Practical polynomial optimization through positivity certificates with and without denominators

Ngoc Hoang Anh Mai

▶ **To cite this version:**

Ngoc Hoang Anh Mai. Practical polynomial optimization through positivity certificates with and without denominators. Optimization and Control [math.OC]. Université Paul Sabatier - Toulouse III, 2022. English. NNT : 2022TOU30165 . tel-03843327v2

HAL Id: tel-03843327

https://laas.hal.science/tel-03843327v2

Submitted on 26 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *21/09/2022* **par :**

Ngoc Hoang Anh MAI

Practical polynomial optimization through positivity certificates with and without denominators.

JURY

Amir Ali AHMADI	Professeur d'Université	Rapporteur
Monique LAURENT	Professeur d'Université	Membre du Jury
Jean-Bernard LASSERRE	Directeur de recherche	Co-directeur de thèse
David Russell LUKE	Professeur d'Université	Membre du Jury
Victor MAGRON	Chargé de recherche	Directeur de thèse
Thorsten THEOBALD	Professeur d'Université	Rapporteur
Kim-Chuan TOH	Professeur d'Université	Président du Jury
Jie WANG	Chercheur associé	Membre du Jury

École doctorale et spécialité :
MITT : Domaine Mathématiques : Mathématiques appliquées
Unité de Recherche :
Équipe POP, LAAS-CNRS (UPR 8001)
Directeur(s) de Thèse :
Victor MAGRON et *Jean-Bernard LASSERRE*
Rapporteurs :
Amir Ali AHMADI et *Thorsten THEOBALD*

Résumé

Les certificats de positivité ou Positivstellensätze fournissent des représentations de polynômes positifs sur des ensembles semialgébriques de basiques, c'est-à-dire des ensembles définis par un nombre fini d'inégalités polynomiales. Le célèbre Positivstellensatz de Putinar stipule que tout polynôme positif sur un ensemble semialgébrique basique fermé S peut être écrit comme une combinaison pondérée linéaire des polynômes décrivant S, sous une certaine condition sur S légèrement plus forte que la compacité. Lorsqu'il est écrit comme ceci, il devient évident que le polynôme est positif sur S, et donc cette description alternative fournit un certificat de positivité sur S.

De plus, comme les poids polynomiaux impliqués dans le Positivstellensatz de Putinar sont des sommes de carrés (SOS), de tels certificats de positivité permettent de concevoir des relaxations convexes basées sur la programmation semidéfinie pour résoudre des problèmes d'optimisation polynomiale (POP) qui surviennent dans diverses applications réelles, par exemple dans la gestion des réseaux d'énergie et l'apprentissage automatique pour n'en citer que quelques unes. Développée à l'origine par Lasserre, la hiérarchie des relaxations semidéfinies basée sur le Positivstellensatz de Putinar est appelée la *hiérarchie Moment-SOS*.

Dans cette thèse, nous proposons des méthodes d'optimisation polynomiale basées sur des certificats de positivité impliquant des poids SOS spécifiques, sans ou avec dénominateurs.

La première partie de ce manuscrit se concentre sur les méthodes sans dénominateurs, et basées sur le Positivstellensatz de Putinar.

Nous proposons une nouvelle hiérarchie Moment-SOS pour résoudre des problèmes d'optimisation de polynômes creux à grande échelle. Sa nouveauté est d'exploiter simultanément la parcimonie corrélative et la parcimonie des termes en combinant les avantages de deux cadres existants pour l'optimisation des polynômes creux. Ce faisant, nous obtenons (i) une hiérarchie à deux niveaux de relaxations de programmation semidéfinies avec la propriété cruciale de comporter des *blocs* de matrices SDP (au lieu d'une seule grande matrice), et (ii) la garantie de convergence vers l'optimum global sous certaines conditions. Nous démontrons son efficacité et son évolutivité sur plusieurs instances à grande échelle du célèbre problème Max-Cut et sur certaines instances de l'important problème industriel de flux de puissance optimal (OPF), impliquant jusqu'à six mille variables et des dizaines de milliers de contraintes.

Ensuite, nous prouvons que chaque relaxation moment semidéfinie d'un POP contraint peut être reformulée comme un programme semi-défini faisant intervenir une matrice ayant la *propriété de trace constante* (CTP). En conséquence, de telles relaxations peuvent être résolues efficacement par des méthodes du premier ordre qui exploitent le CTP, par exemple, la méthode de Lagrangien augmenté basée sur le gradient conditionnel. Nous étendons également ce cadre d'exploitation de CTP aux POPs à grande échelle avec différentes structures de parcimonie. L'efficacité de cette méthode est illustrée sur des relaxations de moment de second ordre pour divers programmes quadratiques à contrainte quadratique générés aléatoirement.

La deuxième partie de ce manuscrit porte sur les méthodes avec dénominateurs, basées sur les certificats de positivité dus à Putinar et Vasilescu, Reznick, ainsi qu'à Pólya.

Nous revisitons deux certificats de positivité sur des ensembles semialgébriques basiques (éventuellement non compacts) dus à Putinar et Vasilescu. Nous utilisons la technique de Jacobi pour fournir une preuve alternative avec un degré effectif lié aux poids SOS dans de tels certificats. En conséquence, nous pouvons définir une hiérarchie de relaxations semidéfinies pour les POPs généraux. La convergence vers un voisinage de la valeur optimale et la dualité forte sont garanties. Dans une seconde contribution, nous introduisons une nouvelle méthode numérique pour résoudre des systèmes d'inégalités et d'égalités polynomiales avec éventuellement *un nombre indénombrable* de solutions. En prime, on peut appliquer cette méthode pour obtenir des optimiseurs globaux approchés en optimisation polynomiale.

Nous fournissons ensuite une nouvelle borne sur le degré des poids SOS dans le Positivstellensatz de Putinar–Vasilescu et obtenons le nouveau Positivstellensatz suivant:

Si f est un polynôme de degré au plus $2d_f$, positif sur l'ensemble semi-algébrique $S := \{\mathbf{x} :$

$g_i(\mathbf{x}) \geq 0,\ i \in [m]\}$ à intérieur non vide (et avec $g_1 := R - \|\mathbf{x}\|_2^2$ avec $R > 0$), alors il existe des constantes positives $\bar{\mathfrak{c}}$ et $\mathfrak{c}$ dépendantes de f, g_i telles que pour tout $\varepsilon > 0$, pour tout $k \geq \bar{\mathfrak{c}}\varepsilon^{-\mathfrak{c}}$,

$$(1 + \|\mathbf{x}\|_2^2)^k (f + \varepsilon) = \sigma_0 + \sum_{i=1}^{m} \sigma_i g_i \,,$$

pour des polynômes SOS σ_i avec $\deg(\sigma_0)$ et $\deg(\sigma_i g_i)$ au plus $2(d_f + k)$. Ici $\|\cdot\|_2$ désigne la norme vectorielle ℓ_2. En conséquence, nous obtenons une hiérarchie convergente de relaxations semidéfinies pour les bornes inférieures en optimisation polynomiale sur des ensembles semi-algébriques compacts basiques. La complexité de cette hiérarchie est $\mathcal{O}(\varepsilon^{-\mathfrak{c}})$ pour une précision prescrite $\varepsilon > 0$. En particulier, si $m = L = 1$ alors $\mathfrak{c} = 65$, ce qui donne une complexité de calcul $\mathcal{O}(\varepsilon^{-65})$ pour minimiser un polynôme sur la boule unité.

Dans une autre contribution, nous dérivons une variante creuse du Positivstellensatz de Reznick. Si f est une forme définie positive, le Positivstellensatz de Reznick indique qu'il existe $k \in \mathbb{N}$ tel que $\|\mathbf{x}\|_2^{2k} f$ est un SOS. Si nous supposons maintenant que $f = \sum_{c=1}^{p} f_c$, où chaque forme f_c dépend d'un sous-ensemble des variables initiales, et supposons que ces sous-ensembles satisfont la *propriété d'intersection courante* (RIP). Alors il existe $k \in \mathbb{N}$ tel que $f = \sum_{c=1}^{p} \sigma_c / H_c^k$, où σ_c est une somme de carrés de polynômes, H_c est un dénominateur polynomial uniforme, et les deux polynômes σ_c, H_c dépendent uniquement des mêmes variables que f_c, pour chaque $c \in [p]$. En d'autres termes, le modèle de parcimonie de f se reflète également dans cette version parcimonieuse du certificat de positivité de Reznick. Nous utilisons ensuite ce résultat pour obtenir également des certificats de positivité pour (i) les polynômes non négatifs sur tout l'espace et (ii) les polynômes positifs sur un ensemble semialgébrique basique (éventuellement non compact), en supposant que les données d'entrée satisfont la RIP. Les deux sont des versions parcimonieuses du Positivstellensatz de Putinar-Vasilescu.

Enfin, nous considérons la minimisation d'un polynôme sur un ensemble semialgébrique contenu dans l'orthant positif. Il peut être converti en un POP équivalent en élevant au carré chaque variable. En utilisant la parité et le concept de *largeur de facteur*, nous proposons une hiérarchie de relaxations semidéfinies basée sur l'extension du Positivstellensatz de Pólya par Dickinson–Povh. Comme caractéristique distinctive et cruciale, la taille maximale de la matrice de chaque relaxation semidéfinie résultante peut être choisie arbitrairement. De plus, la suite de valeurs renvoyée par la nouvelle hiérarchie converge vers la valeur optimale du POP d'origine au taux $\mathcal{O}(\varepsilon^{-\mathfrak{c}})$ si l'ensemble semialgébrique a un intérieur non vide. Nous appliquons la même idée pour une extension du Positivstellensatz de Handelman pour obtenir une autre hiérarchie de relaxations semidéfinies avec une taille de matrice maximale prescrite. Lorsqu'elle est appliquée à la certification de la robustesse des réseaux de neurones multicouches et au calcul de valeurs singulières maximales positives, notre méthode basée sur le Positivstellensatz de Pólya fournit de meilleures bornes et s'exécute plusieurs centaines de fois plus rapidement que la hiérarchie Moment-SOS standard.

Mots-clés: optimisation polynomiale, hiérarchie Moment-SOS, taux de convergence, programmation semidéfinie, Positivstellensatz de Putinar, Positivstellensatz de Putinar–Vasilescu, Positivstellensatz de Reznick, Positivstellensatz de Pólya, propriété de trace constante, parcimonie des variables et des termes

Abstract

Positivity certificates or Positivstellensätze provide representations of polynomials positive on basic semialgebraic sets, i.e., sets defined by finitely many polynomial inequalities. The famous Putinar's Positivstellensatz states that every positive polynomial on a basic closed semialgebraic set S can be written as a linear weighted combination of the polynomials describing S, under a certain condition on S slightly stronger than compactness. When written in this it becomes obvious that the polynomial is positive on S, and therefore this alternative description provides a certificate of positivity on S.

Moreover, as the polynomial weights involved in Putinar's Positivstellensatz are sums of squares (SOS), such Positivity certificates enable to design convex relaxations based on semidefinite programming to solve polynomial optimization problems (POPs) that arise in various real-life applications, e.g., in management of energy networks and machine learning to cite a few. Originally developed by Lasserre, the hierarchy of semidefinite relaxations based on Putinar's Positivstellensatz is called the *Moment-SOS hierarchy*.

In this thesis, we provide polynomial optimization methods based on positivity certificates involving specific SOS weights, without or with denominators.

The first part of this manuscript focuses on methods without denominators, and based on Putinar's Positivstellensatz.

We propose a new Moment-SOS hierarchy for solving large-scale sparse polynomial optimization problems. Its novelty is to exploit simultaneously correlative sparsity and term sparsity by combining advantages of two existing frameworks for sparse polynomial optimization. In doing so we obtain (i) a two-level hierarchy of semidefinite programming relaxations with the crucial property to involve *blocks* of SDP matrices (instead of a single big matrix), and (ii) the guarantee of convergence to the global optimum under certain conditions. We demonstrate its efficiency and scalability on several large-scale instances of the celebrated Max-Cut problem and some instances of the important industrial optimal power flow problem (OPF), involving up to six thousand variables and tens of thousands of constraints.

Next, we prove that every semidefinite moment relaxation of a constrained POP can be reformulated as a semidefinite program involving a matrix with *constant trace property* (CTP). As a result, such relaxations can be solved efficiently by first-order methods that exploit CTP, e.g., the conditional gradient-based augmented Lagrangian method. We also extend this CTP-exploiting framework to large-scale POPs with different sparsity structures. Efficiency and scalability are illustrated on second-order moment relaxations for various randomly generated quadratically constrained quadratic programs.

The second part of this manuscript focuses on methods with denominators, based on positivity certificates due to Putinar and Vasilescu, Reznick, as well as Pólya.

We revisit two certificates of positivity on (possibly noncompact) basic semialgebraic sets due to Putinar and Vasilescu. We use Jacobi's technique to provide an alternative proof with an effective degree bound on the SOS weights in such certificates. As a consequence, we can define a hierarchy of semidefinite relaxations for general POPs. Convergence to a neighborhood of the optimal value as well as strong duality and analysis are guaranteed. In a second contribution, we introduce a new numerical method for solving systems of polynomial inequalities and equalities with possibly *uncountably* many solutions. As a bonus, one can apply this method to obtain approximate global optimizers in polynomial optimization.

We next provide a new degree bound on the SOS weights in Putinar–Vasilescu's Positivstellensatz and obtain the following new Positivstellensatz:

If f is a polynomial of degree at most $2d_f$, nonnegative on the semialgebraic set $S := \{\mathbf{x} : g_i(\mathbf{x}) \geq 0,\ i \in [m]\}$ with nonempty interior (and with $g_1 := R - \|\mathbf{x}\|_2^2$ for some $R > 0$), then there

exist positive constants $\bar{c}$ and c depending on f, g_i such that for any $\varepsilon > 0$, for all $k \geq \bar{\mathfrak{c}}\varepsilon^{-\mathfrak{c}}$,

$$(1 + \|\mathbf{x}\|_2^2)^k (f + \varepsilon) = \sigma_0 + \sum_{i=1}^{m} \sigma_i g_i \,,$$

for some SOS polynomials σ_i with $\deg(\sigma_0)$ and $\deg(\sigma_i g_i)$ at most $2(d_f + k)$. Here $\|\cdot\|_2$ denotes the ℓ_2 vector norm. As a consequence, we obtain a converging hierarchy of semidefinite relaxations for lower bounds in polynomial optimization on basic compact semialgebraic sets. The complexity of this hierarchy is $\mathcal{O}(\varepsilon^{-\mathfrak{c}})$ for prescribed accuracy $\varepsilon > 0$. In particular, if $m = L = 1$ then $\mathfrak{c} = 65$, which yields a $\mathcal{O}(\varepsilon^{-65})$ computational complexity for minimizing a polynomial on the unit ball.

In another contribution we derive a sparse variant of Reznick's Positivstellensatz. If f is a positive definite form, Reznick's Positivstellensatz states that there exists $k \in \mathbb{N}$ such that $\|\mathbf{x}\|_2^{2k} f$ is an SOS. Namely, assume that $f = \sum_{c=1}^{p} f_c$, where each form f_c depends on a subset of the initial variables, and assume that these subsets satisfy the so-called *running intersection property* (RIP). Then there exists $k \in \mathbb{N}$ such that $f = \sum_{c=1}^{p} \sigma_c / H_c^k$, where σ_c is a sum of squares of polynomials, H_c is a uniform polynomial denominator, and both polynomials σ_c, H_c involve the same variables as f_c, for each $c \in [p]$. In other words, the sparsity pattern of f is also reflected in this sparse version of Reznick's certificate of positivity. We then use this result to also obtain positivity certificates for (i) polynomials nonnegative on the whole space and (ii) polynomials nonnegative on a (possibly noncompact) basic semialgebraic set, assuming that the input data satisfy the RIP. Both are sparse versions of Putinar–Vasilescu's Positivstellensatz.

Finally, we consider the minimization of a polynomial on a semialgebraic set contained in the nonnegative orthant. It can be converted to an equivalent POP by squaring each variable. Using even symmetry and the concept of *factor width*, we propose a hierarchy of semidefinite relaxations based on the extension of Pólya's Positivstellensatz by Dickinson–Povh. As its distinguishing and crucial feature, the maximal matrix size of each resulting semidefinite relaxation can be chosen arbitrarily. Moreover, the sequence of values returned by the new hierarchy converges to the optimal value of the original POP at the rate $\mathcal{O}(\varepsilon^{-\mathfrak{c}})$ if the semialgebraic set has nonempty interior. When applied to (i) robustness cerification of multi-layer neural networks and (ii) computing positive maximal singular values, our method based on Pólya's Positivstellensatz provides better bounds and runs several hundred times faster than the standard Moment-SOS hierarchy.

Keywords: polynomial optimization, Moment-SOS hierarchy, convergence rate, semidefinite programming, Putinar's Positivstellensatz, Putinar–Vasilescu's Positivstellensatz, Reznick's Positivstellensatz, Pólya's Positivstellensatz, constant trace property, correlative and term sparsity

Acknowledgement

Firstly I would like to thank my supervisors for guiding and supporting me over the three years. I appreciate their constant enthusiasm and encouragement. They have set an example of excellence as a researcher, mentor, instructor, and role model. Secondly, I would also like to thank the Jury members for their guidance through the defense process. Their discussion, ideas, and feedback have been invaluable. Thirdly I thank my friends, lab mates, colleagues, and research team for a cherished time spent together in the lab and social settings. In addition, I am also grateful for my teachers' imparted knowledge and shared experiences with me. Finally, I would like to express my gratitude to my wife, parents, and brother. Without their tremendous understanding and encouragement over the past three years, it would be impossible for me to complete my study.

Contents

List of Figures

List of Tables

Table 1: Table of Notation

Usual sets

$\mathbb{N}$	set of natural integers
$[m]$	set $\{1, \ldots, m\}$ of m first consecutive positive integers with $[0] := \emptyset$
$\mathbb{R}$	set of real numbers
$\mathbb{R}_+$	set of nonnegative real numbers
$\mathbb{N}^{*d}$	set $\{u \in \mathbb{N} : u * d\}$, for $* \in \{\geq, \leq, >, <\}$

Linear algebra

$\mathbb{R}^{n \times m}$	space of real matrices with n rows and m columns
$\mathbf{I}_n$	identity matrix in $\mathbb{R}^{n \times n}$
$\text{trace}(\mathbf{M})$	trace of a square matrix $\mathbf{M}$
$\mathbf{M}^\top$	transposition of a matrix $\mathbf{M}$
$\mathcal{S}^n$	space of real symmetric matrices with n rows
$\mathcal{S}^n_+$	cone of symmetric positive semidefinite matrices
$\mathbf{M} \succeq 0$	$\mathbf{M} \in \mathcal{S}^n_+$
$\mathcal{S}^n_{++}$	open cone of symmetric positive definite matrices
$\mathbf{M} \succ 0$	$\mathbf{M} \in \mathcal{S}^n_{++}$
$\mathbf{M} \preceq 0$	$-\mathbf{M} \succeq 0$
$\mathbf{M} \prec 0$	$-\mathbf{M} \succ 0$
$\mathbf{M} \preceq \mathbf{N}$	$\mathbf{N} - M \ \succeq 0$

Euclidean geometry

$\mathbf{x} \in \mathbb{R}^n$	a real vector of n entries $(x_1, \ldots, x_n)$
$\mathbf{0}$	zero finite dimensional vector
$\mathbf{x} \geq 0$	$x_j \geq 0\,, \forall j \in [n]$
$\mathbf{x}^\top \mathbf{y}$	inner product of two finite dimensional real vectors
$\|\mathbf{x}\|_2$	Euclidean norm of a real vector $\mathbf{x} \in \mathbb{R}^n$
$\mathbb{B}^n$	the closed unit ball of $\mathbb{R}^n$
$\mathbb{S}^{n-1}$	the unit sphere of $\mathbb{R}^n$
Δ^n	the unit simplex $\{\mathbf{x} \in \mathbb{R}^n_+ : 1 \geq \sum_{j=1}^n x_j\}$
$B(\mathbf{a}, r)$	the closed ball of center $\mathbf{a} \in \mathbb{R}^n$ and radius $r \geq 0$
$\partial B(\mathbf{a}, r)$	the sphere of center $\mathbf{a} \in \mathbb{R}^n$ and radius $r \geq 0$
$V^\top$	vector space orthogonal to a given linear subspace V of $\mathbb{R}^n$

Analysis

$\mathcal{M}_+(\Omega)$	set of nonnegative Borel measures over $\Omega \subset \mathbb{R}^n$
$\mathcal{C}(\Omega)$	space of continuous functions on $\Omega \subset \mathbb{R}^n$
$\int_\Omega f dx$	Lebesgue integral of $f \in \mathcal{C}(\Omega)$ over $\Omega \subset \mathbb{R}^n$
$\int f d\lambda$	integral of $f \in \mathcal{C}(\Omega)$ w.r.t. $\lambda \in \mathcal{M}_+(\Omega)$
$\int_A f d\lambda$	integral of $f \in \mathcal{C}(A)$ over $A \subset \Omega \subset \mathbb{R}^n$ w.r.t. $\lambda \in \mathcal{M}_+(\Omega)$

Algebraic geometry

$\boldsymbol{\alpha} \in \mathbb{N}^n$	a multi-index vector of nonnegative integers $(\alpha_1, \ldots, \alpha_n)$
$\mathbf{e}_j$	multi-index with j-th coordinate equal to 1, all others being 0
$\mathbf{e}$	multi-index with each coordinate equal to 1
$\|\boldsymbol{\alpha}\|$	degree of $\boldsymbol{\alpha} \in \mathbb{N}^n$ defined by $\alpha_1 + \cdots + \alpha_n$
$\mathbb{N}^n_d$	index set with bounded degree $\{\boldsymbol{\alpha} \in \mathbb{N}^n : \|\boldsymbol{\alpha}\| \leq d\}$
$\mathbf{x}^{\boldsymbol{\alpha}}$	power $\boldsymbol{\alpha} \in \mathbb{N}^n$ of a vector $\mathbf{x} \in \mathbb{R}^n$ defined by $x_1^{\alpha_1} \cdots x_n^{\alpha_n}$
$\mathbb{R}[\mathbf{x}]$	ring of polynomials in a vector of variables $\mathbf{x} = (x_1, \ldots, x_n)$

$\Sigma[\mathbf{x}]$	cone of sums of squares of polynomials $f_1^2 + \ldots, f_r^2$
$\mathbb{R}[\mathbf{x}]_d$	space of polynomials of degree at most d
$\Sigma[\mathbf{x}]_d$	cone of SOS polynomials of degree at most $2d$
$q = \sum_{\boldsymbol{\alpha}} q_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}}$	polynomial in $\mathbf{x}$ with coefficients $q_{\boldsymbol{\alpha}} \in \mathbb{R}$
$\deg(q)$	the degree of $q \in \mathbb{R}[\mathbf{x}]$ defined by $\max\{\lvert\boldsymbol{\alpha}\rvert : q_{\boldsymbol{\alpha}} \neq 0\}$
$\mathbf{v}_d$	vector of monomials up to degree d, i.e., $(\mathbf{x}^{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in \mathbb{N}^n_d}$
$b(n,d)$, $b(d)$	$\binom{n+d}{d}$

Polynomial optimization

k	relaxation order
t	sparse order
s	the upper bound on the maximal block size
p	number of cliques
R	the positive parameter in the ball constraint
f	objective polynomial
$\mathfrak{g}$	set of polynomials $\{g_1, \ldots, g_m\}$ involved in the inequality constraints
$\mathfrak{h}$	set of polynomials $\{h_1, \ldots, h_l\}$ in the equality constraints
$S(\mathfrak{g})$	basic closed semialgebraic set defined by $\{\mathbf{x} \in \mathbb{R}^n : g_i(\mathbf{x}) \geq 0, i \in [m]\}$
$V(\mathfrak{h})$	variety defined by $\{\mathbf{x} \in \mathbb{R}^n : h_j(\mathbf{x}) = 0 : j \in [l]\}$
$S(\mathfrak{g}, \mathfrak{h})$	$S(\mathfrak{g}) \cap V(\mathfrak{h})$
$\mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in \mathbb{N}^n}$	real vector indexed by $\mathbb{N}^n$
$\lceil q \rceil$	$\lceil \deg(q)/2 \rceil$
$L_{\mathbf{y}}(q)$	the Riesz linear functional $L_{\mathbf{y}} : \mathbb{R}[\mathbf{x}] \to \mathbb{R}$ at q defined by $L_{\mathbf{y}}(q) := \sum_{\boldsymbol{\alpha}} q_{\boldsymbol{\alpha}} y_{\boldsymbol{\alpha}}$
$\mathbf{M}_d(\mathbf{y})$	the moment matrix of order d defined by $(y_{\boldsymbol{\alpha}+\boldsymbol{\beta}})_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{N}^n_d}$
$\mathbf{M}_d(q\mathbf{y})$	the localizing matrix of order d defined by $(\sum_{\boldsymbol{\gamma}} q_{\boldsymbol{\gamma}} y_{\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\gamma}})_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{N}^n_d}$
$\mathcal{Q}_d(\mathfrak{g})$	the truncated quadratic module of order d associated with $\mathfrak{g}$
$\mathcal{I}_d(\mathfrak{h})$	the truncated ideal of order d associated with $\mathfrak{h}$
$\mathcal{Q}_d(\mathfrak{g}, \mathfrak{h})$	$\mathcal{Q}_d(\mathfrak{g}) + \mathcal{I}_d(\mathfrak{h})$

Programming and softwares

LP	linear program
SDP	semidefinite program
IP	interior point method
ADMM	the alternating direction method of multipliers
SBM	spectral bundle methods
SM	spectral methods without bundle
CGAL	conditional gradient-based augmented Lagrangian

Chapter 1

Introduction

1.1 A brief history of positivity certificates

Deciding nonnegativity of a polynomial is an important and attractive problem throughout history of the development of real algebraic geometry. In his famous and seminal work [79], Hilbert characterized all cases where nonnegative polynomials are sums of squares (SOS) of polynomials. They are the first positivity certificates without denominators. Later Blekherman showed in [19] that there are significantly more nonnegative polynomials than SOS. In 1927 Artin proved in [7] that every nonnegative polynomial can be decomposed as a sum of squares of rational functions, thereby solving Hilbert's 17th problem. Namely, f is nonnegative if and only if $\sigma_D f = \sigma_N$ for some SOS polynomials σ_N and $\sigma_D \neq 0$. Accordingly, Hilbert–Artin's Positivstellensatz has a non-prescribed denominator. Nevertheless, in his celebrated work [175] Reznick provides a representation that involves a *uniform* denominator for positive definite forms. Later on, positivity certificates on a general semialgebraic set, involving denominators, have been proposed by Stengle [192] (see also Krivine [101]). A *basic semialgebraic* set $S(\mathfrak{g},\mathfrak{h})$ can be written as

$$S(\mathfrak{g},\mathfrak{h}) := \{\, \mathbf{x} \in \mathbb{R}^n : \, g_i(\mathbf{x}) \geq 0\,,\; i \in [m];\; h_j(\mathbf{x}) = 0\,,\; j \in [l]\,\}\,, \tag{1.1.1}$$

where $n \in \mathbb{N}$ is the dimension of the ambient space, $m, l \in \mathbb{N}$ are the number of inequality and equality constraints, and (g_i, h_j) are polynomials, $[m] := \{1, \ldots, m\}$. Stengle and Krivine rely on

$$\mathcal{P}(\mathfrak{g},\mathfrak{h}) := \left\{ \sum_{\boldsymbol{\alpha}\in\{0,1\}^m} \sigma_{\boldsymbol{\alpha}} g_1^{\alpha_1} \ldots g_m^{\alpha_m} + \sum_{j=1}^{l} \phi_j h_j : \, \sigma_{\boldsymbol{\alpha}} \in \Sigma[\mathbf{x}],\; \phi_j \in \mathbb{R}[\mathbf{x}] \right\}, \tag{1.1.2}$$

a tool from real algebraic geometry called *preordering*, associated with the polynomials (g_i, h_j). Here $\mathbb{R}[\mathbf{x}]$ denotes the ring of real polynomials and $\Sigma[\mathbf{x}] \subset \mathbb{R}[\mathbf{x}]$ stands for the set of SOS polynomials. Krivine–Stengle's Positivstellensatz (or *representation of positive polynomials*) states that

$$f \geq 0 \text{ on } S(\mathfrak{g},\mathfrak{h}) \quad \Leftrightarrow \quad \exists q_1, q_2 \in \mathcal{P}(\mathfrak{g},\mathfrak{h}),\; s \in \mathbb{N} :\; q_1 f = f^{2s} + q_2\,, \tag{1.1.3}$$

$$f > 0 \text{ on } S(\mathfrak{g},\mathfrak{h}) \quad \Leftrightarrow \quad \exists q_1, q_2 \in \mathcal{P}(\mathfrak{g},\mathfrak{h}) :\; q_1 f = 1 + q_2\,. \tag{1.1.4}$$

Notice that the above representations involve a multiplier q_1 for f as well as cross-products of the g_i's in (1.1.2).

When $S(\mathfrak{g},\mathfrak{h})$ is compact, Schmüdgen proves in [185] that if $f > 0$ on $S(\mathfrak{g},\mathfrak{h})$, then $f \in \mathcal{P}(\mathfrak{g},\mathfrak{h})$, yielding the first Positivstellensatz without denominators over a general compact semialgebraic set.

Let

$$\mathcal{Q}(\mathfrak{g},\mathfrak{h}) := \left\{ \sigma_0 + \sum_{i=1}^{m} \sigma_i g_i + \sum_{j=1}^{l} \phi_j h_j : \, \sigma_i \in \Sigma[\mathbf{x}],\, \phi_j \in \mathbb{R}[\mathbf{x}] \right\}. \tag{1.1.5}$$

The set $\mathcal{Q}(\mathfrak{g},\mathfrak{h})$ is called the *quadratic module* associated with the polynomials (g_i, h_j). In 1993 Putinar [168] refined Schmüdgen's Positivstelensatz [185] for compact basic semialgebraic sets (1.1.1) that satisfies an *Archimedean* condition. The latter states that $R - \|\mathbf{x}\|_2^2$ belongs to $\mathcal{Q}(\mathfrak{g},\mathfrak{h})$

Table 1.1: Some known Positivstellensätze.

Year	Author(s)	Statement	Compact	Denominator
1888	Hilbert-Artin [79, 7]	If f is globally nonnegative, then $\sigma_D f = \sigma_N$ for some $\sigma_D, \sigma_N \in \Sigma[\mathbf{x}]$.	no	yes (non-prescribed)
1964	Krivine–Stengle [101, 192]	If a polynomial f is nonnegative on $S(\mathfrak{g})$, then $\sigma f = \sum_{\boldsymbol{\alpha}\in\{0,1\}^m}(\sigma_{\boldsymbol{\alpha}} \prod_{i=1}^m g_i^{\alpha_i})$ for some $\sigma, \sigma_{\boldsymbol{\alpha}} \in \Sigma[\mathbf{x}]$.	no	yes (non-prescribed)
1974	Pólya [166]	If f is a form and $f > 0$ on $\mathbb{R}^n_+\backslash\{\mathbf{0}\}$, then $(\sum_j x_j)^k f$ has nonnegative coefficients for some $k \in \mathbb{N}$.	no	yes (prescribed)
1991	Schmüdgen [185]	If f is positive on $S(\mathfrak{g})$ and $S(\mathfrak{g})$ is compact, then $f = \sum_{\boldsymbol{\alpha}\in\{0,1\}^m}(\sigma_{\boldsymbol{\alpha}} \prod_{i=1}^m g_i^{\alpha_i})$ for some $\sigma_{\boldsymbol{\alpha}} \in \Sigma[\mathbf{x}]$.	yes	no
1993	Putinar [168]	If a polynomial f is positive on $S(\mathfrak{g})$ satisfying the Archimedean assumption, then $f = \sigma_0 + \sum_{i=1}^m \sigma_i g_i$ for some $\sigma_i \in \Sigma[\mathbf{x}]$.	yes	no
1995	Reznick [175]	If f is a positive definite form, then $\|\mathbf{x}\|_2^{2k} f \in \Sigma[\mathbf{x}]$ for some $k \in \mathbb{N}$.	no	yes (prescribed)
1999	Putinar–Vasilescu [170]	If a polynomial f is nonnegative on $S(\mathfrak{g})$, then for every $\varepsilon > 0$, there exists $k \in \mathbb{N}$ such that $\theta^k(f + \varepsilon\theta^d) = \sigma_0 + \sum_{i=1}^m \sigma_i g_i$ for some $\sigma_i \in \Sigma[\mathbf{x}]$, where $d := 1 + \lfloor \deg(f)/2 \rfloor$ and $\theta := \|\mathbf{x}\|_2^2 + 1$.	no	yes (prescribed)
2015	Dickinson–Povh [45]	If a polynomial f is nonnegative on $S(\mathfrak{g}) \subset \mathbb{R}^n_+$, then for every $\varepsilon > 0$, there exists $k \in \mathbb{N}$ such that $\theta^k(f + \varepsilon\theta^d) = \sigma_0 + \sum_{i=1}^m \sigma_i g_i$ for some σ_i being SOS of monomials, where $d := 1 + \lfloor \deg(f)/2 \rfloor$ and $\theta := \|\mathbf{x}\|_2^2 + 1$.	no	yes (prescribed)

for some $R > 0$; this can be automatically ensured by including the additional redundant constraint $g_{m+1}(\mathbf{x}) := R - \|\mathbf{x}\|_2^2 \geq 0$ in the definition of $S(\mathfrak{g}, \mathfrak{h})$. It avoids a multiplier for f and no cross-product of the g_i's, a highly desirable feature for optimization purposes. Explicitly, Putinar's Positivstellensatz states the following result:

Theorem 1.1. *(Putinar [168]) Let $f, g_1, \ldots, g_m, h_1, \ldots, h_l \in \mathbb{R}[\mathbf{x}]$. Assume that $\mathcal{Q}(\mathfrak{g}, \mathfrak{h})$ is Archimedean and f is positive on $S(\mathfrak{g}, \mathfrak{h})$. Then f belongs to $\mathcal{Q}(\mathfrak{g}, \mathfrak{h})$.*

In Table 1.1 we list some Positivstellensätze.

SOS decompositions of nonnegative polynomials have a distinguishing feature with important practical implications: Indeed they are *tractable* because they can be obtained by solving a *semidefinite program*. Semidefinite programming (SDP) is an important class of convex conic optimization problems that can be solved efficiently, up to arbitrary precision, fixed in advance; the interested reader is referred to, e.g., [16, Chapter 4]. Namely, writing a polynomial $f \in \mathbb{R}[\mathbf{x}]_{2d}$ as an SOS boils down [161] to computing the entries of a symmetric (Gram) matrix $\mathbf{G}$ with only nonnegative eigenvalues (denoted by "$\mathbf{G} \succeq 0$") such that $f = \mathbf{v}_d^\top \mathbf{G} \mathbf{v}_d$, with $\mathbf{v}_d$ being the vector of all monomials of degree at most d. This connection between SOS and SDP promotes many important applications of optimization, operations research, signal processing, computational geometry, probability and statistics, control, PDEs, let alone recent applications in quantum information and computer vision. For more details the interested reader is referred to, e.g., [208, 202, 214, 184, 183, 36, 34, 194, 154, 193] and references therein.

1.2 A brief history on polynomial optimization

Optimization of polynomials on semialgebraic sets is an important area of applied mathematics, which initially motivated the introduction of the Moment-SOS hierarchy, based on positivity certificates with SOS weights. Indeed since the pioneer works of Lasserre [102] and Parrilo

[157], SOS have now become a powerful tool in polynomial optimization. Given polynomials $f \in \mathbb{R}[\mathbf{x}], \mathfrak{g} = \{g_i\}_{i=1}^m \subset \mathbb{R}[\mathbf{x}], \mathfrak{h} = \{h_j\}_{j=1}^l \subset \mathbb{R}[\mathbf{x}]$, consider the following polynomial optimization problem (POP):

$$f^\star := \inf_{\mathbf{x} \in S(\mathfrak{g},\mathfrak{h})} f(\mathbf{x}), \tag{1.2.1}$$

with n variables, m inequality constraints, l equality constraints, and where $S(\mathfrak{g},\mathfrak{h})$ is defined as in (1.1.1). In general POP (1.2.1) is non-convex and NP-hard (see Laurent [110]).

Consider first the unconstrained case of POP (1.2.1) corresponding to $S(\mathfrak{g},\mathfrak{h}) = \mathbb{R}^n$. If $f - f^\star$ (≥ 0 on $\mathbb{R}^n$) is an SOS polynomial then $f^\star$ can be obtained by solving a single semidefinite program (SDP). Obviously denominators are not required in this case. However in general $f - f^\star$ is an SOS of rational functions (not polynomials) due to Krivine–Stengle's Positivstellensatz, and therefore denominators are required, which yields:

$$f^\star = \sup_{\lambda, \sigma_N, \sigma_D} \{\lambda : \sigma_D (f - \lambda) = \sigma_N ; \quad \sigma_N, \sigma_D \in \Sigma[\mathbf{x}] ; \quad \sigma_D(0) = 1 \}. \tag{1.2.2}$$

By fixing in advance a bound d on the degree of the denominator σ_D, one may solve (1.2.2) by SDP combined with bisection search on λ, and let d increase if there is no solution. The normalization constraint $\sigma_D(0) = 1$ ensures that neither the denominator σ_D nor the numerator σ_N is the zero polynomial.

In the constrained case, a basic idea is to rather consider

$$f^\star = \sup\{\lambda \in \mathbb{R} : f - \lambda \geq 0 \text{ on } S(\mathfrak{g},\mathfrak{h})\} \tag{1.2.3}$$

and replace the difficult constraint "$f - \lambda \geq 0$ on $S(\mathfrak{g},\mathfrak{h})$" with the stronger but more tractable certificate of positivity on $S(\mathfrak{g},\mathfrak{h})$ for $f - \lambda$. For instance, if $S(\mathfrak{g},\mathfrak{h})$ in (1.1.1) is compact and assuming with no loss of generality that the Archimedean assumption holds, Putinar's Positivstellensatz (Theorem 1.1) provides the decomposition $f - \lambda = \sigma_0 + \sum_{i=1}^m \sigma_i g_i$, with $\sigma_i \in \Sigma[\mathbf{x}]$. Then one obtains the monotone non-decreasing sequence $(\rho_k)_{k \in \mathbb{N}}$ of lower bounds on $f^\star$ defined by

$$\rho_k := \sup_{\lambda, \sigma_i} \{ \lambda : f - \lambda = \sigma_0 + \sum_{i=1}^m \sigma_i g_i, \ \sigma_i \in \Sigma[\mathbf{x}], \deg(\sigma_0) \leq 2k, \deg(\sigma_i g_i) \leq 2k \}, \tag{1.2.4}$$

where denominators are not needed. Moreover, by invoking Putinar's Positivstellensatz, one obtains the convergence $\rho_k \uparrow f^\star$ as k increases. Introduced by Lasserre in [102], for each fixed k, (1.2.4) is a semidefinite program and is an *SOS strengthening* of (1.2.3) (as we restrict the feasible set). The dual of (1.2.4) is also a semidefinite program which is a *Moment relaxation* of (1.2.1). The Moment-SOS hierarchy is the sequence (indexed by k) of semidefinite programs (1.2.4) and their associated duals. As proved by Nie [147], convergence of $(\rho_k)_{k \in \mathbb{N}}$ to $f^\star$ is *finite* for generic constraints $S(\mathfrak{g},\mathfrak{h})$, and with the numerical procedure of Henrion and Lasserre [77] one can extract global minimizers from an optimal solution of the (exact) semidefinite relaxation (dual to (1.2.4)) in the hierarchy. It relies on the flat extension condition of Curto and Fialkow [40, 111]. In the above-mentioned frameworks, compactness of $S(\mathfrak{g},\mathfrak{h})$ is crucial.

In this thesis, we treat the case where $S(\mathfrak{g},\mathfrak{h})$ is *not* compact by an appropriate use of Putinar–Vasilescu's Positivstellensatz stated in Table 1.1 with the uniform denominator $(1 + \|\mathbf{x}\|_2^2)^k$. Furthermore, we discover and exploit that with this uniform denominator one may obtain different types of Moment-SOS hierarchies with smaller computational complexity than standard ones, hence with a better efficiency for solving POPs.

1.3 Overview of degree bound and convergence rate

The convergence rate of the Moment-SOS hierarchy to the optimal value of a POP inherently depends on the complexity of the representation of positive polynomials. Roughly speaking, obtaining a lower degree bound on the SOS polynomials involved in the positivity certificate allows one to improve the convergence rate of the corresponding Moment-SOS hierarchy. How to find such lower degree bound is an interesting question and goes hand in hand with the quest of improving the convergence analysis of the Moment-SOS hierarchy. Let us review some of the standard

results on degree bounds of positivity certificates and the corresponding convergence rates of the associated Moment-SOS hierarchy.

We denote by $\mathbb{S}^{n-1}$ the unit sphere in $\mathbb{R}^n$. For each $q \in \mathbb{R}[\mathbf{x}]$, let

$$\delta(q) := \frac{\sup\left\{q(\mathbf{x}) : \mathbf{x} \in \mathbb{S}^{n-1}\right\}}{\inf\left\{q(\mathbf{x}) : \mathbf{x} \in \mathbb{S}^{n-1}\right\}}\,.$$

In [175] Reznick provides a Positivstellensatz involving a uniform denominator for positive definite forms with an explicit degree bound:

Theorem 1.2. *(Reznick [175, Theorem 3.12]) Suppose that $q \in \mathbb{R}[\mathbf{x}]$ is a positive definite form of degree $2d$, for some $d \in \mathbb{N}$. Then for $k \in \mathbb{N}$ and*

$$k \geq \frac{2nd(2d-1)}{4\log 2}\delta(q) - \frac{n+2d}{2}\,, \tag{1.3.1}$$

$\|\mathbf{x}\|_2^{2k} q$ is an SOS of polynomials.

In [175, Theorem 3.12], Reznick guarantees that the SOS decomposition of $\|\mathbf{x}\|_2^{2k} q$ is actually a sum of powers of linear forms. The degree bound of Reznick's Positivstellensatz yields a linear convergence rate of $\mathcal{O}(\varepsilon^{-1})$ for the minimization of a polynomial (see [130, Theorem 6]).

Powers and Reznick [167] improve the existing degree bound available for Pólya's Positivstellensatz [165] which is associated with another uniform denominator. Explicitly, if q is a homogeneous polynomial of degree d positive on the simplex

$$\Delta_n = \{\mathbf{x} \in \mathbb{R}^n \,:\, x_j \geq 0\,,\, j \in [n]\,,\, \textstyle\sum_{j\in[n]} x_j = 1\}\,, \tag{1.3.2}$$

then for all $k \in \mathbb{N}$ satisfying

$$k \geq \frac{d(d-1)\|q\|}{2\min_{\mathbf{x}\in\Delta_n} q(\mathbf{x})} - d\,, \tag{1.3.3}$$

$(\sum_{j\in[n]} x_j)^k q$ has positive coefficients. Here for each $h = \sum_{\boldsymbol{\alpha}} h_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}} \in \mathbb{R}[\mathbf{x}]$, we note $\|h\| := \max_{\boldsymbol{\alpha}} \frac{|h_{\boldsymbol{\alpha}}|}{c_{\boldsymbol{\alpha}}}$ with $c_{\boldsymbol{\alpha}} := \frac{|\boldsymbol{\alpha}|!}{\alpha_1!\ldots\alpha_n!}$ for each $\boldsymbol{\alpha} \in \mathbb{N}^n$. This yields a linear convergence rate of $\mathcal{O}(\varepsilon^{-1})$ for the minimization of a homogeneous polynomial on the simplex.

Applying the result of Powers and Reznick, Schweighofer [188] obtains a degree bound for Schmüdgen's Positivstellensatz [185] claiming that given a semialgebraic set

$$S(\mathfrak{g}) = \{\mathbf{x} \in \mathbb{R}^n \,:\, g_i(\mathbf{x}) \geq 0\,,\, i \in [m]\}\,, \tag{1.3.4}$$

which is a subset of $(-1,1)^n$ and a polynomial f having the minimal value $f^\star > 0$ on $S(\mathfrak{g})$, then there exists a real $\mathfrak{c} > 0$ depending on $\mathfrak{g} = \{g_i\}_{i\in[m]}$ such that for all $k \in \mathbb{N}$ satisfying

$$k \geq \mathfrak{c} d_f^2 \left(1 + \left(d_f^2 n^{d_f} \frac{\|f\|}{f^\star}\right)^{\mathfrak{c}}\right)\,, \tag{1.3.5}$$

one has $f \in \mathcal{P}_k(\mathfrak{g})$, where $\mathcal{P}_k(\mathfrak{g})$ is the truncated preordering of order $k \in \mathbb{N}$ associated with $S(\mathfrak{g})$:

$$\mathcal{P}_k(\mathfrak{g}) := \{\textstyle\sum_{\boldsymbol{\alpha}\in\{0,1\}^m} \sigma_{\boldsymbol{\alpha}} g_1^{\alpha_1}\ldots g_m^{\alpha_m} \,:\, \sigma_{\boldsymbol{\alpha}} \in \Sigma[\mathbf{x}]\,,\, \deg(\sigma_{\boldsymbol{\alpha}} g_1^{\alpha_1}\ldots g_m^{\alpha_m}) \leq k\}\,. \tag{1.3.6}$$

Consequently, the corresponding SOS hierarchy of lower bounds $(\rho_k^{\text{pre}})_{k\in\mathbb{N}}$, with

$$\rho_k^{\text{pre}} := \sup_{\lambda\in\mathbb{R}}\{\lambda \,:\, f - \lambda \in \mathcal{P}_k(\mathfrak{g})\}\,, \quad k \in \mathbb{N}\,, \tag{1.3.7}$$

converges to $f^\star$ with the rate $\mathcal{O}(\varepsilon^{-\mathfrak{c}})$. Observe that no denominator is required in this case. Nevertheless, the representation of $f - \lambda$ in $\mathcal{P}_k(\mathfrak{g})$ involves 2^m SOS polynomials.

