).$$

Similarly we observe that

$$\boldsymbol{x} \in \mathrm{int}_1(X) \Rightarrow \boldsymbol{x} + e_1 \in X,$$

from which it follows that $\mathrm{int}_1(X) = \mathrm{int}(X)$ for a positive standard minimizer X.

By definition we have $\mathrm{int}(X) \subseteq \mathrm{int}_n(X)$. The converse is not true, however we get the following

Lemma 5.23. *Let X be a positive standard minimizer. Then*

$$|\mathrm{int}_n(X) \setminus \mathrm{int}(X)| \leq 1.$$

Proof. We know that

$$\boldsymbol{y} \in \mathrm{int}_n(X) \setminus \mathrm{int}(X) \Rightarrow \boldsymbol{y} + e_j \in X \text{ for all } 1 \leq j \leq n-1 \text{ and } \boldsymbol{y} + e_n \notin X.$$

Assume there are $\boldsymbol{y}, \boldsymbol{y}' \in \mathrm{int}_n(X) \setminus \mathrm{int}(X)$ with $\boldsymbol{y} \neq \boldsymbol{y}'$. Then there must be a smallest $i \in \{1, \ldots, n-1\}$ with $y_i \neq y_i'$, and say $y_i < y_i'$. Then $\boldsymbol{y} + e_j \leq \boldsymbol{y}' + e_n$ for $i < j \leq n-1$, and therefore $\boldsymbol{y}' + e_n \in X$ by (2) of the definition of standard minimizers. □

Note that a point $\boldsymbol{y}$ as in the proof then defines the size of the standard minimizer in question.

Observation 5.24. *An n-level of a standard minimizer $X \subseteq \mathbb{Z}_+^n$ is itself a positive standard minimizer in the first $(n-1)$ dimensions. Also note that the j^{th} n-level fits into the interior of the $(j-1)^{st}$ n-level, i.e.,*

$$L_{n,j}^{n-1}(X) \subseteq \mathrm{int}(L_{n,j-1}^{n-1}(X)).$$

This means that X is almost completely determined by its n-base, as each n-level is the interior of the next lower one, with the possible exception of one point.

So far we have described positive standard minimizers. But actually we can use this knowledge to identify a set as a (special) standard minimizer.

Lemma 5.25. *Suppose a set $X \subseteq \mathbb{Z}_+^n$ has a positive standard minimizer of dimension $n-1$ as n-base, and every n-level is the interior of the next lower level (as sets in $n-1$ dimensions). Then X is a positive standard minimizer.*

Proof. By construction, the largest point among the ones with maximal coordinate sum must lie in in the n-base. The rest follows by the preceding remarks. □

Observing that the n-levels 'nest' like this, we can state the following

Observation 5.26. *The neighborhood in the same level as a given n-level (except the n-base) is also part of the neighborhood in the next lower n-level of X.*

This is one of the reasons for the neighborhood of a standard minimizer to be small.

This fact can be used to count the neighborhood by only looking at the n-base.

Lemma 5.27. *Let X be a positive standard minimizer with n-base B. Then $n(X)$ is $|B|$ plus the size of the $(n-1)$-dimensional neighborhood of B.*

More formally we have

$$|\mathsf{Nbhd}^n(X)| = |L_{n,0}(X)| + |\mathsf{Nbhd}^{n-1}(L_{n,0}^{n-1}(X))|$$

for a standard minimizer X in $\mathbb{Z}_+^n$. □

Another evident but important fact about standard minimizers is the following

Lemma 5.28. *The sizes of* $\mathrm{int}(X)$ *and* $\mathsf{Nbhd}^n(X)$ *increase monotone in the size of the minimizers: If X and Y are positive standard minimizers and*

$$|X| \geq |Y|, \quad \textit{then}$$
$$|\mathrm{int}(X)| \geq |\mathrm{int}(Y)| \quad \textit{and} \quad |\mathsf{Nbhd}(X)| \geq |\mathsf{Nbhd}(Y)|.$$

Proof. The last inequality follows by induction on the dimension, using that $|L_{n,0}(X)|$ is a monotone function of $|X|$ for standard minimizers. □

For the standard minimizer in $\mathbb{Z}^n$, order the n-quadrants and then define the outer shell with respect to this order in every n-quadrant (almost) as above. A more rigid description is given below.

We define $\mathrm{int}_j(X)$ as the set of points in X, such that all neighbors in the directions normal to the j-axis are in X, i.e.,

$$\mathrm{int}_j(X) = \{\boldsymbol{x} \in X \ : \ \boldsymbol{x} \pm e_i \in X \text{ for all } i \neq j\},$$

where e_i is the i^{th} unit vector.

