Geometric and algebraic approaches to mixed-integer polynomial optimization using sos programming

Dissertation

zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades
„Doctor rerum naturalium“
der Georg-August-Universität Göttingen

im Promotionsprogramm „PhD School of Mathematical Sciences“ (SMS)
der Georg-August University School of Science (GAUSS)

vorgelegt von

Sönke Behrends
aus Marburg/Lahn

Göttingen, 2017

Betreuungsausschuss
Prof. Dr. Anita Schöbel, Institut für Numerische und Angewandte Mathematik, Georg-August-Universität Göttingen
Prof. Dr. Russell Luke, Institut für Numerische und Angewandte Mathematik, Georg-August-Universität Göttingen

Mitglieder der Prüfungskommission
Referentin: Prof. Dr. Anita Schöbel, Institut für Numerische und Angewandte Mathematik, Georg-August-Universität Göttingen
Korreferent: Prof. Dr. Oliver Stein, Institut für Operations Research, Karlsruher Institut für Technologie

Weitere Mitglieder der Prüfungskommission
Prof. Dr. Carsten Damm, Institut für Informatik, Georg-August-Universität Göttingen
Prof. Dr. Anja Fischer, Juniorprofessur Management Science, Technische Universität Dortmund
Prof. Dr. Ralf Meyer, Mathematisches Institut, Georg-August-Universität Göttingen
Prof. Dr. Gerlind Plonka-Hoch, Institut für Numerische und Angewandte Mathematik, Georg-August-Universität Göttingen

Tag der mündlichen Prüfung: 23. Oktober 2017

Preface

Mixed-integer nonlinear programs, or MINLPs for short, constitute a large class of optimization problems. The name also reveals its two most challenging features: Nonlinearity and integrality. Concerning nonlinearity, neither the objective nor the constraint functions need to be linear in MINLP, and these nonlinearities are difficult to handle theoretically. For example, ignoring integrality for the moment, the continuous problem need not be convex, which makes global optimization difficult, if not impossible. Nonlinearities also present significant numerical challenges; already the restricted class of polynomials leads to notorious numerical instabilities. Concerning integrality, even "easy" linear optimization problems with integrality requirements are NP-hard in varying dimensions (Theorem 1.35), and the classical optimization methods from "the smooth world" – roughly speaking, compute the gradient and solve for critical points; if you can, use the Hessian or an approximation of it cleverly, too – are, naively carried out, pointless in face of integer variables. Even worse, in the general form of MINLP, it is not even guaranteed that the functions at hand are differentiable, so even first-order methods (exploiting and applying gradient information in the solution process) cannot be applied to continuous subproblems or relaxations. Combining both difficulties, the situation gets worse. We end up with a class of optimization problems that are so difficult that, even when we restrict to the all-integer special case of a polynomial objective, discard all constraint functions, it can still be shown that there cannot exist an algorithm solving the problems in this subclass (Theorem 1.39). In short, the class of MINLP is too large to hope for a general solution. So why should one even try to approach MINLP? Because, by all means, this is not the end of the story. Many important subclasses are well-understood and solvable, and for the remaining ones, a plethora of techniques exist that may allow to solve a given problem. And the fact that many problems in science and industry can be accurately modeled in the form of MINLP ensures that the demand for progress in this field remains high.

Our contribution is the presentation of geometric approaches that assist in the solution process of MINLP. Amongst others, we compute half-spaces, seminorm balls and ellipsoids that contain the relaxed feasible set. We also compute norm balls that contain all optimal solutions, which is formalized in the concept of *norm bounds*. We then investigate how integrality arguments can be used to shrink these sets and potentially cut off continuously feasible points. The norm and seminorm balls as well as the ellipsoids make the integer part of the problem (given some assumptions) accessible to branch and bound. For the branch and bound part, we also propose a class of underestimators that yield tight lower bounds. The approaches always involve the task to optimally choose a geometric object out of a whole class of similar geometric objects – for example, to choose an ellipsoid of minimal volume containing the relaxed feasible set out of the col-

lection of all axis-parallel ellipsoids. For polynomial objective and constraint functions, these auxiliary problems, or approximations thereof, become tractable by using sum of squares programming and tools from real algebra. These auxiliary problems can then be implemented and assist in the solution process of a given problem. We also present results that guarantee that the approximate solutions converge eventually to the optimal solution of the auxiliary problem. A subset of the approaches has been implemented and we demonstrate that they work in computer experiments.

Contents

1. Introduction

In this chapter, we introduce MINLP, or mixed-integer nonlinear programming, a large class of optimization problems, with a focus on polynomials. We motivate our contributions and supply the necessary preliminaries from various fields of mathematics that are used in the chapters to follow.

Section 1.1 gives the formal definition of the problem class. The associated feasible set and its relaxed variant are the point of entry for our geometric approaches. We motivate polynomial constraint and objective functions.

Section 1.2 outlines the structure of this work. The aim of every chapter is summarized in a few sentences, followed by a brief summary of the chapter's sections.

Section 1.3 gives an overview on the literature on mixed-integer nonlinear programming. We refer to surveys as well as negative and positive results.

Section 1.4 presents the tools that are necessary for our approaches. Starting off with the basics on numbers, norms and topology, we proceed with polynomial algebra and give the definition of a quadratic module. We settle terminology from optimization and introduce coercivity, a key property in this work. Furthermore, matrices, valid inequalities and cuts, convexity, polyhedra and spectrahedra as well as ellipsoids are introduced.

Section 1.5 repeats sos programming, a class of optimization problems. Almost all of our approaches result in an sos program, hence this technique is central for our work. Together with results from real algebraic geometry, we outline how sos programming can be used to approximate continuous polynomial programs. Many of the auxiliary programs we consider are continuous polynomial programs. We also show how sos programs translate to semidefinite programs.

Section 1.6 closes this chapter. We repeat fundamental results from complexity theory which help us judge throughout the following chapters which auxiliary problems are tractable and which are not. Furthermore, we recall how various geometric, algebraic and analytic properties relate to the existence of optimal solutions to MINLP.

1.1. Mixed-Integer Nonlinear Programming

Enter MINLP. In the most general form that we consider in this work, a mixed-integer nonlinear program reads

$$\begin{aligned} \min \quad & f(x) & & \\ \text{s.t.} \quad & g_k(x) \geq 0, & k = 1, \ldots, r, & \qquad \text{(MINLP)} \\ & x_i \in \mathbb{Z}, & i \in \mathcal{I}, & \qquad \text{(MIPP)} \\ & x \in \mathbb{R}^n, & & \end{aligned}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is the objective function and $g_1, \ldots, g_r : \mathbb{R}^n \to \mathbb{R}$ are the constraint functions. If f and g_i are arbitrary functions, we refer to the program as MINLP. The variant with polynomial data, i.e., f and g_i are polynomials in n variables, is referred to by MIPP. The integer variables are indexed by the set $\mathcal{I} \subset \{1, \ldots, n\}$.

The program and its geometry

Let us define two sets associated with MINLP. The set of all *feasible solutions with integrality relaxed*, namely,

$$F := \{x \in \mathbb{R}^n : g_1(x) \geq 0, \ldots, g_r(x) \geq 0\}$$

which we denote the *relaxed feasible set* for short, and the set of all *feasible solutions* with integrality enforced, that is,

$$F_{\mathcal{I}} := \{x \in F : x_i \in \mathbb{Z}\ \forall i \in \mathcal{I}\},$$

the *feasible set* for short.

A key step in solving MINLP is to understand and exploit the geometry of the sets F and $F_{\mathcal{I}}$, and then to approximate and finally dissect them, with the objective to simplify a complicated problem into – hopefully – easier subproblems. Actually, there are several classical [BP03] techniques for and approaches to the solution of MINLP that rely on geometric and combinatorial properties of F and $F_{\mathcal{I}}$. Three of the most common techniques are branch and bound, or B&B for short, outer approximation and (linear) cuts. In the following, we show how our approaches contribute to each of them.

Branch and bound. For the integer variables, a recurring solution ingredient is some form of branch and bound (see, e.g., Chapter 3 in [BKL+13], Preface in [LL12]). Enumeration of the integer part of candidate optimal solutions can be depicted as a rooted tree, and branch and bound examines its branches, trying to discard branches once it can be concluded the branch does not contain an optimal solution by using lower and upper bounds on the objective. If a branch and bound approach is to succeed, two factors are decisive: A small search tree (the fewer nodes the better) and tight bounds on the objective (to prune branches as early as possible).

We address both in this work. Concerning the search tree, we explore how to compute sets of minimal size that contain F (or F intersected with a suitable sublevel set in the case of a norm bound). These sets come in the form of norm and seminorm balls as well as ellipsoids, and allow for easy enumeration of the integer variables. Since finding these sets involves a choice, we cast these problems as auxiliary optimization problems. Using integrality arguments, we can shrink the enclosing sets further so that they still contain $F_{\mathcal{I}}$ (or $F_{\mathcal{I}}$ intersected with a sublevel set) solutions but potentially cut off points from F. This information helps to keep the number of enumerated solutions and hence the search tree small. For the unconstrained case and a polynomial objective, we compare our norm bounds with a norm bound from the literature [Mar03]. We show that our norm bound is never worse and, for dense instances, better. Computer experiments on random instances show that our norm bounds are smaller by orders of magnitude, allowing to solve problems which would have been previously unsolvable. Regarding the lower bounds on the objective function, we derive the lower bounds from *underestimators*, where the objective f is underestimated on a subset U of $\mathbb{R}^n$ by g if $g(x) \leq f(x)$ for all $x \in U$. To this end we introduce a class of easy-to-minimize underestimators for mixed-integer minimization problems. Experiments on random instances show that they perform well.

Outer approximations. A further recurring component to the solution of MINLP are outer approximations. The meaning of outer approximation is, in the literature on optimization, twofold: It is the name of a celebrated solution method for a special class of MINLP [DG86; FL94], and also describes the process of relaxing a complicated set to a larger set (hence *outer* approximation) that is easier to handle. An important special case of outer approximation is outer linearization [Geo70], where the set F is approximated by a polyhedron. This results in giving *valid inequalities* for F, which can be derived, e.g., by using gradient information if the associated constraint functions are convex. Geometrically, valid linear inequalities correspond to *half-spaces* containing F.

In this work, we also study the problem to find tight valid linear inequalities for F. More generally, such outer approximations by linear functions can be seen as a convexification [TS02] of the problem, where a non-convex feasible set or objective is approximated by convex sets or a convex objective, respectively. In this view, our enclosing half-spaces, norm and seminorm balls as well as ellipsoids are convexifications of the problem.

Cuts. The third solution method that relies on geometric properties of MINLP and that has proved successful are linear cuts, also known as cutting planes in the literature. The famous article "Solution of a Large-Scale Traveling-Salesman Problem" by Dantzig, Fulkerson and Johnson [DFJ54] can be seen as the first linear cut algorithm (p. 7–9 in [JLN+10]), and since then, linear cuts had a tremendous impact on integer programming, or as the authors in in [JLN+10], p. 9, put it, "Great new ideas may transform the discipline they came from [...] profoundly [...]. The cutting-plane method of George Dantzig, Ray Fulkerson, and Selmer Johnson had the same kind of impact on the dis-

cipline of mathematical programming." A linear cut for F, the relaxed feasible set, is a valid linear inequality for the feasible set $F_{\mathcal{I}}$. Roughly speaking, cutting planes are often used in a solution framework as follows: If an optimal solution of the continuous relaxation – the task to minimize the objective over F instead of $F_{\mathcal{I}}$ – has been found which is infeasible for the original problem (else an optimal solution of the original problem is found), add a linear cut which cuts off this solution, and solve again. Cutting planes have been extended to convex programming. Even more, the concept of a cut has been generalized to nonlinear cuts, sometimes called "nonlinear cutting planes", see, e.g., [BF76], [LS00], [MB09].

The geometric approaches we consider provide – linear or nonlinear – cuts as follows. We search for half-spaces, seminorm balls or ellipsoids containing the set of feasible solutions $F_{\mathcal{I}}$ (and norm bounds on the optimal solutions). Since these problems are themselves mixed-integer nonlinear programs with possibly infinitely many constraints and too difficult in practice, we eventually relax integrality and require containment of F, the relaxed feasible set, in the half-space, seminorm ball, or ellipsoid instead (or, containment of F intersected with a suitable sublevel set in the case of norm bounds). Once a, say, seminorm ball containing F is chosen, we can think of it as a valid – nonlinear in this case – inequality for F. It is then, in a second step, often possible to use integrality arguments to "tighten" the inequality further, so that the inequality still holds for all (non-relaxed) feasible solutions, that is, $F_{\mathcal{I}}$, but is possibly violated at a point in F. Thus, we have a cut for the feasible set.

Real algebra for real polynomials

So far we have seen that in solving MINLP, geometry is omnipresent. Since the title is "Geometric and algebraic approaches to mixed-integer polynomial optimization using sos programming", let us explain where the polynomials and algebra is in all of this.

Our different geometric approaches all involve auxiliary optimization problems. These problems can usually be formulated in very general terms and in a first step, we attempt to infer information about existence of feasible and optimal solutions of these auxiliary problems with as few assumptions as possible. These results are important from a theoretical perspective, as they justify the chosen auxiliary problem and show that it is well-posed.

However, it is of little use to replace a difficult problem (MINLP in this case) by another problem (the auxiliary problem), if the latter is not easier to solve. The path that we pursue in this work to make the auxiliary problems tractable is to restrict the problem data – not immediately, but at some point – to polynomials.

The primary motivation for polynomials is that they constitute a large class of nonlinear functions. In a sense, polynomials are prototypes of (continuous) nonlinear functions. To make this precise, we can refer to the Stone-Weierstrass theorem (Theorem 1.1): Every continuous real function on a compact subset of $\mathbb{R}^n$ can be uniformly approximated by polynomials in n variables. Also, the set of polynomials is closed under some common operations: They can be added, multiplied by scalars and multiplied by other polynomials, without leaving the class (formally, the polynomial functions are a subalgebra of

the algebra of all continuous functions on a subset of $\mathbb{R}^n$), which is very handy for algebraic manipulations. Moreover, polynomials have been extensively studied in real and complex algebra as well as in algebraic geometry, which allows us to connect algebra and geometry. We make extensive use of concepts and results from real algebra (amongst others, the *Positivstellensatz*, Theorem 1.20).

Polynomials are also easy to handle with a computer, as they have the nice property that they are completely parameterized by their coefficients. In contrast, the implementation of functions involving limits (holomorphic functions as sin, exp, ln..., parameter integrals, ...) is rather involved. If the data is rational, it is moreover possible to implement them exactly, at least in principle. The fact that polynomials are described by their coefficients also allows to draw random samples from families of polynomials. In contrast, it is rather difficult to sample from the space of all continuous functions on a subset of $\mathbb{R}^n$.

Another good reason for polynomials is a comparably new result by Lasserre [Las01] that allows to approximately, sometimes even accurately, solve continuous polynomial optimization problems. The result has lead to a technique known as sum of squares programming, which has been powerfully influenced by above-mentioned Positivstellensatz from real algebra, since the Positivstellensatz guarantees under an additional assumption the convergence of an approximating hierarchy towards the actual solution. These sum of squares programs, or sos programs for short, reduce to semidefinite programs, and the latter are well understood. We refer to Section 1.5 for details. Throughout this work, we rely on sos methods to (approximately) solve the auxiliary problems and continuous relaxations.

The last advantage of polynomials that we wish to mention is the following: If a polynomial has integer coefficients, it is integrality preserving – at integer points, it attains integer values. This information can and is used in our approaches to deduce nonlinear cuts.

1.2. Structure of this work

Chapter 1 is the introduction to this work. In Section 1.1 we state the program class, outline some geometric solution approaches and makes the case for polynomials. Section 1.2 is this overview. A literature review is given in Section 1.3. Notation and results from various fields of mathematics are settled in Section 1.4. The approximation of continuous polynomial optimization by sos programming is repeated in Section 1.5. The chapter repeats complexity results for MINLP in Section 1.6, along with a collection of known existence results on optimal solutions to MINLP.

Chapter 2 introduces half-spaces for MINLP, our first geometric approach. In Section 2.1 we motivate half-spaces with known results from linear and convex programming. In Section 2.2 we formulate the task to find tight a half-space as an auxiliary program. For polynomial constraints, we show in Section 2.3 how the problem can be approximated by sos methods. The chapter closes by exploring ways to turn the valid linear inequalities into linear cuts in Section 2.4.

Chapter 3 describes the computation of norm bounds for MINLP, our second approach. Under a coercivity condition, we compute upper bounds on the norm of all optimal solutions. This is formalized in the concept of *norm bounds* in Section 3.1. In the unconstrained case and for a polynomial objective, we compare our norm bound with a norm bound from the literature and show that our norm bound is never worse and better for dense instances. In Section 3.2 we discuss norm bounds for the convex quadratic case and give an application to systems of polynomial equations. We evaluate the norm bounds numerically in Section 3.3.

Chapter 4 analyzes seminorm balls containing the feasible set. In Section 4.1 we motivate the seminorm bounds and show the relation norm bounds. We formulate the auxiliary program in Section 4.2 and give an approximating hierarchy of sos programs along with a convergence result. Section 4.3 explores how Diophantine arguments can be used to derive a cut.

Chapter 5 outlines the fourth and last geometric approach to MINLP: ellipsoids. In Section 5.1, we give an auxiliary program that computes ellipsoids of minimal volume containing the feasible set. Here, we may optimize the shape as well as the center of the ellipsoid. In Section 5.2 we use ideas from the literature [ND05] to linearize the constraints. If the constraint functions are polynomials, we show in Section 5.3 that the problem reduces to a hierarchy of concave semidefinite minimization problems.

Chapter 6 is devoted to the solution of unconstrained (mixed-)integer polynomial optimization, a special case of MINLP. Properties of this subclass are discussed in Section 6.1. In Section 6.2, a class of suitable underestimators is proposed. We implemented

a full branch and bound framework, allowing for the computational solution of a given instance provided its leading form is positive definite. The solution algorithm is presented in Section 6.3. Using the norm bounds from Chapter 3, we evaluate our underestimators on random instances.

Chapter 7 takes a closer look at coercive polynomials. Coercivity appears in one form or another throughout this work, and that chapter explores coercivity in terms of the so-called order of coercivity. We give additional motivation for coercivity in Section 7.1. Then, we recall the order and stability of coercivity in Section 7.2. Section 7.3 contains the main result and relates both concepts. In Section 7.4, we present families with small order of coercivity. We introduce the minimal order of coercivity in Section 7.5. The chapter closes with a look towards the decision problem whether a given polynomial is coercive in Section 7.6.

Chapter 8 summarizes this work and points towards extensions of the presented results. Section 8.1 contains the summary. Section 8.2 yields ideas for underestimation of quadratic functions using the S-lemma and discusses a subgradient-type approach. We also look at robust polynomial optimization problems using quantifier elimination and extensions of sos programming.

Appendix. The work closes with an appendix. We explore the role of sublevel sets and tight inequalities in Chapter A. Supplementary proofs are given in B.

1.3. Literature review

Surveys. The literature on nonlinear mixed-integer programming is vast. For an overview, a presentation of key techniques and complexity results as well as numerous references for further reading, see the article [HKLW10], which comes as chapter of [JLN+10]. For a recent survey on nonlinear mixed-integer programming, see [LL12]. We also wish to mention the recent book [BKL+13], which also covers modeling, convex methods, heuristics and software for MINLP.

Complexity. Integrality turns even seemingly simple problems incomputable: Based on results of Matiyasevich, Jeroslow [Jer73] proved that there cannot be an algorithm for integer minimization of a linear form subject to quadratic constraints (Theorem 1.40), a negative result. But there are also many positive results, see, e.g., [Hei05], [LHKW06], [DPHWZ16], [HWZ16]. For a collection of further complexity results we refer to [Köp12], which is part of [LL12].

Algorithms for polynomial integer programming in special cases. But substantial special cases are solvable, for example, every integer problem with a bounded feasible set is solvable. More specifically, an important case is Boolean programming, see [BH02] for a survey. A classic approach is linearization by introducing new variables and constraints (for early results see, e.g., [For60]). In theory, also a general bounded integer polynomial optimization problem can be reduced to the binary case [Wat67], but this is not practicable since the number of variables grows too much. Another technique for Boolean polynomial programming is the reduction to a quadratic problem which can be done with significantly fewer variables and constraints [Ros75; BR07]. Another substantial case that gained attention are (quasi-)convex problems, as the incomputability results do not hold for this case [Kha83; KP00]. [HK13] present a Lenstra-type algorithm for quasiconvex integer polynomial optimization.

Unconstrained mixed-integer polynomial optimization. Unconstrained quadratic integer minimization is considered by [BHS15]. We did not find results in the literature that consider the unconstrained mixed-integer minimization problem for multivariate polynomials of arbitrary degree.

Branch and bound. As indicated before, a common solution ingredient for the integer variables in MINLP is branch and bound as proposed (originally only for convex functions) by [GR85]. A popular method is to calculate convex underestimators (see, e.g., [LT11] for polynomial functions) to obtain lower bounds. As a different approach, if the feasible set is a box and the objective a polynomial, [BD14] compute separable underestimators which give lower bounds that are easy to obtain. In contrast, [LHKW06] directly compute lower and upper bounds, i.e., no underestimators, for nonnegative polynomials on polytopes.

Sos programming. Throughout our work we rely on methods from constrained *continuous* polynomial optimization. Based on work of Shor [Sho87; SS97], Parrilo [Par00] suggested a method now known as *sos programming* that makes continuous polynomial optimization accessible to semidefinite programming (see, e.g., [WSV00] for the latter), whilst Lasserre [Las01] published the dual approach, based on moment sequences. Since the emergence of the two ground-breaking publications by Parrilo and Lasserre, many results on continuous polynomial optimization via sos techniques and its theoretical background have been published, we only name a few: The expository paper [PS03] shows that existing algebraic techniques are outperformed by the sos method. As in-depth treatments, we refer to [AL12] for the interplay of semidefinite, conic and polynomial optimization, and [BPT13] for a focus on the geometry involved. For an algebraic treatment, we mention Marshall's book [Mar08]. We point out Laurent's elegant survey [Lau09], which treats, among other aspects, the duality of the sos and moment approach.

Chapter-specific literature. We end our literature review remarking that we present additional references to the literature in each of the following chapters that are specific to the topic at hand.

1.4. Preliminaries

In this section we introduce basic concepts and notation that is used throughout this work.

1.4.1. Numbers

The natural, integer, rational, real and complex numbers are denoted by $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ respectively. In this work, the natural numbers do not contain 0, and we denote $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. For $n \in \mathbb{N}$, we let

$$[n] := \{1, \dots, n\}.$$

As is well-known, the supremum and infimum of a subset of the real numbers always exist but may be infinite. Especially, $\inf \emptyset = +\infty$ and $\sup \emptyset = -\infty$. Let R be any of the rings $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$. We use the notation $R_{\geq 0} := \{x \in R : x \geq 0\}$, and $R_{>0} := \{x \in R : x > 0\}$.

1.4.2. Norms, seminorms, topology

A seminorm N on a real or complex vector space V is a real-valued function such that

1. $N(x + y) \leq N(x) + N(y)$
2. $N(ax) = |a|N(x)$

for all $x, y \in V$ and scalars a, see, e.g., Definitions 1.33 in [Rud91]. Property 1 is called *subadditivity*, whilst property 2 is called *absolute homogeneity*. If, additionally *definiteness* holds, that is,

3. $N(x) = 0$ implies $x = 0$

for all $x \in V$, the seminorm N is a *norm*, and usually denoted by $\|\cdot\|$.

The *(open) seminorm ball* with center $p \in V$ and radius $R \in \mathbb{R}$ is given by

$$B_R(p; N) := B_R^N(p) := \{x \in V : N(x - p) < R\}.$$

The *(closed) seminorm ball* with center $p \in V$ and radius $R \in \mathbb{R}$ is given as

$$\mathbb{B}_R(p; N) := \mathbb{B}_R^N(p) := \{x \in V : N(x - p) \leq R\}.$$

We define *norm balls*, open or closed, analogously.

A set $M \subset V$ of a normed space V is *bounded* if $M \subset B_R(0)$ for some $R > 0$.[1]

A *unit sphere*, or *sphere* for short, is the set $\{x \in \mathbb{R}^n : \|x\| = 1\}$ for a norm $\|\cdot\|$ on $\mathbb{R}^n$. For the important special cases of p-norms, we introduce the notation

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

where the p-norm on $\mathbb{R}^n$, $p \in [1, \infty]$, is given by $\|x\|_p = \sqrt[p]{\sum_{j=1}^n |x_j|^p}$ for $x \in \mathbb{R}^n$.

Now let (X, τ) be a topological space. Then, for $S \subset X$, $\operatorname{cl} X$ denotes the *closure* of S with respect to τ.

[1] As all norms are equivalent on $\mathbb{R}^n$, boundedness does not depend on the choice of the norm on $\mathbb{R}^n$.

1.4.3. Ring of polynomials

Let R be a ring with unit. We denote the ring of polynomials in n unknowns $X_1, \ldots, X_n$ and coefficients in R by $R[X_1, \ldots, X_n]$, which we abbreviate by $R[\underline{X}]$. We use $\underline{X}$ here in order to distinguish the multivariate from the univariate case. Using multi-index notation, we write a polynomial $f \in R[X_1, \ldots, X_n]$ as

$$f = \sum_{\alpha \in A(f)} a_\alpha \, X^\alpha = \sum_{\alpha \in A(f)} a_\alpha X_1^{\alpha_1} \cdots X_n^{\alpha_n}$$

where $A(f) \subset \mathbb{N}_0^n$ indexes the terms $c_\alpha X^\alpha$, the *monomials*, appearing in the definition of f. We use the notation $X^\alpha := X_1^{\alpha_1} \cdots X_n^{\alpha_n}$ for $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}_0^n$. The $a_\alpha \in R$, $\alpha \in A(f)$, are the coefficients of f. We assume that the set $A(f)$ is chosen minimally in the sense that $A(f) = \{\alpha \in \mathbb{N}_0^n : a_\alpha \neq 0\}$; especially, $A(f)$ is always finite. The degree of f is defined as $\deg(f) := \sup_{\alpha \in A(f)} |\alpha|$ with $|\alpha| = \sum_{i=1}^n \alpha_i$ being the *modulus* of α. Notably, $\deg(0) = -\infty$. The set of all polynomials in $R[X_1, \ldots, X_n]$ of degree at most d, some $d \in \mathbb{N}_0 \cup \{-\infty\}$, deserves special attention: We denote it by

$$R[X_1, \ldots, X_n]_d := \{f \in R[X_1, \ldots, X_n]_d : \deg(f) \leq d\},$$

and note that this set forms a vector space.

As an example, for

$$f = X_1^3 + 4X_1^5 X_2^6 \in \mathbb{Z}[X_1, X_2],$$

we have $A(f) = \{(3,0), (5,6)\}$, $c_{(3,0)} = 1$, $c_{(5,6)} = 4$ and $\deg(f) = |(5,6)| = 5 + 6 = 11$.

We express the evaluation of f at some $x \in R^n$ by

$$f(x) = \sum_{\alpha \in A(f)} a_\alpha x^\alpha,$$

where $x^\alpha := x_1^{\alpha_1} \cdots x_n^{\alpha_n}$. The map $R^n \to R$, $x \mapsto f(x)$, is the *polynomial function* defined by f.

1.4.4. Real polynomials

Now consider the ring of real polynomials $\mathbb{R}[X_1, \ldots, X_n]$. A polynomial f is *homogeneous of degree i* if f is a sum of monomials of degree i or the zero polynomial. Equivalently, f is homogeneous of degree i if and only if $f = \sum_{|\alpha|=i} a_\alpha X^\alpha$. A homogeneous polynomial is also called a *form*. Any polynomial $f \in \mathbb{R}[\underline{X}]$ can be uniquely decomposed as a sum of forms

$$f = \sum_{j=0}^{d} f_j$$

where $d := \deg f$ and the f_j are homogeneous polynomials of degree j, called the *homogeneous components* of f. The highest degree component, f_d, is the *leading form* of f. Concerning the dimension, we have by, e.g., Remark 1.2.5 in [Mar08]

$$\dim \mathbb{R}[X_1, \ldots, X_n]_d = \tbinom{n+d}{d} \tag{1.1}$$

and

$$\dim\{f \in \mathbb{R}[X_1, \ldots, X_n]_d : f = f_d\} = \tbinom{n+d-1}{d}. \tag{1.2}$$

For a polynomial f that is homogeneous of degree $d \in \mathbb{N}_0$, one has

$$f(\lambda x) = \lambda^d f(x), \quad x \in \mathbb{R}^n,\ \lambda \in \mathbb{R}.$$

This implies that a homogeneous polynomial is uniquely determined by its values on a sphere.

A homogeneous polynomial f is *positive definite* if $f(x) > 0$ for $x \neq 0$. Similarly, a (possibly non-homogeneous) polynomial f is *positive semidefinite* if $f(x) \geq 0$ for all $x \in \mathbb{R}^n$, for short $f > 0$ and $f \geq 0$. For a homogeneous polynomial f, we often use the following equivalent characterization (which does not depend on the choice of the sphere):

$$\begin{aligned} f \geq 0 &\iff \exists c \geq 0 :\ f(x) \geq c \text{ for all } x \in \mathbb{S}^{n-1}, \\ f > 0 &\iff \exists c > 0 :\ f(x) \geq c \text{ for all } x \in \mathbb{S}^{n-1}. \end{aligned} \tag{1.3}$$

We define the following norms for polynomials:

$$\begin{aligned} \|f\|_1 &:= \sum_{\alpha \in A(f)} |a_\alpha|, \quad f \in \mathbb{R}[\underline{X}] \\ \|f\|_\infty &:= \sup_{\alpha \in A(f)} |a_\alpha|, \quad f \in \mathbb{R}[\underline{X}],\ f \neq 0, \end{aligned}$$

and $\|0\|_\infty := 0$. We furthermore define the "norm"

$$\|f\|_0 := \sum_{\alpha \in A(f)} 1, \qquad f \in \mathbb{R}[\underline{X}]$$

which is, of course, not a norm, but counts the monomials in f.

The following result is from functional analysis and constitutes a special case of the Stone-Weierstrass theorem in a more general form.

Theorem 1.1 (Stone-Weierstrass, see, e.g., Corollary 1.3 in Chapter 3.1 of [Lan93])**.** *Let S be a compact subset of $\mathbb{R}^n$. Any real continuous function on S can be uniformly approximated by polynomial functions in n variables.*

1.4.5. Real algebra

We now present the concept of quadratic modules, following [Mar08]. Here, we take the algebraic point of view which is convenient for algebraic manipulations, before we consider a more geometric approach to quadratic modules in Section 1.5.3. In the following, R is a commutative ring with unit.

Definition 1.2. A subset M of R is a *quadratic module* if

$$M + M \subset M, \quad a^2 M \subset M \text{ for all } a \in R, \text{ and } 1 \in M.$$

A quadratic module M of R is *Archimedean* if for each $f \in R$ there is $k \in \mathbb{N}$ with $f + k \in M$.

Hence, $\sum R^2$ is the smallest (with respect to set inclusion) quadratic module of R, where $\sum R^2$ denote the set of all finite sums $\sum a_i^2$, $a_i \in R$. By a classical hull argument, the quadratic module *generated* by $h_1, \ldots, h_s \in R$, that is, the smallest quadratic module in R containing the h_i, is given by

$$M(h_1, \ldots, h_s) := \left\{ \sum_{i=0}^{s} \sigma_i h_i : \sigma_0, \ldots, \sigma_s \in \sum R^2 \right\},$$

with $h_0 := 1$.

Remark 1.3. The name *Archimedean* is related to the Archimedean property for the real numbers as follows: The latter states that for every $x > 0$ there exists $n \in \mathbb{N}$ with $n > x$. Equivalently, for every $x \in \mathbb{R}$ there exists $n \in \mathbb{N}$ with $x + n \geq 0$, which is, in our new terminology, equivalent to the quadratic module $\sum \mathbb{R}^2 = \mathbb{R}_{\geq 0}$ of $\mathbb{R}$ being Archimedean.

1.4.6. Rounding

Given $x \in \mathbb{R}$, we let $\lfloor x \rfloor$ and $\lceil x \rceil$ denote the integer below and above x. The number $\lfloor x \rceil$ denotes the integer obtained by rounding x to its nearest integer. In case this is not unique, we use the *round the half up* rule, although none of our results depends on the exact nature of the tiebreaker. These definitions extend to vectors $x \in \mathbb{R}^n$ by componentwise application. In the mixed-integer setting, the following notation is useful, too: Given $\mathcal{I} \subset [n]$ indexing the integer variables in MINLP, let $\lfloor x \rceil_{\mathcal{I}}$ denote the vector with components

$$(\lfloor x \rceil_{\mathcal{I}})_i := \begin{cases} \lfloor x_i \rceil, & i \in \mathcal{I}, \\ x_i, & \text{else.} \end{cases}$$

For $S \subset \mathbb{R}^n$, let us also introduce the notation

$$S_{\mathcal{I}} = \{x \in S : x_i \in \mathbb{Z} \text{ for all } i \in \mathcal{I}\}.$$

Whilst using integrality arguments, it is useful to know that a function attains integer values if evaluated at a mixed-integer point. This motivates the following definition: Let $f : S \to \mathbb{R}$ be a function defined on $S \subset \mathbb{R}^n$. The function is called *$\mathcal{I}$-integrality preserving* if $f(S_{\mathcal{I}}) \subset \mathbb{Z}$. It is *integrality preserving* if it is $\mathcal{I}$-integrality preserving for $\mathcal{I} = [n]$. As an example, any $f \in \mathbb{Z}[X_1, \ldots, X_n]$ is integrality preserving. Also, the modulus on $\mathbb{R}$ is integrality preserving. The squared seminorm $\sum_{i \in \mathcal{I}} x_i^2$ is $\mathcal{I}$-integrality preserving on $\mathbb{R}^n_{\mathcal{I}}$ for $\mathcal{I} \subset [n]$.

1.4.7. Notions from optimization

Let us establish some terminology of optimization problems. Consider a minimization problem of the form

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in S \end{array} \tag{OPT}$$

where $f : S' \to \mathbb{R}$ is a function, and S' is some set with $S \subset S'$. The function f is the *objective* or *objective function*, and the number given by evaluating f at $x \in S$ is the *objective value* of f at x. Any point in S' is a *solution* of OPT. The set S is the set of all *feasible solutions*.

The number $\inf\{f(x) : x \in S\} \in [-\infty, +\infty]$ is the *optimal value*.[2] Any point $\bar{x}$ in $\arg\min_{x \in S} f(x)$ is an *optimal solution* (or *minimizer*). Note that, with these definitions, OPT always has an optimal value but might not have optimal solutions (or not even feasible solutions).

It is sometimes convenient to use the following special terms in the presence of integrality constraints. If $F \subset \mathbb{R}^n$, and we consider the minimization problem

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in \mathbb{Z}^n \\ & x \in F, \end{array} \tag{1.4}$$

then the problem

$$\begin{array}{ll} \min & f(x) \\ & x \in F \end{array} \tag{1.5}$$

is the *continuous relaxation*, the set F is the *relaxed feasible set*, the optimal value of Program 1.5 is the *continuous minimum*, and any optimal solution of Program 1.5 is a *continuous minimizer*. The optimal value of Program 1.4 is the *integer minimum*, and any optimal solution of Program 1.4 is an *integer minimizer*. The notion of a mixed-integer minimum and minimizer is defined analogously to the notion of the integer minimum and minimizer.

Two optimization problems are *equivalent* if feasible and optimal solutions coincide (but the objective need not be the same). Finally, if

$$\begin{array}{ll} \min & f_1(x) \\ \text{s.t.} & x \in S_1' \end{array} \qquad \begin{array}{ll} \min & f_2(x) \\ \text{s.t.} & x \in S_{12}' \end{array}$$

[2] In accordance with the traditional notation in optimization, we write max or min (instead of the formally more correct notation sup or inf) next to an optimization problem, even if it might not be clear if the maximum or minimum is attained. Note that, however, in this work *we shall not implicitly assume that the supremum (or infimum) is attained.* Existence of optimal solutions will either be explicitly assumed or proved.

are two optimization problems OPT1, OPT2, with $f_1 : S_1 \to \mathbb{R}$, $f_2 : S_1 \times S_2 \to \mathbb{R}$, respectively, and $S_1' \subset S_1$, and $S_{12}' \subset S_1 \times S_2$, we say that program OPT1 is a *projection* of program OPT2 (and OPT2 a *lift* of program OPT1) if, firstly, every feasible solution (x_1, x_2) of OPT2 yields a feasible solution x_1 of OPT1, and similarly, for every feasible solution $x_1 \in S_1'$ of OPT1 exists $x_2 \in S_2'$ such that (x_1, x_2) is feasible for OPT2, and, secondly, the same holds true for optimal solutions of both programs.

We need the notion of a *sublevel set*: For a function $f : U \to \mathbb{R}$ from some set U, the *sublevel set* of *level* $z \in \mathbb{R}$ is defined by

$$\mathcal{L}_{\leq}^{f}(z) = \{x \in U : f(x) \leq z\}.$$

Similarly, a *suplevel set* of level z is defined, with "$\leq$" replaced by "$\geq$". A *level set* of level z is the intersection of the sub- and suplevel set (of level z).

1.4.8. Coercivity

A function $f : S \to \mathbb{R}$, defined on a subset $S \subset \mathbb{R}^n$, is *coercive* if for all $c \in \mathbb{R}$ there exists some $M \in \mathbb{R}$ such that for all $x \in S$ the implication

$$\|x\| \geq M \Rightarrow f(x) \geq c \tag{1.6}$$

holds. Since all norms are equivalent on $\mathbb{R}^n$, this does not depend on the choice of norm. Let us recall the following well-known fact: A function is coercive function if and only if all sublevel sets are bounded.[3]

Proposition 1.4 (see, e.g., Chapter 12.3 in [Lan13])**.** *Let $S \subset \mathbb{R}^n$. Then $f : S \to \mathbb{R}$ is coercive if and only if every sublevel set $\mathcal{L}_{\leq}^{f}(z)$, $z \in \mathbb{R}$, is bounded.*

Proof. If f is coercive, every sublevel set is bounded by (1.6). Suppose now that all sublevel sets are bounded. Let $s = \liminf_{|x|\to+\infty} f(x)$. Suppose to the contrary that $s \in [-\infty, +\infty)$, and pick $z \in (s, +\infty)$. There must be a sequence $x_k \in \mathbb{R}^n$, $\|x_k\|_2 \to \infty$ as $k \to \infty$, such that $f(x_k) \leq z$ for all k. Put differently, $x_k \in \mathcal{L}_{\leq}^{f}(z)$ for all k, hence the sublevel set is unbounded, a contradiction. □

1.4.9. Matrices

Let R be a ring with unit. The set of $m \times n$ matrices over R, $m, n \in \mathbb{N}$, is denoted by $R^{m\times n}$. The $n \times n$-unit matrix over R is denoted by I_n, where we may drop the subscript n if no confusion seems possible. A matrix $A \in R^{m\times n}$ is *square* if $m = n$. For square matrices $A_1, \ldots, A_m$ with $A_i \in R^{n_i \times n_i}$, let $\operatorname{diag}(A_1, \ldots, A_m)$ denote the $(\sum_i n_i) \times (\sum_i n_i)$-block diagonal matrix arising from the A_i. A matrix is *diagonal* if it is block-diagonal with $n_i = 1$ for all i.

[3] In the literature, the statement usually requires some form of continuity, and results in the function being coercive if and only if all sublevel sets are *compact*. However, the proof is the same, which we give for completeness.

A square matrix is *symmetric* if $A = A^T$, where A^T is the transposed of A. We denote the set of real symmetric $n \times n$-matrices by $\mathcal{S}^n$. A real $n \times n$-square matrix V is *orthogonal* if $V^T V = I_n$. The famous spectral theorem holds for symmetric matrices:

Theorem 1.5 (Spectral theorem, see, e.g., Corollary 2.5.11 in [HJ12]). *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Then A has n real eigenvalues and is unitarily diagonalizable: A has a spectral decomposition*

$$A = VDV^T$$

with V orthogonal and D diagonal.

Proof. □

By the spectral theorem, any real symmetric matrix $A \in \mathcal{S}^n$ has n real eigenvalues $\lambda_i(A)$, $i \in [n]$, which we may assume to be nondecreasing, and we write

$$\lambda_{\min}(A) := \lambda_n(A) \leq \lambda_{n-1}(A) \leq \ldots \leq \lambda_2(A) \leq \lambda_1(A) =: \lambda_{\max}(A).$$

Using the spectral decomposition, we may estimate the associated quadratic form. This result is the Rayleigh-Ritz theorem, we give the version for real matrices.

Theorem 1.6 (Rayleigh-Ritz, see, e.g., Theorem 4.2.2 in [HJ12]). *Let $A \in \mathcal{S}^n$. Then*

$$\lambda_{\min}(A)x^T x \leq x^T Ax \leq \lambda_{\max}(A)x^T x, \quad x \in \mathbb{R}^n.$$

Furthermore,

$$\lambda_{\min}(A) = \min_{x\in\mathbb{S}_2^{n-1}} x^T A, \quad \lambda_{\max}(A) = \max_{x\in\mathbb{S}_2^{n-1}} x^T Ax.$$

A matrix $A \in \mathcal{S}^n$ is *positive semidefinite* if $x^T Ax \geq 0$ for all $x \in \mathbb{R}^n$, *positive definite* if $x^T Ax > 0$ for all nonzero $x \in \mathbb{R}^n$, and *indefinite* if there are $x, y \in \mathbb{R}^n$ with $(x^T Ax)(y^T Ay) < 0$. We abbreviate positive semidefiniteness by $A \succeq 0$ and denote the set of all such matrices by $\mathcal{S}^n_+$; similarly, we abbreviate positive definiteness by $A \succ 0$ and denote the set of all such matrices by $\mathcal{S}^n_{++}$. Positive semidefinite matrices play a central role in semidefinite programming, which we consider in more detail in Section 1.5.

The following characterization of positive semidefinite matrices is useful for our purposes:

Proposition 1.7 (see, e.g., Proposition A.1 in [BPT13]). *Let $A \in \mathcal{S}^n$ be a symmetric matrix. Then, the following are equivalent:*

1. *A is positive semidefinite ($A \succeq 0$).*
2. *For all $x \in \mathbb{R}^n$, $x^T Ax \geq 0$.*
3. *All eigenvalues of A are nonnegative.*
4. *There exists a factorization $A = BB^T$ with $B \in \mathbb{R}^{n\times r}$ and r is the rank of A.*

There is a similar characterization for positive definite matrices:

Proposition 1.8 (see, e.g., Proposition A.2 in [BPT13]). *Let $A \in \mathcal{S}^n$ be a symmetric matrix. Then, the following are equivalent:*

1. *A is positive definite ($A \succ 0$).*
2. *For all nonzero $x \in \mathbb{R}^n$, $x^T Ax > 0$.*
3. *All eigenvalues of A are positive.*
4. *There exists a factorization $A = BB^T$ with $B \in \mathbb{R}^{n \times n}$ nonsingular.*

Further characterizations are available in the given reference.

The notion of positive semidefiniteness induces the so-called *Loewner partial order* on $\mathcal{S}^n$: For $A, B \in \mathcal{S}^n$, we write $A \succeq B$ if $A - B$ is positive semidefinite, and, for completeness, define $B \preceq A$ if and only if $A \succeq B$. The following fact is easy to see but useful in arguments involving semidefinite matrices:

Observation 1.9 (see, e.g., p. 6 in [WSV00]). *Let $A, B \in \mathcal{S}^n$, $C, D \in \mathcal{S}^m$. Then*

$$A \preceq B \text{ and } C \preceq D \quad \text{if and only if} \quad \operatorname{diag}(A, C) \preceq \operatorname{diag}(B, D).$$

Moreover, positive (semi-)definiteness is invariant under basis transformations:

Proposition 1.10. *Let $B \in \mathbb{R}^{n \times n}$ be invertible. Then*

$$A \in \mathcal{S}^n_+ \text{ if and only if } B^T AB \in \mathcal{S}^n_+.$$

Similarly,

$$A \in \mathcal{S}^n_{++} \text{ if and only if } B^T AB \in \mathcal{S}^n_{++}.$$

If $B \in \mathbb{R}^{n \times m}$, then

$$A \in \mathcal{S}^n_+ \text{ implies } B^T AB \in \mathcal{S}^m_+.$$

Proof. The equivalences are shown, e.g., in Proposition 1.1.7 in [Hel00]. For the last implication, the proof is identical: Let $y \in \mathbb{R}^m$. Then

$$y^T (B^T AB) y = (By)^T A (By) = x^T Ax \geq 0$$

for $x = By$. □

With the following observation, matrix inversion can be modeled in semidefinite programs:

Theorem 1.11 (Schur complement, see, e.g., Theorem 1.1.9 in [Hel00]). *Let $A \in \mathcal{S}^m_{++}$, $C \in \mathcal{S}^n$, $B \in \mathbb{R}^{m \times n}$. Then*

$$\begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \succ 0 \quad \Longleftrightarrow \quad C - B^T A^{-1} B \succ 0 \tag{1.7}$$

and

$$\begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \succeq 0 \quad \Longleftrightarrow \quad C - B^T A^{-1} B \succeq 0. \tag{1.8}$$

Finally, the space of matrices $\mathbb{R}^{n\times n}$ carries a natural inner product structure. Given $X, Y \in \mathbb{R}^{n\times n}$, the inner product of X and Y is given by

$$\langle X, Y\rangle := \operatorname{tr}(X^T Y) = \sum_{i,j=1}^{n} X_{ij} Y_{ij}.$$

With this inner product, the set of all real $n \times n$ matrices is linear-topologically isomorph to $\mathbb{R}^{n^2}$.

1.4.10. Half-spaces, valid inequalities and cuts

A *half-space* in $\mathbb{R}^n$ is a set of the form $\{x \in \mathbb{R}^n : a^T x \leq b\}$ for some $a \in \mathbb{R}^n$, $a \neq 0$, $b \in \mathbb{R}$. The corresponding *hyperplane* is denoted by

$$H(a, b) = \{x \in \mathbb{R}^n : a^T x = b\}.$$

We may refer to inequalities using the notation $(a^T x \leq b)$. So let an inequality $(a^T x \leq b)$ be given. We say the inequality is

- a *valid inequality* for $S \subset \mathbb{R}^n$ if it is satisfied for all $x \in S$, that is, $a^T x \leq b$ holds for all $x \in S$.
- *tight for* S if it is valid for S and for any $b' < b$, the inequality $(a^T x \leq b')$ is violated by some $x \in S$, where the inequality is
- *violated by* some $x \in S$ if $a^T x > b$.
- is *tight at* $q \in S$ if the inequality is tight for S and $a^T q = b$.
- a *cut for* S if it is a valid inequality for $S_{\mathcal{I}}$.

Abusing notation, we may also use $(a^T x \leq b)$ to denote the set $\{x \in \mathbb{R}^n : a^T x \leq b\}$.

Similarly, if $V : \mathbb{R}^n \to \mathbb{R}$ is a function, $b \in \mathbb{R}$, we say the inequality $(V(x) \leq b)$ is a *valid inequality for* $S \subset \mathbb{R}^n$ if $V(x) \leq b$ for all $x \in S$. The notions *tight for* S, *tight at* $q \in S$, *violated by* $x \in S$ and *cut for* S are defined analogously to the case of a linear function $\mathbb{R}^n \to \mathbb{R}$, $x \mapsto a^T x$.

We use the attributes *linear* and *nonlinear* if we wish to stress the linearity or non-linearity of the function defining the inequality.

1.4.11. Convexity, affine dimension and cones

Turning to convexity, a set $C \subset \mathbb{R}^n$ is *convex* if for all $x, y \in C$, $\lambda \in [0, 1]$, the point $\lambda x + (1 - \lambda)y$ lies in C. The convex hull $\operatorname{conv}(M)$ of $M \subset \mathbb{R}^n$ is the intersection of all convex sets containing M and hence the smallest (with respect to set inclusion) convex set that contains M.

Theorem 1.12 (see, e.g., Corollary 11.5.1 in [Roc70]). *Let $A \subset \mathbb{R}^n$. Then*

$$\operatorname{cl}\operatorname{conv} A = \bigcap \{V : V \text{ is a half-space of } \mathbb{R}^n,\ A \subset V\}.$$

In particular, the closure of a convex set is convex.

Disjoint convex sets can be separated by a hyperplane:

Theorem 1.13 (Separating Hyperplane, see, e.g., Theorem 4.4 in [Gru07]). *Let $C_1, C_2 \subset \mathbb{R}^n$ be disjoint convex sets. Then there is $a \in \mathbb{R}^n \setminus \{0\}$, $b \in \mathbb{R}$ with*

$$C_1 \subset (a^T x \leq b) \quad \text{and} \quad C_2 \subset (a^T x \geq b).$$

In this case, the hyperplane $H(a, b)$ is called a *separating hyperplane*.

Also, points on the boundary of a convex set have a special exposure property:

Theorem 1.14 (Supporting Hyperplane, see, e.g., Chapter 2.5.2, p. 51 in [BV04]). *Let $C \subset \mathbb{R}^n$ be a convex set. Then, for every $x_0 \in \operatorname{bd} C$ there is $a \in \mathbb{R}^n \setminus \{0\}$ with $C \subset (a^T x \leq a^T x_0)$.*

Then, the hyperplane $H(a, a^T x_0)$ is called a *supporting hyperplane to C at x_0*.

Now let $f : C \to \mathbb{R}$ be a function, where $C \subset \mathbb{R}^n$ is a convex set. Then f is *convex* if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y), \quad x, y \in C,\ \lambda \in [0, 1],$$

and f is *quasiconvex* if all sublevel sets of f are convex.

The affine hull $\operatorname{aff}(M)$ of a set $M \subset \mathbb{R}^n$ is the intersection of all affine subspaces containing M and hence the smallest (with respect to set inclusion) affine subspace containing M. The *affine dimension* of $M \subset \mathbb{R}^n$ is the dimension of its affine hull.

A set $K \subset \mathbb{R}^n$ is a *cone* if for all $a \in K$, $\lambda > 0$, the point λa lies in K.

1.4.12. Polyhedra and spectrahedra

A *polyhedron* is a finite intersection of closed half-spaces in some $\mathbb{R}^n$. If $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, we write

$$P(A, b) := \{x \in \mathbb{R}^n : Ax \leq b\}$$

for the polyhedron generated by A and b as well as

$$P_+(A, b) := \{x \in \mathbb{R}^n_{\geq 0} : Ax \leq b\}$$

for the polyhedron generated by A and b with nonnegativity constraints on x.

A *polytope* is a polyhedron that is also bounded. The *integer hull* of a polyhedron $P \subset \mathbb{R}^n$ is the set $\operatorname{conv}(P \cap \mathbb{Z}^n)$ and is denoted by P_I.

Finally, we turn to geometric objects that arise as solution sets of *linear matrix inequalities*: A set $M \subset \mathbb{R}^n$ is a *spectrahedron* if it can be described as

$$M = \{x \in \mathbb{R}^n : A_0 + x_1 A_1 + \ldots + x_n A_n \succeq 0\}$$

for $A_0, \ldots, A_n \in \mathcal{S}^m$ and some $m \in \mathbb{N}$. More generally, a set $M \subset \mathbb{R}^n$ is a *projected spectrahedron* if there exists $k \in \mathbb{N}$ and a spectrahedron $P \subset \mathbb{R}^{n+k}$ with

$$M = \left\{x \in \mathbb{R}^n : (x, y) \in P \text{ for some } y \in \mathbb{R}^k\right\}.$$

In this equation, y is a *lifting vector* and P a *lifting spectrahedron* of S. Equivalently, M is a projected spectrahedron if there are $k, m \in \mathbb{N}$ and

$$M = \left\{x \in \mathbb{R}^n : A_0 + \sum_{i=1}^{n} x_i A_i + \sum_{i=1}^{k} y_j B_j \succeq 0 \text{ for some } y \in \mathbb{R}^k\right\}$$

holds for some $A_i, B_i \in \mathcal{S}^m$. Our definition follows Chapter 6 in [BPT13], and we also point to this reference for more theory on as well as many examples of spectrahedra and projected spectrahedra.

1.4.13. Ellipsoids

If $Q \in \mathcal{S}^n$, $Q \succeq 0$, $x_0 \in \mathbb{R}^n$, we denote the *ellipsoid* corresponding to Q centered at x_0 by

$$E(Q, x_0) := \{x \in \mathbb{R}^n : (x - x_0)^T Q (x - x_0) \leq 1\}.$$

If $Q \not\succ 0$, $E(Q, x_0)$ is *degenerated*. In case $x_0 = 0$, we may write $E(Q)$ instead of $E(Q, 0)$. We use the relation

$$rE(Q) = E\left(\frac{1}{r^2} Q\right) \tag{1.9}$$

for all $Q \succeq 0$ and $r > 0$, which follows directly from the definition.

The volume (Lebesgue measure) of an ellipsoid $E(Q, x_0)$ is proportional to the root of the determinant of the inverse of Q, where $\mathrm{vol}(A)$ denotes the volume of a Lebesgue measurable set $A \subset \mathbb{R}^n$:

Observation 1.15. *Let $Q \succeq 0$, $x_0 \in \mathbb{R}^n$. Then*

$$\mathrm{vol}(E(Q, x_0)) = \frac{\mathrm{vol}(B_n)}{\sqrt{\det(Q)}},$$

where B_n is the n-dimensional unit ball $B_n := \{x \in \mathbb{R}^n : \|x\| \leq 1\}$.

Sketch of a proof. This fact is mathematical folklore. For completeness, let us remark that in case $x_0 = 0$ and Q is non-degenerated, the volume is given, e.g., in [Han96]. Note that, by translation invariance of the Lebesgue measure, the formula holds for any $x_0 \in \mathbb{R}^n$. Let us verify the degenerated case $Q \succeq 0$ but not $Q \succ 0$. Let k be the

nullity of Q. Under an orthogonal transformation, $E(Q)$ takes the form $E(Q') \times \mathbb{R}^k$ where $Q' \in \mathcal{S}_{++}^{n-k}$. Now, the fact that the Lebesgue measure on $\mathbb{R}^n$ is invariant under orthogonal transformations and moreover a product measure, we find

$$\mathrm{vol}(E(Q)) = \mathrm{vol}(E(Q') \times \mathbb{R}^k) = \mathrm{vol}(E(Q')) \cdot \mathrm{vol}(\mathbb{R}^k) = +\infty$$

as $\mathrm{vol}(E(Q')) > 0$ by $Q' \succ 0$. □

1.5. Sum of squares programming

1.5.1. Nonnegativity, sums of squares and tractability

Throughout this work, continuous polynomial programming problems appear – as a relaxation of the original problem, as a subproblem, or as another auxiliary problem. By a continuous polynomial optimization problem, we mean a program of the form

$$\begin{aligned} \min \quad & p(x) \\ \text{s.t.} \quad & h_k(x) \geq 0, \qquad k \in [s], \\ & x \in \mathbb{R}^n \end{aligned} \tag{CPOP}$$

where $p, h_1, \ldots, h_s \in \mathbb{R}[X_1, \ldots, X_n]$.

It is therefore highly desirable to have a method for these problems that has a solid, well-known theory but also works in practice, at least for moderately sized problems. One of the most successful methods can be traced back to the seminal works in the early 2000s of Lasserre [Las01] and Parrilo [Par00]: This method is known as *sos programming*, where sos abbreviates *sum of squares*. Let us motivate the method. A polynomial $p \in \mathbb{R}[\underline{X}]$ is a *sum of squares* or *sos* for short, if it has a representation as a sum of squared polynomials. Formally, $p \in \mathbb{R}[\underline{X}]$ is sos if there are $q_1, \ldots, q_l \in \mathbb{R}[\underline{X}]$ with

$$p = q_1^2 + \ldots + q_l^2. \tag{1.10}$$

What makes this notion so useful? An immediate consequence of a representation of p as in (1.10) is that that p is nonnegative, and the q_i *certify* nonnegativity. It turns out that deciding if a polynomial is a sum of squares can be reformulated as a semidefinite program, and semidefinite programs in turn are well-understood and can be solved efficiently, see, e.g., [WSV00; VB96].

But even more is possible: It can be shown that the considerably more powerful problem of optimizing a linear form over the cone of all sos polynomials (with a bound on the degree) can be rewritten as a semidefinite program. This extension allows to approximate polynomial programming problems with a hierarchy of sos programs. Under additional assumptions, finite convergence holds. We sketch below how semidefinite and sos programming are related.

We should, however, stress that not every nonnegative polynomial is a sum of squares – this holds for almost all number of variables and degrees, except in the three cases outlined in the next theorem, where $P_{n,d}$ denotes the set of nonnegative polynomials in n unknowns of degree at most d, and similarly $\Sigma_{n,d}$ the set of sos polynomials for n and d. The proof is due to David Hilbert. Let us note that it was not until the 1960s that an explicit example of a polynomial $p \in P_{n,d} \setminus \Sigma_{n,d}$ was found [Mot67].[4]

Theorem 1.16 (see, e.g., p. 59 in [BPT13])**.** *Let $n \in \mathbb{N}$ and $d \in 2\mathbb{N}$. We have*

$$P_{n,d} = \Sigma_{n,d}$$

[4]The *Motzkin polynomial* is $p = 1 - 3X^2Y^2 + X^2Y^4 + X^4Y^2$. This degree six polynomial in two unknowns is nonnegative on all of $\mathbb{R}^2$, yet cannot be represented as a sum of squares.

if and only if (n, d) *satisfies*

1. $n = 1$, *d arbitrary, or*
2. $d = 2$, *n arbitrary, or*
3. $(n, d) = (2, 4)$.

In contrast to deciding if a polynomial is a sum of squares, deciding nonnegativity of a given polynomial is a NP hard problem, even if one fixes the degree to $d = 4$ (Theorem 1.38). However, being able to decide if a polynomial is globally nonnegative is important in this work: For example, in Chapter 6 underestimators play a crucial role. A function f is (globally) underestimated by g if $f(x) \geq g(x)$ holds for all $x \in \mathbb{R}^n$. Hence, deciding if f is underestimated by g means deciding nonnegativity of $f - g$, which we approximate with a sufficient criterion by searching for an sos decomposition of $f - g$.

More generally, as deciding nonnegativity is NP hard, continuous minimization of a polynomial must be NP hard, too. Therefore, if one wants to solve a polynomial programming problem, it is common practice to not solve the problem directly but to approximate it with an *sos program*. It is the aim of this section to introduce sos programming in its classical form.

1.5.2. Optimization over the cone of sos polynomials

We saw in the introduction of this section that a polynomial $p \in \mathbb{R}[\underline{X}]$ is sos if there are $q_1, \ldots, q_l \in \mathbb{R}[\underline{X}]$ such that $p = q_1^2 + \ldots + q_l^2$. The set of all sos polynomials is a convex cone in $\mathbb{R}[X_1, \ldots, X_n]$ which we denote by

$$\Sigma_n := \left\{ p \in \mathbb{R}[X_1, \ldots, X_n] : \exists q_1, \ldots, q_l \in \mathbb{R}[X_1, \ldots, X_n] \text{ s.t. } p = \sum_{i=1}^{l} q_i^2 \right\},$$

where we may write Σ instead of Σ_n if n is known by the context. Recall that $\Sigma_{n,d}$ is the set $\{p \in \Sigma_n : \deg(p) \leq d\}$.

It is possible to optimize a linear form over the cone Σ_n subject to affine constraints, and this is *sos programming*. Specifically, for given costs $b \in \mathbb{R}^m$ as well as fixed polynomials $c_i, a_{ij} \in \mathbb{R}[X_1, \ldots, X_n]$, $i \in [k]$, $j \in [m]$, an sos program has the form

$$\begin{aligned} \max \quad & b^T y \\ \text{s.t.} \quad & c_i + y_1 a_{i1} + \ldots + y_m a_{im} \in \Sigma_n, \quad i = 1, \ldots, k, \\ & y \in \mathbb{R}^m \end{aligned} \tag{SOSP}$$

and $y \in \mathbb{R}^m$ are the decision variables. As mentioned before, such an *sos optimization problem* or *sos program* is tractable, as it is equivalent to a semidefinite program. For a detailed introduction to sos programming, we refer to [AL12; BPT13].