Recently, by relying on polynomial kernel methods, Fang and Fawzi [57] have explicited the exponent $\mathfrak{c} := \frac{1}{2}$ in (1.3.5) when $S(\mathfrak{g})$ is the unit sphere $\mathbb{S}^{n-1}$. It yields the optimal convergence rate $\mathcal{O}(1/k^2)$ for the minimization of a polynomial on this set by using Schmüdgen's Positivstellensatz. Following Fang–Fawzi's method, Laurent and Slot obtain in [112, 191] similar convergence rates when $S(\mathfrak{g})$ is one of the following sets: the standard hypercube $[-1,1]^n$, the unit ball $\mathbb{B}^n$ and the

Table 1.2: Recent fastest convergence rates in polynomial optimization.

Year	Author(s) for complexity	Author(s) for Positivstellensatz	Semialgebraic set	Cone	Programming	Rate
1995	Reznick [175]	Reznick [175]	whole space	SOS	SDP	$\mathcal{O}(\varepsilon^{-1})$
2001	Powers–Reznick [167]	Pólya [166]	simplex	nonnegative orthant	LP	$\mathcal{O}(\varepsilon^{-1})$
2004	Schweighofer [188]	Schmüdgen [185]	compact	preordering	SDP	$\mathcal{O}(\varepsilon^{-\mathfrak{c}})$
2016	Fawzi–Saunderson–Parrilo [58]	Schmüdgen [185]	boolean hypercube	preodering	SDP	$\mathcal{O}(\varepsilon^{-1})$
2020	Fang–Fawzi [57]	Schmüdgen [185]	sphere	preodering	SDP	$\mathcal{O}(\varepsilon^{-1/2})$
2021	Laurent–Slot [112]	Schmüdgen [185]	hypercube	preodering	SDP	$\mathcal{O}(\varepsilon^{-1/2})$
2021	Baldi–Mourrain [13]	Putinar [168]	archimedean	quadratic module	SDP	$\mathcal{O}(\varepsilon^{-\mathfrak{c}})$
2021	Slot [191]	Schmüdgen [185]	ball, simplex	preordering	SDP	$\mathcal{O}(\varepsilon^{-1/2})$
2022	Mai–Magron [132] (Proposition 6.1)	Putinar–Vasilescu [170]	general	quadratic module	SDP	$\mathcal{O}(\varepsilon^{-\mathfrak{c}})$
2022	Mai–Magron–Lasserre–Toh [135]	Dickinson–Povh [45]	contained in the nonnegative orthant	nonnegative orthant, quadratic module	LP, SDP	$\mathcal{O}(\varepsilon^{-\mathfrak{c}})$

unit simplex Δ^n. Let $\mathcal{Q}_k(\mathfrak{g})$ denote the truncated quadratic module of order $k \in \mathbb{N}$ associated with $S(\mathfrak{g})$, that is:

$$\mathcal{Q}_k(\mathfrak{g}) := \{\sigma_0 + \textstyle\sum_{i=1}^m \sigma_i g_i \,:\, \sigma_i \in \Sigma[\mathbf{x}]\,,\, \deg(\sigma_0) \le k\,,\, \deg(\sigma_i g_i) \le k\}\,. \tag{1.3.8}$$

Notice that $\mathcal{P}_k(\mathfrak{g}) = \mathcal{Q}_k(\mathfrak{g})$ if $S(\mathfrak{g})$ is the unit ball or the unit sphere.

Applying the degree bound of Laurent and Slot [112], Baldi and Mourrain obtain in [13] polynomial degree bounds for Putinar's Positivstellensatz which improves the exponential bound given by Nie and Schweighofer [150].

Theorem 1.3. *(Baldi–Mourrain [13]) Let $f, g_1, \ldots, g_m$ be in $\mathbb{R}[\mathbf{x}]$. Assume that $\emptyset \neq S(\mathfrak{g}) \subset (-1,1)^n$ is Archimedean and that the minimal value $f^\star$ of f over $S(\mathfrak{g})$ is positive. Then there exist real numbers $\mathfrak{c}_1, \mathfrak{c}_2, \mathfrak{c}_3 > 0$ depending on $\mathfrak{g}$ such that for all $k \in \mathbb{N}$ satisfying*

$$k \ge \mathfrak{c}_1 d_f^{\mathfrak{c}_2} \left(\frac{\|f\|}{f^\star}\right)^{\mathfrak{c}_3}\,, \tag{1.3.9}$$

one has $f \in \mathcal{Q}_k(\mathfrak{g})$.

Accordingly, the corresponding SOS hierarchy of lower bounds $(\rho_k^{\mathrm{mod}})_{k\in\mathbb{N}}$, with

$$\rho_k^{\mathrm{mod}} := \sup_{\lambda\in\mathbb{R}}\{\lambda \,:\, f - \lambda \in \mathcal{Q}_k(\mathfrak{g})\}\,, \quad k \in \mathbb{N}\,, \tag{1.3.10}$$

converges to $f^\star$ with the rate $\mathcal{O}(\varepsilon^{-\mathfrak{c}})$. Moreover, the representation of $f-\lambda$ in $\mathcal{Q}_k(\mathfrak{g})$ involves only $m+1$ SOS polynomials which is in deep contrast with the exponential number of SOS polynomials involved in the representation in $\mathcal{P}_k(\mathfrak{g})$. Moreover, no denominator is required in this case. Most of recent convergence rates obtained for polynomial optimization are summarized in Table 1.2.

Recent work by Lombardi, Perucci and Roy [119] provides degree bounds in a quite general situation. The best degree bounds for Hilbert's 17th problem in three homogeneous variables are actually due to Hilbert's original 1893 paper [79]. Some lower bounds for Hilbert's 17th problem (quite far away from the upper bounds) were proved in [20] by Blekherman, Gouveia and Pfeiffer. Hilbert's 1893 result and sharp degree bounds for projective curves (i.e., curves defined by homogeneous polynomials) were proved in [21] by Blekherman, Smith and Velasco.

Table 1.3: Complexity comparison (in terms of floating-point operations) of several methods for solving SDP.

Method	Software	SDP type	Convergence rate	The most expensive part per iteration
IP [76] (second-order)	`SDPT3` [196], `Mosek` [6]	arbitrary	$\mathcal{O}(\log(1/\varepsilon))$ [201]	system of linear equations solving with complexity $\mathcal{O}((s^{\max})^6)$ [200, Table 1]
IP with non-symmetric cone [158] (second-order)	`alfonso` [159]	arbitrary	$\mathcal{O}(\log(1/\varepsilon))$	system of linear equations solving with complexity $\mathcal{O}(\eta^3)$
ADMM [24] (first-order)	`SCS` [156], `COSMO` [60]	arbitrary	$\mathcal{O}(\varepsilon^{-1})$ [81]	positive definite system of linear equations solving by $LDL^\top$-decomposition with complexity $\mathcal{O}((s^{\max})^6)$
SBM [75] (first-order)	`ConicBundle` [74]	with CTP	$\mathcal{O}(\log(1/\varepsilon)/\varepsilon)$ [47]	positive definite linear system solving with complexity $\mathcal{O}((s^{\max})^6)$
CGAL [218] (first-order)	`SketchyCGAL` [219]	with CTP	$\mathcal{O}(\varepsilon^{-2})$	smallest eigenvalue computing by the Arnoldi iteration with complexity $\mathcal{O}(s^{\max})$ [113]

1.4 Computational complexity of the Moment-SOS relaxations

So far we have discussed the complexity of each representation in term of degree bounds, which basically translates to convergence rates for the values returned by the related relaxations. We next consider the cost of the numerical resolution of a given relaxation.

Computational cost of moment relaxations. The k-th order moment relaxation for POP (1.2.1) can be rewritten in compact form as the following standard SDP:

$$\tau = \inf_{\mathbf{X}\in\mathcal{S}_+} \{\langle \mathbf{C}, \mathbf{X}\rangle \,:\, \langle \mathbf{A}_j, \mathbf{X}\rangle = \mathbf{b}_j \,,\, j \in [\zeta]\}\,, \tag{1.4.1}$$

where $\langle \mathbf{A}, \mathbf{B}\rangle := \text{trace}(\mathbf{AB})$ is the standard Frobenius inner product, $\mathcal{S}_+$ is the set of positive semidefinite (psd) matrices written in a block diagonal form as follows: $\mathbf{X} = \text{diag}(\mathbf{X}_1, \ldots, \mathbf{X}_\omega)$ with $\mathbf{X}_j$ being a block of size $s^{(j)}$, $j \in [\omega]$, and ζ is the number of affine constraints. We denote the largest block size by $s^{\max} := \max_{j\in[\omega]} s^{(j)}$.

Note that SDP-relaxation (1.4.1) of POP (1.2.1) at step k of the Moment-SOS hierarchy has $\omega = m + 1$ blocks whose largest size is $s^{\max} = \binom{n+k}{n}$ while the number of affine constraints is $\zeta = \mathcal{O}(\binom{n+k}{n}^2)$. Thus the computational cost for solving SDP (1.4.1) grows very rapidly with k. We say that SDP (1.4.1) has *constant trace property* (CTP) if there exists a positive real number a such that $\text{trace}(\mathbf{X}) = a$, for all feasible solution $\mathbf{X}$ of SDP (1.4.1). We also say that POP (1.2.1) has CTP when every moment relaxation of POP (1.2.1) has CTP.

In Table 1.3 are listed several available methods for solving SDP (1.4.1). Here we set $\eta := \binom{n+2k}{n}$. In particular, observe that two of them, CGAL and SBM, are first-order methods that exploit CTP. In [219], the authors combined CGAL with the Nyström sketch (named `SketchyCGAL`), which requires dramatically less storage than other methods and is very efficient for solving Shor's relaxation of large-scale Max-Cut instances.

SDP relaxations of non-convex quadratically constrained quadratic programs. A non-convex quadratically constrained quadratic (QCQP) program is a special instance of POP (1.2.1) for which the degree of the input polynomials is at most two. Famous instances of non-convex

QCQPs include the Max-Cut problem and the optimal power flow (OPF) problem [92]; in addition we recall that linearly constrained quadratic programs have an equivalent Max-Cut formulation [106]. They also have applications in deep learning, e.g., the computation of Lipschitz constants [34] and the stability analysis of recurrent neural networks [54]. On the one hand, local optima of non-convex QCQPs can be obtained by using local solvers; see e.g., [32]. On the other hand, the global optimum of QCQPs can be approximated as closely as desired by using moment (SDP) relaxations. In practice, non-convex QCQPs usually involve a large number of variables (say, $n \geq 1000$) and their associated SDP relaxations (1.4.1) can be classified in two groups as follows:

- **The first order relaxation**: $k = 1$ (also known as Shor's relaxation in the literature). In this case the number of affine constraints in SDP (1.4.1) is typically not exceeding the largest block size, i.e., $\zeta \leq s^{\max}$. It can be efficiently solved by most SDP solvers, in particular with `SketchyCGAL` [219]. Nevertheless the first order relaxation may provide only a lower bound of the optimal value of POP (1.2.1). In this case one needs to solve the second and perhaps even higher order relaxations to obtain tighter bounds on the global optimum.

- **The second and higher order relaxations**: $k \geq 2$. In this case the number of affine constraints in SDP (1.4.1) is typically much larger than the largest block size ($\zeta \gg s^{\max}$). Then unfortunately most SDP solvers cannot handle large-scale SDPs of this form.

Common issues of solving large-scale SDP relaxations. When solving the second and higher order SDP relaxations, SDP solvers often encounter the following issues:

- **Storage**: Interior-point methods are often chosen by users because of their high accuracy. These methods are efficient for solving medium-scale SDPs. However they frequently fail due to lack of memory when solving large-scale SDPs (say, $s^{\max} > 500$ and $\zeta > 2 \times 10^5$ on a standard laptop). First-order methods (e.g., ADMM, SBM, CGAL) provide an alternative to interior-point methods to avoid the memory issue. This is due to the fact that the cost per iteration of first-order methods is much lower than that of interior-point methods.

  At the price of losing convexity one can also rely on heuristic methods and replace the full matrix $\mathbf{X}$ in SDP (1.4.1) by a simpler one, in order to save memory. For instance, the Burer-Monteiro method [28] considers a low rank factorization of $\mathbf{X}$. However, to get correct results the rank cannot be too low [205] and therefore this limitation makes it useless for the second and higher order relaxations of POPs. Not suffering from such a limitation, CGAL not only maintains the convexity of SDP (1.4.1) but also possibly runs with implicit matrix $\mathbf{X}$.

- **Accuracy**: First-order methods have low convergence rates compared to the interior-point methods. Their performance depends heavily on the problem scaling and conditioning. As a result, in solving large-scale SDPs with first-order methods it is often difficult to obtain numerical results with high accuracy. Therefore, we do not expect the relative gap of the approximate value ($\text{val}_{\text{approx}}$) returned by first-order SDP solvers w.r.t. the exact value ($\text{val}_{\text{exact}}$), defined by
  $$\frac{|\text{val}_{\text{approx}} - \text{val}_{\text{exact}}|}{|\text{val}_{\text{exact}}|}, \tag{1.4.2}$$
  to be smaller than 10^{-8} (as for interior-point methods) but at least to be smaller than 1%.

1.5 Recent improvements on scalability and efficiency

Overcoming the scalability and efficiency issues mentioned in the previous section has become a major scientific challenge in polynomial optimization. Many recent efforts in this direction are mainly developed around the following ideas:

1. *SDP-relaxations variants* with *small* maximal matrix size solved efficiently by interior point methods. This includes correlative sparsity [203, 103], term sparsity [212, 211, 213], symmetry exploitation [61, 178], Jordan symmetry reduction [26], sublevel relaxations [35].

2. Exploit *low-rank structures* of SDP-relaxations; see, e.g., [215, 217].

3. *First-order methods* to solve SDP-relaxations involving matrix variables of potentially large size with constant trace [134, 131].

4. Develop *convex relaxations* that are based on alternatives to semidefinite cones. For example this includes linear programming (LP) [109, 3], second-order conic programming (SOCP) [123, 207, 3], copositive programming [162], non-symmetric conic programming [158], relative entropy programming [51, 142], geometric programming [52].

Sparsity exploitation is one of the notable methods without denominators for reducing the size of the Moment-SOS relaxations. For POPs in the form

$$f^\star := \min_{\mathbf{x}\in S(\mathfrak{g})} f(\mathbf{x})\,, \tag{1.5.1}$$

where $S(\mathfrak{g})$ is defined as in (1.3.4), Waki et al. [203] (resp. Wang et al. [212]) have exploited correlative (resp. term) sparsity to define appropriate sparse-variants of the associated standard SOS-relaxations. Roughly speaking, in a given standard SOS-relaxation, they break each matrix variable into many blocks of smaller sizes and solve the new resulting SDP via an interior-point solver (e.g., `Mosek` [5] or `SDPT3` [198]). It is due to the fact that the most expensive part of interior-point methods for a standard SDP:

$$\begin{array}{ll}\min\limits_{\mathbf{z},\mathbf{A}_j^{(t)}} & \mathbf{c}^\top\mathbf{z}\\ \text{s.t.} & \mathbf{z}\in\mathbb{R}^w\,,\ \mathbf{A}_j^{(t)}\in\mathbb{R}^{q\times q}\,,\\ & \mathbf{A}_0^{(t)}+\sum_{j=1}^w z_j\mathbf{A}_j^{(t)}\succeq 0\,,\ t\in[u]\,,\end{array}$$

is solving a square linear system in every iteration. It has the complexity $u(w^3q^3+w^3q^2)$, which mainly depends on the matrix size q. Thus one can solve the above SDP efficiently by using interior-point methods if q, w are small and u is large.

Modern SDP solvers via the interior-point method (e.g. `Mosek` [4]) can solve an SDP problem involving matrices of moderate size (say, $q \le 5,000$) and variables of moderate number (say, $w \le 20,000$) in reasonable time on a standard laptop [195]. The SDP relaxations arising from the Moment-SOS hierarchy typically involve matrices of size $\binom{n+k}{k}$ and variables of number $\binom{n+2k}{2k}$, where k is the relaxation order and n is the number of variables for a given POP. For problems with $n \simeq 200$, it is thus possible to compute the first-order SDP relaxation of a quadratically constrained quadratic problem (QCQP), as one can take $k = 1$, yielding $\binom{n+k}{k} \simeq 200$ and $\binom{n+2k}{2k} \simeq 20,000$ (in this case, this relaxation is also known as Shor's relaxation [189]). However, the quality of the resulting approximation is often not satisfactory and it is then required to go beyond the first-order relaxation. But for solving the second-order relaxation ($k = 2$) one is limited to problems of small size, typically with $\binom{n+4}{4} \le 20,000$ (hence with $n \le 23$) on a standard laptop. Therefore, in view of the current state of SDP solvers, the dense Moment-SOS hierarchy does not scale well enough.

One possible remedy is to rely on alternative weaker positivity certificates, such as the hierarchy of linear programming (LP) relaxations based on Krivine–Stengle's certificates [101, 192, 109] or the second-order cone programming (SOCP) relaxation based on (scaled) diagonally dominant sums of squares (DSOS/SDSOS) [3] to bound from below the minimum of f. Even though modern LP/SOCP solvers can handle much larger size problems by comparison with SDP solvers, they have been shown to provide less accurate bounds, in particular for combinatorial problems [110], and do not have the property of finite convergence for continuous problems (not even for convex QCQP problems [105, Section 9.3]). Below we outline some existing remedies to exploit sparsity in different settings.

Unconstrained POPs. A first option is to exploit *term sparsity* for sparse unconstrained problems, i.e., when $S(\mathfrak{g}) = \mathbb{R}^n$, and the objective f involves a few terms (monomials). The algorithm consists of automatically reducing the size of the corresponding SDP matrix by eliminating the monomial terms which never appear among the support of SOS decompositions [174]. Other classes of positivity certificates have been recently developed with a specific focus on sparse unconstrained problems. Instead of trying to decompose a positive polynomial as an SOS, one can try to decompose it as a sum of nonnegative circuits (SONC), by solving a geometric program [84]

or a second-order cone program [9, 207], or alternatively as a sum of arithmetic-geometric-mean-exponentials (SAGE) [33] by solving a relative entropy program. Despite their potential efficiency on certain sub-classes of POPs (e.g., sparse POPs with a small number of variables and a high degree), these methods share the common drawback of not providing systematic practical modeling frameworks for constrained problems.

Correlative sparsity. In order to reduce the computational burden associated with the dense Moment-SOS hierarchy while keeping its nice convergence properties, one possibility is to take into account the sparsity pattern satisfied by the variables of the POP [103, 203]. The resulting algorithm has been implemented in the `SparsePOP` solver [204] and can handle sparse problems with up to several hundred variables. Many applications of interest have been successfully handled thanks to this framework, for instance certified roundoff error bounds in computer arithmetics [121, 120] with up to several hundred variables and constraints, optimal power flow problems [92] (where a multi-ordered Lasserre hierarchy was proposed) with up to several thousand variables and constraints. More recent extensions have been developed for volume computation of sparse semialgebraic sets [194], approximating regions of attraction of sparse polynomial systems [193], noncommutative POPs [100], Lipschitz constant estimation of deep networks [34] and for sparse positive definite functions [133]. In these applications, the cost polynomial and the constraint polynomials possess a specific *correlative sparsity pattern*. The resulting sparse Moment-SOS hierarchy is obtained by building blocks of SDP matrices with respect to some subsets or *cliques* of the input variables. When the sizes of these cliques are reasonably small, one can expect to handle problems with a large number of variables. For instance, the maximal size of cliques is less than 10 for some unconstrained problems in [203] or roundoff error problems in [121], and is less than 20 for the optimal power flow problems handled in [92]. Even though correlative sparsity has been successfully used to tackle several interesting applications, there are still many POPs that cannot be handled by considering merely correlative sparsity. For instance, there are POPs for which the correlative sparsity pattern is (nearly) dense or which admits a correlative sparsity pattern with variable cliques of large cardinality (say, > 20), yielding untractable SDPs.

Term sparsity To overcome these issues, one can exploit *term sparsity* as described in [206, 212, 211]. The *TSSOS hierarchy* from [212] as well as the complementary *Chordal-TSSOS* from [211] offers some alternative to problems for which the correlative sparsity pattern is dense or nearly dense. In both TSSOS and Chordal-TSSOS frameworks a so-called *term sparsity pattern (tsp) graph* is associated with the POP. The nodes of this tsp graph are monomials (from a monomial basis) needed to construct SOS strengthenings of the POP. Two nodes are connected via an edge whenever the product of the corresponding monomials appears in the supports of polynomials involved in the POP or is a monomial square. Note that this graph differs from the *correlative sparsity pattern (csp) graph* used in [203] where the nodes are the input variables and the edges connect two nodes whenever the corresponding variables appear in the same term of the objective function or in the same constraint. A two-step iterative algorithm takes as input the tsp graph and enlarges it to exploit the term sparsity in (1.5.1). Each iteration consists of two successive operations: (i) a support extension operation and (ii) either a block closure operation on adjacency matrices in the case of TSSOS [212] or a chordal extension operation in the case of Chordal-TSSOS [211]. In doing so one obtains a two-level Moment-SOS hierarchy with blocks of SDP matrices. If the sizes of blocks are relatively small then the resulting SDP relaxations become more tractable as their computational cost is significantly reduced. Another interesting feature of TSSOS is that the block structure obtained at the end of the iterative algorithm automatically induces a partition of the monomial basis, which can be interpreted in terms of sign symmetries of the initial POP. TSSOS and Chordal-TSSOS allow one to solve POPs with several hundred variables for which there is no or little correlative sparsity to exploit; see [212, 211] for numerous numerical examples.

1.6 Organization and summary of contributions

The aim of this thesis is to address the above-mentioned complexity and efficiency issues. It consists of two parts related to positivity certificates without and with denominators, respectively:

1.6.1 Positivity certificates without denominators

- We provide in Chapter 2 an overview of the concepts and mandatory background related to the Moment-SOS hierarchy for polynomial optimization, which is of interest in this thesis.

- In Chapter 3, we exploit simultaneously correlative sparsity and term sparsity to improve scalability issues: (i) Correlative sparsity occurs when a polynomial is a sum of polynomials involving a subset of the initial variables. (ii) Term sparsity occurs when a polynomial involves only a few nontrivial terms. The idea is to exploit these different sparsity structures to derive specific relaxations, where one breaks each matrix variable of the standard (dense) SDP-relaxation into many blocks of smaller sizes and solve the new resulting SDP via an interior-point solver (e.g., `Mosek`, `SDPT3`). We apply this method to provide tight bounds for some large-scale optimal power flow problem involving up to six thousand variables and tens of thousands of constraints.

- In addition to size reduction of the SDP relaxations, as mentioned above, we introduce in Chapter 4 a new method that enables us to speed up the resolution of the SDP relaxations. Explicitly, if an equality constrained POP has a sphere constraint of the form $R - \|\mathbf{x}\|_2^2 = 0$ for some $R > 0$, the matrix variables involved in the relaxations have a constant trace, a property that can be further exploited. In Chapter 5, we extend this property to the case of POPs that have ball constraint of the form $R - \|\mathbf{x}\|_2^2 \geq 0$ for some $R > 0$. The constant trace property boils down to SDPs where the feasible set of solutions is the intersection of the semidefinite cone (matrices with nonnegative eigenvalues) and the hyperplane of trace one matrices. By using first-order methods, this property allows one to solve SDP-relaxations involving matrix variables of potentially large size.

1.6.2 Positivity certificates with denominators

- Chapter 6 focuses on designing a hierarchy of semidefinite relaxations based on Putinar–Vasilescu's Positivstellensatz. It allows one to handle general polynomial optimization problems over possibly noncompact semialgebraic sets. This Positivstellensatz is obtained by combining homogenization techniques with Putinar's Positivstellensatz. Each SOS polynomial weight involved in the representation is replaced by an SOS of rational function with (prescribed) denominator. We also provide a new numerical method to find a point in a given basic semialgebraic set that possibly has positive dimension. This method is utilized to find an approximate optimal solution of the initial POP, yielding more expressiveness.

- We provide in Chapter 7 an improved degree bound for Putinar–Vasilescu's Positivstellensatz, namely a polynomial degree bound in the input degrees. As a result (under some mild conditions) we obtain an $\mathcal{O}(\varepsilon^{-\mathfrak{c}})$ rate of convergence for the sequence of values returned by the corresponding hierarchy. The methodology consists of the following two steps: First, we provide a constructive proof of the Positivstellensatz. Then we provide degree bounds for its explicit SOS weights.

- In Chapter 8, we state a sparse version of Reznick's Positivstellensatz [175]. The dense version states that any positive definite form can be decomposed as an SOS of rational functions with prescribed denominators. In the sparse setting, the form is a sum of forms, where each summand only depends on a subset of the initial variables, and we obtain a decomposition into a sum of sparse rational SOS.

- Chapter 9 is dedicated to the minimization of a polynomial over a semialgebraic set contained in the nonnegative orthant. We provide a hierarchy of semidefinite relaxations based on Pólya's Positivstellensatz and associated with sums of squares of s-nomials, i.e., linear combinations of s monomials with real coefficients, where s is prescribed in advance. The advantage of these SDP relaxations is that the maximal block size s is controllable so that we can solve them efficiently by using interior-point methods. We also obtain a convergence rate for this hierarchy which is similar to the one based on Putinar–Vasilescu's Positivstellensatz.

1.7 List of publications

We conclude this introductory chapter with a list of submitted, accepted and published contributions.

1.7.1 Articles in peer-reviewed international journals

1. N. H. A. Mai, J.-B. Lasserre, and V. Magron. Positivity certificates and polynomial optimization on non-compact semialgebraic sets. *Mathematical Programming*, pages 1–43, 2021

2. Y. Ebihara, H. Waki, V. Magron, N. H. A. Mai, D. Peaucelle, and S. Tarbouriech. l2 induced norm analysis of discrete-time LTI systems for nonnegative input signals and its application to stability analysis of recurrent neural networks. *European Journal of Control*, 62:99–104, 2021

3. N. H. A. Mai and V. Magron. On the complexity of Putinar–Vasilescu's Positivstellensatz. *Journal of Complexity*, page 101663, 2022

4. N. H. A. Mai, V. Magron, and J. Lasserre. A sparse version of Reznick's Positivstellensatz. *Mathematics of Operations Research*, 2022

1.7.2 Articles in the proceedings of peer-reviewed international conferences

1. N. H. A. Mai, A. Bhardwaj, and V. Magron. The constant trace property in noncommutative optimization. In *Proceedings of the 2021 on International Symposium on Symbolic and Algebraic Computation*, pages 297–304, 2021

2. Y. Ebihara, H. Waki, V. Magron, N. H. A. Mai, D. Peaucelle, and S. Tarbouriech. Stability Analysis of Recurrent Neural Networks by IQC with Copositive Mutipliers. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 5098–5103. IEEE, 2021

1.7.3 Submitted publications with peer review proccess

1. N. H. A. Mai, V. Magron, and J.-B. Lasserre. A hierarchy of spectral relaxations for polynomial optimization. *arXiv preprint arXiv:2007.09027*, 2020

2. N. H. A. Mai, J.-B. Lasserre, V. Magron, and J. Wang. Exploiting constant trace property in large-scale polynomial optimization. *arXiv preprint arXiv:2012.08873*, 2020

3. J. Wang, V. Magron, J. B. Lasserre, and N. H. A. Mai. CS-TSSOS: Correlative and term sparsity for large-scale polynomial optimization. *arXiv preprint arXiv:2005.02828*, 2020

4. T. L. Dinh and N. H. A. Mai. Comparing different subgradient methods for solving convex optimization problems with functional constraints. *arXiv preprint arXiv:2101.01045*, 2021

5. V. Magron, N. H. A. Mai, Y. Ebihara, and H. Waki. Tractable semidefinite bounds of positive maximal singular values. *arXiv preprint arXiv:2202.08731*, 2022

6. N. H. A. Mai. Exact polynomial optimization strengthened with Fritz John conditions. *arXiv preprint arXiv:2205.04254*, 2022

7. N. H. A. Mai. On the exactness for polynomial optimization strengthened with Fritz John conditions. *arXiv preprint arXiv:2205.08450*, 2022

8. N. H. A. Mai. Complexity for exact polynomial optimization strengthened with Fritz John conditions. *arXiv preprint arXiv:2205.11797*, 2022

9. N. H. A. Mai. A symbolic algorithm for exact polynomial optimization strengthened with Fritz John conditions. *arXiv preprint arXiv:2206.02643*, 2022

Part I

Positivity certificates without denominators

Chapter 2

Polynomial optimization and the Moment-SOS hierarchy

2.1 General notation

With $\mathbf{x} = (x_1, \ldots, x_n)$, let $\mathbb{R}[\mathbf{x}]$ stand for the ring of real polynomials and let $\Sigma[\mathbf{x}] \subset \mathbb{R}[\mathbf{x}]$ be the subset of sum of squares (SOS) polynomials. Their restrictions to polynomials of degree at most d and $2d$ are denoted by $\mathbb{R}[\mathbf{x}]_d$ and $\Sigma[\mathbf{x}]_d$ respectively. For $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$, let $|\boldsymbol{\alpha}| := \alpha_1 + \cdots + \alpha_n$. For $d \in \mathbb{N}$, let $\mathbb{N}^n_d := \{\boldsymbol{\alpha} \in \mathbb{N}^n : |\boldsymbol{\alpha}| \leq d\}$ and $\mathbb{N}^{*d} := \{u \in \mathbb{N} : u * d\}$, for $* \in \{\geq, \leq, >, <\}$. For $r \in \mathbb{N}^{>0}$, let $[r] := \{1, \ldots, r\}$. Let $(\mathbf{x}^{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in \mathbb{N}^n}$ be the canonical monomial basis of $\mathbb{R}[\mathbf{x}]$ (sorted w.r.t. the graded lexicographic order) and $\mathbf{v}_d(\mathbf{x})$ be the vector of monomials of degree up to d, with length $b(n,d) := \binom{n+d}{n}$. When it is clear from the context, we also write $b(d)$ instead of $b(n,d)$. A polynomial $q \in \mathbb{R}[\mathbf{x}]_d$ can be written as $q(\mathbf{x}) = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^n_d} q_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}} = \mathbf{q}^\top \mathbf{v}_d(\mathbf{x})$, where $\mathbf{q} = (q_{\boldsymbol{\alpha}}) \in \mathbb{R}^{b(d)}$ is its vector of coefficients in the canonical monomial basis. For $q \in \mathbb{R}[\mathbf{x}]$, let $\lceil q \rceil := \lceil \deg(q)/2 \rceil$. The l_1-norm of a polynomial q is given by the l_1-norm of its vector of coefficients $\mathbf{q}$, that is $\|\mathbf{q}\|_1 := \sum_{\boldsymbol{\alpha}} |q_{\boldsymbol{\alpha}}|$. Given $\mathbf{a} \in \mathbb{R}^n$, the l_2-norm of $\mathbf{a}$ is $\|\mathbf{a}\|_2 := (a_1^2 + \cdots + a_n^2)^{1/2}$.

2.2 Riesz linear functional and moment/localizing matrices

Given a real-valued sequence $\mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in \mathbb{N}^n}$, let $L_{\mathbf{y}} : \mathbb{R}[\mathbf{x}] \to \mathbb{R}$ be the Riesz linear functional defined by $q \mapsto L_{\mathbf{y}}(q) := \sum_{\boldsymbol{\alpha}} q_{\boldsymbol{\alpha}} y_{\boldsymbol{\alpha}}$. Let d be a positive integer. A real infinite sequence $(y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in \mathbb{N}^n}$ has a *representing measure* if there exists a finite Borel measure μ such that $y_{\boldsymbol{\alpha}} = \int_{\mathbb{R}^n} \mathbf{x}^{\boldsymbol{\alpha}} d\mu(\mathbf{x})$ for every $\boldsymbol{\alpha} \in \mathbb{N}^n$. In this case, $(y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in \mathbb{N}^n}$ is called the moment sequence of μ. We denote by $\mathrm{supp}(\mu)$ the support of a Borel measure μ.

The moment matrix of order d associated with a real-valued sequence $\mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in \mathbb{N}^n}$ and $d \in \mathbb{N}^{>0}$, is the real symmetric matrix $\mathbf{M}_d(\mathbf{y})$ of size $b(d)$, with entries $(y_{\boldsymbol{\alpha}+\boldsymbol{\beta}})_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{N}^n_d}$. The localizing matrix of order d associated with $\mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in \mathbb{N}^n}$ and $q = \sum_{\boldsymbol{\gamma}} q_{\boldsymbol{\gamma}} \mathbf{x}^{\boldsymbol{\gamma}} \in \mathbb{R}[\mathbf{x}]$, is the real symmetric matrix $\mathbf{M}_d(q\,\mathbf{y})$ of size $b(d)$ with entries $(\sum_{\boldsymbol{\gamma}} q_{\boldsymbol{\gamma}} y_{\boldsymbol{\gamma}+\boldsymbol{\alpha}+\boldsymbol{\beta}})_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{N}^n_d}$.

Example 2.1. *Consider the simple case where $n = 1$, $q = 1 - x^2$ and $\mathbf{y} = (y_0, y_1, y_2, y_3, y_4)$. Then $L_{\mathbf{y}}(q) = y_0 - y_2$,*

$$\mathbf{M}_2(\mathbf{y}) = \begin{bmatrix} y_0 & y_1 & y_2 \\ y_1 & y_2 & y_3 \\ y_2 & y_3 & y_4 \end{bmatrix} \quad \textit{and} \quad \mathbf{M}_1(q\mathbf{y}) = \begin{bmatrix} y_0 - y_2 & y_1 - y_3 \\ y_1 - y_3 & y_2 - y_4 \end{bmatrix}. \tag{2.2.1}$$

2.3 Quadratic module and ideal

Let $\mathfrak{g} := \{g_i\}_{i\in[m]} \subset \mathbb{R}[\mathbf{x}]$ and $\mathfrak{h} := \{h_j\}_{j\in[l]} \subset \mathbb{R}[\mathbf{x}]$. Denote by $S(\mathfrak{g})$ and $V(\mathfrak{h})$ a basic semialgebraic set and a real variety defined respectively by

$$\begin{aligned} S(\mathfrak{g}) &:= \{\,\mathbf{x}\in\mathbb{R}^n :\, g_i(\mathbf{x})\geq 0\,,\, i\in[m]\,\} \text{ and} \\ V(\mathfrak{h}) &:= \{\,\mathbf{x}\in\mathbb{R}^n :\, h_j(\mathbf{x}) = 0\,,\, i\in[l]\,\}\,. \end{aligned} \tag{2.3.1}$$

Set $S(\mathfrak{g},\mathfrak{h}) := S(\mathfrak{g})\cap V(\mathfrak{h})$. In other words, we have

$$S(\mathfrak{g},\mathfrak{h}) = \{\,\mathbf{x}\in\mathbb{R}^n :\, g_i(\mathbf{x})\geq 0\,,\, i\in[m]\,;\, h_j(\mathbf{x}) = 0\,,\, j\in[l]\,\}\,. \tag{2.3.2}$$

The *quadratic module* associated with $\mathfrak{g}$ is defined by

$$\mathcal{Q}(\mathfrak{g}) := \{\sigma_0 + \sum_{i=1}^m \sigma_i g_i \;:\; \sigma_0\in\Sigma[\mathbf{x}]\,,\, \sigma_i\in\Sigma[\mathbf{x}]\} \tag{2.3.3}$$

and for a positive integer k, the set

$$\mathcal{Q}_k(\mathfrak{g}) := \{\sigma_0 + \textstyle\sum_{i=1}^m \sigma_i g_i \;:\; \sigma_0\in\Sigma[\mathbf{x}]_k\,,\, \sigma_i\in\Sigma[\mathbf{x}]_{k-\lceil g_i\rceil}\} \tag{2.3.4}$$

is the truncation of $\mathcal{Q}(\mathfrak{g})$ of order k.

Given $\mathfrak{h} = \{h_j\}_{j\in[l]} \subseteq \mathbb{R}[\mathbf{x}]$, the set

$$\mathcal{I}(\mathfrak{h}) := \{\sum_{j=1}^l \psi_j h_j \;:\; \psi_j\in\mathbb{R}[\mathbf{x}]\} \tag{2.3.5}$$

is the *ideal* generated by $\mathfrak{h}$, and the set $\mathcal{I}_k(\mathfrak{h}) := \{\sum_{j=1}^l \psi_j h_j \;:\; \psi_j\in\mathbb{R}[\mathbf{x}]_{2(k-\lceil h_j\rceil)}\}$ is the truncation of $\mathcal{I}(\mathfrak{h})$ of order k.

Set $\mathcal{Q}(\mathfrak{g},\mathfrak{h}) = \mathcal{Q}(\mathfrak{g}) + \mathcal{I}(\mathfrak{h})$ and $\mathcal{Q}_k(\mathfrak{g},\mathfrak{h}) = \mathcal{Q}_k(\mathfrak{g}) + \mathcal{I}_k(\mathfrak{h})$ for $k\in\mathbb{N}$.

2.4 The Moment-SOS hierarchy

A polynomial optimization problem (POP) is defined as

$$f^\star := \inf_{\mathbf{x}\in S(\mathfrak{g},\mathfrak{h})} f(\mathbf{x})\,, \tag{2.4.1}$$

where $S(\mathfrak{g},\mathfrak{h})$ are defined as in (2.3.1) for some polynomial $f\in\mathbb{R}[\mathbf{x}]$, $\mathfrak{g}$ and $\mathfrak{h}$ defined as in the previous section. We will assume that POP (2.4.1) has at least one global minimizer.

Set

$$k_{\min} := \max_{i,j}\{\lceil f\rceil, \lceil g_i\rceil, \lceil h_j\rceil\}\,. \tag{2.4.2}$$

Given a POP of the form (2.4.1), consider the following associated hierarchy of SOS strengthenings indexed by $k\in\mathbb{N}^{\geq k_{\min}}$:

$$\rho_k := \sup_{\xi\in\mathbb{R}}\{\,\xi \;:\; f-\xi\in\mathcal{Q}_k(\mathfrak{g},\mathfrak{h})\}\,. \tag{2.4.3}$$

For each $\sigma\in\Sigma[\mathbf{x}]_d$, there exists $\mathbf{G}\succeq 0$ such that $\sigma = \mathbf{v}_d^\top\mathbf{G}\mathbf{v}_d$. Thus for each $k\in\mathbb{N}^{\geq k_{\min}}$, (2.4.3) can be rewritten as an SDP:

$$\rho_k = \sup_{\xi,\mathbf{G}_i,\mathbf{q}_j}\left\{\xi \;\middle|\; \begin{array}{rl} \mathbf{G}_i\succeq 0\,,\, f-\xi = & \mathbf{v}_k^\top\mathbf{G}_0\mathbf{v}_k \\ & +\sum_{i\in[m]} g_i\mathbf{v}_{k-\lceil g_i\rceil}^\top\mathbf{G}_i\mathbf{v}_{k-\lceil g_i\rceil} \\ & +\sum_{j\in[l]} h_j\mathbf{v}_{2(k-\lceil h_j\rceil)}^\top\mathbf{q}_j \end{array}\right\}. \tag{2.4.4}$$

For every $k\in\mathbb{N}^{\geq k_{\min}}$, the dual of (2.4.4) reads as

$$\tau_k := \inf_{\mathbf{y}\in\mathbb{R}^{b(2k)}}\left\{L_{\mathbf{y}}(f) \;\middle|\; \begin{array}{l} \mathbf{M}_k(\mathbf{y})\succeq 0\,,\, y_{\mathbf{0}} = 1 \\ \mathbf{M}_{k-\lceil g_i\rceil}(g_i\,\mathbf{y})\succeq 0\,,\, i\in[m] \\ \mathbf{M}_{k-\lceil h_j\rceil}(h_j\,\mathbf{y}) = 0\,,\, j\in[l] \end{array}\right\}. \tag{2.4.5}$$

This primal-dual sequence of semidefinite programs (2.4.5)-(2.4.4) is the so-called Moment-SOS hierarchy for optimization (also known as "Lasserre's hierarchy"). If $\mathcal{Q}(\mathfrak{g},\mathfrak{h})$ is Archimedean, i.e., $R-\|\mathbf{x}\|_2^2 \in \mathcal{Q}(\mathfrak{g},\mathfrak{h})$ for some $R>0$, then $(\rho_k)_{k\in\mathbb{N}\geq k_{\min}}$ converges to $f^\star$ by invoking Putinar's Positivstellensatz (Theorem 1.1).

Here we slightly abuse terminology and say that there is a "zero duality gap" between (2.4.3) and (2.4.5) if $\tau_k=\rho_k$ and $\tau_k\in\mathbb{R}$ (the abuse of terminology is due to the fact that zero duality gap can occur with both values being infinite). Slater's condition, i.e., there exists a feasible point at which inequality constrains are strict (see, e.g., [25, Section 5.2.3]), on either (2.4.3) or (2.4.5) is a well-known sufficient condition to ensure zero duality gap. However, in case of equality constraints in the description (2.3.2) of $S(\mathfrak{g},\mathfrak{h})$, Slater's condition does *not* hold for (2.4.5).

Proposition 2.1. *(Josz-Henrion [91]) Let $f^\star$ be as in* (2.4.1) *with $S(\mathfrak{g},\mathfrak{h})\neq\emptyset$ as in* (2.3.2). *Assume that $R-\|\mathbf{x}\|^2\in\mathfrak{g}$ for some real $R>0$. Zero duality gap between the primal* (2.4.3) *and dual* (2.4.5) *holds for sufficiently large $k\in\mathbb{N}$, i.e., $\rho_k=\tau_k$ and $\tau_k\in\mathbb{R}$. Moreover, SDP* (2.4.5) *has an optimal solution.*

In [91] the authors prove that the set of optimal solutions of (2.4.5) is compact and therefore (2.4.5) has an optimal solution. Although there exist situations where SDP (2.4.3) has no optimal solution (see for instance the end of [147, Section 3]), the following proposition ensures the existence of an optimal solution under mild assumptions:

Proposition 2.2. *(Lasserre [102, Theorem 3.4 (a)]) If $S(\mathfrak{g},\mathfrak{h})$ has nonempty interior, then Slater's condition on the dual* (2.4.5) *holds for $k\geq k_{\min}$, where $k_{\min}$ is defined as in* (2.4.2). *In this case, $\rho_k=\tau_k$, $\tau_k\in\mathbb{R}$ and the primal* (2.4.3) *has an optimal solution.*

The convergence rate of $\mathcal{O}(k^{-\mathfrak{c}})$ for the sequence $(\rho_k)_{k\in\mathbb{N}}$ follows from Theorem 1.3. For more details on the Moment-SOS hierarchy and its various applications, the interested reader is referred to [104].

2.5 Extraction of global minimizers

Let $\delta_{\mathbf{a}}$ stand for the Dirac measure at point $\mathbf{a}\in\mathbb{R}^n$. The following result is a consequence of Curto–Fialkow's Flat Extension Theorem [40, 111].

Proposition 2.3. *Let $\mathbf{y}^\star$ be an optimal solution of the SDP* (2.4.5) *at some order $k\in\mathbb{N}$, and assume that the flat extension condition holds, i.e.,* $\operatorname{rank}(\mathbf{M}_{k-w}(\mathbf{y}^\star))=\operatorname{rank}(\mathbf{M}_k(\mathbf{y}^\star))=:r$, *with $w:=\max_{i,j}\{\lceil g_i\rceil,\lceil h_j\rceil\}$. Then $\mathbf{y}^\star$ has a representing r-atomic measure $\mu=\sum_{t=1}^r\lambda_j\delta_{\mathbf{a}^{(t)}}$, where $(\lambda_1,\ldots,\lambda_r)$ belongs to the standard $(r-1)$-simplex and $\{\mathbf{a}^{(1)},\ldots,\mathbf{a}^{(r)}\}\subset S(\mathfrak{g},\mathfrak{h})$. Moreover, $\tau_k=f^\star$ and $\mathbf{a}^{(1)},\ldots,\mathbf{a}^{(r)}$ are all global minimizers of POP* (2.4.1).