Formally a standard minimizer in $\mathbb{Z}^n$ is defined as follows:

We define our ordering, such that the points are arranged first by the coordinate magnitude sum. Among points with the same sum, we associate a $2n$ digit number to each point and order according to this number (in the usual numerical ordering), taking the largest first.

In the $2n$ digit number, the first n describe the orthant of the number, the last n digits are the total value of the coordinates.

The j^{th} orthant digit is 1, if the $(n-j+1)^{st}$ coordinate of the point is strictly positive, and 0 otherwise.

A *standard minimizer* now is a set of points in $\mathbb{Z}^n$, which correspond to an initial segment in this ordering.

To illustrate the ordering, we list the first thirteen points in two dimensions, the associated four digit number listed beneath the points (see also Figure 5.18).

$(0,0)$	$(0,1)$	$(1,0)$	$(-1,0)$	$(0,-1)$	$(1,1)$	$(-1,1)$
0000	1001	0110	0010	0001	1111	1011

$(0,2)$	$(2,0)$	$(1,-1)$	$(-2,0)$	$(-1,-1)$	$(0,-2)$
1002	0120	0111	0020	0011	0002

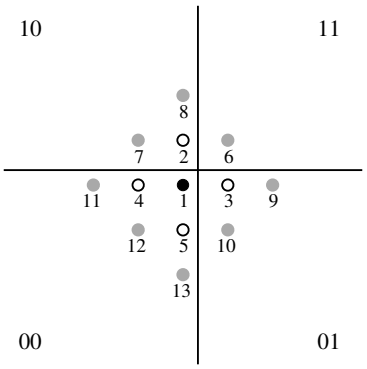

Figure 5.18: Illustration of the ordering. The numbers under the points indicate the position in the order, the shading indicates their coordinate magnitude sum.

It is apparent that within each orthant, the set of points obtained correlates to a standard minimizer in $\mathbb{Z}^n_+$.

It is not difficult to show that every above statement, starting with Observation 5.24, is also true for $\mathbb{Z}^n$, where in Lemma 5.27 we get

$$|\mathsf{Nbhd}^n(X)| = 2|L_{n,0}(X)| + |\mathsf{Nbhd}^{n-1}(L^{n-1}_{n,0}(X))|.$$

Lemma 5.29. *Standard minimizers in $\mathbb{Z}^2$ are optimal.*

Proof. We first consider standard minimizers in Z^2_+. Remember that a positive standard minimizer (in 2 dimensions) consists of the integer points with coordinate sum $\leq j$ for some $j \in \mathbb{N}$, and some (maybe all) points with coordinate sum $j+1$. The size of its neighborhood is then $j+3$.

Obviously any set $X \subseteq \mathbb{Z}^2_+$ of the same size has some point $p = (a,b) \in X$ with $a+b \geq j+1$. As we saw before, it is no restriction to seek optimal sets only among connected sets, and also we may assume that the sets have points on both axis.

For any such set it is easy to see that $n(X) \geq (a+1)+(b+1) \geq j+3$, which proves the lemma for $\mathbb{Z}^2_+$. For the more general case, we center any given X around the origin, minimizing the maximal coordinate sum in the quadrants. Then the reasoning is analogous to the previous. □

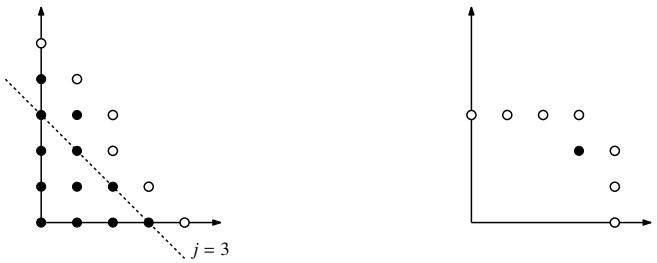

Figure 5.19: A standard minimizer in $\mathbb{Z}_+^2$ with 12 elements (left). Lower bound on the neighborhood (right).

5.7.2 Optimality of Standard Minimizers

Because it makes the methods of the proof more transparent, we will first prove the following:

Theorem 5.30. *Standard minimizers in $\mathbb{Z}_+^n$ are optimal.*

Subsequently, we explain how to extend this result to $\mathbb{Z}^n$:

Theorem 5.31. *Standard minimizers in $\mathbb{Z}^n$ are optimal.*

The proof of Theorem 5.31 is essentially the same as the proof of Theorem 5.30, with some additions. Both will be inductions on the dimension n, where we have already seen the case $n = 2$ in the previous section.