Before we give a generalization of sos programming that suits our needs, we show how sos programming can be used to approximate continuous polynomial optimization

problems. It turns out that central to this is the ability to represent a polynomial that is positive (or, in other variants, nonnegative) on a set given by polynomial constraints as a sum of the constraint polynomials, if one allows to scale the latter by sos polynomials, or more generally, if one allows cross-multiplication of the constraint polynomials. Statements that give sufficient (or sometimes equivalent) conditions that guarantee the existence of such a representation are known in the literature on real algebraic geometry as *Stellensätze*.[5]

1.5.3. Putinar's Positivstellensatz

In this section we introduce Putinar's Positivstellensatz. It holds under a technical condition (a related quadratic module is required to be Archimedean). We explore when this condition holds in our setting before giving the Stellensatz. Our notation follows [Mar08] and [NS07].

Semialgebraic sets and Archimedean quadratic modules

Given a finite collection of multivariate polynomials $h_1, \ldots, h_s \in \mathbb{R}[X_1, \ldots, X_n]$, consider the subset of $\mathbb{R}^n$ where all polynomials h_i attain nonnegative values:

$$K(h_1, \ldots, h_s) := \{x \in \mathbb{R}^n : h_1(x) \geq 0, \ldots, h_s(x) \geq 0\}. \tag{1.11}$$

A subset of $\mathbb{R}^n$ is called *basic closed semi-algebraic* if it is of the form (1.11) for some polynomials $h_1, \ldots, h_s$. The Stellensatz we consider gives a sufficient condition which allows to represent every polynomial $p \in \mathbb{R}[\underline{X}]$ that is positive on $K(h_1, \ldots, h_s)$ as a combination of the h_i and 1 – each multiplied by a sum of squares. The set of these combinations is the quadratic module generated by the h_i (Definition 1.2) and is thus given by

$$M(h_1, \ldots, h_s) := \left\{ \sum_{i=0}^{s} \sigma_i h_i : \sigma_0, \ldots, \sigma_s \in \Sigma \right\} \tag{1.12}$$

where $h_0 := 1$.

For the Positivstellensatz to hold we need to impose a technical condition on the quadratic module $M(h_1, \ldots, h_s)$, namely, the quadratic module needs to be Archimedean (we introduced the algebraic definition of the Archimedean property in Definition 1.2). This is the case if any of the following equivalent conditions hold.

Theorem 1.17 (see, e.g., Corollary 5.2.4 in [Mar08] and Corollary 3 in [NS07])**.** *Let $h_1, \ldots, h_s \in \mathbb{R}[X_1, \ldots, X_n]$. Then, for the quadratic module $M = M(h_1, \ldots, h_s)$, the following are equivalent:*

1. *M is Archimedean, that is, for all p in $\mathbb{R}[\underline{X}]$ exists $k \in \mathbb{N}$ with $p + k \in M$.*

[5] A *Stellensatz* (plural *Stellensätze*) from the German words *Stelle*, meaning: argument of a function, and *Satz*, meaning: theorem. A *Positivstellensatz* is a theorem on the arguments where a function is positive.

2. *There is a number* $k \in \mathbb{N}$ *such that* $k - \sum_{i=1}^{n} X_i^2 \in M$.

3. *There is* $k \in \mathbb{N}$ *with* $k \pm X_i \in M$ *for* $i \in [n]$.

4. *There is a polynomial* $h \in M$ *such that* $K(h)$ *is compact.*

We explore how the Archimedean property, a prerequisite for the Stellensatz, is related to geometric and analytic properties of MIPP at the end of this section.

The Stellensatz

Suppose a polynomial f is positive (or nonnegative) on

$$K(h_1, \ldots, h_s) = \{x \in \mathbb{R}^n : h_1(x) \geq 0, \ldots, h_s(x) \geq 0\}.$$

Can it then be written in terms of the defining inequalities $h_i(x) \geq 0$ of $K(h_1, \ldots, h_s)$? Conditions that guarantee such representations are addressed in Positivstellensätzen (Nichtnegativstellensätzen, respectively). The "converse direction", a geometric conclusion from an algebraic fact, is usually much easier to show. We give a well-known example in the following observation. It states that if a polynomial is written in terms of the defining inequalities (in the forms "allowed" by quadratic modules), it is nonnegative on $K(h_1, \ldots, h_s)$. The proof is immediate from the definition.

Observation 1.18. *Let* $h_1, \ldots, h_s \in \mathbb{R}[X_1, \ldots, X_n]$ *and* $p \in M(h_1, \ldots, h_s)$. *Then* $p \geq 0$ *on* $K(h_1, \ldots, h_s)$.

A straightforward, well-known but useful conclusion is that the basic closed semialgebraic set associated with the polynomials h_i is compact provided the quadratic module they generate is Archimedean.

Corollary 1.19. *Let* $h_1, \ldots, h_s \in \mathbb{R}[X_1, \ldots, X_n]$ *and* $M(h_1, \ldots, h_s)$ *be Archimedean. Then* $K(h_1, \ldots, h_s)$ *is compact.*

Proof. This follows from Observation 1.18 and Theorem 1.17 (2). □

The Positivstellensatz – a (real) algebraic statement implied by a geometric condition – is much more difficult to prove. Note that the theorem requires positivity, a stronger requirement than nonnegativity.[6]

Theorem 1.20 (Putinar's Positivstellensatz, see, e.g., Corollary 6.1.2 in [Mar08] and [NS07])**.** *Let* p, $h_1, \ldots, h_s \in \mathbb{R}[X_1, \ldots, X_n]$ *be given. Furthermore, let the quadratic module* $M(h_1, \ldots, h_s)$ *generated by the* h_i *be Archimedean. Then* $p(x) > 0$ *for all* $x \in K(h_1, \ldots, h_s)$ *implies* $p \in M(h_1, \ldots, h_s)$.

[6] For an overview on Nichtnegativstellensätze, we refer to [Mar08].

In the application of the Stellensatz in sos programs, the σ_i appearing in (1.12) are unknowns that we optimize. As there is no degree bound on the σ_i, this is impractical. Hence, we instead use the *truncated quadratic module of order* $k \in \mathbb{N} \cup \{-\infty\}$, given by

$$M(h_1, \ldots, h_s)[k] := \left\{ \sum_{i=0}^{s} \sigma_i h_i : \sigma_i \in \Sigma,\ \deg(\sigma_i h_i) \leq k,\ i = 1, \ldots, s \right\} \tag{1.13}$$

where, again, $h_0 := 1$.

The Archimedean property in terms of MIPP

The Archimedean property for a quadratic module associated with MIPP can be enforced if the relaxed feasible set F is contained in a 2-norm ball of known radius[7] $R \geq 0$ by adding the redundant constraint $\|x\|_2^2 \leq R^2$ to MIPP.

Proposition 1.21 (see, e.g., [JLL14])**.** *Consider* MIPP*. Suppose F is contained in $B_R(0; \|\cdot\|_2)$ for some $R \geq 0$. Put*

$$g_{s+1} := R^2 - \sum_{i=1}^{n} X_i^2.$$

Then, the quadratic module $M(g_1, \ldots, g_{s+1})$ is Archimedean, and

$$F = K(g_1, \ldots, g_s) = K(g_1, \ldots, g_{s+1}).$$

The task to find R such that F is contained in $B_R(0, \|\cdot\|_2)$ can be approximated with sos programming as outlined in [JLL14]. Let us now explore which geometrical, topological and analytical conditions on MIPP ensure that the Archimedean property for a related quadratic module, an algebraic statement, holds. The results are well-known or at least easy consequences from well-known results, but are nevertheless important in this work.

Proposition 1.22. *Consider* MIPP*. The quadratic module $M := M(g_1, \ldots, g_r)$ is Archimedean if*

$$K(g_i) = \{x \in \mathbb{R}^n : g_i(x) \geq 0\}$$

is compact for some $i \in [s]$. This holds if $-g_i$ is coercive.

Proof. If $K(g_i)$ is compact for some i, then $M(g_1, \ldots, g_r)$ is Archimedean by Theorem 1.17 (4). If $-g_i$ is coercive, all of its sublevel sets are compact by Proposition 1.4, especially $\mathcal{L}_{\leq}^{-g_i}(0) = K(g_i)$. □

If a feasible solution $q \in F_{\mathcal{I}}$ is known, the optimal solutions do not change by adding the constraint $f(x) \leq f(q)$. This motivates the following variant, where we only assume knowledge of an upper bound on a feasible objective value.

[7] In theory, this is obviously equivalent to F being compact. However, to compute R in case of a compact relaxed feasible set F is not straightforward.

Proposition 1.23. *Consider* MIPP. *Let $z \in \mathbb{R}$ with $z \geq f(q)$ for some $q \in F_{\mathcal{I}}$. The quadratic module $M' := (g_1, \ldots, g_r, z - f)$ is Archimedean if $M(g_1, \ldots, g_r)$ or $M(z - f)$ is Archimedean. $M(z - f)$ is Archimedean in turn if and only if $\mathcal{L}^f_{\leq}(z)$ is compact. This holds if f is coercive.*

Proof. Let $M(g_1, \ldots, g_r)$ be Archimedean. By (1.12), we know that $M(g_1, \ldots, g_r) \subset M(g_1, \ldots, g_s, z - f)$, and the Archimedean property for the larger quadratic module follows directly from Definition 1.2. The proof that $M(g_1, \ldots, g_r, z - f)$ is Archimedean provided $M(z - f)$ is Archimedean follows similarly.

Now let $M(z - f)$ be Archimedean. By Corollary 1.19, $K(z - f) = \mathcal{L}^f_{\leq}(z)$ is compact. The remaining arguments to finish the proof are verbatim the same as for Proposition 1.22. □

Proposition 1.22 says that we get convergence if there is a single constraint with compact suplevel set. In case that MIPP has equality constraints[8], this can be generalized to the requirement of only an intersection of sublevel sets being compact.

Proposition 1.24. *Suppose some of the constraints in* MIPP *are equality constraints, the first r', say. Suppose further that the constraint set they generate, i.e.,*

$$K' := \{x \in \mathbb{R}^n : g_i(x) = 0 \ \forall i \in [r']\}$$

is compact. Then $M(\tilde{g}, g_{r'+1}, \ldots, g_r)$ is Archimedean, where $\tilde{g} := \sum_{i=1}^{r'} -g_i^2$.

Proof. An equality constraint $g_i(x) = 0$ is equivalent to $g_i(x) \geq 0$ and $-g_i(x) \geq 0$. Note that

$$K' = \{x \in \mathbb{R}^n : \tilde{g}(x) \geq 0\},$$

which is compact by assumption. The claim follows by Theorem 1.17 (4). □

After having illustrated the Archimedean property – the assumption in Putinar's Positivstellensatz – for MIPP, we turn now to the statement itself.

1.5.4. A note on model building in sos programming

In the following, we recall how common types of constraints can be remodeled as classical sos programming constraints.

Sos variables. Constraints of the form

$$\begin{aligned} \sigma_1 a_1 + \ldots + \sigma_m a_m &\in \Sigma_n \\ \deg \sigma_i &\leq k, \qquad i \in [m] \\ \sigma_i &\in \Sigma_n, \qquad i \in [m] \end{aligned} \tag{1.14}$$

for some $a_i \in \mathbb{R}[X_1, \ldots, X_n]$ and $k \in \mathbb{N}$. The decision variables, sos polynomials with a degree bound, translate directly to classical sos programming constraints of the form SOSP since the σ_i have a bound on the degree and are thus fully parameterized by finitely many scalar decision variables.

[8] This is just to ease notation: Every equality constraint can be modeled by two inequality constraints.

Polynomial variables. Constraints of the form

$$\begin{aligned} p_1 a_1 + \ldots + p_m a_m &\in \Sigma_n \\ \deg p_i &\leq k, \qquad i \in [m] \\ p_i &\in \mathbb{R}[X_1, \ldots, X_n], \qquad i \in [m] \end{aligned} \tag{1.15}$$

for some $a_i \in \mathbb{R}[X_1, \ldots, X_n]$ and $k \in \mathbb{N}$. The decision variables, polynomials with a degree bound, translate to classical sos programming by the same argument as for sos variables.

Truncated quadratic module containment. Constraints of the form

$$\begin{aligned} c + y_1 a_1 + \ldots + y_m a_m &\in M(h_1, \ldots, h_s)[k] \\ y &\in \mathbb{R}^m \end{aligned} \tag{1.16}$$

for some $a_i, h_i \in \mathbb{R}[X_1, \ldots, X_n]$ and $k \in \mathbb{N}$. This translates to a classical sos programming constraints as follows: The statement

$$c + y_1 a_1 \ldots + y_m a_m \in M(h_1, \ldots, h_s)[k]$$

is equivalent to

$$\begin{aligned} c + y_1 a_1 \ldots + y_m a_m - \sum_{j=1} \sigma_j h_j &\in \Sigma_n \\ \deg \sigma_i &\leq k, \quad i \in [s] \\ \sigma_i &\in \Sigma_n, \quad i \in [s] \end{aligned}$$

and is thus a constraint with sos variables as introduced in (1.14).

Linear programming constraints. Constraints of the form

$$\begin{aligned} Ay &\leq b \\ y &\in \mathbb{R}^m \end{aligned} \tag{1.17}$$

for a real matrix A and real vector b. The requirement $u \leq v$ for real numbers u, v is equivalent to $v - u \in \Sigma_n$, as every nonnegative real number is a square. Note that, as sos solvers reformulate an sos program into a semidefinite one, this can be done more efficiently in practice.

Semidefinite constraints. Constraints of the form

$$\begin{aligned} Q &\in \mathcal{S}^n \\ Q &\succeq 0, \end{aligned}$$

that is, a positive definite decision variable, can be modeled using classical sos constraints. The matrix is fully parameterized by finitely many scalars. Symmetry of Q can be enforced by a system of linear equations in the entries of Q, moreover, $Q \succeq 0$ is equivalent to $(X_1, \ldots, X_n)^T Q (X_1, \ldots, X_n) \in \Sigma_n$ by Theorem 1.16. Again we note that this can be done more efficiently in practice.

Combinations of the above. It goes without saying that said constraints can be combined. For example, a constraint of the form

$$\begin{aligned} c + \sum_{j=1}^{m} y_j a_{0j} + \sum_{j=1}^{m_1} \sigma_j a_{1j} + \sum_{j=1}^{m_2} p_j a_{2j} \in M(h_1, \ldots, h_s)[k] \\ y \in \mathbb{R}^m, \quad \deg \sigma_j \leq k',\ \sigma_j \in \Sigma_n, \quad \deg p_j \leq k'',\ p \in \mathbb{R}[X_1, \ldots, X_n] \end{aligned}$$

where $c, a_{ij} \in \mathbb{R}[X_1, \ldots, X_n]$ and $k, k', k'' \in \mathbb{N}$, combining scalar decision variables y, sos decision polynomial variables σ_j (of bounded degree) and polynomial decision variable p_j (of bounded degree) can be modeled in sos programming.

1.5.5. Lower bounds for continuous polynomial optimization problems

We have already remarked that throughout this work we use sos programming to approximate continuous polynomial optimization problems of the form CPOP, that arise, for example, as relaxations of mixed-integer polynomial problems. As these problems are hard, we use sos programming to compute lower bounds on CPOP.

In the following we describe how lower bounds on

$$\begin{aligned} \min \quad & p(x) \\ \text{s.t.} \quad & x \in K(h_1, \ldots, h_s), \end{aligned} \tag{1.18}$$

for $p, h_1, \ldots, h_s \in \mathbb{R}[\underline{X}]$ can be derived by sos programming. Note that (1.18) is just a reformulation of CPOP, since by the defining equation (1.11),

$$K(h_1, \ldots, h_s) = \{x \in \mathbb{R}^n : h_1(x) \geq 0, \ldots, h_s(x) \geq 0\},$$

or put differently, the feasible set of CPOP is basic closed semi-algebraic.

The method we outline follows Schweighofer [Sch05], based on Lasserre's [Las01] work. Consider the hierarchy Q_k, $k = 1, 2, \ldots$, of sos programs

$$\begin{aligned} \max \quad & y & & \\ \text{s.t.} \quad & p - y - \sum_{i=1}^{s} \sigma_i h_i \in \Sigma & & \\ & \deg(\sigma_i h_i) \leq k, & & i = 1, \ldots, s \\ & \sigma_i \in \Sigma, & & i = 1, \ldots, s \\ & y \in \mathbb{R}. & & \end{aligned} \tag{Q_k}$$

In Q_k, the decision variables are $y \in \mathbb{R}$ and the real coefficients of $\sigma_1, \ldots, \sigma_s \in \mathbb{R}[\underline{X}]$. We then have the following result:

Proposition 1.25 (see, e.g., Chapter 3 in [BPT13])**.** *Every feasible solution y to Q_k gives a lower bound on (1.18).*

Proof. Let y be feasible. Then, there are $\sigma_0, \ldots, \sigma_s \in \Sigma$, $\deg(\sigma_i h_i) \leq k$ for $i = 1, \ldots, s$, such that

$$\begin{aligned} & p - y - \sum_{i=1}^{s} \sigma_i h_i = \sigma_0 \\ & \quad \Longrightarrow \; p(x) = y + \sigma_0(x) + \sum_{i=1}^{s} \sigma_i(x) h_i(x) \geq y, \quad x \in K(h_1, \ldots, h_s), \end{aligned}$$

as $\sigma_i \in \Sigma$, hence σ_i are nonnegative, and $h_i(x) \geq 0$ on $K(h_1, \ldots, h_s)$ by definition. Hence p is bounded from below by y on $K(h_1, \ldots, h_s)$, i.e., every feasible solution to Q_k is a lower bound on (1.18). □

Remark 1.26. In view of (1.16), the hierarchy Q_k can be formulated more compactly as

$$\begin{aligned} \max \quad & y \\ \text{s.t.} \quad & p - y \in M(h_1, \ldots, h_s)[k] \\ & y \in \mathbb{R}. \end{aligned} \tag{1.19}$$

A justification for the ansatz Q_k is the following well-known and easy consequence of Putinar's Positivstellensatz (Theorem 1.20). We give the proof to illustrate how to apply the Positivstellensatz in convergence arguments.

Corollary 1.27 (see, e.g., Proposition 10.5.2 in [Mar08])**.** *Denote the minimum of (1.18) by p^* and the minimum of Q_k by $y^{(k)}$. If $M(h_1, \ldots, h_s)$ is Archimedean, then $y^{(k)} \nearrow p^*$ for $k \to \infty$.*

Proof. There is nothing to prove if $p^* = \pm\infty$. So suppose $p^* \in \mathbb{R}$ and let $\varepsilon > 0$. Hence $p - p^* + \varepsilon > 0$ on $K(h_1, \ldots, h_s)$. Since $M := M(h_1, \ldots, h_s)$ is Archimedean, $p - p^* + \varepsilon \in M$. Thus there is $k_\varepsilon \in \mathbb{N}$ with $p - p^* + \varepsilon \in M[k_\varepsilon]$, in other words, $y_\varepsilon := p^* - \varepsilon$ is feasible for Q_k with $k = k_\varepsilon$. As $\varepsilon > 0$ was arbitrary, we have $y_\varepsilon \to p^*$ for $\varepsilon \to 0$. Since moreover $y_\varepsilon \leq y^{(k_\varepsilon)} \leq p^*$, we also have $y^{(k_\varepsilon)} \to p^*$, and a subsequence of $y^{(k)}$ converges.

It remains to show monotonicity. Monotonicity follows from $M[k] \subset M[k+1]$ for all $k \in \mathbb{N}_0$ by the defining equation (1.13). □

Although finite convergence is not guaranteed [Las01], there are cases where an optimal solution $x \in K(h_1, \ldots, h_s)$ to (1.18) can be extracted from Q_k. For example, a sufficient condition for the extraction is given by a rank condition on associated *moment matrices* [HL05]. In the unconstrained case $\min\{p(x) : x \in \mathbb{R}^n\}$ given by $s = 0$ in (1.18) even more is known: Instead of solving Q_k with respect to $K(\emptyset) = \mathbb{R}^n$ which would be given as $\max\{y : p - y \in \Sigma\}$, one can consider the gradient variety[9], resulting in $2n$ constraints corresponding to the equations

$$\partial_{x_1} p = \ldots = \partial_{x_n} p = 0$$

[9]It is well-known that the gradient of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ vanishes at a local optimum, a fortiori at a global minimum.

and solve (1.19) with respect to the constraint polynomials

$$\partial_{x_1} p, \ldots, \partial_{x_n} p, -\partial_{x_1} p, \ldots, -\partial_{x_n} p. \tag{1.20}$$

Then we have:

Theorem 1.28 ([NDS06]). *Consider the set of polynomials of degree at most $d \in \mathbb{N}_0$ that possess a global continuous minimizer:*

$$\mathcal{F}_d := \{p \in \mathbb{R}[\underline{X}] : \deg(p) \leq d \text{ and } \exists x^* \in \mathbb{R}^n \text{ s.t. } p(x^*) = p^* = \inf_{x \in \mathbb{R}^n} p(x)\}.$$

Then, for the sos programs Q'_k with gradient variety constraint polynomials from (1.20)*, finite convergence holds for almost all[10] polynomials $p \in \mathcal{F}_d$. More precisely, there is a $k_0 \in \mathbb{N}_0$ such that for the optimal solutions $y^{(k)}$ of Q'_k one has $y^{(k)} = y^{(k_0)} = p^*$ for $k \geq k_0$. Moreover, a minimizer x^* of* (1.18) *can then be extracted.*

1.5.6. Sos and semidefinite programming

Sos programs translate to semidefinite programs. Semidefinite programs in turn are optimization problems that are well-understood, and many numerical solvers are available. Since semidefinite programs are omnipresent in this work – implicitly in the form as sos programs, or explicitly in Chapter 5 –, we recall their standard formulations in this section. Note that in Section 5.3 we use a more general form of sos programming. The basic idea itself – how extensions of semidefinite programming transfer to extensions of sos programming – is indicated in Section 8.2.4

Semidefinite programs are matrix optimization problems. To end with a scalar objective, the standard inner product on $\mathbb{R}^{n \times n}$ plays a crucial role. A *semidefinite program in primal form* minimizes a linear form over the cone of positive semidefinite matrices, subject to linear constraints. Formally, it is given by[11]

$$\begin{aligned} \min \quad & \langle C, X \rangle \\ \text{s.t.} \quad & \mathcal{A}(X) = b \\ & X \succeq 0 \end{aligned} \tag{SDP-P}$$

for given $C \in \mathcal{S}^n$, $b \in \mathbb{R}^m$ and a linear map $\mathcal{A} : \mathcal{S}^n \to \mathbb{R}^m$. The decision variables are $n \times n$ symmetric matrices X that are required to be positive semidefinite. Note that the constraint $\mathcal{A}(X) = b$ can be rewritten as

$$\langle A_1, X \rangle = b_1, \ \ldots, \ \langle A_m, X \rangle = b_m$$

for some matrices $A_1, \ldots, A_m \in \mathcal{S}^n$.

[10] More precisely, finite convergence holds if the *gradient ideal* $\langle \partial_{x_1} p, \ldots, \partial_{x_n} p \rangle$ is radical and the corresponding *complex gradient variety* consists of finitely many points. These properties are *generic* in the sense of algebraic geometry. See [NDS06] for details.

[11] We follow the presentation of semidefinite programming in Chapter 2 in [Hel00].

The dual program to SDP-P, i.e., a *semidefinite program in dual form*, is given by

$$\begin{aligned} \max \quad & \langle b, y \rangle \\ \text{s.t.} \quad & \mathcal{A}^T(y) + Z = C \\ & y \in \mathbb{R}^m, \quad Z \succeq 0, \end{aligned} \tag{SDP-D}$$

where $\mathcal{A}^T : \mathbb{R}^m \to \mathcal{S}^n$ is the adjoint of $\mathcal{A}$, that is, the unique linear map satisfying

$$\langle \mathcal{A}X, y \rangle = \langle X, \mathcal{A}^T y \rangle$$

for all $X \in \mathcal{S}^n$ and $y \in \mathbb{R}^m$. For the dual, the two constraints $\mathcal{A}^T(y) + Z = C$ and $Z \succeq 0$ can be rewritten as the single constraint

$$\sum_{k=1}^{m} y_k A_k \preceq C$$

with the matrices A_k from above. We call constraints as they appear in SDP-P *primal constraints* and in SDP-D *dual constraints*. For completeness, let us mention the well-known fact that equality constraints and matrix unknowns are perfectly acceptable in a semidefinite program in dual form, and vice versa. We give a proof to illustrate the techniques involved.

Observation 1.29. *Let $A \in \mathcal{S}^n$, $b \in \mathbb{R}^n$. A constraint of the form*

$$\langle A_i, X \rangle = b_i, \quad i = 1, \ldots, m, \quad X \succeq 0$$

with decision variable $X \in \mathcal{S}^n$ can be expressed as a single constraint of the form

$$\sum_{j=1}^{m} y_j B_j \preceq C, \quad y \in \mathbb{R}^{m'},$$

for suitable $C, B_1, \ldots, B'_m \in \mathcal{S}^{n'}$, $n', m' \in \mathbb{N}$, and vice versa.

Proof. We start with a single equation $\langle A, X \rangle = b$, $A \in \mathcal{S}^n$, $b \in \mathbb{R}$. This is equivalent to the two 1×1-matrix constraints $(\sum_{i,j} A_{ij} X_{ij}) \preceq b$, $(-\sum_{i,j} A_{ij} X_{ij}) \preceq -b$. Two matrix constraints $D \preceq E, F \preceq G$ can be written into one block diagonal matrix requirement $\operatorname{diag}(D, F) \preceq \operatorname{diag}(E, G)$ (by Observation 1.9). By the same argument, all m constraints can be written into a single constraint. Similarly, the requirement $X \succeq 0$ – in its equivalent form $0 \preceq X$ – can be appended to said block diagonal matrix. The proof for the converse direction is similar. □

Semidefinite programming has a rich duality theory. We refer to Chapter 4 of the handbook [WSV00] and the references therein for a detailed exposition. Let us note that, as a fundamental difference to linear programming, strong duality does not necessarily hold: It can happen that both the primal and the dual program have an optimal solution,

but the optimal values do not coincide. Put differently, it may happen that the *duality gap* is nonvanishing.

We present now the key theorem that links sum of squares programming and semidefinite programming. The theorem states that, given a polynomial $p \in \mathbb{R}[X_1, \ldots, X_n]$, the polynomial p has an sos decomposition if and only if a semidefinite system involving the coefficients of p has a solution. Before we state the theorem, we introduce the following notation: Let

$$[X]_d := (1, X_1, \ldots, X_n, X_1^2, X_1X_2, \ldots, X_n^d)^T$$

be the vector of all $\binom{n+d}{d}$ monomials in the unknowns $X_1, \ldots, X_n$ of degree at most d. In view of (1.1), $[X]_d$ is an ordered basis of $\mathbb{R}[X_1, \ldots, X_n]_d$.

Theorem 1.30 (see, e.g., Theorem 3.39 in [BPT13])**.** *A polynomial $p = \sum_\alpha p_\alpha X^\alpha$ in n variables of degree $2d$ is a sum of squares if and only if there is $Q \in \mathcal{S}^{\binom{n+d}{d}}$ with*

$$p = [X]_d^T Q [X]_d, \quad Q \succeq 0,$$

where we index the matrix Q by the exponent tuples. Equivalently, this is the case if and only $Q \succeq 0$ and the following system of $\binom{n+2d}{2d}$ linear equations (in the unknown entries in Q) holds:

$$p_\alpha = \sum_{\beta+\gamma=\alpha} Q_{\beta\gamma}.$$

Note that the characterization in Theorem 1.30 explicitly refers to the monomial basis on $\mathbb{R}[X_1, \ldots, X_n]_d$. It is, of course, easily possible to express the coefficients in terms of any other basis of the space of polynomials with a bound on the degree and its dual space. The details can be found in Theorem 3.41 in [BPT13].

With Theorem 1.30, it is easy and well-known to translate sos constraints to semidefinite constraints, and thus an sos program into a semidefinite program.

Corollary 1.31. *Let $c, a_1, \ldots, a_m \in \Sigma$. Put*

$$d_0 := \deg c, \quad d_i := \deg a_j, \ j \in [m], \quad d := \max_{0 \le j \le m} \lceil d_j/2 \rceil.$$

Then, the constraint

$$c + \sum_{j=1}^m y_j a_j \in \Sigma, \quad y \in \mathbb{R}^m,$$

is equivalent to the semidefinite system

$$c_\alpha + \sum_{j=1}^m (a_j)_\alpha y_j = \sum_{\beta+\gamma=\alpha} Q_{\beta\gamma}, \quad y \in \mathbb{R}^m, \quad Q \in \mathcal{S}^{\binom{n+d}{d}}, \ Q \succeq 0.$$

1.6. Existence and hardness results for MIPP

This section characterize sufficient conditions ensuring the existence of optimal solutions to MIPP and gives key hardness results.

1.6.1. Existence of optimal solutions

The aim of this section is to illustrate with the following proposition the interplay of geometrical, algebraical, topological and analytical conditions that force the existence of optimal solutions to MIPP. In the proposition, we assume that MIPP has feasible solutions, that is, $F_{\mathcal{I}} \neq \emptyset$. None of the single observations are new, so the proof consists of pointers to well-known results. We do not address the important but difficult question how to decide whether or not $F_{\mathcal{I}} = \emptyset$. Also, note that for the sake of clarity we omit implications that are easily derived by transitivity.

Proposition 1.32. *Let $f, g_1, \ldots, g_r \in \mathbb{R}[X_1, \ldots, X_n]$ be polynomial data for* MIPP *and suppose $F_{\mathcal{I}} \neq \emptyset$. Let $z \in \mathbb{R}$ with $z \geq f(q)$ for some $q \in F_{\mathcal{I}}$. Let $\mathbb{S}^{n-1}$ be a sphere corresponding to a norm on $\mathbb{R}^n$. Then, the following implications hold:*

$$
\begin{array}{ccccc}
f_{\deg(f)} > 0 & \Longleftrightarrow & \inf\limits_{x \in \mathbb{S}^{n-1}} f_{\deg(f)}(x) > 0 & & \\
\Downarrow & & & & \\
f \text{ is coercive} & \Longleftrightarrow & \text{all } \mathcal{L}^f_{\leq}(z') \text{ compact} & & \\
\Downarrow & & & & \\
M(z-f) \text{ Ar.} & \Longleftrightarrow & \mathcal{L}^f_{\leq}(z) \text{ compact} & & \\
\Downarrow & & & & \\
M(g_1, \ldots, g_r, z-f) \text{ Ar.} & \Longrightarrow & F \cap \mathcal{L}^f_{\leq}(z) \text{ compact} & \Longrightarrow & F_{\mathcal{I}} \cap \mathcal{L}^f_{\leq}(z) \text{ compact} \Rightarrow \text{opt: MIPP} \\
\Uparrow & & \Uparrow & & \Uparrow \\
M(g_1, \ldots, g_r) \text{ Ar.} & \Longrightarrow & F \text{ compact} & \Longrightarrow & F_{\mathcal{I}} \text{ compact} \\
\Uparrow & & & & \\
\exists i : K(g_i) \text{ compact} & & & & \\
\Uparrow & & & & \\
\exists i : -g_i \text{ coercive} & & & & \\
\Uparrow & & & & \\
\exists i : -(g_i)_{\deg(g_i)} > 0 & \Longleftrightarrow & \sup\limits_{x \in \mathbb{S}^{n-1}} g_{\deg(g_i)}(x) < 0 & &
\end{array}
$$

In the diagram, "Ar." abbreviates "is Archimedean" and "opt: MIPP*" abbreviates "*MIPP *has optimal solutions".*

Proof. The proof iterates as outer loop from top to bottom and as inner loop from left to right. The equivalence of f having a positive definite leading form f_d and f_d having a positive infimum on the sphere is from (1.3) To see that $f_{\deg(f)} > 0$ implies coercivity, let $d := \deg(f)$, and decompose f into its homogeneous components, $f = \sum_{j=0}^{d} f_j$. Let $\mathbb{S}$

denote the Euclidean unit sphere in $\mathbb{R}^n$. Let $c_j = \min_{x\in\mathbb{S}} f_j(x)$ for all j. By compactness, the c_j are finite, and by positive definiteness of f_d, we know that $c_d > 0$ by (1.3). By homogeneity,

$$f(x) = \sum_{j=0}^{d} f_j(x) = \sum_{j=0}^{d} f_j\left(\frac{x}{\|x\|_2}\right)\|x\|_2^j \geq \sum_{j=0}^{d} c_j^*\|x\|_2^j,$$

and f is bounded from below by a coercive univariate polynomial in $\|x\|_2$. The claim follows.

If f is coercive, boundedness of all sublevel sets follows from Proposition 1.4. As f is continuous, the sublevel sets are moreover closed. Compactness of all sublevel sets follows. Suppose now all sublevel sets are compact. Then they are a fortiori bounded, and the claim follows from Proposition 1.4. The topmost equivalence is proved.

Coercivity of f implies $M(z-f)$ Archimedean by Proposition 1.23. To see that coercivity implies that $M(z-f)$ is Archimedean, we know by the above that coercivity implies all sublevel sets are compact, especially $\mathcal{L}^f_{\leq}(z) = K(z-f)$ is compact. Then by Theorem 1.17 (4), $M(z-f)$ is Archimedean. On the other hand, if $M(z-f)$ is Archimedean, by Corollary 1.19, $K(z-f) = \mathcal{L}^f_{\leq}(z)$ is compact.

Now, let $M(z-f)$ be Archimedean. By Proposition 1.23, $M(g_1,\ldots,g_r,z-f)$ is Archimedean. The implication "$M(g_1,\ldots,g_r,z-f) \Rightarrow F$ compact " results from

$$K(g_1,\ldots,g_r,z-f) = F \cap \mathcal{L}^f_{\leq}(z)$$

and Corollary 1.19. That this implies compactness of $F_{\mathcal{I}}$ in turn follows from the fact that $F_{\mathcal{I}} = F \cap \mathbb{R}^n_{\mathcal{I}}$ and $\mathbb{R}^n_{\mathcal{I}}$ is closed. The final implication in this row is based on the subsequent observation: We know that $z \geq f(q)$ for some feasible q, hence $F_{\mathcal{I}} \cap \mathcal{L}^f_{\leq}(z)$ is nonempty. Minimizing f, a continuous function, over a nonempty, compact set yields an optimal solution.

Now, if $M(g_1,\ldots,g_r)$ is Archimedean, $M(g_1,\ldots,g_r,z-f)$ is Archimedean by Proposition 1.23. We refer to Corollary 1.19 again to see that $M(g_1,\ldots,g_r)$ Archimedean implies $F = K(g_1,\ldots,g_r)$ is compact. Also, $F_{\mathcal{I}}$ is compact if F is, by the argument $F_{\mathcal{I}} = F \cap \mathbb{R}^n_{\mathcal{I}}$ again (and $\mathbb{R}^n_{\mathcal{I}}$ is closed). The implications F compact ($F_{\mathcal{I}}$ compact) implies $F \cap \mathcal{L}^f_{\leq}(z)$ compact ($F \cap \mathcal{L}^f_{\leq}(z)$ compact) follow from the fact that $\mathcal{L}^f_{\leq}(z)$ is a closed set by continuity of f.

The next two implications – $K(g_i)$ compact for some $i \in [r]$ implies $M(g_1,\ldots,g_r)$ and $-g_i$ coercive implies $K(g_i)$ compact – were proved in Proposition 1.22.

The proof of the claim, "$-g_i$ has positive definite leading form implies $-g_i$ is coercive, is verbatim identical to the proof of "$f_d > 0$ implies f coercive".

Also, the last equivalence is proved as the first. □

Let us note that in the unconstrained case, the existence of a mixed-integer minimum enforces positive semidefiniteness of the leading form.

Proposition 1.33. *Let $f \in \mathbb{R}[X_1,\ldots,X_n]$ of degree d. Suppose that*

$$\inf_{x\in\mathbb{R}^n_{\mathcal{I}}} f(x) > -\infty.$$

Then $f_d \geq 0$; especially, d is even.

Proof. It suffices to show the claim for $\mathcal{I} = [n]$, that is, integer minimization. Suppose there is $x \in \mathbb{R}^n$ such that $f_d(x) < 0$. By homogeneity, we may assume $x \in \mathbb{S}_\infty^{n-1}$. By continuity, there is a whole neighborhood W of x such that $f_d(y) < 0$ for all $y \in W$. As $W \cap \mathbb{S}_\infty^{n-1} \neq \emptyset$, there is a point $r \in W \cap \mathbb{S}_\infty^{n-1}$ with rational coordinates $r_i = \frac{z_i}{n_i}$, $z_i \in \mathbb{Z}$, $n_i \in \mathbb{N}$, $i = 1, \ldots, n$. Now for all $\lambda \in \mathbb{R}$,

$$f(\lambda r) = \sum_{j=0}^{d} f_j(r)\lambda^j,$$

and since $f_d(r) < 0$, we have $f(\lambda r) \to -\infty$ as $\lambda \to \infty$. Since $r_i = \frac{z_i}{n_i}$, $i = 1, \ldots, n$, there is a lowest common denominator $l \in \mathbb{N}$ of the r_i. For $k \in \mathbb{N}$, we have especially $f(klr) \to -\infty$ as $k \to \infty$. But since $klr \in \mathbb{Z}^n$, f is unbounded from below on $\mathbb{Z}^n$. The conclusion that d must be even follows from homogeneity. □

1.6.2. Hardness results

In this section we state complexity results for important special cases of MINLP. We start with binary programming.

Theorem 1.34 ([Kar72])**.** *Let $A \in \mathbb{Z}^{m\times n}$, $b \in \mathbb{Z}^m$. The decision problem:*

$$\textit{Does} \quad Ax = b \quad \textit{have a solution } x \in \{0,1\}^n \textit{?}$$

is NP-complete.

The more general problem of linear integer programming is in a precise sense not more difficult than binary programming:

Theorem 1.35 (see, e.g., Problem MP1 in [GJ79])**.** *Let $A \in \mathbb{Z}^{m\times n}$, $b \in \mathbb{Z}^m$. The decision problem:*

$$\textit{Does} \quad Ax \leq b \quad \textit{have a solution } x \in \mathbb{Z}^n \textit{?}$$

is NP-complete.

We turn now to the complexity of continuous polynomial optimization. Nesterov [Nes00] showed how to encode a binary optimization problem as a continuous polynomial optimization problem. We repeat a simplified variant, which seems mathematical folklore (see, e.g., Lecture 11 in [Tod12]). Let $A \in \mathbb{Z}^{m\times n}$, $b \in \mathbb{Z}^m$, and consider the following continuous polynomial optimization problem:

$$\begin{aligned} \min \quad & p(x) := \|Ax - b\|_2^2 + \sum_{i=1}^{n} (x_i - x_i^2)^2 \\ & x \in \mathbb{R}^n \end{aligned} \tag{1.21}$$

For $x \in \mathbb{R}^n$, $p(x) = 0$ if and only if $Ax = b$ and $x_i = x_i^2$ for all $i \in [n]$, equivalently, $x \in \mathbb{R}^n$ satisfies $Ax = b$ and $x \in \{0,1\}^n$. By Theorem 1.34, we have:

Theorem 1.36 ([Nes00])**.** *Deciding nonnegativity of $f \in \mathbb{Z}[X_1, \ldots, X_n]$ is NP-hard, even if one restricts to polynomials of degree four.*

Positive definite leading forms of polynomials play a central role, so we also cite:

Theorem 1.37 ([Nes00])**.** *Deciding positive definiteness of $f \in \mathbb{Z}[X_1, \ldots, X_n]$ is NP-hard, even if one restricts to homogeneous polynomials of degree four.*

The following is a newer, stronger variant.

Theorem 1.38 (see, e.g., p. 459 in [AOPT13])**.** *Deciding nonnegativity of a polynomial $f \in \mathbb{Z}[X_1, \ldots, X_n]$ is strongly NP-hard, even if one restricts to biquadratic forms.*[12]

We end this section with two important incomputability results for all-integer problems and a hardness result for continuous optimization. The all-integer special case of MIPP with no constraint functions and a polynomial objective has the form

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in \mathbb{Z}^n \end{array} \tag{1.22}$$

and are studied in Chapter 6. Even this very restricted subclass of MIPP is already incomputable in general: Hilbert's tenth problem asks if there exists an algorithm that decides whether for a given polynomial f with integer coefficients the equation $f(x) = 0$ has a solution $x \in \mathbb{Z}^n$. Seventy years later it was proved by Matiyasevich [Mat70] that no such algorithm can exist. So if there was an algorithm to solve (1.22), we would also get an algorithm to decide whether $f(x) = 0$ has an integer solution by minimizing f^2 over $\mathbb{Z}^n$. Let us state this well-known, important consequence of Matiyasevich's result:

Theorem 1.39. *There cannot be an algorithm that solves* (1.22)*.*

As a final hardness result, we show that integer optimization of a linear function subject to quadratic constraints is also incomputable. Consider the program

$$\begin{array}{lll} \min & c^T x & \\ \text{s.t.} & g_i(x) \geq 0, & i = 1, \ldots, n \\ \text{s.t.} & x \in \mathbb{Z}^n & \end{array} \tag{1.23}$$

where $r = n$ and $g_i \in \mathbb{Z}[X_1, \ldots, X_n]$ are quadratic polynomials with no mixed terms. Extending ideas of Matiyasevich, Jeroslow proved the following theorem.

Theorem 1.40 ([Jer73])**.** *There cannot be an algorithm that solves* (1.23)*.*

[12] A subclass of homogeneous polynomials of degree four.

2. Half-spaces containing the feasible set

In this chapter we consider half-spaces that contain the feasible set F or the relaxed feasible set $F_{\mathcal{I}}$. We find these half-spaces with the help of gauges – a generalization of a norm on $\mathbb{R}^n$ – and a known feasible point $q \in F_{\mathcal{I}}$. We explore when half-spaces containing F can be used to find half-spaces containing $F_{\mathcal{I}}$; the latter half-space is called a cut for F. The main motivation is that these cuts have proved quite successful for linear and convex programming.

Section 2.1 gives a more detailed motivation of cuts. We repeat some results from linear and convex programming and outline our gauge-based approach. In that section we also recall the definition of gauges as well as their basic properties and special types of gauges that prove useful for our ansatz.

Section 2.2 formulates the task to find a half-space containing F as an auxiliary program. We give geometric characterizations that ensure the existence of feasible and optimal solutions.

Section 2.3 is concerned with the task to compute solutions to the auxiliary program. With additional assumptions, the task can be formulated as an sos program. Several linearization steps are necessary. We end with an approximating hierarchy and give convergence conditions.

Section 2.4 closes the chapter with an investigation when a half-space containing F can be used to derive a cut for F, that is, a half-space containing $F_{\mathcal{I}}$.

2.1. Motivating half-spaces and preliminaries on gauges

In this chapter we consider, as outlined in the introduction, half-spaces that contain the feasible set or the relaxed feasible set. Yet the task to find a half-space containing a given set is an interesting task by itself. We will thus, for additional generality, state our results using a deputy set $S \subset \mathbb{R}^n$ instead. We keep in mind that candidates for this set S are, amongst others, the feasible set $F_{\mathcal{I}}$, the relaxed feasible set F, and the set of all optimal solutions of MINLP.

At first, we discuss the problem of finding a valid, and tight if possible, inequality. The formulation we use is quite general and assumes knowledge of a point $q \in S$. Our auxiliary problem finds a valid inequality that minimizes the distance to q, where our measure for the distance comes from a gauge. In later sections we restrict to polynomial constraints and polyhedral gauges and derive an approximating hierarchy, which makes the problem tractable. Finally, we show how techniques from mixed-integer linear programming can, in certain cases, yield linear cuts for MINLP.

The geometric and algebraic perspectives in this section are closely related. Recall from Section 1.4.10 that half-spaces correspond to linear inequalities, and a half-space containing S corresponds to a valid linear inequality for S. A linear cut for S is a valid linear inequality for $S_{\mathcal{I}}$. Before we start with the math, we present some results on valid linear inequalities and cuts in linear and convex programming.

2.1.1. Motivation from linear and convex programming

Valid linear inequalities are of prime importance in the theory and history of linear, linear integer and linear mixed-integer programming.[1] One reason is the following. For a given linear function $c^T x$ to be optimized over $F_{\mathcal{I}}$ – where the constraints in linear programming are affine-linear functions –, it is not difficult to see that it is sufficient to optimize it over the convex hull of $F_{\mathcal{I}}$ instead, see, e.g., [CCZ10]. The convex hull of $F_{\mathcal{I}}$ in turn is in this case, for rational data, polyhedral. This fact was first proved by Meyer [Mey74] and is also known as the *fundamental theorem of integer programming* [CCZ10]. Hence, the convex hull is completely described by finitely many valid inequalities, the so-called facet-defining inequalities. Knowledge of all facet-defining inequalities of $F_{\mathcal{I}}$ thus reduces mixed-integer linear programming to linear programming, at least in principle.[2]

We saw that in mixed-integer linear programming, the search for half-spaces that contain $F_{\mathcal{I}}$ but do not contain a given point $q \in F$ is the search for linear cuts. The first result in this direction was Gomory's algorithmic approach to linear cuts [Gom58]. Later, Chvátal and Schrijver showed that the repeated application of all Gomory-type linear cuts yields the convex hull of $F_{\mathcal{I}}$ (for linear constraints), see Theorem 1.1 in [MMWW02]. Nowadays, the underlying theory of linear cuts has become quite deep and the number

[1] For example, the seminal work on integer and combinatorial optimization [NW88] devotes a whole chapter on the theory of valid inequalities.

[2] However, by complexity considerations, it must be NP-hard to compute all facet-defining inequalities of $F_{\mathcal{I}}$.

of different types of linear cuts is – even though the underlying ideas of the cuts are often related – vast. The article [MMWW02] is a modern presentation of the most influential linear cuts, and the article [CL01] explores the relationships in the linear cut zoo. Recently, *maximal lattice free polyhedra* have attracted attention:[3] It can be shown that, as the authors in [AWW11] put it, the strongest linear cuts are derived from maximal lattice-free polyhedra.

Linear cuts also play a fundamental role in the actual solution process in a modern, state-of-the art linear solver. This seems like a generally agreed-upon fact in the optimization community, but as a reference, let us mention a talk given in 2010 [BGR10] by Robert E Bixby, a mathematician who also conceived two of the three large commercial linear solvers of today, presented computer experiments on which single feature in the solution process of mixed-integer linear programs was most helpful (the feature list consisted of cuts, presolve, variable selection, heuristics and node presolve). In his experiments, cuts contributed by far the most significant part.

Cutting plane methods also exist in convex programming. Soon after the first linear cut method was published by Gomory in 1958 [Gom58], Kelley published a version for the more general, convex program in 1960 [Kel60]. A modern introduction into the key ideas on linear cuts for the continuous convex problems (absence of integrality constraints) is given in [BV11]. For an overview on linear cuts for mixed-integer convex problems, we refer to Chapter 4 in [BKL+13].

2.1.2. A note on our approach

With this background from linear and convex programming in mind, we approach the task to find valid linear inequalities for MINLP, and take a look at some ideas on how to generate cuts from these valid inequalities. Now seems the right moment to mention that we must not expect a general algorithm of the following type: Given a point $q \in F \setminus F_{\mathcal{I}}$, compute a valid inequality for $F_{\mathcal{I}}$ that is violated by F. Why is this? Algorithms of this type are *separation* algorithms. A fundamental result from convex optimization states that, roughly speaking, any separation algorithm for convex problems yields an optimization algorithm (Corollary 4.2.7 in [GLS93]). Hence, if we could give an algorithm that in polynomial time separates $q \in F$ from $F_{\mathcal{I}}$, we have a polynomial time optimization algorithm for MINLP, which in view of Theorem 1.35 cannot be expected (unless P = NP).

[3] A convex set $K \subset \mathbb{R}^n$ with non-empty interior is *lattice-free* if $\mathbb{Z}^n$ does not intersect the interior of K. It is *maximal lattice-free* if K is inclusion maximal in the set of all lattice-free convex sets. If the data is rational, it can then be shown that every maximal lattice-free set K is actually a polyhedron and each facet of K has an integer point in its relative interior [BCCZ10]. We refer to [AWW11] and the references therein for further material on the topic.

2.1.3. Gauges

Measuring distances via gauges

A *gauge* is a function $\gamma : \mathbb{R}^n \to \mathbb{R}$ of the form

$$\gamma(x) = \gamma(x; A) = \inf\{t \geq 0 : x \in tA\} \tag{2.1}$$

for $A \subset \mathbb{R}^n$ compact, convex with 0 in its interior. Let us note some well-known properties of gauges:

Proposition 2.1 (see, e.g., p. 14 in [Hol75]; also p. 25 and Theorem 1.39 in [Rud91])**.** *Let $\gamma : \mathbb{R}^n \to \mathbb{R}$ be a gauge. Then, the gauge γ*

- *is nondegenerated, $\gamma(x) = 0$ if and only if $x = 0$ and $\gamma(x) < \infty$ for all $x \in \mathbb{R}^n$,*
- *attains the infimum in the defining equation (2.1) for all $x \in \mathbb{R}^n$,*
- *is positively homogeneous, $\gamma(\lambda x) = \lambda\gamma(x)$ for all $x \in \mathbb{R}^n$ and $\lambda > 0$,*
- *is absolutely homogeneous, $\gamma(\lambda x) = |\lambda|\gamma(x)$ for all $x \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$ if A in (2.1) is symmetric ($A = -A$),*
- *is subadditive, $\gamma(x + y) \leq \gamma(x) + \gamma(y)$ for all $x, y \in \mathbb{R}^n$.*

Remark on the proof. For the actual proof we refer to the given references. To explain the assumptions, we note the following. The property $\gamma(x) < \infty$ follows from A having 0 in its interior. The property $\gamma(x) = 0$ if and only if $x = 0$ follows from A being bounded. The infimum is attained due to compactness. Positive homogeneity is immediate from the definition. Absolute homogeneity follows from definition and symmetry of A. Subadditivity follows from convexity. □

The *(closed) gauge ball* at $p \in \mathbb{R}^n$ of radius $R \geq 0$ is the set

$$\mathbb{B}_R^\gamma(p) := \mathbb{B}_R(p; \gamma) := \{x \in \mathbb{R}^n : \gamma(x - p) \leq R\}. \tag{2.2}$$

Given $A, B \subset \mathbb{R}^n$ and a gauge γ on $\mathbb{R}^n$, the *distance from A to B* is

$$d(A, B) = \inf\{\gamma(b - a) : a \in A, b \in B\}. \tag{2.3}$$

For a singleton set $A = \{a\}$ we also write $d(a, B)$.

In the following Lemma, we note that the distance measured by gauges between two sets A, K is attained if A is closed and K compact. Since the proof is elementary and similar to the proof for the case of norms, we expect the result to be known. As we could not find a proof in the literature, we give one in the appendix for completeness.

Lemma 2.2. *Let γ be a gauge on $\mathbb{R}^n$, $A \subset \mathbb{R}^n$ be closed, $K \subset \mathbb{R}^n$ be compact and A, $K \neq \emptyset$. Then there are $a^* \in A$ and $k^* \in K$ with*

$$d(a^*, k^*) = d(A, K).$$

Proof. See Chapter B in the appendix. □

Polar gauges

The function $\gamma^\circ(x) = \sup\{x^T y : y \in \mathbb{R}^n, \gamma(y) \leq 1\}$ is the *polar* of γ. The polar is again a gauge, see, e.g., Theorem 15.1 in [Roc70], and if we introduce the notion of a *polar* of a set $A \subset \mathbb{R}^n$ as

$$A^\circ = \{x \in \mathbb{R}^n : x^T y \leq 1 \ \forall y \in A\}, \tag{2.4}$$

it can be shown that

$$\gamma^\circ(x, A) = \gamma(x, A^\circ), \tag{2.5}$$

see, e.g., Theorem 15.1 in [Roc70]. We also consider gauges γ that are *polyhedral*: $\gamma(\cdot, A)$ is called polyhedral if A is a polyhedron. As the polar of a polyhedron is a polyhedron (see, e.g., Corollary 19.2.2 in [Roc70]), it is clear in view of (2.5) that the polar of a polyhedral gauge is again polyhedral. For a polyhedral gauge $\gamma(\cdot, A)$, the extreme points of A of are called *fundamental directions* of γ.

2.2. Finding valid inequalities using gauges

Suppose you are given a set $S \subset \mathbb{R}^n$ and your task is to compute an inequality that is tight for S. Can this be done? This task seems too difficult without additional assumptions. Our approach to generate valid, possibly tight, inequalities for the set S is governed by the following assumption: You are given a point q in S. Is it now possible to compute a valid inequality for S with minimum distance to q, in the sense of (2.3)? And then, secondly, is the inequality thus computed also tight for S? A somewhat abstract answer to the first question is that the task to find a valid inequality can be cast as an auxiliary program (see V1 below). A constructive answer to the first question, i.e., to compute a valid inequality for a given set S in practice, requires additional assumptions, and is deferred to Section 2.3. There, provided S is semi-algebraic, we approximate Program V1 with a hierarchy of sos programs, that yield feasible solutions and, under additional assumptions, converge towards the optimal solution. The second question can be answered in the affirmative (Proposition 2.3).

Now consider the following program. It computes a valid inequality for S that is as close as possible to $q \in S$.

$$\begin{array}{rl} \min & d(q, H(a,b)) \\ \text{s.t.} & a \neq 0 \\ & a^T x \leq b \quad \text{for } x \in S \\ & a \in \mathbb{R}^n,\ b \in \mathbb{R} \end{array} \tag{V1}$$

In our setting, natural candidates for S are F or $F \cap \mathcal{L}^f_{\leq}(f(q))$. Before we prove some properties of this program, let us note that Program V1 can also be considered as a hyperplane location problem. For an overview over such problems, we refer to Chapter 7 in [Sch99]. Let us now interpret solutions of Program V1 geometrically.

Proposition 2.3. *For Program* V1*,*

1. *Every feasible solution (a, b) yields a valid inequality $(a^T x \leq b)$ for S.*
2. *Every optimal solution (a, b) yields an inequality $(a^T x \leq b)$ that is tight.*
3. *Every optimal solution (a, b) with objective value 0 yields an inequality that is tight at q.*

Proof. Claim (1) is clear. To see Claim (2), let (a, b) be an optimal solution and assume the contrary, i.e., there is $b' < b$ with $(a^T x \leq b')$ valid for S. By Lemma 2.2 there is $p \in H(a, b)$ with $\gamma(p - q) = d(q, H(a, b))$. Using $q \in S$, we get the inequalities

$$a^T q \leq b' < b = a^T p, \tag{2.6}$$

and hence $a^T(p - q) > 0$. Put

$$\hat{p} := p + \lambda(q - p) \text{ with } \lambda := \frac{b - b'}{a^T(p - q)}.$$

Observe that $\hat{p}$ is a point on $H(a, b')$:

$$b' = a^T p + b' - b = a^T p + \frac{a^T(q-p)}{a^T(q-p)}(b' - b) = a^T p + \lambda a^T(q-p) = a^T \hat{p}.$$

Note that (2.6) implies $\lambda > 0$ and $\lambda \leq \frac{b-b'}{b-b'} = 1$. Since $\hat{p} - q = (1-\lambda)(p-q)$, all our observations combine to

$$d\,(q, H(a, b')) \leq \gamma(\hat{p} - q) = (1-\lambda)\gamma(p-q) < \gamma(p-q) = d(q, H(a, b)).$$

Hence (a, b') is a feasible solution to V1 with better objective value, contradicting optimality of (a, b).

To see Claim 3, note that if the objective value is 0 at (a, b) we know from Lemma 2.2 that $d(q, p) = 0$ for some $p \in H(a, b)$, hence $q = p$ and we conclude $q \in H(a, b)$. The claim follows. □

It turns out that feasibility of V1 is sufficient for the existence of optimal solutions.

Theorem 2.4. *Let $S \subset \mathbb{R}^n$ and $q \in S$. Program* V1 *is feasible if and only if* $\operatorname{conv} S \subsetneq \mathbb{R}^n$. *In that case optimal solutions exist.*

Proof. Assume first that Program V1 is feasible. Then $S \subset (a^T x \leq b)$ for some $a \in \mathbb{R}^n$, $a \neq 0$, $b \in \mathbb{R}$. Now

$$\operatorname{conv} S \subset \operatorname{conv}(a^T x \leq b) = (a^T x \leq b) \subsetneq \mathbb{R}^n$$

follows.

To see the converse implication, let $z \in \mathbb{R}^n \setminus \operatorname{conv} S$. By Theorem 1.13, we may separate z from $\operatorname{conv} S$ by a hyperplane $H(a, b)$ with $a^T x \leq b$ for all $x \in \operatorname{conv} S$, and this hyperplane is feasible to Program V1.

To see that feasibility implies existence of optimal solutions, we construct an optimal solution that corresponds to a supporting hyperplane at a suitably chosen point on the boundary of the closure of the convex hull of S. So let $(a^T x \leq b)$ be an inequality that is valid for S and thus $\operatorname{conv} S$. Moreover, as half-spaces are closed, $(a^T x \leq b)$ remains valid for $C := \operatorname{cl}\operatorname{conv} S$, and we conclude $C \subsetneq \mathbb{R}^n$. Also, C is convex by Theorem 1.12. As $q \in S \subset C$, C is a nonempty, proper subset of $\mathbb{R}^n$, so its boundary $B := \operatorname{bd} C$ is nonempty. As B is closed, Lemma 2.2 ensures the existence of $x_1 \in B$ with $d(q, x_1) = d(q, B)$. By Theorem 1.14, there is a supporting hyperplane $H_1 = H(a_1, b_1)$ to C at x_1. We claim that (a_1, b_1) is optimal.

Suppose it is not, so there are (a_2, b_2) such that $(a_2^T x \leq b_2)$ is valid for S and the corresponding hyperplane $H_2 := H(a_2, b_2)$ suffices $d_2 := d(q, H_2) < d(q, H_1) =: d_1$. Again there is $x_2 \in H_2$ with $d(q, x_2) = d_2$. We now distinguish all three possible locations for x_2 and derive a contradiction in every case.

1. x_2 in $\mathbb{R}^n \setminus C$. As $q \in C$, the line segment from x_2 to q crosses the boundary B of C at a point x_3. But then $d(x_3, q) \leq d_2 < d_1$, contradicting minimality of x_1.

2. $x_2 \in \operatorname{bd} C = B$. As in the case $x_2 \notin C$, this contradicts minimality of x_1.

3. $x_2 \in \operatorname{int} C$. Hence there is $\varepsilon > 0$ with $x_2 + \varepsilon a \in C$, and $a_2^T(x_2 + \varepsilon a_2) = b_2 + \varepsilon a_2^T a_2 > b_2$ as $x_2 \in H(a_2, b_2)$. Consequently, $(a_2^T x \leq b_2)$ is not a valid inequality for C. On the other hand, $S \subset \{x \in \mathbb{R}^n : a_2^T x \leq b_2\}$ by assumption on $(a_2^T x \leq b_2)$. But Theorem 1.12 then entails $C = \operatorname{cl}\operatorname{conv} S \subset (a_2^T x \leq b_2)$, contradicting our observation that $(a_2^T x \leq b_2)$ is not valid for $x_2 + \varepsilon a_2 \in C$.