Henrion and Lasserre [77] provide a numerical algorithm to extract the r minimizers $\mathbf{a}^{(1)},\ldots,\mathbf{a}^{(r)}$ from $\mathbf{M}_k(\mathbf{y}^\star)$ when the assumptions of Proposition 2.3 hold.

2.6 Finite convergence

Second-order sufficient condition. Given $(\lambda_i)_{i\in[m]}$ and $(\gamma_j)_{i\in[l]}$, let:

$$\mathbf{x}\mapsto\mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\gamma}) := f(\mathbf{x})-\sum_{i\in[m]}\lambda_i\,g_i(\mathbf{x})-\sum_{j\in[l]}\gamma_j\,h_j(\mathbf{x}),\quad \mathbf{x}\in\mathbb{R}^n.$$

Given $x\in S(\mathfrak{g},\mathfrak{h})$, let $J(\mathbf{x}):=\{\,i\in[m]:g_i(\mathbf{x})=0\,\}$.

Definition 2.1. *(see [152, Chapter 2]) The second-order sufficient condition (S2) holds at $\mathbf{x}^\star\in S(\mathfrak{g},\mathfrak{h})$ under the three following conditions.*

- **KKT-Lagrange multipliers:** *There exist $\lambda_i^\star\geq 0$, $i\in[m]$, and $\gamma_j\in\mathbb{R}$, $j\in[l]$, such that $\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^\star,\boldsymbol{\lambda}^\star,\boldsymbol{\gamma}^\star)=0$ and $\lambda_i^\star\,g_i(\mathbf{x}^\star)=0$ for all $i\in[m]$.*

- **Linear independence constraint qualification:** *The family*

$$\{\nabla g_i(\mathbf{x}^\star), \nabla h_j(\mathbf{x}^\star)\}_{i\in J(\mathbf{x}^\star), j\in[l]}$$

is linearly independent.

- **Strict complementarity:** $\lambda_i^\star + g_i(\mathbf{x}^\star) > 0$, *for all* $i \in [m]$.
- $\mathbf{u}^\top \nabla^2_{\mathbf{x}} \mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\gamma}^\star)\, \mathbf{u} > 0$ *for all* $\mathbf{u} \neq 0$ *such that* $\mathbf{u}^\top \nabla g_i(\mathbf{x}^\star) = 0$ *and* $\mathbf{u}^\top \nabla h_j(\mathbf{x}^\star) = 0$, $i \in J(\mathbf{x}^\star)$, *for all* $j \in [l]$.

We say that $\mathcal{I}(\mathfrak{h})$ is real radical if

$$\mathcal{I}(\mathfrak{h}) = \{f \in \mathbb{R}[\mathbf{x}] \,:\, \exists m \in \mathbb{N} \,:\, -f^{2m} \in \Sigma[\mathbf{x}] + \mathcal{I}(\mathfrak{h})\}\,. \tag{2.6.1}$$

The following proposition provides a sufficient condition to ensure finite convergence of the sequence $(\tau_k)_{k\in\mathbb{N}}$.

Proposition 2.4. *The following statements are true:*

1. *(Nie [147]) The equality* $\tau_k = f^\star$ *occurs generically for some* $k \in \mathbb{N}$.
2. *(Lasserre [105, Theorem 7.5]) If (i)* $\mathcal{Q}(\mathfrak{g}, \mathfrak{h})$ *is Archimedean, (ii) the ideal* $\mathcal{I}(\mathfrak{h})$ *is real radical, and (iii) the second-order sufficient conditions (see Definition 2.1) hold at every global minimizer of POP* (2.4.1)*, then* $\tau_k = \rho_k = f^\star$ *for some* $k \in \mathbb{N}$ *and both primal-dual* (2.4.3)-(2.4.5) *have optimal solutions.*
3. *(Lasserre et al. [108, Proposition 1.1] and [105, Theorem 6.13]) If* $V(\mathfrak{h})$ *defined as in* (2.3.1) *is finite,* $\tau_k = \rho_k = f^\star$ *for some* $k \in \mathbb{N}$ *and both primal-dual* (2.4.3)-(2.4.5) *have optimal solutions. In this case, the flatness condition holds at order* k.

The first statement of Proposition 2.4 means that with fixed $f \in \mathbb{R}[\mathbf{x}]$, the equality $\tau_k = f^\star$ holds for $k \in \mathbb{N}$ sufficiently large on a Zariski open set (the complement of the zeros of a polynomial) in the space of the coefficients of g_i, h_j with given degrees. Note that the real radical property is not generic and so the condition "$\mathcal{I}(\mathfrak{h})$ is real radical" must be checked case by case. On the other hand, if $V(\mathfrak{h})$ is the real zero set of a squared system of polynomial equations, i.e., $l = n$, then generically $V(\mathfrak{h})$ has a finite number of points.

Chapter 3

Exploiting correlative and term sparsity

Most of the content of this chapter is from [213].

This chapter is concerned with solving large-scale polynomial optimization problems. As is often the case, the polynomials in the problem description involve only a few monomials of low degree and the ultimate goal is to exploit this crucial feature to provide semidefinite relaxations that are computationally much cheaper than those of the standard SOS-based hierarchy [102] or its sparse version [103, 203] based on correlative sparsity.

Throughout this chapter, we consider large-scale instances of the following polynomial optimization problem (POP):

$$f^\star = \inf_{\mathbf{x}} \{ f(\mathbf{x}) : \mathbf{x} \in S(\mathfrak{g}) \}, \tag{3.0.1}$$

where the objective function f is assumed to be a polynomial in n variables $\mathbf{x} = (x_1, \ldots, x_n)$ and the feasible set $S(\mathfrak{g}) \subseteq \mathbb{R}^n$ is assumed to be defined by a finite conjunction of m polynomial inequalities, namely

$$S(\mathfrak{g}) := \{\mathbf{x} \in \mathbb{R}^n : g_1(\mathbf{x}) \geq 0, \ldots, g_m(\mathbf{x}) \geq 0\}, \tag{3.0.2}$$

for some $\mathfrak{g} = \{g_1, \ldots, g_m\} \subset \mathbb{R}[\mathbf{x}]$. Here "large-scale" means that the magnitude of the number of variables n and the number of inequalities m can be both proportional to several thousands.

To tackle large-scale POPs, a natural idea is to simultaneously benefit from correlative and term sparsity patterns. This is the spirit of our contribution. Also in the same vein the work in [141] combines the (S)DSOS framework [3] with the TSSOS hierarchy but does not provide systematic convergence guarantees.

Contribution. Our main contribution in this chapter is as follows:

I. For large-scale POPs with a correlative sparsity pattern, we first apply the usual sparse polynomial optimization framework [103, 203] to get a coarse decomposition in terms of cliques of variables. Next we apply the term sparsity strategy (either TSSOS or Chordal-TSSOS) to each subsystem (which involves only one clique of variables) to reduce the size of SDPs even further. While the overall strategy is quite clear and simple, its implementation is not trivial and needs some care. Indeed for its coherency one needs to take extra care of the monomials which involve variables that belong to intersections of variable cliques (those obtained from correlative sparsity). The resulting combination of correlative sparsity (CS for short) and term sparsity produces what we call the *CS-TSSOS* hierarchy – a two-level hierarchy of SDP relaxations with blocks of SDP matrices, which yields a converging sequence of certified approximations for POPs. Under certain conditions, we prove that the corresponding sequence of optimal values converges to the global optimum of the POP.

II. Our algorithmic development of the CS-TSSOS hierarchy is fully implemented in the `TSSOS` tool [124]. The most recent version of `TSSOS` has been released within the Julia programming

language, which is freely available online and documented.[1] With `TSSOS`, the accuracy and scalability of the CS-TSSOS hierarchy are evaluated on several large-scale benchmarks coming from the continuous and combinatorial optimization literature. In particular, numerical experiments demonstrate that the CS-TSSOS hierarchy is able to handle challenging Max-Cut instances and optimal power flow instances with several thousand ($\simeq 6,000$) variables on a laptop whenever appropriate sparsity patterns are accessible. We remark that the CS-TSSOS framework has been recently extended to handle noncommutative polynomial optimization [210] and complex polynomial optimization [209].

The rest of the chapter is organized as follows: in Section 3.1, we provide preliminary background on correlative and term sparsity. In Section 3.2, we explain how to combine them to obtain a two-level CS-TSSOS hierarchy. Its convergence is analyzed in Section 3.3. Eventually, we provide numerical experiments for large-scale POP instances in Section 6.3.

3.1 Preliminaries

A polynomial $f \in \mathbb{R}[\mathbf{x}]$ can be written as $f(\mathbf{x}) = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} f_{\boldsymbol{\alpha}}\mathbf{x}^{\boldsymbol{\alpha}}$ with $\mathcal{A} \subseteq \mathbb{N}^n$ and $f_{\boldsymbol{\alpha}} \in \mathbb{R}, \mathbf{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1}\cdots x_n^{\alpha_n}$. The support of f is defined by $\operatorname{supp}(f) := \{\boldsymbol{\alpha} \in \mathcal{A} \mid f_{\boldsymbol{\alpha}} \neq 0\}$. We use $|\cdot|$ to denote the cardinality of a set. For a finite set $\mathcal{A} \subseteq \mathbb{N}^n$, let $\mathbf{x}^{\mathcal{A}}$ be the $|\mathcal{A}|$-dimensional column vector consisting of elements $\mathbf{x}^{\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \mathcal{A}$ (fix any ordering on $\mathbb{N}^n$). For a positive integer r, the set of $r \times r$ symmetric matrices is denoted by $\mathcal{S}^r$ and the set of $r \times r$ positive semidefinite (PSD) matrices is denoted by $\mathcal{S}^r_+$. A matrix $\mathbf{A} \in \mathcal{S}^r_+$ is written as $\mathbf{A} \succeq 0$. For matrices $\mathbf{A}, \mathbf{B} \in \mathcal{S}^r$, let $\langle \mathbf{A}, \mathbf{B}\rangle \in \mathbb{R}$ denote the trace inner-product, defined by $\langle \mathbf{A}, \mathbf{B}\rangle = \operatorname{trace}(\mathbf{A}^\top\mathbf{B})$, and let $\mathbf{A} \circ \mathbf{B} \in \mathcal{S}^r$ denote the Hadamard product, defined by $[\mathbf{A} \circ \mathbf{B}]_{ij} = A_{ij}B_{ij}$. For $d \in \mathbb{N}$, let For $\boldsymbol{\alpha} \in \mathbb{N}^n, \mathcal{A}, \mathcal{B} \subseteq \mathbb{N}^n$, let $\boldsymbol{\alpha} + \mathcal{B} := \{\boldsymbol{\alpha} + \boldsymbol{\beta} \mid \boldsymbol{\beta} \in \mathcal{B}\}$ and $\mathcal{A} + \mathcal{B} := \{\boldsymbol{\alpha} + \boldsymbol{\beta} \mid \boldsymbol{\alpha} \in \mathcal{A}, \boldsymbol{\beta} \in \mathcal{B}\}$. For $m \in \mathbb{N}\backslash\{0\}$, let $[m] := \{1, 2, \ldots, m\}$.

3.1.1 Chordal graphs and sparse matrices

In this subsection, we recall some basic results on chordal graphs and sparse matrices which are crucial for our subsequent development. Our notation and definitions here mostly follow from [199].

An *(undirected) graph* $G(V, E)$ or simply G consists of a set of nodes V and a set of edges $E \subseteq \{\{\mathbf{v}_i, \mathbf{v}_j\} \mid \mathbf{v}_i \neq \mathbf{v}_j, (\mathbf{v}_i, \mathbf{v}_j) \in V \times V\}$. For a graph G, we use $V(G)$ and $E(G)$ to indicate the node set of G and the edge set of G, respectively. The *adjacency matrix* of a graph G is denoted by $\mathbf{B}_G$ for which we put ones on its diagonal. For two graphs G, H, we say that G is a *subgraph* of H, denoted by $G \subseteq H$, if both $V(G) \subseteq V(H)$ and $E(G) \subseteq E(H)$ hold.

Definition 3.1. *A graph is called a* chordal graph *if all its cycles of length at least four have a chord*[2]*.*

The notion of chordal graphs plays an important role in sparse matrix theory. Any non-chordal graph $G(V, E)$ can be always extended to a chordal graph $G'(V, E')$ by adding appropriate edges to E, which is called a *chordal extension* of $G(V, E)$. As an example, in Figure 3.1 the two dashed edges are added to obtain a chordal extension. The chordal extension of G is usually not unique and the symbol G' is used to represent any specific chordal extension of G throughout the chapter. For graphs $G \subseteq H$, we assume that $G' \subseteq H'$ always holds in this chapter.

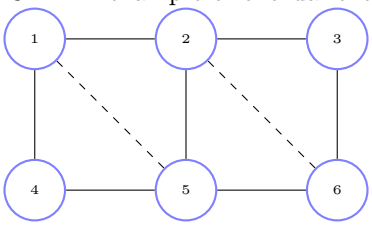
A *complete graph* is a graph in which any two nodes have an edge. A *clique* of a graph is a subset of nodes that induces a complete subgraph. A *maximal clique* is a clique that is not contained in any other clique. It is known that for a chordal graph, its maximal cliques can be enumerated efficiently in linear time in terms of the number of nodes and edges. See e.g. [18, 59, 65] for the details.

From now on we consider graphs with the node set $V \subseteq \mathbb{N}^n$. Given a graph $G(V, E)$, a symmetric matrix $\mathbf{G}$ with rows and columns indexed by V is said to have sparsity pattern G if $G_{\boldsymbol{\beta}\boldsymbol{\gamma}} = G_{\boldsymbol{\gamma}\boldsymbol{\beta}} = 0$ whenever $\boldsymbol{\beta} \neq \boldsymbol{\gamma}$ and $\{\boldsymbol{\beta}, \boldsymbol{\gamma}\} \notin E$. Let $\mathcal{S}_G$ be the set of symmetric matrices with

[1] `https://github.com/wangjie212/TSSOS`

[2] A chord is an edge that joins two nonconsecutive nodes in a cycle.

Figure 3.1: An example of chordal extension



sparsity pattern G. For a matrix in $\mathcal{S}_G$, its submatrices/blocks indexed by the maximal cliques of G play a crucial role, especially in the case when G is a chordal graph (see Theorems 3.1 and 3.2). The maximal size of blocks is the maximal size of maximal cliques of G, namely, the *clique number* of G.

Remark 3.1. *For a graph G, among all chordal extensions of G, there is a particular one G' which makes every connected component of G to be a complete subgraph. Accordingly, the matrix with sparsity pattern G' is block diagonal (after an appropriate permutation on rows and columns): each block corresponds to a connected component of G. We call this chordal extension the* maximal *chordal extension. In this chapter, we only consider chordal extensions that are subgraphs of the maximal chordal extension.*

Given a graph $G(V, E)$, the PSD matrices with sparsity pattern G form a convex cone

$$\mathcal{S}_+^{|V|} \cap \mathcal{S}_G = \{\mathbf{G} \in \mathcal{S}_G \mid \mathbf{G} \succeq 0\}. \tag{3.1.1}$$

Once the sparsity pattern graph $G(V, E)$ is a chordal graph, the cone $\mathcal{S}_+^{|V|} \cap \mathcal{S}_G$ can be decomposed as a sum of simple convex cones thanks to the following theorem and hence the related optimization problem can be solved more efficiently.

Theorem 3.1 ([2], Theorem 2.3)**.** *Let $G(V, E)$ be a chordal graph and assume that $C_1, \ldots, C_t$ are the list of maximal cliques of $G(V, E)$. Then a matrix $\mathbf{G} \in \mathcal{S}_+^{|V|} \cap \mathcal{S}_G$ if and only if $\mathbf{G}$ can be written as $\mathbf{G} = \sum_{i=1}^t \mathbf{G}_i$, where $\mathbf{G}_i \in \mathcal{S}_+^{|V|}$ has nonzero entries only with row and column indices coming from C_i for $i \in [t]$.*

Given a graph $G(V, E)$, let Π_G be the projection from $\mathcal{S}^{|V|}$ to the subspace $\mathcal{S}_G$, i.e., for $\mathbf{G} \in \mathcal{S}^{|V|}$,

$$\Pi_G(\mathbf{G})_{\boldsymbol{\beta}\boldsymbol{\gamma}} = \begin{cases} G_{\boldsymbol{\beta}\boldsymbol{\gamma}}, & \text{if } \boldsymbol{\beta} = \boldsymbol{\gamma} \text{ or } \{\boldsymbol{\beta}, \boldsymbol{\gamma}\} \in E, \\ 0, & \text{otherwise.} \end{cases} \tag{3.1.2}$$

The set $\Pi_G(\mathcal{S}_+^{|V|})$ denotes matrices that are projections of PSD matrices onto $\mathcal{S}_G$. More precisely,

$$\Pi_G(\mathcal{S}_+^{|V|}) = \{\Pi_G(\mathbf{G}) \mid \mathbf{G} \in \mathcal{S}_+^{|V|}\}. \tag{3.1.3}$$

One can easily check that the cone $\Pi_G(\mathcal{S}_+^{|V|})$ and the cone $\mathcal{S}_+^{|V|} \cap \mathcal{S}_G$ form a pair of dual cones in $\mathcal{S}_G$ (see [199, Chapter 10]). Moreover, for a chordal graph G, the decomposition result for matrices in $\mathcal{S}_+^{|V|} \cap \mathcal{S}_G$ given in Theorem 3.1 leads to the following characterization of matrices in the cone $\Pi_G(\mathcal{S}_+^{|V|})$.

Theorem 3.2 ([69], Theorem 7)**.** *Let $G(V, E)$ be a chordal graph and assume that $C_1, \ldots, C_t$ are the list of maximal cliques of $G(V, E)$. Then a matrix $\mathbf{G} \in \Pi_G(\mathcal{S}_+^{|V|})$ if and only if $\mathbf{G}[C_i] \succeq 0$ for $i \in [t]$, where $\mathbf{G}[C_i]$ denotes the principal submatrix of $\mathbf{G}$ indexed by the clique C_i.*

By Theorem 3.2, to check $\mathbf{G} \in \Pi_G(\mathcal{S}_+^{|V|})$, it suffices to check the positive semidefiniteness of certain blocks of $\mathbf{G}$. For more details on chordal graphs and sparse matrices, the reader may refer to [199].

3.1.2 Correlative sparsity

To exploit correlative sparsity in the Moment-SOS hierarchy for POPs, one proceeds in two steps: 1) decompose the set of variables into cliques according to the links between variables emerging in the input polynomial system, and 2) construct a sparse Moment-SOS hierarchy with respect to the former decomposition of variables [203].

More concretely, we define the *correlative sparsity pattern (csp) graph* associated with POP (3.0.1) to be the graph G^{csp} with nodes $V = [n]$ and edges E satisfying $\{i,j\} \in E$ if one of following holds:

(i) there exists $\boldsymbol{\alpha} \in \mathrm{supp}(f)$ s.t. $\alpha_i > 0, \alpha_j > 0$;

(ii) there exists $k \in [m]$ such that $x_i, x_j \in \mathrm{var}(g_k)$, where $\mathrm{var}(g_k)$ is the set of variables involved in g_k.

Let $(G^{\mathrm{csp}})'$ be a chordal extension of G^{csp} and $\{I_c\}_{c=1}^p$ be the list of maximal cliques of $(G^{\mathrm{csp}})'$ with $n_c := |I_c|$. Let $\mathbb{R}[\mathbf{x}(I_c)]$ denote the ring of polynomials in the n_c variables $\mathbf{x}(I_c) = \{x_j \mid j \in I_c\}$. We then partition the constraint polynomials $g_1, \ldots, g_m$ into groups $\{g_i \mid i \in J_c\}, c \in [p]$ which satisfy

(i) $J_1, \ldots, J_p \subseteq [m]$ are pairwise disjoint and $\cup_{c=1}^p J_c = [m]$;

(ii) for any $i \in J_c$, $\mathrm{var}(g_i) \subseteq I_c$, $c \in [p]$.

Next, with $c \in \{1, \ldots, p\}$ fixed, for $d \in \mathbb{N}$ and $g \in \mathbb{R}[\mathbf{x}(I_c)]$, let $\mathbf{M}_d(\mathbf{y}, I_c)$ (resp. $\mathbf{M}_d(g\mathbf{y}, I_c)$) be the moment (resp. localizing) submatrix obtained from $\mathbf{M}_d(\mathbf{y})$ (resp. $\mathbf{M}_d(g\mathbf{y})$) by retaining only those rows and columns indexed by $\boldsymbol{\beta} = (\beta_i) \in \mathbb{N}^n_d$ of $\mathbf{M}_d(\mathbf{y})$ (resp. $\mathbf{M}_d(g\mathbf{y})$) with $\mathrm{supp}(\boldsymbol{\beta}) \subset I_c$, where $\mathrm{supp}(\boldsymbol{\beta}) := \{i \mid \beta_i \neq 0\}$.

Assume that $f \in \mathbb{R}[\mathbf{x}]$ can be written as $f = f_1 + \cdots + f_p$, for some $f_c \in \mathbb{R}[\mathbf{x}(I_c)]$. We denote $k_{\min} := \max\{\lceil f \rceil, \lceil g_i \rceil\}$. Then with $k \geq k_{\min}$, the moment hierarchy based on correlative sparsity for POP (3.0.1) is defined as

$$\begin{aligned} \rho_k := \inf \quad & L_{\mathbf{y}}(f) \\ \text{s.t.} \quad & \mathbf{M}_k(\mathbf{y}, I_c) \succeq 0, \quad c \in [p], \\ & \mathbf{M}_{k-\lceil g_i \rceil}(g_i\mathbf{y}, I_c) \succeq 0, \quad i \in J_c, c \in [p], \\ & y_{\mathbf{0}} = 1. \end{aligned} \tag{3.1.4}$$

In the following, we refer to (3.1.4) as the *CSSOS* hierarchy for POP (3.0.1).

Remark 3.2. *As shown in [103] under some compactness assumption, the sequence* $(\rho_k)_{k \geq k_{\min}}$ *monotonically converges to the global optimum* $f^\star$ *of POP* (3.0.1).

3.1.3 Term sparsity

In contrast to the correlative sparsity pattern which focuses on links between *variables*, the term sparsity pattern focuses on links between *monomials* (or terms). To exploit term sparsity in the Moment-SOS hierarchy one also proceeds in two steps: 1) decompose each involved monomial basis into blocks according to the links between monomials emerging in the input polynomial system, and 2) construct a sparse Moment-SOS hierarchy with respect to the former decomposition of monomial bases [212, 211].

More concretely, let $\mathcal{A} = \mathrm{supp}(f) \cup \bigcup_{i=1}^m \mathrm{supp}(g_i)$ and $\mathbb{N}^n_{k-\lceil g_i \rceil}$ be the standard monomial basis for $i = 0, \ldots, m$. Fixing a relaxation order $k \geq k_{\min}$, we define the *term sparsity pattern (tsp) graph* associated with POP (3.0.1) or the support set $\mathcal{A}$, to be the graph G^{tsp}_k with node set $V_{k,0} := \mathbb{N}^n_k$ and edge set

$$E(G^{\mathrm{tsp}}_k) := \{\{\boldsymbol{\beta}, \boldsymbol{\gamma}\} \mid \boldsymbol{\beta} \neq \boldsymbol{\gamma} \in V_{k,0}, \boldsymbol{\beta} + \boldsymbol{\gamma} \in \mathcal{A} \cup (2\mathbb{N})^n\}, \tag{3.1.5}$$

where $(2\mathbb{N})^n := \{2\boldsymbol{\alpha} \mid \boldsymbol{\alpha} \in \mathbb{N}^n\}$.

For a graph $G(V, E)$ with $V \subseteq \mathbb{N}^n$, let $\mathrm{supp}(G) := \{\boldsymbol{\beta} + \boldsymbol{\gamma} \mid \boldsymbol{\beta} = \boldsymbol{\gamma}$ or $\{\boldsymbol{\beta}, \boldsymbol{\gamma}\} \in E\}$. We define the graphs $G^{(0)}_{k,0} := G^{\mathrm{tsp}}_k$ and, for $i \in [m]$, $G^{(0)}_{k,i}$ is the empty graph with node set $V_{k,i} := \mathbb{N}^n_{k-\lceil g_i \rceil}$ and empty edge set. Note that $\mathrm{supp}(G^{(0)}_{k,0}) = \mathcal{A} \cup 2\mathbb{N}^n_k$ and $\mathrm{supp}(G^{(0)}_{k,i}) = \emptyset$ for $i \geq 1$. Now for each

$i \in \{0\} \cup [m]$, we iteratively define an ascending chain of graphs $(G_{k,i}^{(t)}(V_{k,i}, E_{k,i}^{(t)}))_{t\geq 1}$. To this end, we start with the initial graph $G_{k,i}^{(0)}$ and each iteration consists of two successive operations:
1) **support extension.** Define $F_{k,i}^{(t)}$ to be the graph with nodes $V_{k,i}$ and with (recall $g_0 = 1$)

$$\begin{aligned} E(F_{k,i}^{(t)}) =& \{\{\boldsymbol{\beta}, \boldsymbol{\gamma}\} \mid \boldsymbol{\beta} \neq \boldsymbol{\gamma} \in V_{k,i}, \\ & (\boldsymbol{\beta} + \boldsymbol{\gamma} + \operatorname{supp}(g_i)) \cap (\cup_{j=0}^{m} \operatorname{supp}(G_{k,j}^{(t-1)})) \neq \emptyset\}, \quad i \in \{0\} \cup [m]. \end{aligned} \tag{3.1.6}$$

2) **chordal extension.** Let

$$G_{k,i}^{(t)} := (F_{k,i}^{(t)})', \quad i \in \{0\} \cup [m]. \tag{3.1.7}$$

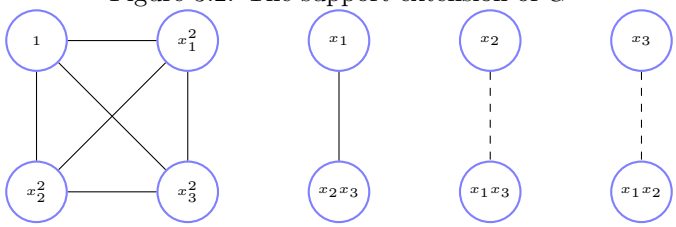
Note that $F_{k,0}^{(1)}$ has edges $\{\boldsymbol{\beta}, \boldsymbol{\gamma}\}$ with $\boldsymbol{\beta} + \boldsymbol{\gamma} \in \mathcal{A} \cup (2\mathbb{N})^n$. To summarise, the iterative process is

$$G_{k,i}^{(0)} \to \cdots \to G_{k,i}^{(t-1)} \xrightarrow{\text{support extension}} F_{k,i}^{(t)} \xrightarrow{\text{chordal extension}} G_{k,i}^{(t)} \to \cdots,$$

for each $i \in \{0\} \cup [m]$.

Example 3.1 (support extension)**.** *Assume $m = 0, k = 2$, and consider the graph G with solid edges shown in Figure 3.2. Then by support extension, the two dashed edges are added to G for $x_1x_2x_3 \in \operatorname{supp}(G)$.*

Figure 3.2: The support extension of G



Let $r_i := |\mathbb{N}^n_{k-\lceil g_i \rceil}| = \binom{n+k-\lceil g_i \rceil}{k-\lceil g_i \rceil}, i = 0, \ldots, m$. Then with $k \geq k_{\min}$ and $t \geq 1$, the moment hierarchy based on term sparsity for POP (3.0.1) is defined as

$$\begin{aligned} \inf \quad & L_{\mathbf{y}}(f) \\ \text{s.t.} \quad & \mathbf{B}_{G_{k,0}^{(t)}} \circ \mathbf{M}_k(\mathbf{y}) \in \Pi_{G_{k,0}^{(t)}}(\mathcal{S}_+^{r_0}), \\ & \mathbf{B}_{G_{k,i}^{(t)}} \circ \mathbf{M}_{k-\lceil g_i \rceil}(g_i\mathbf{y}) \in \Pi_{G_{k,i}^{(t)}}(\mathcal{S}_+^{r_i}), \quad i \in [m], \\ & y_{\mathbf{0}} = 1. \end{aligned} \tag{3.1.8}$$

The notation $\mathbf{B}_G \circ \mathbf{A}$ in (3.1.8) refers to a matrix whose $(\boldsymbol{\beta}, \boldsymbol{\gamma})$-entry is $A_{\boldsymbol{\beta}\boldsymbol{\gamma}}$ if $\boldsymbol{\beta} = \boldsymbol{\gamma}$ or $\{\boldsymbol{\beta}, \boldsymbol{\gamma}\} \in E(G)$, and 0 otherwise. We call t the *sparse order* and in the remainder of this chapter, the *TSSOS* hierarchy for POP (3.0.1) refers to the hierarchy (3.1.8).

Remark 3.3. *In* (3.1.8)*, one has the freedom to choose a specific chordal extension for any involved graph $G_{k,j}^{(t)}$. As shown in [212], if one chooses the maximal chordal extension then with k fixed, the resulting sequence of optimal values of the TSSOS hierarchy (as t increases) monotonically converges in finitely many steps to the optimal value of the corresponding dense moment relaxation for POP* (3.0.1).

3.2 The CS-TSSOS Hierarchy

When applicable, one can significantly improve the scalability of the Moment-SOS hierarchy by exploiting correlative sparsity or term sparsity. For large-scale POPs, it is then natural to ask whether one can combine correlative sparsity and term sparsity to further reduce the size of SDPs involved in the Moment-SOS hierarchy and to improve its scalability even more. As we shall see in the following sections, the answer is affirmative.

3.2.1 The CS-TSSOS Hierarchy for general POPs

Let us continue considering POP (3.0.1)[3]. A first natural idea to combine correlative sparsity and term sparsity would be to apply the TSSOS hierarchy for each subsystem (involving one variable clique) *separately*, once the cliques have been obtained from the csp graph of POP (3.0.1). However, with this naive approach convergence may be lost and in the following we take extra care to avoid this annoying consequence.

Let $G^{\rm csp}$ be the csp graph associated with POP (3.0.1), $(G^{\rm csp})'$ a chordal extension of $G^{\rm csp}$ and $\{I_c\}_{c=1}^p$ be the list of maximal cliques of $(G^{\rm csp})'$ with $n_c := |I_c|$. As in Section 3.1.2, the set of variables x is decomposed into $\mathbf{x}(I_1), \mathbf{x}(I_2), \ldots, \mathbf{x}(I_p)$. Let $J_1, \ldots, J_p$ be defined as in Section 3.1.2.

Now we apply the term sparsity pattern to each subsystem involving variables $\mathbf{x}(I_c)$, $c \in [p]$ respectively as follows. Let

$$\mathcal{A} := \mathrm{supp}(f) \cup \bigcup_{i=1}^m \mathrm{supp}(g_i) \text{ and } \mathcal{A}_c := \{\boldsymbol{\alpha} \in \mathcal{A} \mid \mathrm{supp}(\boldsymbol{\alpha}) \subset I_c\} \tag{3.2.1}$$

for $c \in [p]$. As before, we set $k_{\min} := \max\{\lceil f \rceil, \lceil g_1 \rceil, \ldots, \lceil g_m \rceil\}$ and $g_0 := 1$. Fix a relaxation order $k \geq k_{\min}$ and let $\mathbb{N}^{n_c}_{k-\lceil g_i \rceil}$ be the standard monomial basis for $i \in \{0\} \cup J_c$, $c \in [p]$. Let $G^{\rm tsp}_{k,c}$ be the tsp graph with nodes $\mathbb{N}^{n_c}_k$ associated with $\mathcal{A}_c$ defined as in Section 3.1.3, i.e., its node set is $\mathbb{N}^{n_c}_k$ and $\{\boldsymbol{\beta}, \boldsymbol{\gamma}\}$ is an edge if $\boldsymbol{\beta} + \boldsymbol{\gamma} \in \mathcal{A}_c \cup (2\mathbb{N})^{n_c}$. Note that we embed $\mathbb{N}^{n_c}$ into $\mathbb{N}^n$ via the map $\boldsymbol{\alpha} = (\alpha_i) \in \mathbb{N}^{n_c} \mapsto \boldsymbol{\alpha}' = (\alpha'_i) \in \mathbb{N}^n$ which satisfies

$$\alpha'_i = \begin{cases} \alpha_i, & \text{if } i \in I_c, \\ 0, & \text{otherwise.} \end{cases}$$

Let us define $G^{(0)}_{k,c,0} := G^{\rm tsp}_{k,c}$ and $G^{(0)}_{k,c,i}$, $i \in J_c$, $c \in [p]$ are all empty graphs with nodes $\mathbb{N}^{n_c}_{k-\lceil g_i \rceil}$. Next, for an integer $k \geq 1$, for each $i \in \{0\} \cup J_c$, $c \in [p]$, we iteratively define an ascending chain of graphs $(G^{(t)}_{k,c,i}(V_{k,c,i}, E^{(t)}_{k,c,i}))_{t \geq 1}$ with $V_{k,c,i} := \mathbb{N}^{n_c}_{k-\lceil g_i \rceil}$ via two successive operations:

1) **support extension.** Define $F^{(t)}_{k,c,i}$ to be the graph with nodes $V_{k,c,i}$ and with

$$E(F^{(t)}_{k,c,i}) = \{\{\boldsymbol{\beta}, \boldsymbol{\gamma}\} \mid \boldsymbol{\beta} \neq \boldsymbol{\gamma} \in V_{k,c,i}, (\boldsymbol{\beta} + \boldsymbol{\gamma} + \mathrm{supp}(g_i)) \cap \mathcal{C}^{(t-1)}_k \neq \emptyset\}, \tag{3.2.2}$$

where

$$\mathcal{C}^{(t-1)}_k := \bigcup_{c=1}^p (\cup_{i \in \{0\} \cup J_c} (\mathrm{supp}(g_i) + \mathrm{supp}(G^{(t-1)}_{k,c,i}))). \tag{3.2.3}$$

2) **chordal extension.** Let

$$G^{(t)}_{k,c,i} := (F^{(t)}_{k,c,i})', \quad i \in \{0\} \cup J_c\,,\ c \in [p]. \tag{3.2.4}$$

Example 3.2. *Let $f = 1 + x_1^2 + x_2^2 + x_3^2 + x_1x_2 + x_2x_3 + x_3$ and consider the unconstrained POP: $\min\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n\}$. We have $n = 3, m = 0$ and take the relaxation order $k = k_{\min} = 1$. The variables are decomposed into two cliques: $\{x_1, x_2\}$ and $\{x_2, x_3\}$. The tsp graphs with respect to these two cliques are illustrated in Figure 3.3. The left graph corresponds to the first clique: x_1 and x_2 are connected because of the term x_1x_2. The right graph corresponds to the second clique: 1 and x_3 are connected because of the term x_3; x_2 and x_3 are connected because of the term x_2x_3. If we apply the TSSOS hierarchy (using the maximal chordal extension in (3.2.4)) separately for each clique, then the graph sequences $(G^{(t)}_{1,c})_{t \geq 1}, c = 1, 2$ (the subscript j is omitted here since there is no constraint) stabilize at $t = 1$. However, the added (dashed) edge in the right graph corresponds to the monomial x_2, which only involves the variable x_2 belonging to the first clique. Hence we need to add the edge connecting 1 and x_2 to the left graph in order to get the guarantee of convergence as we shall see in Section 3.3.1. Consequently, the graph sequences $(G^{(t)}_{1,c})_{t \geq 1}, c = 1, 2$ stabilize at $t = 2$.*

[3]Though we only include inequality constraints in the definition of $S(\mathfrak{g})$ (3.0.2) for the sake of simplicity, equality constraints can be treated in a similar way.

Figure 3.3: The tsp graphs of Example 3.2. The dashed edge is added after the maximal chordal extension.

Let $r_{c,i} := b(n_c, k - \lceil g_i \rceil)$ for all c, i. Then with $k \geq 1$, the moment hierarchy based on correlative-term sparsity for POP (3.0.1) is defined as

$$\begin{aligned} \rho_k^{(t)} := \inf \quad & L_{\mathbf{y}}(f) \\ \text{s.t.} \quad & \mathbf{B}_{G_{k,c,0}^{(t)}} \circ \mathbf{M}_k(\mathbf{y}, I_c) \in \Pi_{G_{k,c,0}^{(t)}}(\mathcal{S}_+^{r_{c,0}}), \quad c \in [p], \\ & \mathbf{B}_{G_{k,c,i}^{(t)}} \circ \mathbf{M}_{k-\lceil g_i \rceil}(g_i \mathbf{y}, I_c) \in \Pi_{G_{k,c,i}^{(t)}}(\mathcal{S}_+^{r_{c,i}}), \quad i \in J_c\,,\, c \in [p], \\ & \mathbf{y_0} = 1. \end{aligned} \tag{3.2.5}$$

Proposition 3.1. *Fixing a relaxation order $k \geq k_{\min}$, the sequence $(\rho_k^{(t)})_{t\geq 1}$ is monotonically non-decreasing and $\rho_k^{(t)} \leq \rho_k$ for all t.*

Proof. By construction, we have $G_{k,c,j}^{(t)} \subseteq G_{k,c,j}^{(t+1)}$ for all k, c, j and all t. It follows that each maximal clique of $G_{k,c,j}^{(t)}$ is a subset of some maximal clique of $G_{k,c,j}^{(t+1)}$. Hence by Theorem 3.2, (3.2.5) with value $\rho_k^{(t)}$ is a relaxation of (3.2.5) with value $\rho_k^{(t+1)}$ and is clearly also a relaxation of (3.1.4) with value ρ_k. Therefore, $(\rho_k^{(t)})_{t\geq 1}$ is monotonically non-decreasing and $\rho_k^{(t)} \leq \rho_k$ for all t. □

Proposition 3.2. *Fixing a sparse order $t \geq 1$, the sequence $(\rho_k^{(t)})_{k\geq k_{\min}}$ is monotonically non-decreasing.*

Proof. The conclusion follows if we can show that $G_{k,c,j}^{(t)} \subseteq G_{k+1,c,j}^{(t)}$ for all k, c, j, t since by Theorem 3.2 this implies that (3.2.5) is a relaxation of (3.2.5) with value $\rho_{k+1}^{(t)}$. Let us prove $G_{k,c,j}^{(t)} \subseteq G_{k+1,c,j}^{(t)}$ by induction on t. For $t = 1$, from (3.1.5), we have $G_{k,c,0}^{(0)} = G_{k,c}^{\text{tsp}} \subseteq G_{k+1,c}^{\text{tsp}} = G_{k+1,c,0}^{(0)}$, which together with (3.2.2)-(3.2.3) implies that $F_{k,c,j}^{(1)} \subseteq F_{k+1,c,j}^{(1)}$ for $j \in \{0\} \cup J_c, c \in [p]$. It then follows that $G_{k,c,j}^{(1)} = (F_{k,c,j}^{(1)})' \subseteq (F_{k+1,c,j}^{(1)})' = G_{k+1,c,j}^{(1)}$. Now assume that $G_{k,c,j}^{(t)} \subseteq G_{k+1,c,j}^{(t)}$, $j \in \{0\} \cup J_c, c \in [p]$, holds for some $t \geq 1$. Then by (3.2.2)-(3.2.3) and by the induction hypothesis, we have $F_{k,c,j}^{(t+1)} \subseteq F_{k+1,c,j}^{(t+1)}$ for $j \in \{0\} \cup J_c, c \in [p]$. Thus $G_{k,c,j}^{(t+1)} = (F_{k,c,j}^{(t+1)})' \subseteq (F_{k+1,c,j}^{(t+1)})' = G_{k+1,c,j}^{(t+1)}$ which completes the induction. □

From Proposition 3.1 and Proposition 3.2, we deduce the following two-level hierarchy of lower bounds for the optimum $f^\star$ of (3.0.1) (3.0.1):

$$\begin{array}{ccccccc} \rho_{k_{\min}}^{(1)} & \leq & \rho_{k_{\min}}^{(2)} & \leq & \cdots & \leq & \rho_{k_{\min}} \\ \wedge| & & \wedge| & & & & \wedge| \\ \rho_{k_{\min}+1}^{(1)} & \leq & \rho_{k_{\min}+1}^{(2)} & \leq & \cdots & \leq & \rho_{k_{\min}+1} \\ \wedge| & & \wedge| & & & & \wedge| \\ \vdots & & \vdots & & \vdots & & \vdots \\ \wedge| & & \wedge| & & & & \wedge| \\ \rho_k^{(1)} & \leq & \rho_k^{(2)} & \leq & \cdots & \leq & \rho_k \\ \wedge| & & \wedge| & & & & \wedge| \\ \vdots & & \vdots & & \vdots & & \vdots \end{array} \tag{3.2.6}$$

The array of lower bounds (3.2.6) (and its associated SDP relaxations (3.2.5)) is what we call the *CS-TSSOS* hierarchy associated with (3.0.1) (3.0.1).

Example 3.3. *Let* $f = 1+\sum_{i=1}^{6} x_i^4 + x_1x_2x_3 + x_3x_4x_5 + x_3x_4x_6 + x_3x_5x_6 + x_4x_5x_6$*, and consider the unconstrained POP:* $\min\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n\}$*. We have* $n = 6, m = 0$*. Let us apply the CS-TSSOS hierarchy (using the maximal chordal extension in* (3.2.4)*) to this problem by taking the relaxation order* $k = k_{\min} = 2$ *and the sparse order* $t = 1$*. First, according to the csp graph (see Figure 3.4), we decomposes the variables into two cliques:* $\{x_1, x_2, x_3\}$ *and* $\{x_3, x_4, x_5, x_6\}$*. Figure 3.5 and Figure 3.6 illustrate the tsp graphs for the first clique and the second clique, respectively. For the first clique one obtains four blocks of SDP matrices with respective sizes* $4, 2, 2, 2$*. For the second clique one obtains two blocks of SDP matrices with respective sizes* $5, 10$*. As a result, the original SDP matrix of size* 28 *has been reduced to six blocks of maximal size* 10*.*

If one applies the TSSOS hierarchy (using the maximal chordal extension in (3.1.7)*) directly to the problem by taking* $k = k_{\min} = 2, t = 1$ *(i.e., without decomposing variables), then the tsp graph is illustrated in Figure 3.7. One obtains* 11 *SDP blocks with respective sizes* $7, 2, 2, 2, 1, 1, 1, 1, 1, 1, 10$*. Compared to the CS-TSSOS case, there are six additional blocks of size one and the two blocks with respective sizes* $4, 5$ *are replaced by a single block of size* 7*.*

Figure 3.4: The csp graph of Example 3.3

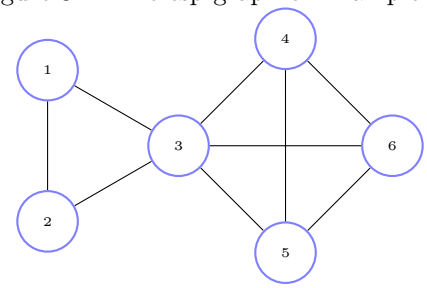


Figure 3.5: The tsp graph for the first clique of Example 3.3

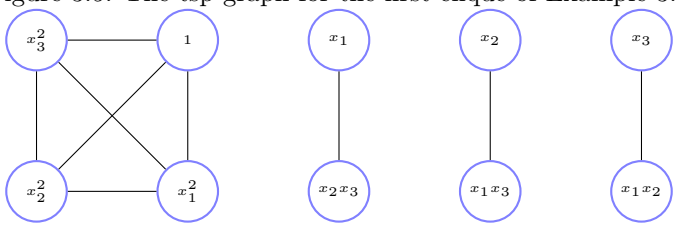
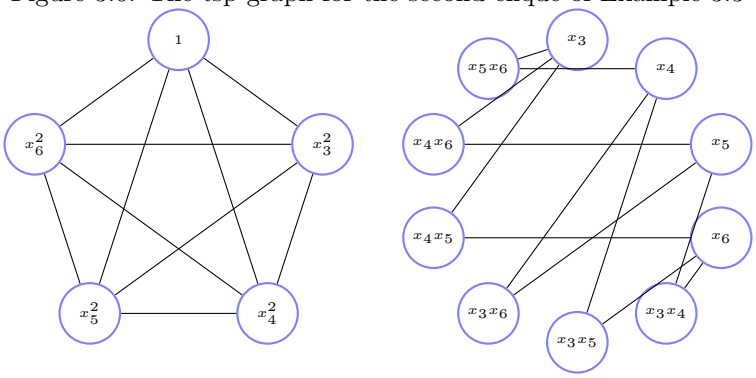
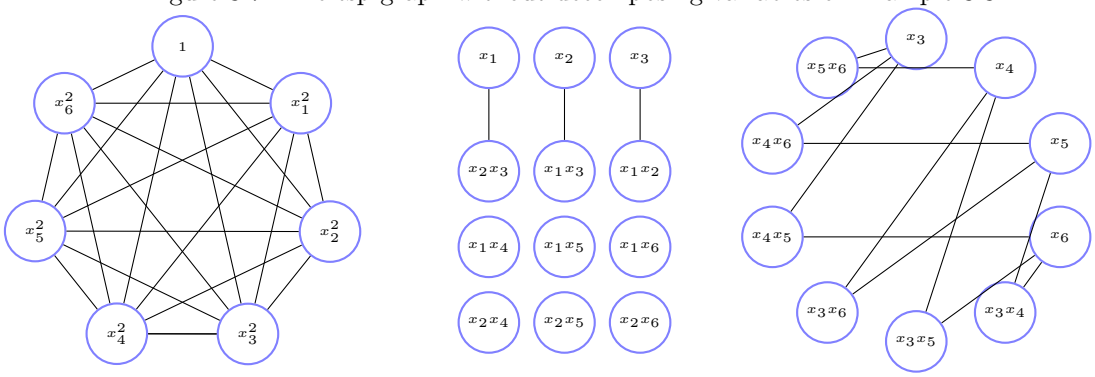


Figure 3.6: The tsp graph for the second clique of Example 3.3



The CS-TSSOS hierarchy entails a trade-off. Indeed, one has the freedom to choose a specific chordal extension for any graph involved in (3.2.5). This choice affects the resulting size of blocks of SDP matrices and the quality of optimal values of corresponding relaxations. Intuitively, chordal extensions with small clique numbers lead to blocks of small size and optimal values of (possibly)

Figure 3.7: The tsp graph without decomposing variables of Example 3.3



low quality while chordal extensions with large clique numbers lead to blocks of large size and optimal values of (possibly) high quality.