We observe that the proof of Theorem 5.30 can be obtained by the following steps:

1. Prove that the result holds in two dimensions.
2. Define 'normalization' in the k^{th} direction.
3. Prove that such normalization cannot increase the neighborhood.
4. Prove that alternating normalization in the 1^{st} and n^{th} direction yields a stable set w.r.t. these operations.
5. Prove that in such a stable set the interior of the j^{th} n-level includes the $(j+1)^{st}$.

Lemma 5.29 is the first step. Now recall the *k-normalization* $N_k(X)$ of a set $X \subseteq \mathbb{Z}_+^n$:

- replace all (nonempty) k-levels of X by $(n-1)$-dimensional positive standard minimizers of the same cardinality, and then
- change the order of the levels, such that the largest one is the k-base and the remaining ones are arranged above the k-base in increasing order, i.e $|L_{k,i}| \geq |L_{k,i+1}|$ for $i \geq 0$.

Also, we define a rank-function $r : \mathbb{N}^n \to \mathbb{N}^n \setminus \{0\}$ with

$$r(a) < r(b) :\Leftrightarrow \text{ There is a standard minimizer } X \text{ with } a \in X \text{ and } b \notin X.$$

Remember that for $X \subseteq \mathbb{K}^n$ we defined $n(X) = |\mathsf{Nbhd}^n(X)|$.

Lemma 5.32. *Let $X \subseteq \mathbb{Z}_+^n$. Then $n(N_k(X)) \leq n(X)$.*

Proof. We will without loss of generality assume that in all sets considered, the k-levels $L_{k,j}(X)$ are monotonically decreasing in size with increasing j. This is no loss, because we can map any other set X to a set X' with this property, by reducing the k-components of all its points as much as possible. It is easy to see that the neighborhood of X' has at most the size of the neighborhood of X.

By induction on n, standard minimizers have the minimum $(n-1)$-dimensional neighborhood within each k-level, and therefore

$$n(L^{n-1}_{k,j}(X)) \geq n(L^{n-1}_{k,j}(N_k X)).$$

By $n_{k,j}(X)$ we denote the size of the set of neighborhood points of X with k^{th} coordinate equal to j, i.e., $n_{k,j}(X) := |\{x \in \mathsf{Nbhd}^n(X) \ : \ x_k = j\}|$.

Then obviously $n_{k,j}(X)$ is at least as big as the difference between the sizes of $L_{k,j-1}(X)$ and $L_{k,j}(X)$, and also at least the size of (the $(n-1)$-dimensional set) $\mathsf{Nbhd}(L^{n-1}_{k,j}(X))$. More formally:

$$n_{k,j}(X) \geq \max\left\{n(L^{n-1}_{k,j}(X)); |L_{k,j-1}(X)| - |L_{k,j}(X)|\right\}.$$

On the other hand, for the k-normalization we have equality:

$$n_{k,j}(N_k(X)) = \max\left\{n(L^{n-1}_{k,j}(N_k X)); |L_{k,j-1}(N_k X)| - |L_{k,j}(N_k X)|\right\},$$

so that $n_{k,j}(X) \geq n_{k,j}(N_k X)$ for every j, and therefore the proof is finished by summation on j. □

Lemma 5.33. *$\sum_{a\in N_k(X)} r(a) \leq \sum_{a\in X} r(a)$, where we get equality only for $N_k(X) = X$.*

Proof. This can be done by the definition of the standard minimizer and is left to the reader. □

From Lemma 5.33 we can deduce that for any (finite) $Y \subseteq \mathbb{Z}^n_+$ there is an $\alpha \in \mathbb{N}$, such that $Z := (N_1 N_d)^\alpha(Y)$ satisfies $Z = N_1(Z) = N_d(Z)$.

Now $L^{n-1}_{n,j+1}(Z) \subseteq L^{n-1}_{n,j}(Z)$, unless for some $y \in Z$ we have

(i) $y + e_n \in Z$,

(ii) $y + e_j \notin Z$ for some $j \neq n$.

Since $Z = N_1(Z)$, and by the properties of standard minimizers in $n-1$ dimensions, we know that (i) implies that $y + e_i \in Z$ for all $2 \leq i \leq n$. Therefore we must have $y + e_1 \notin Z$. But since $Z = N_n(Z)$, this implies that $y + e_j \notin Z$ for all $1 \leq j \leq n-1$, which is a contradiction to the above.

We conclude that the $(j+1)^{st}$ n-level of Z fits into the interior of the j^{th} n-level of Z.