We conclude that x_2 cannot exist, so neither can (a_2, b_2). Hence (a_1, b_1) is an optimal solution to Program V1. □

So far, we have not yet taken into account the presence of an objective function. Suppose the objective function is given as $f : S \to \mathbb{R}$. Now, if $q \in S$ and q is not a local maximizer of f, it can be shown (Proposition A.3) that q lies on the boundary (with respect to S) of $\mathcal{L}^f_{\leq}(f(q)) \cap S$. In other words and under above-said assumption, adding the inequality $f(x) \leq f(q)$ (note that no optimal solutions get lost) pushes q to the boundary of the (altered) feasible set $\mathcal{L}^f_{\leq}(f(q)) \cap S$. For the convex setting, this has interesting consequences: By Corollary A.4, there is a supporting hyperplane to $S \cap \mathcal{L}^f_{\leq}(f(q))$ which is tight at q. We have thus proved:

Proposition 2.5. *Let S be a convex set, $q \in S$, $f : S \to \mathbb{R}$ quasiconvex. If q is not a local maximum of f, then Program* V1 *with S replaced by $S \cap \mathcal{L}^f_{\leq}(f(q))$ has an optimal solution with objective value* 0.

2.3. Computing valid inequalities

We now want to make Program V1 tractable through a relaxation that can be solved with a computer. Several obstacles need to be addressed: We lack an analytic, preferably linear, expression for $d(q, H(a,b))$, the distance of q to the hyperplane. This is, as shown in Section 2.3.1, possible if we enforce an additional constraint. This constraint, however, is non-convex, but using disjunctive arguments, we linearize the constraint nevertheless in Section 2.3.2. Finally, the requirement that the inequality is valid for $x \in S$ possibly corresponds to infinitely many linear constraints, which we sidestep by restricting to semi-algebraic S as explained in Section 2.3.3.

2.3.1. Linearization of the objective

The aim of this section is to linearize the objective function $d(q, H(a,b))$ in Program V1. The following result is the first step.

Theorem 2.6 (Theorem 1.1 in [PC01])**.** *Let γ be a gauge on $\mathbb{R}^n$ and denote its polar by γ°. Furthermore, let $0 \neq a \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Then*

$$d(q, H(a,b)) = \begin{cases} (b - a^T q)/\gamma^\circ(a), & a^T q \leq b, \\ (a^T q - b)/\gamma^\circ(-a), & a^T q > b. \end{cases} \tag{2.7}$$

The variable a enters the fractions in (2.7) in a nonlinear fashion. Moreover, the constraint $a \neq 0$ is not closed. Now compare Program V1 with the following program that avoids a constraint of the form $a \neq 0$:

$$\begin{aligned} \min \quad & b - a^T q \\ \text{s.t.} \quad & a^T x \leq b \quad \text{for } x \in S \\ & \gamma^\circ(a) = 1 \\ & a \in \mathbb{R}^n,\ b \in \mathbb{R}, \end{aligned} \tag{V2}$$

It turns out that Programs V1 and V2 are closely related.

Proposition 2.7. *Let $q \in S \subset \mathbb{R}^n$ and γ be any gauge on $\mathbb{R}^n$. Then, solutions of Program* V1 *are in correspondence to solutions of Program* V2 *as follows: The optimal values coincide. Moreover, if (a,b) is feasible/optimal to* V1*, then $\frac{(a,b)}{\gamma^\circ(a)}$ is feasible/optimal to* V2*. If (a,b) is feasible/optimal to* V2*, then (a,b) is feasible/optimal to* V1*.*

Proof. By (2.7) and using the fact that $a^T q \leq b$, Program V1 is the same as

$$\begin{aligned} \min \quad & (b - a^T q)/\gamma^\circ(a) \\ \text{s.t.} \quad & a^T x \leq b \quad \text{for } x \in S \\ & a \neq 0 \\ & a \in \mathbb{R}^n,\ b \in \mathbb{R}. \end{aligned} \tag{2.8}$$

Let (a, b) be feasible for Program (2.8). To ease the presentation write $a' = a/\gamma^\circ(a)$ and $b' = b/\gamma^\circ(a)$. As $a \neq 0$, $H(a, b) = H(a', b')$ and the inequality $[a', b']$ is valid for S. Also, the tuple (a', b') is feasible to V2, and the objective values coincide. On the other hand, every feasible solution to V2 is a feasible solution to (2.8), with coinciding objective values. This implies the claim about optimal solutions, too. □

To summarize, instead of solving V1 we may solve V2.

2.3.2. Linearization of a constraint

Program V2 contains the non-convex constraint

$$\gamma^\circ(a) = 1.$$

This non-convexity can be circumvented by considering polyhedral gauges: There are well-known results on the facets of their unit balls which allow us to linearize the constraint $\gamma^\circ(a) = 1$. We start with a characterization of the faces of a unit ball in terms of the extreme points of the polar polyhedron defined in (2.4).

Theorem 2.8 (see, e.g., Proposition 3.2, and Theorems 5.3 and 5.5 in Chapter I.4 in [NW88])**.** *If P is a full-dimensional and bounded polyhedron and 0 is an interior point of P then*

$$P = \bigcap_{k \in K} \left\{ x \in \mathbb{R}^n : \pi_k^T x \leq 1 \right\}$$

where $\{\pi_k\}_{k \in K}$ are the extreme points of P°. The inequalities $\pi_k^T x \leq 1$ describe exactly the facets of P.

Now suppose γ is a polyhedral gauge. Denote its fundamental directions – that is, the extreme points of its unit ball B – by $v_1, \ldots, v_l \in \mathbb{R}^n$, and the unit ball of the polar gauge γ° by B°. Then, it is well-known that the facets of B° are $E_j = \left\{ x \in B^\circ : v_j^T x = 1 \right\}$, $j = 1, \ldots, l$. This can be seen as follows:

As γ is a polyhedral gauge, so is the polar gauge γ°. This implies that B° is a polyhedron which is is full-dimensional, bounded, with 0 in its interior. Thus, B° satisfies the assumptions of Theorem 2.8, so every facet of B° is given by one of the sets $\{x \in B^\circ : \hat{\pi}_k^T x = 1\}$ where $\{\hat{\pi}_k\}_{\hat{K}}$ are the extreme points of $B^{\circ\circ}$. But using the fact that $B^{\circ\circ} = B$ – this holds for all closed, convex sets containing the origin [Roc70, Th. 14.5] – we have $\{\hat{\pi}_k\}_{k \in \hat{K}} = \{v_1, \ldots, v_l\}$, so every facet of B° corresponds to one of the E_j.

We use this characterization to write the boundary of B°, in other words the non-convex set $\{a \in \mathbb{R}^n : \gamma^\circ(a) = 1\}$, with disjunctive linear constraints.

Corollary 2.9. *Let γ be a polyhedral gauge and denote its fundamental directions by $v_1, \ldots, v_l \in \mathbb{R}^n$. Denote the unit ball of γ° by B°. Then*

$$\operatorname{bd} B^\circ = \{x \in \mathbb{R}^n : \gamma^\circ(x) = 1\} = \bigcup_{j=1}^{l} E_j, \tag{2.9}$$

where $E_j = \left\{ x \in B^\circ : v_j^T x = 1 \right\}$, $j = 1, \ldots, l$.

Proof. The first equality in (2.9) is clear. As the boundary of every polyhedron is the union of its facets, which are the E_j in this case, the second equality in (2.9) follows. □

With this corollary, we note that V2 can be disjunctified. Here, a disjunctification of a program $\min\{f(x) : x \in \cup_{k\in K} M_k\}$ – where K and M_k are sets and $f : \cup_{k\in K} M_k \to \mathbb{R}$ is any function – means rewriting it as $|K|$ programs $\min\{f(x) : x \in M_k\}$, $k \in K$.

Proposition 2.10. *Let $q \in S \subset \mathbb{R}^n$ and γ be a polyhedral gauge on $\mathbb{R}^n$ with fundamental directions $v_1, \ldots, v_l \in \mathbb{R}^n$. Consider the $j = 1, \ldots, l$ programs*

$$\begin{array}{ll} \min & b - a^T q \\ s.t. & v_i^T a \leq 1 \quad \textit{for } i \in \{1, \ldots, l\},\ i \neq j \\ & v_j^T a = 1 \\ & a^T x \leq b \quad \textit{for } x \in S \\ & a \in \mathbb{R}^n,\ b \in \mathbb{R}. \end{array} \tag{V3$_j$}$$

Then the programs V3$_j$ *are a disjunctification of* V2.

Proof. Immediate from Corollary 2.9. □

Remark 2.11. Let the assumptions of Proposition 2.10 hold. Disjunctification entails as usual the following statements:

1. Program V2 has feasible solutions if and only if V3$_j$ has for some $j \in [l]$.
2. Denote the optimal value of Program V2 by z^* and for $j \in [l]$ denote the optimal value of Program V3$_j$ by z_j^*. Then $z^* = \min_{j\in[l]} z_j^*$.
3. If (a, b) is an optimal solution of V2, there is $j \in [l]$ such that (a, b) is an optimal solution to V3$_j$.
4. If $(a, b; j_0)$ is an optimal solution to V3$_j$ with $z_{j_0}^* = \min_{j\in[l]} z_j^*$, then (a, b) is an optimal solution to V2 (with the same objective value).

2.3.3. An approximative solution to semi-algebraic sets

The aim of this section is to state an approximating hierarchy to the Programs V3$_j$ in terms of sos programming. Note that the constraint

$$a^T x \leq b \quad \text{for } x \in S$$

is semi-infinite if S contains infinitely many points. There is much literature on semi-infinite programming problems. Classical overview articles are, e.g., [HK93], [RR98]; a more recent survey is [Ste12]. A bi-level approach is explored in [Ste13]. Also, several numerical solution methods exist, for an overview, we refer to [RG98], [LS07] [VRSS08].

However, in this work we take a different route. Let us explore how semi-infinite constraints can be sidestepped by the requirement of semi-algebraic S and a polyhedral

gauge γ. For example when considering MIPP, S is semi-algebraic if $S = F$ or if $S = F \cap \mathcal{L}^f_{\leq}(z)$ for some $z \in \mathbb{R}$. To avoid such case distinctions, we use an arbitrary basic closed semi-algebraic set $S = K(h_1, \ldots, h_s)$ instead of the one given by the constraints and the objective, say, $S = K(g_1, \ldots, g_r)$ or $S = K(g_1, \ldots, g_r, z - f)$ and so forth.

With this in mind, we consider the following hierarchy of programs.

$$\begin{aligned} \min \quad & b - a^T q \\ \text{s.t.} \quad & v_i^T a \leq 1 \quad \text{for } i \in [l],\ i \neq j \\ & v_j^T a = 1 \\ & b - \sum_{i=1}^{n} a_i X_i \in M\left(h_1, \ldots, h_s\right)[k] \\ & a \in \mathbb{R}^n,\ b \in \mathbb{R}. \end{aligned} \tag{$\mathrm{VR}_{j,k}$}$$

where $k \in \mathbb{N}$, $h_1, \ldots, h_s \in \mathbb{R}[X_1, \ldots, X_n]$ and $q \in S := K(h_1, \ldots, h_s)$, and $v_1, \ldots, v_l \in \mathbb{R}^n$ be given.

Before we proceed, let us convince ourselves that this formulation is within the scope of sos programming.

Observation 2.12. *Program* $\mathrm{VR}_{j,k}$ *is a valid sos program.*

Proof. Linear programming constraints can be used in sos programming constraints, cf. (1.17). Also, the containment constraint involving the quadratic module is allowed in sos programming, cf. (1.16). □

The next proposition shows that feasible solutions to Program $\mathrm{VR}_{j,k}$ yield feasible solutions to Program $\mathrm{V3}_j$.

Proposition 2.13. *Let* $h_1, \ldots, h_s \in \mathbb{R}[X_1, \ldots, X_n]$, $S := K(h_1, \ldots, h_s)$ *and* $v_1, \ldots, v_l \in \mathbb{R}^n$ *be given. If* $(a, b; j, k)$ *is feasible to* $\mathrm{VR}_{j,k}$, *some* $j \in [l]$, $k \in \mathbb{N}$, *then* $(a, b; j)$ *is feasible to* $\mathrm{V3}_j$.

Proof. Let $(a, b; j, k)$ feasible to $\mathrm{VR}_{j,k}$. Feasibility implies that

$$b - \sum_{i=1}^{n} a_i X_i \in M\left(h_1, \ldots, h_s\right)[k],$$

which entails $a^T x \leq b$ on $S = K(h_1, \ldots, h_s)$ by Observation 1.18. Hence $(a, b; j)$ is feasible to $\mathrm{V3}_j$. □

The next theorem shows that we get valid inequalities that are, at least asymptotically, tight if the corresponding quadratic module $M(h_1, \ldots, h_s)$ is Archimedean. Note that for the important special cases $S = F$ or $S = F \cap \mathcal{L}^f_{\leq}(z))$ for some suitable $z \in \mathbb{R}$, we have given characterizations in terms of the MIPP domain that ensure the Archimedean property in Propositions 1.22, 1.23 and 1.24.

Theorem 2.14. *Let $h_1, \dots, h_s \in \mathbb{R}[X_1, \dots, X_n]$ and $q \in S := K(h_1, \dots, h_s)$. Suppose further that $M(h_1, \dots, h_s)$ is Archimedean. Also, let a polyhedral gauge γ on $\mathbb{R}^n$ with fundamental directions $v_1, \dots, v_l \in \mathbb{R}^n$ be given. Then, it holds:*

1. *There is $j \in [l]$ such that Program $\mathrm{VR}_{j,k}$ is feasible for eventually all k.*
2. *Fix $j \in [l]$. Denote the optimal value of $\mathrm{V3}_j$ by z_j^* and put $z^* = \min_j z_j^*$. For $k \in \mathbb{N}$ denote the optimal value of $\mathrm{VR}_{j,k}$ by $z_{j,k}$ and put $z_k = \min_j z_{j,k}$. Then $z_k \searrow z^*$ for $k \to \infty$.*

Proof. To see Claim 1, we first show that $\mathrm{V3}_j$ has a feasible solution. To this end note that $M(h_1, \dots, h_s)$ Archimedean implies that $S = K(h_1, \dots, h_s)$ is compact (Corollary 1.19). Consequently, $\operatorname{conv} S$ is a proper subset of $\mathbb{R}^n$ and by Theorem 2.4 an optimal solution (a, b) to V1 exists. By scaling the optimal solution if necessary, we may further assume that $\gamma^\circ(a) = 1$, hence (a, b) is optimal for V2 by Proposition 2.7. By Remark 2.11, there is $j \in [l]$ such that $(a, b; j)$ is optimal for $\mathrm{V3}_j$. Fix $\varepsilon > 0$. Then $b + \varepsilon - \sum_{i=1}^n a_i x_i > 0$ for $x \in S = K(h_1, \dots, h_s)$, and by Theorem 1.20 there is $k_0 \in \mathbb{N}$ with $b + \varepsilon - \sum_{i=1}^n a_i X_i \in M(h_1, \dots, h_s)[k]$ for all $k \geq k_0$. Hence $(a, b + \varepsilon; j, k)$ is feasible for $\mathrm{VR}_{j,k}$ for all $k \geq k_0$, and Claim 1 is proved.

To see Claim 2, fix $\varepsilon > 0$. We have just proved that $z^* = z_{j_0}^*$, some $j_0 \in [l]$, is finite. Denote some corresponding optimal solution by $(a, b; j_0)$. Furthermore we have just proved that there is $k_\varepsilon \in \mathbb{N}$ such that $(a, b + \varepsilon; j_0, k_\varepsilon)$ is a feasible solution of Claim $\mathrm{VR}_{j,k}$. Hence

$$z_{k_\varepsilon} = \min_j z_{j,k_\varepsilon} \leq z_{j_0,k_\varepsilon} \leq b + \varepsilon - a^T q = z_{j_0}^* + \varepsilon = z^* + \varepsilon$$

and as $z_j^* \leq z_{j,k_\varepsilon}$ by Proposition 2.13, we have $z^* \leq z_{k_\varepsilon} \leq z^* + \varepsilon$, hence $z_{k_\varepsilon} \to z^*$ for $\varepsilon \to 0$.

Since the sets $M(h_1, \dots, h_s)[k]$ are increasing in k, the values $z_{j,k}$ are monotonically decreasing in k for j fixed. Hence the values $z_k = \min_j z_{j,k}$ are monotonically decreasing. Together with the observation $z_{k_\varepsilon} \to z^*$ for $\varepsilon \to 0$ we find $z_k \searrow z^*$ for $k \to \infty$. □

To summarize, we have shown that the problem to find a tight valid inequality for S, using a gauge γ and a known feasible point $q \in S$ as stated in V1, can be approximated with sos programming if the set S is semi-algebraic. S is (basic closed) semi-algebraic if $S = M(h_1, \dots, h_s)$ for some polynomials $h_1, \dots, h_s \in \mathbb{R}[X_1, \dots, X_n]$. The approximating hierarchy is guaranteed to converge if the quadratic module $M(h_1, \dots, h_s)$ is Archimedean. This in turn is the case if f or $-g_i$ are coercive, related sub- and suplevel sets are compact.

2.4. Cuts from valid linear inequalities

In this section we approach the question on how to generate linear cuts from given valid linear inequalities. Again, we consider the problem in more generality by not restricting our considerations to the feasible set F or F intersected with a sublevel set, but more generally to some deputy set $S \subset \mathbb{R}^n$. Our focus is on the all-integer case $\mathcal{I} = [n]$.

2.4.1. A single valid linear inequality

The first approach for the all-integer variant is based on the following observation that allows to cut off q if it is sufficiently exposed and a tight valid inequality with rational data is available. As a notation, let $\gcd(a)$ of $a \in \mathbb{Z}^n$ denotes the greatest common divisor of a.

Proposition 2.15. *Let $S \subset \mathbb{R}^n$ and $(a^T x \leq b)$ be a valid inequality for S, where $a \in \mathbb{Z}^n$ and $b \in \mathbb{R}$. Then*

$$\left(\frac{a}{\gcd(a)}\right)^T x \leq \left\lfloor \frac{b}{\gcd(a)} \right\rfloor \quad \text{for } x \in S \cap \mathbb{Z}^n. \tag{2.10}$$

Prior to the proof, let us motivate division by $\gcd(a)$ in Proposition 2.15 before rounding down by showing that it strengthens the cut.

Lemma 2.16. *Let $b \in \mathbb{R}$ and $g \in \mathbb{N}$. Then*

$$\left\lfloor \frac{b}{g} \right\rfloor \leq \frac{\lfloor b \rfloor}{g}.$$

Proof. Clearly, $\lfloor b/g \rfloor \leq b/g$. Thus, $g\lfloor b/g \rfloor \leq b$, and since the right hand side is integer, flooring on both sides yields

$$g\lfloor b/g \rfloor = \lfloor g\lfloor b/g \rfloor \rfloor \leq \lfloor b \rfloor$$

and hence division by g gives $\lfloor b/g \rfloor \leq \lfloor b \rfloor / g$, proving the claim. □

For the proof of Proposition 2.15, let us recall the following fact:

Lemma 2.17 (Integer hull of a rational half-space, see, e.g., p. 15 in [Eis00])**.** *Let $a \in \mathbb{Z}^n$ with $\gcd(a) = 1$ and $b \in \mathbb{R}$. Consider the polyhedron*

$$P := P(a, b) = \{x \in \mathbb{R}^n : a^T x \leq b\}.$$

Then

$$P_I = \{x \in \mathbb{R}^n : a^T x \leq \lfloor b \rfloor\}$$

We can now prove the proposition.

Proof of Proposition 2.15. Let $(a^T x \leq b)$ be a valid inequality for S. Equivalently,

$$\left(\frac{a}{\gcd(a)}\right)^T x \leq \frac{b}{\gcd(a)} \quad \text{for } x \in S. \tag{2.11}$$

Put $P := P(a, b)$. Since $\gcd(a/\gcd(a)) = 1$, Lemma 2.17 applies to the half-space defined by the valid inequality (2.11). On the other hand, since the inequality is valid, $S \subset P$, hence

$$\operatorname{conv}(S \cap \mathbb{Z}^n) \subset \operatorname{conv}(P \cap \mathbb{Z}^n) = P_I = P\left(\frac{a}{\gcd(a)}, \left\lfloor \frac{b}{\gcd(a)} \right\rfloor\right).$$

Hence, the inequality in (2.10) is valid for $S \cap \mathbb{Z}^n$. The claim follows. □

This gives the first linear cut result.

Proposition 2.18. *Let $S \subset \mathbb{R}^n$. Let $(a^T x \leq b)$ be a valid inequality for S, $a \in \mathbb{Z}^n$ with $\gcd(a) = 1$, $b \in \mathbb{R}$. If $q \in S$ satisfies $a^T q > \lfloor b \rfloor$, then $(a^T x \leq \lfloor b \rfloor)$ is a linear cut that cuts q from $S \cap \mathbb{Z}^n$.*

Proof. This follows readily from Proposition 2.15. □

We consider now how the presence of additional linear constraints can be exploited to generate a cut.

2.4.2. Presence of nonnegativity constraints

If S is moreover contained in the positive orthant, e.g. if the constraint $x \geq 0$ is present in the description of S, then for separating $q \in S$ from $S \cap \mathbb{Z}^n$, given a valid inequality $(a^T x \leq b)$ for S, it is enough to separate q from the integer hull of

$$P_+(a, b) = \{x \in \mathbb{R}^n : a^T \leq b, x \geq 0\}.$$

It seems that no explicit description of the integer hull of $P_+(a, b)$ is available. This is not too surprising: In [Eis00], the explicit linear description of the integer hull of a half-space (Lemma 2.17) and the integer hull of the intersection of two half-spaces is outlined (p. 15 in [Eis00]).[4] However, Eisenbrand concludes (p. 16 in [Eis00]) "There does not seem to exist an elementary method to construct the linear description of the integer hull formed by three or more half spaces in polynomial time. It is possible though with an application of Lenstra's method ([LJ83]) as proposed by Cook, Hartmann, Kannan & McDiarmid ([CHKM92])."

This motivates to look at a more elementary approach. Since we have the additional constraint $x \geq 0$, we do not need to restrict to the case of half-spaces with integer normal as in Proposition 2.15 , but may round down the coefficients. Formally, we have the following fact, which uses the well-known key observation leading towards Gomory cuts:

[4]Note that $P_+(a, b)$ is an intersection of $n + 1$ half-spaces.

Proposition 2.19. *Let $S \subset \mathbb{R}^n_{\geq 0}$ and let $(a^T x \leq b)$ be valid for S. Then*

$$(\lfloor a \rfloor)^T x \leq \lfloor b \rfloor \quad \text{for } x \in S \cap \mathbb{Z}^n.$$

Proof. Let $x \in \mathbb{R}^n_{\geq 0}$. Hence we find $\sum_{i=1}^n (\lfloor a_i \rfloor - a_i) x_i \leq 0$. On the other hand, if x is also in S, we have the inequality $\sum_{i=1}^n a_i x_i \leq b$, and adding both yields

$$\sum_{i=1}^n \lfloor a_i \rfloor x_i \leq b \qquad \text{for } x \in S.$$

Since $\lfloor a_i \rfloor$ are integers and $\mathbb{Z}$ is a ring, the sum on the left is integer for integer x. Hence

$$\sum_{i=1}^n \lfloor a_i \rfloor x_i \leq \lfloor b \rfloor \quad \text{for } x \in S \cap \mathbb{Z}^n,$$

proving the claim. □

Given our valid inequality $(a^T x \leq b)$, we saw in (2.10) and Lemma 2.16 that it can be advantageous to scale it before rounding the inequality down. This idea is formalized in the following corollary.

Corollary 2.20. *Let $S \subset \mathbb{R}^n_{\geq 0}$ and let $(a^T x \leq b)$ be valid for S. Then*

$$(\lfloor ua \rfloor)^T x \leq \lfloor ub \rfloor \quad \text{for } x \in S \cap \mathbb{Z}^n \text{ and } u \in \mathbb{R}_{>0}.$$

Proof. Let $u \in \mathbb{R}_{>0}$. The inequality $(a^T x \leq b)$ is valid for S if and only if $((ua)^T x \leq ub)$ is valid for S. The claim follows from Proposition 2.19. □

This yields the next linear cut result.

Proposition 2.21. *Let $S \subset \mathbb{R}^n_{\geq 0}$ and $q \in S$. Let $(a^T x \leq b)$ be a valid inequality for S. If there is $u \in \mathbb{R}_{>0}$ such that q satisfies*

$$(\lfloor ua \rfloor)^T q > \lfloor ub \rfloor \tag{2.12}$$

then $(\lfloor ua \rfloor)^T x \leq \lfloor b \rfloor)$ is a linear cut that cuts q from $S \cap \mathbb{Z}^n$.

Now if we are given $q \in S$ that we wish to cut off with a linear cut, and since Corollary 2.20 holds for all $u \in \mathbb{R}_{>0}$, it is only natural to ask for the choice of u with the "deepest cut" for q. Let us formulate this an optimization problem. Since we have a cut if and only if $(\lfloor ua \rfloor)^T q - \lfloor ub \rfloor$ is positive, we maximize this expression. It turns out that the following optimization problem models this task (Section 2 in [FL07]).

$$\begin{aligned} \max \quad & z^T q - y \\ & z_i \leq u a_i, \quad i \in [n] \\ & ub - 1 < y \\ & u > 0 \\ & (z, y) \in \mathbb{Z}^n \times \mathbb{Z} \\ & u \in \mathbb{R} \end{aligned} \tag{2.13}$$

The variables can be motivated as follows: Essentially, $z_i = \lfloor ua_i \rfloor$ and $y = \lfloor ub \rfloor$. The authors show [FL07] that q can be cut off if and only if the maximum of Program (2.13) is positive. Hence, we find:

Proposition 2.22. *Let $S \subset \mathbb{R}^n_{\geq 0}$ and $q \in S$. Let $(a^T x \leq b)$ be a valid inequality for S. There is $u \in \mathbb{R}_{>0}$ which cuts off q as in* (2.12) *if and only if the Program* (2.13) *has a positive maximum.*

We close this chapter with a note on possible directions of further research. Suppose some of the constraint functions in MINLP are linear. Extending the idea of Proposition 2.22, it makes sense to additionally use these linear constraints for the cut generation by incorporating them in a separation program similar to (2.13). This amounts to separation from the Chvátal closure and is also treated in [FL07].

Another interesting question is how linear cuts can be generated in the presence of mixed-integer constraints. The associated separation problem from the so-called MIR closure is, however, more difficult, since the objective of the auxiliary program is nonlinear [DGL10].

3. Norm bounds on optimal solutions

This chapter is all about the computation of upper bounds on the norm of optimal solutions of MINLP. This idea is formalized in the concept of a norm bound and uses a known feasible point. Computer experiments show that they work in practice, and we indicate that there are other areas in mathematics for which they can be used.

Section 3.1 gives our definition of norm bounds. We present a result from the literature for polynomial objective functions with positive definite leading form, a strong form of coercivity, that can be interpreted as a norm bound. For this class of objective functions, we give a new norm bound. It can be proved that our bound is never worse and strictly better in dense instances. Our norm bound makes extensive use of lower bounds on the homogeneous components of the objective restricted to a sphere, and we discuss in detail how such lower bounds can be derived. In this section we also explore how integrality information can be used to tighten the norm bounds.

Section 3.2 is about special cases and applications to other fields. It starts with a discussion of the convex quadratic case. The applications we give are so-called search bounds for Diophantine equations and a ball containing a real variety, both illustrated with an example.

Section 3.3 ends the chapter, where we evaluate our norm bounds on random instances. We start this with a discussion on how to generate polynomials with positive definite leading form, a necessary input to our norm bound and the one from the literature. Then, we proceed by comparing the volume of the corresponding norm balls. The experiments show that our norm bound outperforms the norm bound from the literature by orders of magnitude.

3.1. Introducing norm bounds

If a norm ball that contains all optimal solutions of MINLP is known, the problem is accessible to branch and bound, at least in the all-integer case. This is the primary motivation for this chapter.

A ball containing all optimal solutions corresponds to a valid nonlinear inequality for the set of optimal solutions. In this regard they supply *norm bounds*. The name norm bound catches the notion that they provide a bound on the norm of all feasible solutions with a smaller objective value than a known feasible solution, a fortiori on all optimal solutions. The formal definition allows for a "reference point" $\bar{x}$ other than zero and is the following.

Definition 3.1. Let $q \in F_{\mathcal{I}}$, a norm $\|\cdot\|$ on $\mathbb{R}^n$ and $\bar{x} \in \mathbb{R}^n$ be given. A number $R(q, \bar{x}) \in \mathbb{R}$ is a *norm bound for f with respect to the feasible solution q, the norm $\|\cdot\|$ and the point $\bar{x}$* if the following holds: For all $x \in F_{\mathcal{I}}$,

$$f(x) \leq f(q) \quad \Longrightarrow \quad \|\bar{x} - x\| \leq R(q, \bar{x}). \tag{3.1}$$

If $\|\cdot\|$ is merely a seminorm, it is equally possible to define a *seminorm bound*, but in this work we focus on norm bounds. Also, most results are concerned with the case $\bar{x} = 0$, and we write $R(q)$ instead of $R(0, q)$. In this case, if no confusion seems possible, we will furthermore write R instead of $R(q)$ and also suppress the norm in the notation.

Norm bounds boil down to a valid inequality for MINLP with an additional constraint. Explicitly, a norm bound $R(\bar{x}, q)$ gives rise to the valid inequality $\|\bar{x} - x\| \leq R(\bar{x}, q)$ for the program

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g_k(x) \geq 0, && k = 1, \ldots, r, \\ & f(q) \geq f(x), \\ & x_i \in \mathbb{Z}, && i \in \mathcal{I}, \\ & x \in \mathbb{R}^n. \end{aligned} \tag{3.2}$$

Hence, any feasible solution with norm larger than R cannot be optimal, as the objective value is worse than $f(q)$, and said feasible solution may thus be discarded in the solution process.

The smallest possible norm bound, given $\bar{x} \in \mathbb{R}^n$, $q \in F_{\mathcal{I}}$ and a norm, is thus given by solving the mixed-integer nonlinear program

$$\begin{aligned} \max \quad & \|\bar{x} - x\| \\ \text{s.t.} \quad & g_k(x) \geq 0, && k = 1, \ldots, r, \\ & f(q) \geq f(x), \\ & x_i \in \mathbb{Z}, && i \in \mathcal{I}, \\ & x \in \mathbb{R}^n. \end{aligned} \tag{3.3}$$

Since this program can be difficult to solve, we compute upper bounds on the optimal value of Program (3.3) in this chapter. We note that any finite upper bound on the

optimal value of Program (3.3) is a norm bound for f with respect to q, the norm $\|\cdot\|$ and $\bar{x}$.

A suitable form of coercivity

It turns out that it is quite easy to compute norm bounds on MINLP if f satisfies a strong type of coercivity condition, namely, if f has a positive definite leading form. In the remainder of Section 3.1, we assume that f has a positive definite leading form (and that a feasible solution is known). In other words, we impose a growth condition on f, which is a stronger form of coercivity as the former implies the latter, cf. Proposition 1.32.

Now let $f \in \mathbb{R}[\underline{X}]$ have a positive definite leading form f_d. Thus, by (1.3), this is equivalent to the minimum of f_d on the sphere $\mathbb{S}_p^{n-1}$, given by $c_d^* = \min_{x \in \mathbb{S}_p^{n-1}} f_d(x)$, being positive. In the following we do not assume knowledge of c_d^*, which may be difficult to compute as deciding nonnegativity of a polynomial is NP hard in general (Theorem 1.38), but assume knowledge of a lower bound c_d on the minimum with $0 < c_d \leq c_d^*$ instead. It is then possible to give norm bounds.

3.1.1. Norm bounds, old and new

A norm bound from the literature

We only found one result in the literature that can be interpreted as a norm bound. The norm under consideration in the result is the 2-norm; the result itself is a special case of a more general theorem, which assumes the lower bound c_d on c_d^* holds on $\mathbb{S}_2^{n-1}$ and that $q = 0$ is feasible.

Theorem 3.2 (see Conclusions 5.5 in [Mar03])**.** *Let $f \in \mathbb{R}[\underline{X}]$ satisfy $f_d(x) \geq c_d > 0$ for all $x \in \mathbb{S}_2^{n-1}$. Put*

$$R_{lit} := \max\left(1, \frac{1}{c_d} \sum_{j=1}^{d-1} \|f_j\|_1\right) \tag{3.4}$$

Then $f(x) \leq f(0)$ implies $\|x\|_2 \leq R_{lit}$ for all $x \in \mathbb{R}^n$.

Laurent gives a more elementary proof for Marshall's bound (3.4) by showing $f(x) > f(0)$ for $\|x\|_2 > R_{\text{lit}}$ directly, see Lemma 7.12 in [Lau09] As 0 is not necessarily a feasible solution, we generalize her argument so that it takes the form of (3.1) (with $\bar{x} = 0$). As the constant term of f does not enter R_{lit}, we have to modify the estimate in [Lau09].

Theorem 3.3. *Let $f \in \mathbb{R}[\underline{X}]$ satisfy $f_d(x) \geq c_d > 0$ for all $x \in \mathbb{S}_2^{n-1}$. Also, let $q \in \mathbb{R}^n$. Put*

$$R'_{lit} := \max\left(1, \frac{1}{c_d}\left(f(q) - f(0) + \sum_{j=1}^{d-1} \|f_j\|_1\right)\right). \tag{3.5}$$

Then $f(x) \leq f(q)$ implies $\|x\| \leq R'_{lit}$ for all $x \in \mathbb{R}^n$.

Proof. Introduce the notation $f_{<d} := f - f_d$. Let $f = \sum_\alpha a_\alpha X^\alpha$ and $x \in \mathbb{R}^n$ with $f(x) \leq f(q)$. If $\|x\|_2 \leq 1$, we are done, so we may assume $\|x\|_2 > 1$. Then

$$\begin{aligned} c_d \cdot \|x\|_2^d \leq f_d\left(\frac{x}{\|x\|_2}\right)\|x\|_2^d = f_d(x) = f(x) - f_{<d}(x) \leq f(q) - f_{<d}(x) \\ = f(q) - f(0) - \sum_{0<|\alpha|<d} a_\alpha x^\alpha \leq f(q) - f(0) + \sum_{0<|\alpha|<d} |a_\alpha||x^\alpha| \end{aligned}$$

As $|x^\alpha| \leq \|x\|_2^{|\alpha|}$ for $\|x\|_2 \geq 1$, division by $\|x\|_2^{d-1}$ yields

$$\begin{aligned} c_d \cdot \|x\|_2 &\leq \frac{f(q) - f(0)}{\|x\|_2^{d-1}} + \sum_{0<|\alpha|<d} |a_\alpha| \frac{\|x\|_2^{|\alpha|}}{\|x\|_2^{d-1}} \\ &\leq f(q) - f(0) + \sum_{0<|\alpha|<d} |a_\alpha| = f(q) - f(0) + \sum_{j=1}^{d-1} \|f_j\|_1. \end{aligned}$$

□

It is immediate that (3.5) specializes to (3.4) for $q = 0$.

A new bound

However, for non-sparse polynomials, this bound may get quite large. Within branch and bound approaches it is crucial to find a small bound R to reduce the number of feasible solutions – scaling R by a constant $C > 0$, the number of integer points that satisfy the norm bound scales with a factor of (roughly) C^n. We hence suggest a different approach: In the following theorem, we still compute $R \geq 0$ with $f(x) > f(q)$ for $\|x\| > R$, but instead of bounding all homogeneous components simultaneously, we compute constants c_j such that

$$c_j \leq c_j^* = \min_{x \in \mathbb{S}_p^{n-1}} f_j(x)$$

on a sphere $\mathbb{S}^{n-1}$ corresponding to a given norm, that is, $\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}$.

Theorem 3.4. *Let $f \in \mathbb{R}[X_1, \ldots, X_n]$ with $\deg f = d > 0$ and $q \in \mathbb{R}^n$. Moreover, let $\|\cdot\|$ be a norm on $\mathbb{R}^n$ with unit sphere $\mathbb{S}^{n-1}$. There are $c_j \in \mathbb{R}$ with*

$$f_j(x) \geq c_j \quad \text{for all} \quad x \in \mathbb{S}^{n-1},\ j \in [n].$$

Suppose that $c_d > 0$. Define a univariate polynomial $P : \mathbb{R} \to \mathbb{R}$ via

$$P(\lambda) := \sum_{j=0}^{d} c_j \lambda^j,$$

where $c_0 := f(0) - f(q)$. Then P has nonnegative real roots, and denote the largest of them by $R(q)$.

1. *Then $f(x) \leq f(q)$ implies $\|x\| \leq R(q)$ for all $x \in \mathbb{R}^n$.*

Suppose further that x is integer and $\|\cdot\|$ is a p-norm for some $p \in \mathbb{N}$.

2. *Then $f(x) \leq f(q)$ implies*

$$\|x\|_p \leq \sqrt[p]{\lfloor R(q)^p \rfloor},$$

as well as

$$|x_i| \leq \lfloor R(q) \rfloor \quad \text{for } i \in [n].$$

Proof. We begin with Claim 1. By compactness of the sphere and continuity, every f_j is bounded below by some $c_j \in \mathbb{R}$. We observed in (1.3) that $c_d > 0$ implies positive definiteness of f_d, hence d is even. Using homogeneity, we find for $x \in \mathbb{R}^n$ that $f_j(x) \geq c_j \|x\|^j$, $j \in [n]$, and

$$f(x) - f(q) = \sum_{j=0}^{d} f_j(x) - f(q) = \sum_{j=1}^{d} f_j(x) + f(0) - f(q) \geq \sum_{j=0}^{d} c_j \|x\|^j = P\left(\|x\|\right). \quad (3.6)$$

This implies $P(\|q\|) \leq 0$. The assumption $c_d > 0$ also yields $\lim_{\lambda \to +\infty} P(\lambda) = +\infty$. These two observations, together with the intermediate value theorem, imply that P has nonnegative real roots as well as $P(\lambda) > 0$ for $\lambda > R(q)$. Thus, eq. (3.6) forces $f(x) > f(q)$ for $\|x\| > R(q)$.

To see Claim 2, suppose further that x is integer and $\|\cdot\| = \|\cdot\|_p$ for some $p \in \mathbb{N}$. Let $x \in \mathbb{R}^n$ with $f(x) \leq f(q)$. We have just seen that then $\|x\|_p \leq R(q)$ holds, equivalently, $\sum_{i=1}^n |x_i|^p \leq R(q)^p$. The p-th power of the p-norm is integrality preserving, hence

$$\sum_{i=1}^{n} |x_i|^p \leq \lfloor R(q)^p \rfloor.$$

The claim $\sqrt[p]{\sum_{i=1}^n |x_i|^p} \leq \sqrt[p]{\lfloor R(q)^p \rfloor}$ follows. Finally, since $\|x'\|_\infty \leq \|x'\|_p$ for all $x' \in \mathbb{R}^n$, we have $\|x\|_\infty \leq R(q)$, or $|x_i| \leq R(q)$ for $i \in [n]$. Since the modulus is integrality preserving, the last claim follows. □

Note that Statement (2) of Theorem 3.4 shows how integrality can be used to tighten the bounds. It turns out that there are more sophisticated arguments that, based on integrality, can be used to strengthen the norm bound further. We defer this discussion to Section 4.3.

Let us state that the larger the c_j the smaller the resulting norm bound $R(q)$.

Proposition 3.5. *Let $P = \sum_{j=0}^n c_j \lambda^j$, $\tilde{P} = \sum_{j=0}^n \tilde{c}_j \lambda^j$, such that $c_j \geq \tilde{c}_j$, and define $R(q)$ and $\tilde{R}(q)$ as in Theorem 3.4. Then $R(q) \leq \tilde{R}(q)$.*

Proof. W.l.o.g., it suffices to consider the case that $c_j = \tilde{c}_j$ for $j \neq k$ and $c_k > \tilde{c}_k$ for some $k \in \{1, \ldots, n\}$. Now $P(\lambda) - \tilde{P}(\lambda) = (c_k - \tilde{c}_k)\lambda^k > 0$ for $\lambda > 0$ and by assumption on c_k, $\tilde{c}_k$. Thus, $P(\lambda) > \tilde{P}(\lambda)$ for $\lambda > 0$, hence $R(q) < \tilde{R}(q)$ – unless $\tilde{R}(q) = 0$. In this case $R(q) = \tilde{R}(q) = 0$. □

Proposition 3.5 has two consequences: The norm bound gets better the smaller the incumbent upper bound $f(q)$ on the minimum gets, since $c_0 = f(0) - f(q)$. Moreover, the smallest norm bound computable with the method of Theorem 3.4 is attained if $c_j = c_j^*$, $j \in [n]$, where $c_j^* = \min_{x \in \mathbb{S}} f_j(x)$.

Comparison of the norm bounds

Before we present different methods of computing valid c_j, we compare R and R_{lit}. In the experiments in Section 3.3, we show that our norm bound R is drastically smaller than R_{lit}. We prove in the next proposition that our norm bounds are never larger and, except for special cases, are actually strictly smaller than the bound from the literature. To this end let us fix a basic estimate.

Lemma 3.6. *Let $f \in \mathbb{R}[X_1, \ldots, X_n]$. Fix $p \in [1, \infty]$, and denote by f^* the optimal value of the program*

$$\begin{aligned} \min \quad & f(x) \\ & x \in \mathbb{S}_p^{n-1}. \end{aligned}$$

Then

$$f^* \geq -\|f\|_1.$$

Proof. Write f in multi-index notation as

$$f = \sum_{\alpha \in A(f)} a_\alpha X^\alpha.$$

As $\|x\|_p \leq 1$ implies $\|x\|_\infty \leq 1$ and hence $|x^\alpha| \leq 1$ for $\alpha \in \mathbb{N}_0^n$, one has

$$f(x) = \sum_{\alpha \in A(f)} a_\alpha x^\alpha \geq \sum_{\alpha \in A(f)} -|a_\alpha||x^\alpha| \geq \sum_{\alpha \in A(f)} -|a_\alpha| = -\|f\|_1 \tag{3.7}$$

for $x \in \mathbb{S}_p^{n-1}$. □

We can now show the relation between R and R_{lit}.

Proposition 3.7. *Let f with $\deg f = d > 0$ and $c_d \in \mathbb{R}$ with*

$$f_d(x) \geq c_d > 0 \quad \text{for all } x \in \mathbb{S}_2^{n-1},$$

and $q \in \mathbb{R}^n$. Compute $R \in [0, \infty)$ as in Theorem 3.4 for

$$c_j := -\|f_j\|_1, \ j \in [d-1], \ \text{and } c_0 := f(q) - f(0), \tag{3.8}$$

and compute $R'_{lit} \in [1, \infty)$ as in (3.5). Then $R \leq R'_{lit}$. If moreover $d > 2$ and there is a coefficient $a_\alpha \neq 0$ of f with $|\alpha| < d-1$, then $R < R'_{lit}$ for $R \neq 1$ and $R = R'_{lit}$ for $R = 1$.

Proof. Observe that by Lemma 3.6, the numbers $c_j = -\|f_j\|_1$ in (3.8) are indeed valid lower bounds for f_j on $\mathbb{S}_2^{n-1}$, $j \in [n-1]$. We prove the case $d > 2$ and $a_\alpha \neq 0$ for some α with $|\alpha| < d-1$. The claim obviously holds in case $R < 1$. For the cases $R = 1$ and $R > 1$, define $P(\lambda) = \sum_{j=0}^{d} c_j \lambda^j$ as before and let $\tilde{P}(\lambda) = c_d\lambda^d + \left(\sum_{j=0}^{d-1} c_j\right)\lambda^{d-1}$. Then we have

$$P(\lambda) > \tilde{P}(\lambda) \text{ for } \lambda > 1, \quad P(\lambda) = \tilde{P}(\lambda) \text{ for } \lambda = 1 \tag{3.9}$$

as $c_j \leq 0$ for $j = 1, \ldots, d-1$ and one $c_k < 0$ for some $k \in \{1, \ldots, d-2\}$ by the assumption on a_α. By definition, the largest nonnegative real root of P is R, and the largest nonnegative real root $\tilde{R}$ of $\tilde{P}$ is

$$\tilde{R} = -\frac{1}{c_d}\sum_{j=0}^{d-1} c_j = \frac{1}{c_d}\sum_{0<|\alpha|<d} |a_\alpha| = \frac{1}{c_d}\left(f(q) - f(0) + \sum_{j=1}^{n-1}\|f_j\|\right)$$

and, by definition, $R'_{\text{lit}} = \max(1, \tilde{R})$. If $R = 1$, we infer from (3.9) that $0 = P(1) = \tilde{P}(1)$, so $R'_{\text{lit}} = 1$. In case $R > 1$, we infer from (3.9) that $0 = P(R) > \tilde{P}(R)$, so $R < R'_{\text{lit}}$ as $\tilde{P}(\lambda) \to +\infty$ for $\lambda \to +\infty$. The proof for the two remaining cases, $d = 2$ or all $a_\alpha = 0$ for $|\alpha| < d-1$, is similar as $P = \tilde{P}$ in these cases. □

3.1.2. Computation of the constants c_j

We now present different ways of computing bounds c_j on

$$c_j^* = \min_{x \in \mathbb{S}^{n-1}} f_j(x).$$

The first approach is the only one that yields c_j^* – the others yield, in virtually all cases, only lower bounds $c_j \leq c_j^*$ – and consists in solving a hard continuous problem. The second approach gives lower bounds by sos programming. In the third approach we consider sos-free methods.

Constrained polynomial optimization

By Proposition 3.5, the best possible bounds are found by solving the constrained polynomial optimization problem

$$\begin{aligned} c_j^* = \min \quad & f_j(x) \\ & x \in \mathbb{S}^{n-1} \end{aligned}$$

for $j \in [n]$, for a sphere $\mathbb{S}^{n-1}$ corresponding to a norm on $\mathbb{R}^n$. This problem is NP-hard in general: If $c_j^* > 0$, then $f_j > 0$ by homogeneity. So this approach involves solving several hard problems, as, e.g., deciding positive definiteness of f_j, j even, is NP-hard (Theorem 1.37).

Approximations via sos programming

Standard approximation of the constrained program. If the sphere is given by a p-norm with $p \in 2\mathbb{N}$, then the arguably easiest way to find the c_j, $j \in [n]$, by sos programming is to minimize f_j on the sphere $\mathbb{S}_p^{n-1}$. To this end, consider the following program with parameter $k' \in \mathbb{N}$:

$$\begin{aligned} \max \quad & y \\ \text{s.t.} \quad & f_j - y - q \cdot \left(1 - \sum_{i=1}^n X_i^p\right) \in \Sigma \\ & q \in \mathbb{R}[\underline{X}],\ \deg q \leq k' \\ & y \in \mathbb{R}. \end{aligned} \tag{3.10}$$

Note that this is a valid sos program by (1.15). We use an arbitrary polynomial q instead of an sos polynomial as we deal with equality constraints, which is a standard technique in sos programming, see, e.g., Section 2 in [PPP02]. It turns out that the optimal values of the hierarchy converge to the constant c_j^*.

Proposition 3.8. *For $f \in \mathbb{R}[X_1, \ldots, X_n]$ and $j \in [n]$, consider the hierarchy* (3.10) *and denote the optimal values by $y^{(k')}$ and let*

$$c_j^* = \min_{x \in \mathbb{S}_p^{n-1}} f_j(x).$$

Then

$$y^{(k')} \nearrow c_j^*$$

for $k' \to \infty$.

Proof. For $p \in 2\mathbb{N}$, the constraint $\|x\|_p = 1$ is equivalent to two semi-algebraic constraints given by $g_1 := 1 - \sum_{i=1}^n X_i^p$ and $g_2 := \sum_{i=1}^n X_i^p - 1$. Now, note that the usual quadratic module containment constraint in (1.19) rewrites via

$$\sigma_1 g_1 + \sigma_2 g_2 = (\sigma_1 - \sigma_2) g_1 = q g_1,$$

some $q \in \mathbb{R}[\underline{X}]$. Note that moreover any polynomial can be written as the *difference* of sums of squares, e.g. using $4q = (q+1)^2 - (q-1)^2$.

As $p \in 2\mathbb{N}$, the quadratic module $M(1 - \sum_{i=1}^n X_i^p, \sum_{i=1}^n X_i^p - 1)$ is Archimedean by Theorem 1.17 (4). Hence, from Corollary 1.27, the optimal objective values of (3.10) converge, for $k' \to \infty$, to c_j^*. □

For completeness, let us note that this is, in principle, also possible for odd p, but requires to consider 2^n constraints as the modulus occurring in odd p-norms cannot be modeled as easily as for even p, that is, by squaring. We will not pursue this idea further.

A lower bound for one of the forms. If the norm under consideration is $\|\cdot\|_p$, where $p \in 2\mathbb{N}$ and $p \leq \deg(f)$, a lower bound on the form f_p can be computed via the program

$$\begin{aligned} \max \quad & \gamma \\ \text{s.t.} \quad & f_p - \gamma \cdot \sum_{i=1}^{n} X_i^p \in \Sigma, \end{aligned}$$

cf. [Nie12].

Sos-free methods

In the remainder we present lower bounds c_j on the best constants c_j^* that do not rely on sos programming. We start with a somewhat rough estimate which we successively refine.

Using a basic estimate. We saw in (3.7) that the choice

$$c_j = -\|f_j\|_1 \tag{3.11}$$

gives valid lower bounds for any $p \in [1, \infty]$. However, this bound is rather rough and only useful for the lower order forms, that is those f_j with $j < d$.

Improvement by term-wise optimization. The first improvement on (3.11) comes from the observation that we can replace the estimate $|x^\alpha| \leq 1$ on $\mathbb{S}_p^{n-1}$ in Lemma 3.6 with $|x^\alpha| \leq \hat{x}^\alpha$, where $\hat{x}$ is a continuous maximizer of the function $\mathbb{S}_p^{n-1} \to \mathbb{R}$, $x \mapsto x^\alpha$. We give a closed form for $\hat{x}$ in the following lemma.

Lemma 3.9. *Let $0 \neq \alpha \in \mathbb{N}_0^n$ and $p \in [1, \infty)$. Then, the monomial X^α attains its maximum on $\mathbb{S}_p^{n-1}$ at $\hat{x}$ with coordinates*

$$\hat{x}_i = \sqrt[p]{\frac{\alpha_i}{\sum_{i=1}^n \alpha_i}}, \quad i = 1, \ldots, n. \tag{3.12}$$

Proof. Maximizers exist by compactness of the sphere. At first, we show the auxiliary claim that it suffices to prove the assertion for $\alpha_i \geq 1$ for all i. To this end, denote the zero entries of α by $z_1, \ldots, z_r$ and the positive entries of α by $p_1, \ldots, p_s$. For every maximizer $\tilde{x}$, $\tilde{x}_{z_1} = \ldots = \tilde{x}_{z_r} = 0$. Hence, we may as well optimize X^α over the set

$$\begin{aligned} &\{x \in \mathbb{R}^n : \|x\|_p = 1,\ x_{z_1} = \ldots = x_{z_r} = 0\} \\ =&\{x \in \mathbb{R}^n : \|(x_{p_1}, \ldots, x_{p_s})\|_p = 1, x_{z_1} = \ldots = x_{z_r} = 0\} \\ \cong&\{x \in \mathbb{R}^s : \|x\|_p = 1\} = \mathbb{S}^{s-1} = \mathbb{S}^{n-1-r}. \end{aligned}$$

This means optimization of X^α over $\mathbb{S}_p^{n-1}$ is equivalent to optimization of $X^{(\alpha_{p_1}, \ldots, \alpha_{p_s})}$ over $\mathbb{S}_p^{n-1-r}$, and the auxiliary claim is proved. By symmetry of the sphere, we may further assume that $x \geq 0$. Even more, we may exclude equality since by the auxiliary

claim, all $\alpha_i > 0$, hence $\tilde{x}_i > 0$ for every maximizer $\tilde{x}$. So it suffices to find a maximizer of X^α over

$$\mathbb{S}^+ := \{x \in \mathbb{S}_p^{n-1} : x_1 > 0, \dots, x_n > 0\}$$

– and in this form, the problem is accessible to Lagrange multipliers: At a critical point $x \in \mathbb{S}^+$ there is $\lambda \in \mathbb{R}$ such that

$$\nabla(x^\alpha) = (\alpha_1 x_1^{\alpha_1 - 1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}, \dots, \alpha_n x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n - 1}) = \lambda(p x_1^{p-1}, \dots, p x_n^{p-1})$$

where we used the equivalence of $\|x\|_p = 1$ to $\sum_{i=1}^n x_i^p = 1$ if all $x_i > 0$. Equivalently,

$$\lambda = \alpha_1 x_1^{\alpha_1 - p} x_2^{\alpha_2} \cdots x_n^{\alpha_n} = \cdots = \alpha_n x_1^{\alpha_1} \cdots x_n^{\alpha_n - p}. \tag{3.13}$$

Fix some $k \in \{1, \dots, n\}$, and (3.13) gives $x_i^p = \frac{\alpha_i}{\alpha_k} x_k^p$ for $i \neq k$. As x suffices the constraint,

$$1 = x_k^p + \sum_{i \neq k} \frac{\alpha_i}{\alpha_k} x_k^p \iff x_k^p = \frac{\alpha_k}{\sum_i \alpha_i}.$$

This proves that the only critical point of X^α on $\mathbb{S}^+$ is $\hat{x}$, which must be the maximizer. □

Observation 3.10. *Denote by $\hat{x}_{(\alpha)}$ the maximizer of X^α on $\mathbb{S}_p^{n-1}$ as in (3.12). Hence for $x \in \mathbb{S}_p^{n-1}$ we have*

$$f_j(x) = \sum_{|\alpha|=j} a_\alpha x^\alpha \geq \sum_{|\alpha|=j} -|a_\alpha| \cdot (\hat{x}_{(\alpha)})^\alpha =: c_j. \tag{3.14}$$

This c_j is as least as large as the c_j from Proposition 3.7, since, for $0 \neq \alpha$, $(\hat{x}_{(\alpha)})^\alpha < 1$ – unless $X^\alpha \in \mathbb{R}[X_i]$ for some i, in which case $\hat{x}_{(\alpha)} = e_i$, the i-th unit vector, and thus $(\hat{x}_{(\alpha)})^\alpha = 1$.

Further improvement by orthant distinction. In a second improvement step, we consider all orthants separately to furthermore get rid of approximately half of the terms in the estimate in Lemma 3.6. This last approach for computing c_j is different to the ones before, as we actually compute 2^n norm bounds: We restrict f to each of the 2^n orthants

$$H_\tau = \{x \in \mathbb{R}^n : \tau_i x_i \geq 0\} \text{ for } \tau \in \{-1, 1\}^n \tag{3.15}$$

and compute a norm bound on minimizers for every $f|_{H_\tau}$. This has the advantage that, roughly speaking, we may neglect half of the terms of $f = \sum a_\alpha X^\alpha$. Also, minimization on H_τ can be reduced to minimization on $H_{(1,\dots,1)}$, i.e., the set of those $x \in \mathbb{R}^n$ with $x \geq 0$, as we shall see in a moment.

Introducing the notation $|a|^- = |\min(a, 0)|$ for $a \in \mathbb{R}$ and with $\hat{x}$ from (3.12), we have for every term

$$a_\alpha x^\alpha \geq -|a_\alpha|^- x^\alpha \geq -|a_\alpha|^- \hat{x}^\alpha$$

as $x \geq 0$, thus

$$f_j(x) = \sum_{|\alpha|=j} a_\alpha x^\alpha \geq \underbrace{\sum_{|\alpha|=j} -|a_\alpha|^- \hat{x}^\alpha}_{=:c_j^{(1,\dots,1)}}, \quad x \in \mathbb{S}_p^{n-1} \text{ and } x \geq 0,$$

which means about half of the coefficients are neglected in comparison to (3.14), if signs are distributed equally among the a_α. Now let $R^{(1,\dots,1)}$ be the largest real root of

$$q^{(1,\dots,1)}(\lambda) := c_d\lambda^d + \sum_{j=1}^{d-1} c_j^{(1,\dots,1)}\lambda^j.$$

The verbatim argument of Theorem 3.4 shows that $f(x) > f(q)$ for $\|x\|_p > R^{(1,\dots,1)}$ and $x \geq 0$. This gives norm bounds on minimizers in $H_{(1,\dots,1)}$. Bounding the norm of minimizers of f on H_τ, $\tau \in \{-1,1\}^n$, can be reduced to bounding the norm of minimizers on $H_{(1,\dots,1)}$ by a simple change of coordinates. To this end, let $\tau(x) = (\tau_1 x_1, \dots, \tau_n x_n)$, $x \in \mathbb{R}^n$, and f^τ be the polynomial

$$f^\tau(x) := f(\tau(x)) = \sum_\alpha a_\alpha \tau^\alpha x^\alpha, \quad \tau \in \{-1,1\}^n.$$

As $\tau^\alpha \in \{-1,1\}$, f and f^τ merely differ in the sign of their coefficients, and $f_d^\tau(x) \geq c_d$ still holds for $x \in \mathbb{S}_p^{n-1}$ as the sphere is τ-invariant, that is $\tau(\mathbb{S}_p^{n-1}) = \mathbb{S}_p^{n-1}$. Similarly to before, denote by R^τ the largest real root of

$$q^\tau(\lambda) = c_d\lambda^d + \sum_{j=1}^{d-1} c_j^\tau \lambda^j,$$

with $c_j^\tau = -|a_\alpha\tau^\alpha|^-\hat{x}^\alpha$ for $j \in [d-1]$ and $c_0 = f(0) - f(q)$. It is now clear that $f^\tau(x) > f(q)$ for $\|x\|_p > R^\tau$ and $x \geq 0$, equivalently, $f(x) > f(q)$ for $\|x\|_p > R^\tau$ and $x \in H_\tau$.

This results in more effort in the preprocessing, but reduces the number of feasible solutions.

3.2. Special cases and applications

In this section we consider norm bounds for the special case of a strictly convex quadratic objective and show an application of norm bounds to systems of polynomial equations.

3.2.1. The convex quadratic case

We consider norm bounds for the following variant of MINLP:

$$\begin{aligned} \min \quad & f(x) = x^T Q x + L^T x + c \\ \text{s.t.} \quad & x \in F_{\mathcal{I}} \end{aligned} \tag{MIQP}$$

for some $Q \in \mathbb{R}^{n \times n}$ with $Q \succ 0$, $L \in \mathbb{R}^n$ and $c \in \mathbb{R}$. It is a well-known fact that the condition $Q \succ 0$ is equivalent to f being strictly convex. The all-integer variant without constraint functions and linear constraints have been studied recently. We refer to the two articles [BCL12; BHS15] and the references therein for an introduction into this class.

It is well-known that a strictly convex quadratic has a unique continuous minimizer. The following characterization is useful for our purposes.

Lemma 3.11 (Lemma 2.1 in [BHS15])**.** *Let $f(x) = x^T Q x + L^T x + c$ be strictly convex. The unique continuous minimizer of f is given by*

$$\bar{x} = -\frac{1}{2} Q^{-1} L.$$

The optimal value is

$$f(\bar{x}) = c - \frac{1}{4} L^T Q^{-1} L.$$

Moreover,

$$f(x) = (x - \bar{x})^T Q (x - \bar{x}) + f(\bar{x}). \tag{3.16}$$

The sublevel sets of f are of the form

$$\mathcal{L}^f_{\leq}(z) = \sqrt{z - f(\bar{x})} E(Q, \bar{x}) \tag{3.17}$$

for $z \geq f(\bar{x})$ and empty otherwise.

Optimal constants for our norm bounds

The homogeneous components of the quadratic objective f are $f_2(x) = x^T Q x$, $f_1(x) = L^T x$ and $f_0 = c$. In contrast to the general case, computation of constants c_2 and c_1, bounding the homogeneous components on the sphere $\mathbb{S}_2^{n-1}$ from below, are easier to compute. In fact, the largest possible lower bound, i.e., the minimum c_2^* of all admissible bounds c_2, is given by the smallest eigenvalue of Q. This follows from the famous Rayleigh-Ritz Theorem (Theorem 1.6). In the quadratic case, c_1^* can easily be computed as well.

Proposition 3.12. *Let $L \in \mathbb{R}^n$. Then the optimal value c_1^* of*

$$\begin{aligned} \min \quad & L^T x \\ s.t. \quad & x \in \mathbb{S}_2^{n-1} \end{aligned}$$

satisfies $c_1^* = -\|L\|_2$.