For all c, i, write $\mathbf{M}_{k-\lceil g_i \rceil}(g_i\mathbf{y}, I_c) = \sum_{\boldsymbol{\alpha}} \mathbf{D}_{\boldsymbol{\alpha}}^{c,i} y_{\boldsymbol{\alpha}}$ for appropriate symmetry matrices $\{\mathbf{D}_{\boldsymbol{\alpha}}^{c,i}\}$. Then for each $k \geq 1$, the dual of (3.2.5) reads as:

$$\begin{aligned} \sup \quad & \rho \\ \text{s.t.} \quad & \sum_{c=1}^{p} \sum_{j\in\{0\}\cup J_c} \langle \mathbf{G}_{c,j}, \mathbf{D}_{\boldsymbol{\alpha}}^{c,j} \rangle + \rho\delta_{\mathbf{0}\boldsymbol{\alpha}} = f_{\boldsymbol{\alpha}}, \quad \forall \boldsymbol{\alpha} \in \mathcal{C}_k^{(t)}, \\ & \mathbf{G}_{c,j} \in \mathcal{S}_+^{r_{c,j}} \cap \mathcal{S}_{G_{k,c,j}^{(t)}}, \quad j \in \{0\} \cup J_c\,,\, c \in [p], \end{aligned} \tag{3.2.7}$$

where $\mathcal{C}_k^{(t)}$ is defined in (3.2.3).

Proposition 3.3. *Let $f \in \mathbb{R}[\mathbf{x}]$ and $S(\mathfrak{g})$ be as in (3.0.2). Assume that $S(\mathfrak{g})$ has a nonempty interior. Then there is no duality gap between (3.2.5) and (3.2.7) for any $k \geq k_{\min}$ and $t \geq 1$.*

Proof. By the duality theory of convex programming, this easily follows from Theorem 3.6 of [103] and Theorem 3.2. □

Note that the number of equality constraints in (3.2.7) is equal to the cardinality of $\mathcal{C}_k^{(t)}$. We next give a description of the elements in $\mathcal{C}_k^{(t)}$ in terms of sign symmetries.

3.2.2 Sign symmetries

Definition 3.2. *Given a finite set $\mathcal{A} \subseteq \mathbb{N}^n$, the* sign symmetries *of $\mathcal{A}$ are defined by all vectors $\mathbf{r} \in \mathbb{Z}_2^n := \{0,1\}^n$ such that $\mathbf{r}^\top \boldsymbol{\alpha} \equiv 0 \pmod 2$ for all $\boldsymbol{\alpha} \in \mathcal{A}$.*

For any $\boldsymbol{\alpha} \in \mathbb{N}^n$, we define $(\boldsymbol{\alpha})_2 := (\alpha_1 (\mathrm{mod}\ 2), \ldots, \alpha_n (\mathrm{mod}\ 2)) \in \mathbb{Z}_2^n$. We also use the same notation for any subset $\mathcal{A} \subseteq \mathbb{N}^n$, i.e., $(\mathcal{A})_2 := \{(\boldsymbol{\alpha})_2 \mid \boldsymbol{\alpha} \in \mathcal{A}\} \subseteq \mathbb{Z}_2^n$. For a subset $S \subseteq \mathbb{Z}_2^n$, the *orthogonal complement space* of S in $\mathbb{Z}_2^n$, denoted by $S^\perp$, is the set $\{\boldsymbol{\alpha} \in \mathbb{Z}_2^n \mid \boldsymbol{\alpha}^\top \mathbf{s} \equiv 0 \,(\mathrm{mod}\ 2)\,, \forall \mathbf{s} \in S\}$.

Remark 3.4. *By definition, the set of sign symmetries of $\mathcal{A}$ is exactly the orthogonal complement space $(\boldsymbol{\alpha})_2^\perp$ in $\mathbb{Z}_2^n$, which therefore can be essentially represented by a basis of the subspace $(\mathcal{A})_2^\perp$ in $\mathbb{Z}_2^n$.*

For a subset $S \subseteq \mathbb{Z}_2^n$, we say that S is *closed* under addition modulo 2 if $\mathbf{s}_1, \mathbf{s}_2 \in S$ implies $(\mathbf{s}_1 + \mathbf{s}_2)_2 \in S$. The minimal set containing S with elements which are closed under addition modulo 2 is denoted by $\langle S \rangle_{\mathbb{Z}_2}$. It is easy to prove $\langle S \rangle_{\mathbb{Z}_2} = \{(\sum_i \mathbf{s}_i)_2 \mid \mathbf{s}_i \in S\}$ which is the subspace spanned by S in $\mathbb{Z}_2^n$.

Lemma 3.1. *Let $S \subseteq \mathbb{Z}_2^n$. Then $(S^\perp)^\perp = \langle S \rangle_{\mathbb{Z}_2}$.*

Proof. It is immediate from the definitions. □

Lemma 3.2. *Suppose G is a graph with $V(G) \subseteq \mathbb{N}^n$. Then it holds $(\mathrm{supp}(G'))_2 \subseteq \langle (\mathrm{supp}(G))_2 \rangle_{\mathbb{Z}_2}$.*

Proof. By definition, we need to show $(\boldsymbol{\beta}+\boldsymbol{\gamma})_2 \in \langle(\mathrm{supp}(G))_2\rangle_{\mathbb{Z}_2}$ for any $\{\boldsymbol{\beta},\boldsymbol{\gamma}\} \in E(G')$. Since in the process of chordal extensions, edges are added only if two nodes belong to the same connected component, for any $\{\boldsymbol{\beta},\boldsymbol{\gamma}\} \in E(G')$ there is a path connecting $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in G: $\{\boldsymbol{\beta},\boldsymbol{v}_1,\ldots,\boldsymbol{v}_r,\boldsymbol{\gamma}\}$ with $\{\boldsymbol{\beta},\boldsymbol{v}_1\},\{\boldsymbol{v}_r,\boldsymbol{\gamma}\} \in E(G)$ and $\{\boldsymbol{v}_i,\boldsymbol{v}_{i+1}\} \in E(G), i \in [r-1]$. From $(\boldsymbol{\beta}+\boldsymbol{v}_1)_2, (\boldsymbol{v}_1+\boldsymbol{v}_2)_2 \in (\mathrm{supp}(G))_2$, we deduce that $(\boldsymbol{\beta}+\boldsymbol{v}_2)_2 \in \langle(\mathrm{supp}(G))_2\rangle_{\mathbb{Z}_2}$ because $\langle(\mathrm{supp}(G))_2\rangle_{\mathbb{Z}_2}$ is closed under addition modulo 2. Likewise, we can prove $(\boldsymbol{\beta}+\boldsymbol{v}_i)_2 \in \langle(\mathrm{supp}(G))_2\rangle_{\mathbb{Z}_2}$ for $i = 3,\ldots,r+1$ with $\boldsymbol{v}_{r+1} := \boldsymbol{\gamma}$. Hence $(\boldsymbol{\beta}+\boldsymbol{\gamma})_2 \in \langle(\mathrm{supp}(G))_2\rangle_{\mathbb{Z}_2}$ as desired. □

Proposition 3.4. *Let $\mathcal{A}$ be defined as in* (3.2.1)*, $\mathcal{C}_k^{(t)}$ be defined as in* (3.2.3) *and assume that the sign symmetries of $\mathcal{A}$ are represented by the column vectors of a binary matrix, denoted by $\mathbf{R}$. Then for any $k \geq 1$ and any $\boldsymbol{\alpha} \in \mathcal{C}_k^{(t)}$, it holds $\mathbf{R}^\top\boldsymbol{\alpha} \equiv 0\,(mod\ 2)$. In other words, $(\mathcal{C}_k^{(t)})_2 \subseteq R^\perp$, where we regard R as a set of the column vectors of $\mathbf{R}$.*

Proof. By Lemma 3.1, we only need to prove $(\mathcal{C}_k^{(t)})_2 \subseteq \langle(\mathcal{A})_2\rangle_{\mathbb{Z}_2}$. Let us do induction on $t \geq 0$. For $t = 0$, by (3.2.3), $\mathcal{C}_k^{(0)} = \bigcup_{c=1}^p \mathrm{supp}(G_{k,c,0}^{(0)}) = \bigcup_{c=1}^p \mathrm{supp}(G_{k,c}^{\mathrm{tsp}}) \subseteq \bigcup_{c=1}^p (\mathcal{A}_c \cup (2\mathbb{N})^{n_c}) \subseteq \mathcal{A} \cup (2\mathbb{N})^n$. Hence $(\mathcal{C}_k^{(0)})_2 \subseteq \langle(\mathcal{A})_2\rangle_{\mathbb{Z}_2}$. Now assume that $(\mathcal{C}_k^{(t)})_2 \subseteq \langle(\mathcal{A})_2\rangle_{\mathbb{Z}_2}$ holds for some $t \geq 0$. By (3.2.2), for any c,i and any $\{\boldsymbol{\beta},\boldsymbol{\gamma}\} \in E(F_{k,c,i}^{(t+1)})$, we have $(\mathrm{supp}(g_i)+\boldsymbol{\beta}+\boldsymbol{\gamma}) \cap \mathcal{C}_k^{(t)} \neq \emptyset$, i.e., there exists $\boldsymbol{\alpha} \in \mathrm{supp}(g_i)$ such that $\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\gamma} \in \mathcal{C}_k^{(t)}$, which implies $(\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\gamma})_2 \in (\mathcal{C}_k^{(t)})_2$. Hence by the induction hypothesis, $(\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\gamma})_2 \in \langle(\mathcal{A})_2\rangle_{\mathbb{Z}_2}$. Since $\langle(\mathcal{A})_2\rangle_{\mathbb{Z}_2}$ is closed under addition modulo 2 and $(\boldsymbol{\alpha})_2 \in (\mathcal{A})_2$, we have $(\boldsymbol{\beta}+\boldsymbol{\gamma})_2 \in \langle(\mathcal{A})_2\rangle_{\mathbb{Z}_2}$. It follows $(\mathrm{supp}(F_{k,c,j}^{(t+1)}))_2 \subseteq \langle(\mathcal{A})_2\rangle_{\mathbb{Z}_2}$. Because $G_{k,c,j}^{(t+1)} = (F_{k,c,j}^{(t+1)})'$, by Lemma 3.2, we have $(\mathrm{supp}(G_{k,c,j}^{(t+1)}))_2 \subseteq \langle(\mathrm{supp}(F_{k,c,j}^{(t+1)}))_2\rangle_{\mathbb{Z}_2} \subseteq \langle(\mathcal{A})_2\rangle_{\mathbb{Z}_2}$. From this, (3.2.3) and the fact that $\langle(\mathcal{A})_2\rangle_{\mathbb{Z}_2}$ is closed under addition modulo 2, we then deduce the inclusion $(\mathcal{C}_k^{(t+1)})_2 \subseteq \langle(\mathcal{A})_2\rangle_{\mathbb{Z}_2}$ which completes the induction. □

Remark 3.5. *Proposition 3.4 actually indicates that the block structure produced by the CS-TSSOS hierarchy is consistent with the sign symmetries of the POP.*

3.3 Convergence analysis

3.3.1 Global convergence

We next prove that if for any graph involved in (3.2.5), the chordal extension is chosen to be *maximal*, then for any relaxation order $k \geq k_{\min}$ the sequence of optimal values $(\rho_k^{(t)})_{t\geq 1}$ of the CS-TSSOS hierarchy converges to the optimal value ρ_k of the corresponding CSSOS hierarchy (3.1.4). In turn, as the relaxation order k increases, the latter sequence converges to the global optimum $f^\star$ of the original POP (3.0.1) (after adding some redundant quadratic constraints) as shown in [103].

Obviously, the sequences of graphs $(G_{k,c,j}^{(t)}(V_{k,c,j}, E_{k,c,j}^{(t)}))_{t\geq 1}$ stabilize for all $j \in \{0\}\cup J_c, c \in [p]$ after finitely many steps. We denote the resulting stabilized graphs by $G_{k,c,j}^{(*)}, j \in \{0\}\cup J_c, c \in [p]$ and the corresponding SDP (3.2.5) with value ρ_k^*.

Theorem 3.3. *Assume that the chordal extension in* (3.2.4) *is the maximal chordal extension. Then for any $k \geq k_{\min}$, the sequence $(\rho_k^{(t)})_{t\geq 1}$ converges to ρ_k in finitely many steps.*

Proof. Let $\mathbf{y} = (y_{\boldsymbol{\alpha}})$ be an arbitrary feasible solution of (3.2.5) with value ρ_k^*. Note that $\{y_{\boldsymbol{\alpha}} \mid \boldsymbol{\alpha} \in \bigcup_{c=1}^p (\cup_{i\in\{0\}\cup J_c}(\mathrm{supp}(g_i)+\mathrm{supp}(G_{k,c,i}^{(*)})))\}$ is the set of decision variables involved in $(\mathrm{Q}_{k,*}^{\text{cs-ts}})$. Let $\mathcal{R}$ be the set of decision variables involved in (3.1.4). We then define a vector $\overline{\mathbf{y}} = (\overline{y}_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathcal{R}}$ as follows:

$$\overline{y}_{\boldsymbol{\alpha}} = \begin{cases} y_{\boldsymbol{\alpha}}, & \text{if } \boldsymbol{\alpha} \in \bigcup_{c=1}^p (\cup_{i\in\{0\}\cup J_c}(\mathrm{supp}(g_i)+\mathrm{supp}(G_{k,c,i}^{(*)}))), \\ 0, & \text{otherwise.} \end{cases}$$

By construction and since $G_{k,c,i}^{(*)}$ stabilizes under support extension for all c,i, we have $\mathbf{M}_{k-\lceil g_i\rceil}(g_i\overline{\mathbf{y}}, I_c) = \mathbf{B}_{G_{k,c,i}^{(*)}} \circ \mathbf{M}_{k-\lceil g_i\rceil}(g_i\mathbf{y}, I_c)$. As we use the maximal chordal extension in (3.2.4), the matrix $\mathbf{B}_{G_{k,c,i}^{(*)}} \circ \mathbf{M}_{k-\lceil g_i\rceil}(g_i\mathbf{y}, I_c)$ is block diagonal up to permutation (see Remark 3.1). So from $\mathbf{B}_{G_{k,c,i}^{(*)}} \circ \mathbf{M}_{k-\lceil g_i\rceil}(g_i\mathbf{y}, I_c) \in$

$\Pi_{G^{(*)}_{k,c,i}}(\mathcal{S}_+^{r_{c,i}})$ it follows $\mathbf{M}_{k-\lceil g_i\rceil}(g_i\overline{\mathbf{y}}, I_c) \succeq 0$ for $i \in \{0\}\cup J_c, c \in [p]$. Therefore $\overline{\mathbf{y}}$ is a feasible solution of (3.1.4) and so $L_{\mathbf{y}}(f) = L_{\overline{\mathbf{y}}}(f) \geq \rho_k$. Hence $\rho_k^* \geq \rho_k$ since $\mathbf{y}$ is an arbitrary feasible solution of (3.2.5) with value ρ_k^*. By Proposition 3.1, we already have $\rho_k^* \leq \rho_k$. Therefore, $\rho_k^* = \rho_k$. □

To guarantee the global optimality, we need the following compactness assumption on the feasible set $S(\mathfrak{g})$.

Assumption 1. Let $S(\mathfrak{g})$ be as in (3.0.2). There exists an $M > 0$ such that $\|\mathbf{x}\|_\infty < M$ for all $x \in S(\mathfrak{g})$.

Because of Assumption 1, one has $\|\mathbf{x}(I_c)\|_2^2 \leq n_c M^2$, $c \in [p]$. Therefore, we can add the p redundant quadratic constraints

$$g_{m+c}(\mathbf{x}) := n_c M^2 - \|\mathbf{x}(I_c)\|_2^2 \geq 0, \quad c \in [p] \tag{3.3.1}$$

in the definition (3.0.2) of $S(\mathfrak{g})$ and set $m' = m + p$, so that $S(\mathfrak{g})$ is now defined by

$$S(\mathfrak{g}) := \{\mathbf{x} \in \mathbb{R}^n \mid g_i(\mathbf{x}) \geq 0, \quad i \in [m']\}. \tag{3.3.2}$$

Note that $g_{m+c} \in \mathbb{R}[\mathbf{x}(I_c)]$ for $c \in [p]$.

Then by Theorem 3.6 in [103], the sequence $(\rho_k)_{k\geq k_{\min}}$ converges to the globally optimal value $f^\star$ of POP (3.0.1). So this together with Theorem 3.3 gives the global convergence of the CS-TSSOS hierarchy.

3.3.2 A sparse representation theorem

Proceeding along Theorem 3.3, we are able to provide a *sparse representation* theorem for a polynomial positive on a compact basic semialgebraic set.

Theorem 3.4 (sparse representation)**.** *Let $f \in \mathbb{R}[\mathbf{x}]$ and $S(\mathfrak{g})$ be as in* (3.3.2) *with the additional quadratic constraints* (3.3.1)*. Let I_c, J_c be defined as in Section 3.2.1 and $\mathcal{A} = \operatorname{supp}(f) \cup \bigcup_{i=1}^{m'} \operatorname{supp}(g_i)$. Assume that the sign symmetries of $\mathcal{A}$ are represented by the column vectors of the binary matrix $\mathbf{R}$. If f is positive on $S(\mathfrak{g})$, then*

$$f = \sum_{c=1}^{p}\left(\sigma_{c,0} + \sum_{i\in J_c}\sigma_{c,i}g_i\right), \tag{3.3.3}$$

for some polynomials $\sigma_{c,i} \in \Sigma[\mathbf{x}(I_c)], i \in \{0\}\cup J_c, c \in [p]$, satisfying $\mathbf{R}^\top\boldsymbol{\alpha} \equiv 0 \pmod 2$ for any $\boldsymbol{\alpha} \in \operatorname{supp}(\sigma_{c,i})$, i.e., $(\operatorname{supp}(\sigma_{c,i}))_2 \subseteq R^\perp$, where we regard R as a set of its column vectors.

That is, (3.3.3) *provides a certificate of positivity of f on $S(\mathfrak{g})$.*

Proof. By Corollary 3.9 of [103] (or Theorem 5 of [68]), there exist polynomials $\sigma'_{c,i} \in \Sigma[\mathbf{x}(I_c)], i \in \{0\}\cup J_c, c \in [p]$ such that

$$f = \sum_{c=1}^{p}\left(\sigma'_{c,0} + \sum_{i\in J_c}\sigma'_{c,i}g_i\right). \tag{3.3.4}$$

Let $k = \max\{\lceil \deg(\sigma'_{c,i}g_i)/2\rceil : i \in \{0\}\cup J_c, c \in [p]\}$. Let $\mathbf{G}'_{c,i}$ be a PSD Gram matrix associated with $\sigma'_{c,i}$ and indexed by the monomial basis $\mathbb{N}^{n_c}_{k-\lceil g_i\rceil}$. Then for all c, i, we define $\mathbf{G}_{c,i} \in \mathcal{S}^{r_{c,i}}$ with $r_{c,i} = b(n_c, k - \lceil g_i\rceil)$ (indexed by $\mathbb{N}^{n_c}_{k-\lceil g_i\rceil}$) by

$$[\mathbf{G}_{c,i}]_{\boldsymbol{\beta}\boldsymbol{\gamma}} := \begin{cases} [\mathbf{G}'_{c,i}]_{\boldsymbol{\beta}\boldsymbol{\gamma}}, & \text{if } \mathbf{R}^\top(\boldsymbol{\beta}+\boldsymbol{\gamma}) \equiv 0 \pmod 2, \\ 0, & \text{otherwise,}\end{cases}$$

and let $\sigma_{c,i} = (\mathbf{x}^{\mathbb{N}^{n_c}_{k-\lceil g_i\rceil}})^\top \mathbf{G}_{c,i}\mathbf{x}^{\mathbb{N}^{n_c}_{k-\lceil g_i\rceil}}$. One can easily verify that $\mathbf{G}_{c,i}$ is block diagonal up to permutation (see also [212]) and each block is a principal submatrix of $\mathbf{G}'_{c,i}$. Then the positive semidefiniteness of $\mathbf{G}'_{c,i}$ implies that $\mathbf{G}_{c,i}$ is also positive semidefinite. Thus $\sigma_{c,i} \in \Sigma[\mathbf{x}(I_c)]$.

By construction, substituting $\sigma'_{c,j}$ with $\sigma_{c,j}$ in (3.3.4) boils down to removing the terms with exponents $\boldsymbol{\alpha}$ that do not satisfy $\mathbf{R}^\top\boldsymbol{\alpha} \equiv 0 \pmod 2$ from the right hand side of (3.3.4). Since any

$\boldsymbol{\alpha} \in \mathrm{supp}(f)$ satisfies $\mathbf{R}^\top \boldsymbol{\alpha} \equiv 0 \pmod 2$, this does not change the match of coefficients on both sides of the equality. Thus we obtain

$$f = \sum_{c=1}^{p} \left(\sigma_{c,0} + \sum_{i \in J_c} \sigma_{c,i} g_i \right)$$

with the desired property. □

3.3.3 Extracting a solution

In the case of dense Moment-SOS relaxations, there is a standard procedure described in [77] to extract globally optimal solutions when the so-called flatness condition for the moment matrix is satisfied. This procedure was partially generalized to the correlative sparsity setting in [103, § 3.3]. However, in the combined sparsity setting, the corresponding procedure cannot be applied because we do not have complete information on the moment matrix associated with each clique. In order to extract a solution in this case, we may add a dense moment matrix of order one for each clique in (3.2.5):

$$\begin{aligned} \inf \quad & L_{\mathbf{y}}(f) \\ \text{s.t.} \quad & \mathbf{B}_{G^{(t)}_{k,c,0}} \circ \mathbf{M}_k(\mathbf{y}, I_c) \in \Pi_{G^{(t)}_{k,c,0}} (\mathcal{S}_+^{r_{c,0}}), \quad c \in [p], \\ & \mathbf{M}_1(y, I_c) \succeq 0, \quad c \in [p], \\ & \mathbf{B}_{G^{(t)}_{k,c,i}} \circ \mathbf{M}_{k-\lceil g_i \rceil}(g_i \mathbf{y}, I_c) \in \Pi_{G^{(t)}_{k,c,i}} (\mathcal{S}_+^{r_{c,i}}), \quad i \in J_c\,,\, c \in [p], \\ & y_{\mathbf{0}} = 1. \end{aligned} \tag{3.3.5}$$

Let $\mathbf{y}^\star$ be an optimal solution of (3.3.5). Typically, $\mathbf{M}_1(\mathbf{y}^\star, I_c)$ (after identifying sufficiently small entries with zeros) is a block diagonal matrix (up to permutation). If for all c, every block of $\mathbf{M}_1(\mathbf{y}^\star, I_c))$ is of rank one, then a globally optimal solution $\mathbf{x}^\star$ to (3.0.1) which is unique up to sign symmetries can be extracted, and the global optimality is certified (see [103, Theorem 3.2]). Otherwise, the relaxation might be not exact or yield multiple global solutions.

Remark 3.6. *Note that* (3.3.5) *is a tighter relaxation of* (3.0.1) *than* (3.2.5) *and so might provide a better lower bound for* (3.0.1).

3.4 Applications and numerical experiments

In this section, we conduct numerical experiments for the proposed CS-TSSOS hierarchy and apply it to two important classes of POPs: Max-Cut problems and AC optimal power flow (AC-OPF) problems. Depending on specific problems, we consider two types of chordal extensions for the term sparsity pattern: maximal chordal extensions and approximately smallest chordal extensions[4]. The tool `TSSOS` which executes the CS-TSSOS hierarchy (as well as the CSSOS hierarchy and the TSSOS hierarchy) is implemented in Julia. For an introduction to `TSSOS`, one could refer to [124]. `TSSOS` is available on the website:

`https://github.com/wangjie212/TSSOS`.

In the following subsections, we compare the performances of the CSSOS approach, the TSSOS approach, the CS-TSSOS approach and the SDSOS approach [3] (implemented in `SPOT` [140]). `Mosek` [6] is used as an SDP (in the CSSOS, TSSOS, CS-TSSOS cases) or SOCP (in the SDSOS case) solver. All numerical examples were computed on an Intel Core i5-8265U@1.60GHz CPU with 8GB RAM memory. The timing includes the time required to generate the SDP/SOCP and the time spent to solve it. The notations used in this section are listed in Table 3.1.

[4] A smallest chordal extension is a chordal extension with the smallest clique number. Computing a smallest chordal extension is generally NP-complete. So in practice we compute approximately smallest chordal extensions instead with efficient heuristic algorithms.

Table 3.1: Notation

var	number of variables
cons	number of constraints
mc	maximal size of variable cliques
mb	maximal size of SDP blocks
opt	optimal value
time	running time in seconds
gap	optimality gap
CE	type of chordal extensions used in (3.2.4)
min	approximately smallest chordal extension
max	maximal chordal extension
0	a number whose absolute value less than 1e-5
-	an out of memory error

3.4.1 Benchmarks for unconstrained POPs

The Broyden banded function is defined as

$$f_{\mathrm{Bb}}(\mathbf{x}) = \sum_{i=1}^{n}(x_i(2+5x_i^2)+1-\sum_{j\in J_i}(1+x_j)x_j)^2,$$

where $J_i = \{j \mid j \neq i, \max(1,i-5) \leq j \leq \min(n,i+1)\}$.

The task is to minimize the Broyden banded function over $\mathbb{R}^n$ which is formulated as an unconstrained POP. Using the relaxation order $k = 3$, we solve the CSSOS hierarchy (3.1.4), the TSSOS hierarchy (3.1.8) with $t = 1$ and the CS-TSSOS hierarchy (3.2.5) with $t = 1$. In the latter two cases, approximately smallest chordal extensions are used. We also solve the POP with the SDSOS approach. The results are displayed in Table 3.2.

It can be seen from the table that CS-TSSOS significantly reduces the maximal size of SDP blocks and is the most efficient approach. CSSOS, TSSOS and CS-TSSOS all give the exact minimum 0 while SDSOS only gives a very loose lower bound -13731 when $n = 20$. Due to the limitation of memory, CSSOS scales up to 180 variables; TSSOS scales up to 40 variables; SDSOS scales up to 20 variables. On the other hand, CS-TSSOS can easily handle instances with up to 500 variables.

Table 3.2: The result for Broyden banded functions ($k = 3$)

var	CSSOS			TSSOS			CS-TSSOS			SDSOS	
	mb	opt	time	mb	opt	time	mb	opt	time	opt	time
20	120	0	21.7	33	0	4.39	19	0	2.24	-13731	374
40	120	0	44.6	52	0	231	19	0	6.95	-	-
60	120	0	81.8	-	-	-	19	0	13.0	-	-
80	120	0	116	-	-	-	19	0	19.6	-	-
100	120	0	151	-	-	-	19	0	27.0	-	-
120	120	0	195	-	-	-	19	0	34.4	-	-
140	120	0	249	-	-	-	19	0	43.1	-	-
160	120	0	298	-	-	-	19	0	50.2	-	-
180	120	0	338	-	-	-	19	0	63.8	-	-
200	120	-	-	-	-	-	19	0	72.9	-	-
250	120	-	-	-	-	-	19	0	106	-	-
300	120	-	-	-	-	-	19	0	132	-	-
400	120	-	-	-	-	-	19	0	220	-	-
500	120	-	-	-	-	-	19	0	313	-	-

3.4.2 Benchmarks for constrained POPs

- The generalized Rosenbrock function

$$f_{\mathrm{gR}}(\mathbf{x}) = 1 + \sum_{i=2}^{n}(100(x_i - x_{i-1}^2)^2 + (1-x_i)^2).$$

- The Broyden tridiagonal function

$$\begin{aligned} f_{\mathrm{Bt}}(\mathbf{x}) =&((3-2x_1)x_1 - 2x_2 + 1)^2 + \sum_{i=2}^{n-1}((3-2x_i)x_i - x_{i-1} - 2x_{i+1} + 1)^2 \\ &+ ((3-2x_n)x_n - x_{n-1} + 1)^2. \end{aligned}$$

- The chained Wood function

$$\begin{aligned} f_{\mathrm{cW}}(\mathbf{x}) =&1 + \sum_{i\in J}(100(x_{i+1} - x_i^2)^2 + (1-x_i)^2 + 90(x_{i+3} - x_{i+2}^2)^2 \\ &+ (1-x_{i+2})^2 + 10(x_{i+1} + x_{i+3} - 2)^2 + 0.1(x_{i+1} - x_{i+3})^2), \end{aligned}$$

where $J = \{1, 3, 5, \ldots, n-3\}$ and $4|n$.

With the generalized Rosenbrock (resp. Broyden tridiagonal or chained Wood) function as the objective function, we consider the following constrained POP:

$$\begin{cases} \inf & f_{\mathrm{gR}} \quad (\text{resp. } f_{\mathrm{Bt}} \text{ or } f_{\mathrm{cW}}) \\ \text{s.t.} & 1 - (\sum_{i=20j-19}^{20j} x_i^2) \geq 0, \quad j \in [n/20], \end{cases} \tag{3.4.1}$$

where $20|n$. The generalized Rosenbrock function, the Broyden tridiagonal function and the chained Wood function involve cliques with 2 or 3 variables, which can be efficiently handled by the CSSOS hierarchy; see [203]. For them, the CS-TSSOS hierarchy gives almost the same results with the CSSOS hierarchy. Hence we add the sphere constraints in (3.4.1) to increase the clique size and to show the difference.

For these problems, the minimum relaxation order $k = 2$ is used. As in the unconstrained case, we solve the CSSOS hierarchy (3.1.4), the TSSOS hierarchy (3.1.8) with $t = 1$ and the CS-TSSOS hierarchy (3.2.5) with $t = 1$, and use approximately smallest chordal extensions. We also solve these POPs with the SDSOS approach. The results are displayed in Tables 3.3–3.5.

From these tables, one can see that CS-TSSOS significantly reduces the maximal size of SDP blocks and is again the most efficient approach. For the generalized Rosenbrock function, CSSOS, TSSOS and CS-TSSOS give almost the same optimum while SDSOS gives a slightly loose lower bound (only for $n = 40$); for the Broyden tridiagonal function, CSSOS, TSSOS and CS-TSSOS all give the same optimum while SDSOS gives a very loose lower bound (only for $n = 40$); for the chained Wood function, CSSOS, TSSOS and CS-TSSOS all give the same optimum while SDSOS gives a slightly loose lower bound (only for $n = 40$). Due to the limitation of memory, CSSOS scales up to 180 variables; TSSOS scales up to 180 or 200 variables; SDSOS scales up to 40 variables. On the other hand, CS-TSSOS can easily handle these instances with up to 1000 variables.

3.4.3 The Max-Cut problem

The Max-Cut problem is one of the basic combinatorial optimization problems, which is known to be NP-hard. Let $G(V, E)$ be an undirected graph with $V = [n]$ and with edge weights w_{ij} for $\{i, j\} \in E$. Then the Max-Cut problem for G can be formulated as a QCQP in binary variables:

$$\begin{cases} \max & \frac{1}{2}\sum_{\{i,j\}\in E} w_{ij}(1 - x_i x_j) \\ \text{s.t.} & 1 - x_i^2 = 0, \quad i \in [n]. \end{cases} \tag{3.4.2}$$

The property of binary variables in (3.4.2) can be also exploited to reduce the size of SDPs arising from the Moment-SOS hierarchy, which has been implemented in `TSSOS`.

Table 3.3: The result for the generalized Rosenbrock function ($k = 2$)

var	CSSOS			TSSOS			CS-TSSOS			SDSOS	
	mb	opt	time	mb	opt	time	mb	opt	time	opt	time
40	231	38.051	126	41	38.049	0.61	21	38.049	0.23	37.625	115
60	231	57.849	232	61	57.845	3.31	21	57.845	0.32	-	-
80	231	77.647	306	81	77.641	11.7	21	77.641	0.41	-	-
100	231	97.445	377	101	97.436	31.3	21	97.436	0.54	-	-
120	231	117.24	408	121	117.23	75.4	21	117.23	0.60	-	-
140	231	137.04	495	141	137.03	190	21	137.03	0.75	-	-
160	231	156.84	570	161	156.82	367	21	156.82	0.90	-	-
180	231	176.64	730	181	176.62	628	21	176.62	1.09	-	-
200	231	-	-	201	196.41	1327	21	196.41	1.27	-	-
300	231	-	-	-	-	-	21	295.39	2.26	-	-
400	231	-	-	-	-	-	21	394.37	3.36	-	-
500	231	-	-	-	-	-	21	493.35	4.65	-	-
1000	231	-	-	-	-	-	21	988.24	15.8	-	-

Table 3.4: The result for the Broyden tridiagonal function ($k = 2$)

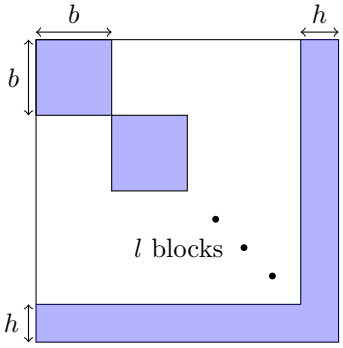
var	CSSOS			TSSOS			CS-TSSOS			SDSOS	
	mb	opt	time	mb	opt	time	mb	opt	time	opt	time
40	231	31.234	168	43	31.234	1.95	23	31.234	0.64	-5.8110	138
60	231	47.434	273	63	47.434	8.33	23	47.434	1.14	-	-
80	231	63.634	413	83	63.634	33.9	23	63.634	1.50	-	-
100	231	79.834	519	103	79.834	104	23	79.834	1.96	-	-
120	231	96.034	671	123	96.034	199	23	96.034	2.30	-	-
140	231	112.23	872	143	112.23	490	23	112.23	2.94	-	-
160	231	128.43	1002	163	128.43	783	23	128.43	3.67	-	-
180	231	144.63	1066	183	144.63	1329	23	144.63	4.46	-	-
200	231	-	-	-	-	-	23	160.83	4.88	-	-
300	231	-	-	-	-	-	23	241.83	8.67	-	-
400	231	-	-	-	-	-	23	322.83	13.3	-	-
500	231	-	-	-	-	-	23	403.83	19.9	-	-
1000	231	-	-	-	-	-	23	808.83	57.5	-	-

Table 3.5: The result for the chained Wood function ($k = 2$)

var	CSSOS			TSSOS			CS-TSSOS			SDSOS	
	mb	opt	time	mb	opt	time	mb	opt	time	opt	time
40	231	574.51	164	41	574.51	0.81	21	574.51	0.26	518.11	110
60	231	878.26	254	61	878.26	3.61	21	878.26	0.40	-	-
80	231	1182.0	393	81	1182.0	15.3	21	1182.0	0.57	-	-
100	231	1485.8	505	101	1485.8	43.2	21	1485.8	0.73	-	-
120	231	1789.5	516	121	1789.5	88.4	21	1789.5	0.93	-	-
140	231	2093.3	606	141	2093.3	195	21	2093.3	1.16	-	-
160	231	2397.0	700	161	2397.0	403	21	2397.0	1.39	-	-
180	231	2700.8	797	181	2700.8	867	21	2700.8	1.54	-	-
200	231	-	-	201	3004.5	1238	21	3004.5	1.91	-	-
300	231	-	-	-	-	-	21	4523.6	3.39	-	-
400	231	-	-	-	-	-	21	6042.0	5.72	-	-
500	231	-	-	-	-	-	21	7560.7	7.88	-	-
1000	231	-	-	-	-	-	21	15155	23.0	-	-

For the numerical experiments, we construct random instances of Max-Cut problems with a block-band sparsity pattern (illustrated in Figure 3.8) which consists of l blocks of size b and two bands of width h. Here we select $b = 25$ and $h = 5$. For a given l, we generate a random sparse binary matrix $\mathbf{A} \in \mathcal{S}^{lb+h}$ according to the block-arrow sparsity pattern: the entries out of the blue area take zero; the entries in the block area take one with probability 0.16; the entries in the band area take one with probability $2/\sqrt{l}$. Then we construct the graph G with A as its adjacency matrix. For each edge $\{i, j\} \in E(G)$, the weight w_{ij} randomly takes values 1 or -1 with equal probability. Doing so, we build 10 Max-Cut instances with $l = 20, 40, 60, 80, 100, 120, 140, 160, 180, 200$, respectively[5]. The largest number of nodes is 5005.

Figure 3.8: The block-band sparsity pattern



l: the number of blocks; b: the size of blocks; h: the width of bands.

For each instance, we solve the first-order Moment-SOS relaxation (Shor's relaxation), the CSSOS hierarchy with $k = 2$, and the CS-TSSOS hierarchy with $k = 2, t = 1$ for which the maximal chordal extension is used. The results are displayed in Table 3.6. From the table we can see that for each instance, both CSSOS and CS-TSSOS significantly improve the bound obtained by Shor's relaxation. Meanwhile, CS-TSSOS is several times faster than CSSOS at the cost of possibly providing a sightly weaker bound. In addition, CS-TSSOS yields smaller block sizes than CSSOS.

Table 3.6: The result for Max-Cut instances

instance	nodes	edges	mc	Shor	CSSOS			CS-TSSOS		
				opt	mb	opt	time	mb	opt	time
g20	505	2045	14	570	120	488	51.2	92	488	19.6
g40	1005	3441	14	1032	120	885	134	92	893	41.1
g60	1505	4874	14	1439	120	1227	183	92	1247	71.3
g80	2005	6035	15	1899	136	1638	167	106	1669	84.8
g100	2505	7320	14	2398	120	2073	262	92	2128	112
g120	3005	8431	14	2731	120	2358	221	79	2443	127
g140	3505	9658	13	3115	105	2701	250	79	2812	153
g160	4005	10677	14	3670	120	3202	294	79	3404	166
g180	4505	12081	13	4054	105	3525	354	79	3666	246
g200	5005	13240	13	4584	105	4003	374	79	4218	262

In this table, only the integer part of optimal values is preserved.

[5]The instances are available at https://wangjie212.github.io/jiewang/code.html.

3.4.4 The AC-OPF problem

The AC optimal power flow (AC-OPF) is a central problem in power systems. It can be formulated as the following POP in complex variables V_i, S_q^g, S_{ij}:

$$\begin{cases} \inf\limits_{V_i, S_q^g, S_{ij}} & \sum_{q\in G}(\mathbf{c}_{2q}(\Re(S_q^g))^2 + \mathbf{c}_{1q}\Re(S_q^g) + \mathbf{c}_{0q}) \\ \text{s.t.} & \angle V_r = 0, \\ & \mathbf{S}_q^{gl} \le S_q^g \le \mathbf{S}_q^{gu}, \quad \forall q \in G, \\ & \boldsymbol{\upsilon}_i^l \le |V_i| \le \boldsymbol{\upsilon}_i^u, \quad \forall i \in N, \\ & \sum_{q\in G_i} S_q^g - \mathbf{S}_i^d - \mathbf{Y}_i^s |V_i|^2 = \sum_{(i,j)\in E_i\cup E_i^R} S_{ij}, \quad \forall i \in N, \\ & S_{ij} = (\mathbf{Y}_{ij}^* - \mathbf{i}\frac{\mathbf{b}_{ij}^c}{2})\frac{|V_i|^2}{|\mathbf{T}_{ij}|^2} - \mathbf{Y}_{ij}^* \frac{V_i V_j^*}{\mathbf{T}_{ij}}, \quad \forall (i,j) \in E, \\ & S_{ji} = (\mathbf{Y}_{ij}^* - \mathbf{i}\frac{\mathbf{b}_{ij}^c}{2})|V_j|^2 - \mathbf{Y}_{ij}^* \frac{V_i^* V_j}{\mathbf{T}_{ij}^*}, \quad \forall (i,j) \in E, \\ & |S_{ij}| \le \mathbf{s}_{ij}^u, \quad \forall (i,j) \in E \cup E^R, \\ & \boldsymbol{\theta}_{ij}^{\Delta l} \le \angle(V_i V_j^*) \le \boldsymbol{\theta}_{ij}^{\Delta u}, \quad \forall (i,j) \in E. \end{cases} \tag{3.4.3}$$

The meaning of the symbols in (3.4.3) is as follows: N - the set of buses, G - the set of generators, G_i - the set of generators connected to bus i, E - the set of *from* branches, E^R - the set of *to* branches, E_i and E_i^R - the subsets of branches that are incident to bus i, $\mathbf{i}$ - imaginary unit, V_i - the voltage at bus i, S_q^g - the power generation at generator q, S_{ij} - the power flow from bus i to bus j, $\Re(\cdot)$ - real part of a complex number, $\angle(\cdot)$ - angle of a complex number, $|\cdot|$ - magnitude of a complex number, $(\cdot)^*$ - conjugate of a complex number, r - the voltage angle reference bus. All symbols in boldface are constants ($\mathbf{c}_{0q}, \mathbf{c}_{1q}, \mathbf{c}_{2q}, \boldsymbol{\upsilon}_i^l, \boldsymbol{\upsilon}_i^u, \mathbf{s}_{ij}^u, \boldsymbol{\theta}_{ij}^{\Delta l}, \boldsymbol{\theta}_{ij}^{\Delta u} \in \mathbb{R}$,$\mathbf{S}_q^{gl}, \mathbf{S}_q^{gu}, \mathbf{S}_i^d, \mathbf{Y}_i^s, \mathbf{Y}_{ij}, \mathbf{b}_{ij}^c, \mathbf{T}_{ij} \in \mathbb{C}$). For a full description on the AC-OPF problem, the reader may refer to [10]. By introducing real variables for both real and imaginary parts of each complex variable, we can convert the AC-OPF problem to a POP involving only real variables[6].

To tackle an AC-OPF instance, we first compute a locally optimal solution with a local solver and then rely on an SDP relaxation to certify the global optimality. Suppose that the optimal value reported by the local solver is AC and the optimal value of the SDP relaxation is opt. The *optimality gap* between the locally optimal solution and the SDP relaxation is defined by

$$\text{gap} := \frac{\text{AC} - \text{opt}}{\text{AC}} \times 100\%.$$

If the optimality gap is less than 1.00%, then we accept the locally optimal solution as globally optimal. For many AC-OPF instances, the first-order Moment-SOS relaxation (Shor's relaxation) is already able to certify the global optimality (with an optimality gap less than 1.00%). Therefore, we focus on the more challenging AC-OPF instances for which the optimality gap given by Shor's relaxation is greater than 1.00%. We select such instances from the AC-OPF library *PGLiB* [10]. Since we shall go to the second-order Moment-SOS relaxation, we can replace the variables S_{ij} and S_{ji} by their right-hand side expressions in (3.4.3) and then convert the resulting problem to a POP involving real variables. The data for these selected AC-OPF instances are displayed in Table 3.7, where the AC values are taken from *PGLiB*.

We solve Shor's relaxation, the CSSOS hierarchy with $k = 2$ and the CS-TSSOS hierarchy with $k = 2, t = 1$ for these AC-OPF instances and the results are displayed in Tables 3.7–3.8. For instances 162_ieee_dtc, 162_ieee_dtc_api, 500_tamu, 1888_rte, with maximal chordal extensions `Mosek` ran out of memory and so we use approximately smallest chordal extensions. As the tables show, CS-TSSOS is more efficient and scales much better with the problem size than CSSOS. In particular, CS-TSSOS succeeds in reducing the optimality gap to less than 1.00% for all instances.

[6]The expressions involving angles of complex variables can be converted to polynomials by using $\tan(\angle z) = y/x$ for $z = x + \mathbf{i}y \in \mathbb{C}$.