Lemma 5.34. *If the n-levels of Z (which is defined as above) are standard minimizers and the $(j+1)^{st}$ n-level of Z fits into the interior of the j^{th} n-level, then $|L_{n,0}(Z)| \geq |L_{n,0}(Q)|$ for a standard minimizer Q of size $|Z|$.*

Proof. Suppose the n-base of Z is strictly smaller that the n-base of the standard minimizer Q. Since the $(j+1)^{st}$ n-level of Z is not larger than the interior of the j^{th}, Z can be no larger than the set obtained by choosing as $(j+1)^{st}$ n-level the exact interior of the j^{th}, which by Lemma 5.25 yields a standard minimizer Q'. Then Q' would have to be strictly smaller than Q or Z by the definition of the standard minimizer, but also $|Q'| \geq |Z| = |Q|$ by the construction of Q', so the lemma is proven. □

By Lemma 5.32 we know that $n(Y) \geq n(Z)$ and it is obvious that $n(Z) \geq |L_{n,0}(Z)| + n(L_{n,0}^{n-1}(Z))$. By the monotonicity in the size of the neighborhood of a standard minimizer, we therefore get

$$n(Y) \geq n(Z) \geq |L_{n,0}(Z)| + n(L_{n,0}^{n-1}(Z)) \geq |Q| + n(L_{n,0}^{n-1}(Q)) = n(Q),$$

where Q is a positive standard minimizer with $|Q| = |Z|$. This proves Theorem 5.30.

Remember the list of steps to prove Theorem 5.30. The first four steps are the same for Theorem 5.31, and then we have:

5'. Prove that in such a stable set the interior of the j^{th} n-level includes the $(j+1)^{st}$, and the interior of the of the $(-j)^{th}$ includes the $-(j+1)^{st}$.

6. Prove that one can rearrange such a set into a standard minimizer without increasing the neighborhood.

The last step is necessary, as normalization does not always terminate with a standard minimizer, and in fact will only do so, if either the n-base or the 1^{st} n-level is a (n,m)-ball for some m.

We will show that the set after the normalization can be replaced by a potentially larger set, which has this property, and such that the neighborhood is not increased.

Now define k-normalization as before, only now we 'stack' the standard minimizers with alternating positive and negative sign in the k^{th} coordinate.

Since standard minimizers nest in one another and since by induction we assume they are optimal in $(n-1)$ dimensions, normalization will again not increase the neighborhood.

That alternating normalization yields a stable set, in which the level fit into the interior of the adjacent level in the direction of the base, is also analogous to the non-negative setting.

Step 6. In the light of step 5', a stable set will have as neighborhood size the sum of the sizes of its n-base L_0, its 1^{st} n-level L_1 (meaning the set of points with n^{th} coordinate equal to 1) and their $(n-1)$-dimensional neighborhoods. Both these levels are $(n-1)$-dimensional standard minimizers.

By the monotonicity in the size of the standard minimizers we may assume that the i^{th} n-level (for $i > 1$) of the stable set has the same size as the interior of the $(i-1)^{st}$ n-level; and that the j^{th} n-level (for $j < 0$) has the size of the interior of the $(j+1)^{st}$ n-level. Thus, the stable set X obtained is a (n,m)-ball in all but two orthants.

If

$$L_0 = \left\{ x \in X \ : \ \sum_{i=1}^{n-1} |x_i| \leq k, \ x_n = 0 \right\} \quad \text{and}$$

$$L_1 = \left(\left\{ x \in X \ : \ \sum_{i=1}^{n-1} |x_i| \leq k-1, \ x_n = 1 \right\} \cup S \right),$$

where S consists of some points with coordinate magnitude sum k without the n^{th} coordinate (which equals 1), then X defines a standard minimizer.

Also, if L_0 and L_1 both contain all points having first $n-1$ coordinate magnitudes sum to k, and L_0 contains others as well, we again get a standard minimizer.

However, it is also possible that L_1 contains some but not all points with this sum = k and L_0 some but not all with this sum = $k+1$; or both contain some but not all points with this sum = $k+1$. In either case it is necessary to prove that there is a standard minimizer at least as large with no larger neighborhood.

We proceed by showing that there are four cases that need to be considered, and present arguments to handle them.

Let X be a standard minimizer in $n-1$ dimensions, and let B_k^{n-1} be its core. Then we define k to be its *core size*.

Now for the cases:

Case 1 L_0 has core size k and L_1 has core size $k-1$; The shell of L_0 is only nonempty in the first orthant and the shell of L_1 is full (i.e., is equal to the shell of a ball) in all but the last two orthants.