Proof. We only prove the case $L \neq 0$. By the Cauchy-Schwarz inequality, for every $x \in \mathbb{S}_2^{n-1}$,

$$|L^T x| \leq \|L\|_2 \cdot \|x\|_2 = \|L\|_2$$

with equality if and only if x and L are parallel. This is the case if and only if $x = L/\|L\|_2$ or $x = -L/\|L\|_2$. The latter x minimizes the objective. □

We can now express the norm bound in closed form in case of a strictly convex quadratic objective by solving a quadratic equation.

Corollary 3.13. *Let $f(x) = x^T Q x + L^T x + c$ be strictly convex quadratic and $q \in \mathbb{R}^n$. The norm bound (for the Euclidean norm) as defined in Theorem 3.4 that arises from the best possible choice of c_1 and c_2 is given by*

$$R = -\frac{c_1^*}{2c_2^*} + \sqrt{\left(\frac{c_1^*}{2c_2^*}\right)^2 - \frac{f(q) - f(0)}{c_2^*}} = \frac{\|L\|_2}{2\lambda_{\min}(Q)} + \sqrt{\left(\frac{\|L\|_2}{2\lambda_{\min}(Q)}\right)^2 - \frac{f(q) - c}{\lambda_{\min}(Q)}}$$

Proof. The optimal choice (cf. Proposition 3.5) for c_1 is $c_1^* = -\|L\|_2$ by Proposition 3.12, the optimal choice for c_2 is $c_2^* = \lambda_{\min}(Q)$ by Theorem 1.6. Hence, the claim follows by solving $c_2^* \lambda^2 + c_1^* \lambda + f(q) - f(0) = 0$ for the largest nonnegative root, using $f(0) = c$. □

A further improvement: Optimal norm bounds for a convex objective

It turns out that smallest possible norm bound for a strictly convex quadratic objective function can actually be derived analytically. More specifically, given a feasible solution $q \in F_{\mathcal{I}}$, our result is an explicit formula for the maximal Euclidean distance of all feasible solutions to the unique continuous minimizer. Arguments from [BHS15] are central to the proof.

Theorem 3.14. *Let $f(x) = x^T Q x + L^T x + c$ be strictly convex and $q \in \mathbb{R}^n$. Denote the continuous minimizer by $\bar{x}$. The smallest possible R such that*

$$f(x) \leq f(q) \text{ implies } \|x - \bar{x}\|_2 \leq R$$

is given by

$$R = \sqrt{\frac{f(q) - f(\bar{x})}{\lambda_{\min}(Q)}} = \sqrt{\frac{(q - \bar{x})^T Q (q - \bar{x})}{\lambda_{\min}(Q)}}.$$

To prove the theorem, we need the following lemma.

Lemma 3.15 (see, e.g., Section 3.3 in [BHS15]). *Let $Q \in \mathcal{S}^n$, $Q \succ 0$, $x_0 \in \mathbb{R}^n$ and $r > 0$. Then*

$$E(Q, x_0) \subset B_r(x_0)$$

if and only if

$$r \geq \frac{1}{\lambda_{\min}(Q)}.$$

Proof of Theorem 3.14. Let $z = f(q)$ and $r \geq 0$. By Lemma 3.11, f has a unique continuous minimizer $\bar{x}$. Moreover, Lemmata 3.11 and 3.15 imply that

$$\mathcal{L}^f_{\leq}(z) = \sqrt{z - f(\bar{x})}E(Q, \bar{x}) \subset \sqrt{z - f(\bar{x})}B_r(\bar{x})$$

if and only if $r \geq \frac{1}{\lambda_{\min}(Q)}$. In other words, the smallest $R \geq 0$ such that $x \in \mathcal{L}^f_{\leq}(z)$ implies $x \in B_r(\bar{x})$ is given by $R = \frac{\sqrt{z-f(\bar{x})}}{\lambda_{\min}(Q)}$. □

The theorem has a nice consequence: The distance in norm of all integer minimizers towards the continuous minimizer can be bounded from above by an expression only depending on the eccentricity of the ellipsoid $E(Q)$, that is, the ratio of the smallest to the largest eigenvalue of Q, and the dimension.

Corollary 3.16. *Let $f(x) = x^TQx + L^Tx + c$ be strictly convex. Denote the continuous minimizer by $\bar{x}$. Then we have the following a priori norm bound for all integer minimizers x^*:*

$$\|x^* - \bar{x}\|_2 \leq \sqrt{\frac{n}{4} \cdot \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)}}$$

Proof. The point $\lfloor \bar{x} \rceil$ is a feasible solution to the unconstrained integer minimization problem with objective f. By Theorem 3.14,

$$\|x^* - \bar{x}\|_2 \leq \sqrt{\frac{(\lfloor \bar{x} \rceil - \bar{x})^TQ(\lfloor \bar{x} \rceil - \bar{x})}{\lambda_{\min}(Q)}}.$$

Then, by Rayleigh-Ritz (Theorem 1.6),

$$(\lfloor \bar{x} \rceil - \bar{x})^TQ(\lfloor \bar{x} \rceil - \bar{x}) \leq \lambda_{\max}(Q) \, \|\lfloor \bar{x} \rceil - \bar{x}\|_2^2 \leq \lambda_{\max}(Q) \left\| \left(\frac{1}{2}, \ldots, \frac{1}{2} \right) \right\|_2^2 \leq \lambda_{\max}(Q) \frac{n}{4},$$

and the claim follows. □

In the proof of Corollary 3.16, the point $\lfloor \bar{x} \rceil$, the rounded continuous minimizer of $f = x^TQx + L^Tx + c$, played a central role. It has been investigated when this point coincides with the integer minimizer of f, and sufficient conditions where given in [BHS15]:

Proposition 3.17 (Lemma 3.2 in [BHS15]). *Let $f = x^TQx + L^Tx + c$ be strictly convex and let $\bar{x}$ be its continuous minimizer. Then $\lfloor \bar{x} \rceil$ is the integer minimizer of f in either of the following cases:*

- *The matrix Q is diagonal.*
- *The sublevel sets of f satisfy a so-called quasi round property, namely, if*

$$\frac{1}{\lambda_{\min}(Q)} - \frac{1}{\lambda_{\max}(Q)} \leq \frac{\alpha(\bar{x})}{\sqrt{f(\lfloor \bar{x} \rceil) - f(\bar{x})}},$$

where

$$\alpha(\bar{x}) := \inf\{\|\bar{x} - y\|_2 : y \in \mathbb{Z}^n \setminus \{\lfloor \bar{x} \rceil\}\} - \inf\{\|\bar{x} - y\|_2 : y \in \mathbb{Z}^n\}.$$

3.2.2. Application to systems of polynomial equations

As a further application of norm bounds, we consider systems of polynomial equations in this section. It is a common approach to solve a system of polynomial equations

$$h_1(x) = 0, \ldots, h_s(x) = 0, \quad x \in \mathbb{K}^n,$$

with solutions restricted to, say, $\mathbb{K} \in \{\mathbb{Z}, \mathbb{Q}, \mathbb{R}\}$, by minimizing $f = h_1^2 + \ldots h_s^2$ over the integers, rationals or reals, respectively. If the minimum is 0 at some x, the equations have a solution at x; if the minimum is nonzero, there cannot be any solution.

Diophantine equations

As an example, does the system

$$\begin{aligned} -3x_1^3 + x_1^2x_2 - x_1^2 + 2x_1x_2 + x_1 - 2x_2^2 - 2x_2 + 4 &= 0 \\ 2x_2^3 + x_1x_2^2 + 4x_2 - 5 &= 0 \end{aligned}$$

possess an integer solution? Denote the polynomials in $\mathbb{Z}[X_1, X_2]$ on the left hand side in the first and second equation by h_1 and h_2, respectively, and consider $f := h_1^2 + h_2^2$. The homogeneous components of f are bounded from below on $\mathbb{S}_6^1$ by

$$(c_1, \ldots, c_6) = (-60.49, -13.03, -41.76, -7.85, -24.45, 2.59),$$

we found the values by solving (3.10) numerically.[1] Using the feasible point $q = 0$, the univariate polynomial $P(\lambda) = \sum_{j=1}^6 c_j\lambda^j$ has only two real roots: 0 and $R \approx 9.90$. Thus, by Theorem 3.4, integer minimizers exist and must be in the box $[-9, 9]^2$. Iterating over all integer points in the box one finds $f(x_1, x_2) = 0$ at $(x_1, x_2) = (-1, 1)$. From the perspective of number theory, our method provides *search bounds* on solutions of a system of Diophantine equations if the leading form of $f = \sum_{j=1}^s h_j^2$ is positive definite.

[1] See Section 3.3 for details on the software.

Bounds on algebraic varieties

Similarly to the systems of Diophantine equations, our bounds apply to real algebraic varieties: Given $h_1, \ldots, h_s \in \mathbb{R}[\underline{X}]$, the variety of the h_i is

$$V(h_1, \ldots, h_s) = \{x \in \mathbb{R}^n : h_1(x) = 0, \ldots, h_s(x) = 0\}.$$

If the leading form of $f = \sum_{j=1}^{s} h_j^2$ is positive definite, we may give a norm bound on all points of the variety. As an example, let us consider the polynomial system:

$$\begin{aligned} x^2 + y^2 + z^2 &= 1 \\ x^2 + z^2 &= y \\ x &= z \\ x, y, z &\in \mathbb{C} \end{aligned} \tag{3.18}$$

from Example 2, Section 2 § 8 in [CLO07]. Computing the c_j by solving (3.10) for $p = 2$ yields $(c_1, \ldots, c_4) = (0, -2.0, -0.77, 1.0)$ and gives us for $q = 0$ the value $R \approx 1.86$ as a 2-norm bound on all points in the variety. It is known that the variety consists of exactly four points: The system has two real and two complex solutions (x, y, z_i) with $z_i \in \{\pm\frac{1}{2}\sqrt{\pm\sqrt{5} - 1}\}$, where the real solutions suffice $\|(x, y, z)\|_2 = 1$ by (3.18). We conclude that in this case our bound is not far off.

3.3. Experimental comparison of norm bounds

To evaluate our norm bounds on random instances, we ran computer experiments.[2] In this section we outline the algorithm to decide positive definiteness and if applicable proceed to compute the norm bounds. We moreover elaborate the choice of a distribution from which we sample random instances. Later, in Section 6.3, we use such samples as objective functions to an unconstrained optimization problem and solve it by branch and bound.

3.3.1. Sampling, positive definiteness and norm bounds

Once we have agreed upon a distribution to sample our random polynomials from, we have to decide if the instance has a positive definite leading form and, if this is the case, proceed to compute the norm bounds. We present these steps in algorithmic form (Algorithm 1). In summary, we solve a hierarchy of sos programs to test positive definiteness. Either we can certify positive definiteness by (1.3) or extract a point x with $f_d(x) \leq 0$ or positive definiteness cannot be decided at this level of the hierarchy.[3] In case of positive definiteness, we compute the constants c_j with the methods discussed in Section 3.1 and proceed to compute a p-norm bound on all integer minimizers.

In the experiments, we computed the following five norm bounds:

a) the bound R_{lit} from the literature, defined in (3.4), denoted **Lit** in the plots,

b) the new norm bound with the rather rough estimate $c_j = -\|f_j\|$ from Proposition 3.7, denoted **Drct** in the plots,

c) the norm bound with the refined constants c_j from Observation 3.10, denoted **Mx** in the plots,

d) the orthant based bound as outline in Section 3.1.2, denoted **Or** in the plots,

e) the norm bound where the c_j are the optimal solutions of the sos program (3.10) with $k' = 2$, denoted **Sos** in the plots.

We compare the volume, i.e., the Lebesgue measure λ, of the resulting sets, as it approximates the number of integer points, i.e., potentially optimal solutions, contained

[2] We use MATLAB 2014b 64-bit (MATLAB is a registered trademark of The MathWorks Inc., Natick, Massachusetts), SOSTOOLS 3.00 [PAV+] to translate the sos programs into semidefinite programs and CSDP 6.1.0 [Bor99]/SDPT3 [TTT99] to solve the latter. The experiments were conducted on a GNU/Linux machine running on 2 Intel® Xeon®X5650 CPUs, 6 cores each, with a total of 96 GB RAM.

[3] We shortly discussed the extraction of x after Corollary 1.27; in our setup, this corresponds to a non-empty third return argument of SOSTOOLS's `findbound.m`-routine.

Algorithm 1 Norm bound on mixed-integer minimizers

input $f \in \mathbb{R}[X_1, \ldots, X_n]$ with $\deg f \in 2\mathbb{N}$, parameters $p \in 2\mathbb{N}$, $k'_{\max} \in \mathbb{N}_0$
$k' \leftarrow 0$
$c_d \leftarrow -\infty$
$x \leftarrow$ `NULL`
while $k' \leq k_{\max}$ and $c_d < 0$ and $x =$ `NULL` **do**
  solve program (3.10) for $j = d$ and parameter k'
  $c_d \leftarrow$ optimal value
  if corresponding optimal solution can be extracted **then**
    $x \leftarrow$ optimal solution
  end if
  $k' \leftarrow k' + 1$
end while
if $c_d < 0$ and $x \neq$ `NULL` **then** *// by Proposition 1.33,* $\inf_{x \in \mathbb{Z}^n} f(x) = -\infty$
  print f `has no mixed-integer minimizers.`
  output x
else if $c_d = 0$ and $x \neq$ `NULL` **then** *// as* $x \in \mathbb{S}_p^{n-1}$, $x \neq 0$, *so* f_d *is not positive def.*
  print `Cannot decide existence of mixed-integer minimizers.`
  output x
else if $c_d \leq 0$ and $x =$ `NULL` **then**
  print `Cannot decide` $f_d > 0$ `for` $k \leq k_{\max}$`.`
else *//* $c_d > 0$ *in the following*
  print f `has mixed-integer minimizers.` *//* $f_d > 0$ *by (1.3)*
  for $j = 1, \ldots, d-1$ **do**
    compute valid c_j *// via methods* (3.8), (3.10) *or* (3.14)
  end for
  define $q : \mathbb{R} \to \mathbb{R}$, $q(\lambda) = \sum_{j=1}^{d} c_j \lambda^j$
  $R \leftarrow$ largest root of q in $\mathbb{R}$ *//* $R \geq 0$ *by Theorem 3.4*
  print `The minimizers` x' `suffice` $\|x'\|_p \leq R$`.` *// also by Theorem 3.4*
  output R
end if

in these sets. More precisely, we compare the values $\lambda(\{x \in \mathbb{R}^n : \|x\|_p \leq R\})$ for methods a), b), c) and e) as well as

$$\lambda\left(\bigcup_{\tau\in\{-1,1\}^n} \{x \in H_\tau : \|x\|_p \leq R_\tau\}\right) = \frac{1}{2^n} \sum_{\tau\in\{-1,1\}^n} \lambda(\{x \in \mathbb{R}^n : \|x\|_p \leq R_\tau\})$$

for method d) with H_τ from (3.15).

3.3.2. The first approach to sampling

Our first approach to sampling is the following: For a fixed number of variables n and an even degree d, we sample from the family

$$f = \sum_{|\alpha|\leq d} a_\alpha X^\alpha = \sum_{|\alpha|\leq d} a_\alpha X_1^{\alpha_1} \cdots X_n^{\alpha_n}, \quad a_\alpha \sim \begin{cases} \mathcal{U}(-1,1), & \alpha \neq de_i, \\ \mathcal{U}(0,1), & \alpha = de_i, \end{cases} \tag{F1}$$

where the e_i are unit vectors. We make the case distinction on α as we are only interested in polynomials with positive definite leading form, and a leading form that does not satisfy the condition

$$a_{(d,0,\ldots,0)} > 0,\ a_{(0,d,0,\ldots,0)} > 0,\ \ldots,\ a_{(0,\ldots,0,d)} > 0 \tag{A}$$

cannot be positive definite. Then, we solve Program (3.10) at the level $k' = d$ in the hierarchy to compute a lower bound c_d on $\min_{x\in\mathbb{S}_2^{n-1}} f_d(x)$ to determine whether f indeed has a positive definite leading form. If $c_d \leq 0$, we discard the instance, else we know that f_d is positive definite.

For every tuple (n, d) with $n = 2, 3, 4$ and $d = 2, 4, 6, 8, 10$, we created 1000 random instances of polynomials from family (F1). In Figure 3.1, we plot how many of these have been detected to satisfy $f_d > 0$. For these instances that satisfy $f_d > 0$, we compute the norm bounds, which are depicted in Figure 3.2 for a selection of n and d. By construction, the bound b) is improved by c), which is in turn improved by d), and the plot shows that the difference is significant. The plot also shows that the sos-based norm bound e) yields a further improvement and gives the best results. Our bounds improve the bound from the literature a) by several orders of magnitude (even more so with higher n and d, cf. the next section).

However, Figure 3.1 also reveals that, as d and n increase, less instances with positive definite leading form get detected. This is to be expected, as the ratio of terms we control through condition (A), linear in n, to the total number of terms of the leading form, $\binom{n+d-1}{d}$ by (1.2), decreases quickly in n and d, and so does the probability that the leading form only attains positive values. Moreover, the conversion from an sos program to an SDP and its following solution process takes much longer with increasing n and d. It is thus not practicable to generate enough instances with positive definite leading form via (F1) in reasonable time for higher n and d. We hence sample from a second family.

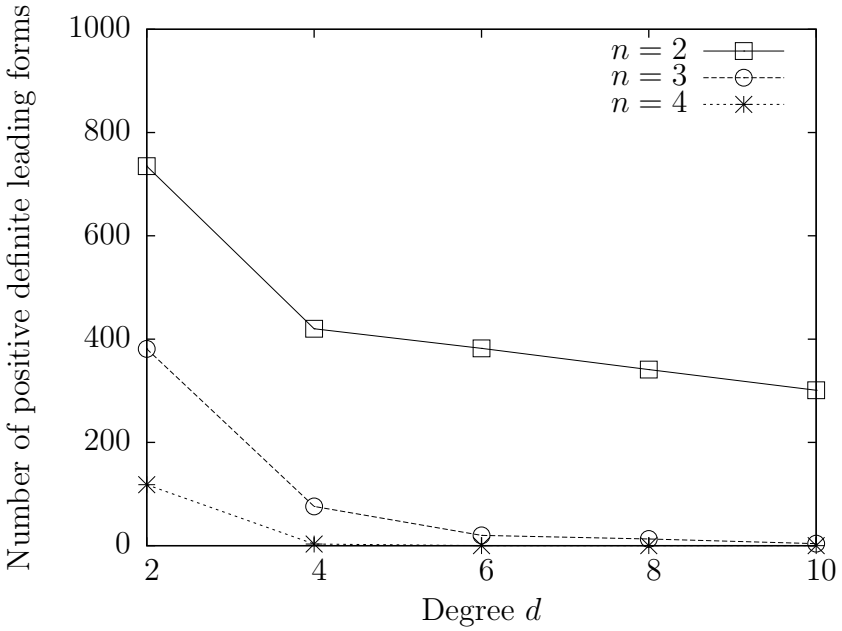


Figure 3.1.: From family (F1), 1000 instances are sampled for different dimension n and degree d. The plot shows the number of instances with detected positive definite leading form.

3.3.3. Sampling from the Parrilo-Sturmfels distribution

The following distribution on a subset of all polynomials is based on a suggestion by Parrilo and Sturmfels [PS03]:

$$f = X_1^d + \ldots + X_n^d + \sum_{|\alpha|<d} a_\alpha X^\alpha, \quad a_\alpha \sim \mathcal{U}(-K, K), \tag{F2}$$

where $K > 0$ is a constant. The advantage is that every polynomial from this family has a positive definite leading form. As parameters for the experiments, we chose the (n, d)-tuples $(2, 6)$, $(3, 4)$, $(3, 6)$, $(4, 4)$, $(5, 8)$ and $(7, 4)$, $p = \deg f$ and sampled 50 instances from the distribution (F2). Five of these (n, d) tuples are of the order of the B&B-experiments to follow, and $(5, 8)$ illustrates the volume for a high degree and a moderate number of variables. We chose $K = 2$ since then the coefficients are of the same size on average, i.e. the expected value of the modulus of all present coefficients is 1.

The resulting volumes are plotted in Figure 3.2. In comparison with family (F1), there is a smaller variance in the values of each norm bound for family (F2). This can be seen as follows: The largest real zero of the polynomial $q(\lambda) = \sum_{j=1}^{d} c_j \lambda^j$ is a valid norm bound on integer minimizers of f if $c_d > 0$ and $c_j \leq c_j^* = \min_{x \in \mathbb{S}_p^{n-1}} f_j(x)$ for all j (Theorem 3.4). For family (F1), the leading form f_d contains random terms and the value c_d^* (and thus c_d) can, in principle, be positive and arbitrarily close to zero. Positive but arbitrarily small c_d lead to arbitrarily large real zeros of q and thus arbitrarily large norm bounds. On the other hand, for family (F2), the leading form f_d and hence c_d^*

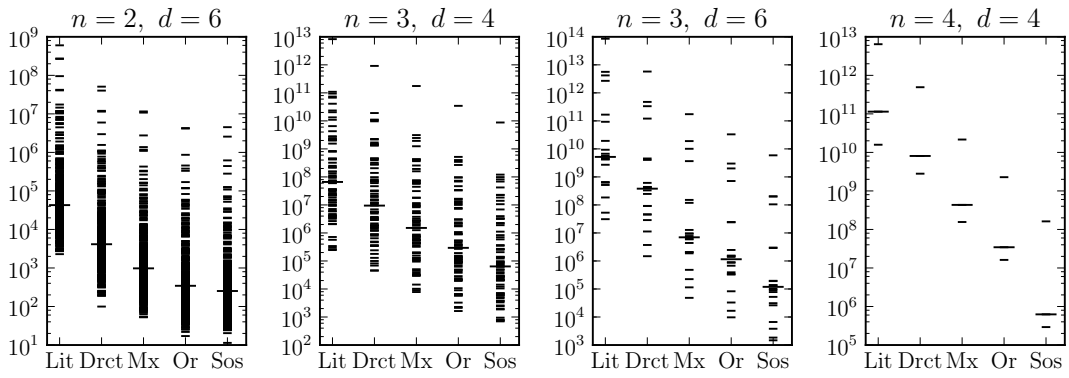


Figure 3.2.: From family (F1), 1000 instances are sampled for different dimension n and degree d. If the leading form is detected to be positive definite, the 5 norm bounds are computed. We scatter plot the volume of the sets they confine (logarithmic scale). The larger line is the median of the depicted values.

and c_d are constant, which results in less variance in the largest zero of q and thus in the norm bounds. As for family (F1), the plot shows that the sos-based bound e) improves on our other bounds b), c), d). Concerning the norm bound a), the new norm bounds outperform the one from the literature on all instances, and this even more so with increasing n and d. Most prominently in the plot, for $(n, d) = (5, 8)$, the norm bound based on the c_j from e) reduces, on average, the number of potentially optimal solutions by a factor of approximately 18 orders of magnitude by comparison with the classic norm bound.

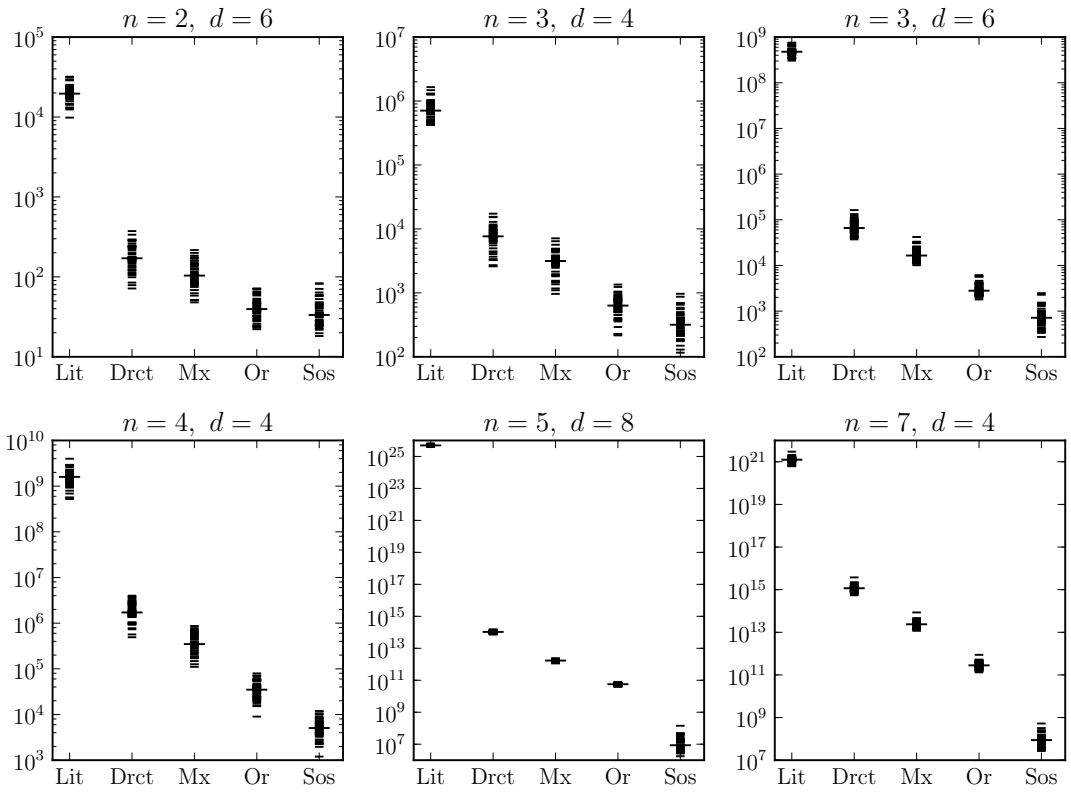


Figure 3.3.: From family (F2), 50 instances are sampled for different dimension n and degree d. We compute the 5 norm bounds and scatter plot the volume corresponding to the sets they confine (logarithmic scale). The larger line is the median of the depicted values.

4. Seminorm balls containing the feasible set

In this chapter we compute seminorm balls that contain all feasible solutions, that is, the set $F_{\mathcal{I}}$, or the relaxed feasible set F. Similar to norm bounds, one motivation is to make the integer variables of MINLP accessible to branch and bound. Contrasting to the norm bounds in the previous chapter (Chapter 3), we do not assume knowledge of a feasible point q.

Section 4.1 starts with a motivation and illustrates the relation to norm bounds.

Section 4.2 formulates the task to find a seminorm ball containing $F_{\mathcal{I}}$ as an auxiliary program. We go on by discussing how the geometry of $F_{\mathcal{I}}$ is related to the existence of feasible solutions of the auxiliary program. For polynomial constraint functions, relaxed integrality and suitable seminorms, we show that the auxiliary program can be approximated with sos programming.

Section 4.3 shows that if a seminorm ball containing F is known, it is possible to use purely arithmetic arguments to shrink the ball further such that it still contains $F_{\mathcal{I}}$, but not necessarily F. In other words, we explore how we can infer a nonlinear cut for F.

4.1. Motivation

Norm bounds have been introduced in (3.3) in Section 3.1 as upper bounds on the mixed-integer nonlinear problem

$$\begin{aligned} \max \quad & \|\bar{x} - x\| \\ \text{s.t.} \quad & f(x) \leq f(q) \\ & x \in F_{\mathcal{I}}. \end{aligned}$$

Similar to norm bounds, the primary motivation is to make the integer variables of MINLP accessible to branch and bound. We have seen that norm bounds can be explicitly computed with various methods provided the leading form of f is positive definite (Theorem 3.4).

The norm bound approach works well if a feasible point q is known. This chapter generalizes norm bounds to the case that no such q is known, that is, we consider the problem

$$\begin{aligned} \max \quad & \|\bar{x} - x\| \\ \text{s.t.} \quad & x \in F_{\mathcal{I}}, \end{aligned}$$

where our auxiliary program also allows for a seminorm N instead of a norm $\|\cdot\|$.

We explore in this chapter in which ways the assumptions necessary for the computation of the norm bound can be weakened. Firstly, we show how the strong coercivity assumption on the objective f, namely, positive definiteness of the leading form f_d, can be weakened or even replaced by suitable assumptions involving only the constraint functions. Secondly, the new approach takes the underlying geometry of the feasible set into account explicitly by considering nonlinear valid inequalities that are deduced from the constraint functions (the quadratic module generated by the constraint polynomials), and not only implicitly (by, e.g., a feasible point q). Thirdly, we allow for seminorms instead of norms, which is of particular interest in the mixed-integer case, as we explain below. Let us note that, however, these extensions do not belittle the results on norm bounds. The results in this chapter rely strongly on sos methods, whilst the norm bounds work without sos methods, too.

In the next section, we analyze the auxiliary problem and discuss an approximating hierarchy.

4.2. Finding tight enclosing seminorm balls

In the following, $N : \mathbb{R}^n \to \mathbb{R}$ is a seminorm defined on $\mathbb{R}^n$. For the sake of generality, we consider again a deputy set $S \subset \mathbb{R}^n$; natural candidates for S are $F_{\mathcal{I}}$ and F. The auxiliary problem to find a seminorm ball centered at $\bar{x} \in \mathbb{R}^n$ containing S can be formulated as

$$\begin{aligned} \max \quad & N(\bar{x} - x) \\ \text{s.t.} \quad & x \in S \end{aligned} \tag{S1}$$

where $\bar{x}$ may be thought of as a reference point.

We can interpret interpret Program S1 in two ways: Geometrically, and in terms of valid inequalities.

Observation 4.1. *Let $N : \mathbb{R}^n \to \mathbb{R}$ be a seminorm and denote the optimal value of* S1 *by z^*.*

1. a) *Let $S \supseteq F_{\mathcal{I}}$. If z^* is finite, $(N(\bar{x} - x) \leq z^*)$ is a valid inequality for $F_{\mathcal{I}}$. Put differently, $F_{\mathcal{I}}$ is contained in the seminorm ball $\mathbb{B}^N_{z^*}(\bar{x})$ in this case.*
   b) *Let $S \subset F_{\mathcal{I}}$. If z^* is infinite, $F_{\mathcal{I}}$ is unbounded.*
2. *Suppose $q \in F_{\mathcal{I}}$ is given.*
   a) *Let $S \supseteq \left(F_{\mathcal{I}} \cap \mathcal{L}^f_{\leq}(f(q))\right)$. If z^* is finite, $(N(\bar{x}-x) \leq z^*)$ is a valid inequality for all optimal solutions of* MINLP*. Put differently, all optimal solutions are contained in the seminorm ball $\mathbb{B}^N_{z^*}(\bar{x})$ in this case.*
   b) *Let $S \subset \left(F_{\mathcal{I}} \cap \mathcal{L}^f_{\leq}(f(q))\right)$. If z^* is infinite, the set of feasible solutions with objective value at least as good as $f(q)$ is unbounded.*

For the proof, we need the following result:

Lemma 4.2 (see, e.g., "Proposition" in [Gol17])**.** *Let V be a finite-dimensional real or complex vector space, equipped with a seminorm N and a norm $\|\cdot\|$. Then N is left-equivalent to $\|\cdot\|$, i.e., there is a constant $C > 0$ with*

$$N(x) \leq C\|x\|, \quad x \in V.$$

Proof of Observation 4.1. To see 1a, note that, by optimality of z^*, $N(\bar{x}-x) \leq z^*$ holds for all $x \in S$, thus for all $x \in F_{\mathcal{I}}$. For a proof of 1b, note that, by Lemma 4.2, there is $C > 0$ such that $N(\bar{x}-x) \leq C\|\bar{x}-x\|_2$ for all $x \in \mathbb{R}^n$. Hence, as $N(\bar{x}-\cdot)$ is unbounded on S, so is $\|\bar{x}-\cdot\|_2$, and hence $\|\bar{x}-\cdot\|$ on $F_{\mathcal{I}}$, and the claim follows. The proofs for the remaining claims follow analogously. □

To get tractable relaxations, we restrict our attention from now on weighted p-seminorms, that is, seminorms of the form

$$N(x) = \sqrt[p]{\sum_{i=1}^{n} a_i |x_i|^p}, \quad x \in \mathbb{R}^n, \tag{4.1}$$

for some fixed $a_1, \ldots, a_n \in \mathbb{R}_{\geq 0}$ and $p \in [1, \infty)$. Clearly, N is a norm if and only if all a_i are positive. Then, as N is nonnegative, maximizing N over S is equivalent to maximizing N^p over S, that is, S1 reads

$$\begin{aligned} \max \quad & \sum_{i=1}^{n} a_i |\bar{x}_i - x_i|^p \\ \text{s.t.} \quad & x \in S. \end{aligned} \tag{S2}$$

If p is even, lower bounds on S2 can be computed with the help of sos programming, provided S is given by polynomial constraints. So suppose $S = K(h_1, \ldots, h_s)$ for some $h_1, \ldots, h_s \in \mathbb{R}[X_1, \ldots, X_n]$. This leads to the following hierarchy of sos programs, parameterized by $k \in \mathbb{N}$:

$$\begin{aligned} \min \quad & \lambda \\ \text{s.t.} \quad & \lambda - \sum_{i=1}^{n} a_i (\bar{x} - X_i)^p \quad \in M\left(h_1, \ldots, h_s\right)[k] \\ & \lambda \in \mathbb{R} \end{aligned} \tag{SR_k}$$

where $p \in 2\mathbb{N}$, $a_1, \ldots, a_n \in \mathbb{R}_{\geq 0}$. Let us note that the formulation is within the scope of sos programming.

Observation 4.3. *Program* SR_k *is a valid sos program for* $p \in 2\mathbb{N}$.

Proof. This follows from (1.16), using the fact that $p \in 2\mathbb{N}$. □

Let us also note that feasible solutions of SR_k can be used to derive feasible solutions of S2.

Observation 4.4. *Every feasible solution* λ *of* SR_k *yields the upper bound* $\sqrt[p]{\lambda}$ *on* S2 *with* $S = M(h_1, \ldots, h_s)$.

Proof. Feasibility of λ for SR_k implies

$$\lambda - \sum_{i=1}^{n} a_i (\bar{x}_i - x_i)^p \geq 0 \text{ for } x \in S = K(h_1, \ldots, h_s)$$

by Observation 1.18. Thus, as p is even and $p > 0$, this means $\lambda \geq N(x)^p$ (where N was defined in (4.1)) on S, or $\sqrt[p]{\lambda} \geq N(x)$ on S, and the claim follows. □

We may now prove, under the usual assumption that the corresponding quadratic module is Archimedean, convergence of the optimal values of the approximating hierarchy towards the optimal value of the auxiliary problem. We have outlined in detail in Section 1.5.3 several sufficient conditions in terms of the optimization problem MINLP that ensure the Archimedean property.

Theorem 4.5. *Suppose $M(h_1, \dots, h_s)$ is Archimedean. Then* S2 *has an optimal solution $x \in S = K(h_1, \dots, h_s)$. Denote the optimal value of* S2 *by z^* and of* SR_k *by $\lambda^{(k)}$. Then*

$$\sqrt[p]{\lambda^{(k)}} \searrow z^* \text{ for } k \to \infty.$$

Proof. As $M(h_1, \dots, h_s)$ is Archimedean, the set $S = K(h_1, \dots, h_s)$ is compact by Corollary 1.19. Fix a norm $\|\cdot\|$ on $\mathbb{R}^n$. By Lemma 4.2, there is $C > 0$ such that

$$N(x) \leq C\|x\|$$

for $x \in \mathbb{R}^n$. As any norm is bounded on the compact set S, so is $N(x)$, and $\max_{x \in S} N(x)$ is finite. As N is continuous, the maximum is attained at some x in the compact set S, in other words z^* is finite and attained.

Let $\max_{x \in S} N^p(x) = \lambda^*$. We have just shown that λ^* is finite. By Corollary 1.27,

$$\lambda^{(k)} \searrow \lambda^* \text{ for } k \to \infty.$$

The convergence claim now follows from the fact that $(z^*)^p = \lambda^*$. □

Theorem 4.5 can be strengthened in case that N is a norm.

Proposition 4.6. *Let a norm $N(x) = \sqrt[p]{\sum_{i=1}^n a_i x_i^p}$ with $a_i > 0$ for all $i \in [n]$ be given. Then* SR_k *is feasible if and only if $M(h_1, \dots, h_s)$ is Archimedean.*

Proof. Note that $a_i > 0$ for all i implies that N is a norm. Let SR_k be feasible. Thus there is $\lambda \in \mathbb{R}$ such that

$$q := \lambda - \sum_{i=1}^n a_i(\bar{x}_i - X)^p \in M(h_1, \dots, h_s).$$

Clearly,

$$K(q) = \{x \in \mathbb{R}^n : q(x) \geq 0\} = B_{\sqrt[p]{\lambda}}(\bar{x})$$

where $B_R(p)$ denotes the N-norm ball with radius R centered at p. Since λ is finite and N a norm, the set $K(q)$ is bounded. Also, $K(q)$ is closed, hence $K(q)$ is compact. By Theorem 1.17 (4), $M(h_1, \dots, h_s)$ is Archimedean.

The converse direction is immediate from Theorem 4.5, since the existence of optimal solutions implies the existence of feasible solutions. □

We have given many sufficient conditions in terms of MIPP that ensure $M(h_1, \dots, h_s)$ is Archimedean in Section 1.5.3. It is future research to compare the performance of the norm bounds from Chapter 3 and the (semi)norm ball approach from this chapter on random instances.

4.3. Exploiting integrality

We end this section by considering how integrality information can be used to tighten the seminorm balls. This takes the idea in Theorem 3.4 (2) further. Clearly, the arguments to follow work for norm bounds, too.

For our arguments, the p-th power of the seminorm needs to be $\mathcal{I}$-integrality preserving, that is $N(x)^p \in \mathbb{Z}$ whenever x in $\mathbb{R}^n_{\mathcal{I}}$. To this end let we consider seminorms N of the type $N^p(x) = \sum_{i=1}^n a_i|x_i|^p$, $a_i \geq 0$. As a first step, let us consider when N^p is $\mathcal{I}$-integrality preserving.

Observation 4.7. *Let $p \in [1, \infty)$ and*

$$N(x) = \sqrt[p]{\sum_{i=1}^n a_i|x_i|^p}, \quad x \in \mathbb{R}^n,$$

with $a_i \geq 0$ and not all $a_i = 0$. Then, the following are equivalent:

1. *N^p is $\mathcal{I}$-integrality preserving.*
2. *$p \in \mathbb{N}$, $a_i \in \mathbb{Z}_{\geq 0}$ for $i \in \mathcal{I}$ and $a_i = 0$ for $i \in [n] \setminus \mathcal{I}$.*

Proof. Let N^p be $\mathcal{I}$-integrality preserving. We derive a contradiction if any of the statements in Condition 2 does not hold. Let $i_0 \in [n]$ with $a_{i_0} > 0$ and e_i be the i-th unit vector. If p is not a natural number, then $N^p(\lambda e_{i_0}) = a_{i_0}|\lambda|^p$ is not integer for any prime λ. Hence $p \in \mathbb{N}$. Suppose $a_i \notin \mathbb{Z}_{\geq 0}$ for some $i \in [n]$. Then $N^p(e_i) = a_i \notin \mathbb{Z}$, and N^p is not $\mathcal{I}$-integrality preserving. Now suppose $a_i > 0$ for some $i \in [n] \setminus \mathcal{I}$. Then, there is $\lambda \in \mathbb{R}$ such that $N(\lambda e_i) = a_i|\lambda|^p$ is not integer. For the converse direction, if Condition 2 holds, N^p is obviously $\mathcal{I}$-integrality preserving. □

We also require that the reference point $\bar{x} \in \mathbb{R}^n$ from Program S2 in Section 4.2 is a mixed-integer point with respect to $\mathcal{I}$, that is, $\bar{x} \in \mathbb{R}^n_{\mathcal{I}}$. To derive a cut, suppose we have an upper bound λ on S2, i.e., an upper bound on

$$\begin{aligned} \max \quad & \sum_{i \in \mathcal{I}} a_i|\bar{x}_i - x_i|^p \\ \text{s.t.} \quad & x \in S. \end{aligned} \tag{4.2}$$

To solve MINLP, it is however more useful to have an upper bound for $S_{\mathcal{I}}$, that is, an upper bound on

$$\begin{aligned} \max \quad & \sum_{i \in \mathcal{I}} a_i|\bar{x}_i - x_i|^p \\ \text{s.t.} \quad & x \in S_{\mathcal{I}}. \end{aligned} \tag{4.3}$$

To use integrality arguments, denote the optimal value of (4.3) by z^*. We can now shrink the seminorm ball containing S by decreasing the upper bound λ towards z^*, which geometrically corresponds to cutting off points in $S \setminus S_{\mathcal{I}}$, using integer rounding. Integer rounding is a common technique in integer programming [NW88], and for integrality preserving maps, the argument can be formulated as follows:

Observation 4.8. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be $\mathcal{I}$-integrality preserving and $\lambda \in \mathbb{R}$ with $\lambda \geq \max_{x \in S} f(x)$. Then $\lfloor \lambda \rfloor \geq \max_{x \in S_\mathcal{I}} f(x)$.*

Proof. This follows from the fact that $f(S_\mathcal{I}) \subset \mathbb{Z}$. □

For our seminorm, we have:

Observation 4.9. *Let N be a seminorm defined as in* (4.1) *which is $\mathcal{I}$-integrality preserving. Given $\bar{x} \in \mathbb{R}^n_\mathcal{I}$, let λ be a finite upper bound on* (4.2) *and z^* be the optimal value of* (4.3)*. Then*

$$z^* \leq \lfloor \lambda \rfloor.$$

In decreasing the bound from λ to $\lfloor \lambda \rfloor$, we have not used any geometric property of S or $S_\mathcal{I}$, but only relied on arithmetic arguments. Pursuing this path further, we may ask more generally, given a seminorm N as in (4.1) with $p \in 2\mathbb{N}$ (to end with a polynomial) and furthermore N^p $\mathcal{I}$-integrality preserving:

$$\text{Is there } x \in \mathbb{R}^n_\mathcal{I} \text{ with } \quad \sum_{i \in \mathcal{I}} a_i (\bar{x}_i - x_i)^p = \lfloor \lambda \rfloor ? \tag{4.4}$$

Equations such as (4.4) are in the field of number theory known as *Diophantine equations*, see, e.g., p. 10 in [Sma98]. Clearly, these equations do not always have a solution, as the example below shows. Let us introduce some notation. For $p \in 2\mathbb{N}$, $N \in \mathbb{N}$ and $\lambda \in \mathbb{R}$ nonnegative and $a \in \mathbb{Z}^k_{\geq 0}$ put

$$L_{p,a}(\lambda) := \max \left\{ \mu \in \mathbb{Z} : \mu \leq \lambda \text{ and there is } x \in \mathbb{Z}^k \text{ with } \sum_{i=1}^k a_i x_i^p = \mu \right\}.$$

Example 4.10. We consider an instance with $p = 2$, $\mathcal{I} = [2]$, $N^2(x) = x_1^2 + x_2^2$ and $\lambda = 96.2$ as upper bound for N^2 on some feasible set S. Using, e.g., enumeration, it turns out that the largest integer μ below λ that can be written as $\mu = x_1^2 + x_2^2$ with integer x_i is $90 = 3^2 + 9^2$, i.e., $L_{2,(1,1)}(96.2) = 90$.

With the new notation, we have:

Proposition 4.11. *Let N be a seminorm defined as in* (4.1) *with $p \in 2\mathbb{N}$. Suppose further that N^p is $\mathcal{I}$-integrality preserving. Given $\bar{x} \in \mathbb{R}^n_\mathcal{I}$, let λ be a finite upper bound on* (4.2) *and z^* be the optimal value of* (4.3)*. Then*

$$z^* \leq L_{p,a}(\lambda).$$

Proof. If $S_\mathcal{I}$ is empty, $z^* = -\infty$ and there is nothing to prove. Suppose now that $S_\mathcal{I}$ is nonempty. As before, z^* is bounded from below by any feasible objective value and bounded from above by λ. Since N^p is $\mathcal{I}$-integrality preserving, N^p attains integer values on $\mathbb{R}^n_\mathcal{I}$ only, hence there is an optimal solution $x \in S_\mathcal{I}$ to (4.3) where z^* is attained, i.e., $z^* = \sum_{i \in \mathcal{I}} a_i x_i^p$. On the other hand we have $z^* \leq \lambda$. By definition of $L_{p,a}$, the inequality $z^* \leq L_{p,a}(\lambda)$ follows. □

We have thus shown how one may find a cut by using number-theoretic arguments which are isolated from the geometry at hand (and can thus, e.g., be preprocessed). This is similar to the arguments leading to Gomory-type cuts in integer programming [MMWW02]. In the following, we discuss the important special case of unit weights for the seminorm N.

Unit weights

We now look at p-seminorms (which are a special case of (4.1) with unit weights), that is,

$$N^p(x) = \sum_{i \in \mathcal{I}} x_i^p, \quad x \in \mathbb{R}^n.$$

For this case, the Diophantine equation in (4.4) is closely related to Waring's problem [VW02]. It asks the following: Given an integer k, is there an integer s such that every natural number is the sum of at most s k-th powers (of natural numbers)? The smallest such s is denoted by $g(k)$ in the literature and it was shown by Hilbert [Hil09] that $g(k)$ always exists. The first value is $g(2) = 4$: This is a restatement of Lagrange's classical four-square theorem – every natural number is the sum of four integer squares [HWHBS08, Theorem 369]. The next values are $g(3) = 9$ and $g(4) = 19$. This results in the following consequence:

Observation 4.12. *Let $p \in 2\mathbb{N}$. If $\mathcal{I}$ satisfies $|\mathcal{I}| \geq g(p)$, then $L_{p,(1,\dots,1)}(\lambda) = \lfloor \lambda \rfloor$ for all $\lambda \geq 0$.*

Proof. Let $\mathcal{I}$ satisfy $|\mathcal{I}| \geq g(p)$ and $\mu \in \mathbb{Z}_{\geq 0}$. By definition of $g(p)$,

$$\mu = \sum_{i=1}^{g(p)} x_i^p$$

for some $x \in \mathbb{Z}^{g(p)}$. A fortiori,

$$\mu = \sum_{i \in \mathcal{I}} x_i^p$$

for some $x \in \mathbb{Z}^{\mathcal{I}}$. Thus, for $\lambda \in [\mu, \mu + 1)$, we have

$$L_{p,a}(\lambda) = \mu = \lfloor \lambda \rfloor$$

where $a_i = 1$ for $i \in \mathcal{I}$. As μ was arbitrary, the claim follows from the fact that $\mathbb{R}_{\geq 0} = \bigcup_{\mu \in \mathbb{Z}_{\geq 0}} [\mu, \mu + 1)$. □

In other words, if we allow for sufficiently many summed-up powers, the cut in Proposition 4.11 reduces to the standard cut in Observation 4.9.

As a small digression, let us take a look at some examples as to how deep such a cut "on average" will be.[1] That is, given $p \in 2\mathbb{N}$ and unit weights $a_i = 1$ for $i \in \mathcal{I}$, how

[1] Note that, geometrically, we are only able to shrink the seminorm ball until the first integer points appear on the boundary of the ball.

often can we expect a cut $L_{p,a}(\lambda)$ that is deeper than $\lfloor\lambda\rfloor$ for varying λ? By means of number theory, questions as these can be approached with *densities*: For $A \subset \mathbb{N}$, let $A(m) = A \cap \{1, \ldots, m\}$, $a(m) = |A(m)|$. The *upper* and *lower asymptotic densities* of A are

$$\overline{d}(A) = \limsup_{m\to\infty} a(m)/m, \quad \underline{d}(A) = \liminf_{m\to\infty} a(m)/m.$$

In case the upper and lower asymptotic densities coincide, A has *asymptotic density* $d(A) = \underline{d}(A) = \overline{d}(A)$. For our setting, we define

$$A_{p,k} = \left\{ a \in \mathbb{N} : \text{there are } x_1, \ldots, x_k \in \mathbb{Z} \text{ with } \sum_{i=1}^{k} x_i^p = a \right\}, \qquad p \in 2\mathbb{N},\ k \in \mathbb{N}.$$

The sequence $a_{p,k}(m) = |A_{p,k}(m)|$ thus counts the natural numbers below m for which there is an integer point on a $(k-1)$ dimensional p-sphere with integer radius not larger than m. For some p and k, analytic expressions for the densities are known. Clearly, $a_{p,1}(m)$ counts the perfect p-th powers below m. Hence $a_{p,1}(m) = \lfloor\sqrt[p]{m}\rfloor$, and $d(A_{p,1}) = 0$ follows by elementary means. Also, it can be shown [Tho73] that $d(A_{2,2}) = 0$, $d(A_{2,3}) = \frac{5}{6}$. The identities $d(A_{2,4}) = d(A_{2,k}) = 1$ for $k \geq 4$ follow again by Lagrange's four-square theorem [HWHBS08, Theorem 369]. For a graphical visualization of $a_{p,k}(m)/m$ for some further values of p and k, we refer to Figure 4.1.

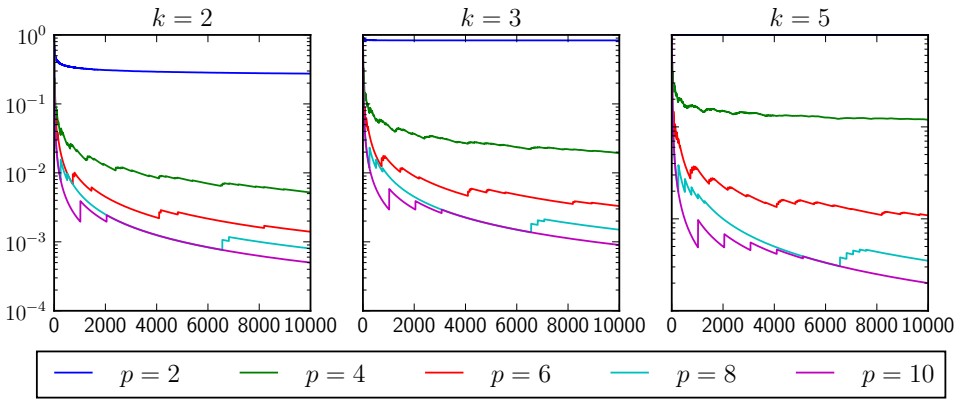


Figure 4.1.: First values of the sequence $a_{p,k}(m)/m$ (vertical axis) depending on m (horizontal axis) for different values of p and k.

5. Ellipsoids containing the feasible set

This chapter is about ellipsoids of minimal volume containing $F_{\mathcal{I}}$, the feasible set of MINLP. The primary motivation is again to make the integer variables in MINLP accessible to branch and bound. There are two prominent differences to the seminorm balls approach from Chapter 4: The shape (and not only the size) of the ellipsoid enters the auxiliary program, and, most importantly, we treat the center of the ellipsoid as a decision variable. Additionally, we allow for restrictions on the shape and position of the center.

Section 5.1 formulates the task to find an ellipsoid of minimal volume – with shape confined to a shape class and with a center restricted to a certain region – containing F as an auxiliary program. We analyze geometric conditions on $F_{\mathcal{I}}$ that guarantee the existence of feasible and optimal solutions of the auxiliary program. In this general form, the auxiliary program is not yet tractable.

Section 5.2 provides the first steps towards tractability: We use results from the literature to linearize some of the constraints and end up with a semidefinite program with possibly infinitely many constraints, a semi-infinite program.

Section 5.3 circumvents semi-infinite constraints by restricting to polynomial constraint functions, that is, to MIPP. This makes the problem accessible to methods from real algebra. We give an approximating hierarchy and convergence results. Under suitable assumptions on the shape class and the region confining the center of the ellipsoid, the approximating hierarchy is a semidefinite program with concave objective.

5.1. An auxiliary program to find ellipsoids of minimal volume

5.1.1. Motivation

The aim of this chapter is, as indicated in the introduction, to find ellipsoids containing the feasible set. Similarly to the norm bounds and seminorm balls we considered, ellipsoids are well-understood objects and highly useful in the actual solution process of MINLP: For example, they make the integer part of solutions enumerable and thus accessible to branch and bound.

However, the task to find an ellipsoid of minimal volume containing a given set is interesting in its own right, and for additional generality, we state our results using a deputy set $S \subset \mathbb{R}^n$ instead. We keep in mind that natural candidates for this set S are the relaxed feasible set F or the feasible set $F_{\mathcal{I}}$ itself, or even the set of all optimal solutions.

The reason for considering ellipsoids instead of norm or seminorm balls is the following: The norm and seminorm balls that we studied in the previous chapters have a fixed shape, more precisely their unit spheres are fixed. However, it may happen that the set S is large (in diameter, say) but flat (in some unknown direction). In applications, this is a useful information as it allows, amongst other things, earlier branching on the integer variables. Hence it makes sense to detect this flatness, and this is what we attempt with enclosing ellipsoids of minimal volume. It turns out that, in the ellipsoidal setting, we can also optimize for the center of the ellipsoid. Once an ellipsoid containing S is found, we are confident that integrality arguments can be used to shrink the ellipsoids further so that it still contains $S_{\mathcal{I}}$ but not necessarily S, in other words, resulting in a nonlinear cut, but leave this as interesting future research.

5.1.2. The auxiliary program

In this section we formulate the auxiliary program that computes an ellipsoid of minimal volume containing the deputy set $S \subset \mathbb{R}^n$. The shape of the ellipsoid is determined by its defining matrix Q and its center x_0.[1] We additionally allow restrictions on the shape of the ellipsoid by defining a *shape class* $\mathcal{Q} \subset \mathcal{S}^n_+$, that is, a set over which Q may range and on the position of the center x_0 of the ellipsoid, ranging over $\mathcal{X} \subset \mathbb{R}^n$.

$$\begin{aligned} \min \quad & \operatorname{vol}(E(Q, x_0)) \\ \text{s.t.} \quad & S \subset E(Q, x_0) \\ & Q \in \mathcal{Q} \\ & x_0 \in \mathcal{X} \end{aligned} \tag{E1}$$

For future reference, let us state the assumptions on $\mathcal{X}$ and $\mathcal{Q}$ that shall tacitly hold henceforth.

[1] Deviating from our conventions in Section 1.4.7, we allow in Program E1 for feasible (and possibly optimal) solutions with objective value $+\infty$.

Assumption 5.1. In Program E1 and in the programs to follow in this chapter,

- $\mathcal{X}$ is a closed, nonempty subset of $\mathbb{R}^n$,
- $\mathcal{Q}$, the shape class, is a closed, nonempty cone in $\mathcal{S}^n_+$, with $\mathcal{Q} \cap \mathcal{S}^n_{++} \neq \emptyset$.

As a note on the assumptions, we require nonemptiness of $\mathcal{X}$ and $\mathcal{Q}$ to avoid irrelevant special cases. The assumption that $\mathcal{Q}$ contains a positive definite matrix is to ensure existence of feasible solutions in case S is bounded (Theorem 5.3). Closedness is necessary for compactness arguments, which in turn gives a sequence converging to an optimal solution (Theorem 5.6). Finally, the requirement of $\mathcal{Q}$ to be a cone is to stay feasible when scaling the ellipsoids. In [BHS15], the authors consider related shape classes; amongst others, the set of diagonal matrices – corresponding to axis-parallel ellipsoids – and the set of those matrices that yield quadratic functions having the so-called strong rounding property.

Before we analyze Program E1, let us also note that it contains several interesting special cases. For the shape of the ellipsoids, we may chose $\mathcal{Q} = \mathcal{S}^n_+$ and impose no constraints on the shape. We may choose $\mathcal{Q} = \{Q \in \mathcal{S}^n_+ : Q \text{ is diagonal}\}$ to only consider axis-parallel ellipsoids. As a final important example, we may choose $\mathcal{Q} = \{rI_n : r \in \mathbb{R}_{\geq 0}\}$ to restrict the investigation to norm balls. For the center x_0 of the ellipsoid, we may choose no restriction $\mathcal{X} = \mathbb{R}^n$ to end up with the smallest enclosing ellipsoid in the shape class $\mathcal{Q}$, mixed-integer restriction $\mathcal{X} = \mathbb{R}^n_{\mathcal{I}}$ which is useful for integrality arguments, or a singleton set $\mathcal{X} = \{x_0\}$ for some $x_0 \in \mathbb{R}^n$ to fix the center of the ellipsoid.

Also, let us note the suitability of E1. We omit the trivial proof.

Observation 5.2. *Let $S \subset \mathbb{R}^n$ be arbitrary. If (Q, x_0) is a feasible solution of Program* E1*, it yields an ellipsoid $E(Q, x_0)$ that contains S.*

5.1.3. Existence of feasible solutions

We proceed by characterizing the existence of feasible to Program E1 with finite volume. Our attention is restricted to solutions of finite volume as the solution $Q = 0$ of little insight is always feasible: the zero matrix is in the shape class $\mathcal{Q}$, since the latter is conic, closed, and nonempty. It should not be surprising that the existence of said solutions is tied – with equivalence – to a natural geometric assumption on S.

Theorem 5.3. *Let $S \subset \mathbb{R}^n$. Then Program* E1 *has feasible solutions with finite objective value if and only if S is bounded.*

Before we give the proof, let us remark some easy observations. The following observation allows nesting of scaled unit balls, possibly degenerated, and ellipsoids.

Observation 5.4. *Let $Q \in \mathcal{S}^n$. Then*

$$\lambda_{\min}(Q) \cdot I_n \preceq Q \preceq \lambda_{\max}(Q) \cdot I_n.$$

Proof. This is immediate by Rayleigh-Ritz (Theorem 1.6). □

Furthermore, the Loewner order of positive semidefinite matrices and containment of the corresponding ellipsoids are in order-reversing correspondence:

Observation 5.5 (see, e.g., Section 4 in [BHS15])**.** *Let $A, B \in \mathcal{S}_+^n$ and $x_0 \in \mathbb{R}^n$. Then*

$$A \preceq B \quad \textit{if and only if} \quad E(A, x_0) \supset E(B, x_0).$$

We can now prove the theorem.

Proof of Theorem 5.3. Let (Q, x_0) be a feasible solution with finite objective value. Hence $Q \succ 0$ by Observation 1.15, and the smallest eigenvalue is positive. By Observation 5.4, there is $r > 0$ with $rI_n \preceq Q$, thus

$$E(Q, x_0) \subset E(rI_n, x_0) = x_0 + E(rI_n) = x_0 + \frac{1}{\sqrt{r}} E(I_n) = B_{1/\sqrt{r}}(x_0)$$

by Observation 5.5 and the relation (1.9), hence S is bounded.

For the converse implication, let S be bounded, hence $\operatorname{cl} S$ is compact. By Assumption 5.1, there are $x \in \mathcal{X}$ and $Q \in \mathcal{Q}$ with $Q \succ 0$. Let

$$z' := \sup_{y \in \operatorname{cl} S} (y - x_0)^T Q (y - x_0).$$

By continuity and compactness, $z' < \infty$, by $Q \succ 0$ we also have $z' \geq 0$. Now put $z := \max(1, z')$. Observe that

$$S \subset E(\mu Q, x_0)$$

for all $\mu < \frac{1}{z}$, since

$$(y - x_0)^T \mu Q (y - x_0) \leq \frac{(y - x_0)^T Q (y - x_0)}{z} \leq 1$$

for $y \in S$. Since $\mu Q \in \mathcal{Q}$, we have found a feasible solution $(\mu Q, x_0)$. □

5.1.4. Existence of optimal solutions

Similarly to the characterization of feasible solutions of Program E1 with finite volume (Theorem 5.3), we now characterize optimal solutions of Program E1. Again, the existence of said solutions is tied to mild, natural geometric assumptions on S. We discuss a converse statement at the end of this section.

Theorem 5.6. *Let $S \subset \mathbb{R}^n$. Then for Program* E1*, optimal solutions with finite objective exist if S is bounded and has maximal affine dimension.*

Again we break the proof down into a sequence of observations and lemmata. Firstly, we need the fact that a sequence of ellipsoids with centers diverging in norm that contain a fixed ball must have a diverging volume as well. This seems completely evident, however, we give a proof for completeness. Instead of tedious volume computations we give a simple argument that suffices to push the volume towards infinity. This observation is used in the proof of Theorem 5.6 to bound a sequence of feasible x_k, which results in a converging subsequence by a compactness argument.