Table 3.7: The data for AC-OPF instances

case name	var	cons	mc	AC	Shor	
					opt	gap
3_lmbd_api	12	28	6	1.1242e4	1.0417e4	7.34%
5_pjm	20	55	6	1.7552e4	1.6634e4	5.22%
24_ieee_rts_api	114	315	10	1.3495e5	1.3216e5	2.06%
24_ieee_rts_sad	114	315	14	7.6943e4	7.3592e4	4.36%
30_as_api	72	297	8	4.9962e3	4.9256e3	1.41%
73_ieee_rts_api	344	971	16	4.2263e5	4.1041e5	2.89%
73_ieee_rts_sad	344	971	16	2.2775e5	2.2148e5	2.75%
162_ieee_dtc	348	1809	21	1.0808e5	1.0616e5	1.78%
162_ieee_dtc_api	348	1809	21	1.2100e5	1.1928e5	1.42%
240_pserc	766	3322	16	3.3297e6	3.2818e6	1.44%
500_tamu_api	1112	4613	20	4.2776e4	4.2286e4	1.14%
500_tamu	1112	4613	30	7.2578e4	7.1034e4	2.12%
793_goc	1780	7019	18	2.6020e5	2.5636e5	1.47%
1888_rte	4356	18257	26	1.4025e6	1.3748e6	1.97%
3022_goc	6698	29283	50	6.0143e5	5.9278e5	1.44%

Table 3.8: The result for AC-OPF instances

case name	CSSOS				CS-TSSOS				
	mb	opt	time	gap	mb	opt	time	gap	CE
3_lmbd_api	28	1.1242e4	0.21	0.00%	22	1.1242e4	0.09	0.00%	max
5_pjm	28	1.7543e4	0.56	0.05%	22	1.7543e4	0.30	0.05%	max
24_ieee_rts_api	66	1.3442e5	5.59	0.39%	31	1.3396e5	2.01	0.73%	max
24_ieee_rts_sad	120	7.6943e4	94.9	0.00%	39	7.6942e4	14.8	0.00%	max
30_as_api	45	4.9927e3	4.43	0.07%	22	4.9920e3	2.69	0.08%	max
73_ieee_rts_api	153	4.2246e5	758	0.04%	44	4.2072e5	96.0	0.45%	max
73_ieee_rts_sad	153	2.2775e5	504	0.00%	44	2.2766e5	71.5	0.04%	max
162_ieee_dtc	253	–	–	–	34	1.0802e5	278	0.05%	min
162_ieee_dtc_api	253	–	–	–	34	1.2096e5	201	0.03%	min
240_pserc	153	3.3072e6	585	0.68%	44	3.3042e6	33.9	0.77%	max
500_tamu_api	231	4.2413e4	3114	0.85%	39	4.2408e4	46.6	0.86%	max
500_tamu	496	–	–	–	31	7.2396e4	410	0.25%	min
793_goc	190	2.5938e5	563	0.31%	33	2.5932e5	66.1	0.34%	max
1888_rte	378	–	–	–	27	1.3953e6	934	0.51%	min
3022_goc	1326	–	–	–	76	5.9858e5	1886	0.47%	max

Chapter 4

Exploiting the constant trace property: Equality constraints

Most of the content of this chapter is from [134].

In the previous chapter, we have combined correlative and term sparsity exploitation to improve the scalability of the Moment-SOS hierarchy for POP.

Another complementary workaround to improve further the scalability is by exploiting a *Constant Trace Property* (CTP) of semidefinite relaxations associated with POPs coming from combinatorial optimization [75, 219]. This permits to solve a given semidefinite relaxation with ad-hoc methods, like, e.g., limited-memory bundle methods, instead of the costly interior-point methods.

The present chapter is part of this effort.

Background on SDP with CTP

One way to exploit the CTP of matrices in SDPs is to consider the dual which reduces to minimize the maximum eigenvalue of a symmetric matrix pencil [75]. For problems of moderate size one may solve the latter problem with interior-point methods [16]. However for larger-scale instances, running a single iteration becomes computationally too demanding and therefore one has to use an alternative method, and in particular first-order methods.

To solve large-scale instances of this maximal eigenvalue minimization problem, two types of first-order methods can be used: subgradient descent or variants of the mirror-prox algorithm [144], and spectral bundle methods [75]. In other methods of interest based on non-convex formulations [27, 93], the problem is directly solved over the set of low rank matrices. These latter approaches are particularly efficient for problems where the solution is low rank, e.g., for matrix completion or combinatorial relaxations.

Despite their empirical efficiency, the computational complexity of spectral bundle and low rank methods is still not completely understood. This is in contrast with methods based on stochastic smoothing results for which explicit computational complexity estimates are available. For instance in [42] smooth stochastic approximations of the maximum eigenvalue function are obtained via rank-one Gaussian perturbations. In [155] Newton's method is used, assuming that the multiplicity of the maximal eigenvalue is known in advance.

By combining quasi-Newton methods (e.g. Broyden-Fletcher-Goldfarb-Shanno (BFGS) method or its so-called "Limited-memory" version (L-BFGS) [151]) with adaptive gradient sampling [30, 99], convergence guarantees are obtained for certain non smooth problems while keeping good empirical performance [116, 39].

Another hybrid method is the Limited-Memory Bundle Method (LMBM) which combines L-BFGS with bundle methods [71, 70]: Briefly, L-BFGS is used in the line search procedure to determine the step sizes in the bundle method. LMBM enjoys global convergence for locally Lipschitz continuous functions which are not necessarily differentiable.

Finally the more recent `SketchyCGAL` algorithm [219] also uses limited memory and arithmetic. It combines a primal-dual optimization scheme together with a randomized sketch for low-rank matrix approximation. Assuming that zero duality holds, it provides a near-optimal low-rank

approximation. A variant of `SketchyCGAL` can handle SDPs with *bounded* (instead of *constant*) trace property.

Concerning SDPs coming from relaxations in polynomial optimization, Malick and Henrion [78, Section 3.2.3] have used the CTP to provide an efficient algorithm for unconstrained polynomial optimization problems. At last but not least, the CTP trivially holds for Shor's relaxation [189] of combinatorial optimization problems formulated as linear-quadratic POPs on the discrete hypercube $\{-1,1\}^n$. This fact has been exploited in Helmberg and Rendl [75] to avoid solving the associated SDP via interior-point methods.

Contribution

A novelty with respect to previous efforts is to show that *every* POP on a compact basic semialgebraic set has an equivalent equality constrained POP formulation on an Euclidean sphere (possibly after adding some artificial variables) such that each of its semidefinite relaxations in the Moment-SOS hierarchy has the CTP. We call CTP-POP such a formulation of POPs. Therefore to solve each semidefinite relaxation of a CTP-POP one may avoid the computationally costly interior-point methods in some cases. Indeed as the dual reduces to minimize the largest eigenvalue of a matrix pencil, one may rather use efficient ad-hoc non smooth methods as those invoked above.

I. In Section 4.2.1, we prove that each semidefinite moment relaxation indexed by $k \in \mathbb{N}$:

$$-\tau_k = \sup_{\mathbf{X} \in \mathcal{S}^{(k)}} \{ \langle \mathbf{C}_k, \mathbf{X} \rangle \,:\, \mathcal{A}_k \mathbf{X} = \mathbf{b}_k \,,\, \mathbf{X} \succeq 0 \} \,, \tag{4.0.1}$$

of the Moment-SOS hierarchy associated with an equality constrained POP on an Euclidean sphere of $\mathbb{R}^n$ has CTP (see Lemma 4.4), i.e.,

$$\forall\, \mathbf{X} \in \mathcal{S}^{(k)} \,,\, \mathcal{A}_k \mathbf{X} = \mathbf{b}_k \Rightarrow \operatorname{trace}(\mathbf{X}) = a_k \,,$$

where $\mathcal{A}_k : \mathcal{S}^{(k)} \to \mathbb{R}^{m_k}$ is a linear operator with $\mathcal{S}^{(k)}$ being the set of real symmetric matrices of size $\binom{n+k}{n}$, $\mathbf{C}_k \in \mathcal{S}^{(k)}$ and $\mathbf{b}_k \in \mathbb{R}^{m_k}$ with $m_k = \mathcal{O}\left(\binom{n+k}{n}^2\right)$. Following the framework by Helmberg and Rendl [75], SDP (4.0.1) boils down to minimizing the largest eigenvalue of a matrix pencil:

$$-\tau_k = \inf_{\mathbf{z} \in \mathbb{R}^{m_k}} a_k \lambda_1(\mathbf{C}_k - \mathcal{A}_k^\top \mathbf{z}) + \mathbf{b}_k^\top \mathbf{z} \,, \tag{4.0.2}$$

where $\lambda_1(\mathbf{A})$ stands for the largest eigenvalue of $\mathbf{A}$ and $\mathcal{A}_k^\top$ denotes the adjoint operator of $\mathcal{A}_k$.

Hence (4.0.2) form what we call a hierarchy of (non smooth, convex) *spectral relaxations* of the equality constrained POP on a sphere. Convergence of $(\tau_k)_{k \in \mathbb{N}}$ to the optimal value $f^\star$ of the initial POP is guaranteed with rate at least $\mathcal{O}(k^{-\mathfrak{c}})$ (see Theorem 4.1), where c depends only on the polynomials describing the cost and constraints of the POP.

In addition, existence of an optimal solution of the spectral relaxation (4.0.2) is guaranteed for sufficiently large k under certain conditions on the POP (see Proposition 4.1). Finally, when the set of global minimizers of the equality constrained POP on the sphere is finite, we also describe how to obtain an optimal solution $\mathbf{x}^\star$ via an optimal solution $\bar{\mathbf{z}}$ of (4.0.2).

II. In Section 4.2 we prove that any POP on a compact basic semialgebraic set (including a ball constraint $R - \|\mathbf{x}\|_2^2 \geq 0$) has an equivalent equality constrained POP (called CTP-POP) on a sphere of $\mathbb{R}^{n+m+1}$, where m is the number of inequality constraints of the initial POP. This CTP-POP can be solved by using spectral relaxations (4.0.2).

III. We describe Algorithm 3 to handle a given equality constrained POP on the sphere. It consists of handling each semidefinite relaxation (4.0.1) by solving the spectral formulation (4.0.2), with a nonsmooth optimization procedure chosen in advance by the user in our software library, called SpectralSOS. This library supports the three optimization subroutines LMBM [71, 70], proximal bundle (PB) [75], and `SketchyCGAL` [219]. Our default method in Algorithm 3 is LMBM.

IV. Finally, efficiency and robustness of SpectralPOP are illustrated in Section 4.3 on extensive benchmarks. We solve several (randomly generated) dense equality constrained QCQPs on the unit sphere by running Algorithm 3 and compare results with those obtained with the standard Moment-SOS hierarchy. Suprisingly SpectralPOP can provide the optimal value as well as an optimal solution with high accuracy, and up to twenty five times faster than the semidefinite

hierarchy. For instance, SpectralPOP can solve the first relaxation of minimization problem of dense quadratic polynomials on the unit sphere with up to $n = 500$ variables in about 35 seconds and up to 1500 variables in about 7000 seconds on a standard laptop computer. We emphasize that for some problems not randomly generated and scaled so as to fit our optimization framework on the unit sphere, we could observe a lack of high precision after transferring results (of the scaled formulation) back to the unscaled initial formulation.

We also provide extended applications of spectral relaxations to the following three decision problems: deciding nonnegativity of even degree forms, deciding convexity of even degree forms and deciding copositivity of real symmetric matrices, with very satisfactory results.

In [75], Helmberg and Rendl propose a spectral bundle method (based on Kiwiel's proximal bundle method [97]) to solve an SDP relying on the maximal eigenvalue minimization problem of the form (4.0.2). This method works better than interior-point algorithms for very large-scale SDPs, when the number of trace equality constraints is not larger than the size of the positive semidefinite matrix (e.g., Shor's relaxation of MAXCUT problems). However this method is not always more efficient than interior-point solvers (e.g., `SDPT3`) for instance when the SDPs involve a number of trace equality constraints which is larger than the size of the positive semidefinite matrix, as reported in [74, Table 1-6]. Unfortunately this latter type of SDP is the generic form of Moment-SOS relaxations for POPs and thus is not suitable to be solved by Helmberg-Rendl's spectral bundle method. By contrast with previous works, our numerical results show that the combination between Helmberg-Rendl's spectral formulation and LMBM is cheaper and faster than `Mosek` (the currently fastest SDP solver based on interior-point method) while maintaining the same accuracy when solving moment relaxations of equality constrained POPs on a sphere.

4.1 Background and Preliminary Results

4.1.1 General POPs on basic compact semialgebraic sets

We recall that a POP is of the form

$$f^\star := \inf\{f(\mathbf{x}) \ : \ \mathbf{x} \in S(\mathfrak{g},\mathfrak{h})\}\,, \tag{4.1.1}$$

where $S(\mathfrak{g},\mathfrak{h})$ is a basic semialgebraic set defined as follows:

$$S(\mathfrak{g},\mathfrak{h}) \ := \ \{\,\mathbf{x} \in \mathbb{R}^n : \ g_i(\mathbf{x}) \geq 0\,,\, i \in [m]\,;\, h_j(\mathbf{x}) \ = \ 0\,,\, j \in [l]\,\} \tag{4.1.2}$$

for some polynomials $f, g_i, h_j \in \mathbb{R}[\mathbf{x}]$. Here $\mathfrak{g} := \{g_i\}_{i\in[m]}$ and $\mathfrak{h} := \{h_j\}_{j\in[l]}$. Set $\lceil g_i \rceil := \lceil \deg(g_i)/2 \rceil$ and $\lceil h_j \rceil := \lceil \deg(h_j)/2 \rceil$.

If $S(\mathfrak{g},\mathfrak{h})$ is nonempty and compact, then $f^\star < \infty$ and POP (4.1.1) has at least one global minimizer. Next, as we are concerned with POPs on compact feasible sets, we assume that $S(\mathfrak{g},\mathfrak{h}) \subset B(\mathbf{0},R)$, where $B(\mathbf{0},R) := \{\mathbf{x} \in \mathbb{R}^n : R - \|\mathbf{x}\|_2^2 \geq 0\}$. In addition, if $m \neq 0$ then we may and will assume that $g_1 := R - \|\mathbf{x}\|_2^2$.

4.1.2 POPs on a variety contained in a sphere

We consider a special form of POP (4.1.1) which is of the form

$$f^\star := \inf\{\, f(\mathbf{x}) \ : \ \mathbf{x} \in V(\mathfrak{h})\}\,, \tag{4.1.3}$$

where $V(\mathfrak{h})$ is the real variety defined by:

$$V(\mathfrak{h}) \ := \ \{\,\mathbf{x} \in \mathbb{R}^n : \ h_j(\mathbf{x}) \ = \ 0\,;\, j \in [l]\,\}\,, \tag{4.1.4}$$

for some set of polynomials $\mathfrak{h} := \{h_j\}_{j=1}^l \subset \mathbb{R}[\mathbf{x}]$. We assume that $h_1 := \bar{R} - \|\mathbf{x}\|_2^2$ for some $\bar{R} > 0$, so that $V(\mathfrak{h}) \subset \partial B(\mathbf{0},\bar{R})$, where $\partial B(\mathbf{0},\bar{R}) := \{\mathbf{x} \in \mathbb{R}^n : \bar{R} - \|\mathbf{x}\|_2^2 = 0\}$. By assuming that $V(\mathfrak{h}) \neq \emptyset$, $f^\star < \infty$ and POP (4.1.3) has at least one global minimizer.

Given $k \in \mathbb{N}$, define the *truncated preordering* of order k associated with the variety $V(\mathfrak{h})$ in (4.1.4) as follows:

$$\mathcal{P}_k(\mathfrak{h}) := \left\{ \sigma_0 + \sum_{j=1}^{l} \psi_j h_j \ : \ \sigma_0 \in \Sigma[\mathbf{x}]_k\,,\, \psi_j \in \mathbb{R}[\mathbf{x}]_{2(k-\lceil h_j \rceil)}\,,\, j \in [l] \right\}.$$

Remark 4.1. *For every $k \in \mathbb{N}$, $\mathcal{P}_k(\mathfrak{h})$ is also the truncated quadratic module $\mathcal{Q}_k(\mathfrak{h})$ associated with the semialgebraic set $V(\mathfrak{h}) = S(\emptyset, \mathfrak{h})$.*

As a consequence of Schweighofer's main result in [188, Theorem 4], one obtains the following result:

Lemma 4.1. *Let $f^\star$ be as in (4.1.3) with $V(\mathfrak{h})$ as in (4.1.4). There exists $\mathfrak{c} > 0$ depending on V such that for $k \in \mathbb{N}$ with $k \geq \mathfrak{c} d^{\mathfrak{c}} n^{\mathfrak{c} d}$, one has*

$$(f - f^\star) + \mathfrak{c} d^4 n^{2d} \|f\| k^{-1/\mathfrak{c}} \in \mathcal{P}_k(\mathfrak{h}) \,.$$

Note that in the case of polynomial optimization on the sphere (i.e., $\mathfrak{h} = \{R - \|\mathbf{x}\|_2^2\}$ for some $R > 0$), one can take $\mathfrak{c} = \frac{1}{2}$ in Lemma 4.1, as a consequence of the convergence result from [57].

Next, consider the hierarchy of semidefinite programs (SDP) indexed by $k \in \mathbb{N}$:

$$\rho_k := \sup \{ \xi \in \mathbb{R} \ : \ f - \xi \in \mathcal{P}_k(\mathfrak{h}) \} \,. \tag{4.1.5}$$

For every $k \in \mathbb{N}$, the dual of (4.1.5) reads

$$\begin{aligned} \tau_k := \inf_{\mathbf{y} \in \mathbb{R}^{b(n,2k)}} \quad & L_{\mathbf{y}}(f) \\ \text{s.t.} \quad & \mathbf{M}_k(\mathbf{y}) \succeq 0 \,;\, y_{\mathbf{0}} = 1 \\ & \mathbf{M}_{k - \lceil h_j \rceil}(h_j \, \mathbf{y}) = 0 \,,\, j \in [l] \,. \end{aligned} \tag{4.1.6}$$

By invoking Lemma 4.1 and Proposition 2.4, one obtains the convergence behavior of the sequence $(\rho_k)_{k \in \mathbb{N}}$ in the following result.

Theorem 4.1. *Let $f^\star$ be as in (4.1.3) with $V(\mathfrak{h}) \neq \emptyset$ as in (4.1.4). Then:*

1. *For all $k \in \mathbb{N}$, $\rho_k \leq \rho_{k+1} \leq f^\star$.*
2. *The sequence $(\rho_k)_{k \in \mathbb{N}}$ converges to $f^\star$ with rate at least $\mathcal{O}(k^{-1/\mathfrak{c}})$.*
3. *If the ideal $I(\mathfrak{h})$ is real radical and the second-order sufficiency conditions (Definition 2.1) hold at every global minimizer of POP (4.1.3) then $\tau_k = \rho_k = f^\star$ for some k and (4.1.5) has an optimal solution, i.e., $f - f^\star \in \mathcal{P}_k(\mathfrak{h})$.*
4. *If $V(\mathfrak{h})$ defined as in (4.1.4) is finite, $\tau_k = \rho_k = f^\star$ for some $k \in \mathbb{N}$ and both primal-dual (2.4.3)-(2.4.5) have optimal solutions. In this case, the flatness condition holds at order k.*

With $V(\mathfrak{h})$ in lieu of $S(\mathfrak{g}, \mathfrak{h})$, zero duality gap as well as analogues of Proposition 2.1 and 2.3, also hold.

4.1.3 Spectral minimizations of SDP

Let $s, l, s^j \in \mathbb{N}^{>0}$, $j \in [l]$, be fixed such that $s = \sum_{j=1}^{l} s^{(j)}$. Let $\mathcal{S}$ be the set of real symmetric matrices of size s in a block diagonal form:

$$\mathbf{X} = \operatorname{diag}(\mathbf{X}_1, \ldots, \mathbf{X}_l) \,, \tag{4.1.7}$$

such that $\mathbf{X}_j$ is of size $s^{(j)}$, $j \in [l]$. Let $\mathcal{S}^+$ be the set of all $\mathbf{X} \in \mathcal{S}$ such that $\mathbf{X} \succeq 0$, i.e., $\mathbf{X}$ has only nonnegative eigenvalues. Then $\mathcal{S}$ is a Hilbert space with scalar product $\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{trace}(\mathbf{B}^\top \mathbf{A})$ and $\mathcal{S}_+$ is a self-dual cone.

Let us consider the following SDP:

$$-\tau = \sup_{\mathbf{X} \in \mathcal{S}} \{ \langle \mathbf{C}, \mathbf{X} \rangle \ : \ \mathcal{A}\mathbf{X} = \mathbf{b} \,,\, \mathbf{X} \succeq 0 \} \,, \tag{4.1.8}$$

where $\mathcal{A} : \mathcal{S} \to \mathbb{R}^m$ is a linear operator of the form

$$\mathcal{A}\mathbf{X} = [\langle \mathbf{A}_1, \mathbf{X} \rangle, \ldots, \langle \mathbf{A}_m, \mathbf{X} \rangle] \,,$$

with $\mathbf{A}_i \in \mathcal{S}$, $i \in [m]$, $\mathbf{C} \in \mathcal{S}$ is the cost matrix and $\mathbf{b} \in \mathbb{R}^m$ is the right-hand-side vector.

The dual of SDP (4.1.8) reads:

$$-\rho = \inf_{\mathbf{z}} \{ \mathbf{b}^\top \mathbf{z} \,:\, \mathcal{A}^\top \mathbf{z} - \mathbf{C} \succeq 0 \}, \tag{4.1.9}$$

where $\mathcal{A}^\top : \mathbb{R}^m \to \mathcal{S}$ is the adjoint operator of $\mathcal{A}$, i.e., $\mathcal{A}^\top \mathbf{z} = \sum_{i=1}^m z_i \mathbf{A}_i$. The following assumption will be used in the next two sections:

Assumption 4.1. *Consider the following conditions:*

1. *Zero duality gap of primal-dual* (4.1.8)-(4.1.9) *holds, i.e., $\tau = \rho$ and $\tau \in \mathbb{R}$.*
2. *Primal attainability: SDP* (4.1.8) *has an optimal solution.*
3. *Dual attainability: SDP* (4.1.9) *has an optimal solution.*
4. *Constant trace property (CTP): There exists $a > 0$ such that*

$$\forall\ \mathbf{X} \in \mathcal{S}\,,\ \mathcal{A}\mathbf{X} = \mathbf{b} \Rightarrow \operatorname{trace}(\mathbf{X}) = a\,. \tag{4.1.10}$$

5. *Bounded trace property (BTP): There exists $a > 0$ such that*

$$\forall\ \mathbf{X} \in \mathcal{S}\,,\ \mathcal{A}\mathbf{X} = \mathbf{b} \Rightarrow \operatorname{trace}(\mathbf{X}) \leq a\,. \tag{4.1.11}$$

In Assumption 5.3, conditions 1 and 5 (or conditions 1 and 4) imply condition 2. Indeed, if condition 5 holds, the feasible set of (5.3.25) is compact and if condition 1 holds, the feasible set of (5.3.25) is nonempty. Moreover, conditions 2 and 5 (or conditions 2 and 4) imply condition 1. Indeed, if conditions 2 and 5 hold, the set of optimal solutions of (5.3.25) is nonempty and bounded. Then Trnovska's result [197, Corollary 1] yields condition 1.

Remark 4.2. *If condition 5 of Assumption 4.1 holds, by adding a slack variable* $\mathbf{y}$ *and noting* $\mathbf{Y} = \operatorname{diag}(\mathbf{X}, y)$*, we obtain an equivalent SDP of* (4.1.8) *as follows:*

$$-\tau = \sup_{\mathbf{Y} \in \hat{\mathcal{S}}} \{ \langle \hat{\mathbf{C}}, \mathbf{Y} \rangle \,:\, \langle \hat{\mathbf{A}}_i, \mathbf{Y} \rangle = \mathbf{b}_i\,,\ \mathbf{Y} \succeq 0\,,\ \operatorname{trace}(\mathbf{Y}) = a \}, \tag{4.1.12}$$

where $\hat{\mathcal{S}} = \{\operatorname{diag}(\mathbf{X}, y) : \mathbf{X} \in \mathcal{S}\,,\ y \in \mathbb{R}\}$*,* $\hat{\mathbf{C}} = \operatorname{diag}(\mathbf{C}, 0)$ *and* $\hat{\mathbf{A}}_i = \operatorname{diag}(\mathbf{A}_i, 0)$*. Obviously, SDP* (4.1.12) *has CTP.*

SDP with Constant Trace Property (CTP)

Recall that $\lambda_1(\mathbf{A})$ stands for the largest eigenvalue of a real symmetric matrix $\mathbf{A}$.

Lemma 4.2. *Let conditions 1 and 4 of Assumption 4.1 hold and let $\varphi : \mathbb{R}^m \to \mathbb{R}$ be the function:*

$$\mathbf{z} \mapsto \varphi(\mathbf{z}) \,:=\, a\lambda_1(\mathbf{C} - \mathcal{A}^\top \mathbf{z}) + \mathbf{b}^\top \mathbf{z}\,. \tag{4.1.13}$$

Then:

$$-\tau = \inf_{\mathbf{z}} \{ \varphi(\mathbf{z}) \ :\ \mathbf{z} \in \mathbb{R}^m \}. \tag{4.1.14}$$

Moreover if condition 3 of Assumption 4.1 holds, i.e., SDP (4.1.9) *has an optimal solution then problem* (4.1.14) *has an optimal solution.*

The proof of Lemma 4.2 is available in [134, Appendix].

Next, we describe Algorithm 1 to solve SDP (4.1.8), which is based on nonsmooth first-order optimization methods (e.g., LMBM [71, Algorithm 1]). As shown later on in Section 4.3, this algorithm works well in almost all cases and with significantly lower computational cost when compared to the (currently fastest) SDP solver `Mosek` 9.1.

For $\mathbf{X} \in \mathcal{S}$, the Frobenius norm of $\mathbf{X}$ is defined by $\|\mathbf{X}\|_F := \sqrt{\langle \mathbf{X}, \mathbf{X} \rangle}$. We denote by $\|\mathcal{A}\|$ the operator norm of $\mathcal{A}$, i.e., $\|\mathcal{A}\| := \max_{\mathbf{X} \in \mathcal{S}} \|\mathcal{A}\mathbf{X}\|_2 / \|\mathbf{X}\|_F$.

Remark 4.3. *Before running Algorithm 1, we scale the problem's input as follows:* $\|\mathbf{C}\|_F = \|\mathcal{A}\| = a = 1$ *and* $\|\mathbf{A}_1\|_F = \cdots = \|\mathbf{A}_m\|_F$.

Algorithm 1 SDP-CTP

Input: SDP (4.1.8) with unknown optimal value and optimal solution;
method (T) for solving convex nonsmooth unconstrained optimization problems (NSOP).
Output: optimal value $-\tau$ and optimal solution $\mathbf{X}^\star$ of SDP (4.1.8).

1: Compute the optimal value $-\tau$ and an optimal solution $\bar{\mathbf{z}}$ of the NSOP (4.1.14) by using method (T);
2: Compute a normalized eigenvector $\mathbf{u}$ corresponding to $\lambda_1(\mathbf{C} - \mathcal{A}^\top \bar{\mathbf{z}})$ and set $\mathbf{X}^\star = a\mathbf{u}\mathbf{u}^\top$.

The fact that Algorithm 1 is well-defined under certain conditions is a corollary of Lemma 4.2 and [134, Lemma A.22].

Corollary 4.1. *Let conditions 1 and 4 of Assumption 4.1 hold. Assume that the method (T) is globally convergent for NSOP* (4.1.14) *(e.g., (T) is LMBM). Then output $-\tau$ of Algorithm 1 is well-defined. Moreover, if condition 3 of Assumption 4.1 holds, the vector $\bar{\mathbf{z}}$ mentioned at Step 1 of Algorithm 1 exists. In this case, if $\lambda_1(\mathbf{C} - \mathcal{A}^\top \bar{\mathbf{z}})$ has multiplicity* 1*, the output $\mathbf{X}^\star$ of Algorithm 1 is well-defined.*

When $\lambda_1(\mathbf{C}-\mathcal{A}^\top \bar{\mathbf{z}})$ has multiplicity larger than 1, one can obtain a dual matrix $\mathbf{G}$ corresponding to an optimal solution $\mathbf{X}^\star$ of SDP (4.1.8) as in the following corollary:

Corollary 4.2. *Let condition 4 of Assumption 4.1 hold. Let $\bar{\mathbf{z}}$ be an optimal solution of the NSOP* (4.1.14)*. Define*

$$\begin{aligned} \mathbf{U} &:= \mathbf{C} - \mathcal{A}^\top \bar{\mathbf{z}}\,, \\ \mathbf{G} &:= \lambda_1(\mathbf{U})\mathbf{I} - \mathbf{U}\,, \end{aligned} \tag{4.1.15}$$

where $\mathbf{I}$ is the identity matrix. Then $\mathbf{G}$ is positive semidefinite and satisfies

$$\mathbf{G} = \mathcal{A}^\top \mathbf{z}^\star - \mathbf{C}\,, \tag{4.1.16}$$

for some optimal solution $\mathbf{z}^\star$ of (4.1.9)*.*

Proof. It is not hard to prove that $\mathbf{G} \succeq 0$. Let us prove the other statement. By using Farkas' lemma, there exists $\mathbf{y} \in \mathbb{R}^m$ such that $\mathcal{A}^\top \mathbf{y} = \mathbf{I}$ and $\mathbf{y}^\top \mathbf{b} = a$ (see [74, Section 2]). Then $\mathbf{G} = \lambda_1(\mathbf{U})\mathcal{A}^\top \mathbf{y} - \mathbf{C} + \mathcal{A}^\top \bar{\mathbf{z}} = \mathcal{A}^\top(\lambda_1(\mathbf{U})\mathbf{y} + \bar{\mathbf{z}}) - \mathbf{C} = \mathcal{A}^\top \mathbf{z}^\star - \mathbf{C}$, where $\mathbf{z}^\star := \lambda_1(\mathbf{U})\mathbf{y} + \bar{\mathbf{z}}$. Since $\mathbf{b}^\top \mathbf{z}^\star = \lambda_1(\mathbf{U})\mathbf{b}^\top \mathbf{y} + \mathbf{b}^\top \bar{\mathbf{z}} = \varphi(\bar{\mathbf{z}}) = -\tau$, $\mathbf{z}^\star$ is an optimal solution of (4.1.9). □

Largest eigenvalue computation: Step 1 of Algorithm 1 requires the largest eigenvalue and corresponding eigenvectors of $\mathbf{C} - \mathcal{A}^\top \mathbf{z}$ to evaluate the function φ (resp. ψ) and a subgradient of the subdifferential $\partial\varphi$ (resp. $\partial\psi$) given in [134, Proposition A.1] (resp. [134, Proposition A.2]) at $\mathbf{z}$. Fortunately, solving the eigenvalue problem for $\mathbf{C} - \mathcal{A}^\top \mathbf{z} \in \mathcal{S}$ can be done on every block of $\mathbf{C} - \mathcal{A}^\top \mathbf{z}$. Indeed, with $\mathbf{X} \in \mathcal{S}$ as in (4.1.7),

$$\lambda(\mathbf{X}) = \lambda(\mathbf{X}_1) \cup \cdots \cup \lambda(\mathbf{X}_l)\,,$$

where $\lambda(\mathbf{A})$ is the set of all eigenvalues $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_t(\mathbf{A})$ for every real symmetric matrix $\mathbf{A}$ of size t. In particular,

$$\lambda_1(\mathbf{X}) = \max\{\lambda_1(\mathbf{X}_1), \ldots, \lambda_1(\mathbf{X}_l)\}\,.$$

If $\mathbf{u} \in \mathbb{R}^{s^{(j)}}$ is an eigenvector of $\mathbf{X}_j$ corresponding to the eigenvalue $\lambda_i(\mathbf{X}_j)$ for some $i \in [s^{(j)}]$ and $j \in [l]$, by adding zero entries in $\mathbf{u}$,

$$\bar{\mathbf{u}} = (\mathbf{0}_{\mathbb{R}^{s^{(1)}+\cdots+s^{(j-1)}}}, \mathbf{u}, \mathbf{0}_{\mathbb{R}^{s^{(j+1)}+\cdots+s^{(l)}}})$$

is an eigenvector of $\mathbf{X} = \mathrm{diag}(\mathbf{X}_1, \ldots, \mathbf{X}_l)$ corresponding to $\lambda_i(\mathbf{X}_j)$.

The interested reader can refer to [181, 115] to solve largest eigenvalue problems of symmetric matrices of large sizes.

Remark 4.4. *Let conditions 1, 2 and 5 of Assumption 4.1 hold. We keep all notation from Remark 4.2. By applying Lemma 4.2 for SDP* (4.1.12) *with CTP, one has*

$$-\tau = \inf\{\ a\lambda_1(\hat{\mathbf{C}} - \hat{\mathcal{A}}^\top \mathbf{z}) + \mathbf{b}^\top \mathbf{z}\ :\ \mathbf{z} \in \mathbb{R}^m\}\,, \tag{4.1.17}$$

where $\hat{\mathcal{A}}^\top \mathbf{z} = \sum_{i=1}^m z_i \hat{\mathbf{A}}_i$. *Note that* $\hat{\mathbf{C}} - \hat{\mathcal{A}}^\top \mathbf{z} = \operatorname{diag}(\mathbf{C} - \mathcal{A}^\top \mathbf{z}, 0)$. *It implies that* $\lambda_1(\hat{\mathbf{C}} - \hat{\mathcal{A}}^\top \mathbf{z}) = \max\{\lambda_1(\mathbf{C} - \mathcal{A}^\top \mathbf{z}), 0\}$. *Thus,* (4.1.17) *can be rewritten as*

$$-\tau = \inf\{\ a\max\{\lambda_1(\mathbf{C} - \mathcal{A}^\top \mathbf{z}), 0\} + \mathbf{b}^\top \mathbf{z}\ :\ \mathbf{z} \in \mathbb{R}^m\}\,. \tag{4.1.18}$$

In the next section, we consider the spectral formulation (4.1.18) introduced by Ding et al. in [48, Section 6].

SDP with Bounded Trace Property (BTP)

In the last subsection, we have seen that SDPs with CTP can be solved efficiently with first-order methods. Similar results can be obtained for the larger class of SDPs with the weaker *bounded trace property* (BTP). In particular the semidefinite relaxations of the Moment-SOS hierarchy associated with a POP on a compact semialgebraic set have the BTP. So in principle there is no need to add auxiliary "slack" variables to obtain an equivalent CTP-POP, as shown in Remark 4.2. However, numerical experiments of Section 4.3 suggest that the CTP is a highly desirable property that justifies addition of auxiliary variables.

The analogue of Lemma 4.2 for BTP reads:

Lemma 4.3. *Let conditions 1, 2 and 5 of Assumption 4.1 hold, and let* $\psi : \mathbb{R}^m \to \mathbb{R}$ *be the function:*

$$\mathbf{z} \mapsto \psi(\mathbf{z}) := a\max\{\lambda_1(\mathbf{C} - \mathcal{A}^\top \mathbf{z}), 0\} + \mathbf{b}^\top \mathbf{z}\,. \tag{4.1.19}$$

Then

$$-\tau = \inf_{\mathbf{z}}\{\ \psi(\mathbf{z})\ :\ \mathbf{z} \in \mathbb{R}^m\}\,. \tag{4.1.20}$$

Moreover if condition 3 of Assumption 4.1 holds, then problem (4.1.20) *has an optimal solution.*

The proof of Lemma 4.3 can be found in [134, Appendix].

We next describe Algorithm 2 to solve SDP (4.1.8). As Algorithm 1, it is also based on nonsmooth optimization methods such as LMBM.

Algorithm 2 SDP-BTP

Input: SDP (4.1.8) with unknown optimal value and optimal solution;
method (T) for solving convex NSOP.

Output: optimal value $-\tau$ and optimal solution $\mathbf{X}^\star$ of SDP (4.1.8).

1: Compute the optimal value $-\tau$ and an optimal solution $\bar{\mathbf{z}}$ of NSOP (4.1.14) by using method (T);
2: Compute $\lambda_1(\mathbf{C} - \mathcal{A}^\top \bar{\mathbf{z}})$ and a corresponding normalized eigenvector $\mathbf{u}$;
3: Let $\bar{\xi} > 0$ such that

$$\bar{\xi} = \begin{cases} 0 & \text{if } \lambda_1(\mathbf{C} - \mathcal{A}^\top \bar{\mathbf{z}}) < 0\,, \\ \zeta a & \text{if } \lambda_1(\mathbf{C} - \mathcal{A}^\top \bar{\mathbf{z}}) = 0\,, \\ a & \text{otherwise}\,, \end{cases} \tag{4.1.21}$$

for some $\zeta \in [0, 1]$ such that $\mathbf{X}^\star = \bar{\xi}\mathbf{u}\mathbf{u}^\top$ satisfies $\mathcal{A}\mathbf{X}^\star = \mathbf{b}$.

The next result is a consequence of Lemma 4.3 and [134, Lemmas A.3, A.4].

Corollary 4.3. *Let conditions 1, 2 and 5 of Assumption 4.1 hold. Assume that method (T) is globally convergent for NSOP* (4.1.20) *(e.g., (T) is LMBM). Then the output* $-\tau$ *of Algorithm 2 is well-defined. Moreover, if condition 3 of Assumption 4.1 holds, the vector* $\bar{\mathbf{z}}$ *from Step 1 of Algorithm 2 exists. In this case, if* $\lambda_1(\mathbf{C} - \mathcal{A}^\top \bar{\mathbf{z}})$ *has multiplicity* 1, *the output* $\mathbf{X}^\star$ *of Algorithm 2 is well-defined.*

4.2 Application to polynomial optimization

We consider the following POP:

$$f^\star := \inf\{f(\mathbf{x}) \ : \ \mathbf{x} \in S(\mathfrak{g},\mathfrak{h})\}\,, \tag{4.2.1}$$

where $S(\mathfrak{g},\mathfrak{h})$ is defined as in (4.1.2) with m (resp. l) being the number of inequality (resp. equality) constraints. Assume that $S(\mathfrak{g},\mathfrak{h}) \subset B(\mathbf{0},R)$.

Remark 4.5. *By setting* $\mathbf{X} := \operatorname{diag}(\mathbf{M}_k(\mathbf{y}), \mathbf{M}_{k-\lceil g_1\rceil}(g_1\mathbf{y}), \ldots, \mathbf{M}_{k-\lceil g_m\rceil}(g_m\mathbf{y}))$ *and using the upper bound* $\operatorname{trace}(\mathbf{X}) \le \bar{a}_k$ *with*

$$\bar{a}_k := R^k \left(b(n,k) + \sum_{i=1}^m \|g_i\|_1 b(n, k-\lceil g_i\rceil) \right)\,, \tag{4.2.2}$$

SDP (2.4.5) *can be converted to an equivalent SDP with BTP, thanks to the absolute upper bound for each moment variable* $|y_{\boldsymbol{\alpha}}| \le R^{|\boldsymbol{\alpha}|/2}$, $\boldsymbol{\alpha} \in \mathbb{N}^n$. *In principle, we can solve this SDP by applying directly Algorithm 2. However, in our experiments presented in Section 4.3 this method is not only inefficient but also provides output with low accuracy.*

In order to overcome the accuracy issue mentioned in Remark 4.5, we convert every POP to a CTP-POP (i.e., a new POP formulation with CTP) by adding slack variables associated with inequality constraints. In the sequel, we consider three particular cases: equality constrained POPs on a sphere in Section 4.2.1, constrained POPs with single inequality (ball) constraint in Section 4.2.2, and constrained POPs on a ball in Section 4.2.3.

4.2.1 Equality constrained POPs on a sphere

Assume that $m = 0$ and $h_1 = \bar{R} - \|\mathbf{x}\|_2^2$. Note that $\|\mathbf{x}\|_2^2 = x_1^2 + \cdots + x_n^2$ is a quadratic polynomial. In this case, we consider equality constrained POPs on a sphere, presented in Section 4.1.2. We propose to reduce SDP (4.1.6) to an NSOP. For each $k \in \mathbb{N}$, let $(\theta_{k,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_k}$ be a sequence of positive real numbers such that

$$(1 + \|\mathbf{x}\|_2^2)^k = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^n_k} \theta_{k,\boldsymbol{\alpha}} \mathbf{x}^{2\boldsymbol{\alpha}}\,,$$

and define the diagonal matrix

$$\mathbf{P}_{n,k} := \operatorname{diag}((\theta_{k,\boldsymbol{\alpha}}^{1/2})_{\boldsymbol{\alpha}\in\mathbb{N}^n_k})\,. \tag{4.2.3}$$

For every $k \in \mathbb{N}$, since $\mathbf{P}_{n,k} \succ 0$, SDP (4.1.6) is equivalent to SDP:

$$\begin{aligned} \tau_k = \inf_{\mathbf{y}\in\mathbb{R}^{b(n,2k)}} \quad & L_{\mathbf{y}}(f) \\ \text{s.t.} \quad & y_{\mathbf{0}} = 1\,;\, \mathbf{P}_{n,k}\mathbf{M}_k(\mathbf{y})\mathbf{P}_{n,k} \succeq 0\,, \\ & \mathbf{M}_{k-\lceil h_j\rceil}(h_j\,\mathbf{y}) = 0\,,\ j \in [l]\,. \end{aligned} \tag{4.2.4}$$

For every $k \in \mathbb{N}$, let $a_k := (\bar{R}+1)^k$. We will use the following lemma:

Lemma 4.4. *For all* $k \in \mathbb{N}$,

$$\left.\begin{array}{l} \mathbf{M}_{k-1}((\bar{R} - \|\mathbf{x}\|_2^2)\,\mathbf{y}) = 0\,, \\ y_{\mathbf{0}} = 1 \end{array}\right\} \Rightarrow \operatorname{trace}(\mathbf{P}_{n,k}\mathbf{M}_k(\mathbf{y})\mathbf{P}_{n,k}) = a_k\,.$$

Proof. Let $k \in \mathbb{N}$ be fixed. From $\mathbf{M}_{k-1}((\bar{R} - \|\mathbf{x}\|_2^2)\ \mathbf{y}) = 0$, $L_{\mathbf{y}}(p(\bar{R} - \|\mathbf{x}\|_2^2)) = 0$, for every $p \in \mathbb{R}[\mathbf{x}]_{2(k-1)}$. For every $r \in \mathbb{N}^{\le k-1}$, by choosing $p = \|\mathbf{x}\|_2^{2r}$,

$$L_{\mathbf{y}}(\|\mathbf{x}\|_2^{2(r+1)}) = -L_{\mathbf{y}}(\|\mathbf{x}\|_2^{2r}(\bar{R} - \|\mathbf{x}\|_2^2)) + \bar{R}L_{\mathbf{y}}(\|\mathbf{x}\|_2^{2r}) = \bar{R}L_{\mathbf{y}}(\|\mathbf{x}\|_2^{2r})\,.$$

By induction, $L_{\mathbf{y}}(\|\mathbf{x}\|_2^{2r}) = \bar{R}L_{\mathbf{y}}(\|\mathbf{x}\|_2^{2(r-1)}) = \cdots = \bar{R}^k L_{\mathbf{y}}(\|\mathbf{x}\|_2^{2\times 0}) = \bar{R}^k y_{\mathbf{0}} = \bar{R}^r$, for every $r \in \mathbb{N}^{\leq k}$. Thus,

$$\begin{aligned}\operatorname{trace}(\mathbf{P}_{n,k}\mathbf{M}_k(\mathbf{y})\mathbf{P}_{n,k}) &= \sum_{\boldsymbol{\alpha}\in\mathbb{N}^n_k} \theta_{k,\boldsymbol{\alpha}}^{1/2} y_{2\boldsymbol{\alpha}} \theta_{k,\boldsymbol{\alpha}}^{1/2} = L_{\mathbf{y}}\left(\sum_{\boldsymbol{\alpha}\in\mathbb{N}^n_k} \theta_{k,\boldsymbol{\alpha}}\mathbf{x}^{2\boldsymbol{\alpha}}\right)\\ &= L_{\mathbf{y}}((1+\|\mathbf{x}\|_2^2)^k) = L_{\mathbf{y}}\left(\sum_{r=0}^{k}\binom{k}{r}\|\mathbf{x}\|_2^{2r}\right)\\ &= \sum_{r=0}^{k}\binom{k}{r} L_{\mathbf{y}}(\|\mathbf{x}\|_2^{2r}) = \sum_{r=0}^{k}\binom{k}{r}\bar{R}^r = (\bar{R}+1)^k\,.\end{aligned}$$

□

For each $k \in \mathbb{N}$, we denote by $\mathcal{S}^{(k)}$ the set of symmetric matrices of size $b(k)$ and let $\langle \mathbf{A}, \mathbf{B}\rangle = \operatorname{trace}(\mathbf{B}^\top\mathbf{A})$ be the usual scalar product on $\mathcal{S}^{(k)}$. For every $k \in \mathbb{N}$, letting

$$\mathbf{X} = \mathbf{P}_{n,k}\mathbf{M}_k(\mathbf{y})\mathbf{P}_{n,k}\,, \tag{4.2.5}$$

(4.2.4) can be written in the form:

$$-\tau_k = \sup_{\mathbf{X}\in\mathcal{S}^{(k)}} \{\langle \mathbf{C}_k, \mathbf{X}\rangle \,:\, \mathcal{A}_k\mathbf{X} = \mathbf{b}_k\,,\, \mathbf{X} \succeq 0\}\,, \tag{4.2.6}$$

where $\mathcal{A}_k : \mathcal{S}^{(k)} \to \mathbb{R}^{m_k}$ is a linear operator of the form

$$\mathcal{A}_k\mathbf{X} = [\langle \mathbf{A}_{k,1}, \mathbf{X}\rangle, \ldots, \langle \mathbf{A}_{k,m_k}, \mathbf{X}\rangle]\,,$$

with $\mathbf{A}_{k,i} \in \mathcal{S}^{(k)}$, $i \in [m_k]$, $\mathbf{C}_k \in \mathcal{S}^{(k)}$ is the cost matrix and $\mathbf{b}_k \in \mathbb{R}^{m_k}$ is the right-hand-side vector. [134, Appendix A.2] describes how to reduce SDP (4.2.4) to the form (4.2.6).