Case 2 L_0 has core size k and L_1 has core size $k-1$; The shell of L_0 is only nonempty in the first two orthants and the shell of L_1 is full in all but the last orthant.

Case 3 L_0 has core size k and L_1 has core size k; The shell of L_0 is full in all but the last two orthants and the shell of L_1 is only nonempty in the first orthant.

Case 4 L_0 has core size k and L_1 has core size k; The shell of L_0 is full in all but the last orthant and the shell of L_1 is only nonempty in the first two orthants.

First note that these cases are exhaustive, because

(a) $L_0 \supseteq L_1 \supseteq \text{int}_n(L_0)$, where the first two sets are considered without the n^{th} coordinate; and

(b) it is not possible to have the next to last orthant non-full in L_1 (resp. L_0) and the second orthant nonempty in L_0 (resp. L_1) if the core sizes differ (resp. are the same) among L_0 and L_1.

Both these statements follow from the fact that our set is stable under normalization in the first direction. For (b) note that if a point in the second orthant with $2n$-coordinate number

$$(0,1,1,\ldots,1,0,\alpha,\beta,\gamma,\ldots,0)$$

is in L_0, then the point with $2n$-number

$$(1,0,0,\ldots,0,0,\alpha,\beta-1,\gamma,\ldots,1)$$

must be in L_1; while if

$$(1,1,\ldots,1,0,\alpha,\beta,\gamma,\ldots,1)$$

is in L_1, then

$$(0,0,\ldots,0,0,\alpha,\beta,\gamma,\ldots,1)$$

must be in L_0.

Finally, we can transform the set in each case step by step into one that more closely resembles a standard minimizer, without increasing the neighborhood, until we terminate with a standard minimizer. This is done by transferring points between the orthants, until the shell is nonempty in only one of L_0 and L_1. Some more details of this are given in [81].

Bibliography

[1] K. Aardal and F. Eisenbrand. Integer programming, lattices, and results in fixed dimension. In G.L. Nemhauser K. Aardal and R. Weismantel, editors, *Discrete Optimization*, volume 12 of *Handbooks in Operations Research and Management Science*, pages 171 – 243. Elsevier, 2005.

[2] K. Aardal, C.A.J. Hurkens, and A.K. Lenstra. Solving a system of linear Diophantine equations with lower and upper bounds on the variables. *Math. Oper. Res.*, 25(3):427–442, 2000.

[3] K. Aardal and A.K. Lenstra. Hard equality constrained integer knapsacks. *Math. Oper. Res.*, 29(3):724–738, 2004. Erratum: Math. Oper. Res. **31**(2006), no. 4, p 846.

[4] K. Aardal and F. von Heymann. On the structure of reduced kernel lattice bases. In M. Goemans and J. Correa, editors, *Integer Programming and Combinatorial Optimization*, volume 7801 of *Lecture Notes in Computer Science*, pages 1–12. Springer-Verlag, Berlin Heidelberg, 2013.

[5] K. Aardal and L.A. Wolsey. Lattice based extended formulations for integer linear equality systems. *Math. Prog.*, 121:337–352, 2010.

[6] K. Aardal and L.A. Wolsey. More on lattice reformulations of integer programs. Working paper, 2010.

[7] A. Akhavi. The optimal LLL algorithm is still polynomial in fixed dimension. *Theoret. Comput. Sci.*, 297(1-3):3–23, 2003. Latin American theoretical informatics (Punta del Este, 2000).

[8] K. Azuma. Weighted sums of certain dependent random variables. *Tôhoku Math. J. (2)*, 19(3):357–367, 1967.

[9] A. Bachem and R. Kannan. Applications of polynomial Smith normal form calculations. In *Numerische Methoden bei graphentheoretischen und kombinatorischen Problemen, Band 2 (Tagung, Math. Forschungsinst., Oberwolfach, 1978)*, volume 46 of *Internat. Ser. Numer. Math.*, pages 9–21. Birkhäuser, Basel, 1979.

[10] E. Balas. Intersection cuts—a new type of cutting planes for integer programming. *Operations Res.*, 19:19–39, 1971.

[11] E. Balas. Disjunctive programming. *Ann. Discrete Math.*, 5:3–51, 1979.

[12] E. Balas, S. Ceria, and G. Cornuéjols. A lift-and-project cutting plane algorithm for mixed 0-1 programs. *Math. Programming*, 58(3, Ser. A):295–324, 1993.