Observation 5.7. *Let a ball $B_r(p)$ corresponding to a norm on $\mathbb{R}^n$ be given, $r > 0$ and $p \in \mathbb{R}^n$, that is contained in every member of a sequence of ellipsoids $E(Q_k, x_k)$, $Q_k \in \mathcal{S}^n_+$, $x_k \in \mathbb{R}^n$, $k \in \mathbb{N}$. Let the sequence of centers $\{x_k\}_{k\in\mathbb{N}}$ satisfy $\|x_k\| \to +\infty$. Then $\operatorname{vol} E(Q_k, x_k) \to +\infty$.*

This observation in turn is easy to see with the following lemma.

Lemma 5.8. *Let $A, B \subset \mathbb{R}^n$, $x_0 \in \mathbb{R}^n$ with*

$$x_0 + B \subset A.$$

1. *If A is symmetric ($A = -A$), then $-x_0 - B \subset A$.*
2. *If A and B are symmetric, then $-x_0 + B \subset A$.*
3. *If A and B are symmetric and A is convex, then*

$$\operatorname{conv}\left((x_0 + B) \cup (-x_0 + B)\right) \subset A.$$

Proof. To see Claim 1, note that $x_0 + B \subset A$ if and only if $-x_0 - B \subset -A$. Since $A = -A$, the claim follows. To see Claim 2, note that $B = -B$ if B is symmetric, and the claim follows from Claim 1. Claim 3 is immediate by the fact that $x_0 + B$ and $-x_0 + B$ are subsets of A. □

Proof of Observation 5.7. By equivalence of norms on $\mathbb{R}^n$, we may assume that the norm under consideration is the 2-norm. Fix $k_0 \in \mathbb{N}$. Note that $B_r(0)$ and $E(Q_{k_0})$ are symmetric, convex sets, and that $B_r(p) \subset E(Q_{k_0}, x_{k_0})$ is equivalent to $p + B_r(0) \subset x_{k_0} + E(Q_{k_0})$, equivalently, $p - x_{k_0} + B_r(0) \subset E(Q_{k_0})$. Hence by Lemma 5.8,

$$\operatorname{conv}((p - x_{k_0} + B_r(0)) \cup (x_{k_0} - p + B_r(0)) \subset E(Q_{k_0}).$$

Thus, $E(Q_{k_0})$ – and hence $E(Q_{k_0}, x_{k_0})$, too – contains a cylinder of radius r and height $2\|p - x_{k_0}\|$. Since x_k diverges, so does the volume of the cylinder contained in $E(Q_k, x_k)$, and hence the volume of $E(Q_k, x_k)$ itself. □

In the proof of Theorem 5.6, we also consider a sequence of feasible Q_k. To exhibit convergence for the matrices as well, we rely again on a compactness argument. Compactness enters our setting as follows:

Lemma 5.9. *Let $Q^* \in \mathcal{S}^n_+$. Then, the set*

$$B(Q^*) := \{Q \in \mathcal{S}^n_+ : Q \preceq Q^*\}$$

is compact in $\mathcal{S}^n_+$.

Proof. Firstly, the set is closed. Indeed, let $Q \in \mathcal{S}^n_+$ and $\{Q_k\}_{k\in\mathbb{N}}$ a sequence in $B(Q^*)$ with $Q_k \to Q$. We must show that $Q \in B(Q^*)$. The condition $Q_k \in B(Q^*)$ is equivalent to

$$x^T(Q^* - Q_k)x \geq 0, \quad \text{all } x \in \mathbb{R}^n.$$

Taking the limit in the equation (matrix multiplication is continuous) shows that

$$x^T(Q^* - Q)x \geq 0$$

for all $x \in \mathbb{R}^n$, hence $Q^* \succeq Q$ and $Q \in B(Q^*)$ follows.

Secondly, it is also bounded. Suppose the contrary, and let $\{Q_k\}_{k\in\mathbb{N}}$ be a sequence in $B(Q^*)$ with $\rho(Q_k) \to +\infty$, where ρ is the spectral norm on $\mathbb{R}^{n\times n}$. As all Q_k are symmetric, $\rho(Q_k) = \lambda_{\max}(Q_k)$. On the other hand, by Observation 5.4, $\lambda_{\max}(Q^*) \succeq Q^* \succeq Q_k$ for all k, bounding the maximum eigenvalue of Q_k and thus the spectral norm, a contradiction. □

Now, in the proof of Theorem 5.6, if we can find an upper bound $Q^* \succ 0$ with $Q \preceq Q^*$ for all feasible solutions (Q, x_0) of E1, we can apply Lemma 5.9. Finding Q^* is achieved using convexity and a dimensionality argument by the fact that there is a fixed ball $B_r(p)$ contained in all ellipsoids that are feasible for E1.

Lemma 5.10. *Let $S \subset \mathbb{R}^n$. Assume further that S is bounded and of affine dimension n. Then, there is a ball $B_r(p)$, where $r > 0$ and $p \in \mathbb{R}^n$, such that every feasible ellipsoid $E(Q, x_0)$ to Program* E1 *satisfies $B_r(p) \subset E(Q, x_0)$. Moreover, there is $Q^* \in \mathcal{S}^n_{++}$ with the property:*

If (Q, x_0) is feasible for Program E1*, then $Q \preceq Q^*$.*

Proof. Let $C := \operatorname{conv}(S)$. Since S has affine dimension n, so does the supset C. Any full-dimensional convex set has nonempty interior, hence there is $p \in C$, $r > 0$ with $B_r(p) = rE(I_n, p) = E(Q^*, p) \subset C$ where $Q^* := \frac{1}{r^2} I_n$. Now let (Q, x_0) be any feasible solution. The proof is finished if we can show that $Q \preceq Q^*$. As ellipsoids are convex, the ellipsoid $E(Q, x_0)$ must also contain the convex hull of $C = \operatorname{conv}(S)$, especially, $B_r(p) \subset E(Q, x_0)$. Note that by Lemma 5.8 (3), the ball $B_r(p)$ still lies in $E(Q, x_0)$ if we shift it towards x_0, that is, $B_r(x_0) = E(Q^*, x_0) \subset E(Q, x_0)$. Hence, by Observation 5.5, $Q^* \succeq Q$. □

We can finally prove the theorem.

Proof of Theorem 5.6. Let S be bounded and of affine dimension n. By Theorem 5.3, the program has feasible solutions, and we denote the infimum of E1 by v^*. There is a sequence of feasible solutions $(Q_k, x_k) \in \mathcal{Q} \times \mathcal{X}$ with the property

$$\operatorname{vol} E(Q_k, x_k) \downarrow v^* \quad \text{for } k \to \infty. \tag{5.1}$$

We show now that (a subsequence of) the x_k converge. By Lemma 5.10, there is $r > 0$, $p \in \mathbb{R}^n$ such that $B_r(p) \subset E(Q_k, x_k)$ for all k, and by Observation 5.7, we may assume that the x_k are bounded (the volume of the $E(Q_k, x_k)$ is eventually bounded). Every bounded sequence in $\mathbb{R}^n$ has a converging subsequence, and we denote the limit by $x_0 \in \mathbb{R}^n$. Since $\mathcal{X}$ was assumed to be closed, $x_0 \in \mathcal{X}$ follows. By passing to this subsequence if necessary, we may assume that x_k converges to $x_0 \in \mathcal{X}$.

We show next that (a subsequence of) the Q_k converge. By Lemma 5.10, there is Q^* such that for all feasible solutions Q, we have

$$Q \in B(Q^*) = \{Q' \in \mathcal{S}^n_+ : Q' \preceq Q^*\},$$

especially $Q_k \in B(Q^*)$ for all k. By Lemma 5.9, $B(Q^*)$ is compact. Since $\mathcal{Q}$ is closed, the intersection $B' := \mathcal{Q} \cap B(Q^*)$ is compact, too, thus a subsequence of the Q_k converges to some $Q_0 \in B'$. By passing to a subsequence if necessary, we may assume that Q_k converges to $Q_0 \in \mathcal{Q}$.

It remains to show that (Q_0, x_0) is an optimal solution. To this end we need to verify feasibility of (Q_0, x_0) and $\operatorname{vol} E(Q_0, x_0) = v^*$. We have just seen that $Q_0 \in \mathcal{Q}$ and $x_0 \in \mathcal{X}$. Feasibility follows if $S \subset E(Q_0, x_0)$. Indeed, let $x \in S$. Then, as all (Q_k, x_k) are feasible,

$$(x - x_k)^T Q_k (x - x_k) \leq 1$$

holds for all $k \in \mathbb{N}$. By taking the limit and using the fact that matrix multiplication is continuous, we find

$$(x - x_0)^T Q_0 (x - x_0) \leq 1,$$

or $x \in E(Q_0, x_0)$. Feasibility of (Q_0, x_0) for E1 follows. Now, using the explicit formula for the volume of an ellipsoid (Observation 1.15), we find

$$v^* = \lim_{k\to\infty} \operatorname{vol} E(Q_k, x_k) = \lim_{k\to\infty} \frac{\operatorname{vol} B_n}{\sqrt{\det(Q_k)}} = \frac{\operatorname{vol}(B_n)}{\sqrt{\det(Q_0)}} = \operatorname{vol} E(Q_0, x_0),$$

where continuity of the determinant was used. □

Remark 5.11. The task to exhibit minimal assumptions for a converse statement of Theorem 5.6 is left as future research. For example, if $\mathcal{Q} = \mathcal{S}^n_+$ and $\mathcal{X} = \mathbb{R}^n$, then optimal solutions exist only if S is bounded and has maximal affine dimension (i.e., affine dimension n). Note that these assumptions are surely not minimal. To see the claim, note the following. Suppose S does not satisfy the assumptions (S is bounded and has affine dimension n). In case S is unbounded, obviously no optimal solutions with finite objective exist, as by Theorem 5.3, no feasible solutions with finite objective exist. It remains to show that even if S is bounded, an affine dimension of $d < n$ implies that no optimal solutions exist. Let A be the affine hull of S, and let V be a subspace of $\mathbb{R}^n$ with $A = v_0 + V$ (and hence, $\dim V = d$). The case $d = 0$, corresponding to an empty or singleton set S, is trivial. By the proved necessity in Theorem 5.6 and the assumption that $\mathcal{Q} = \mathcal{S}^n_+$, there is a nondegenerated, d-dimensional ellipsoid $E(Q, x_0)$ in A of minimal volume containing S. Let $\{b_1, \ldots, b_{n-d}\}$ be an orthonormal basis of the

orthogonal complement $V^\perp$ of V, that is, $\mathbb{R}^n = V \oplus V^\perp$. The ellipsoid can be turned into a nondegenerated, n-dimensional ellipsoid E' by adding the axis $\{rb_1, \ldots, rb_{n-d}\}$ for some fixed $r > 0$, with volume proportional to r^{n-d}. Hence, the infimum objective is 0. If there was an optimal solution to this infimum objective of 0, it would, in view of Observation 1.15 correspond to a matrix with determinant $+\infty$, which is absurd.

5.2. Towards semidefinite constraints

Program E1 computes an ellipsoid of minimal volume containing a given set S, with additional constraints on the shape and center, and has nice theoretical properties. However, it is not clear how to solve it for some given S.

The first step to make this program tractable is to switch to a simpler, computationally advantageous[2] objective. Secondly, we rewrite the set inclusion constraint as a nonnegativity constraint. Thirdly, we restrict our attention to nondegenerated ellipsoids, which we justify formally in Proposition 5.12. After these steps, the program still contains the matrix Q and its inverse Q^{-1}; furthermore, one constraint is quadratic in x_0. To avoid matrix inversion, we fourthly use a linearization technique from the literature to rewrite the program with (possibly infinitely many) semidefinite constraints. Fifthly, the quadratic constraint can be circumvented with another linearization technique from the literature. After the first three steps, program E1 reads

$$\begin{array}{ll} \min & \log\det(Q^{-1}) \\ \text{s.t.} & 1-(x-x_0)^T Q(x-x_0) \geq 0 \quad \text{for } x \in S \\ & Q \in \mathcal{Q}_{\succ 0} \\ & x_0 \in \mathcal{X} \end{array} \tag{E2}$$

where $\mathcal{Q}_{\succ 0} := \mathcal{Q} \cap \mathcal{S}^n_{++}$. Let us state formally how both programs are related and that it is enough to consider nondegenerated ellipsoids.

Proposition 5.12. *The feasible solutions of Program* E2 *are feasible solutions of* E1*. Conversely, if a feasible solution of* E1 *has finite objective, it is feasible for* E2*. Moreover, if* E1 *has optimal solutions of finite objective, the optimal solutions of both programs coincide.*

Proof. Note at first that $S \subset E(Q, x_0)$ is equivalent to $x \in E(Q, x_0)$ for $x \in S$, or $(x-x_0)^T Q(x-x_0) \leq 1$ for $x \in S$. By Observation 1.15, the volume of $E(Q, x_0)$ equals $C_n/\sqrt{\det(Q)}$, where $C_n > 0$ is a constant depending on n. This shows that feasible solutions of finite objective coincide, and moreover, as the volume is nonnegative and the square root is monotonic, minimizing the volume over the feasible set is equivalent to minimizing $1/\det(Q)$ over the feasible set. Now the fact

$$\det(A^{-1}) = 1/\det A$$

for all invertible $n \times n$-matrices A and that the logarithm is a monotonic function implies the claim. □

For semidefinite modeling reasons preparing the fourth step, we substitute $P := Q^{-1}$, which yields the formulation

[2]The function $X \mapsto -\log\det X$ is self-concordant on $\mathcal{S}^n_{++}$; see, e.g., Example 9.5 in [BV04].

$$\begin{aligned} \min \quad & \log\det P \\ \text{s.t.} \quad & 1-(x-x_0)^T P^{-1}(x-x_0) \geq 0 \quad \text{for } x \in S \\ & P \in \mathcal{Q}' \\ & x_0 \in \mathcal{X} \end{aligned} \tag{E2$'$}$$

where

$$\mathcal{Q}' := \{Q^{-1} : Q \in \mathcal{Q}_{\succ 0}\} \tag{5.2}$$

Observation 5.13. *Feasible and optimal solutions of Programs* E2 *and* E2$'$ *are in bijection under the map* $(Q, x_0) \mapsto (Q^{-1}, x_0)$.

For now, the variable x_0 enters the constraint in E2 in a nonlinear fashion. We now explore the fifth step, i.e., how this constraint may be linearized using tools from semidefinite programming. The linearization technique is based on [ND05] (where the authors in turn rely on methods from [EGC99] and [CEG04]). The important difference of our approach to [ND05] is that in the reference, the authors do not minimize the volume directly, which is proportional to the determinant of Q^{-1}, but minimize the trace of Q^{-1} through a hierarchy of programs. However, the trace of Q^{-1} is only a coarse approximation of the volume of an ellipsoid.

The first observation from [ND05] is that in the constraint

$$1-(x-x_0)^T P^{-1}(x-x_0) \geq 0 \quad \text{for } x \in S$$

the variables x and x_0 can be separated into different factors, as the constraint is equivalent to

$$1-\begin{pmatrix} x \\ 1 \end{pmatrix}^T \left(I_n, -x_0\right)^T P^{-1} \left(I_n, -x_0\right) \begin{pmatrix} x \\ 1 \end{pmatrix} \geq 0 \quad \text{for } x \in S.$$

where $\left(I_n, -x_0\right)$ is an $n \times (n+1)$ matrix. As we are minimizing, the latter is equivalent to

$$\begin{aligned} 1-\begin{pmatrix} x \\ 1 \end{pmatrix}^T A \begin{pmatrix} x \\ 1 \end{pmatrix} \geq 0 \quad & \text{for } x \in S \\ A \succeq \left(I_n, -x_0\right)^T P^{-1} \left(I_n, -x_0\right) & \end{aligned} \tag{5.3}$$

where $A \in \mathcal{S}^{n+1}$ is an additional matrix variable.[3] The second observation from [ND05] that we use is that now the Schur complement (Theorem 1.11) can be taken to rewrite equation (5.3) in fully semidefinite form, completing the fourth step, as

$$\begin{pmatrix} P & \left(I_n, -x_0\right) \\ \left(I_n, -x_0\right)^T & A \end{pmatrix} \succeq 0,$$

[3] All reformulations are verified formally in Proposition 5.14.

which is the very reason for the reformulation of E2 into E2′. Collecting our observations, we get the following variant of Program E1:

$$\begin{aligned}
\min \quad & \log\det P && \text{(E3)}\\
\text{s.t.} \quad & 1 - \begin{pmatrix} x \\ 1 \end{pmatrix}^T A \begin{pmatrix} x \\ 1 \end{pmatrix} \geq 0 \quad \text{for } x \in S && \text{(5.4)}\\
& \begin{pmatrix} P & \left(I_n, -x_0\right) \\ \left(I_n, -x_0\right)^T & A \end{pmatrix} \succeq 0 && \text{(5.5)}\\
& A \in \mathcal{S}^{n+1} \\
& P \in \mathcal{Q}' \\
& x_0 \in \mathcal{X}
\end{aligned}$$

Since we have added additional variables in a nontrivial manner, we verify that we can still extract the relevant information from E3. This motivates the following proposition.

Proposition 5.14. *Program* E2′ *is a projection of Program* E3*.*

Proof. Let (P, x_0) be a feasible solution of E2′. We have to show that the solution extends to a feasible solution of E3. Choose $A := \left(I_n, -x_0\right)^T Q \left(I_n, -x_0\right)$. From the definition of A and feasibility of P, (5.4) holds. By the Schur complement (Theorem 1.11), (5.5) holds. Thus, the feasible solution lifts to a feasible solution (P, x_0, A) of E3.

For the converse direction, let (P, x_0, A) be a feasible solution to E2. Especially, $P \succ 0$ and P is invertible. Feasibility of (P, x_0) for E2 follows if we can show that the constraint

$$1 - (x - x_0)^T P^{-1} (x - x_0) \geq 0 \quad \text{for } x \in S \tag{5.6}$$

of E2′ holds. By the Schur complement again, (5.5) is equivalent to

$$A \succeq \left(I_n, -x_0\right)^T P^{-1} \left(I_n, -x_0\right),$$

and hence

$$\begin{pmatrix} x \\ 1 \end{pmatrix}^T A \begin{pmatrix} x \\ 1 \end{pmatrix} \geq \begin{pmatrix} x \\ 1 \end{pmatrix}^T \left(I_n, -x_0\right)^T P^{-1} \left(I_n, -x_0\right) \begin{pmatrix} x \\ 1 \end{pmatrix} = (x - x_0)^T P^{-1} (x - x_0) \tag{5.7}$$

for all $x \in \mathbb{R}^n$. Now equations (5.4) and 5.7 immediately imply (5.6).

It remains to show projection and lift of optimal solutions. But since the objective does not depend on A, this follows readily from the first part of this proof. □

5.3. Computational formulation as an approximating hierarchy

5.3.1. The hierarchy of approximations

As S may be infinite, the constraint (5.4) parameterized by all $x \in S$ can be semi-infinite (cf. the discussion in Section 2.3.3). To sidestep this, we restrict to a constraint set S given in the form of polynomial inequalities, which makes the problem accessible to (generalized) sos methods. In [ND05], the authors use a similar approach for a different objective function. Precisely, we assume that S is basic closed semialgebraic, i.e., S is given by polynomial constraints, $S = K(h_1, \ldots, h_s)$, for polynomials $h_1, \ldots, h_s \in \mathbb{R}[X_1, \ldots, X_n]$. Now let us consider the following program with parameter $k \in \mathbb{N}$:

$$\begin{aligned}
\min \quad & \log\det P && (\mathrm{ER}_k)\\
\text{s.t.} \quad & 1 - \begin{pmatrix} X \\ 1 \end{pmatrix}^T A \begin{pmatrix} X \\ 1 \end{pmatrix} \in M(h_1, \ldots, h_s)[k] && (5.8)\\
& \begin{pmatrix} P & (I_n, -x_0) \\ (I_n, -x_0)^T & A \end{pmatrix} \succeq 0 && (5.9)\\
& A \in \mathcal{S}^{n+1}\\
& P \in \mathcal{Q}'\\
& x_0 \in \mathcal{X}
\end{aligned}$$

where $\mathcal{Q}'$ was defined in (5.2) and, to be sure, $\begin{pmatrix} X \\ 1 \end{pmatrix} = (X_1, \ldots, X_n, 1)^T$. Let us verify that this formulation yields feasible solutions to E3.

Observation 5.15. *Any feasible solution to Program* ER_k *yields a feasible solution to* E3 *for* $S = K(h_1, \ldots, h_s)$ *and given polynomials* $h_1, \ldots, h_s$.

Proof. Let $M := M(h_1, \ldots, h_s)$. Let (P, x_0, A) be a feasible solution to ER_k for some $k \in \mathbb{N}$. By feasibility,

$$p_1 := 1 - \begin{pmatrix} X \\ 1 \end{pmatrix}^T A \begin{pmatrix} X \\ 1 \end{pmatrix} \in M[k]. \tag{5.10}$$

Hence $p_1 \in M$. By Observation 1.18, if $p_1 \in M = M(h_1, \ldots, h_s)$, then $p_1(x) \geq 0$ on $S = K(h_1, \ldots, h_s)$. The claim follows. □

The next theorem shows that we have convergence of the optimal values of the hierarchy to the optimal value of the original program, provided the quadratic module is Archimedean.

Theorem 5.16. *Let* $h_1, \ldots, h_s \in \mathbb{R}[X_1, \ldots, X_n]$ *be given, and consider Program* ER_k. *The quadratic module* $M(h_1, \ldots, h_s)$ *is Archimedean if and only if* ER_k *is feasible for*

some $k \in \mathbb{N}$. In this case, denote the optimal values of ER_k *by* $\lambda^{(k)}$ *and the optimal value of* E3 *by* z^**. Then*

$$\lim_{k\to\infty} \lambda^{(k)} \searrow z^*.$$

Proof. Let $M := M(h_1, \ldots, h_s)$ be Archimedean. By assumption on $\mathcal{Q}$ there is $Q \in \mathcal{S}^n_{++}$ with $Q \in \mathcal{Q}$, and hence $P' := Q^{-1} \succ 0$. Pick $x_0 \in \mathcal{X}$. By Theorem 1.17 (1), there is $N \in \mathbb{N}$ with $N - (\underline{X} - x_0)^T (P')^{-1} (\underline{X} - x_0) \in M$, hence, by the Definition 1.2 of quadratic modules, we find that we may divide by N and still find that

$$p_2 := 1 - (\underline{X} - x_0)^T P^{-1} (\underline{X} - x_0) \in M \tag{5.11}$$

for $P := NP'$. Note that P is still in $\mathcal{Q}'$ as $\mathcal{Q}$ is conic by assumption. Put

$$A := \begin{pmatrix} I_n, -x_0 \end{pmatrix}^T P^{-1} \begin{pmatrix} I_n, -x_0 \end{pmatrix}$$

and note that

$$p_2 = 1 - \begin{pmatrix} \underline{X} \\ 1 \end{pmatrix}^T A \begin{pmatrix} \underline{X} \\ 1 \end{pmatrix}.$$

Moreover, by the definition of the truncated quadratic module, there is $k \in \mathbb{N}$ such that $p_2 \in M[k]$. Thus, constraint (5.8) is satisfied. With the Schur complement (Theorem 1.11) in mind, it is easily seen that (5.9) holds, too. Hence, (P, x_0, A) is a feasible solution to ER_k.

For the converse direction, let (P, x_0, A) be a feasible solution to ER_k, some $k \in \mathbb{N}$. Thus, by feasibility, p_1 defined as in (5.10) is in $M[k]$ and hence in M. Furthermore, the Schur complement and (5.9) imply that

$$H := A - \begin{pmatrix} I_n, -x_0 \end{pmatrix}^T P^{-1} \begin{pmatrix} I_n, -x_0 \end{pmatrix}$$

is positive semidefinite. Therefore, the spectral theorem (Theorem 1.5) implies the existence of a factorization $H = V^T D V$ with V orthogonal and D diagonal, and by Proposition 1.7, D has nonnegative entries. This implies that

$$q := \begin{pmatrix} \underline{X} \\ 1 \end{pmatrix}^T H \begin{pmatrix} \underline{X} \\ 1 \end{pmatrix} = \left(V \begin{pmatrix} \underline{X} \\ 1 \end{pmatrix} \right)^T D \left(V \begin{pmatrix} \underline{X} \\ 1 \end{pmatrix} \right)$$

is a sum of squares and, by Definition 1.2, lies in M. Again by Definition 1.2, so does the sum of p_1 and q:

$$q' := p_1 + q = 1 - (\underline{X} - x_0)^T P^{-1} (\underline{X} - x_0) \in M.$$

But $K(q') = \{x \in \mathbb{R}^n : q'(x) \geq 0\} = E(P^{-1}, x_0)$ is a nondegenerated ellipsoid – by feasibility $P \succ 0$ – and thus a compact set. By Theorem 1.17 (4), M is Archimedean.

To show convergence of the objective values, let M be Archimedean. We have just shown that feasible solutions to ER_k exist. Hence, by Observation 5.15, Program E3 is feasible, so by Proposition 5.14, also Program E2′ is feasible. This forces the infimum

v^* of Program E2′ to be finite, yet it may vanish. Fix $\varepsilon > 0$. Then, there is a feasible solution $(P, x_0) \in \mathcal{Q}' \times \mathcal{X}$ of E2 with

$$\log\det P \leq v^* + \varepsilon. \tag{5.12}$$

We show in the following that the family of solutions $((1+\delta)P_\delta, x_0, A)$ parameterized by $\delta > 0$ is feasible for ER_k and approximates P in volume. To this end, given $M \in \mathbb{R}^{n\times n}$, put

$$p_M := (\underline{X} - x_0)^T M(\underline{X} - x_0) \in \mathbb{R}[X_1, \ldots, X_n],$$

and by feasibility we have for all $x \in S = K(h_1, \ldots, h_s)$ that

$$\begin{aligned} & 1 - p_{P^{-1}}(x) \geq 0 \\ \iff & 1 + \delta - p_{P^{-1}}(x) \geq \delta \\ \iff & 1 - \frac{1}{1+\delta} p_{P^{-1}}(x) \geq \frac{\delta}{1+\delta} \\ \iff & 1 - p_{P^{-1}/(1+\delta)}(x) \geq \frac{\delta}{1+\delta} \end{aligned}$$

where $\delta > 0$ is arbitrary, and hence $q_\delta(x) := 1 - p_{P^{-1}/(1+\delta)}(x) > 0$ on S for $\delta > 0$. Since M is Archimedean, $q_\delta \in M$ by the Positivstellensatz (Theorem 1.20), and there is $k_\delta \in \mathbb{N}$ with $q_\delta \in M[k_\delta]$. As P is part of a feasible solution of E2, we have $P^{-1} \succ 0$, hence $P^{-1}/(1+\delta) = ((1+\delta)P)^{-1} \succ 0$ for all $\delta > 0$, and as $\mathcal{Q}$ is conic by Assumption 5.1, $((1+\delta)P, x_0)$ is feasible for Program E2. By Proposition 5.12, $((1+\delta)P, x_0)$ lifts to a feasible solution $((1+\delta)P, x_0, A_\delta)$ of Program E3, which we have just seen is a feasible solution to ER_k for $k = k_\delta$. The objective value depending on $\delta > 0$ is

$$\log\det((1+\delta)P) = n \cdot \log(1+\delta) + \log\det(P) \leq n\log(1+\delta) + v^* + \varepsilon \xrightarrow{\delta\to 0} v^* + \varepsilon,$$

and as $\varepsilon > 0$ was arbitrary in (5.12), the claim follows. □

5.3.2. Solving the hierarchy

We show now that Program ER_k can be reformulated as a (generalized version of) a so-called *log-det* minimization program. With semidefinite constraints, log-det minimization programs have the form

$$\begin{aligned} \min \quad & \log\det X \\ \text{s.t.} \quad & \mathcal{A}(X) = b \\ & X \succeq 0 \end{aligned} \tag{LOGDET}$$

with matrix decision variable X, a linear map $\mathcal{A} : \mathcal{S}^n \to \mathbb{R}^m$, and $b \in \mathbb{R}^m$[FHB03]. It turns out that Program ER_k is slightly more general and we need to allow the minimiza-

tion of the determinant of a submatrix of X. The program takes the form

$$\begin{aligned} \min \quad & \log\det X_1 \\ \text{s.t.} \quad & \mathcal{A}(X) = b \\ & X = \begin{pmatrix} X_1 & X_2 \\ X_2^T & X_3 \end{pmatrix} \succeq 0 \end{aligned} \tag{LOGDET+}$$

Before we explore the relation to Program ER_k, let us note that LOGDET+ is a concave program.[4] The objective is common in convex optimization and known to be concave:

Lemma 5.17 (see, e.g., Lemma 1.4.2 in [Hel00])**.** *The function*

$$\mathcal{S}^n_{++} \to \mathbb{R}, \quad X \mapsto \log\det X$$

is concave.

Observation 5.18. *Program* LOGDET+ *is a concave minimization problem.*

Proof. By Lemma 5.17, the objective is concave. Since the feasible set is an intersection of half-spaces with the positive semidefinite cone, it is convex. The claim follows. □

The aim of the remainder of this section is to show that Program ER_k is essentially of the form LOGDET+, if $\mathcal{Q}'$ and $\mathcal{X}$ are spectrahedra (see Section 1.4.12). But first we need to address a technical detail: The constraint set

$$\mathcal{Q}' = \{Q^{-1} : Q \in \mathcal{Q}_{\succ 0}\} = \{Q^{-1} : Q \in \mathcal{Q}, Q \succ 0\}$$

is not closed since $0 \notin Q'$, but spectrahedra are. We thus need to require $\overline{\mathcal{Q}'}$ to be a spectrahedron – that is, if we identify $\mathbb{R}^{n\times n} \cong \mathbb{R}^{n^2}$, then $\overline{\mathcal{Q}'}$ is a projected spectrahedron in $\mathbb{R}^{n^2}$ –, and optimize over $\overline{\mathcal{Q}'}$ instead of over $\mathcal{Q}'$. Let us note that this does not change the problem:

Observation 5.19. *Any matrix $P \in \overline{\mathcal{Q}'} \setminus \mathcal{Q}'$ is singular.*

Proof. Let $P \in \overline{\mathcal{Q}'} \setminus \mathcal{Q}'$ be given. Hence there is a sequence $\{P_n\}_{n\in\mathbb{N}} \subset \mathcal{Q}'$ with $P_n \to P$ for $n \to \infty$. Suppose to the contrary P is nonsingular. Thus $P \succ 0$, since P lies in the closed set $\mathcal{S}^n_+$. Since inversion is a continuous operation over the set of all nonsingular matrices, $P_n^{-1} \to P^{-1}$. It follows $P^{-1} \in \mathcal{Q}$, since the latter is closed by Assumption 5.1. As $P^{-1} \succ 0$, we have $(P^{-1})^{-1} = P \in \mathcal{Q}'$, in contradiction to the assumption. □

In other words, any $P \in \overline{\mathcal{Q}'}$ has objective value $\log\det(P) = -\infty$, and can geometrically be interpreted as an "ellipsoid" of vanishing volume (corresponding to Program E1 with optimal value 0); this can only happen in the degenerated case that S does not have maximal affine dimension.

We may now state above-mentioned equivalence:

[4] A convex (concave) programming problem is a minimization problem of a convex (concave) objective subject to convex constraints.

Theorem 5.20. *Suppose $\overline{\mathcal{Q}'}$ and $\mathcal{X}$ are projected spectrahedra. Then Program* ER_k *can be reformulated as a program of the form* LOGDET+.

Proof. By assumption on $\overline{\mathcal{Q}'}$, there are $m_{\mathcal{Q}'}, n_{\mathcal{Q}'} \in \mathbb{N}$ and matrices $B_{ij}, B'_i \in \mathcal{S}^{k_{\mathcal{Q}'}}$ with

$$\overline{\mathcal{Q}'} = \left\{ P \in \mathbb{R}^{n \times n} : \sum_{i,j}^{n} P_{ij} B_{ij} + \sum_{i=1}^{n_{\mathcal{Q}'}} y_i B'_i \succeq 0 \text{ for some } y \in \mathbb{R}^{n_{\mathcal{Q}'}} \right\}. \tag{5.13}$$

Similarly, by assumption on $\mathcal{X}$, there are $m_{\mathcal{X}}, n_{\mathcal{X}} \in \mathbb{N}$ as well as matrices $C_i, C'_i \in \mathcal{S}^{k_{\mathcal{X}}}$ with

$$\mathcal{X} = \left\{ x_0 \in \mathbb{R}^n : \sum_{i=1}^{n} x_{0,i} C_i + \sum_{i=1}^{n_{\mathcal{X}}} y'_i C'_i \succeq 0 \text{ for some } y' \in \mathbb{R}^{n_{\mathcal{X}}} \right\}. \tag{5.14}$$

We have just seen (Observation 5.19) that in Program ER_k, the set $\mathcal{Q}'$ can be replaced by $\overline{\mathcal{Q}'}$. The constraint

$$P \in \overline{\mathcal{Q}'}$$

is thus by (5.13) equivalent to the semidefinite constraint (in dual form, see Section 1.5.6)

$$\begin{aligned} \sum_{i,j}^{n} P_{ij} B_{ij} + \sum_{i=1}^{n_{\mathcal{Q}'}} y_i B'_i &\succeq 0 \\ P \in \mathbb{R}^{n \times n}, \quad y &\in \mathbb{R}^{n_{\mathcal{Q}'}}. \end{aligned}$$

Similarly, the constraint $x_0 \in \mathcal{X}$ can be rewritten using (5.14) as a semidefinite constraint (in dual form).

The constraint in (5.8), given by

$$1 - \begin{pmatrix} X \\ 1 \end{pmatrix}^T A \begin{pmatrix} X \\ 1 \end{pmatrix} \in M(h_1, \ldots, h_s)[k]$$

involving the quadratic module is equivalent to an sos constraint and additional decision variables as detailed in (1.16). By Corollary 1.31 again, the constraint is thus equivalent to a semidefinite constraint (in primal form). Constraint 5.9, that is,

$$\begin{pmatrix} P & (I_n, -x_0) \\ (I_n, -x_0)^T & A \end{pmatrix} \succeq 0$$

is a semidefinite constraint (in dual form).

Now, by Observation 1.29, all constraints in dual form can be equivalently reformulated in primal form, and aggregated into a single constraint (Observation 1.9). The claim follows. □

6. Unconstrained mixed-integer polynomial optimization

This chapter is devoted to a subclass of MIPP: Mixed-integer polynomial optimization in absence of constraints. We review a condition that ensures the existence of minimizers for this subclass. In case this condition holds, we compute norm bounds from Chapter 3 on the optimal solution, which makes the problem accessible to branch and bound, especially the all-integer case. Furthermore, we derive a new class of underestimators of the polynomial objective function. Using a result from real algebraic geometry and again sos programming, we optimize over this class to get a strong lower bound on the mixed-integer minimum. Our lower bounds are evaluated experimentally for the all-integer case. They show good performance, in particular within a branch and bound framework.

Section 6.1 introduces the problem class, outlines the solution process and recalls the condition that ensures the existence of minimizers.

Section 6.2 motivates, introduces and illustrates our class of underestimators. We show that the lower bound they provide on the mixed-integer minimum gives rise to a linear function that can be used as objective in an sos program. We proceed by casting the task to find the best global underestimator as an sos program. It turns out that it is sufficient for our purposes to use underestimators on sublevel sets, and refine the sos program accordingly.

Section 6.3 is about the solution of random instances with branch and bound. To this end, we give the solution process in algorithmic form. To evaluate the performance of our underestimators in a branch and bound framework, we present other lower bounds from the literature. We implemented our underestimators and the lower bounds from the literature and report on the runtimes.

6.1. Introductory remarks

6.1.1. The problem class

In this section, we consider the unconstrained mixed-integer optimization problem of a polynomial objective function $f \in \mathbb{R}[X_1, \ldots, X_n]$, that is, the problem

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x_i \in \mathbb{Z}, \quad i \in \mathcal{I}, \\ & x \in \mathbb{R}^n, \end{aligned} \tag{UMIPP}$$

and the all-integer variant

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in \mathbb{Z}^n. \end{aligned} \tag{UIPP}$$

In the following, we outline the solution process.

6.1.2. Discussion of the solution process

We have seen that, even though Program UIPP is a restricted special case of MIPP, it is already undecidable in general (Theorem 1.39). We therefore concentrate on a class of objective functions that satisfy a coercivity condition which is sufficient for the existence of minimizers. If the leading form of f is positive definite, f has mixed-integer minimizers – a fortiori, f has integer minimizers – by Proposition 1.32. As deciding positive definiteness is NP-hard (Theorem 1.38), we approximate this problem by sos programming. We outline the details below.

However, positive definiteness of f only tells us that f has mixed-integer minimizers, but not where they are located. We locate the minimizers by computing the radius of a p-norm ball that contains all minimizers, that is, the norm bound on the minimizers of f as introduced in Chapter 3. Since norm bounds require a feasible solution, we choose the point $q = 0$ which is always feasible in the absence of constraints.

This section is devoted to the actual solution of UMIPP for a given polynomial f, where, for simplicity, we restrict minimization at some point to the all-integer case UIPP. To this end, we present algorithms and demonstrate that they are actually implementable by sampling from a family of random polynomials and carrying out all described algorithmic steps.

In principle, once a norm bound is known, UIPP is solvable by enumeration. However, to find the optimal solution to UIPP, instead of enumeration, we use a branch and bound scheme. As indicated in the introduction (Chapter 1), an effective branch and bound procedure has two key ingredients: Tight lower bounds and a small search tree. To find tight lower bounds, we introduce a class of polynomials with obvious integer minimizer that serve as underestimators to f. Using sos programming, we may choose the underestimator g with the strongest lower bound. Firstly, we search for a global underestimator which is later refined to underestimation on sublevel sets, yielding stronger bounds. This refinement further allows to prove that, provided f has a positive

definite leading form, there always are underestimators in our class that can be found by sos programming. Concerning the small search tree, we compute our norm bounds, and show that they outperform the bounds from the literature by orders of magnitude.

The branch and bound scheme consists of two algorithms. The first algorithm decides whether f suffices the coercivity condition, and if f does, computes a norm bound on f, making the problem accessible to branch and bound. The second algorithm computes a suitable underestimator g for f from our class of underestimators. It then proceeds to the actual branch and bound part and uses the underestimator g, in conjunction with the norm bound for f, to minimize f over the integer lattice.

In our experiments, we compare the performance of our underestimators with underestimators from the literature as well as the classical approaches, namely, continuous relaxation and brute force enumeration.

6.1.3. A note on the coercivity condition

We indicated that we rely on the sufficient condition $f_d > 0$ to ensure the existence of integer minimizers. As deciding nonnegativity of the leading form f_d is NP hard, even for degree four (Theorem 1.38), we compute a lower bound c_d on the leading form f_d restricted to the sphere, i.e.,

$$c_d \leq c_d^* = \min_{x \in \mathbb{S}_p^{n-1}} f_d(x). \tag{6.1}$$

If $c_d > 0$ we know from (1.3) that $f_d > 0$, so integer minimizers exist by Proposition 1.32. Our approach fails if $c_d \leq 0$ unless we find a point $x \in \mathbb{S}_p^{n-1}$ with $f_d(x) < 0$ which certifies that f_d is not positive semidefinite, and hence f cannot have minimizers by Proposition 1.33.

6.2. Underestimation

Given $f, g : \mathbb{R}^n \to \mathbb{R}$ and $U \subset \mathbb{R}^n$, we say[1] g is an *underestimator of f on U* if

$$g(x) \leq f(x) \quad \text{for all } x \in U.$$

Let us mention two important special cases that appear in this chapter: In case $U = \mathbb{R}^n$, the function g is a *global underestimator*; if $U = \mathcal{L}^f_{\leq}(z)$ for some $z \in \mathbb{R}$, g is an *underestimator on a sublevel set*.

The primary motivation for global underestimators is that they yield *lower bounds* on the mixed-integer minimum, since

$$(\forall x \in \mathbb{R}^n : g(x) \leq f(x)) \Longrightarrow \inf_{x \in F_{\mathcal{I}}} g(x) \leq \inf_{x \in F_{\mathcal{I}}} f(x) \tag{6.2}$$

where $\inf_{x \in F_{\mathcal{I}}} g(x)$ gives a stronger bound on the mixed-integer minimum of f than $\inf_{x \in \mathbb{R}^n} g(x)$. Using the mixed-integer minimum of g to derive a lower bound on the mixed-integer minimum of f makes only sense if mixed-integer minimization of g is easy compared to mixed-integer minimization of f. In Section 6.2.3, we show that lower bounds on the mixed-integer minimum of f can still be computed by considering functions g that underestimate f only on a sublevel set (of f) and possibly not all of $\mathbb{R}^n$.

We outline now our class of underestimators and show that the problem to choose the underestimator with the strongest lower bound can be cast as an sos problem.

6.2.1. A class of underestimators

We motivate our class of easy-to-minimize underestimators g with an observation on monomials with a shift in the argument which shall serve as the building blocks to the more general underestimators.

Observation 6.1. *For some $h \in \mathbb{R}^n$ and $\alpha \in \mathbb{N}^n_0$, let*

$$g = (X - h)^\alpha = \prod_{j=1}^{n} (X_j - h_j)^{\alpha_j}$$

be a shifted monomial. *If all α_i are even, g has a mixed-integer minimizer at $\lfloor h \rceil_{\mathcal{I}}$ (and thus a continuous minimizer at h and an integer minimizer at $\lfloor h \rceil$). If one α_i is odd, g is not bounded from below and does not have mixed-integer minimizers.*

Our underestimators are conic combinations of shifted monomials with even α_j, $j = 1, \ldots, n$, as the combinations inherit the mixed-integer minimizer $\lfloor h \rceil_{\mathcal{I}}$. More precisely:

Proposition 6.2. *Let a polynomial $g \in \mathbb{R}[\underline{X}]$ be given as $g = \sum_\alpha b_\alpha (X - h)^{2\alpha}$ with $b_\alpha \geq 0$ for $\alpha \neq 0$, and $h \in \mathbb{R}^n$.*

[1] Synonymously, we may say *g underestimates f on U*.

1. *The restriction of g to $\prod_{i=1}^{k-1}\{x_i\} \times \mathbb{R} \times \prod_{i=k+1}^{n}\{x_i\}$ that is, the univariate function $y \mapsto g(x_1, \ldots, x_{k-1}, y, x_{k+1}, \ldots, x_n)$ for fixed $x \in \mathbb{R}^n$ is nonincreasing for $y \leq h_k$ and nondecreasing for $y \geq h_k$, $k \in \{1, \ldots, n\}$.*

2. *We have*
$$g(x_1, \ldots, x_n) \geq g(x_1, \ldots, x_{k-1}, \lfloor h_k \rceil, x_{k+1}, \ldots, x_n)$$
for every $x \in F_{\mathcal{I}} = \{x \in \mathbb{R}^n : x_i \in \mathbb{Z} \text{ for } i \in \mathcal{I}\}$ and $k \in \mathcal{I}$.

3. *The point $\lfloor h \rceil_{\mathcal{I}}$ is a mixed-integer minimizer of g.*

Proof. It is enough to show the claimed properties for a term of the form $(X - h)^{2\alpha}$, as g is a conic combination of such terms. Claim 1 is elementary and immediately implies Claim 2. Applying the latter repeatedly gives Claim 3. □

Three properties make these polynomials g useful underestimators: Firstly, mixed-integer minimization is trivial – this is inspired by the so-called *rounding property* [HS14], i.e. rounding of a continuous minimizer yields an integer minimizer. Secondly, all nonlinearity is confined to the parameter h. Thirdly, the fact that the expression is linear in the b_α makes them accessible to optimization.

Proposition 6.2 motivates

Notation 6.3. We denote the set of conic combinations of monomials with a shift of h by

$$\mathscr{C}(h) := \left\{ g \in \mathbb{R}[\underline{X}] : g = \sum_{\alpha \in J} b_\alpha (X - h)^{2\alpha},\ b_\alpha \in \mathbb{R}_{\geq 0} \text{ for all } \alpha \neq 0, J \subset \mathbb{N}_0^n \text{ finite} \right\}.$$

As an example, the polynomial

$$g = (X_1 - 1.5)^4 (X_2 - 2)^6 + 0.3(X_1 - 1.5)^2 (X_3 - 3.2)^8 - 1 \in \mathscr{C}(1.5, 2, 3.2)$$

with $J = \{(2, 3, 0),\ (1, 0, 4),\ (0, 0, 0)\}$ has an integer minimizer at $(1, 2, 3)$.

Proposition 6.4. *Let $g \in \mathscr{C}(h)$ satisfy $g(x) \leq f(x)$ for all $x \in \mathbb{R}^n$. Then*

$$g(\lfloor h \rceil_{\mathcal{I}}) \leq \inf_{x \in F_{\mathcal{I}}} f(x)$$

Proof. This follows from (6.2) and Proposition 6.2. □

For determining an underestimator g we still have to choose h and the coefficients b_α. This is described next.

Choice of h: In principle, every $h \in \mathbb{R}^n$ may be chosen. Heuristically, we chose an approximate continuous minimizer of f since g has its continuous minimizer at h. In fact, every nontrivial g looks like an elliptic paraboloid or a parabolic cylinder near h, as does f near every local minimizer. For almost all f, the continuous minimizer of f can be found using sos methods (Theorem 1.28).

Choice of b_α: We choose the b_α so that the lower bound $g(\lfloor h \rceil_{\mathcal{I}})$ is maximized. In other words, we wish to maximize the expression

$$g(\lfloor h \rceil_{\mathcal{I}}) = \sum_{\alpha \in J} b_\alpha (\lfloor h \rceil_{\mathcal{I}} - h)^{2\alpha}$$

subject to $g \leq f$. The higher order terms in g ensure a certain aggressiveness in the growth behavior away from h, even for small coefficients b_α, which leads to strong bounds.

6.2.2. Global underestimation

Using the notation $w_\alpha := (\lfloor h \rceil_{\mathcal{I}} - h)^{2\alpha}$, we get the following optimization problem:

$$\begin{aligned} \max_{J,\ b_\alpha} \quad & \sum_{\alpha \in J} w_\alpha b_\alpha \\ \text{s.t.} \quad & f(x) - \sum_{\alpha \in J} b_\alpha (x-h)^{2\alpha} \geq 0 \quad \forall x \in \mathbb{R}^n \\ & b_\alpha \geq 0 \quad \text{for } \alpha \neq 0 \end{aligned}$$

with decision variables $b_\alpha \in \mathbb{R}$, $\alpha \in J$ and $J \subset \mathbb{N}_0^n$ finite. Since this program is not tractable in general, we consider the following sos version instead:

$$\begin{aligned} y = \max \quad & \sum_{\alpha \in J} w_\alpha b_\alpha && && \text{(GLOB)} \\ \text{s.t.} \quad & f - \sum_{\alpha \in J} b_\alpha (X-h)^{2\alpha} && \text{is sos in } \mathbb{R}[X_1, \ldots, X_n], \\ & b_\alpha && \text{is sos in } \mathbb{R}[X_1, \ldots, X_n] \text{ for } \alpha \neq 0. \end{aligned}$$

The decision variables are the real b_α, $\alpha \in J$. Note that $b_\alpha \in \Sigma$ is equivalent to $b_\alpha \geq 0$. Once J is fixed, GLOB is a valid sos program. We show in Corollary 6.9 that it is sufficient to choose $J = \{\alpha \in \mathbb{N}_0^n : |\alpha| \leq \deg(f)/2\}$.

In the following we identify a solution b_α, $\alpha \in J$, with the polynomial g it defines, that is with $g = \sum_{\alpha \in J} b_\alpha (X-h)^{2\alpha}$, and hence may say that a polynomial is a feasible or optimal solution to GLOB. We note that every feasible solution to GLOB (for any choice of h) gives valid lower bounds on UMIPP:

Theorem 6.5. *Let $f \in \mathbb{R}[\underline{X}]$, $h \in \mathbb{R}^n$ and $g = \sum_{\alpha \in J} b_\alpha (X-h)^{2\alpha} \in \mathscr{C}(h)$ be a feasible solution to* GLOB *for some J. Then*

1. *$g(\lfloor h \rceil_{\mathcal{I}}) \leq \inf_{x \in F_{\mathcal{I}}} f(x)$.*

If moreover $f - f(h) \in \Sigma$ holds and g is an optimal solution to GLOB*, then*

2. $g(\lfloor h \rceil_{\mathcal{I}}) \geq f(h)$.

Proof. Claim 1 holds as g being feasible to GLOB implies $f - g \in \Sigma$, hence $f - g \geq 0$, and the claim follows by Proposition 6.4. Concerning Claim 2, observe that $f - f(h) \in \Sigma$ implies that h is a continuous minimizer of f and that the constant polynomial $\tilde{g} = f(h)$ is a feasible solution to GLOB, hence $g(\lfloor h \rceil_{\mathcal{I}}) \geq \tilde{g}(\lfloor h \rceil_{\mathcal{I}}) = f(h)$ for every optimal solution $g \in \mathscr{C}(h)$. □

We note that GLOB is feasible if and only if f is sos-bounded from below, i.e. if there is $r \in \mathbb{R}$ with $f - r \in \Sigma$. Indeed, if $g = \sum_{\alpha \in J} b_\alpha (X - h)^{2\alpha}$ is feasible for GLOB, then $f - g \in \Sigma$, and hence $f - g + \sum_{\alpha \in J, \alpha \neq 0} b_\alpha (X - h)^{2\alpha} = f - b_0 \in \Sigma$. For the converse direction, if $f - r \in \Sigma$, then $g := r$ is a feasible solution to GLOB.

6.2.3. Underestimation on sublevel sets

Motivation

A quite restrictive condition in GLOB is that it requires $g(x) \leq f(x)$ globally, i.e., for all $x \in \mathbb{R}^n$. Actually, this is not necessary for our purposes. It is enough to require $g(x) \leq f(x)$ only for those $x \in \mathbb{R}^n$ that satisfy $f(x) \leq f(q)$ for some $q \in F_{\mathcal{I}}$. That is, for all $q \in F_{\mathcal{I}}$, we have

$$\Big(\forall x \in \mathcal{L}^f_{\leq}(f(q)) : g(x) \leq f(x)\Big) \Longrightarrow \inf_{x \in F_{\mathcal{I}}} g(x) \leq \inf_{x \in F_{\mathcal{I}}} f(x), \tag{6.3}$$

in other words, the mixed-integer minimum of g is a lower bound on the mixed-integer minimum of f even if g is an underestimator of f only on a sublevel set $\mathcal{L}^f_{\leq}(f(q))$. If we make use of this in our sos program, the lower bound can only improve.

But before we delve into the details, let us consider the potential payoff by taking a look at the integer minimization example in Figure 6.1a. The plot depicts the univariate polynomial

$$f = 0.2 \cdot (X - 0.3)^6 - 5 \cdot (X - 0.3)^4 + 32 \cdot (X - 0.3)^2.$$

along with two underestimators g_{GLOB}, g_{SLS}. A short calculation shows that f has five local extrema at 0.3 and $0.3 \pm \sqrt{\frac{25 \pm \sqrt{145}}{3}}$, and that the local minimizers are at $x = 0.3$ and at $x_\pm = 0.3 \pm \sqrt{\frac{25+\sqrt{145}}{3}} \approx 0.3 \pm 3.51$. Considering that f has a positive definite leading form, one of the local minimizers must be a global one, and comparing the function values shows that $x = 0.3$ is the continuous minimizer. Moreover, f must have its integer minimizer in $[-3, 3]$ as $\min\{f(x_+), f(x_-)\} > f(0)$; comparing the function values shows that f has a single integer minimizer at $x = 0$ with value $f(0) \approx 2.84$. The underestimator $g_{\text{GLOB}} \in \mathscr{C}(h)$, computed as optimal solution to GLOB is given by[2]

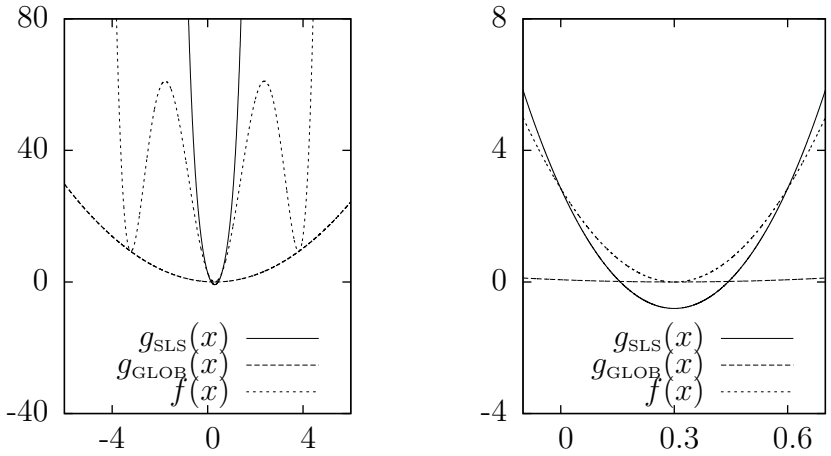
$$g_{\text{GLOB}} \approx 8.71 \cdot 10^{-11} \cdot (X - 0.3)^6 + 1.09 \cdot 10^{-09} \cdot (X - 0.3)^4 + 0.75 \cdot (X - 0.3)^2 - 1.22 \cdot 10^{-09},$$

[2]For this example we solved GLOB for $h = 0.3$ and $\deg g = 6$, using SOSTOOLS 3.00 and CSDP 6.1.0.

is globally below f. To find an underestimator on a sublevel set, we first fix the level $z = f(q)$ heuristically. Note that any $q \in \mathbb{Z}$ is a feasible solution to UIPP and hence an upper bound; any integer minimizer must be contained in $\mathcal{L}^f_{\leq}(f(q))$. As $h = 0.3$ is the global minimizer, we choose $q = \lfloor h \rceil_{\mathcal{I}} = 0$ here. The polynomial g_{SLS}, given by

$$g_{\text{SLS}} \approx 9.09 \cdot (X - 0.3)^6 + 11.80 \cdot (X - 0.3)^4 + 39.36 \cdot (X - 0.3)^2 - 0.81,$$

is an underestimator on the sublevel set $\mathcal{L}^f_{\leq}(f(0)) = [0, 0.6]$, as can be seen in Figure 6.1b. In the next section, we show show how this function can be found. The plot reveals the shortcomings of global underestimation: Any *global* underestimator in $\mathscr{C}(0.3)$ cannot go above the local minimizers of f. This "barrier" from above turns g_{GLOB} in this example essentially into a quadratic underestimator for small x as the ratio of the higher order coefficients and the one in front of the quadratic term is of order 10^{-10}. The underestimator g_{SLS} however is a degree 6 polynomial whose higher order coefficients are not small at all. Note that g_{GLOB} is much closer to f near 0.3 compared to the new underestimator g_{SLS}. However, the quality of the resulting lower bound depends on the function values at 0 and there g_{SLS} is closer to f than g_{GLOB}. The lower bounds the two underestimators provide are $g_{\text{GLOB}}(0) \approx 0.07$ and $g_{\text{SLS}}(0) \approx 2.84$. In this case, we are lucky as the lower bound on the integer minimum and $f(0)$ coincide, showing once more that f has its integer minimizer at 0.



(a) From far (b) From close

Figure 6.1.: Global underestimator g_{GLOB} and an underestimator g_{SLS} on a sublevel set.

6.2.4. The sos program for computing the improved underestimator

How do we compute the improved underestimator? At first, we observe that every sublevel set $\mathcal{L}^f_{\leq}(z)$, $z \in \mathbb{R}$, of f is semi-algebraic. Indeed, with the notation from (1.11), we have

$$\mathcal{L}^f_{\leq}(z) = \{x \in \mathbb{R}^n : z - f(x) \geq 0\} = K(z - f).$$

Moreover, $\mathcal{L}^f_{\leq}(z)$ is compact if the leading form of f is positive definite (see Proposition 1.32). Compactness of $\mathcal{L}^f_{\leq}(z)$ in turn implies that the quadratic module $M(f-g, z-f)$, for any $g \in \mathbb{R}[X_1, \ldots, X_n]$, is Archimedean by Theorem 1.17 (4). Hence, for every feasible underestimator $g \in \mathscr{C}(h)$ the existence of a representation for $f-g$ as in Putinar's Positivstellensatz (Theorem 1.20) is guaranteed. This motivates the following program:

$$
\begin{aligned}
y^{(k)} = \max \quad & \sum_{\alpha \in J} w_\alpha b_\alpha && && \text{(SLS)}\\
\text{s.t.} \quad & f - \sum_{\alpha \in J} b_\alpha (X-h)^{2\alpha} - \sigma(z-f) && \text{is sos in } \mathbb{R}[X_1, \ldots, X_n],\\
& b_\alpha \text{ for } \alpha \neq 0,\ \sigma && \text{are sos in } \mathbb{R}[X_1, \ldots, X_n],\\
& \deg \sigma \leq k
\end{aligned}
$$

The decision variables are the real b_α as for GLOB and, additionally, the real coefficients of the polynomial σ. As before, we use the notation $w_\alpha := (\lfloor h \rceil_{\mathcal{I}} - h)^{2\alpha}$. SLS is a valid sos program once J and the degree of σ are fixed.

Theorem 6.6. *Let $f \in \mathbb{R}[\underline{X}]$, $h \in \mathbb{R}^n$ and $g \in \mathscr{C}(h)$ be a feasible solution to* SLS *with $z \geq f(q)$ for some $q \in F_{\mathcal{I}}$.*

1. *Then $g(\lfloor h \rceil_{\mathcal{I}}) \leq \inf_{x \in F_{\mathcal{I}}} f(x)$.*
2. *If J is fixed, $y^{(-\infty)} \leq y^{(0)} \leq y^{(2)} \leq y^{(4)} \leq \ldots$*[3]
3. *If f is coercive, there is $k_0 \in \mathbb{N}_0$ such that* SLS *is feasible for all $k \geq k_0$.*
4. SLS *with $k = -\infty$ is* GLOB.
5. *If $f - f(h) \in \Sigma$ and g is optimal, then $g(\lfloor h \rceil_{\mathcal{I}}) \geq f(h)$.*

Proof. Statement 1 holds as g feasible implies $f-g-\sigma(z-f) \in \Sigma$. Hence $f(x)-g(x) \geq 0$ for those x with $f(x) \leq z$, especially for those x with $f(x) \leq f(q)$ as $f(q) \leq z$ by assumption. The claim follows by (6.3).

Statement 2 is clear as we only allow more coefficients for σ.

To see Statement 3, note that $\mathcal{L}^f_{\leq}(z)$ is nonempty as $z \geq f(q)$ and moreover compact (Proposition 1.32), so $f(x) > c$ for some $c \in \mathbb{R}$ and all $x \in \mathcal{L}^f_{\leq}(z)$. Hence $f-c \in M(z-f)$ by Putinar's Positivstellensatz (Theorem 1.20). This means $f - c = \sigma_0 + \sigma(z-f)$ for some sos $\sigma_0, \sigma \in \mathbb{R}[\underline{X}]$. Thus $g := c$ is a feasible solution, and $k_0 := \deg \sigma$.

To see Statement 4, we note that $k = -\infty$ corresponds to $\sigma = 0$, in which case SLS is GLOB.

Statement 5 is a consequence of Statements 2 and 4 and Theorem 6.5. □

[3]Note that every sos polynomial $\sigma \neq 0$ has even degree.

We have not yet addressed the degree of g in GLOB and SLS nor the degree of σ in SLS. The following proposition shows that once the degree of σ in SLS is fixed, the degree of g in any feasible solution is bounded from above in terms of $\deg f$ and $\deg \sigma$. Let us note at first the following technical lemma that we need for its proof.