For every $k \in \mathbb{N}$, the dual of SDP (4.2.6) reads:

$$-\rho_k = \inf_{\mathbf{z}} \{\, \mathbf{b}_k^\top\mathbf{z} \,:\, \mathcal{A}_k^\top\mathbf{z} - \mathbf{C}_k \succeq 0\,,\}\tag{4.2.7}$$

where $\mathcal{A}_k^\top : \mathbb{R}^{m_k} \to \mathcal{S}^{(k)}$ is the adjoint operator of $\mathcal{A}_k$, i.e., $\mathcal{A}_k^\top\mathbf{z} = \sum_{i=1}^{m_k} z_i\mathbf{A}_{k,i}$.

From Lemma 4.4 and since $h_1 = \bar{R} - \|\mathbf{x}\|_2^2$, it implies that for every $k \in \mathbb{N}$,

$$\forall\ \mathbf{X} \in \mathcal{S}^{(k)}\,,\ \mathcal{A}_k\mathbf{X} = \mathbf{b}_k \Rightarrow \operatorname{trace}(\mathbf{X}) = a_k\,. \tag{4.2.8}$$

We guarantee zero duality gap, primal attainability, and dual attainability for primal-dual (4.2.6)-(4.2.7) in the following proposition:

Proposition 4.1. *Let $f^\star$ be as in* (4.1.3)*. Then:*

1. *Zero duality gap holds for primal-dual* (4.2.6)-(4.2.7) *for large enough $k \in \mathbb{N}$.*
2. *SDP* (4.2.6) *has an optimal solution for large enough $k \in \mathbb{N}$.*
3. *Assume that one of the following two conditions holds:*
   (a) *$I(\mathfrak{h})$ is real radical and the second-order sufficiency conditions (Definition 2.1) hold at every global minimizer of* (4.1.3)*;*
   (b) *$V(\mathfrak{h})$ is finite.*

   Then SDP (4.2.7) *has an optimal solution for large enough $k \in \mathbb{N}$. In this case, $\underline{\tau}_k = \underline{\rho}_k = f^\star$.*

Proof. Since (4.1.6) (resp. (4.1.5)) and (4.2.6) (resp. (4.2.7)) are equivalent, the first and second statements follow from Proposition 2.1. The third statement is due to Theorem 4.1. □

By replacing $(\mathcal{A}_k, \mathbf{A}_{k,i}, \mathbf{b}_k, \mathbf{C}_k, \mathcal{S}^{(k)}, b(k), m_k, \tau_k, \rho_k, a_k)$ by $(\mathcal{A}, \mathbf{A}_i, \mathbf{b}, \mathbf{C}, \mathcal{S}, s, m, \tau, \rho, a)$, primal-dual (4.2.6)-(4.2.7) becomes primal-dual (4.1.8)-(4.1.9), we then go back to Section 4.1.3 with $l = 1$.

We illustrate the conversion from SDP (4.1.6) to SDP (4.2.6) in the following example.

Example 4.1. *Consider a simple example of POP* (4.1.3) *with* $n = 1$*:*

$$-1 = \inf\{x \ : \ 1 - x^2 = 0\}\,.$$

Then the second order moment relaxation ($k = 2$) *has the form:*

$$\begin{aligned} \tau_2 = \inf_{\mathbf{y}} \quad & y_1 \\ s.t. \quad & \begin{bmatrix} y_0 & y_1 & y_2 \\ y_1 & y_2 & y_3 \\ y_2 & y_3 & y_4 \end{bmatrix} \succeq 0\,, \ \begin{bmatrix} y_0 - y_2 & y_1 - y_3 \\ y_1 - y_3 & y_2 - y_4 \end{bmatrix} = 0\,, \ y_0 = 1\,. \end{aligned}$$

It can be rewritten as

$$\begin{aligned} \tau_2 = \inf_{\mathbf{y}} \quad & y_1 \\ s.t. \quad & \begin{bmatrix} 1 & y_1 & 1 \\ y_1 & 1 & y_1 \\ 1 & y_1 & 1 \end{bmatrix} \succeq 0\,, \end{aligned}$$

by removing equality constraints. Obviously, the positive semidefinite matrix of this form has trace 3.

In a different way, according to [134, Appendix A.2], let us note

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_0 & y_1 & y_2 \\ y_1 & y_2 & y_3 \\ y_2 & y_3 & y_4 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}\,,$$

to obtain

$$-\tau_2 = \sup_{\mathbf{X} \in \mathcal{S}_2} \{\langle \mathbf{C}, \mathbf{X} \rangle \ : \ \langle \mathbf{A}_i, \mathbf{X} \rangle = b_i\,,\ i \in [5]\,,\ \mathbf{X} \succeq 0\}\,,$$

where $b_1 = \cdots = b_4 = 0$, $b_5 = 1$ *and*

$$\begin{aligned} \mathbf{C} &= -\tfrac{\sqrt{2}}{4} \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\,,\ \mathbf{A}_1 = \tfrac{\sqrt{2}}{2} \begin{bmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix}\,,\ \mathbf{A}_2 = \tfrac{1}{2} \begin{bmatrix} 2 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}\,, \\ \mathbf{A}_3 &= \tfrac{\sqrt{2}}{4} \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix}\,,\ \mathbf{A}_4 = \tfrac{1}{2} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & -2 \end{bmatrix}\,,\ \mathbf{A}_5 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\,. \end{aligned}$$

Remark that for any $\mathbf{X} \in \mathcal{S}^{(2)}$*,*

$$(\langle \mathbf{A}_i, \mathbf{X} \rangle = b_i\,,\ i \in [5]) \Rightarrow \operatorname{trace}(\mathbf{X}) = 4\,.$$

Next, we present an alternative iterative method, stated in Algorithm 3, to solve (4.1.3), based on nonsmooth optimization methods, e.g., LMBM. It performs well in practice for most cases and with significantly lower computational cost when compared to the (currently fastest) SDP solver `Mosek` 9.1.

Algorithm 3 SpectralPOP-CTP

Input: POP (4.1.3) with unknown optimal value $f^\star$ and optimal solutions;
method (D) for solving SDP with CTP.

Output: increasing real sequence $(\tau_k)_{k \in \mathbb{N}}$ and $\mathbf{x}^\star \in \mathbb{R}^n$.

1: **for** $k \in \mathbb{N}$ **do**
2: Compute the optimal value $-\tau_k$ and an optimal solution $\mathbf{X}^\star$ of SDP (4.2.6) by using method (D);
3: Set $\mathbf{M}_k(\mathbf{y}^\star) := \mathbf{P}_{n,k}^{-1} \mathbf{X}^\star \mathbf{P}_{n,k}^{-1}$ (relying on (4.2.5)) and extract an atom $\mathbf{x}^\star$ by using Henrion-Lasserre's algorithm in [77] from $\mathbf{M}_k(\mathbf{y}^\star)$;
4: If $\mathbf{x}^\star$ exists, set $\tau_{k+j} = \tau_k$, $j \in \mathbb{N}^{>0}$, and terminate.

Note that one can choose method (D) in Algorithm 3 as Algorithm 1 with LMBM solver or `SketchyCGAL`.

Remark 4.6. *In practice, to verify that an atom $\mathbf{x}^\star$ extracted in Step 3 of Algorithm 3 is an approximate optimal solution of POP* (4.1.3)*, with given $\varepsilon \in (0,1)$, we check the following inequalities:*

$$|f(\mathbf{x}^\star) - \tau_k| \le \varepsilon \|f\|_{\max} \text{ and } |h_j(\mathbf{x}^\star)| \le \varepsilon \|h_j\|_{\max}\,,\ j \in [m]\,,$$

where $\|p\|_{\max} := \max_{\boldsymbol{\alpha}} |p_{\boldsymbol{\alpha}}|$ for any $p \in \mathbb{R}[\mathbf{x}]$. We take $\varepsilon = 0.01$ for the experiments in Section 4.3.

Following Proposition 2.3, Corollary 4.1 and Proposition 4.1, we obtain the following corollary:

Corollary 4.4. *(i) Sequence $(\tau_k)_{k\in\mathbb{N}}$ of Algorithm 3 is well defined and $\tau_k \uparrow f^\star$ as $k \to \infty$.*
(ii) Assume that condition (a) or (b) of Proposition 4.1.3 holds. If there exists an optimal solution $\mathbf{y}^\star$ of SDP (4.1.6) *for some order $k \in \mathbb{N}$ such that the flat extension condition holds, $\mathbf{x}^\star$ exists at the k-th iteration of Algorithm 3. In this case, if $\mathbf{X}^\star$ in the first step of Algorithm 3 is well-defined, Algorithm 3 terminates at the k-th iteration, $\mathbf{x}^\star$ is an optimal solution of POP* (4.1.3) *and $f^\star = \tau_k$.*

In Corollary 4.4, the flat extension condition implies that the SOS problem (4.1.5) has an optimal solution (due to [102, Theorem 3.4 (b)] and $\tau_k = \rho_k$), so that SDP (4.2.7) has an optimal solution. In this case, $\mathbf{X}^\star$ exists, which in turn implies the existence of $\mathbf{x}^\star$. In Step 4 of Algorithm 3, if the atom $\mathbf{x}^\star$ exists, then we do not need to increase the relaxation order k. It is due to the fact that $f^\star = \tau_k \le \tau_{k+1} \le \cdots \le f^\star$.

Remark 4.7. *When Algorithm 1 with LMBM solver is used for method (D) in Algorithm 3, we have the following cases:*

1. *If the SDP relaxation* (4.1.6) *is exact then the value is $f^\star$ and one indeed may expect that generically the moment matrix is rank-one, which will yield $\mathbf{X}^\star = \mathbf{u}\mathbf{u}^\top$ for some $\mathbf{u}$. Thus, $\mathbf{X}^\star$ in the first step of Algorithm 3 is well-defined.*
2. *If the SDP relaxation* (4.1.6) *is not exact then we only use the relaxation value as a (supposedly accurate) lower bound on the global minimum $f^\star$.*

Remark 4.8. *In practice, we use the following heuristic extraction algorithm when method (D) in Algorithm 3 is Algorithm 1 with LMBM solver:*

1. *Obtain a dual matrix $\mathbf{G}$ corresponding to an optimal solution $\mathbf{X}^\star$ of SDP* (4.2.6) *based on Corollary 4.2;*
2. *Set $\bar{\mathbf{G}} := \mathbf{P}_{n,k}\mathbf{G}\mathbf{P}_{n,k}$;*
3. *Obtain an atom $\mathbf{x}^\star$ by using the extraction algorithm of Henrion and Lasserre in [77], where the matrix $\mathbf{V}$ in [77, (6)] is taken such that the columns of $\mathbf{V}$ form a basis of the null space $\{\mathbf{u} \in \mathbb{R}^{b(k)} : \bar{\mathbf{G}}\mathbf{u} = 0\}$;*
4. *Verify that $\mathbf{x}^\star$ is an approximate optimal solution of POP* (4.1.3) *as in Remark 4.6.*

This heuristic extraction algorithm performs practically well when the moment matrices are not rank-one. Note that $\bar{\mathbf{G}}$ obtained in Step 2 is a Gram matrix corresponding to some moment matrix $\mathbf{M}_k(\mathbf{y}^\star)$. Step 3 is well-defined when the complementary slackness, i.e., $\bar{\mathbf{G}}\ \mathbf{M}_k(\mathbf{y}^\star) = 0$, and the strict complementarity, i.e., $\text{rank}\bar{\mathbf{G}} + \text{rank}\mathbf{M}_k(\mathbf{y}^\star) = b(k)$, hold (see [48, Section 1.3]). In this case, the range of $\mathbf{M}_k(\mathbf{y}^\star)$, which is the linear span of the columns of $\mathbf{V}$ in [77, (6)], is identical with the null space of $\bar{\mathbf{G}}$.

In the two following subsections, we consider POPs on general compact sets as stated in Section 4.1.1.

4.2.2 Constrained POPs with a single inequality (ball) constraint

Assume that $m = 1$ and $g_1 = R - \|\mathbf{x}\|_2^2$. In this case, $\mathfrak{g} = \{R - \|\mathbf{x}\|_2^2\}$. Let us show that POP (4.1.1) can be reduced to an equality constrained POP on a sphere. By adding one slack variable x_{n+1}, the inequality constraint $R - \|\mathbf{x}\|_2^2 \ge 0$ can be rewritten as an equality constraint $R - \|\mathbf{x}\|_2^2 - x_{i+n}^2 = 0$ and so

$$f^\star := \inf\{\, f(\mathbf{x}) \ :\ (\mathbf{x}, x_{n+1}) \in V(\bar{\mathfrak{h}})\}\,, \tag{4.2.9}$$

where $\bar{\mathfrak{h}} := \mathfrak{h} \cup \{R - \|\mathbf{x}\|_2^2 - x_{n+1}^2\} \subset \mathbb{R}[\mathbf{x}, x_{n+1}]$.

Notice that:

- If $\bar{\mathbf{x}}^\star = (\mathbf{x}^\star, x_{n+1}^\star)$ is an optimal solution of POP (4.2.9), $\mathbf{x}^\star$ is an optimal solution of POP (4.1.1).
- Conversely, if $\mathbf{x}^\star$ is an optimal solution of POP (4.1.1), then $\bar{\mathbf{x}}^\star := \left(\mathbf{x}^\star, \sqrt{R - \|\mathbf{x}^\star\|_2^2}\right)$ is an optimal solution of POP (4.2.9).

Let us define $\bar{n} := n + 1$ and $\bar{\mathbf{x}} := (\mathbf{x}, x_{n+1})$ to ease notation. For every $k \in \mathbb{N}$, consider the order k moment relaxation of (4.2.9):

$$\begin{aligned} \bar{\tau}_k = \inf_{\mathbf{y} \in \mathbb{R}^{b(\bar{n}, 2k)}} \quad & L_{\mathbf{y}}(f) \\ \text{s.t.} \quad & y_{\mathbf{0}} = 1\,,\, \mathbf{M}_k(\mathbf{y}) \succeq 0\,, \\ & \mathbf{M}_{k-1}((R - \|\bar{\mathbf{x}}\|_2^2)\,\mathbf{y}) = 0\,, \\ & \mathbf{M}_{k-\lceil h_j \rceil}(h_j\,\mathbf{y}) = 0\,,\, j \in [l]\,. \end{aligned} \tag{4.2.10}$$

The corresponding dual SOS problem indexed by $k \in \mathbb{N}$ reads:

$$\bar{\rho}_k \;:=\; \sup\{\, \xi \in \mathbb{R} \;:\; f - \xi \in \mathcal{P}_k(\bar{\mathfrak{h}})\}\,, \tag{4.2.11}$$

where $\mathcal{P}_k(\bar{\mathfrak{h}})$ is the truncated preodering of all polynomials of the form

$$\sigma_0 + \psi_0(R - \|\bar{\mathbf{x}}\|_2^2) + \sum_{j=1}^{l} \psi_j h_j\,,$$

with $\sigma_0 \in \Sigma[\bar{\mathbf{x}}]_k$, $\psi_0 \in \mathbb{R}[\bar{\mathbf{x}}]_{2(k-1)}$, and $\psi_j \in \mathbb{R}[\bar{\mathbf{x}}]_{2(k-\lceil h_j \rceil)}$, $j \in [l]$.

The following lemma will be used later on:

Lemma 4.5. *If $f - f^\star \in \mathcal{Q}_k(\mathfrak{g}, \mathfrak{h})$ for some $k \in \mathbb{N}$ then $f - f^\star \in \mathcal{P}_k(\bar{\mathfrak{h}})$.*

Proof. By assumption, there exist $\sigma_0 \in \Sigma[\mathbf{x}]_k$, $\sigma_1 \in \Sigma[\mathbf{x}]_{k-1}$, and $\psi_j \in \mathbb{R}[\mathbf{x}]_{2(k-\lceil h_j \rceil)}$, $j \in [l]$ such that

$$f - f^\star = \sigma_0 + \sigma_1(R - \|\mathbf{x}\|_2^2) + \sum_{j=1}^{l} \psi_j h_j = \sigma_0 + \sigma_1 x_{n+1}^2 + \sigma_1(R - \|\bar{\mathbf{x}}\|_2^2) + \sum_{j=1}^{l} \psi_j h_j\,,$$

yielding the result. □

Zero duality gap, primal attainability, and dual attainability for primal-dual (4.2.10)-(4.2.11) are guaranteed in the following proposition:

Proposition 4.2. *Let $f^\star$ be as in (4.1.1) with $g = \{R - \|\mathbf{x}\|_2^2\}$. Then:*

1. *Zero duality gap holds for primal-dual (4.2.10)-(4.2.11) for large enough $k \in \mathbb{N}$.*
2. *SDP (4.2.10) has an optimal solution for large enough $k \in \mathbb{N}$.*
3. *Assume that one of the following two conditions holds:*
   (a) *$\mathcal{Q}(\mathfrak{g}, \mathfrak{h})$ is Archimedean, the ideal $I(\mathfrak{h})$ is real radical, and the second-order sufficiency conditions (Definition 2.1) hold at every global minimizer of POP (4.1.1);*
   (b) *$V(\mathfrak{h})$ is finite.*

   Then SDP (4.2.11) has an optimal solution for large enough $k \in \mathbb{N}$. In this case, $\bar{\tau}_k = \bar{\rho}_k = f^\star$.

Proof. The first and second statement follow from Proposition 2.1, after replacing $S(\mathfrak{g}, \mathfrak{h})$ by $V(\bar{\mathfrak{h}})$. The third statement is due to Proposition 2.4 and Lemma 4.5. □

For every $k \in \mathbb{N}$, according to Lemma 4.4, if $\mathbf{M}_{k-1}((R - \|\bar{\mathbf{x}}\|_2^2)\ \mathbf{y}) = 0$ and $y_{\mathbf{0}} = 1$, then one has

$$\operatorname{trace}(\mathbf{P}_{\bar{n},k}\mathbf{M}_k(\mathbf{y})\mathbf{P}_{\bar{n},k}) = (R+1)^k\,, \tag{4.2.12}$$

where $\mathbf{P}_{\bar{n},k}$ is defined as in (4.2.3) after replacing n by $\bar{n}$. Thus SDP (4.2.10) has the CTP. We now do a similar process as in Section 4.2.1.

Next, we present an iterative method, stated in Algorithm 4, to solve (4.1.1) with $\mathfrak{g} = \{R - \|\mathbf{x}\|_2^2\}$, based on a nonsmooth optimization method such as LMBM.

Algorithm 4 SpectralPOP-CTP-WithSingleBallConstraint

Input: POP (4.1.1) with $\mathfrak{g} = \{R - \|\mathbf{x}\|_2^2\}$, unknown optimal value $f^\star$ and optimal solutions; method (D) for solving SDP with CTP.

Output: increasing real sequence $(\bar{\tau}_k)_{k\in\mathbb{N}}$ and $\mathbf{x}^\star \in \mathbb{R}^n$.

1: **for** $k \in \mathbb{N}$ **do**
2: Compute the optimal value $-\bar{\tau}_k$ and an optimal solution $\mathbf{y}^\star$ of SDP (4.2.10) with CTP (4.2.12) by using method (D);
3: Extract an atom $\bar{\mathbf{x}}^\star = (\mathbf{x}^\star, x^\star_{n+1})$ by using Henrion-Lasserre's algorithm in [77] from $\mathbf{M}_k(\mathbf{y}^\star)$;
4: If $\bar{\mathbf{x}}^\star$ exists, set $\bar{\tau}_{k+j} = \bar{\tau}_k$, $j \in \mathbb{N}^{>0}$, and terminate.

Note that one can choose method (D) in Algorithm 4 as Algorithm 1 with LMBM solver or `SketchyCGAL`.

Following Proposition 2.3, Corollary 4.1 and Proposition 4.2, we obtain the following corollary:

Corollary 4.5. *(i) Sequence $(\bar{\tau}_k)_{k\in\mathbb{N}}$ of Algorithm 4 is well defined and $\bar{\tau}_k \uparrow f^\star$ as $k \to \infty$.*
(ii) Assume that condition (a) or (b) of Proposition 4.2.3 holds. If there exists an optimal solution $\mathbf{y}^\star$ of SDP (4.2.10) *for some order $k \in \mathbb{N}$ such that the flat extension condition holds, $\mathbf{x}^\star$ exists at the k-th iteration of Algorithm 4. In this case, if Algorithm 4 terminates at the k-th iteration, $\mathbf{x}^\star$ is an optimal solution of POP* (4.1.1) *and $f^\star = \bar{\tau}_k$.*

4.2.3 Constrained POPs on a ball

Assume that $m > 1$ and $g_1 = R - \|\mathbf{x}\|_2^2$. Let us show that POP (4.1.1) can be reduced to an equality constrained POP on a sphere. After adding m slack variables x_{n+i}, $i \in [m]$, every inequality constraint $g_i(\mathbf{x}) \geq 0$ can be rewritten as an equality constraint $g_i(\mathbf{x}) = x_{i+n}^2$ and so

$$f^\star := \inf\{\, f(\mathbf{x}) \ :\ (\mathbf{x}, x_{n+1}, \ldots, x_{n+m}) \in V(\hat{\mathfrak{h}})\}\,,$$

where $\hat{\mathfrak{h}} := \mathfrak{h} \cup \{g_i - x_{i+n}^2 : i \in [m]\} \subset \mathbb{R}[\mathbf{x}, x_{n+1}, \ldots, x_{n+m}]$.

Let us take upper bounds $b_i \geq \sup\{g_i(\mathbf{x}) : x \in S(\{g_1\}, \mathfrak{h})\}$, $i \in [m]$. For every $i \in [m]$, the bound b_i can be computed by solving the order k moment relaxation:

$$\begin{aligned} -b_i = \inf_{\mathbf{y}\in\mathbb{R}^{b(n+1,2k)}} \quad & L_{\mathbf{y}}(-g_i) \\ \text{s.t.} \quad & y_{\mathbf{0}} = 1\,,\ \mathbf{M}_k(\mathbf{y}) \succeq 0\,, \\ & \mathbf{M}_{k-1}((R - \|(\mathbf{x}, x_{n+1})\|_2^2)\ \mathbf{y}) = 0\,, \\ & \mathbf{M}_{k-\lceil h_j\rceil}(h_j\ \mathbf{y}) = 0\,,\ j \in [l]\,, \end{aligned} \tag{4.2.13}$$

based on the spectral minimization method presented in the previous section.

For every $(\mathbf{x}, x_{n+1}, \ldots, x_{n+m}) \in V(\hat{\mathfrak{h}})$, $x \in S(\mathfrak{g}, \mathfrak{h})$ and $x_{n+i}^2 = g_i(\mathbf{x}) \leq b_i$, $i \in [m]$, since $S(\mathfrak{g}, \mathfrak{h}) \subset S(\{g_1\}, \mathfrak{h})$. Therefore

$$\|\mathbf{x}\|_2^2 + \sum_{i=1}^m x_{n+i}^2 \leq \bar{R} \quad \text{with} \quad \bar{R} := R + \sum_{i=1}^m b_i\,. \tag{4.2.14}$$

Equivalently $V(\hat{\mathfrak{h}}) \subset B_{\bar{R}}^{n+m}$ and after adding one more slack variable x_{n+m+1}:

$$f^\star := \inf\{f(\mathbf{x}) \ :\ \bar{\mathbf{x}} \in V(\bar{\mathfrak{h}})\}\,, \tag{4.2.15}$$

where $\bar{\mathbf{x}} := (\mathbf{x}, x_{n+1}, \ldots, x_{n+m+1})$ and

$$\bar{\mathfrak{h}} := \hat{\mathfrak{h}} \cup \{\bar{R} - \|\bar{\mathbf{x}}\|_2^2\} \;=\; \mathfrak{h} \cup \{g_i - x_{i+n}^2 : i \in [m]\} \cup \{\bar{R} - \|\bar{\mathbf{x}}\|_2^2\} \;\subset \mathbb{R}[\bar{\mathbf{x}}]\,.$$

Notice that:

- If $\bar{\mathbf{x}}^\star = (\mathbf{x}^\star, x_{n+1}^\star, \ldots, x_{n+m+1}^\star)$ is an optimal solution of POP (4.2.15), $\mathbf{x}^\star$ is an optimal solution of POP (4.1.1).
- Conversely, if $\mathbf{x}^\star$ is an optimal solution of POP (4.1.1), then

$$\bar{\mathbf{x}}^\star := \left(\mathbf{x}^\star, \sqrt{g_1(\mathbf{x}^\star)}, \ldots, \sqrt{g_m(\mathbf{x}^\star)}, \sqrt{\bar{R} - \sum_{i=1}^m g_i(\mathbf{x}^\star) - \|\mathbf{x}^\star\|_2^2}\right)$$

  is an optimal solution of POP (4.2.15).

Note $\bar{n} := n + m + 1$ for simplicity. For every $k \in \mathbb{N}$, consider the order k moment relaxation of (4.2.15):

$$\begin{array}{rll} \bar{\tau}_k = \inf\limits_{\mathbf{y} \in \mathbb{R}^{b(\bar{n},2k)}} & L_{\mathbf{y}}(f) & \\ \text{s.t.} & y_{\mathbf{0}} = 1\,,\, \mathbf{M}_k(\mathbf{y}) \succeq 0\,, & \\ & \mathbf{M}_{k-\lceil g_i \rceil}((g_i - x_{n+i}^2)\,\mathbf{y}) = 0\,,\, i \in [m]\,, & \\ & \mathbf{M}_{k-1}((\bar{R} - \|\bar{\mathbf{x}}\|_2^2)\,\mathbf{y}) = 0\,, & \\ & \mathbf{M}_{k-\lceil h_j \rceil}(h_j\,\mathbf{y}) = 0\,,\, j \in [l]\,. & \end{array} \tag{4.2.16}$$

The corresponding dual SOS problem indexed by $k \in \mathbb{N}$ reads:

$$\bar{\rho}_k \;:=\; \sup\{\,\xi \in \mathbb{R} \;:\; f - \xi \in \mathcal{P}_k(\bar{\mathfrak{h}})\}\,, \tag{4.2.17}$$

where $\mathcal{P}_k(\bar{\mathfrak{h}})$ is the truncated preodering of all polynomials of the form

$$\sigma_0 + \sum_{i=1}^m \psi_i(g_i - x_{n+i}^2) + \psi_{m+1}(\bar{R} - \|\bar{\mathbf{x}}\|_2^2) + \sum_{j=1}^l \psi_{m+1+j} h_j$$

with $\sigma_0 \in \Sigma[\bar{\mathbf{x}}]_k$, $\psi_i \in \mathbb{R}[\bar{\mathbf{x}}]_{2(k-\lceil g_i \rceil)}$, $i \in [m]$, $\psi_{m+1} \in \mathbb{R}[\bar{\mathbf{x}}]_{2(k-1)}$, and $\psi_{m+1+j} \in \mathbb{R}[\bar{\mathbf{x}}]_{2(k-\lceil h_j \rceil)}$, $j \in [l]$.

We will use the following lemma later on:

Lemma 4.6. *If $f - f^\star \in \mathcal{Q}_k(\mathfrak{g}, \mathfrak{h})$ for some $k \in \mathbb{N}$ then $f - f^\star \in \mathcal{P}_k(\bar{\mathfrak{h}})$.*

Proof. By assumption, there exist $\sigma_0 \in \Sigma[\mathbf{x}]_k$, $\sigma_i \in \Sigma[\mathbf{x}]_{k-\lceil g_i \rceil}$, $i \in [m]$, and $\psi_j \in \mathbb{R}[\mathbf{x}]_{2(k-\lceil h_j \rceil)}$, $j \in [l]$ such that

$$f - f^\star = \sigma_0 + \sum_{i=1}^m \sigma_i g_i + \sum_{j=1}^l \psi_j h_j\,.$$

It implies that

$$f - f^\star = \sigma_0 + \sum_{i=1}^m \sigma_i x_{i+n}^2 + \sum_{i=1}^m \sigma_i(g_i - x_{i+n}^2) + 0 \times (\bar{R} - \|\bar{\mathbf{x}}\|_2^2) + \sum_{j=1}^l \psi_j h_j\,,$$

yielding the result. □

Zero duality gap, primal attainability, and dual attainability for primal-dual (4.2.16)-(4.2.17) are guaranteed in the following proposition:

Proposition 4.3. *Let $f^\star$ be as in (4.1.1). Then:*

1. *Zero duality gap holds for primal-dual (4.2.16)-(4.2.17) for large enough $k \in \mathbb{N}$.*
2. *SDP (4.2.16) has an optimal solution for large enough $k \in \mathbb{N}$.*

3. Assume one of the following two conditions holds:

(a) $\mathcal{Q}(\mathfrak{g},\mathfrak{h})$ is Archimedean, the ideal $\mathcal{I}(\mathfrak{h})$ is real radical, and the second-order sufficiency conditions (Definition 2.1) hold at every global minimizer of POP (4.1.1)*;*

(b) $V(\mathfrak{h})$ is finite.

Then SDP (4.2.17) *has an optimal solution for large enough $k \in \mathbb{N}$. In this case, $\bar{\tau}_k = \bar{\rho}_k = f^\star$.*

Proof. The first and second statement follow from to Proposition 2.1 after replacing $S(\mathfrak{g},\mathfrak{h})$ by $V(\bar{\mathfrak{h}})$. The third statement is due to Proposition 2.4 and Lemma 4.6. □

For every $k \in \mathbb{N}$, according to Lemma 4.4, if $\mathbf{M}_{k-1}((\bar{R} - \|\bar{\mathbf{x}}\|_2^2)\,\mathbf{y}) = 0$ and $y_{\mathbf{0}} = 1$,

$$\operatorname{trace}(\mathbf{P}_{\bar{n},k}\mathbf{M}_k(\mathbf{y})\mathbf{P}_{\bar{n},k}) = (\bar{R}+1)^k\,, \tag{4.2.18}$$

where $\mathbf{P}_{\bar{n},k}$ is defined as in (4.2.3) with n replaced by $\bar{n}$. Thus SDP (4.2.16) has the CTP. It remains to follow a process which is similar to the one from Section 4.2.1.

Next, we present an iterative method, stated in Algorithm 5, to solve POP (4.1.1) with $\mathfrak{g} = \{R - \|\mathbf{x}\|_2^2\}$, based on nonsmooth optimization methods such as LMBM.

Algorithm 5 SpectralPOP-CTP-WithBallConstraint

Input: POP (4.1.1) with $g_1 = R - \|\mathbf{x}\|_2^2$, unknown optimal value $f^\star$ and optimal solutions; method (D) for solving SDP with CTP.
Output: increasing real sequence $(\bar{\tau}_k)_{k\in\mathbb{N}}$ and $\mathbf{x}^\star \in \mathbb{R}^n$.

1: **for** $k \in \mathbb{N}$ **do**
2: Compute the optimal value b_i of SDP (4.2.13) with CTP, $i \in [m]$, by using method (D) and set $\bar{R} := R + \sum_{i=1}^m b_i$;
3: Compute the optimal value $-\bar{\tau}_k$ and an optimal solution $\mathbf{y}^\star$ of SDP (4.2.16) with CTP (4.2.18) by using method (D);
4: Extract an atom $\bar{\mathbf{x}}^\star = (\mathbf{x}^\star, x^\star_{n+1}, \ldots, x^\star_{n+m+1})$ by using Henrion-Lasserre's algorithm in [77] from $\mathbf{M}_k(\mathbf{y}^\star)$;
5: If $\bar{\mathbf{x}}^\star$ exists, set $\bar{\tau}_{k+j} = \bar{\tau}_k$, $j \in \mathbb{N}^{>0}$, and terminate.

As in the single (ball) constraint case, one can choose method (D) in Algorithm 5 as Algorithm 1 with LMBM solver or `SketchyCGAL`.

Following Proposition 2.3, Corollary 4.1 and Proposition 4.3, we obtain the following corollary:

Corollary 4.6. *(i) The sequence $(\bar{\tau}_k)_{k\in\mathbb{N}}$ of Algorithm 5 is well defined and $\bar{\tau}_k \uparrow f^\star$ as $k \to \infty$. (ii) Assume that either condition (a) or condition (b) of Proposition 4.3.3 holds. If there exists an optimal solution $\mathbf{y}^\star$ of SDP* (4.2.16) *at order $k \in \mathbb{N}$ such that the flat extension condition holds, then $\mathbf{x}^\star$ exists at the k-th iteration of Algorithm 5 . In this case, Algorithm 5 terminates at the k-th iteration, $\mathbf{x}^\star$ is an optimal solution of POP* (4.1.1) *and $f^\star = \bar{\tau}_k$.*

4.3 Numerical experiments

Let us report numerical results obtained while relying on algorithms from Section 4.2 to solve equality constrained QCQPs on a sphere, quartic minimization problems on the unit sphere as well as further applications to three well-known NP-hard optimization problems on the unit sphere : deciding nonnegativity/convexity of even degree forms and copositivity of real symmetric matrices.

The experiments are performed in Julia 1.3.1 with the following packages:

- SumOfSquare.jl [216] is a modeling library to write and solve SDP relaxations of POPs, based on JuMP.jl and the SDP solver `Mosek` 9.1.
- LMBM.jl solves unconstrained NSOPs with the limited-memory bundle method of Haarala et al. [71, 70]. LMBM.jl calls Karmitsa's Fortran implementation of LMBM algorithm [95].

Table 4.1: Notation

n	the number of variables of the POP
m	the number of inequality constraints of the POP
l	the number of equality constraints of the POP
k	the order of the Moment-SOS relaxation or the iteration of Algorithm 3
s	the size of the positive semidefinite matrix involved in the SDP relaxation
m	the number of trace equality constraints of the SDP relaxation
`SumOfSquares`	SDP relaxation modeled by SumOfSquares.jl and solved by `Mosek` 9.1
CTP	the method described either in Section 4.2.1, Section 4.2.2 or Section 4.2.3
BTP	the method described in Remark 4.5
LMBM	SDP relaxation solved by spectral minimization, described in Section 4.1.3 with the LMBM solver
`SketchyCGAL`	SDP relaxation solved by `SketchyCGAL`
SpectralPOP	SDP relaxation handled by CTP or BTP method, with LMBM or `SketchyCGAL` solver
val	the optimal value of the SDP relaxation
gap	the relative optimality gap w.r.t. `SumOfSquares`, defined by $\text{gap} = \frac{\lvert \text{val} - \text{val}(\texttt{SumOfSquares})\rvert}{\lvert \text{val}(\texttt{SumOfSquares})\rvert}$
*	there exists at least one optimal solution of the POP, which can extracted by Henrion-Lasserre's algorithm in [77]
time	the total computation time of the SDP relaxation in seconds
$-$	the calculation did not finish in 3000 seconds or ran out of memory

- `SketchyCGAL` is a MATLAB package to handle SDP problems with CTP/BTP, implemented by Yurtsever et al. [219]. We have implemented a Julia version (`SketchyCGAL`.jl) of `SketchyCGAL` to ensure fair comparison with LMBM.jl and SumOfSquare.jl. In this section, `SketchyCGAL` is used as a solver for SDP (4.2.6) in Algorithm 3 instead of Algorithm 1 or 2.

We also use the package Arpack.jl, which is based on the implicitly restarted Lanczos's algorithm, to compute the largest eigenvalues and the corresponding eigenvectors of real symmetric matrices of (potentially) large size.

When POPs have equality constraints, SumOfSquare.jl uses reduced forms with Groebner basis instead of creating SOS multipliers, in order to reduce solving time.

The implementation of algorithms described in Section 4.2 can be downloaded from the link: https://github.com/maihoanganh/SpectralSOS.

We use a desktop computer with an Intel(R) Core(TM) i7-8665U CPU @ 1.9GHz $\times$ 8 and 31.2 GB of RAM. The notation for our numerical results are given in Table 4.1.

4.3.1 Random dense equality constrained QCQPs on the unit sphere

Test problems: We construct several instances of POP (4.1.3) as follows:

1. Take $h_1 = 1 - \|\mathbf{x}\|_2^2$ and choose f, h_j, $j \in [l]\backslash\{1\}$ with degrees at most 2;
2. Each coefficient of the objective function f is taken randomly in $(-1, 1)$ with respect to the uniform distribution;
3. Select a random point $\mathbf{a} \in \mathbb{R}^n$ in the unit sphere;
4. For every $j \in [l]\backslash\{1\}$, all non-constant coefficients of h_j are taken randomly in $(-1, 1)$ with respect to the uniform distribution, and the constant coefficient of h_j is chosen such that $h_j(\mathbf{a}) = 0$.

Note that by construction, $\mathbf{a}$ is a feasible solution. We use the method presented in Section 4.2.1 (actually the k-th iteration of Algorithm 3) to solve these problems. Numerical results are displayed

Table 4.2: Numerical results for random dense equality constrained QCQPs on the unit sphere, described in Section 4.3.1, with $(m, l) = (0, 1)$ and $k = 1$.

POP size	`SumOfSquares` (`Mosek`)		SpectralPOP (CTP)			
			LMBM		`SketchyCGAL`	
n	val	time	val	time	val	time
50	-6.24844*	0.3	-6.24844*	0.7	-6.11351	0.7
75	-7.25323*	2	-7.25326*	0.7	-6.95325	0.8
100	-7.00957*	8	-7.00957*	0.9	-6.75991	1
125	-9.76963*	23	-9.76963*	1	-9.39907	1
150	-8.49449*	64	-8.49449*	1	-8.15382	2
175	-10.7286*	140	-10.72866*	1	-10.1323	2
200	-11.3521*	300	-11.3521*	2	-10.4724	3
250	-13.8881*	1152	-13.8881*	4	-13.5571	5
300	-13.9957*	3708	-13.9958*	6	-13.8327	12
400	–	–	-15.7584*	15	-15.5036	28
500	–	–	-17.5838*	35	-17.2513	65
700	–	–	-22.3584*	218	-22.0710	355
900	–	–	-25.6117*	621	-25.2435	947
1200	–	–	-28.3170*	1401	-27.8270	2074
1500	–	–	-30.8475*	7120	-30.2347	9020

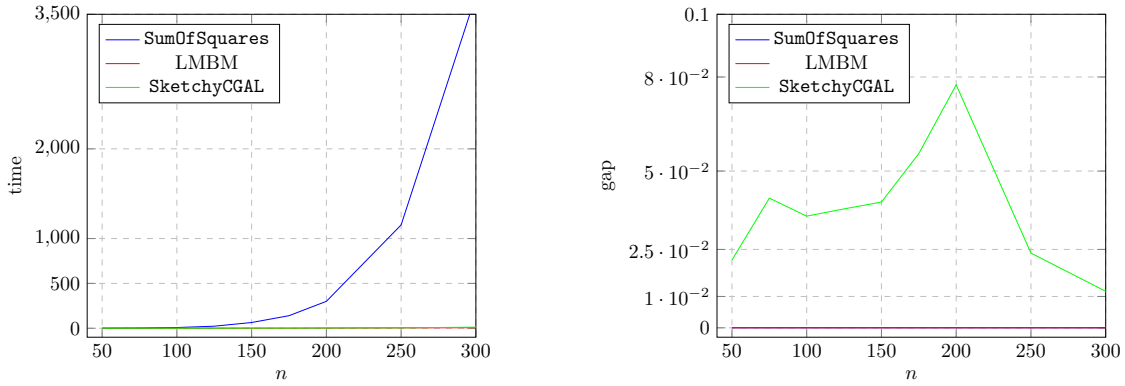


Figure 4.1: Efficiency and accuracy comparison for Table 4.2.

in Table 4.2 for the case $l = 1$ and Table 4.3, 4.4 for the case $l = \lceil n/4 \rceil$. For these results, we use the Julia version of `SketchyCGAL`, which runs much faster than the MATLAB version without compromising accuracy.

Efficiency comparison: In Table 4.2, we minimize quadratic polynomials on the unit sphere. The SDP relaxation for a POP in n variables involves a matrix of size $n + 1$ and 2 trace equality constraints. In this table, LMBM is the fastest SDP solver while `Mosek` (the SDP solver used by `SumOfSquares`) is the slowest. It is due to the fact that `Mosek` relies on interior-point methods based on second order conditions to solve SDP while LMBM and `SketchyCGAL` only rely on algorithms based on first order conditions. Note that we use the same modeling technique to generate the SDP-CTP relaxation solved with either `SketchyCGAL` or LMBM, so both related modeling times are the same. The solving time of `SketchyCGAL` is a bit larger than the one of LMBM in this case.

In Table 4.3 and Table 4.4, we consider random equality constrained QCQPs and solve their first ($k = 1$) and second ($k = 2$) order moment relaxation, respectively. In Table 4.3, the size of the positive semidefinite matrix (resp. the number of trace equality constraints) involved in the SDP relaxation is equal to $n + 1$ (resp. $l + 1$). In Table 4.4, the matrices involved in the SDP relaxation have size $b(n, 4)$ and the number of trace equality constraints is $\mathcal{O}(b(n, 4)^2)$, due to [134, Appendix A.2]. Thus, the number of trace equality constraints for these SDP relaxations

Table 4.3: Numerical results for random dense equality constrained QCQPs on the unit sphere, described in Section 4.3.1, with $(m, l) = (0, \lceil n/4 \rceil)$ and $k = 1$.

POP size		SumOfSquares (Mosek)		SpectralPOP (CTP)			
				LMBM		SketchyCGAL	
n	l	val	time	val	time	val	time
50	14	-4.26516	0.4	-4.26516	1	-4.24511	1
60	16	-6.42900*	1	-6.42929*	1	-6.36177	2
70	19	-5.08320	3	-5.08322	2	-5.01911	3
80	21	-5.35178	5	-5.35178	2	-5.29900	4
100	26	-7.50097	15	-7.50097	10	-7.42432	11
120	31	-5.89903	33	-5.89903	12	-5.81244	18
150	39	-7.44920	127	-7.44921	26	-7.32154	36
200	51	-8.93976	363	-8.93976	51	-8.79487	71
300	76	-12.4295	3753	-12.4295	530	-12.2180	480
400	101	–	–	-14.7190	2553	-14.4830	2318

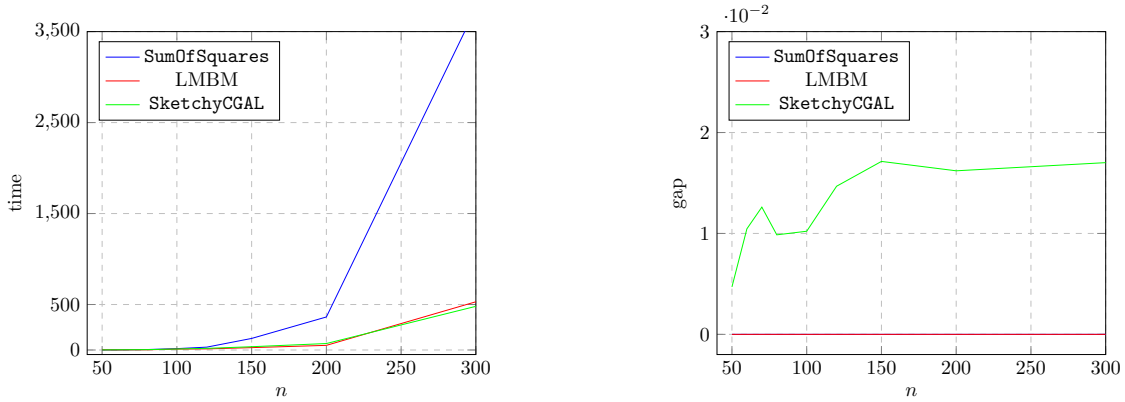


Figure 4.2: Efficiency and accuracy comparison for Table 4.3.

Table 4.4: Numerical results for random dense equality constrained QCQPs on the unit sphere, described in Section 4.3.1, with $(m, l) = (0, \lceil n/4 \rceil)$ and $k = 2$.