[13] E. Balas and R.G. Jeroslow. Strengthening cuts for mixed integer programs. *European J. Oper. Res.*, 4(4):224 – 234, 1980.

[14] E. Balas and M. Perregaard. A precise correspondence between lift-and-project cuts, simple disjunctive cuts, and mixed integer Gomory cuts for 0-1 programming. *Math. Program.*, 94(2-3, Ser. B):221–245, 2003. The Aussois 2000 Workshop in Combinatorial Optimization.

[15] S.L. Bezrukov. Construction of the solutions of a discrete isoperimetric problem in a Hamming space. *Mat. Sb. (N.S.)*, 135(177)(1):80–95, 143, 1988. English translation in Math. USSR-Sb. 63 (1) (1989), 81–96.

[16] S.L. Bezrukov. Isoperimetric problems in discrete spaces. *Extremal Problems for Finite Sets*, 3:59–91, 1994.

[17] S.L. Bezrukov and O. Serra. A local-global principle for vertex-isoperimetric problems. *Discrete Math.*, 257(2–3):285–309, 2002.

[18] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. *Linear matrix inequalities in system and control theory*, volume 15 of *SIAM Studies in Applied Mathematics*. SIAM, Philadelphia, PA, 1994.

[19] J.W.S. Cassels. *An introduction to the geometry of numbers*. Classics in Mathematics. Springer-Verlag, Berlin, 1997. Corrected reprint of the 1971 edition.

[20] V. Chvátal. Edmonds polytopes and a hierarchy of combinatorial problems. *Discrete Math.*, 4:305–337, 1973.

[21] G.F. Clements and B. Lindström. A generalization of a combinatorial theorem of Macaulay. *J. Combin. Theory*, 7(Ser. A):230–238, 1969.

[22] H. Cohen. *A course in computational algebraic number theory*. Springer-Verlag, New York, 1993.

[23] W. Cook, C.R. Coullard, and Gy. Turán. On the complexity of cutting-plane proofs. *Discrete Appl. Math.*, 18(1):25–38, 1987.

[24] W. Cook, R. Kannan, and A. Schrijver. Chvátal closures for mixed integer programming problems. *Math. Program.*, 47(2, (Ser. A)):155–174, 1990.

[25] W. Cook, T. Rutherford, H.E. Scarf, and D. Shallcross. An implementation of the generalized basis reduction algorithm for integer programming. *ORSA J. Comput.*, 5:206–212, 1993.

[26] G. Cornuéjols and M. Dawande. A class of hard small 0-1 programs. *INFORMS J. Comput.*, 11:205–210, 1999.

[27] G. Cornuéjols and Y. Li. Elementary closures for integer programs. *Oper. Res. Lett.*, 28(1):1–8, 2001.

[28] G. Cornuéjols, Y. Li, and D. Vandenbussche. K-cuts: A variation of gomory mixed integer cuts from the lp tableau. *INFORMS J. Comput.*, 15(4):385–396, December 2003.

[29] G. Cornuéjols, R. Urbaniak, R. Weismantel, and L.A. Wolsey. Decomposition of integer programs and of generating sets. In R.E. Burkard and G.J. Woeginger, editors, *Algorithms – ESA '97*, volume 1284 of *Lecture Notes in Computer Science*, pages 92–103. Springer-Verlag, 1997.

[30] M.J. Coster, A. Joux, B.A. LaMacchia, A.M. Odlyzko, C.-P. Schnorr, and J. Stern. Improved low-density subset sum algorithms. *Comput. Complexity*, 2(2):111–128, 1992.

[31] A. Dall, F. von Heymann, and B. Vogtenhuber. Sets with small neighborhood in the integer lattice. In M. Noy and J. Pfeifle, editors, *DocCourse combinatorics and geometry 2009 : discrete and computational geometry*, volume 5.3 of *CRM Documents*, pages 163–181. Centre de Recerca Mathemàtica, 2010.

[32] G.B. Dantzig. Maximization of a linear function of variables subject to linear inequalities. In *Activity Analysis of Production and Allocation*, Cowles Commission Monograph No. 13, pages 339–347. John Wiley & Sons Inc., New York, NY, 1951.

[33] G.B. Dantzig. *Linear programming and extensions*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, corrected edition, 1998.

[34] I. Dinur, G. Kindler, R. Raz, and S. Safra. Approximating CVP to within almost-polynomial factors is NP-hard. *Combinatorica*, 23(2):205–243, 2003.

[35] M. Fischetti, A. Lodi, and A. Tramontani. On the separation of disjunctive cuts. *Math. Program.*, 128(1-2, Ser. A):205–230, 2011.