Lemma 6.7 (see, e.g., Corollary 1.1.3 in [Mar08])**.** *Suppose $v = u_1^2 + \cdots + u_k^2$ for some given $u_1, \ldots, u_k \in \mathbb{R}[\underline{X}]$ and $u_1 \neq 0$. Then $v \neq 0$, and*

$$\deg v = 2 \max_{1 \leq i \leq k} \deg u_i.$$

Proposition 6.8. *Let $f \in \mathbb{R}[\underline{X}]$, $g \in \mathscr{C}(h)$ with $\deg f > 0$, $\deg g > 0$, $z \in \mathbb{R}$ and $\sigma \in \Sigma$ such that*

$$f - g - \sigma(z - f) \text{ is sos.} \tag{6.4}$$

Then

$$\deg(g) \leq \deg(f) + \max\{\deg(\sigma), 0\}.$$

Proof. Eq. (6.4) is equivalent to $f - g - \sigma(z - f) = \sigma_0$ for some $\sigma_0 \in \Sigma$, or

$$g + \sigma_0 = f(1 + \sigma) - z\sigma. \tag{6.5}$$

$$\begin{aligned} \text{Hence} \quad \deg(g) &\leq \max\{\deg(g), \deg(\sigma_0)\} \overset{\text{(I)}}{=} \deg(g + \sigma_0) \overset{\text{(II)}}{=} \deg\left(f(1+\sigma) - z\sigma\right) \\ &\overset{\text{(III)}}{=} \max\{\deg\left(f(1+\sigma)\right), \deg(z\sigma)\} \overset{\text{(IV)}}{=} \deg\left(f(1+\sigma)\right) \\ &\overset{\text{(V)}}{=} \deg(f) + \deg(1+\sigma) \overset{\text{(VI)}}{=} \deg(f) + \max\{\deg(\sigma), 0\}. \end{aligned}$$

As $g - g(h) \in \Sigma$ and $\deg g > 0$, equality (I) follows from Lemma 6.7. Equality in (II) follows from eq. (6.5). Using $\deg f > 0$, the equalities in (III) and (IV) follow from a typical degree argument: If $u, v \in \mathbb{R}[\underline{X}]$, $\deg u \neq \deg v$, we have $u + v \neq 0$ and $\deg(u + v) = \max(\deg u, \deg v)$. Equality in (V) holds as the degree is multiplicative, (VI) follows easily if one distinguishes the cases $\sigma = 0$, $\sigma \in \mathbb{R}_{\geq 0}$ and $\deg \sigma > 0$. □

Corollary 6.9. *Let $g \in \mathscr{C}(h)$ be a feasible solution to* GLOB*. Then $\deg g \leq \deg f$.*

Proof. Use Proposition 6.8 with $\sigma = 0$ and the result follows from Statement 4 of Theorem 6.6. □

6.3. Using and evaluating our underestimators within branch and bound

Note that, for the remainder of this section, we assume the all-integer case: $\mathcal{I} = [n]$. We evaluate the underestimators in a branch and bound framework. Firstly, we present an algorithm that shows how special properties of our underestimators can be exploited to speed up branching and pruning. In the following experiments, we generate random polynomials. For completeness, we sample from families (F1) and (F2) (Sections 3.3.2 and 3.3.3, respectively), but focus on the latter as we get instances with higher n and d: Specifically, for the family (F1) and each of the (n, d)-tuples $(2, 4)$, $(3, 4)$ and $(4, 2)$, we sample instances until we have 50 with detected positive definite leading form (cf. Section 3.3.2); for the second family (F2) and each of the (n, d)-tuples $(2, 6)$, $(3, 4)$, $(3, 6)$, $(4, 4)$ and $(5, 6)$ for $K = 2$, we sample 50 instances (cf. Section 3.3.3). We compare the resulting runtimes in a branch and bound framework with other lower bounds from the literature and conclude with an evaluation of the initial lower bound $g(\lfloor h \rceil)$.

6.3.1. Algorithm

We present the implementation of our underestimators in a branch and bound framework in Algorithm 2. The algorithm can be summarized as follows. New subproblems are chosen depth first. This keeps memory usage small and allows us to quickly obtain good feasible solutions. We do not reorder the variables. Subproblems are collected in a list $\mathcal{L}$; every subproblem $\mathcal{P} \in \mathcal{L}$ is of the form $\mathcal{P} = (m, r_1, \ldots, r_m)$, where $m \in \{0, \ldots, n\}$ encodes the number of fixed variables $(r_1, \ldots, r_m) \in \mathbb{Z}^m$; i.e.,

$$\begin{aligned} \min \quad & f(r_1, \ldots, r_m, x_{m+1}, \ldots, x_n) \\ & x_{m+1}, \ldots, x_n \in \mathbb{Z} \end{aligned} \qquad (\mathcal{P} = (m, r_1, \ldots, r_m))$$

and (0) encodes the initial problem. At every subproblem $(m, r_1, \ldots, r_m)$, we get a univariate underestimator $\tilde{g}$ from the original underestimator g by fixing the first m variables $x_1, \ldots, x_m$ to the values $r_1, \ldots, r_m$ from the subproblem, and by fixing the last $n - m + 1$ free variables $x_{m+2}, \ldots, x_n$ to $\lfloor h_{m+2} \rceil, \ldots, \lfloor h_n \rceil$. Now suppose g is a solution to GLOB or SLS. Then, for some $q \in \mathbb{Z}^n$, $f(x) \geq g(x)$ for those $x \in \mathbb{R}^n$ with $f(x) \leq f(q)$. Using Proposition 6.2,

$$\begin{aligned} f(r_1, \ldots, r_m, x_{m+1}, x_{m+2}, \ldots, x_n) &\geq g(r_1, \ldots, r_m, x_{m+1}, x_{m+2}, \ldots, x_n) \\ &\geq g(r_1, \ldots, r_m, x_{m+1}, \lfloor h_{m+2} \rceil, \ldots \lfloor h_n \rceil) = \tilde{g}(x_{m+1}) \end{aligned}$$

for all $x \in \mathbb{Z}^n$ with $f(x) \leq f(q)$. This means that all subproblems with $\tilde{g}(x_{m+1})$ larger than the current upper bound u can be pruned – and this holds, trivially, also for those x with $f(x) > f(q)$. In short, if $\tilde{g}(x_{m+1}) > u$, the subproblem $\mathcal{P}' = (m+1, r_1, \ldots, r_m, x_{m+1})$ of $\mathcal{P}$ can be pruned. Since $\tilde{g}$ is nondecreasing for $x_{m+1} \geq h_{m+1}$ and nonincreasing for $x_{m+1} \leq h_{m+1}$, the set of all these subproblems of $\mathcal{P}$, whose x_{m+1}-coordinate must lie in a subinterval of $[-R, R]$, can be identified by a binary search:

Proposition 6.10. *1. Algorithm 2 is correct, that is, it always terminates after a finite number of steps with an optimal integer solution x^* that satisfies $f(x^*) = u$.*

2. The integers L_1 and L_2 (in lines 13 & 15) can be found with binary search in $\lceil \log_2(L) \rceil + 2 \leq \lceil \log_2(R) \rceil + 2$ evaluations of $\tilde{g}$ if $L > 0$.

Proof. Let x^* be any optimal solution. To prove 1. it suffices to show that the algorithm terminates and no problem with (n, x^*) as subproblem gets pruned in Step 12 or lost in Step 19. To see termination of the algorithm, we observe that the number of subproblems is finite as the sets $B_m = \{y \in \mathbb{Z}^m : \|y\|_p \leq R\}$, $m = 1, \ldots, n$ are finite, every subproblem $(m, r_1, \ldots, r_m)$ suffices $(r_1, \ldots, r_m) \in B_m$ and no subproblem is inserted into the list $\mathcal{L}$ more than once. To see that x^* does not get discarded in Step 12, define

$$\tilde{g}(x_{m+1}) := g(x_1^*, \ldots, x_m^*, x_{m+1}, \lfloor h_{m+2} \rceil, \ldots, \lfloor h_n \rceil) \tag{6.6}$$

and suppose $\tilde{g}(\lfloor h_{m+1} \rceil) > u$. Hence

$$\tilde{g}(\lfloor h_{m+1} \rceil) > u \geq f(x^*) \geq g(x^*) \geq g(x_1^*, \ldots, x_m^*, \lfloor h_{m+1} \rceil, \ldots, \lfloor h_n \rceil) = \tilde{g}(\lfloor h_{m+1} \rceil),$$

a contradiction, where we used the monotonicity property of g (Proposition 6.2) and that $g(x) \leq f(x)$ for $x \in \mathcal{L}^f_{\leq}(f(q))$, a fortiori for $x \in \mathcal{L}^f_{\leq}(f(x^*))$. Suppose that x^* gets lost in Step 19. Necessarily, $x^*_{m+1} < L_1$ or $x^*_{m+1} > L_2$. We derive a contradiction for $x^*_{m+1} < L_1$, the other case is identical. Observe that $x^*_{m+1} \in [-L, L]$ as every optimal integer solution satisfies $\sum_{j=1}^n |x_j^*|^p \leq R^p$, so we must have $|x^*_{m+1}| = \sqrt[p]{|x^*_{m+1}|^p} \leq \sqrt[p]{R^p - |x_1^*|^p - \ldots - |x_m^*|^p}$. As x^*_{m+1} is integer, we may round down – in other words, $x^*_{m+1} \in [-L, L]$. By definition of L_1 and Proposition 6.2, we have $\tilde{g}(x^*_{m+1}) > u$ with $\tilde{g}$ from (6.6), thus, using Proposition 6.2 again,

$$\tilde{g}(x^*_{m+1}) > u \geq f(x^*) \geq g(x^*) \geq \tilde{g}(x^*_{m+1}),$$

a contradiction.

We finally show that Claim 2 holds. We prove the claim for $h_{k+1} \geq 0$, the proof for $h_{k+1} \leq 0$ is similar. In case $h_{k+1} > L$, L_1 exists if and only if $\tilde{g}(L) \leq u$ as $\tilde{g}(x_{k+1})$ is non-increasing for $x_{k+1} \leq h_{k+1}$ (by Proposition 6.2); necessarily, $L_2 := L$. Using binary search on $[-L, L]$, L_1 can be found using at most $\lceil \log_2(2L) \rceil = \lceil \log_2(L) \rceil + 1$ further evaluations of $\tilde{g}$. In case $0 \leq h_{k+1} \leq L$, L_1 exists as $\tilde{g}(\lfloor h_{k+1} \rceil) \leq u$ in Step 12. Again using binary search, $L_1 \in [-L, \lfloor h_{k+1} \rceil]$ can be found in no more than $\lceil \log_2(2L) \rceil$ evaluations. As $\tilde{g}(x_{k+1}) = \tilde{g}(h_{k+1} - x_{k+1})$, it only needs at most one more evaluation of $\tilde{g}$ to find L_2, so we find both numbers in no more than $\lceil \log_2(L) \rceil + 2$ evaluations of $\tilde{g}$. □

Remark 6.11. Concerning our implementation, we chose $\deg g = \deg f$ for GLOB and SLS since for large x, we expect a similar order of growth of f and g. Also, we chose $\deg \sigma = 2$ for SLS, since $\deg \sigma = 4$ takes too long in the preprocessing. For the parameter $h \in \mathbb{R}^n$ we chose an (approximate) continuous minimizer computed via the SOSTOOLS function `findbound.m` – however, the algorithm accepts arbitrary $h \in \mathbb{R}^n$. We determined R using Algorithm 1.

Algorithm 2 Branch and Bound

input $f \in \mathbb{R}[X_1, \ldots, X_n]$, $h \in \mathbb{R}^n$, p-norm bound R on minimizers, $k \in 2\mathbb{N}_0$
$x^* \leftarrow \lfloor h \rceil$ *// initial guess for integer minimizer*
$u \leftarrow f(x^*)$ *// upper bound on integer minimum*
$\mathcal{L} \leftarrow \{(0)\}$ *// initial list of subproblems*
find underestimator g: solve SLS with h, $\deg g \leq \deg \sigma = k$ *// or* GLOB*, resp.*
while $\mathcal{L} \neq \emptyset$ **do**
  pick $\mathcal{P} = (m, r_1, \ldots, r_m) \in \mathcal{L}$ with m maximal
  $\mathcal{L} \leftarrow \mathcal{L} \setminus \{\mathcal{P}\}$
  if $m < n$ **then**
    $L \leftarrow \left\lfloor \sqrt[p]{R^p - |r_1|^p - \cdots - |r_m|^p} \right\rfloor$
    let $\tilde{g} : \mathbb{R} \to \mathbb{R}$, $\tilde{g}(x_{m+1}) = g(r_1, \ldots, r_m, x_{m+1}, \lfloor h_{m+2} \rceil, \ldots, \lfloor h_n \rceil)$
    if $\tilde{g}(\lfloor h_{m+1} \rceil) \leq u$ **then** *// otherwise prune*
      find $L_1 \in [-L, L] \cap \mathbb{Z}$ minimal with $\tilde{g}(L_1) \leq u$
      if such an L_1 exists **then**
        find $L_2 \in [-L, L] \cap \mathbb{Z}$ maximal with $\tilde{g}(L_2) \leq u$ *// cf. Proposition 6.10*
      else
        $L_1 \leftarrow +\infty$, $L_2 \leftarrow -\infty$.
      end if
      for all $r_{m+1} \in [L_1, L_2] \cap \mathbb{Z}$ **do** *// $[L_1, L_2] = \emptyset$ if $L_1 = +\infty$*
        $\mathcal{L} \leftarrow \mathcal{L} \cup \{(m+1, r_1, \ldots, r_{m+1})\}$ *// actual branching*
      end for
    end if
  else *// all variables x_i were fixed to values r_i*
    if $f(r) < u$ **then** *// update upper bound*
      $x^* \leftarrow r$
      $u \leftarrow f(r)$
    end if
  end if
end while
output x^*, u
print f `attains its integer minimum` u `at` x^*.

6.3.2. Presentation of other lower bounds

It is not straightforward to compare the performance of our underestimators with lower bounds from the literature. In our setting, we compute a *single* underestimator per instance – which is then merely evaluated during the branch and bound process.[4] We could not find other underestimators with this property that give sensible results in branch and bound. However, there are lower bounds in the literature that are more general than ours since they consider restricted polynomial optimization problems and can hence be applied to any polynomial – not only to those with positive definite leading form – and are suitable for branch and bound if computed new at each node. In addition to Algorithm 2 (with GLOB and SLS) we implemented the following four algorithms in a MATLAB framework for solving UIPP: three of them are branch and bound approaches as Algorithm 2 which use other bounds (taken from [BD14], [LHKW06], and the continuous relaxation) while our last algorithm is a simple brute force approach.

- For arbitrary polynomials on boxes, Buchheim and D'Ambrosio [BD14] suggested to compute, for every term of f, the L^1-best separable underestimator. The sum of the underestimators is again separable, so its integer minimization is a univariate problem. For degree $d \leq 4$ and arbitrary n, they provide explicit underestimators. We hard-coded the explicit underestimators, and used the MATLAB built-ins `polyval`, `polyder` and `roots` to evaluate and differentiate the separable underestimators, and to compute their roots, respectively. As a suitable box at the subproblem $\mathcal{P} = (m, r_1, \ldots, r_m)$ we chose the box $[-L, L]^{n-m}$ where $L = \left\lfloor \sqrt[p]{R^p - |r_1|^p - \cdots - |r_m|^p} \right\rfloor$. The authors suggest to successively halve the box into subboxes which does not fit into our scheme. This approach is abbreviated **SEP** in the plots.

- For nonnegative polynomials on polytopes P, De Loera et al. [LHKW06] approximate the maximum of f on $P \cap \mathbb{Z}^n$ by the sequence $\sqrt[k]{\sum_{x \in P \cap \mathbb{Z}^n} f(x)^k}$. Each member of the sequence can be computed in polynomial time, using a reformulation as a limit of a rational function which in turn is based on the generating function of P. We did experiments with $k = 2$ and $k = 4$, the latter taking significantly longer, without giving much better results, so we restricted ourselves to $k = 2$. Note that the suggested implementation uses residue techniques, while we just use symbolic limit computations. On the other hand, we improved the bounds as follows: To make their approach applicable to not necessarily nonnegative polynomials, the authors suggest to add the sufficiently large constant

  $$c := \|f\|_0 \|f\|_\infty M^d$$

  to obtain $\overline{f} = f + c$ nonnegative on P. Here, $M \geq 0$ is a bound on the polyhedron such that $|x_i| \leq M$ for all $x \in P$; the norm $\|\cdot\|_\infty$ (as well as the norm $\|\cdot\|_1$ below)

[4] By fixing some variables at each node and then computing new underestimators, this could be improved but would need additional runtime for the computation of the new underestimator.

and the "norm" $\|\cdot\|_0$ were introduced for polynomials in Section 1.4.4. However, the constant

$$c' := \sum_{j=0}^{d} \|f_j\|_1 M^j$$

suffices to ensure that $f + c'$ is nonnegative on P. A short calculation shows that $c' \leq c$ if $M \geq 1$, and in dense instances one often has $c' \ll c$. As polyhedron we again chose the box $[-L, L]^{n-m}$ from the previous bound. This bound is abbreviated to **POLY** in the plots.

- We compute an sos approximation of the global continuous relaxation (**CR** in the plots) at each subproblem $\mathcal{P} = (m, r_1, \ldots, r_m)$, that is

$$\begin{aligned} \max \quad & \lambda \\ \text{s.t.} \quad & f(r_1, \ldots, r_m, X_{m+1}, \ldots, X_n) - \lambda \text{ is sos in } \mathbb{R}[X_{m+1}, \ldots, X_n] \end{aligned}$$

- Brute force enumeration with no lower bounds, abbreviated **BF**. As f has to be evaluated at each node, we use `matlabFunction` to convert the Symbolic Math Toolbox object that encodes f into a function handle that can be evaluated significantly faster.

- Algorithm 2 using **GLOB** with parameters as described in Remark 6.11.

- Algorithm 2 using **SLS** with parameters as described in Remark 6.11.

6.3.3. Runtime comparison

Implementing the six different lower bounds from Section 6.3.2 into our B&B-framework, we measured the runtime of the B&B routine without preprocessing time, which is evaluated separately at the end of this section. On every instance each of the lower bounds had a maximum of 5 minutes to complete B&B; if this time constraint was violated, the process was interrupted and the lower bound considered as unsuccessful on this instance. If the parameter h could not be found by SOSTOOLS' `findbound.m` function, GLOB and SLS were considered to have violated the time constraint. The results for family (F1) from Section 3.3.2 are plotted in Figure 6.2 and for family (F2) from Section 3.3.3 in Figure 6.3.

We infer from the plots that for a small number of variables, the problem size (i.e., R), is still so small that brute force is competitive with our approach. However, if instances get larger, GLOB and SLS are, on average, faster than brute force and the continuous relaxation. SEP is quite fast in small instances, but for large instances the running time is high as a new underestimator is computed at each node. In our setting, POLY takes too long to be competitive. The continuous relaxation is satisfactory for smaller instances but slower than brute force. For $(n, d) = (5, 6)$, no lower bound except for SLS finished B&B within the time limit of 5 minutes.

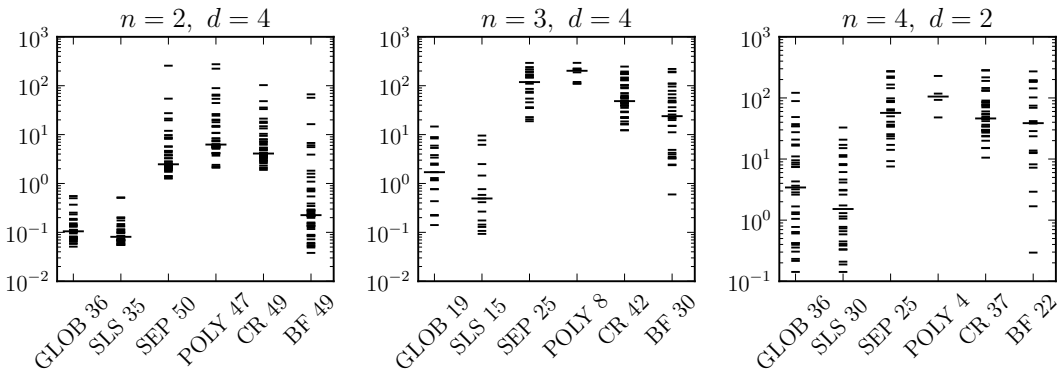


Figure 6.2.: From family (F1), we sample instances until 50 with positive definite leading form are generated. Using these as input to our B&B framework together with the lower bounds from Section 6.3.2, we scatter plot the runtimes in [s] for different dimension n and degree d (logarithmic scale). The larger line is the median of the depicted values. The number following the bound is the amount of instances that are solved within 5 minutes.

To evaluate the underestimators on larger instances, we additionally increased n, the number of variables, for fixed degree $d = 4$ until no instance was solved by any underestimator. For these large instances we allowed a larger time constraint of 2 hours. We improved solvability by alternatively allowing $h = 0$ if a continuous minimizer $\bar{x}$ could not be found, or if GLOB or SLS could not be computed with $h = \bar{x}$.

The largest n for which instances could still be solved for the different underestimators are the following:

GLOB	SLS	SEP	POLY	CR	BF
7	9	6	4	11	6

The results for the largest instances that SLS was still able to solve are plotted in Figure 6.4. The plot and table show that brute force cannot solve larger instances. Only the continuous relaxation can compete with SLS, it solves more instances within the given time limit. However, as Figure 6.4 shows, SLS is significantly faster than CR for $n = 7,\ 8$ and for 9 in most of the instances that SLS solved within the time limit.

Figure 6.5 illustrates the differences in preprocessing time, i.e., the time needed to compute an approximate continuous minimizer h and the underestimator g, and the time needed for the actual branch and bound. It can be seen that the preprocessing time does not vary too much with the instance and is mostly significantly longer than the subsequential branch and bound.

Also, SLS takes longer than GLOB in the preprocessing phase. This is to be expected as the corresponding sos program is larger, and so are preprocessing times. Indeed, Figure 6.5 reveals that GLOB has shorter preprocessing times throughout, but SLS is superior in B&B, as expected.

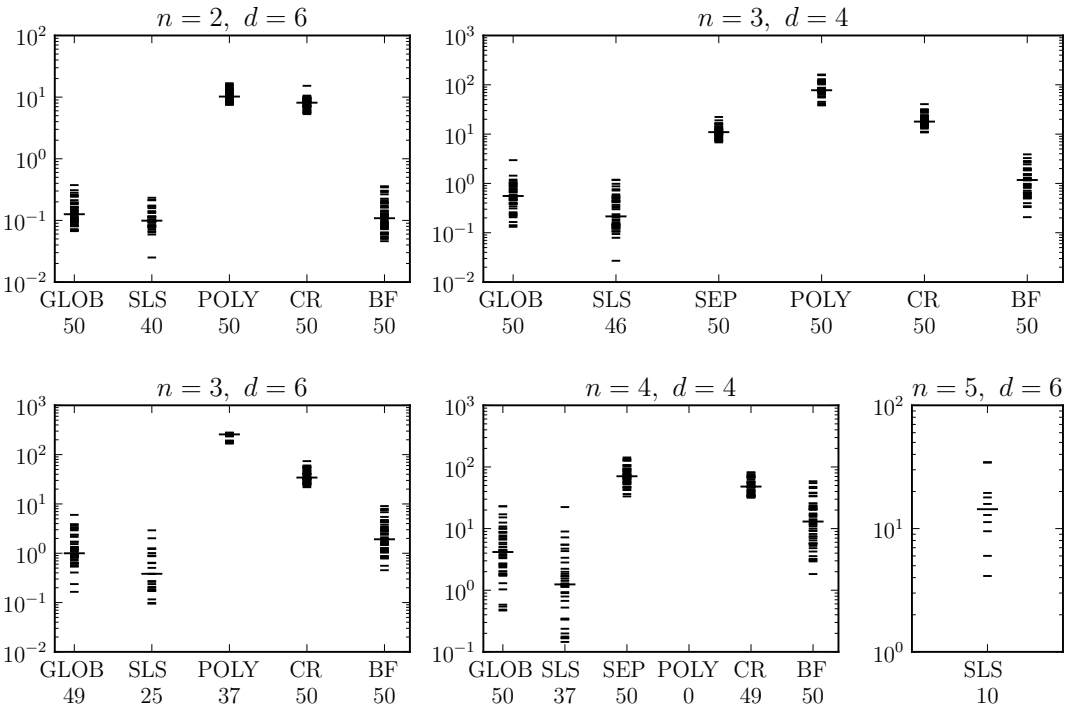


Figure 6.3.: From family (F2), we sample 50 instances. Using these as input to our B&B framework together with the lower bounds from Section 6.3.2, we scatter plot the runtimes in [s] for different dimension n and degree d (logarithmic scale). The larger line is the median of the depicted values. The number below the bound is the amount of instances that are solved within 5 minutes.

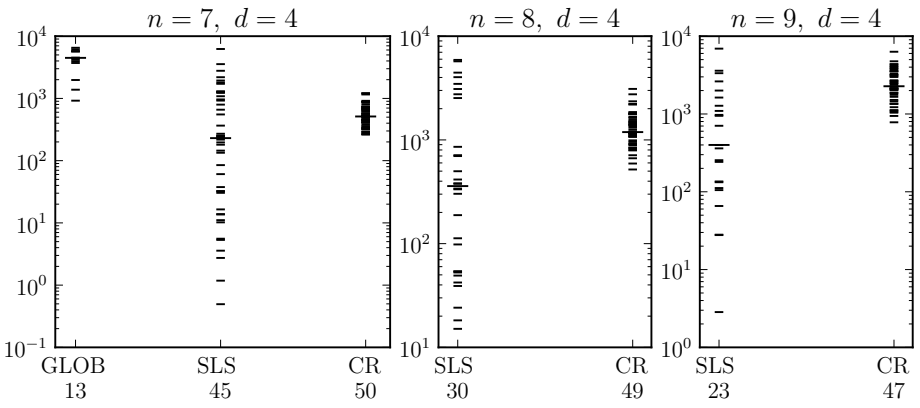


Figure 6.4.: Large instances: From family (F2), we sample 50 instances. Using these as input to our B&B framework together with the lower bounds from Section 6.3.2, we scatter plot the runtimes in [s] for different dimension n and degree d (logarithmic scale). The larger line is the median of the depicted values. The number below the bound is the amount of instances that are solved within 2 hours.

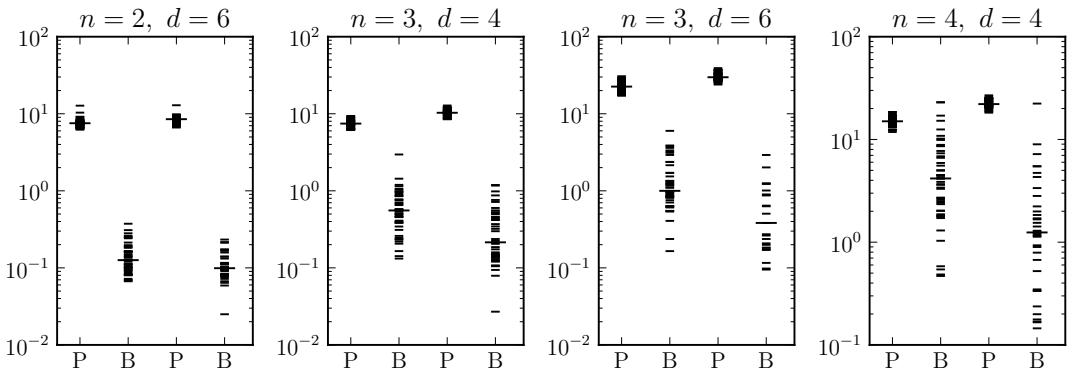


Figure 6.5.: From family (F2), 50 instances are sampled for different dimension n and degree d. The scatter plot shows the Preprocessing (P) and B&B (B) times – GLOB on the left, SLS on the right (logarithmic scale). The larger line is the median of the depicted values.

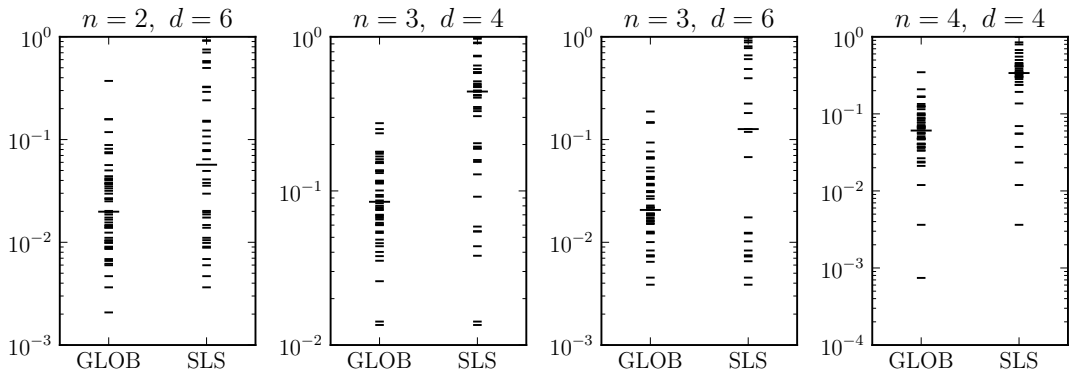


Figure 6.6.: From family (F2), 50 instances are sampled for different dimension n and degree d. We compare the initial lower bound on the integer minimizer using the ratio Q (logarithmic scale).

The initial lower bound on the minimum

At last, we evaluate the initial lower bound $g(\lfloor h \rceil)$ of our underestimators g. To this end, we define a ratio Q as follows: Let h be a continuous minimizer of f (if found by sos methods), x^* an integer minimizer of f found during B&B and g be a solution to GLOB or SLS. Then

$$Q := \frac{g(\lfloor h \rceil) - f(h)}{f(x^*) - f(h)}$$

takes values in $[0, 1]$, is invariant under scaling of f by constants $\lambda > 0$ and addition of constants $c \in \mathbb{R}$ to f – and, needless to say, the larger Q, the tighter the lower bound. See Figure 6.6 for the results.

By Theorem 6.6, SLS gives bounds that are at least as good as GLOB. The plots show that, once the semidefinite program is solved, SLS gives strictly tighter bounds more often than not.

7. Growth and stability properties of coercive polynomials

This chapter is about coercive polynomials, which appeared throughout this work. To characterize coercive polynomials, we investigate how fast a coercive polynomial grows, using the so-called order of coercivity, and how stable the coercivity property is under perturbations of its coefficients.

Section 7.1 motivates the problem and concepts of this chapter, and gives some pointers to the literature.

Section 7.2 recalls, for a coercive multivariate polynomial $f \in \mathbb{R}[X_1, \ldots, X_n]$, the notion and also some of the properties of the so-called order of coercivity $o(f)$ and links them to the so-called Łojasiewicz exponent at infinity. For coercive polynomials f we further recall the definition of the degree of stable coercivity $s(f)$ and introduce the degree of strongly stable coercivity $\tilde{s}(f)$. Also, by studying the order of coercivity of rational functions, we give an alternative proof of the known result that $o(f)$ is always positive for any coercive polynomial f.

Section 7.3 describes, for a coercive polynomial f, how the order of coercivity $o(f)$, the degree of stable and strongly stable coercivity, $s(f)$ and $\tilde{s}(f)$, respectively, are related. One of the two main results, Theorem 7.15, gives an explicit relation between these three numbers. The other main result, Theorem 7.16, shows that coercive polynomials f whose order of coercivity $o(f)$ is maximum possible, that is, $o(f) = \deg(f)$, are exactly the polynomials with a positive definite leading form. We also show that this is equivalent to the fact that their degree of stable coercivity $s(f)$ is maximum possible, that is, $s(f) = \deg(f)$.

Section 7.4 explicitly constructs two families of coercive polynomials with the corresponding order of coercivity being positive but tending to zero. For the first family the number of variables is hold fixed but the degree varies, and, for the second family, the degree is fixed but the number of variables varies.

Section 7.5 addresses the question whether it is possible to determine the minimal possible order of coercivity for a polynomial in n variables of degree not exceeding d.

Section 7.6 is about the decision problem whether a given polynomial is coercive.

7.1. Introduction

7.1.1. Motivation

In almost every of the previous chapters, coercivity played an important role: To guarantee existence of optimal solutions to MINLP given feasibility (Proposition 1.32), as a sufficient condition for the Archimedean property (Section 1.5.3), as an assumption to the norm bound from the literature and our norm bound as outlined in Chapter 3, to guarantee existence of underestimators (Theorem 6.6) and so forth.

To understand multivariate polynomials better, we consider the order of growth at infinity and how this relates to the stability of coercivity with respect to perturbations of the coefficients.

As a motivation, let us first consider the univariate case. A polynomial $f \in \mathbb{R}[X]$ is called coercive on $\mathbb{R}$ if $f(x) \to +\infty$ whenever $|x| \to +\infty$. This is the case if and only if the leading coefficient of f is positive and the degree $\deg(f)$ of f is positive and even. This, in turn, is equivalent to the property $f(x)/|x|^q$ being coercive for all $q \in [0, \deg(f))$. Hence, the number $\deg(f)$ expresses how fast f grows for large x, and, thus, it can be viewed as a meaningful measure for the order of growth of f at infinity. We call this number the *order of coercivity* of f.

We observe further that, in the univariate case, small perturbations of a coercive polynomial f by another univariate polynomial g preserves coercivity. In fact, if f is coercive, so is $f + g$ whenever $\deg(g) \leq \deg(f)$ and if the leading coefficient of g is sufficiently close to zero. On the other hand, $f + g$ is not necessarily coercive if the degree of g exceeds the degree of f, and, thus, the number $\deg(f)$ can also be viewed as a measure expressing how *stable* the coercivity of f on $\mathbb{R}$ is. We call this number the *degree of stable coercivity* of f.

Consequently, for a univariate coercive polynomial f, the order of coercivity coincides with the degree of stable coercivity, and both are equal to the degree of f. Once these two numbers are properly defined in the multivariate setting, it is only natural to ask if the order of coercivity again equals the degree of stable coercivity, and if so, whether these numbers again coincide with the degree of f.

In [BS15b] the first question is answered affirmatively for a broad class of coercive polynomials whereas the authors give a dissenting answer to the second question. More precisely, using properties of the underlying Newton polytopes, a class of coercive polynomials f is identified for which the order of coercivity coincides with the degree of stable coercivity, and both are equal to a so-called *degree of convenience* of f which, in general, differs from $\deg(f)$.

In the present article we shall show that for coercive polynomials f the degree of stable coercivity of f may differ from the order of coercivity of f in general, but not "too much". More precisely, our main results show that for any coercive polynomial, its degree of stable coercivity is always equal to the integral part of the order of coercivity. We shall further characterize the case when the order of coercivity of f is maximum possible by positivity of its leading form. The latter turns out to be equivalent to the degree of stable coercivity of f also being maximum possible (see Theorems 7.15

and 7.16).

7.1.2. Related literature

Coercivity of multivariate polynomials itself is an interesting property for various reasons. In continuous polynomial optimization theory it is a recurring question whether a given polynomial $f \in \mathbb{R}[X_1, \ldots, X_n]$ attains its infimum over $\mathbb{R}^n$ (see, e.g. [BS15a; ED08; GSED14; GSED11; NDS06; Sch06; VP07; VP10]); a similar question is equally relevant in our integer and mixed-integer programming variant (Proposition 1.32). Coercivity of f is a sufficient condition for f having this property, and, thus, it is a natural task to verify or disprove whether f is coercive.

As a further consequence of coercivity, f is bounded below on $\mathbb{R}^n$ by some $v \in \mathbb{R}$, so that $f - v$ is positive semidefinite on $\mathbb{R}^n$. Also, since coercivity of f is equivalent to the boundedness of its lower level sets $\{x \in \mathbb{R}^n : f(x) \leq \alpha\}$ for all $\alpha \in \mathbb{R}$, understanding coercivity can be useful to decide whether a basic semi-algebraic set is bounded. Furthermore, properness of polynomial maps $F : \mathbb{R}^n \to \mathbb{R}^n$ can be characterized by coercivity of the polynomial $\|F\|_2^2$. This is useful to decide whether F is globally invertible (see, e.g. [BS17; BA07; CDTT14]).

Coercivity of polynomials is partially analyzed in [JLL14] and, in the convex setting, in [JPL14], while the coercivity of a polynomial f defined on a basic closed semi-algebraic set and its relation to the Fedoryuk and Malgrange conditions are examined in [VP10]. A connection between coercivity of multivariate polynomials and their Newton polytopes is given in [BS15a]. In [MN14], the authors study how fast – not necessarily coercive – polynomials grow on semi-algebraic sets.

7.2. Order and stability of coercivity

7.2.1. The order of coercivity

We saw in Section 1.4.8 that a function $f : S \to \mathbb{R}$, defined on a subset $S \subset \mathbb{R}^n$, is coercive if

$$\forall c > 0 \quad \exists M \geq 0 \quad \forall x \in S,\ \|x\| \geq M \ :\ f(x) \geq c \tag{7.1}$$

holds. The function f is called q-coercive for some $q \geq 0$ if $f(x)/\|x\|^q$ is coercive. Note that coercivity and q-coercivity are properties that are independent of the choice of the norm on $\mathbb{R}^n$. The following characterization of q-coercivity, $q \geq 0$, turns out to be useful for our later purposes. For completeness, we give its short proof.

Observation 7.1. *Let $f : S \to \mathbb{R}$ defined on a subset $S \subset \mathbb{R}^n$ and $q \geq 0$ be given. Then f is q-coercive if and only if*

$$\forall c > 0 \quad \exists M \geq 0 \quad \forall x \in S,\ \|x\| \geq M \ :\ f(x) \geq c \cdot \|x\|^q \tag{A}$$

holds.

Proof. Let f be q-coercive and $c > 0$. By definition of q-coercivity, there is $M \geq 0$ such that $f(x)/\|x\|^q \geq c$ whenever $x \in S$ and $\|x\| \geq M$. Multiplication by $\|x\|^q$ gives the claim. Now suppose (A) holds and fix $c > 0$ with the corresponding $M \geq 0$. Division by $\|x\|^q$ yields for all nonzero $x \in S$ with $\|x\| \geq M$ the inequality

$$\frac{f(x)}{\|x\|^q} \geq c,$$

and, thus, $\liminf_{\|x\|\to\infty} f(x)/\|x\|^q \geq c$. Since $c > 0$ was arbitrary, f is q-coercive. □

For coercive $f : S \to \mathbb{R}$, the number

$$o(f) := \sup\{q \geq 0 : f \text{ is } q\text{-coercive}\}$$

is called the *order of coercivity* of f. A coercive function f is q'-coercive for all q' with $0 \leq q' < o(f)$, but f need not be $o(f)$-coercive. Now, if property (A) does not hold for all but only some $c > 0$, we may not conclude q-coercivity of f. However, the following holds:

Observation 7.2. *Let $f : S \to \mathbb{R}$ defined on a subset $S \subset \mathbb{R}^n$ and $q > 0$ be given. Then the property*

$$\exists c > 0 \quad \exists M \geq 0 \quad \forall x \in S,\ \|x\| \geq M \ :\ f(x) \geq c \cdot \|x\|^q \tag{B}$$

implies $o(f) \geq q$.

Proof. For any $0 < \varepsilon < q$ and all nonzero $x \in S$ with $\|x\| \geq M$ one has

$$\frac{f(x)}{\|x\|^{q-\varepsilon}} \geq c\|x\|^{\varepsilon},$$

and, as the right hand side is coercive, f is $(q-\varepsilon)-$coercive. Thus the inequality $o(f) \geq q - \varepsilon$ holds and since ε was arbitrary, $o(f) \geq q$ follows. □

Note that the converse statement does not necessarily hold.

The following example shows that for quadratic coercive polynomials, the equality $o(f) = \deg(f)$ is always fulfilled. For the proof and the proofs to come, the following immediate estimate is handy; for completeness, we give a proof.

Observation 7.3. *For $f \in \mathbb{R}[X_1, \ldots, X_n]_d$, where $n \in \mathbb{N}$, $d \in \mathbb{N}_0$, and any $q \in [d, +\infty)$, the following estimate holds:*

$$|f(x)| \leq \binom{n+d}{d} \cdot \|f\|_\infty \cdot (\|x\|_\infty^q + 1), \quad x \in \mathbb{R}^n.$$

Proof. Fix $n \in \mathbb{N}$, $d \in \mathbb{N}_0$, $f \in \mathbb{R}[X_1, \ldots, X_n]$ of degree at most d and $q \geq d$. In multi-index notation, $f = \sum_{\alpha \in A(f)} a_\alpha X^\alpha$, and in view of (1.1), we get $|A(f)| \leq \binom{n+d}{d}$. Also, for all $x \in \mathbb{R}^n$ and $\alpha \in \mathbb{N}_0^d$ with $|\alpha| = \alpha_1 + \ldots + \alpha_n \leq q$, we have

$$|x^\alpha| \leq \|x\|_\infty^{|\alpha|} \leq \max(\|x\|_\infty^q, 1) \leq \|x\|_\infty^q + 1.$$

The estimates combine to

$$\begin{aligned} |f(x)| &= |\sum_{\alpha \in A(f)} a_\alpha x^\alpha| \leq \|f\|_\infty \sum_{\alpha \in A(f)} |x^\alpha| \leq \|f\|_\infty \sum_{\alpha \in A(f)} (\|x\|_\infty^q + 1) \\ &= |A(f)| \cdot \|f\|_\infty \cdot (\|x\|_\infty^q + 1) \leq \binom{n+d}{d} \cdot \|f\|_\infty \cdot (\|x\|_\infty^q + 1). \end{aligned}$$

□

Example 7.4. Let $f \in \mathbb{R}[X_1, \ldots, X_n]$, $f(x) = x^T Q x + L^t x + c$ with $Q \in \mathbb{R}^{n \times n}$ symmetric, $L \in \mathbb{R}^n$ and $c \in \mathbb{R}$ be given. If f is coercive, then $o(f) = 2$. Indeed, as f is coercive, Q must be positive definite. It is well-known that this implies the existence of a unique global minimal point $x_0 \in \mathbb{R}^n$ of f, and one finds $f(x) = (x - x_0)^T Q (x - x_0) + f(x_0)$ (see, e.g. [BHS15]). Denoting the smallest eigenvalue of Q by λ, one obtains that $f(x) \geq \lambda (x - x_0)^T (x - x_0) + f(x_0) = \lambda \|x - x_0\|_2^2 + f(x_0)$ holds for all $x \in \mathbb{R}^n$, and thus, by Observation 7.2, the inequality $o(f) \geq 2$ follows. On the other hand, Observation 7.3 implies $o(f) \leq \deg(f)$, and, due to $\deg(f) = 2$, one obtains $o(f) \leq 2$.

Property (B) shows how the order of coercivity is related to the so-called *Łojasiewicz exponent at infinity* (see, e.g. [Kra07]). For a polynomial map $F : \mathbb{R}^n \to \mathbb{R}^m$ it is defined as

$$\mathcal{L}_\infty(F) := \sup\{\nu \in \mathbb{R} : \exists c, M > 0 \; \forall x \in \mathbb{R}^n : \|x\| \geq M \Rightarrow \|F(x)\| \geq c\|x\|^\nu\}.$$

Indeed, for coercive polynomials, the order of coercivity and Łojasiewicz exponent at infinity coincide:

Observation 7.5. *Let $f \in \mathbb{R}[X_1, \ldots, X_n]$ be coercive. Then*

$$o(f) = \mathcal{L}_\infty(f).$$

Proof. From the definitions, $o(f) \leq \mathcal{L}_\infty(f)$. Note that the coercivity of f implies $\mathcal{L}_\infty(f) \geq 0$ and $o(f) \geq 0$. Suppose at first that $\mathcal{L}_\infty(f) = 0$, then $o(f) = 0$ and hence $o(f) = \mathcal{L}_\infty(f) = 0$ follows. Suppose next that $\mathcal{L}_\infty(f) > 0$. It is enough to show that for any $0 \leq q < \mathcal{L}_\infty(f)$, we also have $q < o(f)$. Let $\varepsilon > 0$ with $q + \varepsilon \leq \mathcal{L}_\infty(f)$. By definition of $\mathcal{L}_\infty(f)$, there is $c > 0$ and $M \geq 0$ with $f(x) \geq c \cdot \|x\|^{q+\varepsilon} = (c \cdot \|x\|^\varepsilon) \cdot \|x\|^q$ whenever $\|x\| \geq M$. As $c \cdot \|x\|^\varepsilon$ grows without bound, this yields $o(f) > q$. □

Since the Łojasiewicz exponent $\mathcal{L}_\infty(f)$ is known to be rational (see [Gor61]), Observation 7.5 yields the following:

Corollary 7.6. *If $f \in \mathbb{R}[X_1, \ldots, X_n]$ is coercive, then $o(f) \in \mathbb{Q}$.*

7.2.2. The stability of coercivity

Given a coercive polynomial $f \in \mathbb{R}[X_1, \ldots, X_n]$ we are interested in how stable this coercivity property is under small perturbations of f by other polynomials. This gives rise to the following definition for stability of coercivity which was already analyzed from the viewpoint of the underlying Newton polytopes in [BS15b] and is inspired by the concept of stable boundedness of polynomials [Mar03].

Definition 7.7 (Stable coercivity)**.** A polynomial $f \in \mathbb{R}[X_1, \ldots, X_n]$ is called q-stably coercive for $q \in \mathbb{N}_0$, if there exists an $\varepsilon > 0$ such that for all $g \in \mathbb{R}[X_1, \ldots, X_n]$ with $\deg g \leq q$ and all coefficients of g bounded in absolute value by ε it holds that $f + g$ is coercive. The degree of stable coercivity $s(f)$ of f is the largest q such that f is q-stable coercive.

We also introduce the following stronger notion for the stability of coercivity.

Definition 7.8 (Strong stable coercivity)**.** A polynomial $f \in \mathbb{R}[X_1, \ldots, X_n]$ is called strongly q-stable coercive for $q \in \mathbb{N}_0$, if for all $g \in \mathbb{R}[x]$ with $\deg g \leq q$ it holds that $f + g$ is coercive. The degree of strongly stable coercivity $\tilde{s}(f)$ of f is the largest q such that f is strongly q-stable coercive.

7.2.3. Observations on the order of coercivity

In this section we collect some preliminary results on the order of coercivity. The following result is not only useful for our purposes but interesting in its own right: It states that any coercive rational function has a positive order of growth. To this end we denote the vanishing set of the polynomial $g \in \mathbb{R}[X_1, \ldots, X_n]$ by $V(g) := \{x \in \mathbb{R}^n : g(x) = 0\}$ and its complement by $V^c(g) := \mathbb{R}^n \setminus V(g)$.

Theorem 7.9. *Let $f, g \in \mathbb{R}[X_1, \ldots, X_n]$, $g \neq 0$, such that $f/g : V^c(g) \to \mathbb{R}$ is coercive. Then*

$$o\,(f/g) > 0.$$

As a corollary, every coercive polynomial $f \in \mathbb{R}[X_1, \ldots, X_n]$ has a positive order of growth, which is a known result (see, e.g. [Gor61]). For the proof of Theorem 7.9, we use the following.

Theorem 7.10 (see Theorem 4.1 in [Gor61])**.** *Let $P(x, z, w)$ be a real polynomial of $n' = n_1 + n_2 + n_3$ variables $x \in \mathbb{R}^{n_1}$, $z \in \mathbb{R}^{n_2}$, $w \in \mathbb{R}^{n_3}$ where n_1, n_2, n_3 are non-negative integers. If the surface M given by the equation*

$$P(x, z, w) = 0$$

is not empty and lies in the domain defined by the inequality

$$\|z\|_2 \geq \varphi(\|x\|_2),$$

where $\varphi(t) \to +\infty$ as $t \to +\infty$, then there exists constants $h > 0$ and b such that this surface also lies in the domain defined by the inequality

$$\|z\|_2 \geq \|x\|_2^h - b.$$

Our choice of a suitable φ is given in the next lemma.

Lemma 7.11. *In the setting of Theorem 7.9, let $\varphi : [0, \infty) \to \mathbb{R}$ be defined as follows: Let*

$$\tilde{\varphi}(t) := \inf \left\{ \left| \frac{f(y)}{g(y)} \right| : y \in V^c(g),\ \|y\|_2 = t \right\}$$

and put

$$\varphi(t) = \begin{cases} \tilde{\varphi}(t), & \tilde{\varphi}(t) < \infty, \\ 0, & \textit{else.} \end{cases}$$

Then φ is coercive.

Proof. Note at first that there can at most be $d := \deg(g)$ many $t \geq 0$ with the property that $g(x) = 0$ whenever $\|x\| = t$ (resulting in $\tilde{\varphi}(t) = \infty$). Indeed, suppose there were $d+1$ points $t_0 < \ldots < t_d$ with that property. Consider the leading form g_d of g and pick $x \in \mathbb{R}^n$ with $g_d(x) \neq 0$. Then $\lambda \mapsto g_d\left(\lambda \cdot \frac{x}{\|x\|}\right)$ is a univariate polynomial of degree d with zeros at $t_0, \ldots, t_d$, which is impossible. Now suppose φ is not coercive. Thus there is $C > 0$ and an increasing sequence $\{\tau_k\}_{k \in \mathbb{N}}$ of reals with $\tau_k \to +\infty$ and $\varphi(\tau_k) \leq C$. We may assume τ_1 is larger than any of the at most d points t_i from above. Fix $\varepsilon > 0$. Thus there is a sequence $\{x_k\}_{k \in \mathbb{N}} \subset V^c(g)$ with $\|x_k\| = \tau_k$ and $\left|\frac{f(x_k)}{g(x_k)}\right| - \varepsilon \leq \varphi(\tau_k) \leq C$, contradicting coercivity of f/g. □

We may now prove Theorem 7.9.

Proof of Theorem 7.9. Let $f, g \in \mathbb{R}[X_1, \ldots, X_n]$ with f/g coercive. To apply the theorem by Gorin, we let $n_1 = n$, $n_2 = n_3 = 1$ and define $P \in \mathbb{R}[X_1, \ldots, X_n, Z, W]$ via

$$P(x, z, w) := (f(x) - zg(x))^2 + (wg(x) - 1)^2, \quad x \in \mathbb{R}^n,\ z, w \in \mathbb{R}.$$

The surface $M = V(P)$ is not empty, as

$$\begin{aligned} M &= \left\{(x, z, w) \in \mathbb{R}^{n+2} : f(x) = zg(x) \text{ and } wg(x) = 1\right\} \\ &= \left\{(x, z, w) \in \mathbb{R}^{n+2} : x \in V^c(g),\ f(x)/g(x) = z \text{ and } w = 1/g(x)\right\}. \end{aligned}$$

Consider the function φ from Lemma 7.11. We now show that M lies in the domain defined by the inequality $\|z\|_2 \geq \varphi(\|x\|_2)$. To this end let a point $(x, z, w) \in M$ be given. Then $g(x) \neq 0$ and so we conclude

$$\|z\|_2 = |z| = \left|\frac{f(x)}{g(x)}\right| \geq \inf\left\{\left|\frac{f(y)}{g(y)}\right| : y \in V^c(g),\ \|y\|_2 = \|x\|_2\right\} = \varphi(\|x\|_2).$$

Hence, we may apply Gorin's theorem, so there are constants $h > 0$ and b such that M also lies in the domain defined by the inequality

$$|z| = \|z\|_2 \geq \|x\|_2^h - b.$$

This means $|f(x)/g(x)| \geq \|x\|_2^h - b$ whenever $g(x) \neq 0$. From Observation 7.2 we conclude $o(|f/g|) \geq h$. Since f/g is coercive, $f(x)/g(x) > 0$ for $x \in V^c(g)$ with $\|x\|_2$ large enough, which implies $o(f/g) \geq h$, too. □

We note that for a q-coercive polynomial f, the number q is strictly bound above by the order of growth of f. This is implicit in [Gor61]; we give an explicit proof in the setting of this article for completeness.

Lemma 7.12. *Let $f \in \mathbb{R}[X_1, \ldots, X_n]$ be coercive. Then f is not $o(f)$-coercive.*

Proof. Suppose the contrary and let f be $o(f)$-coercive. By Corollary 7.6 and Theorem 7.9, the number $o(f)$ is rational and positive, so $o(f) = p/q$, with some $p, q \in \mathbb{N}$ and we may further assume that p is even. Thus by definition, $f(x)/\|x\|_2^{p/q}$ is coercive, and hence $r(x) := f(x)^q/\|x\|_2^p$ is coercive. However, as p is even, r is a coercive rational function, so by Theorem 7.9, there is $h > 0$ such that r is h-coercive. Hence by Observation 7.1 and continuity of f, there are $c_1 > 0$, $c_2 \geq 0$ with

$$\frac{f(x)^q}{\|x\|_2^p} \geq c_1\|x\|_2^h - c_2.$$

Hence, for any fixed $0 < \varepsilon < h$,

$$\frac{f(x)^q}{\|x\|_2^{p+\varepsilon}} \geq c_1\|x\|_2^{h-\varepsilon} - \frac{c_2}{\|x\|_2^{\varepsilon}},$$

which means f^q is $(p + \varepsilon)$-coercive. As f, being coercive, attains positive values for large x, this implies that f is $((p + \varepsilon)/q)$-coercive, and we conclude $o(f) \geq (p + \varepsilon)/q$, contradicting the assumption $o(f) = p/q$. □

Although, by Lemma 7.12, a coercive $f \in \mathbb{R}[X_1, \ldots, X_n]$ is not $o(f)$-coercive, one can still underestimate f by an $o(f)$-power of a norm for large values of $\|x\|$. That is, for coercive polynomials, we have a converse statement to Observation 7.2. Several variants of this result are known; one may argue by Tarski-Seidenberg [Gor61] or, in the complex setting, by curve selection at infinity [Kra07]. Our contribution is a proof by elementary methods.

Lemma 7.13. *Let $f \in \mathbb{R}[X_1, \ldots, X_n]$ be coercive. Then there exist $c > 0$, $M \geq 0$ with*

$$f(x) \geq c \cdot \|x\|^{o(f)}, \quad \|x\| \geq M.$$

Proof. Assume to the contrary that the assertion does not hold. Then for every sequence $\{c_k\}_{k\in\mathbb{N}} \subset \mathbb{R}$ with $c_k \downarrow 0$ there exists a sequence $\{x_k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ with $\|x_k\| \to +\infty$ such that

$$f(x_k) < c_k \|x_k\|^{o(f)}, \quad k \in \mathbb{N}.$$

Since f is coercive and $\|x_k\| \to +\infty$, we may further assume $f(x_k) \geq 0$ for all $k \in \mathbb{N}$, hence

$$0 \leq \frac{f(x_k)}{\|x_k\|^{o(f)}} < c_k, \quad k \in \mathbb{N}.$$

Using the decomposition $f = \sum_{i=0}^{d} f_i$ of f into its homogeneous components $f_i \in \mathbb{R}[X_1, \ldots, X_n]$ of degree $i = 0, \ldots, d$, with $\xi_k := x_k/\|x_k\|$ the latter property yields

$$0 \leq \sum_{i=\lceil o(f)\rceil}^{d} \|x_k\|^{i-o(f)} f_i(\xi_k) < c_k - \sum_{i=0}^{\lceil o(f)\rceil - 1} \|x_k\|^{i-o(f)} f_i(\xi_k), \quad k \in \mathbb{N}. \tag{7.2}$$

Due to $c_k \downarrow 0$ and $i - o(f) < 0$ holding for all $i = 0, \ldots, \lceil o(f)\rceil - 1$, the right hand side in (7.2) converges to zero as k approaches infinity. This implies

$$\lim_{k\to\infty} \sum_{i=\lceil o(f)\rceil}^{d} \|x_k\|^{i-o(f)} f_i(\xi_k) = 0 \tag{7.3}$$

Passing to an appropriate convergent subsequence of the sequence $\xi_k = x_k/\|x_k\|$ with a limit point ξ, due to continuity of each homogeneous component f_i of f, we may assume that $\lim_{k\to\infty} f_i(\xi_k) = f_i(\xi) \in \mathbb{R}$ for all $i = \lceil o(f)\rceil, \ldots, d$. In fact, property (7.3) yields $f_i(\xi) = 0$ for all $i = \lceil o(f)\rceil, \ldots, d$, and, hence again, using the homogeneous decomposition of f one obtains

$$f(t \cdot \xi) = \sum_{i=0}^{d} f_i(t \cdot \xi) = t^{\lceil o(f)\rceil - 1} f_{\lceil o(f)\rceil - 1}(\xi) + \cdots + f_0(\xi) \quad \text{for all } t \in \mathbb{R}$$

resulting in $o(f) \leq \lceil o(f)\rceil - 1$, a contradiction. □

7.3. Main result

In this section we show how the degree of stable and strongly stable coercivity are tied to the order of growth (Theorem 7.15). In case of a positive definite leading form, a stronger characterization is available (Theorem 7.16). We use the following estimate in the proof of both.

Proposition 7.14. *Let $f \in \mathbb{R}[X_1, \ldots, X_n]$ be coercive. Then the following inequalities are fulfilled:*

$$\tilde{s}(f) \leq s(f) \leq o(f) \leq \tilde{s}(f) + 1.$$

Proof. The first inequality $\tilde{s}(f) \leq s(f)$ follows obviously from the Definitions 7.7 and 7.8. To see $s(f) \leq o(f)$, assume $q := s(f) > o(f)$. We introduce polynomials

$$f_{c,\sigma} := f - c \cdot \left(\sum_{j=1}^{n} \sigma_j X_j \right)^q,$$

parameterized by $c \in \mathbb{R}$ and $\sigma \in \Sigma := \{-1, 1\}^n$. As $s(f) = q$, for every $\sigma \in \Sigma$ there is $\varepsilon_\sigma > 0$ such that $f_{c,\sigma}$ is coercive whenever $c \in [-\varepsilon_\sigma, \varepsilon_\sigma]$. Let $\hat{\varepsilon} := \min_{\sigma \in \Sigma} \varepsilon_\sigma$ and fix $\hat{c} \in (0, \hat{\varepsilon})$. Hence $f_{\hat{c},\sigma}$ is coercive for all $\sigma \in \Sigma$ and thus also bounded from below. Boundedness from below means for every σ there is $k_\sigma \geq 0$ with

$$f(x) \geq \hat{c} \left(\textstyle\sum_{j=1}^{n} \sigma_j x_j \right)^q - k_\sigma, \quad x \in \mathbb{R}^n,\ \sigma \in \Sigma.$$

Put $\hat{k} := \max_{\sigma \in \Sigma} k_\sigma$. Then for $x \in \mathbb{R}^n$

$$f(x) \geq \hat{c} \cdot \max_{\sigma \in \Sigma} \left(\textstyle\sum_{j=1}^{n} \sigma_j x_j \right)^q - \hat{k} = \hat{c} \cdot \left(\textstyle\sum_{j=1}^{n} |x_j| \right)^q - \hat{k} = \hat{c} \cdot \|x\|_1^q - \hat{k},$$

so Observation 7.2 implies $o(f) \geq q = s(f)$, a contradiction.

Now we proceed to prove the third inequality $o(f) \leq \tilde{s}(f) + 1$. Assume the contrary: Let $q := \tilde{s}(f)$ and suppose $o(f) > q + 1$. We have arrived at a contradiction if we may show that for any $g \in \mathbb{R}[X_1, \ldots, X_n]$ of degree at most $q + 1$, $f + g$ is coercive, as in this case $\tilde{s}(f) \geq q + 1 = \tilde{s}(f) + 1$. To this end fix an arbitrary $g \in \mathbb{R}[X_1, \ldots, X_n]_{q+1}$. Now choose $c_1 > \binom{n+d}{d} \cdot \|g\|_\infty$. As $o(f) > q + 1$, f is $q + 1$-coercive, therefore, by Observation 7.1 and continuity of f, there is $c_2 \geq 0$ such that $f(x) \geq c_1 \|x\|_\infty^{q+1} - c_2$ holds for $x \in \mathbb{R}^n$, and hence, by Observation 7.3,

$$\begin{aligned} f(x) + g(x) &\geq f(x) - |g(x)| \geq c_1 \|x\|_\infty^{q+1} - c_2 - \binom{n+d}{d} \cdot \|g\|_\infty \left(\|x\|_\infty^{q+1} + 1 \right) \\ &= c_1' \cdot \|x\|_\infty^{q+1} - c_2', \quad x \in \mathbb{R}^n, \end{aligned}$$

for some appropriately chosen $c_1' > 0$, $c_2' \in \mathbb{R}$. Thus $f + g$ is coercive.