POP size		SDP size		SumOfSquares (Mosek)		SpectralPOP (CTP)			
						LMBM		SketchyCGAL	
n	l	s	m	val	time	val	time	val	time
5	3	21	169	-2.43822*	0.04	-2.43823*	1	-2.43982	1
10	4	66	1475	-1.42006*	0.4	-1.42013*	1	-1.41268	1
15	5	136	6121	-2.87129*	7	-2.87142*	2	-2.86744	6
20	6	231	17557	-3.28734*	73	-3.28736*	5	-3.27733	26
25	8	351	40834	-3.32902*	592	-3.32918*	13	-3.31634	65
30	8	496	81345	-4.34398*	4678	-4.34407*	60	-4.32974	294
35	10	666	146521	–	–	-4.77580*	275	-4.75946	450
40	11	861	244812	–	–	-2.95099	390	-2.91856	1225
45	13	1081	386999	–	–	-3.95743	1588	-3.88533	2905
50	14	1326	582115	–	–	-4.01846	6126	–	–

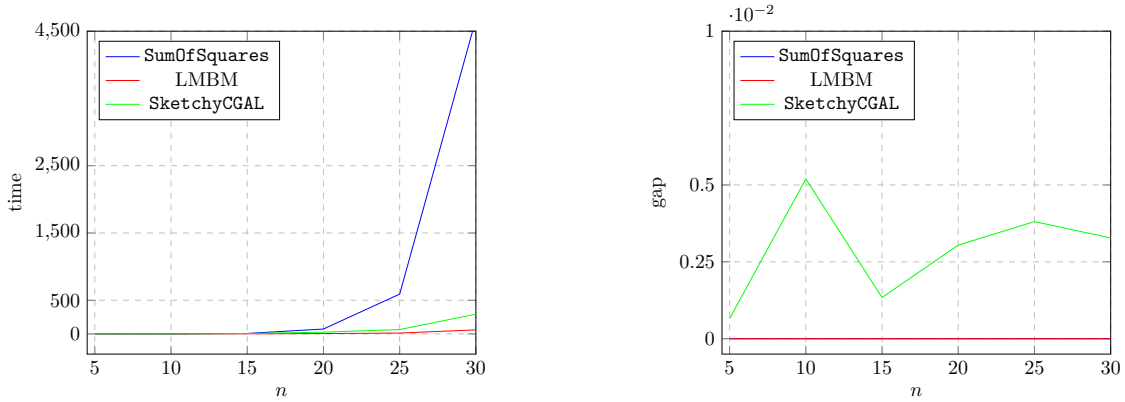


Figure 4.3: Efficiency and accuracy comparison for Table 4.4.

is more than 200 times larger than the matrix size, for almost all instance of Table 4.4. LMBM and `SketchyCGAL` still happen to be faster than `SumOfSquares` in both Table 4.3 and Table 4.4, but LMBM is not much more efficient than `SketchyCGAL`. The most expensive step performed by `Mosek` (used by `SumOfSquares`) is to solve a system of linear equations coming from certain complementarity conditions (see page 13 in [41] for more details). The linear system becomes harder to solve when the number of trace equality constraints is larger. This is in contrast with LMBM, which does not need to solve any such large size linear system of equations. By comparison with LMBM, `SketchyCGAL` may perform a larger number of operations [219, Algorithm 6.1], as emphasized later on.

Accuracy comparison: When $n \leq 300$ in Table 4.2, $n \leq 300$ in Table 4.3 or $n \leq 30$ in Table 4.4, LMBM converges to the exact optimal value of POPs with high accuracy, similarly to `SumOfSquares`. Both LMBM and `SumOfSquares` can extract at least one approximate optimal solution by Henrion-Lasserre's algorithm [77], when $n \leq 300$ in Table 4.2 or $n \leq 35$ in Table 4.4. Moreover, LMBM can provide an approximate optimal solution even for large-scale problems with $n = 1500$ in Table 4.2 (resp. $n = 40$ in Table 4.4) and in a case in Table 4.3. Unfortunately `SketchyCGAL` cannot do the extraction procedure successfully, because of its inaccurate output.

Storage and evaluation comparisons: In Table 4.5 and 4.6, we display some additional information related to `Mosek`, LMBM and `SketchyCGAL`, for the rows $n = 5, 10, 15, 20, 25$ of Table 4.4:

- storage;
- $\#\mathcal{A}$: the number of evaluations of the linear operator $\mathcal{A}$ in SDP (4.1.8);
- $\#\mathcal{A}^\top$: the number of evaluations of the adjoint operator $\mathcal{A}^\top$;
- $s_{\max}$: the largest size of symmetric matrices of which eigenvalues and eigenvectors are computed;
- N_{eig}: the number of symmetric matrices of which eigenvalues and eigenvectors are computed.

Table 4.5 indicates that `SumOfSquares` requires a bit lower storage than LMBM only for the cases $n = 5, 10$. However, `SketchyCGAL` requires a bit smaller storage than LMBM. Note that `SketchyCGAL` performs a large number of evaluations of both $\mathcal{A}$ and $\mathcal{A}^\top$ while relying on three specific primitive computations (see [219, Section 2.3]). Compared to `SketchyCGAL`, LMBM performs a smaller number of evaluations of both $\mathcal{A}$ and $\mathcal{A}^\top$. For instance, the number of evaluations of LMBM is ten times smaller than the one of `SketchyCGAL` for the row $n = 25$ of Table 4.6. Because of the large number m of trace equality constraints, the evaluations of $\mathcal{A}$ and $\mathcal{A}^\top$ in SDP relaxations of POPs is more expensive than the simple one related to the first order SDP relaxation of MAXCUT, which is solved very efficiently by `SketchyCGAL` (see [219, Section 2.5]).

Table 4.5: Storage comparisons for the rows $n = 5, 10, 15, 20, 25$ of Table 4.4.

	`SumOfSquares` (`Mosek`)	SpectralPOP (CTP)	
		LMBM	`SketchyCGAL`
n	storage	storage	storage
5	80.21 k	1.19 M	679.48 k
10	803.00 k	1.02 M	881.75 k
15	3.74 M	2.44 M	2.29 M
20	12.06 M	6.51 M	6.46 M
25	32.67 M	19.04 M	19.02 M

Table 4.6: Evaluation comparisons for the rows $n = 5, 10, 15, 20, 25$ of Table 4.4.

	SpectralPOP (CTP)							
	LMBM				`SketchyCGAL`			
n	$\#\mathcal{A}$	$\#\mathcal{A}^\top$	$s_{\max}$	N_{eig}	$\#\mathcal{A}$	$\#\mathcal{A}^\top$	$s_{\max}$	N_{eig}
5	18	19	21	19	191	2036	12	192
10	26	27	66	27	108	1350	14	109
15	404	405	136	405	205	3383	19	206
20	469	470	231	470	328	6605	24	329
25	336	337	351	337	292	6134	25	293

These specific behaviors mainly come from the subroutines used by LMBM and `SketchyCGAL` to compute eigenvalues and eigenvectors. While LMBM computes directly the largest eigenvector (and corresponding eigenvalue) of the matrix $\mathbf{C}-\mathcal{A}^\top\mathbf{z}$ involved in the nonsmooth function from (4.1.13), `SketchyCGAL` computes indirectly the smallest eigenvalue of the matrix $\mathbf{C}+\mathcal{A}^\top(\mathbf{y}+\beta(\mathbf{z}-\mathbf{b}))$ in Step 8 of [219, Algorithm 6.1] while relying on the so-called "ApproxMinEvec" subroutine. When the ApproxMinEvec subroutine is implemented via [219, Algorithm 4.2], `SketchyCGAL` provides approximations of the smallest eigenvalue and eigenvector of each matrix $\mathbf{C}+\mathcal{A}^\top(\mathbf{y}+\beta(\mathbf{z}-\mathbf{b}))$ by using the randomized Lanczos method. It only requires to compute the smallest eigenvalue and eigenvector of a tridiagonal matrix of small size (e.g. $s_{\max} = 42$ when $n = 25$ in Table 4.6 while the value $s_{\max}$ of LMBM is 351). Besides, `SketchyCGAL` computes $\mathbf{v}_i^\top(\mathbf{C}+\mathcal{A}^\top(\mathbf{y}+\beta(\mathbf{z}-\mathbf{b}))\mathbf{v}_i$ [1] within the loop from Step 5 of [219, Algorithm 4.2] while relying on three primitive computations (see [219, (2.4)] for more details), which yields a large number of evaluations of $\mathcal{A}^\top$. For instance, $\#\mathcal{A}^\top = 6134$ for `SketchyCGAL` when $n = 25$ while the value $\#\mathcal{A}^\top$ is 337 for LMBM. Because of its slow convergence, `SketchyCGAL` also performs a larger number of iterations in Step 6 of [219, Algorithm 6.1].

Based on the above comparison, we emphasize that LMBM is cheaper and faster than `Mosek` or `SketchyCGAL` while LMBM ensures the same accuracy as `Mosek` when solving SDP relaxations of equality constrained QCQPs on the unit sphere.

4.3.2 Random dense QCQPs on the unit ball

Test problems: We construct several samples of POP (4.1.1) as follows:

1. Take $g_1 = 1 - \|\mathbf{x}\|_2^2$ and choose f, g_i, $i \in [m]\backslash\{1\}$, and h_j, $j \in [l]$ with degrees at most 2;
2. Each coefficient of the objective function f is taken randomly in $(-1, 1)$ with respect to the uniform distribution;
3. Select a random point $\mathbf{a} \in \mathbb{R}^n$ in the unit ball, with respect to the uniform distribution;
4. For each $i \in [m]\backslash\{1\}$, all non-constant coefficients of g_i are taken randomly in $(-1, 1)$ with respect to the uniform distribution, and the constant coefficient of g_i is chosen such that

[1] the vector $\mathbf{v}_i$ is updated in Step 6 of [219, Algorithm 4.2]

Table 4.7: Numerical results of random dense QCQPs on the unit ball, described in Section 4.3.2, with $(m, l) = (1, \lceil n/4 \rceil)$, and $k = 2$.

POP size		SDP size (CTP)		`SumOfSquares` (`Mosek`)		SpectralPOP			
						CTP (LMBM)		BTP (LMBM)	
n	l	s	m	val	time	val	time	val	time
5	2	28	281	-0.33125*	0.07	-0.33126*	1	-0.98254	0.7
10	3	78	2029	-2.30410*	0.5	-2.30411*	1	-3.64371	5
15	4	153	7702	-2.26182*	10	-2.26195*	2	-4.21202	134
20	5	253	21000	-2.24031*	112	-2.24033*	4	-5.77860	1722
25	7	378	47251	-2.88952*	1484	-2.88770*	15	-6.81243	16185
30	7	528	92049	-4.15791*	4694	-4.15798*	49	–	–
35	9	703	163097	–	–	-4.10015	150	–	–
40	10	903	269095	–	–	-4.47927	694	–	–
45	12	1128	421121	–	–	-5.50988	849	–	–
50	13	1378	628369	–	–	-5.52884	2086	–	–

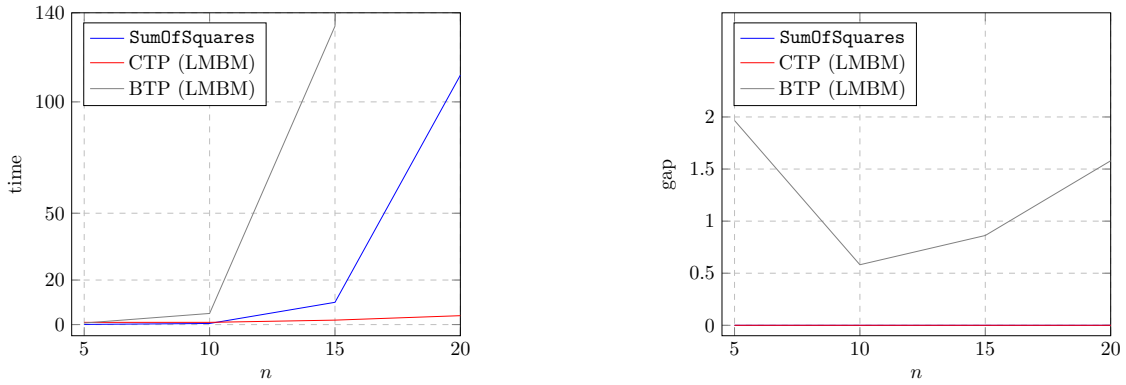


Figure 4.4: Efficiency and accuracy comparison for Table 4.7.

$g_i(\mathbf{a}) > 0$;

5. For $j \in [l]$, all non-constant coefficients of h_j are taken randomly in $(-1, 1)$ with respect to the uniform distribution, and the constant coefficient of h_j is chosen such that $h_j(\mathbf{a}) = 0$.

Numerical results are displayed in Table 4.7 for the case $(m, l) = (1, \lceil n/4 \rceil)$ and Table 4.8 for the case $(m, l) = (\lceil n/8 \rceil, \lceil n/8 \rceil)$. We recall the following notation:

- CTP (LMBM): the SDP relaxation is solved via the method described in Section 4.2.2 (the k-th iteration of Algorithm 4) or Section 4.2.3 (the k-th iteration of Algorithm 5) with the LMBM solver.
- BTP (LMBM): the SDP relaxation is solved via the method described in Remark 4.5 with the LMBM solver (Algorithm 2).

In Table 4.7 and Table 4.8, `SumOfSquares` and BTP solve relaxations involving matrices with the same size, corresponding exactly to the size of the moment relaxation (2.4.5).

Efficiency and accuracy comparisons: In Table 4.7, we consider POPs which involve a single inequality (ball) constraint. In this case, CTP (LMBM) is the most efficient and accurate solver. Numerical results emphasize that `SumOfSquares` and CTP (LMBM) behave in a similar way as in Table 4.4. This indicates that converting a POP with a single inequality (ball) constraint to a CTP-POP by adding one slack variable, and solving the resulting SDP-CTP relaxation by means of spectral methods allows one to reduce the computing time while ensuring the same accuracy as

Table 4.8: Numerical results of random dense QCQPs on the unit ball, described in Section 4.3.2, with $(m, l) = (\lceil n/8 \rceil, \lceil n/8 \rceil)$, and $k = 2$.

POP size			SDP size (CTP)		`SumOfSquares` (`Mosek`)		SpectralPOP			
							CTP (LMBM)		BTP (LMBM)	
n	m	l	s	m	val	time	val	time	val	time
10	2	2	105	3711	-3.17792	0.2	-3.18434	31	-4.71914	6
15	2	2	190	11781	-2.14424	6	-2.16250	69	-4.45435	649
20	3	3	190	11781	-2.92513	190	-3.09124	469	-81.3719	2573

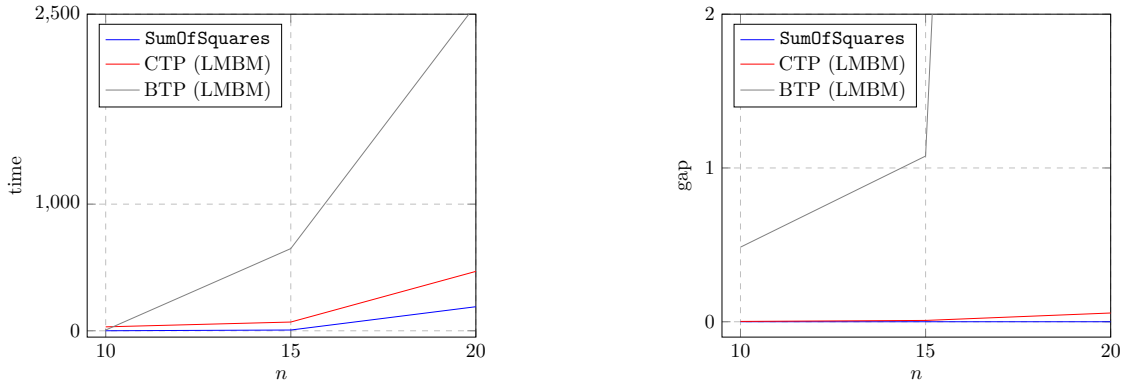


Figure 4.5: Efficiency and accuracy comparison for Table 4.8.

the one obtained with `SumOfSquares` (`Mosek`). Note that when we use the method described in Section 4.2.2, the constant trace in (4.2.12) is always equal to 2^k, which is independent of n.

In Table 4.8, CTP (LMBM) provides inaccurate output as it only yields lower bounds, while `SumOfSquares` still preserves accuracy. Moreover, CTP (LMBM) is less efficient than `SumOfSquares`. We also emphasize that when one relies on the method stated in Section 4.2.3, we obtain a value of $\bar{R}$, in (4.2.14), for the sphere constraint of CTP-POP, which becomes larger when n increases. It implies that the constant trace factor $(\bar{R}+1)^k$ in (4.2.18) has a polynomial growth rate in $\bar{R}$. Thus we minimize a nonsmooth function of the form (4.1.13) with a large constant trace factor a. The norm of the subgradient of this function at a point near its minimizers is rather large, which prevents LMBM to perform properly its minimization, by contrast with Table 4.7. This difference of magnitude is shown in Table 4.9, where we compute the subgradient norms during the last 10 iterations of CTP (LMBM) for the experiments from Table 4.7 and Table 4.8 with $n = 10$.

In both Table 4.7 and Table 4.8, BTP (LMBM) has the worst performance in terms on efficiency and accuracy. The trace bound (4.2.2) obtained in Remark 4.5 is usually much larger than the "exact" trace of the optimal solution of the SDP relaxation. The same issue occurs for the subgradient norm of the nonsmooth function at a point near its minimizers.

According to our experience, LMBM is suitable for spectral minimization of SDP problems with trace bounds which are small enough and close to the exact trace value of the optimal solution.

Table 4.9: Subgradient norms computed during the last 10 iterations of CTP (LMBM) for the experiments from Table 4.7 and Table 4.8 with $n = 10$.

Table 4.7	0.035	0.033	0.028	0.022	0.024	0.019	0.015	0.009	0.007	0.004
Table 4.8	0.871	0.959	0.947	0.792	1.684	0.794	0.579	0.559	0.230	0.916

Table 4.10: Numerical results for random dense quartics on the unit sphere, described in Section 4.3.3, with $(m, l) = (0, 1)$ and $k = 2$.

POP size	SDP size		`SumOfSquares` (`Mosek`)		SpectralPOP (CTP)			
					LMBM		`SketchyCGAL`	
n	s	m	val	time	val	time	val	time
5	21	127	-2.92483*	0.02	-2.92485*	1	-2.84710	1
10	66	1277	-3.59964*	0.4	-3.59964*	1	-3.48501	2
15	136	5577	-4.18773*	7	-4.18778*	12	-4.03882	18
20	231	16402	-3.92438*	88	-3.92440*	35	-3.67857	87
25	351	38377	-6.36891*	711	-6.36894*	74	-5.93774	251

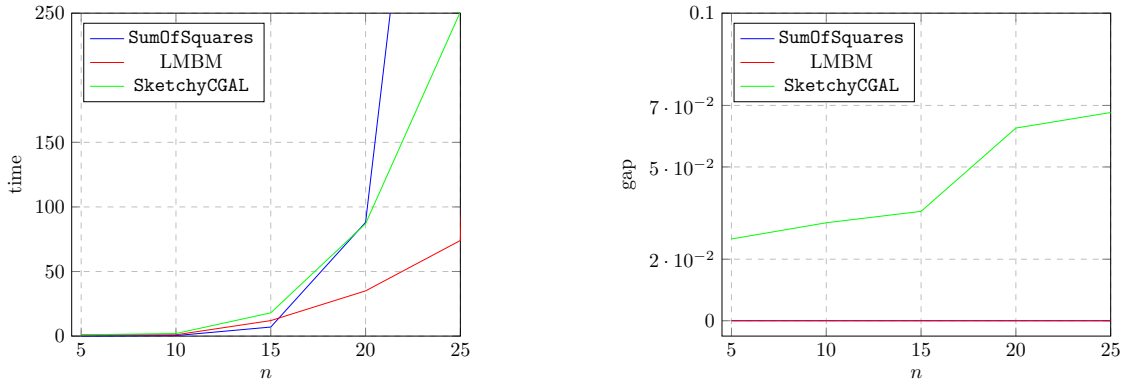


Figure 4.6: Efficiency and accuracy comparison for Table 4.10.

4.3.3 Random dense quartics on the unit sphere

Test problems: We construct several instances of POP (4.1.3) as follows:

1. Take $l = 1$ and $h_1 = 1 - \|\mathbf{x}\|_2^2$ and choose f with degree at most 4;
2. Each coefficient of the objective function f is taken randomly in $(-1, 1)$ with respect to the uniform distribution.

We use the method presented in Section 4.2.1 to solve these problems. The corresponding numerical results are displayed in Table 4.10.

Efficiency and accuracy comparisons: Table 4.10 indicates that LMBM is about twice faster than `SumOfSquares` when $n \geq 20$ as well as `SketchyCGAL`.While `SketchyCGAL` can be rather inaccurate, LMBM has an accuracy which is similar to `SumOfSquares` (`Mosek`), yielding the ability to extract optimal solutions of POPs.

For comparisons in Section 4.3.2 and Section 4.3.3, the coefficients of f have been randomly generated in (-1,1). However, for some non random problems that were *scaled* so as to fit the framework of optimization on the unit sphere, we could observe a lack of precision after transferring results (of the scaled formulation) back to results in the unscaled initial formulation.

In the next three subsections, we consider further applications of the minimization of forms on the unit sphere listed in [107].

4.3.4 Deciding the nonnegativity of even degree forms

Given $q \in \mathbb{R}[\mathbf{x}]$, we recall that q is a form of degree d if $q = \sum_{|\boldsymbol{\alpha}|=d} q_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}}$ for some $d \in \mathbb{N}$ and $q_{\boldsymbol{\alpha}} \in \mathbb{R}$. Given a form $q \in \mathbb{R}[\mathbf{x}]$, q is nonnegative on $\mathbb{R}^n$ iff q is nonnegative on the unit sphere. Moreover, given a polynomial $f \in \mathbb{R}[\mathbf{x}]_{2d}$, f is nonnegative on $\mathbb{R}^n$ iff its homogenization $\bar{f} := x_{n+1}^{2d} f(\frac{\mathbf{x}}{x_{n+1}})$ is nonnegative on $\mathbb{R}^{n+1}$. Note that $\bar{f}$ is a form. Thus, in order to verify the

Table 4.11: Numerical results for deciding the nonnegativity of even degree forms, described in Section 4.3.4, with $k = d$.

POP size		SDP size		`SumOfSquares` (`Mosek`)		SpectralPOP (CTP) (LMBM)	
n	d	s	m	val	time	val	time
5	2	21	127	-2.62353*	0.01	-2.62353*	1
10	2	66	1277	-6.31389*	0.4	-6.31390*	1
15	2	126	5577	-12.1405*	6	-12.1405*	2
20	2	231	16402	-19.9981*	76	-19.9981*	2
25	2	351	38377	-29.4812*	633	-29.4812*	4
30	2	496	77377	-40.6934*	3471	-40.6934*	9
35	2	666	140527	–	–	-54.2561*	24
40	2	861	236202	–	–	-69.4700*	57
45	2	1081	374027	–	–	-86.4113*	127
50	2	1326	564877	–	–	-105.532*	250
5	3	56	1261	-1.56744	0.1	-1.60032	23
5	4	126	7177	-1.35267	1	-1.35315	235

nonnegativity of the polynomial f, we only verify the nonnegativity of its homogenization $\bar{f}$ on the unit sphere in $\mathbb{R}^{n+1}$. Namely, given a form $f \in \mathbb{R}[\mathbf{x}]$ of degree $2d$, we consider the following POP:

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^n} \{f(\mathbf{x}) \,:\, \|\mathbf{x}\|_2^2 = 1\}\,. \tag{4.3.1}$$

Note that if $d = 1$, problem (4.3.1) boils down to computing the minimal eigenvalue of the Gram matrix associated to f. Thus, we consider the case where $d \geq 2$.

Test problems: We construct several instances of the form f of degree $2d$ as follows:

1. Take $f_{\boldsymbol{\alpha}}$ randomly in $(-1, 1)$ with respect to the uniform distribution, for each $\boldsymbol{\alpha} \in \mathbb{N}^n$ with $|\boldsymbol{\alpha}| = 2d$.
2. Set $f := \sum_{|\boldsymbol{\alpha}|=2d} f_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}}$.

We use the method presented in Section 4.2.1 to solve problem (4.3.1). The corresponding numerical results are displayed in Table 4.11.

Efficiency and accuracy comparisons: Table 4.11 shows that LMBM is much faster than `Mosek` when $n \geq 20$ and $d = 2$. In these cases, we were able to extract the solution of the resulting POPs. One can then conclude that f is not globally nonnegative since it has negative value at its atoms. For higher values of $d = 3, 4$, LMBM becomes less efficient and accurate than `Mosek`.

4.3.5 Deciding the convexity of even degree forms

Recall that a polynomial $q \in \mathbb{R}[\mathbf{x}]$ is convex if $q(t\mathbf{x}+(1-t)\mathbf{y}) \leq tq(\mathbf{x})+(1-t)q(\mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $t \in [0, 1]$. Moreover, $q \in \mathbb{R}[\mathbf{x}]$ is convex iff the polynomial $f(\mathbf{x}, \mathbf{y}) := q(\mathbf{y})-q(\mathbf{x})-\nabla q(\mathbf{x})^\top(\mathbf{y}-\mathbf{x})$ is globally nonnegative, where ∇q stands for the gradient of q. If q is a form of degree d, f is also a form of degree d. In this case, the nonnegativity of f can be verified on the unit sphere in $\mathbb{R}^{2n}$.

Given a form $q \in \mathbb{R}[\mathbf{x}]$ of degree $2d$, consider the following POP:

$$\begin{aligned} f^\star := \min_{\mathbf{x},\mathbf{y}\in\mathbb{R}^n} \quad & q(\mathbf{y}) - q(\mathbf{x}) - \nabla q(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) \\ \text{s.t.} \quad & \|(\mathbf{x}, \mathbf{y})\|_2^2 = 1\,. \end{aligned} \tag{4.3.2}$$

From the previous discussion, q is convex iff $f^\star \geq 0$.

Table 4.12: Numerical results for deciding the convexity of even degree forms, described in Section 4.3.5, with $k = d$.

POP size		SDP size		`SumOfSquares` (`Mosek`)		SpectralPOP (CTP) (LMBM)	
n	d	s	m	val	time	val	time
5	2	66	1277	-3.87350*	0.2	-3.87350*	1
7	2	120	4321	-4.44260*	3	-4.44280*	11
10	2	231	16402	-4.98855*	63	-4.98883*	17
12	2	325	32826	-4.47495*	336	-4.49239*	91
5	3	286	34035	-3.89581*	53	-3.89586*	70

Test problems: We construct several instances of the form q of degree $2d$ as follows:

1. Take $q_{\boldsymbol{\alpha}}$ randomly in $(-1, 1)$ with respect to the uniform distribution, for each $\boldsymbol{\alpha} \in \mathbb{N}^n$ with $|\boldsymbol{\alpha}| = 2d$.
2. Set $q := \sum_{|\boldsymbol{\alpha}|=2d} q_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}}$.

We use the method presented in Section 4.2.1 to solve problem (4.3.2). The corresponding numerical results are displayed in Table 4.12.

Efficiency and accuracy comparisons: Table 4.12 shows that LMBM is about 3 times faster than `Mosek` when $n \geq 10$ and $d = 2$. In these cases, one concludes as above that q is nonconvex since $f^\star$ is negative. For $d = 3$ and $n = 5$, LMBM still returns a value with reasonably high accuracy (10^{-5}) even though it is slower than `Mosek`. In this case, one can also certify that q is nonconvex.

4.3.6 Deciding the copositivity of real symmetric matrices

Given a symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, we say that $\mathbf{A}$ is copositive if $\mathbf{u}^\top \mathbf{A}\mathbf{u} \geq 0$ for all $\mathbf{u} \in \mathbb{R}^n_+$. Consider the following POP:

$$f^\star := \min_{\mathbf{x} \in \mathbb{R}^n} \{(\mathbf{x}^2)^\top \mathbf{A}\mathbf{x}^2 \ : \ \|\mathbf{x}\|_2^2 = 1\}, \tag{4.3.3}$$

where $\mathbf{x}^2 := (x_1^2, \ldots, x_n^2)$. The matrix $\mathbf{A}$ is copositive iff $f^\star \geq 0$.

Test problems: We construct several instances of the matrix $\mathbf{A}$ as follows:

1. Take B_{ij} randomly in $(-1, 1)$ with respect to the uniform distribution, for all $i, j \in [n]$.
2. Set $\mathbf{B} := (B_{ij})_{1 \leq i,j \leq n}$ and $\mathbf{A} := \frac{1}{2}(\mathbf{B} + \mathbf{B}^\top)$.

We use the method presented in Section 4.2.1 to solve problem (4.3.3). The corresponding numerical results are displayed in Table 4.13.

Efficiency and accuracy comparisons: Table 4.13 indicates that LMBM is twice faster than `Mosek` when $n \geq 20$. In all cases, we can extract the solutions of the resulting POP and certify that $\mathbf{A}$ is not copositive since $f^\star$ is negative.

Table 4.13: Numerical results for deciding the copositivity of real symmetric matrices, described in Section 4.3.6, with $k = 2$.

POP size	SDP size		`SumOfSquares` (`Mosek`)		SpectralPOP (CTP) (LMBM)	
n	s	m	val	time	val	time
10	66	1277	-0.89102*	0.2	-0.89102*	1
15	136	5577	-0.91701*	6	-0.91701*	9
20	231	16402	-0.98474*	57	-0.98474*	19
25	351	38377	-0.97873*	558	-0.97873*	28

Chapter 5

Exploiting the constant trace property: Inequality constraints

Most of the content of this chapter is from [131].

The goal of this chapter is to extend the CTP-exploiting framework introduced in the previous chapter to the case of POPs with inequality constraints. Based on this, we provide a method which returns the optimal value of the second order moment SDP-relaxation and which is suitable for a class of large-scale non-convex QCQPs with CTP. Ideally (i) it should avoid memory issues, and (ii) the resulting relative gap of the approximate value returned by this method w.r.t. the exact value should be less than 1%.

Contribution. We show that (i) a large class of POPs with inequality constraints have the *constant trace property* and (ii) this property can be exploited for solving their associated semidefinite relaxations via appropriate first-order methods. More precisely our contribution is threefold:

I. In Section 5.1.2 we show that if a positive real number belongs to the interior of every truncated quadratic module associated with the inequality constraints, which is defined later in (2.3.4), then the corresponding POP has CTP. Moreover, we prove that this condition always holds when a ball constraint is present.

II. In Section 5.1.3 we provide a numerical procedure to check whether a POP has CTP. It consists in solving a certain linear program (LP) of the form:

$$\inf_{\mathbf{y}\in\mathbb{R}^a_+} \{\mathbf{c}^\top\mathbf{y} \,:\, \mathbf{A}\mathbf{y} = \mathbf{b}\}, \tag{5.0.1}$$

where $\mathbf{c} \in \mathbb{R}^a$, $\mathbf{A} \in \mathbb{R}^{b\times a}$ and $\mathbf{b} \in \mathbb{R}^b$. With this approach we prove in Section 5.1.4 that several special classes of POPs (including POPs on a ball, annulus, simplex) have CTP.

III. Our final contribution, postponed in Appendices 5.3.1 and 5.3.2, is to handle sparse large-scale POPs by integrating sparsity-exploiting techniques from Chapter 3 into the CTP-exploiting framework.

For practical implementation we provide a software library called ctpPOP. It models each moment SDP-relaxation of POPs as a standard SDP with CTP and then solves this SDP by CGAL or a spectral method (SM), based on nonsmooth optimization solvers (LMBM [95] or the Proximal Bundle Method [98]).

In Section 5.2 and Appendix 5.3.3 we provide extensive numerical experiments to illustrate the efficiency and scalability of ctpPOP with the CGAL solver. In all our randomly generated POPs with different sparsity structures, the relative gap of the optimal value provided by CGAL w.r.t. the optimal value provided by `Mosek` is below 1%. Because of its cheap cost per iteration, CGAL is more suitable for particularly SDPs of form (1.4.1) with $\zeta = \mathcal{O}((s^{\max})^2)$ (such as the second order moment SDP-relaxations of POPs) than other solvers (e.g., `COSMO`).

For instance for minimizing a *dense* quadratic polynomial on the unit ball with 100 variables, CGAL returns the optimal value of the second order moment SDP relaxation within 6 hours on a standard laptop while `Mosek` (considered as the state-of-the-art SDP solver using interior-point

methods) runs out of memory. Similarly, for minimizing a *sparse* quadratic polynomial involving a thousand variables with a ball constraint on each clique of variables, CGAL spends around two thousand seconds to solve the second order moment SDP-relaxation while `Mosek` runs again out of memory. The largest clique of this POP involves 42 variables.

The classical OPF problem without constraints on current magnitudes (as in [90, 63]) can be formulated as a POP with ball and annulus constraints. For many instances Shor's relaxation provides the global optimum. However, for illustration purposes we have compared CGAL and `Mosek` for solving the second order CS-TSSOS relaxation from Chapter 3 of the "case89_pegase_api" instance from the PGLib-OPF database[1]. The largest block size and the number of equality constraints of this SDP are around 1.7 thousands and 8 millions, respectively. While `Mosek` failed because of a memory issue, CGAL still returned the optimal value in two days, and with relative gap w.r.t. a local optimal value being less than 0.6%.

As in the previous chapter, a POP is defined as

$$f^\star := \inf\{f(\mathbf{x}) \ : \ \mathbf{x} \in S(\mathfrak{g}) \cap V(\mathfrak{h})\}\,, \tag{5.0.2}$$

where $S(\mathfrak{g})$ and $V(\mathfrak{h})$ are a basic semialgebraic set and a real variety defined respectively by

$$\begin{array}{rcl} S(\mathfrak{g}) & := & \{\,\mathbf{x} \in \mathbb{R}^n : \, g_i(\mathbf{x}) \geq 0\,,\, i \in [m]\,\} \\ V(\mathfrak{h}) & := & \{\,\mathbf{x} \in \mathbb{R}^n : \, h_j(\mathbf{x}) = 0\,,\, i \in [l]\,\}\,, \end{array} \tag{5.0.3}$$

for some polynomials $f, g_i, h_j \in \mathbb{R}[\mathbf{x}]$ with $\mathfrak{g} := \{g_i\}_{i\in[m]}$, $\mathfrak{h} := \{h_j\}_{j\in[l]}$. Also recall that $\lceil g_i \rceil = \lceil \deg(g_i)/2 \rceil$, $\lceil h_j \rceil = \lceil \deg(h_j)/2 \rceil$ and $k_{\min} = \max_{i,j}\{\lceil f \rceil, \lceil g_i \rceil, \lceil h_j \rceil\}$. We will assume in this chapter that POP (5.0.2) has at least one global minimizer.

5.1 Exploiting CTP for dense POPs

This section is devoted to developing a framework to exploit CTP for dense POPs. We provide a sufficient condition for a POP to have CTP, as well as a series of linear programs to check whether the sufficient condition holds. In addition we show that several special classes of POPs have CTP.

5.1.1 CTP for dense POPs

First let us recall CTP for a POP. To simplify notation, for every $k \in \mathbb{N}^{\geq k_{\min}}$, denote by $\mathcal{S}^{(k)}$ the set of real symmetric matrices
- of size $s_k := b(k) + \sum_{i\in[m]} b(k - \lceil g_i \rceil)$,
- in a block diagonal form $\mathbf{X} = \mathrm{diag}(\mathbf{X}_0, \ldots, \mathbf{X}_m)$, and such that
- $\mathbf{X}_0$ (resp. $\mathbf{X}_i$) is of size $b(k)$ (resp. $b(k - \lceil g_i \rceil)$ for $i \in [m]$).

Letting $\mathbf{D}_k(\mathbf{y}) := \mathrm{diag}(\mathbf{M}_k(\mathbf{y}), \mathbf{M}_{k-\lceil g_1 \rceil}(g_1\mathbf{y}), \ldots, \mathbf{M}_{k-\lceil g_m \rceil}(g_m\mathbf{y}))$, SDP (2.4.5) can be rewritten in the form:

$$\tau_k \ := \ \inf_{\mathbf{y}\in\mathbb{R}^{b(2k)}} \left\{ L_{\mathbf{y}}(f) \ \middle| \ \begin{array}{l} \mathbf{D}_k(\mathbf{y}) \in \mathcal{S}_+^{(k)}\,,\, y_{\mathbf{0}} = 1\,, \\ \mathbf{M}_{k-\lceil h_j \rceil}(h_j\mathbf{y}) = 0\,,\, j \in [l] \end{array} \right\}, \tag{5.1.1}$$

where $\mathcal{S}_+^{(k)}$ is the set of positive semidefinite matrices in $\mathcal{S}^{(k)}$.

Definition 5.1. *(CTP for a POP) We say that POP* (5.0.2) *has CTP if for every $k \in \mathbb{N}^{\geq k_{\min}}$, there exists $a_k > 0$ and a positive definite matrix $\mathbf{P}_k \in \mathcal{S}^{(k)}$ such that for all $\mathbf{y} \in \mathbb{R}^{b(2k)}$,*

$$\left.\begin{array}{l} \mathbf{M}_{k-\lceil h_j \rceil}(h_j\mathbf{y}) = 0\,,\, j \in [l]\,, \\ y_{\mathbf{0}} = 1 \end{array}\right\} \Rightarrow \mathrm{trace}(\mathbf{P}_k\mathbf{D}_k(\mathbf{y})\mathbf{P}_k) = a_k\,. \tag{5.1.2}$$

In other words, we say that POP (5.0.2) has CTP if each moment relaxation (5.1.1) has an equivalent form involving a psd matrix whose trace is constant. In this case, we call a_k the constant trace and $\mathbf{P}_k$ the basis transformation matrix. In the next subsection, we provide a sufficient condition for POP (5.0.2) to have CTP.

[1] https://github.com/power-grid-lib/pglib-opf

Example 5.1. *We recall CTP for equality constrained POPs on a sphere in Chapter 4. If* $g = \emptyset$ *and* $h_1 = R - \|\mathbf{x}\|_2^2$ *for some* $R > 0$*, then POP* (5.0.2) *has CTP with* $a_k = (R+1)^k$ *and* $\mathbf{P}_k := \mathrm{diag}((\theta_{k,\boldsymbol{\alpha}}^{1/2})_{\boldsymbol{\alpha}\in\mathbb{N}^n_k})$*, where* $(\theta_{k,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_k} \subseteq \mathbb{R}^{>0}$ *satisfies* $(1+\|\mathbf{x}\|_2^2)^k = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^n_k} \theta_{k,\boldsymbol{\alpha}}\mathbf{x}^{2\boldsymbol{\alpha}}$*, for all* $k \in \mathbb{N}^{\geq k_{\min}}$.

We now provide a general method to solve a POP with CTP. We first convert the k-th order moment relaxation (5.1.1) of this POP to a standard primal SDP problem with CTP and then leverage appropriate first-order algorithms that exploit CTP to solve the resulting SDP problem.

Suppose POP (5.0.2) has CTP. For every $k \in \mathbb{N}^{\geq k_{\min}}$, letting $\mathbf{X} = \mathbf{P}_k\mathbf{D}_k(\mathbf{y})\mathbf{P}_k$, (5.1.1) can be rewritten as

$$\tau_k = \inf_{\mathbf{X}\in\mathcal{S}_+^{(k)}} \{\langle \mathbf{C}_k, \mathbf{X}\rangle \,:\, \mathcal{A}_k\mathbf{X} = \mathbf{b}_k\}\,, \tag{5.1.3}$$

where $\mathcal{A}_k : \mathcal{S}^{(k)} \to \mathbb{R}^{\zeta_k}$ is a linear operator such that $\mathcal{A}_k\mathbf{X} = (\langle \mathbf{A}_{k,1}, \mathbf{X}\rangle, \ldots, \langle \mathbf{A}_{k,\zeta_k}, \mathbf{X}\rangle)$ with $\mathbf{A}_{k,i} \in \mathcal{S}^{(k)}$, $i \in [\zeta_k]$, $\mathbf{C}_k \in \mathcal{S}^{(k)}$ and $\mathbf{b}_k \in \mathbb{R}^{\zeta_k}$. Appendix 5.3.6 describes how to convert SDP (5.1.1) to the form (5.1.3).

The dual of SDP (5.1.3) reads

$$\rho_k = \sup_{\mathbf{z}\in\mathbb{R}^{\zeta_k}} \{\, \mathbf{b}_k^\top\mathbf{z} \,:\, \mathcal{A}_k^\top\mathbf{z} - \mathbf{C}_k \in \mathcal{S}_+^{(k)}\}\,, \tag{5.1.4}$$

where $\mathcal{A}_k^\top : \mathbb{R}^{\zeta_k} \to \mathcal{S}^{(k)}$ is the adjoint operator of $\mathcal{A}_k$, i.e., $\mathcal{A}_k^\top\mathbf{z} = \sum_{i\in[\zeta_k]} z_i\mathbf{A}_{k,i}$.

After replacing $(\mathcal{A}_k, \mathbf{A}_{k,i}, \mathbf{b}_k, \mathbf{C}_k, \mathcal{S}^{(k)}, \zeta_k, s_k, \tau_k, \rho_k, a_k)$ by $(\mathcal{A}, \mathbf{A}_i, \mathbf{b}, \mathbf{C}, \mathcal{S}, \zeta, s, \tau, \rho, a)$, the primal-dual (5.1.3)-(5.1.4) has an equivalent formulation as the primal-dual (5.3.25)-(5.3.26); see also Appendix 5.3.4 with $\omega = m+1$ and $s^{\max} = s(k)$.

Then two first-order algorithms (CGAL and SM) are leveraged for solving the primal-dual (5.3.25)-(5.3.26); see Appendix 5.3.4 and Appendix 5.3.5.

5.1.2 A sufficient condition for a POP to have CTP

In this section, we provide a sufficient condition for POP (5.0.2) to have CTP.

For $k \in \mathbb{N}^{\geq k_{\min}}$, let $\mathcal{Q}_k^\circ(\mathfrak{g})$ be the interior of the truncated quadratic module $\mathcal{Q}_k(\mathfrak{g})$, i.e., $\mathcal{Q}_k^\circ(\mathfrak{g}) := \{\mathbf{v}_k^\top\mathbf{G}_0\mathbf{v}_k + \sum_{i\in[m]} g_i\mathbf{v}_{k-\lceil g_i\rceil}^\top\mathbf{G}_i\mathbf{v}_{k-\lceil g_i\rceil} \,:\, \mathbf{G}_i \succ 0, \quad i \in \{0\}\cup[m]\}$.

Theorem 5.1. *The following statements hold:*

1. *If one the following equivalent conditions hold for all* $k \in \mathbb{N}^{\geq k_{\min}}$*:*

$$\begin{aligned} \mathbb{R}^{>0} \subseteq \mathcal{Q}_k^\circ(\mathfrak{g}) + \mathcal{I}_k(\mathfrak{h}) &\Leftrightarrow \forall\delta > 0\,,\ \delta \in \mathcal{Q}_k^\circ(\mathfrak{g}) + \mathcal{I}_k(\mathfrak{h}) \\ &\Leftrightarrow 1 \in \mathcal{Q}_k^\circ(\mathfrak{g}) + \mathcal{I}_k(\mathfrak{h})\,, \end{aligned} \tag{5.1.5}$$

   then POP (5.0.2) *has CTP, as in Definition 5.1.*

2. *Assume that* $\mathfrak{h} = \emptyset$ *and* $S(\mathfrak{g})$ *has nonempty interior. Then POP* (5.0.2) *has CTP if and only if*

$$\mathbb{R}^{>0} \subseteq \mathcal{Q}_k^\circ(\mathfrak{g})\,,\ \forall k \in \mathbb{N}^{\geq k_{\min}}\,. \tag{5.1.6}$$

Proof. 1. Let $k \in \mathbb{N}^{\geq k_{\min}}$ and assume that $\mathbb{R}^{>0} \subseteq \mathcal{Q}_k^\circ(\mathfrak{g}) + \mathcal{I}_k(\mathfrak{h})$. Then there exists $a_k > 0$ such that

$$a_k = \mathbf{v}_k^\top\mathbf{G}_0\mathbf{v}_k + \sum_{i\in[m]} g_i\mathbf{v}_{k-\lceil g_i\rceil}^\top\mathbf{G}_i\mathbf{v}_{k-\lceil g_i\rceil} + \sum_{j\in[l]} h_j\mathbf{v}_{2(k-\lceil h_j\rceil)}^\top\mathbf{u}_j\,, \tag{5.1.7}$$

for some $\mathbf{G}_i \succ 0$, $i \in \{0\}\cup[m]$ and real vector $\mathbf{u}_j$, $j \in [l]$. We denote by $\mathbf{G}_i^{1/2}$ the square root of $\mathbf{G}_i$, $i \in \{0\}\cup[m]$. Then $\mathbf{G}_i^{1/2}$ is well-defined and $\mathbf{G}_i^{1/2} \succ 0$. Set $\mathbf{P}_k = \mathrm{diag}(\mathbf{G}_0^{1/2}, \ldots, \mathbf{G}_m^{1/2})$. Let $\mathbf{y} \in \mathbb{R}^{b(2k)}$ such that $\mathbf{M}_k(h_j\mathbf{y}) = 0$, $j \in [l]$, and $y_{\mathbf{0}} = 1$. Then

$$L_{\mathbf{y}}\left(\sum_{j\in[l]} h_j\mathbf{v}_{2(k-\lceil h_j\rceil)}^\top\mathbf{u}_j\right) = \sum_{j\in[l]}\sum_{\boldsymbol{\alpha}\in\mathbb{N}^n_{2(k-\lceil h_j\rceil)}} u_{j,\boldsymbol{\alpha}}L_{\mathbf{y}}(h_j\mathbf{x}^{\boldsymbol{\alpha}}) = 0\,. \tag{5.1.8}$$

From this and (5.1.7),

$$\begin{aligned} a_k &= L_{\mathbf{y}}(\mathbf{v}_k^\top \mathbf{G}_0 \mathbf{v}_k + \textstyle\sum_{i\in[m]} g_i \mathbf{v}_{k-\lceil g_i \rceil}^\top \mathbf{G}_i \mathbf{v}_{k-\lceil g_i \rceil}) \\ &= \operatorname{trace}(\mathbf{M}_k(\mathbf{y})\mathbf{G}_0) + \textstyle\sum_{i\in[m]} \operatorname{trace}(\mathbf{M}_{k-1}(g_i\mathbf{y})\mathbf{G}_i) \\ &= \operatorname{trace}(\mathbf{G}_0^{1/2}\mathbf{M}_k(\mathbf{y})\mathbf{G}_0^{1/2}) + \textstyle\sum_{i\in[m]} \operatorname{trace}(\mathbf{G}_i^{1/2}\mathbf{M}_{k-1}(g_i\mathbf{y})\mathbf{G}_i^{1/2}) \\ &= \operatorname{trace}(\mathbf{P}_k \mathbf{D}_k(\mathbf{y}) \mathbf{P}_k)\,, \end{aligned}$$

yielding the first statement.