[36] M. Fischetti and C. Saturni. Mixed-integer cuts from cyclic groups. *Math. Program.*, 109(1, Ser. A):27–53, 2007.

[37] M.A. Frumkin. Polynomial time algorithms in the theory of linear diophantine equations. In *FCT*, pages 386–392, 1977.

[38] M.A. Frumkin and A.A. Votjakov. An algorithm for finding the general integer solution of a system of linear equations. In *Studies in discrete optimization (Russian)*, pages 128–140. Izdat. "Nauka", Moscow, 1976.

[39] J. von zur Gathen and J. Gerhard. *Modern Computer Algebra*. Cambridge University Press, 2003.

[40] J. von zur Gathen and M. Sieveking. Weitere zum Erfüllungsproblem polynomial äquivalente kombinatorische Aufgaben. In *Komplexität von Entscheidungsproblemen 1976*, pages 49–71, 1976.

[41] R.E. Gomory. Outline of an algorithm for integer solutions to linear programs. *Bull. Amer. Math. Soc.*, 64:275–278, 1958.

[42] R.E. Gomory. Solving linear programming problems in integers. In *Proc. Sympos. Appl. Math., Vol. 10*, pages 211–215. American Mathematical Society, Providence, R.I., 1960.

[43] R.E. Gomory. An algorithm for integer solutions to linear programs. In *Recent advances in mathematical programming*, pages 269–302. McGraw-Hill, New York, 1963.

[44] M. Grötschel, L. Lovász, and A. Schrijver. Geometric methods in combinatorial optimization. In *Progress in combinatorial optimization (Waterloo, Ont., 1982)*, pages 167–183. Academic Press, Toronto, ON, 1984.

[45] M. Hack. Decision problems for petri nets and vector addition systems. In *MAC Techn. Memo 59, MIT*. MIT, 1975.

[46] L.H. Harper. Optimal numberings and isoperimetric problems on graphs. *J. Combin. Theory*, 1:385–393, 1966.

[47] M. Henk. Löwner-John Ellipsoids. In *Optimization Stories*, Documenta Mathematica, pages 95–106, 2012.

[48] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

[49] S. Janson. On concentration of probability. In *Contemporary combinatorics*, volume 10 of *Bolyai Soc. Math. Stud.*, pages 289–301. János Bolyai Math. Soc., Budapest, 2002.

[50] G.O.H. Katona. A theorem of finite sets. In *Theory of Graphs*, pages 187–207, New York, 1968. Academic Press.

[51] L. G. Khachiyan. Polynomial algorithms in linear programming. *Zh. Vychisl. Mat. i Mat. Fiz.*, 20(1):51–68, 260, 1980. English translation in U.S.S.R. Comput. Math. and Math. Phys. 20, (1980), 53–72.

[52] L.G. Khachiyan. A polynomial algorithm in linear programming. *Dokl. Akad. Nauk SSSR*, 244(5):1093–1096, 1979. English translation in Soviet Math. Dokl. 20 (1979), 191–194.

[53] V. Klee and G.J. Minty. How good is the simplex algorithm? In *Inequalities, III (Proc. Third Sympos., Univ. California, Los Angeles, Calif., 1969; dedicated to the memory of Theodore S. Motzkin)*, pages 159–175. Academic Press, New York, 1972.

[54] B. Krishnamoorthy and G. Pataki. Column basis reduction and decomposable knapsack problems. *Discrete Optim.*, 6(3):242–270, 2009.

[55] J.C. Lagarias and A.M. Odlyzko. Solving low-density subset sum problems. *J. Assoc. Comput. Mach.*, 32(1):229–246, 1985.

[56] S. Lang. *Introduction to Linear Algebra*. Springer-Verlag, Berlin Heidelberg, 2012.

[57] A.K. Lenstra, H.W. Lenstra, Jr., and L. Lovász. Factoring polynomials with rational coefficients. *Math. Ann.*, 261(4):515–534, 1982.

[58] H.W. Lenstra, Jr. Integer programming with a fixed number of variables. *Math. Oper. Res.*, 8(4):538–548, 1983.

[59] Q. Louveaux and L.A. Wolsey. Combining problem structure with basis reduction to solve a class of hard integer programs. *Math. Oper. Res.*, 27(3):470–484, 2002.

[60] L. Lovász and H.E. Scarf. The generalized basis reduction algorithm. *Math. Oper. Res.*, 17:751–764, 1992.

[61] L. Lovász and A. Schrijver. Cones of matrices and set-functions and 0-1 optimization. *SIAM J. Optim.*, 1(2):166–190, 1991.