□

We show now how the integer part of the order of growth and our notions of stability are related to each other.

Theorem 7.15. *Let $f \in \mathbb{R}[X_1, \ldots, X_n]$ be coercive.*

1. *If $o(f)$ is integer, then*
$$\tilde{s}(f) + 1 = s(f) = o(f).$$
2. *If $o(f)$ is fractional, then*
$$\tilde{s}(f) = s(f) = \lfloor o(f) \rfloor.$$

Proof. In order to prove 1., we show $\tilde{s}(f) + 1 = o(f)$ first. By integrality of $\tilde{s}(f)$, $o(f)$ and by the property $o(f) \in [\tilde{s}(f), \tilde{s}(f) + 1]$ holding due to Proposition 7.14, it is enough to show that $\tilde{s}(f) < o(f)$. Suppose the contrary, that is $\tilde{s}(f) = o(f) =: q$. Now for $c > 0$ and $\sigma \in \Sigma := \{-1, 1\}^n$, define

$$f_{c,\sigma} := f - c \cdot \left(\sum_{j=1}^{n} \sigma_j X_j \right)^q \in \mathbb{R}[X_1, \ldots, X_n].$$

By definition of $\tilde{s}(f)$, the polynomial $f_{c,\sigma}$ is coercive and hence bounded from below for all $c > 0$ and $\sigma \in \Sigma$. That is, for every $c > 0$ and $\sigma \in \Sigma$, there exists $k_{c,\sigma} \geq 0$ such that

$$f(x) \geq c \cdot \left(\sum_{j=1}^{n} \sigma_j x_j \right)^q - k_{c,\sigma}, \quad x \in \mathbb{R}^n,\ c > 0,\ \sigma \in \Sigma,$$

and hence with $k_c := \max_{\sigma \in \Sigma} k_{c,\sigma}$, we have for all $x \in \mathbb{R}^n$ and $c > 0$ the property

$$f(x) \geq c \cdot \max_{\sigma \in \Sigma} \left(\sum_{j=1}^{n} \sigma_j x_j \right)^q - k_c = c \cdot \left(\sum_{j=1}^{n} |x_j| \right)^q - k_c = c \cdot \|x\|_1^q - k_c.$$

In view of Observation 7.1, the polynomial f is q-coercive. Since $q = \tilde{s}(f) = o(f)$ is holding by assumption, f is $o(f)$-coercive. This is impossible by Lemma 7.12, and we may conclude that $\tilde{s}(f) + 1 = o(f)$.

For the second equality $s(f) = o(f)$, put $q := o(f)$. By Lemma 7.13 and continuity of f, there are constants $c_1, c_2 > 0$ such that

$$f(x) \geq c_1 \|x\|_\infty^q - c_2 \text{ holds for all } x \in \mathbb{R}^n.$$

Define $\varepsilon := \frac{c_1}{2} \cdot \binom{n+q}{q}^{-1}$. Now for any $g \in \mathbb{R}[X_1, \ldots, X_n]_q$ with $\|g\|_\infty \leq \varepsilon$ and all $x \in \mathbb{R}^n$, we have from Observation 7.3

$$\begin{aligned} f(x) + g(x) &\geq f(x) - |g(x)| \\ &\geq c_1 \|x\|_\infty^q - c_2 - \varepsilon \cdot \tbinom{n+q}{q} \left(\|x\|_\infty^q + 1 \right) \\ &= \frac{c_1}{2} \|x\|_\infty^q - c_2 - \frac{c_1}{2}. \end{aligned}$$

To summarize, $f + g$ is coercive whenever $\deg g \leq q$ and $\|g\|_\infty \leq \varepsilon$, that is, f is q-stably coercive, or $s(f) = q = o(f)$.

Statement 2. follows at once from Proposition 7.14. □

Our next result shows that more characterizations are available for a maximal order of coercivity.

Theorem 7.16. *Let $f \in \mathbb{R}[X_1, \ldots, X_n]$ of degree $d \geq 2$. Then, the following assertions are equivalent:*

1. *$f_d(x) > 0$ for all $x \in \mathbb{R}^n$, $x \neq 0$.*
2. *There exists $\delta > 0$ such that $f_d(x) \geq \delta \|x\|^d$ for all $x \in \mathbb{R}^n$.*
3. *$o(f) = d$.*
4. *$o(f) > d - 2$.*
5. *$s(f) = d$.*
6. *$s(f) \geq d - 1$.*
7. *$\tilde{s}(f) = d - 1$.*
8. *$\tilde{s}(f) \geq d - 2$.*

Proof. "1 $\Rightarrow$ 2" For $x = 0$ the assertion is trivial. For nonzero $x \in \mathbb{R}^n$ one obtains

$$f_d(x) = \|x\|^d f_d\left(\frac{x}{\|x\|}\right) \geq \|x\|^d \inf_{y \in \mathbb{S}^{n-1}} f_d(y).$$

The infimum is positive by compactness of the sphere. Now for "2 $\Rightarrow$ 3", let $c_j = \inf_{y \in \mathbb{S}^{n-1}} f_j(y)$ for $j = 0, \ldots, n-1$ and put $c_d = \delta$. Then by homogeneity of the f_j,

$$f(x) = \sum_{j=0}^{d} f_j(x) \geq \sum_{j=0}^{d} c_j \|x\|^j,$$

hence $o(f) = d$. The implication "3 $\Rightarrow$ 4" is trivial. The implication "4 $\Rightarrow$ 1" holds as follows: Suppose $o(f) > d-2$ but $f_d(\tilde{x}) = 0$ for some $\tilde{x} \in \mathbb{R}^n$ with $\tilde{x} \neq 0$. By assumption $o(f)$ is positive, hence f is coercive. Let us show that this implies $f_{d-1}(\tilde{x}) = 0$. Indeed, we find that for all $\lambda \in \mathbb{R}$ it holds

$$f(\lambda\tilde{x}) = \sum_{j=0}^{d} f_j(\lambda\tilde{x}) = \sum_{j=0}^{d-1} \lambda^j f_j(\tilde{x}),$$

which, as a function of λ is unbounded from below unless $f_{d-1}(\tilde{x}) = 0$. In fact, this holds since as $d - 1$ is odd. Hence

$$|f(\lambda\tilde{x})| \leq \sum_{j=0}^{d-2} |f_j(\lambda\tilde{x})| = \sum_{j=0}^{d-2} |\lambda|^j |f_j(\tilde{x})|,$$

implying $o(f) \leq d - 2$, a contradiction, so 1 through 4 are equivalent.

To see "2 $\Rightarrow$ 5", let $g \in \mathbb{R}[X_1, \ldots, X_n]$ of degree d, and let $c' = \max_{x \in \mathbb{S}^{n-1}} g_d(x)$. Then $|g_d(x)| \leq c'\|x\|^d$ by homogeneity, so for $\varepsilon \in [-\frac{\delta}{2c'}, \frac{\delta}{2c'}]$,

$$f_d(x) + \varepsilon g_d(x) \geq f_d(x) - |\varepsilon g_d(x)| \geq \delta\|x\|^d - \frac{\delta}{2}\|x\|^d = \frac{\delta}{2}\|x\|^d,$$

hence $f + \varepsilon g$ is still coercive, and we conclude $s(f) = d$.

We show that 5. implies 6 and 7. The first implication is trivial. To see "5 $\Rightarrow$ 7", note that Proposition 7.14 implies $\tilde{s}(f) \geq d-1$. As $\tilde{s}(f) \geq d$ is not possible for a degree d polynomial, $\tilde{s}(f) = d - 1$. Since both 6 and 7 imply 8 trivially, all equivalences are shown once "8 $\Rightarrow$ 4" holds.

So suppose $\tilde{s}(f) \geq d - 2$. From the definition of strong stable coercivity, this implies coercivity of f, and d must be even. The function $g(x) = \|x\|_2^{d-2}$ is a polynomial of degree $d-2$. The assumption $\tilde{s}(f) \geq d-2$ implies that $f - c_1 g$ is coercive for all $c_1 > 0$. Hence there is M, depending on c_1, such that

$$f(x) - c_1\|x\|_2^{d-2} \geq 0$$

holds for $\|x\| \geq M$. As $d \geq 2$, we may use Observation 7.1 to find that f is $d-2$-coercive. Now Lemma 7.12 states that $o(f) > d - 2$, which finishes the proof. $\square$

7.4. Example families

7.4.1. Introductory remarks

In this section, we give two explicit example families of coercive polynomials with arbitrarily small but positive order of growth. The first family has a bounded number of variables (two) but varying degree and the second family has a bounded degree (four) but a varying number of variables.

There are some examples families of $\{f_i\}_{i\in I} \subset \mathbb{R}[X_1, \ldots, X_n]$, where I is some index set, in the literature where the Łojasiewicz exponents at infinity $\mathcal{L}_\infty(f_i)$ of the f_i – and hence the order of coercivity $o(f_i)$ of f_i, if f_i is coercive – are explicitly computed, e.g., [Gor61], [Kra07]. These example families are extensive in the following sense: For every $q \in \mathbb{Q}$ there is $i \in I$ with $\mathcal{L}_\infty(f_i) = q$. Hence, example polynomials with arbitrarily small order of growth are easily given.

However, these example families from the literature were not created with the objective in mind to keep the number of variables and the degree of the resulting polynomials low. The examples we present are, in this sense, not only some further polynomials with known Łojasiewicz exponents at infinity.

In the literature, the computations are rather terse. We take a different route and carefully prove all assertions. These proofs are simplified by partitioning the domain of definition. Specifically, given $S' \subset S$, we write $o(f|S')$ for the order of coercivity of f restricted to S'. Then, in view of the immediate Observation 7.17, we may compute the order of coercivity on more suitable subsets of $\mathbb{R}^n$ instead of on all of $\mathbb{R}^n$.

Observation 7.17. *Let $S_1, \ldots, S_k \subset \mathbb{R}^n$, $S := \cup_{i=1}^k S_i$ and $f : S \to \mathbb{R}$ coercive. Then*

$$o(f) = \min_{1\le i\le k} o(f|S_i).$$

The following handy observation is immediate.

Observation 7.18. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by $f(x) = \sum_{j=1}^n c_j|x_j|^{\alpha_j}$ for some $c_j > 0$, $\alpha_j > 0$. Then $o(f) = \min_j \alpha_j$.*

7.4.2. Fixed number of variables

We give now an example of a family of coercive polynomials of arbitrarily small (but, of course, positive) order of growth in two variables. The key observation is that the function $\mathbb{R}^2 \to \mathbb{R}$, $(x, y) \mapsto x^2$, is $\frac{1}{k}$-coercive on (the image of) the curve $\gamma : \mathbb{R} \to \mathbb{R}^2$, $t \mapsto (t, t^{2k})$.

Proposition 7.19. *Consider the polynomial $f_k \in \mathbb{R}[X, Y]$, $k \in \mathbb{N}$, given by*

$$f_k = X^2 + \left(Y - X^{2k}\right)^2. \tag{7.4}$$

Then $o(f_k) = \frac{1}{k}$.

Note that by Theorem 7.15, $s(f_k) = \tilde{s}(f_k) = 0$ holds for all $k \geq 2$, thus even small linear perturbations of f_k may lead to the loss of coercivity.

Corollary 7.20. *For any given $\rho > 0$ there is a polynomial that is coercive but not ρ-coercive; this even holds if the number of variables is fixed to 2.*

We split the proof of Proposition 7.19 into two Lemmata.

Lemma 7.21. *For f_k as in (7.4), $o(f_k) \geq \frac{1}{k}$.*

Proof. The proof is by case distinction on a given point $(x, y) \in \mathbb{R}^n$. Put

$$\begin{aligned} S^{\downarrow} &:= \{(x, y) \in \mathbb{R}^2 : y < 0\}, \\ S^{\leftrightarrow} &:= \{(x, y) \in \mathbb{R}^2 : 0 \leq y < 2x^{2k}\}, \\ S^{\uparrow} &:= \{(x, y) \in \mathbb{R}^2 : 2x^{2k} \leq y\}, \end{aligned}$$

and observe that these sets are a partition of $\mathbb{R}^n$.

1. $(x, y) \in S^{\downarrow}$. Then $y < 0$ and hence

   $$f_k(x, y) = x^2 + (-|y| - x^{2k})^2 \geq x^2 + y^2 + x^{4k},$$

   thus $o(f_k|S^{\downarrow}) \geq 2$ by Observation 7.18.

2. $(x, y) \in S^{\leftrightarrow}$. Thus $x^{2k} > \frac{1}{2}|y|$, or $x^2 > \frac{1}{\sqrt[k]{2}}|y|^{1/k}$ and we find

   $$f_k(x, y) \geq \frac{1}{2}x^2 + \frac{1}{2}x^2 \geq \frac{x^2}{2} + \frac{1}{\sqrt[k]{2}}|y|^{1/k},$$

   hence $o(f_k|S^{\leftrightarrow}) \geq \frac{1}{k}$.

3. $(x, y) \in S^{\uparrow}$. Then $y \geq 2x^{2k}$, equivalently, $y - x^{2k} \geq \frac{1}{2}y$. As y is non-negative,

   $$f_k(x, y) \geq x^2 + \left(\frac{y}{2}\right)^2 = x^2 + \frac{y^2}{4},$$

   which yields $o(f_k|S^{\uparrow}) \geq 2$.

The claim follows now from Observation 7.17.

□

Lemma 7.22. *For f_k as in (7.4), $o(f_k) \leq \frac{1}{k}$.*

Proof. Assume $o(f_k) > \frac{1}{k}$. By Observation 7.1 and continuity of f_k, there are $c_1 > 0$, $c_2 \geq 0$ and $\rho > \frac{1}{k}$ with

$$f_k(x, y) \geq c_1 \|(x, y)\|_1^{\rho} - c_2, \quad (x, y) \in \mathbb{R}^2.$$

Let $\{x_n\}_{n\in\mathbb{N}}$ be a sequence of reals with $\lim_{n\to\infty} x_n = +\infty$. We define another sequence $y_n := x_n^{2k}$. Thus $x_n^2 = \sqrt[k]{y_n}$ and

$$f_k(x_n, y_n) = x_n^2 = y_n^{1/k} \geq c_1\|(x_n, y_n)\|_1^\rho - c_2 \geq c_1\|(0, y_n)\|_1^\rho - c_2 = c_1 y_n^\rho - c_2.$$

We shorten the last estimate to the inequality

$$y_n^{1/k} \geq c_1 y_n^\rho - c_2, \quad n \in \mathbb{N},$$

which yields a contradiction: Since $\rho > \frac{1}{k}$, $c_1 > 0$ and $\lim_{n\to\infty} y_n = +\infty$, so this inequality is eventually violated. □

7.4.3. Fixed degree

Our second example is a family of coercive polynomials of arbitrarily small order of growth with a degree fixed to four. The geometric idea behind this family is similar to the one before: The function $\mathbb{R}^n \to \mathbb{R}$, $x \mapsto x_1^2$ is 2^{2-n} coercive on (the image of) the curve

$$\gamma : \mathbb{R} \to \mathbb{R}^n,\ t \mapsto (t, t^2, t^4, t^8, \ldots, t^{2^{n-2}}, t^{2^{n-1}},)$$

To model this curve as the zero set of a single polynomial, we use the fact that for real polynomials $h_1, \ldots, h_s \in \mathbb{R}[X_1, \ldots, X_n]$ and $x \in \mathbb{R}^n$, the following holds:

$$h_1(x) = \ldots = h_s(x) = 0 \iff \sum_{i=1}^{s} h_i(x)^2 = 0.$$

More specifically, the term $\sum_{i=1}^{n-1} (X_{i+1} - X_i^2)^2$ vanishes at x if and only if $x_{i+1} - x_i^2 = 0$ for all $i \in \{2, \ldots, n\}$ if and only if x lies on the curve γ, i.e., if and only if x satisfies $x_n = x_{n-1}^2 = x_{n-2}^4 = x_{n-3}^8 = \cdots = x_1^{2^{n-1}}$.

Proposition 7.23. *Consider the polynomial $g_n \in \mathbb{R}[X_1, \ldots, X_n]$, $n \in \mathbb{N}$, given by*

$$g_n = X_1^2 + \sum_{i=2}^{n} \left(X_i - X_{i-1}^2\right)^2. \tag{7.5}$$

Then $o(g_n) = 2^{2-n}$.

Note that by Theorem 7.15, $s(g_n) = \tilde{s}(g_n) = 0$ holds for all $n \geq 3$, thus even small linear perturbations of g_n may lead to the loss of coercivity.

Corollary 7.24. *For any given $\rho > 0$ there is a polynomial that is coercive but not ρ-coercive; this even holds if the degree is fixed to 4.*

The proof of Proposition 7.23 is divided into three lemmata.

Lemma 7.25. *Let*

$$C := \{x \in \mathbb{R}^n : |x_i| \geq 1 \text{ for all } i \in [n]\}.$$

Then $o(g_n|C) \geq 2^{2-n}$ for g_n as in (7.5).

Proof. We introduce the functions on C

$$T_1(x) := x_1^2, \quad T_i(x) := (x_i - x_{i-1}^2)^2, \quad i = 2, \dots, n,$$
$$Q_i(x) := \frac{1}{8^i}|x_i|^{2^{2-i}}, \quad i = 1, \dots, n,$$

so $g_n(x) = \sum_{i=1}^n T_i(x)$ on C. The claim follows from Observation 7.18 if we can prove by induction

$$\sum_{i=1}^{j} T_i(x) \geq \sum_{i=1}^{j-1} Q_i(x) + 2Q_j(x), \quad x \in C,\ j = 1, \dots, n. \tag{7.6}$$

The claim in 7.6 trivially holds for $j = 1$. Assume it holds for some $j < n$. For the inductive step it suffices to show that for an arbitrary $x \in \mathbb{R}^n$ one of

$$T_{j+1}(x) \geq 2Q_{j+1}(x) \tag{7.7}$$

or

$$Q_j(x) \geq 2Q_{j+1}(x) \tag{7.8}$$

holds. Indeed, in case (7.7) holds at x, then adding this inequality to (7.6) yields

$$\sum_{i=1}^{j+1} T_i(x) \geq \sum_{i=1}^{j-1} Q_i(x) + 2Q_j(x) + 2Q_{j+1}(x) \geq \sum_{i=1}^{j} Q_i(x) + 2Q_{j+1}(x)$$

In the other case, (7.8) holds at x. Then

$$\sum_{i=1}^{j+1} T_i(x) \geq \sum_{i=1}^{j} T_i(x) \geq \sum_{i=1}^{j} Q_i(x) + Q_j(x) \geq \sum_{i=1}^{j} Q_i(x) + 2Q_{j+1}(x).$$

Now let us show by case distinction on $x \in \mathbb{R}^n$ why (7.7) or (7.8) holds. Again, we introduce a partition

$$S_i^{\downarrow} := \{x \in C : x_i < 0\},$$
$$S_i^{\leftrightarrow} := \{x \in C : 0 \leq x_i < 2x_{i-1}^2\},$$
$$S_i^{\uparrow} := \{x \in C : 2x_{i-1}^2 \leq x_i\},$$

for $i = 1, \dots, n-1$. Now fix $x \in C$ and consider the cases

1. $x \in S_{j+1}^{\downarrow}$. Thus $x_{j+1} < 0$, and as $x \in C$, we may use monotonicity of exponentials to find

   $$T_{j+1}(x) = \left(-|x_{j+1}| - x_j^2\right)^2 \geq x_{j+1}^2 \geq |x_{j+1}|^{2^{2-(j+1)}} \geq 2Q_{j+1}(x),$$

   so 7.7 holds.

2. $x \in S^{\leftrightarrow}_{j+1}$. Thus $|x_j|^2 > \frac{1}{2}|x_{j+1}|$ and raising both sides to the $2^{2-(j+1)}$-th power,

$$|x_j|^{2^{2-j}} \geq \frac{1}{2^{2^{2-(j+1)}}} |x_{j+1}|^{2^{2-(j+1)}} \tag{7.9}$$

As $2-(j+1) \leq 0$, we have $\frac{1}{2^{2^{2-(j+1)}}} \geq \frac{1}{4}$. Hence, dividing both sides of 7.9 by 8^j, we see that (7.8) holds.

3. $x \in S^{\uparrow}_{j+1}$. Equivalently, $x_{j+1} - x_j^2 \geq \frac{1}{2}x_{j+1}$, thus by monotonicity again,

$$T_{j+1}(x) \geq \frac{1}{4}x_{j+1}^2 \geq \frac{1}{4}|x_{j+1}|^{2^{2-(j+1)}} \geq 2Q_{j+1}(x),$$

that is, (7.7) holds.

Hence, (7.6) holds for $j+1$ and all $x \in C$, so the induction step is proved. □

Lemma 7.26. *Let*

$$D := \{x \in \mathbb{R}^n : |x_i| < 1 \text{ for some } i \in [n]\}.$$

Then $o(g_n|D) \geq 2^{2-n}$ *for* g_n *as in* (7.5).

Proof. Suppose not. Thus there is a sequence $\{x_m\}_{m\in\mathbb{N}} \subset D$ with $\|x_m\|_\infty \to +\infty$ for $m \to \infty$, and

$$g_n(x_m) = x_{m,1}^2 + \sum_{i=2}^{n} \left(x_{m,i} - x_{m,i-1}^2\right)^2 \leq c\|x_m\|_\infty^{2^{2-n}}, \quad m \in \mathbb{N}.$$

Especially,

$$x_{m,1}^2 \leq c\|x_m\|_\infty^{2^{2-n}}, \quad \left(x_{m,i} - x_{m,i-1}^2\right)^2 \leq c\|x_m\|_\infty^{2^{2-n}}, \quad i = 2, \ldots, n. \tag{7.10}$$

We arrive at a contradiction if we can show by induction that there are $N_{n-1} \geq \ldots \geq N_0$ with

$$|x_{m,n-j}| \geq \frac{1}{2^j}\|x_m\|_\infty^{2^{-j}}, \quad m \geq N_j, \quad j = 0, \ldots, n-1. \tag{7.11}$$

Indeed, once the induction is complete, inequality (7.11) holds for all j and all $m \geq N_{n-1}$, and as $\|x_m\|_\infty$ grows without bound, (7.11) forces x_m to leave the set D, contradicting the assumption $x_m \in D$ for all m.

For the basis of the induction, we use the second inequality in (7.10) to find $|x_{m,i-1}^2 - x_{m,i}| \leq c^{1/2}\|x_m\|_\infty^{2^{1-n}}$ and thus

$$x_{m,i-1}^2 \leq |x_{m,i}| + c^{1/2}\|x_m\|_\infty^{2^{1-n}}$$

by the reverse triangle inequality. Using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$, the last inequality yields

$$|x_{m,i-1}| \leq |x_{m,i}|^{1/2} + c^{1/4}\|x_m\|_\infty^{2^{-n}} \leq \|x_m\|_\infty^{1/2} + c^{1/4}\|x_m\|_\infty^{2^{-n}}$$

By assumption on x_m, $\|x_m\|$ grows without bound, so there is $N_{-1} \in \mathbb{N}$ with $\|x_m\| \geq 1$ for $m \geq N_{-1}$, and then $\|x_m\|_\infty^{2^{-n}} \leq \|x_m\|_\infty^{1/2}$. Thus

$$|x_{m,i-1}| \leq \left(1 + c^{1/4}\right) \|x_m\|_\infty^{1/2}. \tag{7.12}$$

Also, there is $N_0 \geq N_{-1}$ with $\left(1 + c^{1/4}\right)^2 < \|x_m\|_\infty$ for $m \geq N_0$, which together with (7.12) implies $|x_{m,i-1}| < \|x_m\|_\infty$ for $m \geq N_0$ and $i = 2, \ldots, n$, that is,

$$|x_{m,n}| = \|x_m\|_\infty \quad \text{for } m \geq N_0,$$

a rewording of the basis of the induction.

For the inductive hypothesis, suppose (7.11) holds for some $j < n - 1$. We now prove the inductive step. Using the reverse triangle inequality on (7.10) the other way, we find

$$x_{m,j-1}^2 \geq |x_{m,j}| - c^{1/2}\|x_m\|_\infty^{2^{1-n}}, \quad j = 2, \ldots, n. \tag{7.13}$$

With (7.13) and the inductive hypothesis,

$$x_{m,n-(j+1)}^2 \geq |x_{m,n-j}| - c^{1/2}\|x_m\|_\infty^{2^{1-n}} \geq \frac{1}{2^j}\|x_m\|_\infty^{2^{-j}} - c^{1/2}\|x_m\|_\infty^{2^{1-n}} \tag{7.14}$$

On the other hand, as $\|x_m\|_\infty$ grows without bound, there is $N_{j+1} \geq N_j$ with

$$\begin{aligned}
& \|x_m\|_\infty \geq \left(2^{j+1}c^{1/2}\right)^{1/(2^{-j}-2^{1-n})} \\
\iff \quad & \|x_m\|_\infty^{2^{-j}-2^{1-n}} \geq 2^{j+1}c^{1/2} \\
\iff \quad & \|x_m\|_\infty^{2^{-j}} \geq 2^{j+1}c^{1/2} \cdot \|x_m\|_\infty^{2^{1-n}} \\
\iff \quad & \frac{1}{2^{j+1}}\|x_m\|_\infty^{2^{-j}} - c^{1/2} \cdot \|x_m\|_\infty^{2^{1-n}} \geq 0 \\
\iff \quad & \frac{1}{2^j}\|x_m\|_\infty^{2^{-j}} - c^{1/2} \cdot \|x_m\|_\infty^{2^{1-n}} \geq \frac{1}{2^{j+1}}\|x_m\|_\infty^{2^{-j}}
\end{aligned}$$

for $m \geq N_{j+1}$. With (7.14) we deduce

$$x_{m,n-(j+1)}^2 \geq \frac{1}{2^j}\|x_m\|_\infty^{2^{-j}} - c^{1/2}\|x_m\|_\infty^{2^{1-n}} \geq \frac{1}{2^{2(j+1)}}\|x_m\|_\infty^{2^{-j}}$$

and hence

$$|x_{m,n-(j+1)}| \geq \frac{1}{2^{j+1}}\|x_m\|_\infty^{2^{-(j+1)}},$$

proving the induction step. □

Lemma 7.27. *For g_n as in (7.5), $o(g_n) \leq 2^{2-n}$.*

Proof. Assume $o(g_n) > 2^{2-n}$. Using Observation 7.1 and continuity of g_n, there are $c_1 > 0$, $c_2 \geq 0$ and $\rho > 2^{2-n}$ with

$$g_n(x) \geq c_1 \|x\|_1^\rho - c_2 \quad \forall x \in \mathbb{R}^n. \tag{7.15}$$

Let $\{a_m\}_{m\in\mathbb{N}}$ be a sequence of reals with $\lim_{m\to\infty} a_m = +\infty$, and define $\{x_m\}_{m\in\mathbb{N}} \subset \mathbb{R}^n$ with components $(x_m)_1 := a_m$ and

$$(x_m)_2 := ((x_m)_1)^2\,,\ (x_m)_3 := ((x_m)_2)^2\,,\ \ldots,\ (x_m)_n := ((x_m)_{n-1})^2\,.$$

Observe that $(x_m)_1 = ((x_m)_n)^{2^{2-n}}$ and $(x_m)_n \to +\infty$ for $n \to \infty$. Then

$$g_n(x_m) = (x_m)_1^2 + \left(\sum_{i=2}^{n-1} 0^2\right) = (x_m)_1^2 = (x_m)_n^{2^{2-n}} \geq c_1\|x_m\|_1^\rho - c_2$$

by definition of x_m and by (7.15), and we may estimate further

$$\geq c_1\|\,(0,\ldots,0,(x_m)_n)\,\|_1^\rho - c_2 = c_1|(x_m)_n|^\rho - c_2 = c_1(x_m)_n^\rho - c_2$$

which contains the contradictory inequality

$$(x_m)_n^{2^{2-n}} \geq c_1(x_m)_n^\rho - c_2, \quad m \in \mathbb{N}. \tag{7.16}$$

Indeed, as $(x_m)_n \to +\infty$ for $m \to \infty$ and $c_1 > 0$, $\rho > 2^{2-n}$, inequality (7.16) eventually be violated. □

7.5. Minimal order of coercivity and outlook

7.5.1. Minimal order of coercivity

In Section 7.4 we have seen explicit examples of slowly growing coercive polynomials where either the number of variables n or the degree d are fixed. It is thus only a natural question to ask how small the order of growth can get when both the number of variables and the degree of the polynomial are fixed. In other words, we consider for $n \in \mathbb{N}$ and $d \in 2\mathbb{N}$ the number

$$\mathfrak{o}(n,d) = \inf\left\{o(f) : f \in \mathbb{R}[X_1,\dots,X_n]_d \text{ is coercive}\right\}.$$

We call $\mathfrak{o}(n,d)$ the *minimum possible order of coercivity* of a coercive polynomial in n variables of degree d. It is not known to us whether there is a closed formula for $\mathfrak{o}(n,d)$ or if at least $\mathfrak{o}(n,d) > 0$ for all $n \in \mathbb{N}$ and $d \in 2\mathbb{N}$. Also, we do not know if our example families f_k and g_n from Section 7.4 are minimal examples in the sense that $o(f_k) = \mathfrak{o}(2,4k)$ or $o(g_n) = \mathfrak{o}(n,4)$.

Table 7.1 summarizes the special cases and examples discussed in this article. A star $(*)$ indicates arbitrary values; that is, $n \in \mathbb{N}$ or $d \in 2\mathbb{N}$.

n	d	Upper bound on $\mathfrak{o}(n,d)$	Attainment	Reference
$*$	2	2	yes	Example 7.4
1	$*$	d	yes	cf. Section 7.1
2	$4k$	$1/k$	?	Proposition 7.19
≥ 2	4	2^{2-n}	?	Proposition 7.23

Table 7.1.: Upper bounds on the minimum possible order of coercivity $\mathfrak{o}(n,d)$.

7.5.2. Future directions

In [BS15a] a class of coercive polynomials is identified where coercivity can be verified by analyzing properties of the underlying Newton polytopes at infinity. Then, in [BS15b], it is shown that for each polynomial from the aforementioned class one always has $o(f) = s(f) = c(f) \in 2\mathbb{N}$ with $c(f)$ denoting the so-called degree of convenience of f – which is the length of the shortest intercept of the Newton polytope at infinity with the n coordinate axes. So, for coercive polynomials with a fractional order of growth, for example such as those from Section 7.4, it would be an interesting question whether their order of growth is encoded in their Newton polytopes as well.

7.6. Deciding coercivity

This section is devoted to the following problem: Given a polynomial $f \in \mathbb{R}[X_1, \ldots, X_n]$, decide whether it is coercive. Our first result is that the decision problem is decidable. We use quantifier elimination to derive an exponential complexity result. Secondly, we explore how knowledge of a minimal order of coercivity can be used to prove coercivity using sos programming. Thirdly, we explore whether coercivity can be decided on exponential curves. Lastly, we show that we may compose f with a homeomorphism $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ and decide if f is coercive by deciding whether $f \circ \varphi$ is coercive.

7.6.1. Decidability and complexity of the decision problem

In this section we show that coercivity of polynomials is a decidable problem and give a complexity result. To this end, we restate the definition of coercivity from (1.6) with quantifiers and the (squared) 2-norm as follows:

$$\forall c \in \mathbb{R} \quad \exists r \in \mathbb{R} \quad \forall x \in \mathbb{R}^n : \left(\sum_{i=1}^{n} x_i^2 - r \geq 0 \Longrightarrow f(x) - c \geq 0 \right). \tag{7.17}$$

An expression as in (7.17) is an example of a *formula* in the *language of ordered fields* in the variables C, R, X_1, …, X_n; for a precise definition of this language see Ch. 2.3 in [BPR05].[1] As all appearing variables are quantified, this formula does not contain *free variables*. There is a theorem that states that the *language of real closed fields* – real closed fields are a special type of an ordered field that enjoy the property that for every univariate polynomial with coefficients in these fields a sign change at a and b implies a zero in between; $\mathbb{R}$ is an example – admits quantifier elimination:

Theorem 7.28 (see, e.g., Theorem 2.77 in [BPR05])**.** *Suppose we are given a formula $\Phi(Y)$ in the language of ordered fields, where $Y = (Y_1, \ldots, Y_k)$ are the free variables of the formula, over a real closed field R. Then, there is a quantifier free formula $\Psi(Y)$ over R such that for every $y \in R^k$ the formula $\Phi(y)$ is true if and only if the formula $\Psi(y)$ is true.*

As (7.17) does not contain free variables, it is called a *sentence* and, by quantifier elimination, it is $\mathbb{R}$-equivalent to true or false. For such sentences over real closed fields, solution algorithms exist. To specify the running time, we need some more notation. Let $\mathcal{P} \subset R[X_1, \ldots, X_k]$ be a finite set of polynomials. A $\mathcal{P}$-*atom* is one of $P = 0$, $P > 0$, $P < 0$, $P \neq 0$ for a polynomial $P \in \mathcal{P}$. Then, a $\mathcal{P}$-*formula* is a formula (in the language of ordered fields) written with $\mathcal{P}$-atoms, cf. p. 417 in [BPR05]. Moreover, let Π denote a partition of the list of variables $(X_1, \ldots, X_n)$ into blocks $X_{[1]}, \ldots, X_{[\omega]}$, where the block $X_{[i]}$ has size k_i, $1 \leq i \leq \omega$, and $\sum_{i=1}^{\omega} k_i = k$. Then, a $(\mathcal{P}, \Pi)$-*formula* is a formula of the form

$$\Psi(Y) = (\mathrm{Qu}_1 X_{[1]}) \cdots (\mathrm{Qu}_{[w]}) F(X, Y)$$

[1] Roughly speaking, the formulas of this language consist of equations and inequalities of polynomials with coefficients in R, boolean combinations thereof, and existential and universal quantifiers.

where each Qu_i is one of the quantifiers $\exists$ or $\forall$, $Y = (Y_1, \ldots, Y_l)$ and $F(X, Y)$ is a quantifier free $\mathcal{P}$-formula, cf. p. 537 in [BPR05]. Consider the following complexity result for an algorithm that decides the truth of a sentence.

Theorem 7.29 (see, e.g., Theorem 14.14 in [BPR05])**.** *Let $\mathcal{P}$ be a set of at most s polynomials each of degree at most d in k variables with coefficients in a real closed field R, and let Π denote a partition of the list of variables $(X_1, \ldots, X_k)$ into blocks $X_{[1]}, \ldots, X_{[\omega]}$, where the block $X_{[i]}$ has size k_i, $1 \leq i \leq \omega$. Given a $(\mathcal{P}, \Pi)$ sentence Π, there exists an algorithm to decide the truth of Ψ with complexity*

$$s^{(k_\omega+1)\cdots(k_1+1)} d^{O(k_\omega)\cdots O(k_1)}$$

in D, where D is the ring generated by the coefficients of $\mathcal{P}$.

Here, complexity in a structure D is the (worst-case) number of operations in D as a function of the input sizes, see Chapter 8.1 in [BPR05].

We may now state our complexity result.

Proposition 7.30. *The decision problem*

$$\text{Given a polynomial } f \in \mathbb{R}[X_1, \ldots, X_n] \text{ with} \deg f \geq 2, \text{ is it coercive?} \tag{7.18}$$

can be solved by an algorithm with complexity

$$(2 \deg f)^{O(n)}$$

in D, where D is the ring generated by the coefficients of $\mathcal{P}$.

Proof. In our setting, $\mathcal{P} = \{f - C, \sum_{i=1}^n X_i^2 - R\} \subset \mathbb{R}[X_1, \ldots, X_n, C, R]$. For the formula (7.17) we may choose a partition with $\omega = 3$ and $k_1 = n$, $k_2 = k_3 = 1$, i.e., $\Pi = ((X_1, \ldots, X_n), C, R)$. Hence (7.17) is a $(\mathcal{P}, \Pi)$-formula, and with $d = \max\{\deg f, 2\} = \deg f$, $s = |\mathcal{P}| = 2$, we get using Theorem 7.29

$$s^{(1+1)\cdot(1+1)\cdot(n+1)} d^{O(1)\cdot O(1)\cdot O(n)} = s^{O(n)} d^{O(n)} = (sd)^{O(n)} = (2 \deg f)^{O(n)}.$$

□

Note that the addendum "complexity in D" also means that we do not need to worry about the technicalities arising through the encoding of and computing with real numbers if we restrict f to, say, rational coefficients.

7.6.2. Deciding coercivity via the minimal order of coercivity

The worst-case running time of the algorithm in Proposition 7.30 is exponential in n and software-based quantifier elimination is rather slow in practice. However, if the value for $\mathfrak{o}(n, d)$ – the minimal order of coercivity as introduced in Section 7.5 – is known (by a general formula or otherwise) and positive, we can use sos programming to certify coercivity. As a preparatory step, consider the following proposition.

Proposition 7.31. *Let $f \in \mathbb{R}[X_1, \ldots, X_n]$ of degree d. Also, assume that $\mathfrak{o}(n,d) = \rho > 0$. Choose any $p, q \in \mathbb{N}$, p even and q odd, with $\rho > p/q$. Then, the following are equivalent:*

1. *f is coercive.*
2. *f is p/q-coercive.*
3. *$f^q/\|x\|_p^p \to +\infty,\ \|x\|_p \to +\infty$.*
4. *For all $c_1 > 0$ exists $c_2 \geq 0$ such that $f^q(x) \geq c_1\|x\|_p^p - c_2$, $x \in \mathbb{R}^n$.*
5. *There are $c_1 > 0$ and $c_2 \geq 0$ such that $f^q(x) \geq c_1\|x\|_p^p - c_2$, $x \in \mathbb{R}^n$.*
6. *$g := f^q - \sum_{i=1}^n X_i^p \in \mathbb{R}[X_1, \ldots, X_n]$ is bounded from below.*

Proof. Note that, as $\mathfrak{o}(n,d) > 0$ by assumption, such p and q exist. If we can show that $1 \Rightarrow 2 \Rightarrow 3 \Rightarrow 4 \Rightarrow 5 \Rightarrow 1$ and then $4 \Rightarrow 6$ and $6 \Rightarrow 5$, all statements are equivalent. $1 \Rightarrow 2$. As $\mathfrak{o}(n,d) = \rho > 0$ we must have $o(f) \geq \rho$, and so f must be ρ'-coercive for all $0 \leq \rho' < \rho$, especially f is p/q-coercive. $2 \Rightarrow 3$. Since all norms are equivalent on $\mathbb{R}^n$, f is p/q-coercive if and only if $f/\left(\|x\|_p\right)^{\frac{p}{q}} \to +\infty,\ \|x\|_p \to +\infty$. Also, a function $h : \mathbb{R}^n \to \mathbb{R}$ is coercive if and only if h^r is coercive, $r \in \mathbb{N}$ odd. $3 \Rightarrow 4$. This follows from continuity and Lemma 7.1. $4 \Rightarrow 5$. This is trivial. $5 \Rightarrow 1$. By Observation 7.2, $o(f^q) \geq p$, hence f^q is coercive and thus f, too. $4 \Rightarrow 6$. Let $c_1 = 1$ and choose c_2 accordingly. Now g is bounded from below by $-c_2$. $6 \Rightarrow 5$. If g is bounded from below by $C \in \mathbb{R}$, statement 5 holds for $c_1 = 1$ and $c_2 = -C$. □

Now, Statement (6) of Proposition 7.31 can be used as follows:

Corollary 7.32. *Let the assumptions of Proposition 7.31 hold. If the sos program*

$$\begin{aligned} \max \quad & \lambda \\ s.t. \quad & f^q - \sum_{i=1}^n X_i^p - \lambda \text{ is sos in } \mathbb{R}[X_1, \ldots, X_n] \\ & \lambda \in \mathbb{R} \end{aligned}$$

has a feasible solution λ, f is coercive.

Proof. Let $\lambda \in \mathbb{R}$ be a feasible solution. As being sos implies nonnegativity,

$$f^q(x) - \sum_{i=1}^n x_i^p - \lambda \geq 0$$

for all $x \in \mathbb{R}^n$. In other words, $g = f^q - \sum_{i=1}^n X_i^p$ is bounded from below by λ. The claim follows from Proposition 7.31 (6). □

7.6.3. Deciding coercivity on curves

This section solves an open problem in the article [BS15a]. We show that coercivity of a polynomial $f \in \mathbb{R}[X_1, \ldots, X_n]$ cannot be decided by restricting f on certain curves. To this end we introduce some notation from the reference. Let

$$Y := \left\{ y \in \mathbb{R}^n : \prod_{i \in [n]} y_i \neq 0 \right\},$$

and put

$$\mathbb{H} := \{h \in \mathbb{R}^n : h_i \geq 0 \text{ for all } i \in [n]\}.$$

Furthermore, put

$$\Omega := Y \times B.$$

Given $y,\ \beta$ in $\mathbb{R}^n$, define a curve $x_{y,\beta} : \mathbb{R} \to \mathbb{R}^n$ via

$$x_{y,\beta}(t) := (y_1 e^{\beta_1 t}, \ldots, y_n e^{\beta_n t}).$$

Finally, for $f \in \mathbb{R}[X_1, \ldots, X_n]$, we define

$$\Omega_f := \left\{ (y, \beta) \in \mathbb{R}^n \times \mathbb{R}^n : \lim_{t \to +\infty} f(x_{y,\beta}(t)) = +\infty \right\}.$$

Now, we have the following result:

Proposition 7.33 (Lemma 2.2 in [BS15a])**.** *If $f \in \mathbb{R}[X_1, \ldots, X_n]$ is coercive on $\mathbb{R}^n$, then $\Omega \subset \Omega_f$.*

The reverse statement does not hold, as the following proposition shows.

Proposition 7.34. *Let $f \in \mathbb{R}[X_1, X_2]$ given by*

$$f = (X_2 - X_1 - 1)^2 (X_1^2 + X_2^2).$$

Then f is not coercive but $\Omega \subset \Omega_f$.

The proof needs a little preparation. The notion of an *asymptotic direction* of a curve captures some of the behavior of the curve at infinity. Asymptotic directions are elements of the Euclidean unit sphere $\mathbb{S}_2^{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$.

Definition 7.35. Let $\gamma : \mathbb{R} \to \mathbb{R}^n$ be continuous with $\lim_{t \to +\infty} \|\gamma(t)\|_2 = +\infty$. We say the curve γ has the asymptotic direction $\omega \in \mathbb{S}_2^{n-1}$ if

$$\lim_{t \to +\infty} \frac{\gamma(t)}{\|\gamma(t)\|_2} = \omega,$$

or $\mathcal{D}(\gamma) = \omega$ for short.

We can now compute the asymptotic directions of the curves $x_{y,\beta}$ in $n = 2$ dimensions. To this end let $\operatorname{sign}(y)$ be the sign of $y \in \mathbb{R}$ where $\operatorname{sign}(0) := 0$.

Lemma 7.36. *Let $n = 2$ and $(y, \beta) \in \Omega$. Then an asymptotic direction for $x_{y,\beta}$ exists. More precisely, for β we have exactly one of the following cases:*

a) $\beta_1 = 0$. Then $\mathcal{D}(x_{y,\beta}) = (0, \operatorname{sign}(y_2))$ and $x_{y,\beta}$ is a line parallel to the x_2-axis.

b) $\beta_2 = 0$. Then $\mathcal{D}(x_{y,\beta}) = (\operatorname{sign}(y_1), 0)$ and $x_{y,\beta}$ is a line parallel to the x_1-axis.

c) $\beta_1 < 0$. Then $\mathcal{D}(x_{y,\beta}) = (0, \operatorname{sign}(y_2))$.

d) $\beta_2 < 0$. Then $\mathcal{D}(x_{y,\beta}) = (\operatorname{sign}(y_1), 0)$.

e) $\beta_1 = \beta_2 > 0$. Then $\mathcal{D}(x_{y,\beta}) = \frac{y}{\|y\|_2}$ and $x_{y,\beta}$ is a line through the origin.

f) $\beta_1 > \beta_2 > 0$. Then $\mathcal{D}(x_{y,\beta}) = (\operatorname{sign}(y_1), 0)$.

g) $\beta_2 > \beta_1 > 0$. Then $\mathcal{D}(x_{y,\beta}) = (0, \operatorname{sign}(y_2))$.

Proof. Note that by definition of Y, $y_i \neq 0$ for all $i \in [n]$ throughout this proof. As $\beta \in B$, we have $\beta_1 \leq 0 \Rightarrow \beta_2 > 0$ and $\beta_2 \leq 0 \Rightarrow \beta_1 > 0$. Hence, a), b), c) and d) follow by standard arguments, we show c) as an example. Indeed, the first component of $x_{y,\beta}(t)/\|x_{y,\beta}(t)\|_2$ converges to zero for $t \to +\infty$ as the nominator is bounded and the denominator is unbounded. The second component is

$$\frac{y_2 e^{\beta_2 t}}{\sqrt{y_1^2 e^{2\beta_1 t} + y_2^2 e^{2\beta_2 t}}} = \frac{y_2 e^{\beta_2 t}}{|y_2| e^{\beta_2 t}} \frac{1}{\sqrt{\frac{y_1^2}{y_2^2} e^{2(\beta_1 - \beta_2)t} + 1}} = \frac{\operatorname{sign}(y_2)}{\sqrt{\frac{y_1^2}{y_2^2} e^{2(\beta_1 - \beta_2)t} + 1}},$$

and the denominator converges to 1 for $t \to +\infty$, as $\beta_1 < 0$ and $\beta_2 > 0$.

Also, e) is clear. To see f), we observe

$$(y_1 e^{\beta_1 t}, y_2 e^{\beta_2 t}) = e^{\beta_2 t}(y_1 e^{(\beta_1 - \beta_2)t}, y_2),$$

and by absolute homogeneity of the norm, the factor $e^{\beta_2 t}$ has no influence on the asymptotic direction. So we may neglect it, and the asymptotic direction is the same as in b), i.e. $\mathcal{D}(x_{y,(\beta_1,\beta_2)}) = \mathcal{D}(x_{y,(\beta_1-\beta_2,0)}) = (\operatorname{sign}(y_1), 0)$. The proof for g) is similar. □

The following lemma allows us to prove Proposition 7.34. Roughly speaking, the lemma says that if the asymptotic direction of a curve γ exists and is not parallel to one of the two "asymptotic directions" of the zero locus of $g(x_1, x_2) := (x_2 - x_1 - 1)^2$, g cannot get arbitrarily small on $\gamma(t)$ for t large.

Lemma 7.37. *Let $\gamma : \mathbb{R} \to \mathbb{R}^2$ be continuous with $\lim_{t\to+\infty} \|\gamma(t)\|_2 = +\infty$ and $\mathcal{D}(\gamma) = \omega \in \mathbb{S}_2^1$. Put $g := (X_2 - X_1 - 1)^2$. If $\omega \neq \pm\frac{(1,1)}{\|(1,1)\|_2}$ there is $t_0 \in \mathbb{R}$ with*

$$g(\gamma(t)) \geq 1, \quad t \geq t_0.$$

Proof. Suppose the contrary and let $\gamma = (\gamma_1, \gamma_2)$. Thus, for every $n \in \mathbb{N}$ there is $t_n \geq n$ with

$$(\gamma_2(t_n) - \gamma_1(t_n) - 1)^2 < 1. \tag{7.19}$$

Now we add zero to find

$$\frac{(\gamma_1(t_n), \gamma_2(t_n)}{\|\gamma(t_n)\|_2} = \frac{(\gamma_1(t_n), \gamma_2(t_n) - \gamma_1(t_n) - 1 + \gamma_1(t_n) + 1)}{\|\gamma(t_n)\|_2}.$$

We observe, using equation (7.19) and $\gamma(t) \to \infty$ for $t \to \infty$, that

$$\lim_{n\to\infty} \underbrace{\frac{\gamma_2(t_n) - \gamma_1(t_n) - 1}{\|\gamma(t_n)\|_2}}_{=:A_n} \to 0, \quad \lim_{n\to\infty} \underbrace{\frac{1}{\|\gamma(t_n)\|_2}}_{=:B_n} = 0.$$

This implies

$$\begin{aligned}\omega &= \lim_{n\to\infty} \frac{\gamma(t_n)}{\|\gamma(t_n)\|_2} = \lim_{n\to\infty} \left(\frac{\gamma_1(t_n)}{\|\gamma(t_n)\|_2}, \frac{\gamma_1(t_n)}{\|\gamma(t_n)\|_2} + A_n + B_n \right) \\ &= \lim_{n\to\infty} \frac{(\gamma_1(t_n), \gamma_1(t_n))}{\|\gamma(t_n)\|_2} = (\omega_1, \omega_1),\end{aligned}$$

hence $\omega_2 = \omega_1$. However, as $\omega \in \mathbb{S}^1_2$, this forces $\omega = \pm(1,1)/\|(1,1)\|_2$, contradicting the assumption on ω. □

Proof of Proposition 7.34. To see that $f(x_1, x_2) = (x_2 - x_1 - 1)^2(x_1^2 + x_2^2)$, is not coercive, we observe that $f = 0$ on the line $x_2 = x_1 + 1$. To prove that $\Omega \subset \Omega_f$ we need to show

$$\lim_{t\to+\infty} \pi_f(y, \beta, t) = \lim_{t\to+\infty} f(x_{y,\beta}(t)) = +\infty, \quad (y, \beta) \in \Omega.$$

It is enough to show that $(x_2 - x_1 - 1)^2 \geq 1$ on $x_{y,\beta}(t)$ for large t, as the term $x_1^2 + x_2^2$ grows without bound for large t on the curve $x_{y,\beta}(t)$. We make the same case distinction on all possible values of β. In view of Lemmata 7.36 and 7.37, it is now clear that all choices of β except possibly $\beta_1 = \beta_2 > 0$, that is case e), imply coercive behaviour of f on $x_{y,\beta}(t)$. So let $\beta_1 = \beta_2 > 0$, hence $x_{y,\beta}$ suffices $\mathcal{D}(x_{y,\beta}) = y/\|y\|_2$. Lemma 7.37 tells us that we only need to consider the cases $y/\|y\|_2 = \pm(1,1)/\|(1,1)\|_2$. However, we also know that $x_{y,\beta}$ is a line through the origin, more precisely of the form $(x_{y,\beta})_2(t) = (x_{y,\beta})_1(t)$ – hence $|(x_{y,\beta})_2(t) - (x_{y,\beta})_1(t) - 1| = 1$ for all t, and f is coercive along $x_{y,\beta}(t)$ in this case as well. □

7.6.4. Invariance of coercivity under homeomorphisms

The aim of this section is to prove a small observation. When deciding coercivity of a function $f : \mathbb{R}^n \to \mathbb{R}$, polynomial or not, it is sometimes advantageous to change coordinates and then decide coercivity. We show in the following that a general class of changes of coordinates are admissible: Homeomorphisms. Recall that, provided (X, τ_X),

(Y, τ_Y) are topological spaces, a map $\varphi : X \to Y$ is a homeomorphism if φ is bijective and φ and φ^{-1} are continuous.

Note that $f : \mathbb{R}^n \to \mathbb{R}$ is coercive if and only if for every sequence $\{x_k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ with $\|x_k\| \to +\infty$, we have $f(x_k) \to +\infty$. For the proof of the following result, we introduce a notation: Given a sequence $\{x_k\}_{k\in\mathbb{N}}$, the common notation for a subsequence is $\{x_{k_l}\}_{l\in\mathbb{N}}$. However, for readability, we use $x_{k;l}$.

Proposition 7.38. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be any map and $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ be a homeomorphism. Then f is coercive if and only if $f \circ \varphi^{-1}$ is coercive.*

Proof. It is enough to show that for any sequence $\{x_k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$, the following holds: $\|x_k\| \to +\infty$ for $k \to \infty$ if and only if $\|\varphi^{-1}(x_k)\| \to +\infty$ for $k \to \infty$. By a symmetry argument, it is enough to show that $\|x_k\| \xrightarrow{k\to\infty} +\infty$ implies $\|\varphi(x_k)\| \xrightarrow{k\to\infty} +\infty$. Suppose this implication does not hold for one such sequence $\{x_k\}_{k\in\mathbb{N}}$. Thus, the sequence $\varphi(x_k)$ has a bounded subsequence $\varphi(x_{k;l})$. By the Bolzano-Weierstraß theorem, every bounded sequence in $\mathbb{R}^n$ has a convergent subsequence. Hence, there is a further subsequence $x_{k;l;m}$ such that $\varphi(x_{k;l;m})$ converges to some $z \in \mathbb{R}^n$. Let K be a compact neighborhood of z. Hence, K contains $\varphi(x_{k;l;m})$ for eventually all $m \in \mathbb{N}$. In other words, there is a subsequence $x_{k;l;m;p}$ with $\varphi(x_{k;l;m;p}) \in K$ for all $p \in \mathbb{N}$. As φ is bicontinuous, $\varphi^{-1}(K)$ is compact and contains the entire sequence $\{x_{k;l;m;p}\}_{p\in\mathbb{N}}$. However, this means $x_{k;l;m;p}$ is bounded, contradicting $\|x_k\| \to \infty$.

□

8. Summary and extensions

In this chapter, we briefly summarize our contributions and point out future research.

Section 8.1 contains the summary.

Section 8.2 outlines future research. We discuss an application of the S-lemma to find tight underestimators for quadratic integer optimization. Also, a subgradient method approach with the aim of making constrained problems tractable by rewriting them as unconstrained problems with a penalty term is discussed. Then, we digress to robust polynomial optimization problems for which we suggest quantifier elimination methods. Finally, we sketch how extensions of semidefinite programming translate to extensions of sos programs.

8.1. Summary

The nexus of this work consists of half-spaces, seminorm balls and ellipsoids containing the feasible set of MINLP, as well as norm bounds on the optimal solutions of MINLP. Norm bounds can also be thought of as norm balls containing the feasible set intersected with a suitable sublevel set. All four approaches are henceforth referred to as geometric objects.

Our primary motivation for working with norm and seminorm balls as well as ellipsoids is to make the integer variables of MINLP accessible to branch and bound. The primary motivation for using half-spaces is their success in linear and convex (integer) programming. A further advantage of the geometric objects is that they represent an outer approximation of the – complicated – sets F and $F_{\mathcal{I}}$ by an – easy – convex set. In other words, the approaches yield a convexification. We have also investigated how the norm and seminorm balls containing F can be shrunk using Diophantine arguments, resulting in nonlinear cuts for F. For the case of half-spaces, we have analyzed how integrality information can be used to find another half-space that potentially cuts off points in $F \setminus F_{\mathcal{I}}$, that is hence a linear cut for F.

In this work, each of the four geometric objects is to be chosen optimally (with a clear purpose in mind) out of a class of similar objects. In the case of half-spaces, we minimize the distance of the corresponding hyperplane to a known feasible solution to end up with a tight inequality. In the case of norm and seminorm balls, we fix the center and shape (by choosing a norm or seminorm) and minimize the radius – or, equivalently, the volume – to end with as few solutions in it as possible. In the case of ellipsoids, more degrees of freedom were involved since we allowed for a whole shape class and a certain region for the center, and minimized the volume. We have formulated each task as an auxiliary program for MINLP and have given conditions for the existence of feasible and optimal solutions of the auxiliary programs. These results are important from a theoretical perspective. In practice, such auxiliary programs are still difficult optimization problems themselves and can only be solved computationally under additional assumptions. Since the auxiliary problems are interesting in their own right, we have often formulated them with a deputy set $S \subset \mathbb{R}^n$ for additional generality.

For polynomial data, that is, for MIPP, we have approximated the problems as sos programs, and provided convergence results. Since sos methods are rooted in real algebra, convergence usually involves assumptions formulated in the language of that field of mathematics, the Archimedean property in our case. We have used known results to illustrate this property in terms of MIPP and have shown that sufficient conditions arise naturally in the context of optimization.

One such condition is coercivity. A whole chapter is devoted to the study of coercive polynomials in terms of their order of growth, and we have related the order of growth to the stability of the coercivity property with respect to perturbations of its coefficients. Also, we have discussed several approaches to the decision problem itself: Given a polynomial, is it coercive?

For the solution process, we have also outlined a way of finding underestimators for mixed-integer unconstrained polynomial optimization. The underestimators and norm bounds have been implemented in a branch and bound framework for unconstrained integer polynomial optimization and show a good performance.

We have used tools from various mathematical fields. To name some: Analysis for the existence (of feasible and optimal solutions) and approximation results as well as coercivity arguments; Diophantine arguments from number theory for nonlinear cuts; the Positivstellensatz for sos programming and quantifier elimination for complexity results, both from real algebraic geometry; semidefinite programming; point-set topology for boundary and compactness arguments; spectral arguments for matrices and, last but not least, geometric arguments involving volume and convexity.

We want to close this summary with the following thought. The negative results concerning the hardness of seemingly easy special cases of MINLP from the literature in Section 1.6 might at first sight deter from mixed-integer nonlinear programming. However, as indicated in Section 1.3, many encouraging positive results emerged. It is our hope that this work contributes to the feeling that MINLP can be approached with numerous tools from various areas of mathematics.

Attached to this summary is Table 8.1. It compares and contrasts the results for our four geometric objects, listed in the first column, by considering them as valid, possibly nonlinear, inequalities as outlined throughout this work. The next four columns indicate using a ✓ sign whether the result is a valid inequality for F, the relaxed feasible set, $F_{\mathcal{I}}$, the feasible set, or one of the two intersected with a sublevel set $\mathcal{L}^f_{\leq}(f(q))$, abbreviated to $\mathcal{L}_{\leq}$ in the table, where q is a feasible solution. It goes without saying that a method that computes a valid inequality for F yields a valid inequality for $F \cap \mathcal{L}^f_{\leq}(f(q))$, trivially by set containment, or by adding another constraint in the computation – but, and this is an important detail, not vice versa. Similarly for $F_{\mathcal{I}}$ and F – a valid inequality for F is yields a valid inequality for $F_{\mathcal{I}}$, but a valid inequality for $F_{\mathcal{I}}$ only yields a cut for F. The "Section" column points towards the section of the result. The column labeled "Method" reports the type of the auxiliary program or method involved. The column "Problem" states to which problem class the result contributes. All sos methods require polynomial data. "Prerequisite" lists significant assumptions, for example, a known feasible solution. Finally, the "Comment" column comments on the result. If we prove existence of feasible and optimal solutions of the auxiliary program but do not supply a computational method, it is mentioned here, and so we mention approximating hierarchies.

Approach	Valid inequality for F	$F_{\mathcal{I}}$	$F \cap \mathcal{L}_{\leq}$	$F_{\mathcal{I}} \cap \mathcal{L}_{\leq}$	Section	Method	Problem	Prerequisite	Comment
Half-space		✓			2.2	semi-infinite	MINLP	$q \in F_{\mathcal{I}}$	existence result
	✓				2.3	sos	MIPP	$q \in F_{\mathcal{I}}$	approx. hier.
		✓			2.4	gcd, rounding, MIP	MINLP	some linear constr.	
Norm bound			✓		3.1	sos, others	MIPP	$q \in F_{\mathcal{I}}$, $f_{\deg(f)} > 0$	approx. hier.
				✓	3.1	rounding	MINLP	known norm bound	
Semi-norm ball		✓			4.2	MINLP	MINLP		existence result
	✓				4.2	sos	MIPP		approx. hier.
		✓			4.3	Diophantine eq.	MINLP	known seminorm ball	
Ellip-soid		✓			5.1	semi-infinite	MINLP		existence result
	✓				5.3	sos	MIPP		approx. hier.

Table 8.1.: Comparison of the geometric approaches to MINLP. The entries are explained in Section 8.1.

8.2. Extensions

In this section we present some extensions of our work that are left for future research. Note that additionally to the ideas presented here, we have already indicated in some of the previous chapters directions of future research.