[62] F.S. Macaulay. Some Properties of Enumeration in the Theory of Modular Systems. *Proc. London Math. Soc.*, s2-26(1):531–555, 1927.

[63] S. Mehrotra and Z. Li. Branching on hyperplane methods for mixed integer linear and convex programming using adjoint lattices. *J. Global Optim.*, 49(4):623–649, 2011.

[64] D. Micciancio. *On the hardness of the shortest vector problem*. ProQuest LLC, Ann Arbor, MI, 1998. Thesis (Ph.D.)–Massachusetts Institute of Technology.

[65] D. Micciancio. The shortest vector in a lattice is hard to approximate to within some constant. *SIAM J. Comput.*, 30(6):2008–2035, 2001.

[66] D. Micciancio and S. Goldwasser. *Complexity of lattice problems*. The Kluwer International Series in Engineering and Computer Science, 671. Kluwer Academic Publishers, Boston, MA, 2002. A cryptographic perspective.

[67] H. Minkowski. *Geometrie der Zahlen*. BG Teubner, 1910.

[68] G.L. Nemhauser and L.A. Wolsey. A recursive procedure to generate all cuts for $0 - 1$ mixed integer programs. *Math. Program.*, 46(1-3):379–390, 1990.

[69] G.L. Nemhauser and L.A. Wolsey. *Integer and combinatorial optimization*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons Inc., New York, 1999. Reprint of the 1988 original, A Wiley-Interscience Publication.

[70] A. Nemirovski. On self-concordant convex-concave functions. *Optim. Methods Softw.*, 11/12(1-4):303–384, 1999. Interior point methods.

[71] C.H. Papadimitriou and K. Steiglitz. *Combinatorial Optimization: Algorithms and Complexity*. Dover Books on Computer Science Series. DOVER PUBN Incorporated, 1998.

[72] A.O. Pittenger. Sharp mean-variance bounds for Jensen-type inequalities. *Statist. Probab. Lett.*, 10(2):91–94, 1990.

[73] R.L. Rivest. On the optimality of elia's algorithm for performing best-match searches. In *IFIP Congress*, pages 678–681, 1974.

[74] A. Schrijver. On cutting planes. *Ann. Discrete Math.*, 9:291–296, 1980. Combinatorics 79 (Proc. Colloq., Univ. Montréal, Montreal, Que., 1979), Part II.

[75] A. Schrijver. *Theory of linear and integer programming*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons Ltd., Chichester, 1986. A Wiley-Interscience Publication.

[76] Alexander Schrijver. *Combinatorial optimization. Polyhedra and efficiency. Vol. A*, volume 24 of *Algorithms and Combinatorics*. Springer-Verlag, Berlin, 2003. Paths, flows, matchings, Chapters 1–38.

[77] H.D. Sherali and W.P. Adams. A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. *SIAM J. Discrete Math.*, 3(3):411–430, 1990.

[78] N.Z. Shor. Use of the space expansion operation in problems of convex function minimalization. *Kibernetika (Kiev)*, (1):6–12, 1970. English translation in Cybernetics 6 (1970), 7–15.

[79] N.Z. Shor. The cut-off method with stretching of the space for the solution of convex programming problems. *Kibernetika (Kiev)*, (1):94–95, 1977. English translation in Cybernetics 13 (1977), 94–96.

[80] S.P. Tarasov, L.G. Khachiyan, and I.I. Èrlikh. The method of inscribed ellipsoids. *Dokl. Akad. Nauk SSSR*, 298(5):1081–1085, 1988. English translation in Soviet Math. Dokl. 37 (1988), no. 1, 226-230.

[81] D.-L. Wang and P. Wang. Discrete isoperimetric problems. *SIAM J. Appl. Math.*, 32(4):860–870, 1977.

[82] D.-L. Wang and P. Wang. Extremal configurations on a discrete torus and a generalization of the generalized macaulay theorem. *SIAM J. Appl. Math.*, 33(1):55–59, 1977.

[83] D.B. Yudin and A.S. Nemirovski. Estimation of the informational complexity of mathematical programming problems. *Èkonom. i Mat. Metody*, 12(1):128–142, 1976. English translation in Matekon 13 (2) (1976), 3–25.

[84] D.B. Yudin and A.S. Nemirovski. Informational complexity and effective methods for the solution of convex extremal problems. *Èkonom. i Mat. Metody*, 12(2):357–369, 1976. English translation in Matekon 13 (3) (1977), 25–45.

[85] G.M. Ziegler. *Lectures on polytopes*, volume 152 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1995.