8.2.1. Quadratic integer optimization

This extension is considered with the task to find quadratic underestimators for a quadratic objective $f : \mathbb{R}^n \to \mathbb{R}$, an important special case of the underestimation problem in Section 6.2. More precisely, for the minimization problem

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in F_{\mathcal{I}} \end{aligned} \tag{IP}$$

we are interested in finding a quadratic function $g : \mathbb{R}^n \to \mathbb{R}$ that underestimates f on a sublevel set. We saw in Section 6.2 that this is a weaker requirement than global underestimation and potentially leads to stronger bounds which, in turn, are more useful in a branch and bound setting. So suppose $q \in \mathbb{R}^n$ is our incumbent solution, we are interested in g that suffice

$$f(x) \leq f(q) \Rightarrow g(x) \leq f(x) \tag{8.1}$$

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

We use the S-Lemma to derive an underestimator. The S-lemma is a theorem of the alternative for quadratic functions similar to Farkas' lemma in the linear case.

Theorem 8.1 (S-Lemma, see, e.g., Theorem 2.2 in [PT07])**.** *Let $a, b : \mathbb{R}^n \to \mathbb{R}$ quadratic functions and suppose there is $\bar{x} \in \mathbb{R}^n$ with $b(\bar{x}) < 0$. Then, the following statements are equivalent:*

1. *There is no $x \in \mathbb{R}^n$ with*

$$\begin{aligned} a(x) &< 0, \\ b(x) &\leq 0. \end{aligned}$$

2. *There is a nonnegative number $y \geq 0$ s.t.*

$$a(x) + yb(x) \geq 0 \quad \forall x \in \mathbb{R}^n.$$

We remark that characterization 1 of Theorem 8.1 is equivalent to

$$\forall x \in \mathbb{R}^n : \quad (b(x) \leq 0 \Rightarrow a(x) \geq 0)\,.$$

Then, we have the following:

Proposition 8.2. *Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be quadratic functions, $q \in \mathbb{R}^n$ which is not a continuous minimizer of f. Then underestimation of f by g as in* (8.1) *holds if and only if there is $y \geq 0$ with*

$$(1+y)f - g - yf(q) \geq 0 \quad \forall x \in \mathbb{R}^n. \tag{8.2}$$

Proof. Put

$$a := f - g,\ b := f - f(q).$$

To apply the S-Lemma (Theorem 8.1), we need $\bar{x} \in \mathbb{R}^n$ with $b(\bar{x}) = f(\bar{x}) - f(q) < 0$. Since q is not a continuous minimizer, we can choose $\bar{x} = q$. The claim follows from the S-lemma with

$$f - g + y(f - f(q)) = a(x) + yb(x)$$

□

Let us explain why (8.2) is a useful characterization. If the decision variables to choose the (yet unknown polynomial) g enter the equation linearly, the expression (8.2) is linear in the coefficients of g and in y which make g accessible to optimization approaches that require constraints that are linear in the decision variables. We can thus optimize over the class of underestimators on a sublevel set, as was the case for SLS in Section 6.2.

8.2.2. Subgradient methods

Let $f \in \mathbb{R}[X_1, \ldots, X_n]$, and suppose we have a method to solve

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in \mathbb{Z}^n \end{aligned} \tag{UIPP}$$

for example, if f has a positive definite leading form, we may use our branch and bound scheme from Section 6.3. The topic of this extension is to investigate if this extends to linearly constrained problems:

$$\begin{aligned} w^* = \min \quad & f(x) \\ \text{s.t.} \quad & Ax \leq b \\ & x \in \mathbb{Z}^n \end{aligned} \tag{LIPP}$$

where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ such that $P := \{x \in \mathbb{R}^n : Ax \leq b\}$ is a compact polyhedron - hence LIPP has an optimal solution.

General idea

We first rewrite LIPP with equality constraints, introducing slack variables:

$$\begin{aligned} w^* = \min \quad & f(x) \\ \text{s.t.} \quad & Ax + z = b \\ & x \in \mathbb{Z}^n, z \in \mathbb{R}^m_+ \end{aligned} \tag{LIPP$'$}$$

We may now relax LIPP′ as follows:

$$\begin{aligned} w(M,q) = \min \quad & f(x) + M \cdot \|Ax + z - b\|_2^{2q} \\ & x \in \mathbb{Z}^n, z \in \mathbb{R}^m_+ \end{aligned} \tag{R1}$$

with $M \in \mathbb{R}$ and $q \in \mathbb{N}$.

Observation 8.3. *Program R1 is indeed a relaxation to* LIPP′ *for all* $M \in \mathbb{R}$. *In particular,* $w(M,q) \leq w^*$ *for all* $M \in \mathbb{R}$.

Proof. The feasible set of R1 is clearly a supset of LIPP′. Denote the objective of R1 by $\tilde{f} = \tilde{f}_{M,q}$. To see $\tilde{f}(x,z) \leq f(x)$, fix M and q and let $(x,z) \in \mathbb{Z}^n \times \mathbb{R}^m_+$ feasible for LIPP′. Then $Ax + z - b = 0$, so $\tilde{f}(x,z) = f(x)$. Hence $\tilde{f}$ is a relaxation (for all M and q). The second claim is a well-known property of relaxations. □

Lemma 8.4 (Monotonicity of w)**.** *Fix* q *and let* $M \leq M'$. *Then* $w(M,q) \leq w(M',q)$.

Proof. Since $\tilde{f}_{M,q}(x,z) \leq \tilde{f}_{M',q}(x,z)$, minimization preserves the inequality and we get $w(M,q) \leq w(M',q)$. □

Lemma 8.5 (Convergence of the values $w(M)$)**.** *Fix* q. *Then for* $M \to \infty$,

$$w(M) \uparrow \bar{w} \leq w^*$$

Proof. Convergence to $\bar{w}$ follows from monotonicity and boundedness of the function $w(M)$. As $w(M,q) \leq w^*$ by Observation 8.3, $\bar{w} \leq w^*$ follows. □

Note that the polynomial objective $\tilde{f}(x) = f(x) + M \cdot \|Ax + z - b\|_2^{2q}$ has even degree if f has even degree.

Remark 8.6. It turns out that in either case, a problem is that the highest order term of $\|Ax + z - b\|_2^{2q}$ (as a polynomial in x, z) is positive semidefinite, but not positive definite: Indeed, the highest order term is given by $\|Ax + z\|_2^{2q}$ and is positive definite (that is, $\|Ax + z\|_2^{2q} > 0$ for $(x,z) \neq 0$) if and only if

$$\|Ax + z\|_2^2 = (x^T, z^T) \begin{pmatrix} A & 1 \end{pmatrix} \begin{pmatrix} x \\ z \end{pmatrix} > 0$$

for $(x,z) \neq 0$. This is not the case, even for a full rank of A, which can be seen with the choice $z = -Ax$.

One motivation is the following:

Guess 8.7. In R1, $\lim_{M \to +\infty} w(M) = w^*$ and finite convergence might hold, i.e., there is a $\overline{M} \geq 0$ with $w(M) = w^*$ for $M \geq \overline{M}$.

However, both seems not so likely: Program LIPP′ is equivalent to

$$\begin{aligned} w^* = \min \quad & f(x) \\ \text{s.t.} \quad & \|Ax + z - b\|_2^{2q} = 0 \\ & x \in \mathbb{Z}^n, z \in \mathbb{R}^m_+ \end{aligned} \tag{LIPP″}$$

and R1 is then the Lagrangian relaxation of LIPP″. The Lagrangian relaxation could then be approached with a subgradient or bundle method – assuming that we have a method to solve the subproblems for fixed M. It turns out that the maximum value of the minima $w(M)$ of the Lagrangian relaxation is not larger than the minimum value of f on $\text{conv}(P \cap \mathbb{Z}^n)$ by Fact 18 from [Lem01], as in linear integer programming. In other words, we only get a bound on the continuous minimum of f on the set $\text{conv}(P \cap \mathbb{Z}^n)$.

We leave further evaluations of the approach as future work.

8.2.3. Towards robustness

This extension is a first step towards robust mixed-integer polynomial optimization: We show how continuous relaxations of these problems can be solved with quantifier elimination. Quantifier elimination was introduced in Section 7.6.1, and we refer to [BTEGN09] for an introduction into robust optimization.

For polynomials $f, F_1, \ldots, F_k, G_1, \ldots, G_l \in \mathbb{R}[X_1, \ldots, X_n, \Xi_1, \ldots, \Xi_m]$, we consider the family of problems

$$\begin{aligned} \min \quad & f(x, \xi) \\ \text{s.t.} \quad & F_i(x, \xi) \leq 0, \ \forall i \in [k], \\ & x \in \mathbb{R}^n \end{aligned}$$

parameterized by the uncertainty parameter $\xi \in U = \{\xi \in \mathbb{R}^m : G_i(\xi) \leq 0 \ \forall i \in [l]\}$. The so-called robust counterpart (see, e.g., [BTEGN09]), or RC for short, is then

$$\begin{aligned} \min_x \quad & \sup_\xi f(x, \xi) \\ \text{s.t.} \quad & F_i(x, \xi) \leq 0 \ \forall \xi \in U, \ \forall i \in [k], \end{aligned}$$

or, equivalently

$$\begin{aligned} \min_t \quad & t \\ \text{s.t.} \quad & t - f(x, \xi) \geq 0 \ \forall \xi \in U \\ & F_i(x, \xi) \leq 0 \ \forall \xi \in U, \ \forall i \in [k], \end{aligned} \tag{RC}$$

The task is now to rephrase the above using first-order formulas. To this end define

$$\Psi_1(t, x, \xi) = \left(\bigwedge_{j=1}^{l} G_i(\xi) \leq 0 \right) \Rightarrow \left(t - f(x, \xi) \geq 0 \bigwedge_{i=1}^{k} F_i(x, \xi) \leq 0 \right)$$

If Ψ is a formula with free variables $Z_1, \ldots, Z_l$, a *realization* (see, e.g., Chapter 1.1 in [BPR05]) of Ψ is the set

$$\mathrm{Reali}(\Psi, \mathbb{R}^l) := \{y \in \mathbb{R}^l : \Psi(y)\}.$$

The feasible solutions to RC are the realizations of

$$\Psi_2(t, x) := \forall \xi \ \Psi_1(t, x, \xi).$$

We are interested in all $t \in \mathbb{R}$ such that $\Psi_2(t, x)$ holds. In other words, we want to know the realizations of

$$\Psi_3(t) = \exists x \Psi_2(t, x) = \exists x \forall \xi \Psi_1(t, x, \xi).$$

By quantifier elimination, Ψ_3 is equivalent to a quantifier free formula in one variable, thus the realization is a semi-algebraic set over $\mathbb{R}$. In other words, the realization consists of a finite union of single points and intervals!

Theorem 8.8. *Program* RC *can be solved algorithmically.*

Proof. By Theorem 7.29, quantifier elimination terminates after finitely many operations. □

Example 8.9. For parameters $\xi_1, \xi_2 \in [-2, 2]$, consider the problem

$$\begin{aligned} \min_x \quad & f(x, \xi) = \xi_1^4 + (\xi_1^2 - \xi_2)x_1^2 + \xi_1^2 x_2 \\ \text{s.t.} \quad & x_1, x_2 \in [-1, 1]. \end{aligned}$$

To solve the RC, consider the case $x_2 \geq 0$. Since the problem is then strictly convex in both ξ_1 and ξ_2, it follows that at the ξ-supremum, we necessarily have $|\xi_1| = |\xi_2| = 2$. The case $\xi_2 = 2$ can be excluded, thus the problem takes the form $\sup_\xi f(x, \xi) = 16 + (4 - (-2))x_1^2 + 4x_2 = 16 + 6x_1^2 + 4x_2$ for $x_2 \geq 0$, hence the minimum with respect to x is in this case 16.
In the other case $x_2 < 0$, it is still obvious that $\xi_2 = -2$ at the supremum. The other part of the sum is a quartic and may be rewritten as $\xi_1^2(\xi_1^2 + x_1^2 + x_2)$. The quartic is positive iff $\xi_1^2 > x_1^2 + x_2$, and since $|x_i| \leq 1$, this is clearly the case if $|\xi_1| > \sqrt{2}$. Thus, the quartic attains its maximum again at $|\xi_1| = 2$, giving $\sup_\xi f(x, \xi) = 16 + 6x_1^2 + 4x_2$, and the minimum with respect to x is attained at $x = (0, -1)$, with value $16 + 0 - 4 = 12$.

To solve the same problem with Mathematica[1]

$$\begin{aligned} \Psi_1(t, x, \xi) = & -2 \leq \xi_1 \leq 2 \ \wedge \ -2 \leq \xi_2 \leq 2 \\ & \Rightarrow -1 \leq x_1 \leq 1 \ \wedge \ -1 \leq x_2 \leq 1 \ \wedge \ t - \xi_1^4 - (\xi_1^2 - \xi_2)x_1^2 - \xi_1^2 x_2 \geq 0 \end{aligned}$$

and may apply quantifier elimination to $\Psi_2(t, x) = \forall \xi \Psi_1(t, x, \xi)$ by calling the routine `Resolve[`$\Psi(t, x, \xi)$`, Reals]` which gives the equivalent conditions

$$\Psi_2'(t, x) = -1 \leq x_1 \leq 1 \ \wedge \ -1 \leq x_2 \leq 1 \ \wedge \ t \geq 16 + 6x_1^2 + 4x_2$$

[1]We use Mathematica 9.0. Wolfram Research, Inc., Mathematica, Version 9.0, Champaign, IL (2012).

Then $\Psi_3(t) = \exists x \Psi_2(t,x)$ yields, eliminating the quantifier, the equivalent formulae

$$\Psi_3'(t) = t \geq 12 \ \vee \ t \geq 20,$$

which is of course equivalent to $\Psi_3'(t) \Leftrightarrow t \geq 12$. So Mathematica gives the same result: The robust optimal solution is 12.

Degree issues

A further question is whether in special cases – say, linear uncertainties – the degree of the quantifier eliminated problem remains stable. However, the following examples show that this is difficult.

Example 8.10. In this one dimensional example, let

$$f(x,\xi) = -\xi x - \xi^2$$

subject to $x, \xi \in [-1,1]$. The $t-f$ formulation requires $\forall \xi : \xi^2 + \xi x + t \geq 0$ and $\xi, x \in [-1,1]$, which is by QEPCAD[2] equivalent to $x^2 \leq 4t$ and $x \in [-1,1]$.

Proof. This can be seen as follows: If we forget the constraint $\xi \in [-1,1]$ for a second, the condition $\forall \xi : \xi^2 + \xi x + t \geq 0$ can only be satisfied if the discriminant $x^2 - 4t$ is non-negative, which explains the inequality $x^2 \leq 4t$. In fact, this does not change with the constraint $\xi \in [-1,1]$ in place since the quadratic (in ξ) attains its minimum at $\xi = \frac{-x}{2} \in [-1,1]$, which allows us to apply the same reasoning. □

Example 8.11. Consider spherical constrained ξ_i, $\xi_1^2 + \xi_2^2 \leq 1$ and

$$f(x,\xi) = \xi_1 x_1 + \xi_2 x_2 \text{ s.t. } x_i \in [-1,1].$$

Eliminating the ξ_i in the RC in $t - f(x,\xi) \geq 0$ formulation, QEPCAD gives the equivalent conditions

$$t^2 \geq x_1^2 + x_2^2 \text{ and } x_i \in [-1,1]$$

whose realization in $\mathbb{R}^3$ cannot be represented by finitely many linear inequalities.

8.2.4. Extensions of sos programming

We have seen in Section 5.3 that it is possible to transfer generalizations of semidefinite programming to yield a generalization of sos programming. It is the aim of this extension to point out the underlying idea.

For notational simplicity, we allow an arbitrary objective and a set constraint on the decision variables. Finally, we allow for additional ("hidden") decision variables that do not enter the objective. This allows for a simple proof of the transfer statement (Proposition 8.12). The extension reads

[2] We use QEPCADB Version B 1.69, available from https://www.usna.edu/CS/qepcadweb/B/QEPCAD.html.

$$\begin{aligned} \max \quad & f(y) \\ \text{s.t.} \quad & \sum_{j=1}^{m} y_j A_j + \sum_{j=1}^{m'} y'_j B_j \preceq C \\ & (y, y') \in \Omega \times \mathbb{R}^{m'} \end{aligned} \tag{SDP-D+}$$

for matrices $C, A_1, \ldots, A_m, B_1, \ldots, B_{m'} \in \mathcal{S}^n$, a function $f : \Omega' \to \mathbb{R}$ defined on $\Omega' \subset \mathbb{R}^m$ with $\Omega \subset \Omega'$.

Now consider the following similar extension of sos programming

$$\begin{aligned} \max \quad & f(y) \\ \text{s.t.} \quad & c_i + y_1 a_{i1} + \ldots + y_m a_{im} \in \Sigma_n, \quad i = 1, \ldots, k, \\ & y \in \Omega \end{aligned} \tag{SOSP+}$$

where $f : \Omega' \to \mathbb{R}$ is again a function with $\Omega \subset \Omega' \subset \mathbb{R}^n$ and $c_i, a_{ij} \in \Sigma$. Note that the additional decision variables in SDP-D+ do not appear.

The aim of the remainder of this section is to prove the next proposition: Any generalization of semidefinite programming that allows for hidden variables transfers to a generalization of sos programming. The proof uses the standard rewriting of sos constraints as semidefinite constraints (Theorem 1.30).

Proposition 8.12. *Let $f : \Omega' \to \mathbb{R}$ be a function with $\Omega' \subset \mathbb{R}^m$. Furthermore, let a subset Ω of Ω' as well as $k, m, n \in \mathbb{N}$ be given. Also, let c_i, $a_{ij} \in \Sigma$ for $i \in [k], j \in [m]$ be given. Then, the Program* SOSP+ *with objective f and the given data can be reformulated as a program of the form* SDP-D+ *with objective f.*

Proof. In SOSP+, consider the i-th constraint polynomial

$$p_i(y) := c_i + y_1 a_{i1} + \ldots + y_m a_{im}.$$

Let d'_i be the maximum of the degree of the polynomials $c_i, a_{i1}, \ldots, a_{im}$, and put $d_i := \lceil d'_i/2 \rceil$. Then, by Theorem 1.30, $p_i(y)$ is sos if and only if there is $Q_i \in \mathcal{S}^{\binom{n+d_i}{d_i}}$ with $p(y) = [X]_{d_i}^T Q_i [X]_{d_i}$ and $Q_i \succeq 0$. After expanding $p_i(y)$ in the basis $[X]_{d_i}$ and comparing coefficients, this is equivalent to a linear system with unknowns $(Q_i)_{\alpha\beta}$ and y_m (and the requirement $Q_i \succeq 0$). By Observation 1.29, this can be expressed as a single constraint of the form

$$\sum_{j=1}^{m} y_j A_{ij} + \sum_{j=1}^{m'_i} (y'_i)_j E_{ij} \preceq C_i, \quad (y, y'_i) \in \mathbb{R}^m \times \mathbb{R}^{m'_i}$$

for $n'_i, m'_i \in \mathbb{N}$ and matrices $A_{ij}, E_{ij}, C_i \in \mathcal{S}^{n'_i}$. Hence, by the block diagonal argument again (Observation 1.9), all k sos constraints in SOSP+ are equivalent to a constraint

of the form

$$\sum_{j=1}^{m} y_j A_j + \sum_{j=1}^{m'} y'_j E_j \preceq C, \quad (y, y') \in \mathbb{R}^m \times \mathbb{R}^{m'}$$

for $n', m' \in \mathbb{N}$ and matrices $A_j, E_j \in \mathcal{S}^{n'}$. With these preparations, SOSP+ reads

$$\begin{aligned} \max \quad & f(y) \\ \text{s.t.} \quad & \sum_{j=1}^{m} y_j A_j + \sum_{j=1}^{m'} y'_j E_j \preceq C \\ & \sum_{j=1}^{m} y_j B_j \prec D \\ & y \in \Omega \\ & (y, y') \in \mathbb{R}^m \times \mathbb{R}^{m'} \end{aligned}$$

which is of the form SDP-D+. □

From the applied point of view, the following immediate consequence is of interest.

Corollary 8.13. *An algorithm that can solve* SDP-D+ *yields an algorithm that can solve* SOSP+*.*

We expect that these extensions might prove useful in practice, and leave it as interesting future research.

A. Sublevel sets and tight inequalities

Suppose we are given a point $q \in S$, where S is typically one of the sets F or $F_{\mathcal{I}}$. A natural question in our setting is how close a valid inequality for S can get to q. In the best case, the inequality is tight at q, and a necessary condition for being tight at q is that q lies on the boundary of S. On the other hand, it is well-known that the set of optimal solutions is unchanged by intersecting $F_{\mathcal{I}}$ with $\mathcal{L}^f_{\leq}(f(q))$, or equivalently, by adding the constraint $f(x) \leq f(q)$. In applications, this is sensible if it is known that feasible solutions with better objective value that $f(q)$ exist. Alternatively, this is useful if one wants to verify (in a pruning substep or similar) that no feasible solutions with better objective value exist.

In this section we show that, under mild assumptions, q is indeed on the boundary of S intersected with $\mathcal{L}^f_{\leq}(f(q))$. As this is a topological matter, we need a simple fact from topology. The statement in the following lemma relates the boundaries of a space and the boundary of a subspace. This fact seems to be mathematical folklore, we only give a proof for completeness. For a topological space (X, τ) and $Y \subset X$ endowed with the subspace topology, we denote the boundary with respect to X and Y by bd_X and bd_Y, respectively.

Lemma A.1. *Let (X, τ) be a topological space.*

1. *Let $M \subset X$. Then $z \in \mathrm{bd}\, M$ if and only if all neighborhoods U of z contain points of M and $X \setminus M$.*

2. *Let $M \subset Y \subset X$. Then $\mathrm{bd}_Y(M) \subset \mathrm{bd}_X(M) \cap Y$.*

Proof. For a proof of 1, see Definition 2 and Theorem 2 in Chapter 1.2 of [Gaa09]. To see the other assertion, let $x \in \mathrm{bd}_Y(M)$ and V be a neighborhood of x in X. To show $x \in \mathrm{bd}_X(M)$ with assertion 1, we show V intersects M and $X \setminus M$. Now $U := V \cap Y$ is a neighborhood of x in Y. As $x \in \mathrm{bd}_Y(M)$, $U \cap M \neq \emptyset$, and $V \cap M \neq \emptyset$ follows from $U \subset V$. Also, $\emptyset \neq U \cap (Y \setminus M) \subset V \cap (Y \setminus M) \subset V \cap (X \setminus M)$, so V intersects $X \setminus M$. This proves $\mathrm{bd}_Y(M) \subset \mathrm{bd}_X(M)$. As $\mathrm{bd}_Y(M) \subset Y$, the claim follows. □

In the following lemma, we denote the boundary with respect to the subspace $S \subset \mathbb{R}^n$ by bd_S.

Lemma A.2. *Let $S \subset \mathbb{R}^n$, $q \in S$ and $f : S \to \mathbb{R}$ (not necessarily continuous). Then, the following are equivalent:*

1. $q \in \mathrm{bd}_S \left(\mathcal{L}^f_{\leq}(f(q)) \right)$,

2. q is not a local maximizer of f.

Proof. By Lemma A.1 (1), $q \in \mathrm{bd}_S\left(\mathcal{L}^f_{\leq}(f(q)\right)$ if and only if all neighborhoods N of q contain points $x \in S$ with $f(x) > f(q)$ and points $y \in S$ with $f(y) \leq f(q)$. The second requirement is trivial as every neighborhood N of the point q contains q. Thus $q \in \mathrm{bd}_S\left(\mathcal{L}^f_{\leq}(f(q)\right)$ if and only if every neighborhood N of q contains points $x \in S$ with $f(x) > f(q)$. Equivalently, q is not a local maximizer. □

As the boundary with respect to S can be difficult to compute, we may reduce to the boundary with respect to $\mathbb{R}^n$ if we drop the equivalence in the statement.

Proposition A.3. *Let $q \in S \subset \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$ be a function and suppose q is not a local maximizer of $f|_S$. Then $q \in \mathrm{bd}\left(S \cap \mathcal{L}^f_{\leq}(f(q))\right)$, where* bd *is the boundary with respect to $\mathbb{R}^n$.*

Proof. Note first that $q \in \mathrm{bd}\left(S \cap \mathcal{L}^f_{\leq}(f(q))\right)$ if and only if for every neighborhood N of q, N intersects the set $S \cap \mathcal{L}^f_{\leq}(f(q))$ and the complement $\mathbb{R}^n \setminus \left(S \cap \mathcal{L}^f_{\leq}(f(q))\right)$. As q is in each of the sets N, S and $\mathcal{L}^f_{\leq}(f(q))$, the first intersection requirement always holds. Now let q not be a local maximizer of f restricted to S. Equivalently, for every neighborhood N of q, there is $x \in S \cap N$ with $f(x) > f(q)$. But this x lies in the complement of $S \cap \mathcal{L}^f_{\leq}(f(q))$. In other words, the second intersection requirement is fulfilled and hence $q \in \mathrm{bd}\left(S \cap \mathcal{L}^f_{\leq}(f(q))\right)$ follows. □

In a convex setting, intersecting with a sublevel set guarantees a supporting hyperplane to q.

Corollary A.4. *Suppose $S \subset \mathbb{R}^n$ is a convex set, $q \in S$ and $f : S \to \mathbb{R}$ is a quasiconvex function. Suppose q is not a local maximum of f. Then there is a valid inequality for $S \cap \mathcal{L}^f_{\leq}(f(q))$ which is tight at q.*

Proof. Note that $S \cap \mathcal{L}^f_{\leq}(f(q))$ is convex as an intersection of two convex sets. By Proposition A.3, q is on the boundary of $S \cap \mathcal{L}^f_{\leq}(f(q))$. From Theorem 1.14, there is a supporting hyperplane $H = H(a, b)$, $a \in \mathbb{R}^n \setminus \{0\}$, $b \in \mathbb{R}$ to $S \cap \mathcal{L}^f_{\leq}(f(q))$, in other words, a valid inequality for $S \cap \mathcal{L}^f_{\leq}(f(q))$ which is tight at q. □

B. Proofs

We prepare the proof of Lemma 2.2 using two intermediate steps. In the first step we show that gauges on $\mathbb{R}^n$ behave well towards norms.

Lemma B.1. *Let γ be a gauge and $\|\cdot\|$ be a norm on $\mathbb{R}^n$. Then, γ is left-equivalent to $\|\cdot\|$, i.e., there is $\mu > 0$ such that*

$$\gamma(x) \leq \mu \|x\|_2.$$

Furthermore, γ is continuous.

Proof. As 0 is in the interior of the defining set A, there is $\mu' > 0$ such that $\mu \mathbb{B}_1(0; \|\cdot\|) \subset \mathbb{B}_1(0; \gamma)$. Hence $\gamma(x) \leq \frac{1}{\mu'}\|x\|$ for all $x \in \mathbb{R}^n$. The first claim follows with $\mu := 1/\mu'$. To see continuity, fix $\epsilon > 0$ and $x \in \mathbb{R}^n$. Subadditivity implies the inequality $\gamma(x) - \gamma(y) \leq \gamma(x - y)$ for all $y \in \mathbb{R}^n$. By the first part, $\gamma(x - y) \leq \frac{1}{\mu'}\|x - y\|$ for all $x, y \in \mathbb{R}^n$. For all $y \in \mathbb{R}^n$ with $\|x - y\| \leq \varepsilon\mu'$, $\gamma(x) - \gamma(y) \leq \epsilon$. Interchanging x and y and using the fact that norms are absolutely homogeneous we get $|\gamma(x) - \gamma(y)| \leq \epsilon$, and continuity at x follows. □

It can also be shown that γ is right-equivalent to a given norm, but we do not need this property.

We can show now that the distance measured by a gauge towards a fixed set is continuous.

Lemma B.2. *Let γ be a gauge on $\mathbb{R}^n$ and $A \subset \mathbb{R}^n$ some nonempty set. Then, the map $\mathbb{R}^n \to \mathbb{R}$, $x \mapsto d(x, A)$ is continuous.*

Proof. Let $\varepsilon > 0$ and $x \in \mathbb{R}^n$. Let μ as in Lemma B.1 and $y \in \mathbb{R}^n$ with $\|x - y\| \leq \varepsilon/\mu$. There is $a \in A$ with $\gamma(a - x) \leq d(x, A) + \varepsilon$. Now

$$\begin{aligned} d(y, A) \leq \gamma(a - y) &= \gamma(a - x + x - y) \leq \gamma(a - x) + \gamma(x - y) \leq d(x, A) + \epsilon + \gamma(x - y) \\ &= d(x, A) + \varepsilon + \mu\|x - y\| \leq d(x, A) + 2\varepsilon \end{aligned}$$

for all $y \in \mathbb{R}^n$ with $\|x - y\| \leq \varepsilon/\mu$. By estimating $d(x, A)$ analogously, we find

$$d(x, A) \leq d(y, A) + 2\varepsilon,$$

which combines to $|d(x, A) - d(y, A)| \leq 2\varepsilon$ for all $y \in \mathbb{R}^n$ with $\|x - y\| \leq \varepsilon/\mu$. Hence $d(\cdot, A)$ is continuous at x. □

We can now give the proof.

Proof of Lemma 2.2. There is a sequence $(a_n, k_n) \in A \times K$ with $\gamma(k_n - a_n) = d(a_n, k_n) \to d(A, K)$. By compactness, a subsequence $k_{n;p}$ converges to some $k^* \in K$ for $p \to \infty$. Assume first that A is compact, too. Hence, a subsequence $a_{n;p;q}$ converges to some $a^* \in A$ for $q \to \infty$, which results by continuity of γ (Lemma B.1) in

$$d(a^*, k^*) = \gamma(k^* - a^*) = \gamma(\lim_{q\to\infty}(k_{n;p;q} - a_{n;p;q})) = \lim_{q\to\infty} \gamma(k_{n;p;q} - a_{n;p;q}) = d(A, K).$$

Now let us drop the additional assumption that A is compact. Let $\varepsilon > 0$. To reduce this to the compact case, it is enough to show that any pair of points $(a', k') \in A \times K$ with $d(a', k') \le d(A, K) + \varepsilon$, that is, any potential distance minimizer, lies in the compact set $A^* \times K$, where A^* is the set $A^* = A \cap \mathbb{B}^\gamma_R(k^*)$, $\mathbb{B}^\gamma_r(p)$ is the closed gauge ball from (2.2), and R is the number $R := \operatorname{diam}(K) + d(A, K) + \varepsilon$ and $\operatorname{diam}(K)$ is the diameter of the set K, $\operatorname{diam}(K) := \sup_{x,y\in K} \gamma(y - x)$. Let us verify all implicit statements. First note that $a' \in A^*$: Since

$$d(a', k^*) = \gamma(k^* - a') = \gamma(k^* - k' + k' - a') \le \gamma(k^* - k') + \gamma(k' - a') \le \operatorname{diam}(K) + d(A, K) + \varepsilon,$$

it follows that $a' \in \mathbb{B}^\gamma_R(k^*)$. Also note that R is finite: As γ is continuous, the map $K \times K \to \mathbb{R}$, $(x, y) \mapsto \gamma(y - x)$ attains its supremum by compactness of $K \times K$. Finally, we observe that the set $A' := A \cap \mathbb{B}^\gamma_R(k)$ is indeed compact as the balls defined via γ are closed and bounded. □

Summary of contributions

In this chapter we outline which parts of this work are joint work with coauthors and which parts have been published.

Chapter 1. The literature review as well as some of the material in Sections 1.4, 1.5 and 1.6 have been published in the article [BHS17], an article by the author of this work, Anita Schöbel and Ruth Hübner, as well as in the preprint [BB17] by the author of this work and Tomáš Bajbar.

Chapters 2, 4 and 5. These chapters are new and make use of discussions with Anita Schöbel. The proofs are the author's work.

Chapters 3 and 6. Most of the results in these chapters have already been published in the article [BHS17], which is joint work with Anita Schöbel and Ruth Hübner. Theorems 3.3, 3.4, 6.5, 6.6, Proposition 3.5, 3.7, 6.2, 6.4 and Observation 6.1 are stated in a more general form than in the reference. Section 3.2.1 is new. The experiments were conducted by the author.

Chapter 7. The material from the beginning of the chapter until Section 7.5 inclusive is joint work with Tomáš Bajbar, available as the preprint [BB17]. Both authors contributed equally to the writing of this preprint, and both authors are grateful to Lukas Katthän for fruitful discussions on its subject. Section 7.6 is new and the author's sole work.

Section 8.2. The author of this work had the ideas for the material in Sections 8.2.1, 8.2.3 and 8.2.4. Section 8.2.2 makes use of discussions with Christoph Buchheim and Anita Schöbel.

Bibliography

[AL12] M. Anjos and J. B. Lasserre. *Handbook on semidefinite, conic and polynomial optimization*. Springer, 2012.

[AOPT13] A. A. Ahmadi, A. Olshevsky, P. A. Parrilo, and J. N. Tsitsiklis. "NP-hardness of deciding convexity of quartic polynomials and related problems". In: *Mathematical Programming* (2013), pp. 1–24.

[AWW11] G. Averkov, C. Wagner, and R. Weismantel. "Maximal lattice-free polyhedra: finiteness and an explicit description in dimension three". In: *Mathematics of Operations Research* 36.4 (2011), pp. 721–742.

[BA07] C. Bivià-Ausina. "Injectivity of real polynomial maps and Łojasiewicz exponents at infinity". In: *Mathematische Zeitschrift* 257.4 (2007), pp. 745–767.

[BB17] T. Bajbar and S. Behrends. *How fast do coercive polynomials grow?* Tech. rep. Preprint-Reihe, Institut für Numerische und Angewandte Mathematik, Georg-August Universität Göttingen, 2017.

[BCCZ10] A. Basu, M. Conforti, G. Cornuéjols, and G. Zambelli. "Maximal lattice-free convex sets in linear subspaces". In: *Mathematics of Operations Research* 35.3 (2010), pp. 704–720.

[BCL12] C. Buchheim, A. Caprara, and A. Lodi. "An effective branch-and-bound algorithm for convex quadratic integer programming". In: *Mathematical programming* (2012), pp. 1–27.

[BD14] C. Buchheim and C. D'Ambrosio. "Box-Constrained Mixed-Integer Polynomial Optimization Using Separable Underestimators". In: *Integer Programming and Combinatorial Optimization*. Springer, 2014, pp. 198–209.

[BF76] J. W. Blankenship and J. E. Falk. "Infinitely constrained optimization problems". In: *Journal of Optimization Theory and Applications* 19.2 (1976), pp. 261–281.

[BGR10] R. E. Bixby, Z. Gu, and E. Rothberg. *Presolve reductions in mixed integer programming*. `https://symposia.cirrelt.ca/system/documents/000/000/111/Bixby_original.pdf?1441306917`. Talk given at the Spring School on Combinatorial Optimization in Logistics, Université Montreal, May 19th 2010, Montreal, Canda [Accessed: 2017-06-27]. 2010.

[BH02] E. Boros and P. L. Hammer. "Pseudo-boolean optimization". In: *Discrete applied mathematics* 123.1 (2002), pp. 155–225.

[BHS15] C. Buchheim, R. Hübner, and A. Schöbel. "Ellipsoid bounds for convex quadratic integer programming". In: *SIAM Journal on Optimization* 25.2 (2015), pp. 741–769.

[BHS17] S. Behrends, R. Hübner, and A. Schöbel. "Norm Bounds and Underestimators for Unconstrained Polynomial Integer Minimization". In: *Mathematical Methods of Operations Research* (2017). online first, pp. 1–35. ISSN: 1432-5217. DOI: 10.1007/s00186-017-0608-y. URL: https://doi.org/10.1007/s00186-017-0608-y.

[BKL+13] P. Belotti, C. Kirches, S. Leyffer, J. Linderoth, J. Luedtke, and A. Mahajan. "Mixed-integer nonlinear optimization". In: *Acta Numerica* 22 (2013), pp. 1–131.

[Bor99] B. Borchers. "CSDP, A C library for semidefinite programming". In: *Optimization methods and Software* 11.1-4 (1999), pp. 613–623.

[BP03] M. R. Bussieck and A. Pruessner. "Mixed-integer nonlinear programming". In: *SIAG/OPT Newsletter: Views & News* 14.1 (2003), pp. 19–22.

[BPR05] S. Basu, R. Pollack, and M.-F. Roy. *Algorithms in real algebraic geometry*. Vol. 20033. Springer, 2005.

[BPT13] G. Blekherman, P. A. Parrilo, and R. R. Thomas. *Semidefinite optimization and convex algebraic geometry*. Vol. 13. Siam, 2013.

[BR07] C. Buchheim and G. Rinaldi. "Efficient reduction of polynomial zero-one optimization to the quadratic case". In: *SIAM Journal on Optimization* 18.4 (2007), pp. 1398–1413.

[BS15a] T. Bajbar and O. Stein. "Coercive Polynomials and Their Newton Polytopes". In: *SIAM Journal on Optimization* 25.3 (2015), pp. 1542–1570. DOI: 10.1137/140980624.

[BS15b] T. Bajbar and O. Stein. "Coercive polynomials: Stability, order of growth, and Newton polytopes". In: *Optimization Online, Preprint ID 2015-10-5158, 2015* (2015). DOI: 10.1137/140980624.

[BS17] T. Bajbar and O. Stein. "On Globally Diffeomorphic Polynomial Maps via Newton Polytopes and Circuit Numbers". In: *Mathematische Zeitschrift* (2017).

[BTEGN09] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton University Press, 2009.

[BV04] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[BV11] S. Boyd and L. Vandenberghe. *Localization and cutting-plane methods*. Stanford EE 364b lecture notes. 2011.

[CCZ10] M. Conforti, G. Cornuéjols, and G. Zambelli. "Polyhedral approaches to mixed integer linear programming". In: *50 years of integer programming 1958-2008* (2010), pp. 343–385.

[CDTT14] Y. Chen, L. R. G. Dias, K. Takeuchi, and M. Tibar. "Invertible polynomial mappings via Newton non-degeneracy, to appear in Ann". In: *Annales de l'Institut Fourier* 54.5 (2014), pp. 1807–1822.

[CEG04] G. Calafiore and L. El Ghaoui. "Ellipsoidal bounds for uncertain linear equations and dynamical systems". In: *Automatica* 40.5 (2004), pp. 773–787.

[CHKM92] W. Cook, M. Hartmann, R. Kannan, and C. McDiarmid. "On integer points in polyhedra". In: *Combinatorica* 12.1 (1992), pp. 27–37.

[CL01] G. Cornuéjols and Y. Li. "Elementary closures for integer programs". In: *Operations Research Letters* 28.1 (2001), pp. 1–8.

[CLO07] D. A. Cox, J. Little, and D. O'Shea. *Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra*. Springer, 2007.

[DFJ54] G. Dantzig, R. Fulkerson, and S. Johnson. "Solution of a large-scale traveling-salesman problem". In: *Journal of the operations research society of America* 2.4 (1954), pp. 393–410.

[DG86] M. A. Duran and I. E. Grossmann. "An outer-approximation algorithm for a class of mixed-integer nonlinear programs". In: *Mathematical programming* 36.3 (1986), pp. 307–339.

[DGL10] S. Dash, O. Günlük, and A. Lodi. "MIR closures of polyhedral sets". In: *Mathematical Programming* 121.1 (2010), pp. 33–60.

[DPHWZ16] A. Del Pia, R. Hildebrand, R. Weismantel, and K. Zemmer. "Minimizing cubic and homogeneous polynomials over integers in the plane". In: *Mathematics of Operations Research* 41.2 (2016), pp. 511–530.

[ED08] M. S. El Din. "Computing the global optimum of a multivariate polynomial over the reals". In: *Proceedings of the twenty-first international symposium on Symbolic and algebraic computation*. ACM. 2008, pp. 71–78.

[EGC99] L. El Ghaoui and G. Calafiore. "Confidence ellipsoids for uncertain linear equations with structure". In: *Decision and Control, 1999. Proceedings of the 38th IEEE Conference on*. Vol. 2. IEEE. 1999, pp. 1922–1927.

[Eis00] F. Eisenbrand. "Gomory-Chvátal cutting planes and the elementary closure of polyhedra". PhD thesis. Universität des Saarlandes, 2000.

[FHB03] M. Fazel, H. Hindi, and S. P. Boyd. "Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices". In: *American Control Conference, 2003. Proceedings of the 2003*. Vol. 3. IEEE. 2003, pp. 2156–2162.

[FL07] M. Fischetti and A. Lodi. "Optimizing over the first Chvátal closure". In: *Mathematical Programming* 110.1 (2007), pp. 3–20.

[FL94] R. Fletcher and S. Leyffer. "Solving mixed integer nonlinear programs by outer approximation". In: *Mathematical programming* 66.1 (1994), pp. 327–349.

[For60] R. Fortet. "L'algebre de boole et ses applications en recherche opérationnelle". In: *Trabajos de Estadistica y de Investigación Operativa* 11.2 (1960), pp. 111–118.

[Gaa09] S. A. Gaal. *Point set topology*. Dover Publications, 2009.

[Geo70] A. M. Geoffrion. "Elements of large-scale mathematical programming Part I: Concepts". In: *Management Science* 16.11 (1970), pp. 652–675.

[GJ79] M. R. Garey and D. S. Johnson. *Computers and intractability*. W.H. Freeman and Company, 1979.

[GLS93] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric algorithms and combinatorial optimization*. Vol. 2. Springer Science & Business Media, 1993.

[Gol17] M. Goldberg. "Continuity of seminorms on finite-dimensional vector spaces". In: *Linear Algebra and its Applications* 515 (2017), pp. 175–179.

[Gom58] R. E. Gomory. "Outline of an Algorithm for Integer Solutions to Linear Programs". In: *Bull. Am. Math. Soc* 64 (1958), p. 3.

[Gor61] E. A. Gorin. "Asymptotic properties of polynomials and algebraic functions of several variables". In: *Russian mathematical surveys* 16.1 (1961), pp. 93–119.

[GR85] O. Gupta and A. Ravindran. "Branch and bound experiments in convex nonlinear integer programming". In: *Management Science* 31.12 (1985), pp. 1533–1546.

[Gru07] P. Gruber. *Convex and discrete geometry*. Vol. 336. Springer Science & Business Media, 2007.

[GSED11] A. Greuet and M. Safey El Din. "Deciding reachability of the infimum of a multivariate polynomial". In: *Proceedings of the 36th international symposium on Symbolic and algebraic computation*. ACM. 2011, pp. 131–138.

[GSED14] A. Greuet and M. Safey El Din. "Probabilistic algorithm for polynomial optimization over a real algebraic set". In: *SIAM Journal on Optimization* 24.3 (2014), pp. 1313–1343.

[Han96] J. Hannah. "A geometric approach to determinants". In: *American Mathematical Monthly* (1996), pp. 401–409.

[Hei05] S. Heinz. "Complexity of integer quasiconvex polynomial optimization". In: *Journal of Complexity* 21.4 (2005), pp. 543–556.

[Hel00] C. Helmberg. “Semidefinite programming for combinatorial optimization”. Habilitation. 2000.

[Hil09] D. Hilbert. “Beweis für die Darstellbarkeit der ganzen Zahlen durch eine feste Anzahl n^{ter} Potenzen (Waringsches Problem)”. In: *Mathematische Annalen* 67.3 (1909), pp. 281–300.

[HJ12] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge university press, 2012.

[HK13] R. Hildebrand and M. Köppe. “A new Lenstra-type algorithm for quasiconvex polynomial integer minimization with complexity $2^{O(n \log n)}$”. In: *Discrete Optimization* 10.1 (2013), pp. 69–84.

[HK93] R. Hettich and K. O. Kortanek. “Semi-infinite programming: theory, methods, and applications”. In: *SIAM review* 35.3 (1993), pp. 380–429.

[HKLW10] R. Hemmecke, M. Köppe, J. Lee, and R. Weismantel. “Nonlinear integer programming”. In: *50 Years of Integer Programming 1958-2008*. Springer, 2010, pp. 561–618.

[HL05] D. Henrion and J. B. Lasserre. “Detecting global optimality and extracting solutions in GloptiPoly”. In: *Positive polynomials in control*. Springer, 2005, pp. 293–310.

[Hol75] R. B. Holmes. *Geometric functional analysis and its applications*. Vol. 24. Springer Science & Business Media, 1975.

[HS14] R. Hübner and A. Schöbel. “When is rounding allowed in integer nonlinear optimization?” In: *European Journal of Operational Research* 237.2 (2014), pp. 404–410.

[HWHBS08] G. H. Hardy, E. M. Wright, D. R. Heath-Brown, and J. H. Silverman. *An introduction to the theory of numbers*. 6th. Oxford University Press, 2008.

[HWZ16] R. Hildebrand, R. Weismantel, and K. Zemmer. “An FPTAS for minimizing indefinite quadratic forms over integers in polyhedra”. In: *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics. 2016, pp. 1715–1723.

[Jer73] R. C. Jeroslow. “There cannot be any algorithm for integer programming with quadratic constraints”. In: *Operations Research* 21.1 (1973), pp. 221–224.

[JLL14] V. Jeyakumar, J. B. Lasserre, and G. Li. “On polynomial optimization over non-compact semi-algebraic sets”. In: *Journal of Optimization Theory and Applications* 163.3 (2014), pp. 707–718.

[JLN+10] M. Jünger, T. Liebling, D. Naddef, G. L. Nemhauser, W. R. Pulleyblank, G. Reinelt, G. Rinaldi, and L. A. Wolsey. *50 Years of Integer Programming 1958-2008: From the Early Years to the State-of-the-art*. Springer, 2010.

[JPL14] V. Jeyakumar, T. S. Pham, and G. Li. "Convergence of the Lasserre hierarchy of SDP relaxations for convex polynomial programs without compactness". In: *Operations Research Letters* 42.1 (2014), pp. 34–40.

[Kar72] R. M. Karp. "Reducibility among combinatorial problems". In: *Complexity of computer computations*. Springer, 1972, pp. 85–103.

[Kel60] J. E. Kelley Jr. "The cutting-plane method for solving convex programs". In: *Journal of the society for Industrial and Applied Mathematics* 8.4 (1960), pp. 703–712.

[Kha83] L. G. Khachiyan. "Convexity and complexity in polynomial programming". In: *Proceedings of the International Congress of Mathematicians*. 1983, pp. 1569–1577.

[KP00] L. Khachiyan and L. Porkolab. "Integer optimization on convex semialgebraic sets". In: *Discrete & Computational Geometry* 23.2 (2000), pp. 207–224.

[Kra07] T. Krasiński. "On the Łojasiewicz exponent at infinity of polynomial mappings". In: *Acta Math. Vietnam* 32 (2007), pp. 2–3.

[Köp12] M. Köppe. "On the complexity of nonlinear mixed-integer optimization". In: *Mixed-Integer Nonlinear Programming, in: IMA Volumes in Mathematics and its Applications* 154 (2012), pp. 533–558.

[Lan13] K. Lange. *Optimization*. 2nd. Springer Verlag New York, 2013.

[Lan93] S. Lang. "Real and functional analysis, volume 142 of Graduate Texts in Mathematics". In: *Springer-Verlag, New York,* 10 (1993), pp. 11–13.

[Las01] J. B. Lasserre. "Global optimization with polynomials and the problem of moments". In: *SIAM Journal on Optimization* 11.3 (2001), pp. 796–817.

[Lau09] M. Laurent. "Sums of squares, moment matrices and optimization over polynomials". In: *Emerging applications of algebraic geometry*. Springer, 2009, pp. 157–270.

[Lem01] C. Lemaréchal. "Lagrangian relaxation". In: *Computational combinatorial optimization* (2001), pp. 112–156.

[LHKW06] J. A. D. Loera, R. Hemmecke, M. Köppe, and R. Weismantel. "Integer Polynomial Optimization in Fixed Dimension". In: *Mathematics of Operations Research* 31.1 (2006), pp. 147–153.

[LJ83] H. W. Lenstra Jr. "Integer programming with a fixed number of variables". In: *Mathematics of operations research* 8.4 (1983), pp. 538–548.

[LL12] J. Lee and S. Leyffer. *Mixed integer nonlinear programming*. Springer, 2012.

[LS00] Z.-Q. Luo and J. Sun. "A polynomial cutting surfaces algorithm for the convex feasibility problem defined by self-concordant inequalities". In: *Computational Optimization and Applications* 15.2 (2000), pp. 167–191.

[LS07] M. López and G. Still. "Semi-infinite programming". In: *European Journal of Operational Research* 180.2 (2007), pp. 491–518.

[LT11] J. B. Lasserre and T. P. Thanh. "Convex underestimators of polynomials". In: *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*. IEEE. 2011, pp. 7194–7199.

[Mar03] M. Marshall. "Optimization of polynomial functions". In: *Canadian Mathematical Bulletin* 46 (2003), pp. 575–587.

[Mar08] M. Marshall. *Positive polynomials and sums of squares*. Mathematical Surveys and Monographs 146. American Mathematical Soc., 2008.

[Mat70] Y. V. Matiyasevich. "Enumerable sets are diophantine". In: *Doklady Akademii Nauk SSSR* 191.2 (1970), pp. 279–282.

[MB09] A. Mutapcic and S. Boyd. "Cutting-set methods for robust convex optimization with pessimizing oracles". In: *Optimization Methods & Software* 24.3 (2009), pp. 381–406.

[Mey74] R. R. Meyer. "On the existence of optimal solutions to integer and mixed-integer programming problems". In: *Mathematical Programming* 7.1 (1974), pp. 223–235.

[MMWW02] H. Marchand, A. Martin, R. Weismantel, and L. Wolsey. "Cutting planes in integer and mixed integer programming". In: *Discrete Applied Mathematics* 123.1 (2002), pp. 397–446.

[MN14] P. Mondal and T. Netzer. "How fast do polynomials grow on semialgebraic sets?" In: *Journal of Algebra* 413 (2014), pp. 320–344.

[Mot67] T. S. Motzkin. "The arithmetic-geometric inequality". In: *Inequalities (Proc. Sympos. Wright-Patterson Air Force Base, Ohio, 1965)* (1967), pp. 205–224.

[ND05] J. Nie and J. W. Demmel. "Minimum ellipsoid bounds for solutions of polynomial systems via sum of squares". In: *Journal of Global Optimization* 33.4 (2005), pp. 511–525.

[NDS06] J. Nie, J. Demmel, and B. Sturmfels. "Minimizing polynomials via sum of squares over the gradient ideal". In: *Mathematical programming* 106.3 (2006), pp. 587–606.

[Nes00] Y. Nesterov. "Squared functional systems and optimization problems". In: *High performance optimization*. Springer, 2000, pp. 405–440.

[Nie12] J. Nie. "Sum of squares methods for minimizing polynomial forms over spheres and hypersurfaces". In: *Frontiers of mathematics in china* 7.2 (2012), pp. 321–346.

[NS07] J. Nie and M. Schweighofer. "On the complexity of Putinar's Positivstellensatz". In: *Journal of Complexity* 23.1 (2007), pp. 135–150.

[NW88] G. L. Nemhauser and L. A. Wolsey. *Integer and combinatorial optimization*. Wiley-Interscience, 1988.

[Par00] P. A. Parrilo. "Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization". PhD thesis. Citeseer, 2000.

[PAV+] A. Papachristodoulou, J. Anderson, G. Valmorbida, S. Prajna, P. Seiler, and P. A. Parrilo. *SOSTOOLS: Sum of squares optimization toolbox for MATLAB*. Available from `http://www.cds.caltech.edu/sostools`.

[PC01] F Plastria and E Carrizosa. "Gauge distances and median hyperplanes". In: *Journal of Optimization Theory and Applications* 110.1 (2001), pp. 173–182.

[PPP02] S. Prajna, A. Papachristodoulou, and P. A. Parrilo. "Introducing SOSTOOLS: A general purpose sum of squares programming solver". In: *Decision and Control, 2002, Proceedings of the 41st IEEE Conference on*. Vol. 1. IEEE. 2002, pp. 741–746.

[PS03] P. A. Parrilo and B. Sturmfels. "Minimizing polynomial functions". In: *Algorithmic and quantitative real algebraic geometry, DIMACS Series in Discrete Mathematics and Theoretical Computer Science* 60 (2003), pp. 83–99.

[PT07] I. Pólik and T. Terlaky. "A survey of the S-lemma". In: *SIAM review* 49.3 (2007), pp. 371–418.

[RG98] R. Reemtsen and S. Görner. "Numerical methods for semi-infinite programming: a survey". In: *Semi-infinite programming*. Springer, 1998, pp. 195–275.

[Roc70] R. T. Rockafellar. *Convex analysis*. Princeton university press, 1970.

[Ros75] I. G. Rosenberg. "Reduction of bivalent maximization to the quadratic case". In: *Cahiers du Centre d'etudes de Recherche Operationnelle* 17 (1975), pp. 71–74.

[RR98] R. Reemtsen and J.-J. Rückmann. *Semi-infinite programming*. Vol. 25. Springer Science & Business Media, 1998.

[Rud91] W. Rudin. *Functional analysis. International series in pure and applied mathematics*. McGraw-Hill, Inc., New York, 1991.

[Sch05] M. Schweighofer. "Optimization of polynomials on compact semialgebraic sets". In: *SIAM Journal on Optimization* 15.3 (2005), pp. 805–825.

[Sch06] M. Schweighofer. “Global optimization of polynomials using gradient tentacles and sums of squares”. In: *SIAM Journal on Optimization* 17.3 (2006), pp. 920–942.

[Sch99] A. Schöbel. *Locating lines and hyperplanes: theory and algorithms*. Vol. 25. Springer Science & Business Media, 1999.

[Sho87] N. Z. Shor. “Class of global minimum bounds of polynomial functions”. In: *Cybernetics and Systems Analysis* 23.6 (1987), pp. 731–734.

[Sma98] N. P. Smart. *The algorithmic resolution of Diophantine equations: a computational cookbook*. Vol. 41. Cambridge University Press, 1998.

[SS97] N. Z. Shor and P. I. Stetsyuk. “Modified r-algorithm to find the global minimum of polynomial functions”. In: *Cybernetics and Systems Analysis* 33.4 (1997), pp. 482–497.

[Ste12] O. Stein. “How to solve a semi-infinite optimization problem”. In: *European Journal of Operational Research* 223.2 (2012), pp. 312–320.

[Ste13] O. Stein. *Bi-level strategies in semi-infinite programming*. Vol. 71. Springer Science & Business Media, 2013.

[Tho73] G. B. Thomas. “Density properties of primes, squares, and sums of squares”. In: *Advances in Mathematics* 10.3 (1973), pp. 383–386.

[Tod12] M. J. Todd. *Lecture notes on semidefinite programming*. `https://people.orie.cornell.edu/miketodd/orie6327/orie6327.html`. [Accessed: 2017-08-24]. 2012.

[TS02] M. Tawarmalani and N. V. Sahinidis. *Convexification and global optimization in continuous and mixed-integer nonlinear programming: theory, algorithms, software, and applications*. Vol. 65. Springer Science & Business Media, 2002.

[TTT99] K.-C. Toh, M. J. Todd, and R. H. Tütüncü. “SDPT3—a MATLAB software package for semidefinite programming, version 1.3”. In: *Optimization methods and software* 11.1-4 (1999), pp. 545–581.

[VB96] L. Vandenberghe and S. Boyd. “Semidefinite programming”. In: *SIAM review* 38.1 (1996), pp. 49–95.

[VP07] H. H. Vui and T. S. Pham. “Minimizing polynomial functions”. In: *Acta Mathematica Vietnamica* 32.1 (2007), pp. 71–82.

[VP10] H. H. Vui and T. S. Pham. “Representations of positive polynomials and optimization on noncompact semialgebraic sets”. In: *SIAM Journal on Optimization* 20.6 (2010), pp. 3082–3103.

[VRSS08] F. G. Vázquez, J.-J. Rückmann, O. Stein, and G. Still. “Generalized semi-infinite programming: a tutorial”. In: *Journal of computational and applied mathematics* 217.2 (2008), pp. 394–419.

[VW02] R. C. Vaughan and T. D. Wooley. “Waring’s problem: a survey”. In: *Number theory for the millennium* 3 (2002), pp. 301–340.

[Wat67] L. J. Watters. “Reduction of integer polynomial programming problems to zero-one linear programming problems”. In: *Operations Research* 15.6 (1967), pp. 1171–1174.

[WSV00] H. Wolkowicz, R. Saigal, and L. Vandenberghe. *Handbook of semidefinite programming: theory, algorithms, and applications*. Vol. 27. Springer, 2000.

Acknowledgements

First and foremost, I wish to express my gratitude to my supervisor Prof. Dr. Anita Schöbel. Thank you, Anita, for your advice, insights, ideas and optimism that made this work possible. You were open-minded towards my approaches to the topic, however unusual they may have looked at first. Others said: "Your door is always open!", and it is true.

Furthermore, I would like to thank Dr. Tomáš Bajbar for our fruitful joint work which is also reported in Chapter 7 in this work. Tomáš, I look forward to our first publication!

Thanks go to Prof. Dr. Russell Luke for being an inspiring second supervisor.

I am grateful to Prof. Dr. Oliver Stein for being on the thesis committee as well as being supportive of the cooperation with Tomáš.

I express my gratitude to Prof. Dr. Christoph Buchheim and Dr. Lukas Katthän. Thanks, Christoph, thanks Lukas, for inspiring discussions.

The financial support from the Research Training Group 2088 is gratefully acknowledged. Furthermore, many thanks to the OptALI project which financed my stay in New Zealand during three months.

Also, I wish to thank Prof. Dr. Stefan Halverscheid and Dr. Nadine Czempinski, Arne Gerdes, Dr. Susann Graupner, Dr. Britta Schnoor and Marta Ziemba from the Q-Plus-(and Q-Plus++)-team to help me make the most of the introductory courses.

I am indebted to Dr. Jochen Schulz and Dr. Christoph Rügge. Your responsive support and expertise were really helpful, and no-matter-what IT stuff I bothered you with, solutions were found. I learned so much from you guys.

I wish to express my gratitude to Prof. Dr. Ingo Witt. In your lectures and seminars, I learned to use and admire analysis. Even though many years have passed since, with hindsight I believe that I acquired a great deal of my math skills through your teaching.

Also, I am grateful to Mrs Barann. Your empathy is invaluable. Your foresight in all administrative tasks spared, I am convinced, many PhD candidates at our faculty quite some trouble, if not horror.

I am grateful to Prof. Dr. Sven Krumke. We had a good time in New Zealand when I was far from home. Thanks for the good times, Sven!

A word of affection for my mother, my father, and my sisters Meike and Imke. Especially as having a mathematician as son or brother is probably not always that easy. And to Max and Tyge, for sunny days and shelter in times of rain.

Almost last but definitely not least, I am grateful to my colleagues: Alexander (Paris!), Anja (advice and expertise), Corinna (fun games), Jonas H. (awesome New Zealand), Jonas I. (positive vibes), Julius (God of Code, Paris was nice), Lisa (honest thoughts), Marco Be. (very supportive), Marco Bo. (dear socializer, brought chess upon us), Marie (such a powerhouse), Mirko (God of Chess), Morten (sporty discipline), Philine (deep thoughts), Robert (good jokes), Ruth (coauthorship), Sebastian (impressive dedication), Stephan (entertainment), Thorsten (peace of mind).

Finally – Lea! Endless patience, endless support. *Danke!* – with all my love.

Curriculum Vitae

Sönke Behrends

Persönliche Daten

Wohnort	Göttingen
Geburtsort	Marburg/Lahn
Staatsangehörigkeit	deutsch
E-Mail	s.behrends@math.uni-goettingen.de

Akademische Ausbildung

	Wissenschaftlicher Mitarbeiter
Seit 1. Juli 2013	Institut für Numerische und Angewandte Mathematik Universität Göttingen
	M.Sc. Mathematik
Mai 2013	Universität Göttingen
	Auslandsstudium
Oktober 2010 - Juli 2011	University of Warwick United Kingdom
	B.Sc. Mathematik
September 2010	Universität Göttingen
	B.Sc. Physik
April 2010	Universität Göttingen
	Abitur
Mai 2005	Gymnasium Martin-Luther-Schule Marburg