2. The "if" part comes from the first statement. Let us prove the "only if" part. Assume that POP (5.0.2) has CTP (Definition 5.1). Let $\mathbf{a} \in S(\mathfrak{g})$, $\mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n}$ be the moment sequence of the Dirac measure $\delta_{\mathbf{a}}$. Let $k \in \mathbb{N}^{\geq k_{\min}}$ be fixed. Since $\mathbf{P}_k \in \mathcal{S}^{(k)}$, $\mathbf{P}_k = \operatorname{diag}(\mathbf{W}_0, \ldots, \mathbf{W}_m)$. Then $\mathbf{W}_i^2 \succ 0$, $i \in \{0\} \cup [m]$ since $\mathbf{P}_k \succ 0$. Let us define the polynomial $w := \mathbf{v}_k^\top \mathbf{W}_0^2 \mathbf{v}_k + \sum_{i\in[m]} g_i \mathbf{v}_{k-\lceil g_i \rceil}^\top \mathbf{W}_i^2 \mathbf{v}_{k-\lceil g_i \rceil}$. By assumption,

$$\begin{aligned} a_k &= \operatorname{trace}(\mathbf{P}_k \mathbf{D}_k(\mathbf{y}) \mathbf{P}_k) \\ &= \operatorname{trace}(\mathbf{W}_0 \mathbf{M}_k(\mathbf{y}) \mathbf{W}_0) + \textstyle\sum_{i\in[m]} \operatorname{trace}(\mathbf{W}_i \mathbf{M}_{k-1}(g_i\mathbf{y}) \mathbf{W}_i) \\ &= \operatorname{trace}(\mathbf{M}_k(\mathbf{y}) \mathbf{W}_0^2) + \textstyle\sum_{i\in[m]} \operatorname{trace}(\mathbf{M}_{k-1}(g_i\mathbf{y}) \mathbf{W}_i^2) \\ &= L_{\mathbf{y}}(\mathbf{v}_k^\top \mathbf{W}_0^2 \mathbf{v}_k + \textstyle\sum_{i\in[m]} g_i \mathbf{v}_{k-\lceil g_i \rceil}^\top \mathbf{W}_i^2 \mathbf{v}_{k-\lceil g_i \rceil}) = \textstyle\int_{\mathbb{R}^n} w d\delta_{\mathbf{a}}(x) = w(\mathbf{a})\,. \end{aligned}$$

It implies that $w - a_k$ vanishes on $S(\mathfrak{g})$. Since $S(\mathfrak{g})$ has nonempty interior, $w = a_k$, yielding the second statement.

□

The following lemma will be used later on.

Lemma 5.1. *Let $R > 0$. For all $k \in \mathbb{N}^{\geq 1}$, one has*

$$(R+1)^k = (1 + \|\mathbf{x}\|_2^2)^k + (R - \|\mathbf{x}\|_2^2) \sum_{j=0}^{k-1} (R+1)^j (1 + \|\mathbf{x}\|_2^2)^{k-j-1}\,. \tag{5.1.9}$$

Proof. Let $k \in \mathbb{N}^{\geq 1}$. Letting $a = R + 1$ and $b = 1 + \|\mathbf{x}\|_2^2$, the desired equality follows from $a^k - b^k = (a-b) \sum_{j=0}^{k-1} a^j b^{k-1-j}$. □

The next result states that the sufficient condition in Theorem 5.1 holds whenever a ball constraint is present in the POP's description. For a real symmetric matrix $\mathbf{A}$, denote the largest eigenvalue of $\mathbf{A}$ by $\lambda_{\max}(\mathbf{A})$.

Theorem 5.2. *If $R - \|\mathbf{x}\|_2^2 \in \mathfrak{g}$ for some $R > 0$ then the inclusions* (5.1.6) *hold and therefore POP* (5.0.2) *has CTP.*

Proof. Without loss of generality, set $g_m := R - \|\mathbf{x}\|_2^2$ and let $k \in \mathbb{N}^{\geq k_{\min}}$ be fixed. By Lemma 5.1, $(R+1)^k = \Theta + g_m \Lambda$, where $\Theta := (1 + \|\mathbf{x}\|_2^2)^k$ and $\Lambda := \sum_{j=0}^{k-1} (R+1)^j (1 + \|\mathbf{x}\|_2^2)^{k-j-1}$. Note that

- $\Theta = \sum_{\boldsymbol{\alpha}\in\mathbb{N}_k^n} \theta_{\boldsymbol{\alpha}} \mathbf{x}^{2\boldsymbol{\alpha}} = \mathbf{v}_k^\top \mathbf{G}_0 \mathbf{v}_k$ for some $(\theta_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}_k^n} \subseteq \mathbb{R}^{>0}$;
- $\Lambda = \sum_{\boldsymbol{\alpha}\in\mathbb{N}_{k-1}^n} \lambda_{\boldsymbol{\alpha}} \mathbf{x}^{2\boldsymbol{\alpha}} = \mathbf{v}_{k-1}^\top \mathbf{G}_m \mathbf{v}_{k-1}$ for some $(\lambda_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}_{k-1}^n} \subseteq \mathbb{R}^{>0}$.

Here $\mathbf{G}_0 = \operatorname{diag}((\theta_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}_k^n})$ and $\mathbf{G}_m = \operatorname{diag}((\lambda_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}_{k-1}^n})$ are both positive definite. Then we have $(R+1)^k = \mathbf{v}_k^\top \mathbf{G}_0 \mathbf{v}_k + g_m \mathbf{v}_{k-1}^\top \mathbf{G}_m \mathbf{v}_{k-1}$. Denote by $\mathbf{I}_t$ the identity matrix of size $b(t)$ for $t \in \mathbb{N}$.

Let $\mathbf{W}$ be a real symmetric matrix such that $\sum_{i\in[m-1]} g_i \mathbf{v}_{k-\lceil g_i \rceil}^\top \mathbf{I}_{k-\lceil g_i \rceil} \mathbf{v}_{k-\lceil g_i \rceil} = \mathbf{v}_k^\top \mathbf{W} \mathbf{v}_k$. Since $\mathbf{G}_0 \succ 0$, there exists $\delta > 0$ such that $\mathbf{G}_0 - \delta \mathbf{W} \succ 0$. Indeed,

$$\mathbf{G}_0 - \delta \mathbf{W} \succ 0 \Leftrightarrow \mathbf{I}_k \succ \delta \mathbf{G}_0^{-1/2} \mathbf{W} \mathbf{G}_0^{-1/2} \Leftrightarrow 1 > \delta \lambda_{\max}(\mathbf{G}_0^{-1/2} \mathbf{W} \mathbf{G}_0^{-1/2})\,, \tag{5.1.10}$$

yielding the selection $\delta = 1/(|\lambda_{\max}(\mathbf{G}_0^{-1/2} \mathbf{W} \mathbf{G}_0^{-1/2})| + 1)$. Then

$$(R+1)^k = \mathbf{v}_k^\top (\mathbf{G}_0 - \delta \mathbf{W}) \mathbf{v}_k + \delta \sum_{i\in[m-1]} g_i \mathbf{v}_{k-\lceil g_i \rceil}^\top \mathbf{I}_{k-\lceil g_i \rceil} \mathbf{v}_{k-\lceil g_i \rceil} + g_m \mathbf{v}_{k-1}^\top \mathbf{G}_m \mathbf{v}_{k-1}\,,$$

which implies $(R+1)^k \in \mathcal{Q}_k^\circ(\mathfrak{g})$, which in turn yields the desired conclusion. □

The next result is a consequence of Theorem 5.2. It states that if a POP has a ball constraint then the corresponding SOS strengthenings satisfy Slater's condition.

Corollary 5.1. *Assume that $R-\|\mathbf{x}\|_2^2 \in \mathfrak{g}$ for some $R>0$. Then Slater's condition holds for SDP* (2.4.4) *for all $k \geq k_{\min}$. As a consequence, strong duality holds for the primal-dual* (2.4.4)-(2.4.5) *for all $k \geq k_{\min}$.*

Proof. It suffices to prove that SDP (2.4.4) has a strictly feasible solution for all $k \geq k_{\min}$. Let $k \geq k_{\min}$ be fixed. By [139, Proposition 5.8], there exist $\sigma_0 \in \Sigma[\mathbf{x}]_k$, $\sigma \in \Sigma[\mathbf{x}]_{k-1}$ and $\lambda \in \mathbb{R}$ such that $f+\lambda = \sigma_0 + (R-\|\mathbf{x}\|_2^2)\sigma$. Thus $f+\lambda \in \mathcal{Q}_k(\mathfrak{g})$. By Theorem 5.2, $1 \in \mathcal{Q}_k^\circ(\mathfrak{g})$ and therefore $f+1+\lambda \in \mathcal{Q}_k^\circ(\mathfrak{g})$, which yields the desired conclusion. □

Remark 5.1. *From the proofs of Theorem 5.2 and Theorem 5.1, the constant trace a_k and the basis transformation matrix $\mathbf{P}_k$ (Definition 5.1) can be taken as*

$$a_k = (R+1)^k \quad \textit{and} \quad \mathbf{P}_k = \operatorname{diag}((\mathbf{G}_0 - \delta\mathbf{W})^{1/2}, \sqrt{\delta}\mathbf{I}_{k-\lceil g_1 \rceil}, \ldots, \sqrt{\delta}\mathbf{I}_{k-\lceil g_{m-1} \rceil}, \mathbf{G}_m^{1/2}).$$

However, this choice leads to poor numerical properties. In the next section we provide a series of linear programs inspired from the inclusion in (5.1.5)*, to obtain a constant trace a_k and a basis transformation matrix $\mathbf{P}_k$ that achieve better numerical performance.*

5.1.3 Verifying CTP for POPs by solving linear programs

For any $k \in \mathbb{N}^{\geq k_{\min}}$, let $\hat{\mathcal{S}}^{(k)}$ be the set of real diagonal matrices of size $b(k)$ and consider the following LP:

$$\inf_{\xi, \mathbf{G}_i, \mathbf{u}_j} \left\{ \xi \;\middle|\; \begin{array}{l} \mathbf{G}_0 - \mathbf{I}_0 \in \hat{\mathcal{S}}_+^{(k)}, \mathbf{G}_i - \mathbf{I}_i \in \hat{\mathcal{S}}_+^{(k-\lceil g_i \rceil)}, i \in [m], \\ \xi = \mathbf{v}_k^\top \mathbf{G}_0 \mathbf{v}_k + \sum_{i\in[m]} g_i \mathbf{v}_{k-\lceil g_i \rceil}^\top \mathbf{G}_i \mathbf{v}_{k-\lceil g_i \rceil} \\ \qquad + \sum_{j\in[l]} h_j \mathbf{v}_{2(k-\lceil h_j \rceil)}^\top \mathbf{u}_j \end{array} \right\}, \tag{5.1.11}$$

where $\mathbf{I}_i$ is the identity matrix for $i \in \{0\} \cup [m]$.

Lemma 5.2. *If LP* (5.1.11) *has a feasible solution $(\xi_k, \mathbf{G}_{i,k}, \mathbf{u}_{j,k})$ for every $k \in \mathbb{N}^{\geq k_{\min}}$, then POP* (5.0.2) *has CTP with $a_k = \xi_k$ and $\mathbf{P}_k = \operatorname{diag}(\mathbf{G}_{0,k}^{1/2}, \ldots, \mathbf{G}_{m,k}^{1/2})$.*

The proof of Lemma 5.2 is similar to that of Theorem 5.1 with $a_k = \xi_k$ and $\mathbf{G}_i = \mathbf{G}_{i,k}$, $i \in \{0\} \cup [m]$.

Since small constant traces are highly desirable for efficiency of first-order algorithms (e.g. CGAL), we search for an optimal solution of LP (5.1.11) instead of just a feasible solution.

Remark 5.2. *One can extend the classes of diagonal matrices $\hat{\mathcal{S}}^{(k)}$, $\hat{\mathcal{S}}^{(k-\lceil g_i \rceil)}$ in* (5.1.11) *to obtain a smaller constant trace. For instance, one can define $\hat{\mathcal{S}}^{(k)}$, $\hat{\mathcal{S}}^{(k-\lceil g_i \rceil)}$ to be the class of symmetric block diagonal matrices with block size two. As shown in [207, Lemma 4.3],* (5.1.11) *then becomes a second-order cone program which can be also efficiently solved.*

5.1.4 Special classes of POPs with CTP

In this section we identify two classes of POPs whose CTP can be verified by LP (5.1.11).

For $I \subseteq [n]$, let $\mathbf{x}(I) := \{x_j : j \in I\}$. For matrices $\mathbf{A}$ and $\mathbf{B}$ of the same size, the Hadamard product of $\mathbf{A}$ and $\mathbf{B}$, denoted by $\mathbf{A} \circ \mathbf{B}$, is the matrix with entries $[\mathbf{A} \circ \mathbf{B}]_{i,j} = A_{i,j}B_{i,j}$.

POPs with ball or annulus constraints on subsets of variables

Consider the following assumption on the inequality constraints of POP (5.0.2).

Assumption 5.1. *There exists a nonnegative integer $r \leq m/2$ and*

- *$\overline{R}_i > \underline{R}_i > 0$, $T_i \subseteq [n]$ for $i \in [r]$;*
- *$\overline{R}_j > 0$, $T_j \subseteq [n]$ for $j \in [m] \backslash [2r]$*

such that

(1) $(\cup_{i\in[r]}T_i)\cup(\cup_{j\in[m]\setminus[2r]}T_j)=[n]$*;*

(2) $g_i:=\|\mathbf{x}(T_i)\|_2^2-\underline{R}_i$, $g_{i+r}:=\overline{R}_i-\|\mathbf{x}(T_i)\|_2^2$ *for* $i\in[r]$*;*

(3) $g_i:=\overline{R}_i-\|\mathbf{x}(T_i)\|_2^2$ *for* $i\in[m]\setminus[2r]$.

Notice that if Assumption 5.1 holds then POP (5.0.2) has r annulus constraints and $(m-2r)$ ball constraints on subsets of variables. Moreover, $\mathcal{Q}(\mathfrak{g})+\mathcal{I}(\mathfrak{h})$ is Archimedean due to (1-3) in Assumption 5.1.

Example 5.2. *Assumption 5.1 holds in the following cases:*

(1) $m=1$, $r=0$ *and* $g_1:=\overline{R}_1-\|\mathbf{x}\|_2^2$*, i.e.,* $S(\mathfrak{g})$ *is a ball;*

(2) $m=n$, $r=0$ *and* $g_i:=\overline{R}_i-x_i^2$ *for* $i\in[n]$*, i.e.,* $S(\mathfrak{g})$ *is a box;*

(3) $m=2$, $r=1$ *and* $g_1:=\|\mathbf{x}\|_2^2-\underline{R}_1$, $g_2:=\overline{R}_1-\|\mathbf{x}\|_2^2$ *(*$\overline{R}_1>\underline{R}_1>0$*), i.e.,* $S(\mathfrak{g})$ *is an annulus.*

Proposition 5.1. *If Assumption 5.1 holds then LP* (5.1.11) *has a feasible solution for every* $k\in\mathbb{N}^{\geq k_{\min}}$*, and therefore POP* (5.0.2) *has CTP.*

Proof. Let Assumption 5.1 hold. It is sufficient to show that (5.1.11) has a feasible solution for every $k\in\mathbb{N}^{\geq k_{\min}}$.

Let $\mathbf{u}=(u_j)_{j\in[n]}\subseteq\mathbb{N}^{\leq m}$ be defined by

$$u_j:=|\{i\in[r]\ :\ j\in T_i\}|+|\{i\in[m]\setminus[2r]\ :\ j\in T_i\}|\,,\qquad\forall\ j\in[n]\,. \tag{5.1.12}$$

Since $(\cup_{i\in[r]}T_i)\cup(\cup_{i\in[m]\setminus[2r]}T_i)=[n]$, one has $u_j\in\mathbb{N}^{\geq 1}$, $j\in[n]$. Moreover,

$$\|\mathbf{u}\circ\mathbf{x}\|_2^2=\sum_{i\in[r]}\|\mathbf{x}(T_i)\|_2^2+\sum_{i\in[m]\setminus[2r]}\|\mathbf{x}(T_i)\|_2^2\,. \tag{5.1.13}$$

With $R:=\sum_{i\in[r]}(\underline{R}_i+\overline{R}_i)+\sum_{i\in[m]\setminus[2r]}\overline{R}_i$, by replacing $\mathbf{x}$ by $\mathbf{u}\circ\mathbf{x}$ in Lemma 5.1, one obtains that for all $k\in\mathbb{N}^{\geq k_{\min}}$,

$$(R+1)^k=(1+\|\mathbf{u}\circ\mathbf{x}\|_2^2)^k+\Lambda_{k-1}\sum_{i\in[m]}\delta_i g_i\,, \tag{5.1.14}$$

where $\Lambda_{k-1}:=\sum_{j=0}^{k-1}(R+1)^j(1+\|\mathbf{u}\circ\mathbf{x}\|_2^2)^{k-j-1}$ and

$$\delta_i:=\frac{\underline{R}_i}{\overline{R}_i-\underline{R}_i}\,,\ \delta_{i+r}:=\frac{\overline{R}_i}{\overline{R}_i-\underline{R}_i}\,,\ i\in[r]\,,\ \text{ and }\delta_q=1\,,\ q\in[m]\setminus[2r]. \tag{5.1.15}$$

It is due to the fact that

$$R-\|\mathbf{u}\circ\mathbf{x}\|_2^2=\sum_{i\in[r]}(\underline{R}_i+\overline{R}_i-\|\mathbf{x}(T_i)\|_2^2)+\sum_{i\in[m]\setminus[2r]}(\overline{R}_i-\|\mathbf{x}(T_i)\|_2^2)\,, \tag{5.1.16}$$

and $\underline{R}_i+\overline{R}_i-\|\mathbf{x}(T_i)\|_2^2=\delta_i g_i+\delta_{i+r}g_{i+r}$, for all $i\in[r]$. For each $k\in\mathbb{N}^{\geq k_{\min}}$, let $(\theta_{k,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_k}\subseteq\mathbb{R}^{>0}$ and $(\eta_{k-1,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-1}}\subseteq\mathbb{R}^{>0}$ be such that

$$(1+\|\mathbf{u}\circ\mathbf{x}\|_2^2)^k=\sum_{\boldsymbol{\alpha}\in\mathbb{N}^n_k}\theta_{k,\boldsymbol{\alpha}}\mathbf{x}^{2\boldsymbol{\alpha}}\quad\text{and}\quad\Lambda_{k-1}=\sum_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-1}}\eta_{k-1,\boldsymbol{\alpha}}\mathbf{x}^{2\boldsymbol{\alpha}}\,,$$

and define the diagonal matrices

$$\mathbf{G}_k^{(0)}:=\operatorname{diag}((\theta_{k,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_k})\quad\text{and}\quad\mathbf{G}_{k-1}^{(i)}:=\operatorname{diag}((\delta_i\eta_{k-1,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-1}})\,,\ i\in[m]\,. \tag{5.1.17}$$

Then (5.1.14) yields that for every $k\in\mathbb{N}^{\geq k_{\min}}$:

$$(R+1)^k=\mathbf{v}_k^\top\mathbf{G}_k^{(0)}\mathbf{v}_k+\sum_{i\in[m]}g_i\mathbf{v}_{k-1}^\top\mathbf{G}_{k-1}^{(i)}\mathbf{v}_{k-1}\,.$$

Hence $((R+1)^k,\mathbf{G}_k^{(i)},\mathbf{0})$ is a feasible solution of (5.1.11), for every $k\in\mathbb{N}^{\geq k_{\min}}$. □

POPs with inequality constraints of equivalent degree

We say that polynomials $q_1, \ldots, q_t$ are of equivalent degree if $\lceil q_1 \rceil = \cdots = \lceil q_t \rceil$.

Assumption 5.2. *Let $m \geq 3$ and $\{g_i\}_{i\in[m-2]}$ be of equivalent degree. $L > 0$ and $R > 0$ are such that $g_{m-1} = L - \sum_{i\in[m-2]} g_i$ and $g_m = R - \|\mathbf{x}\|_2^2$.*

Proposition 5.2. *If Assumption 5.2 holds then LP (5.1.11) has a feasible solution for every $k \in \mathbb{N}^{\geq k_{\min}}$, and therefore POP (5.0.2) has CTP.*

Proof. Let Assumption 5.2 hold with $u := \lceil g_i \rceil$, $i \in [n+1]$. For every $k \in \mathbb{N}^{\geq k_{\min}}$, letting $\Lambda_{k-1} := \sum_{j=0}^{k-1}(R+1)^j(1+\|\mathbf{x}\|_2^2)^{k-j-1}$ and $\Theta_t := (1+\|\mathbf{x}\|_2^2)^t$, for $t \in \mathbb{N}$, Lemma 5.1 yields $(R+1)^k = \Theta_k + g_m\Lambda_{k-1}$. It implies that for every $k \in \mathbb{N}^{\geq k_{\min}}$,

$$(R+1)^k = (\Theta_k - \frac{L}{L+1}\Theta_{k-u}) + \frac{1}{L+1}\Theta_{k-u}\sum_{i\in[m-1]} g_i + g_m\Lambda_{k-1}\,. \tag{5.1.18}$$

It is due to the fact that $\sum_{i\in[m-1]} g_i = L$. For each $k \in \mathbb{N}^{\geq k_{\min}}$, let us consider the following sequences:

- $(\nu_{k,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_k} \subseteq \mathbb{R}^{>0}$ such that $\Theta_k - \frac{L}{L+1}\Theta_{k-u} = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^n_k} \nu_{k,\boldsymbol{\alpha}}\mathbf{x}^{2\boldsymbol{\alpha}}$;
- $(\theta_{k-u,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-u}} \subseteq \mathbb{R}^{>0}$ such that $\frac{1}{L+1}\Theta_{k-u} = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-u}} \theta_{k-u,\boldsymbol{\alpha}}\mathbf{x}^{2\boldsymbol{\alpha}}$;
- $(\eta_{k-1,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-1}} \subseteq \mathbb{R}^{>0}$ such that $\Lambda_{k-1} = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-1}} \eta_{k-1,\boldsymbol{\alpha}}\mathbf{x}^{2\boldsymbol{\alpha}}$.

For each $k \in \mathbb{N}^{\geq k_{\min}}$, define the diagonal matrices $\mathbf{G}_k^{(0)} := \operatorname{diag}((\nu_{k,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_k})$,

$$\mathbf{G}_{k-u}^{(1)} := \operatorname{diag}((\theta_{k-u,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-u}})\,, \quad \text{and} \quad \mathbf{G}_{k-1}^{(2)} := \operatorname{diag}((\eta_{k-1,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-1}})\,.$$

Then (5.1.18) yields that for every $k \in \mathbb{N}^{\geq k_{\min}}$,

$$(R+1)^k = \mathbf{v}_k^\top\mathbf{G}_k^{(0)}\mathbf{v}_k + \mathbf{v}_{k-u}^\top\mathbf{G}_{k-u}^{(1)}\mathbf{v}_{k-u}\sum_{i\in[m-1]} g_i + \mathbf{v}_{k-1}^\top\mathbf{G}_{k-1}^{(2)}\mathbf{v}_{k-1}g_m\,. \tag{5.1.19}$$

Hence $((R+1)^k, \mathbf{G}_k^{(i)}, \mathbf{0})$ is a feasible solution of (5.1.11), for every $k \in \mathbb{N}^{\geq k_{\min}}$. By using Lemma 5.2, the conclusion follows. □

Example 5.3. *Let $R, L > 0$ satisfy $R \geq L^2$ and*

$$m = n+2\,,\ g_i = x_j \text{ for } j \in [n]\,,\ g_{n+1} = L - \sum_{j\in[n]} x_j \text{ and } g_{n+2} = R - \|\mathbf{x}\|_2^2\,. \tag{5.1.20}$$

Then Assumption 5.2 holds and $S(\mathfrak{g})$ is a simplex.

When $S(\mathfrak{g})$ is compact, we can always reformulate POP (5.0.2) such that Assumption 5.2 holds. Suppose $S(\mathfrak{g}) \subseteq B(\mathbf{0}, R)$ for some $R > 0$. Let $u = \max_{i\in[m]}\lceil g_i \rceil$. Set $\tilde{g}_i := g_i(1+\|\mathbf{x}\|_2^2)^{u-\lceil g_i\rceil}$ for $i \in [m]$. Let L be a positive number such that $\sum_{i\in[m]}\tilde{g}_i \leq L$ on $S(\mathfrak{g})$. Set $\tilde{g}_{m+1} := L - \sum_{i\in[m]}\tilde{g}_i$ and $\tilde{g}_{m+2} := R - \|\mathbf{x}\|_2^2$.

Remark 5.3. *For the latter case, one can choose any positive number $L \geq (R+1)^u\sum_{i\in[m]}\|g_i\|_1$. Indeed, for any $\mathbf{z} \in S(\mathfrak{g})$, and since $\|\mathbf{z}\|_2^2 \leq R$,*

$$|\mathbf{z}^{\boldsymbol{\alpha}}| = \prod_{i\in[n]}|z_i|^{\alpha_i} \leq \prod_{i\in[n]}(1+\|\mathbf{z}\|_2^2)^{\alpha_i/2} = (1+\|\mathbf{z}\|_2^2)^{|\boldsymbol{\alpha}|/2} \leq (1+R)^t\,,\ \forall \boldsymbol{\alpha} \in \mathbb{N}^n_{2t}\,.$$

This implies that for every $i \in [m]$,

$$\tilde{g}_i(\mathbf{z}) \leq (1+R)^{u-\lceil g_i\rceil}\sum_{\boldsymbol{\alpha}\in\mathbb{N}^n_{2\lceil g_i\rceil}}|g_{\boldsymbol{\alpha}}||\mathbf{z}^{\boldsymbol{\alpha}}| \leq (1+R)^{u-\lceil g_i\rceil}(R+1)^{\lceil g_i\rceil}\|g_i\|_1 = (1+R)^u\|g_i\|_1\,.$$

Thus we have $\sum_{i\in[m]}\tilde{g}_i \leq (1+R)^u\sum_{i\in[m]}\|g_i\|_1$ on $S(\mathfrak{g})$.

Corollary 5.2. *With the above notation, $S(\mathfrak{g} \cup \{\tilde{g}_{m+1}, \tilde{g}_{m+2}\}) = S(\mathfrak{g})$ and LP (5.1.11) has a feasible solution when replacing $\mathfrak{g}$ by $\mathfrak{g} \cup \{\tilde{g}_{m+1}, \tilde{g}_{m+2}\}$ for each $k \in \mathbb{N}^{\geq k_{\min}}$. As a result, POP (5.0.2) is equivalent to the following POP*

$$f^\star := \inf\{f(\mathbf{x}) \ : \ \mathbf{x} \in S(\mathfrak{g} \cup \{\tilde{g}_{m+1}, \tilde{g}_{m+2}\}) \cap V(\mathfrak{h})\} \tag{5.1.21}$$

which has CTP.

Proof. Let $\tilde{\mathfrak{g}} := \{\tilde{g}_i\}_{i\in[m+2]}$. Then $\{\tilde{g}_i\}_{i\in[m]}$ are of equivalent degree, i.e., there exists $u \in \mathbb{N}$ such that $\lceil \tilde{g}_i \rceil = u$, for all $i \in [m]$. Thus Assumption 5.2 holds with $\mathfrak{g} \leftarrow \tilde{\mathfrak{g}}$, $m \leftarrow m+2$. By Proposition 5.2, (5.1.11) has a feasible solution with $\mathfrak{g} \leftarrow \tilde{\mathfrak{g}}$ for every order $k \in \mathbb{N}^{\geq k_{\min}}$. It implies that for every $k \in \mathbb{N}^{\geq k_{\min}}$, there exist $\mathbf{u}_k^{(j)} \in \mathbb{R}^{b(2(k-\lceil h_j \rceil))}$, $j \in [l]$, and

$$(\eta^{(0)}_{k,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_k} \subseteq \mathbb{R}^{>0}\,, \quad (\eta^{(i)}_{k-u,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-u}} \subseteq \mathbb{R}^{>0}\,,\, i \in [m+1]\,, \quad (\eta^{(m+2)}_{k-1,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-1}} \subseteq \mathbb{R}^{>0}$$

such that

$$\begin{aligned} 1 = \ & \mathbf{v}_k^\top \operatorname{diag}((\eta^{(0)}_{k,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_k})\mathbf{v}_k + \textstyle\sum_{i\in[m+1]} \tilde{g}_i \mathbf{v}_{k-u}^\top \operatorname{diag}((\eta^{(i)}_{k-u,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-u}})\mathbf{v}_{k-u} \\ & + \tilde{g}_{m+2}\mathbf{v}_{k-1}^\top \operatorname{diag}((\eta^{(m+2)}_{k-1,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-1}})\mathbf{v}_{k-1} + \textstyle\sum_{j\in[l]} h_j \mathbf{v}^\top_{2(k-\lceil h_j\rceil)}\mathbf{u}_k^{(j)}\,. \end{aligned}$$

Let $k \in \mathbb{N}^{\geq k_{\min}}$ be fixed. We define the following polynomials:

- $\sigma_0 := \mathbf{v}_k^\top \operatorname{diag}((\eta^{(0)}_{k,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_k})\mathbf{v}_k = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^n_k} \eta^{(0)}_{k,\boldsymbol{\alpha}}\mathbf{x}^{2\boldsymbol{\alpha}}$,
- $\sigma_i := \mathbf{v}_{k-u}^\top \operatorname{diag}((\eta^{(i)}_{k-u,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-u}})\mathbf{v}_{k-u} = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-u}} \eta^{(i)}_{k-u,\boldsymbol{\alpha}}\mathbf{x}^{2\boldsymbol{\alpha}}$, $i \in [m+1]$,
- $\sigma_{m+2} := \mathbf{v}_{k-1}^\top \operatorname{diag}((\eta^{(m+2)}_{k-1,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-1}})\mathbf{v}_{k-1} = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-1}} \eta^{(m+2)}_{k-1,\boldsymbol{\alpha}}\mathbf{x}^{2\boldsymbol{\alpha}}$,
- $\psi_j := \mathbf{v}^\top_{2(k-\lceil h_j\rceil)}\mathbf{u}_k^{(j)}$, $j \in [l]$.

From this and since $\tilde{g}_i := g_i(1+\|\mathbf{x}\|_2^2)^{u-\lceil g_i\rceil}$, for $i \in [m]$, one has

$$\begin{aligned} 1 = \ & \sigma_0 + \textstyle\sum_{i\in[m]} \sigma_i \tilde{g}_i + \sum_{j\in[l]} \psi_j h_j = \sigma_0 + \sum_{i\in[m]} \sigma_i (1+\|\mathbf{x}\|_2^2)^{u-\lceil g_i\rceil} g_i \\ & + \tilde{g}_{m+1}\sigma_{m+1} + \tilde{g}_{m+2}\sigma_{m+2} + \textstyle\sum_{j\in[l]} \psi_j h_j\,. \end{aligned} \tag{5.1.22}$$

Then there exist $(\theta^{(i)}_{k-\lceil g_i\rceil,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-\lceil g_i\rceil}} \subseteq \mathbb{R}^{>0}$, $i \in [m]$, such that

$$\sigma_i(1+\|\mathbf{x}\|_2^2)^{u-\lceil g_i\rceil} = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-\lceil g_i\rceil}} \theta^{(i)}_{k-\lceil g_i\rceil,\boldsymbol{\alpha}}\mathbf{x}^{2\boldsymbol{\alpha}}\,,\, i \in [m]\,. \tag{5.1.23}$$

Thus (5.1.22) becomes

$$\begin{aligned} 1 = \ & \mathbf{v}_k^\top \operatorname{diag}((\eta^{(0)}_{k,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_k})\mathbf{v}_k + \textstyle\sum_{i\in[m]} g_i \mathbf{v}_{k-\lceil g_i\rceil}^\top \operatorname{diag}((\theta^{(i)}_{k-\lceil g_i\rceil,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-\lceil g_i\rceil}})\mathbf{v}_{k-\lceil g_i\rceil} \\ & + \tilde{g}_{m+1}\mathbf{v}_{k-u}^\top \operatorname{diag}((\eta^{(m+1)}_{k-u,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-u}})\mathbf{v}_{k-u} \\ & + \tilde{g}_{m+2}\mathbf{v}_{k-1}^\top \operatorname{diag}((\eta^{(m+2)}_{k-1,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-1}})\mathbf{v}_{k-1} + \textstyle\sum_{j\in[l]} h_j \mathbf{v}^\top_{2(k-\lceil h_j\rceil)}\mathbf{u}_k^{(j)} \\ & \in \mathcal{Q}_k^\circ(\mathfrak{g} \cup \{\tilde{g}_{m+1}, \tilde{g}_{m+2}\}) + \mathcal{I}_k(\mathfrak{h})\,, \end{aligned}$$

since

- $\operatorname{diag}((\eta^{(0)}_{k,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_k}) \succ 0$, $\operatorname{diag}((\theta^{(i)}_{k-\lceil g_i\rceil,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-\lceil g_i\rceil}}) \succ 0$, $i \in [m]$,
- $\operatorname{diag}((\eta^{(m+1)}_{k-u,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-u}}) \succ 0$, and $\operatorname{diag}((\eta^{(m+2)}_{k-1,\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{k-1}}) \succ 0$.

It yields that (5.1.11) has a feasible solution with $\mathfrak{g} \leftarrow \mathfrak{g} \cup \{\tilde{g}_{m+1}, \tilde{g}_{m+2}\}$, for every order $k \in \mathbb{N}^{\geq k_{\min}}$. □

In case that POP (5.0.2) does not have CTP and $S(\mathfrak{g})$ is compact, Corollary 5.2 provides a way to construct an equivalent POP by including two additional redundant constraints. Then CTP of this new POP can be verified by LP.

Algorithm 6 Approximating the optimal value of a dense POP with CTP

Input: POP (5.0.2) and a relaxation order $k \in \mathbb{N}^{\geq k_{\min}}$
Output: The optimal value τ_k of SDP (5.1.3)

1: Solve LP (5.1.11) with an optimal solution $(\xi_k, \mathbf{G}_{i,k}, \mathbf{u}_{j,k})$;
2: Let $a_k = \xi_k$ and $\mathbf{P}_k = \text{diag}(\mathbf{G}_{0,k}^{1/2}, \ldots, \mathbf{G}_{m,k}^{1/2})$;
3: Compute the optimal value τ_k of SDP (5.1.3) by running an algorithm based on first-order methods and which exploits CTP.

5.1.5 Main algorithm

Algorithm 6 below solves POP (5.0.2) whose CTP can be verified by LP.

Examples of algorithms based on first-order methods and which exploit CTP are CGAL (Algorithm 8 in Appendix 5.3.4) or SM (Algorithm 10 in Appendix 5.3.5).

5.2 Numerical experiments for dense POPs

In this section we report results of numerical experiments obtained by solving the second order Moment-SOS relaxation of various randomly generated instances of QCQPs with CTP. The experiments were performed in Julia 1.3.1 with the following software packages:

- `SumOfSquare` [216] is a modeling library for solving the Moment-SOS relaxations of dense POPs, based on JuMP (with `Mosek` 9.1 as SDP solver).

- `TSSOS` [212, 211, 213] is a modeling library for solving Moment-SOS relaxations of sparse POPs based on JuMP (with `Mosek` 9.1 as SDP solver).

- `LMBM` solves unconstrained non-smooth optimization with the limited-memory bundle method by Haarala et al. [71, 70] and calls Karmitsa's Fortran implementation of the LMBM algorithm [95].

- `Arpack` [115] is used to compute the smallest eigenvalues and the corresponding eigenvectors of real symmetric matrices of (potentially) large size, which is based on the implicitly restarted Arnoldi method.

The implementation of Algorithms 6 and 7 is available online via the link:

https://github.com/maihoanganh/ctpPOP.

We use a desktop computer with an Intel(R) Core(TM) i7-8665U CPU @ 1.9GHz × 8 and 31.2 GB of RAM. The notation for the numerical results is given in Table 5.1.

For the examples tested in this chapter, the modeling time of `SumOfSquares`, `TSSOS` and `ctpPOP` is typically negligible compared to the solving time of the packages `Mosek`, CGAL, and LMBM. Hence the total running time mainly depends on the solvers and we compare their performances below. As mentioned in the introduction, the current framework differs from our previous work [134], where we exploited CTP for equality constrained POPs on a sphere, which could be solved by `LMBM` efficiently. The reason is that the SDP relaxations of such equality constrained POPs involve a single psd matrix. For the benchmarks of this section, we consider POPs involving ball/annulus constraints, and so the resulting relaxations include several psd matrices. Our numerical experiments confirm that for such SDPs, `LMBM` returns inaccurate values since the gap w.r.t. the value of `Mosek` is typically larger than 1% while CGAL (without sketching) performs better for this type of SDPs in terms of accuracy. As showed in Section 5.2.1, the last columns of Table 5.2 and Table 5.3 illustrate how inaccurate `LMBM` can be for large problems ($n \geq 20$), thus we do not report `LMBM` results in the other experiments.

Table 5.1: Notation

n	the number of variables of a POP
m	the number of inequality constraints of a POP
l	the number of equality constraints of a POP
$u^{\max}$	the largest size of variable cliques of a sparse POP
p	the number of variable cliques of a sparse POP
k	the relaxation order of the Moment-SOS hierarchy
t	the sparse order of the sparsity adapted Moment-SOS hierarchy (for TS and CS-TS)
ω	the number of psd blocks of an SDP
$s^{\max}$	the largest size of psd blocks of an SDP
ζ	the number of affine equality constraints of an SDP
$a^{\max}$	the largest constant trace
`Mosek`	the SDP relaxation modeled by `SumOfSquares` (for dense POPs) or `TSSOS` (for sparse POPs) and solved by `Mosek` 9.1
CGAL	the SDP relaxation modeled by our CTP-exploiting method and solved by the CGAL algorithm
LMBM	the SDP relaxation modeled by our CTP-exploiting method and solved by the SM algorithm with the `LMBM` solver
val	the optimal value of the SDP relaxation
gap	the relative optimality gap w.r.t. the value returned by `Mosek`, i.e., gap $= \|\text{val} - \text{val}(\texttt{Mosek})\|/\|\text{val}(\texttt{Mosek})\|$
time	the running time in seconds (including modeling and solving time)
$-$	the calculation runs out of space

5.2.1 Randomly generated dense QCQPs with a ball constraint

Test problems: We construct randomly generated dense QCQPs with a ball constraint as follows:

1. Generate a dense quadratic polynomial objective function f with random coefficients following the uniform probability distribution on $(-1, 1)$;

2. Let $m = 1$ and $g_1 := 1 - \|\mathbf{x}\|_2^2$;

3. Take a random point $\mathbf{a}$ in $S(\mathfrak{g})$ w.r.t. the uniform distribution;

4. For every $j \in [l]$, generate a dense quadratic polynomial h_j by

   (i) for each $\boldsymbol{\alpha} \in \mathbb{N}_2^n \backslash \{\mathbf{0}\}$, taking a random coefficient $h_{j,\boldsymbol{\alpha}}$ for h_j in $(-1, 1)$ w.r.t. the uniform distribution;

   (ii) setting $h_{j,\mathbf{0}} := -\sum_{\boldsymbol{\alpha} \in \mathbb{N}_2^n \backslash \{\mathbf{0}\}} h_{j,\boldsymbol{\alpha}} \mathbf{a}^{\boldsymbol{\alpha}}$.

   Then $\mathbf{a}$ is a feasible solution of POP (5.0.2).

The numerical results are displayed in Table 5.2 and 5.3.

Discussion: As one can see from Table 5.2 and 5.3, CGAL is typically the fastest solver and returns an optimal value of gap within 1% w.r.t. the one returned by `Mosek` when $n \leq 30$. `Mosek` runs out of memory when $n \geq 40$ while CGAL works well up to $n = 100$. We should point out that LMBM is less accurate or even fails to converge to the optimal value when $n \geq 20$. The reason might be that LMBM only solves the dual problem and hence looses information of the primal problem.

Table 5.2: Numerical results for minimizing a dense quadratic polynomial on a unit ball

- POP size: $m = 1$, $l = 0$; Relaxation order: $k = 2$; SDP size: $\omega = 2$, $a^{\max} = 3$.

POP size	SDP size		Mosek		CGAL		LMBM	
n	$s^{\max}$	ζ	val	time	val	time	val	time
10	66	1277	-2.2181	0.3	-2.2170	0.2	-2.2187	0.3
20	231	16402	-3.7973	4	-3.7947	0.6	-3.7096	7
30	496	77377	-3.6876	3474	-3.6858	104	-3.8530	59
40	861	236202	–	–	-4.1718	33	-4.7730	179
50	1326	564877	–	–	-6.3107	1007	-7.3874	139
60	1891	1155402	–	–	-6.5326	1085	-7.4733	674
70	2556	2119777	–	–	-7.3379	1262	-9.5223	1486
80	3321	3590002	–	–	-7.9559	4988	-10.0260	1241
90	4186	5718077	–	–	-7.3425	5187	-9.4477	5313
100	5151	8676002	–	–	-7.7374	22451	-10.684	5355

Table 5.3: Numerical results for randomly generated dense QCQPs with a ball constraint

- POP size: $m = 1$, $l = \lceil n/4 \rceil$; Relaxation order: $k = 2$; SDP size: $\omega = 2$, $a^{\max} = 3$.

POP size		SDP size		Mosek		CGAL		LMBM	
n	l	$s^{\max}$	ζ	val	time	val	time	val	time
10	3	66	1475	-2.0686	1.7	-2.0674	0.8	-2.0874	0.3
20	5	231	17557	-3.0103	61	-3.0075	7	-3.0750	18
30	8	496	81345	-3.3293	4573	-3.3249	80	-3.6863	123
40	10	861	244812	–	–	-4.6977	194	-5.3488	488
50	13	1326	582115	–	–	-4.2394	951	-6.1325	837
60	15	1891	1183767	–	–	-5.7793	1387	-7.5718	3781
70	18	2556	2165785	–	–	-6.1278	4335	-8.1181	15854

5.2.2 Randomly generated dense QCQPs with annulus constraints

Test problems: We construct randomly generated dense QCQPs as in Section 5.2.1, where the ball constraint is now replaced by annulus constraints. Namely, in Step 2 we take $m = 2$, $g_1 := \|\mathbf{x}\|_2^2 - 1/2$ and $g_2 := 1 - \|\mathbf{x}\|_2^2$. The numerical results are displayed in Table 5.4 and 5.5.

Discussion: Same remarks as in Section 5.2.1.

5.2.3 Randomly generated dense QCQPs with box constraints

Test problems: We construct randomly generated dense QCQPs as in Section 5.2.1, where the ball constraint is now replaced by box constraints. Namely, in Step 2 we take $m = n$, $g_j := -x_j^2 + 1/n$, $j \in [n]$.

The numerical results are displayed in Table 5.6 and 5.7.

Discussion: We observe similar behaviors of the solvers as in Section 5.2.1. The important point to note here is that solving a QCQP with box constraints is less efficient than solving the same one with ball constraints. This is because the efficiency of CGAL depends on the number of psd blocks involved in an SDP. For instance, when $n = 50$, CGAL takes around 1000 seconds to solve the second order moment relaxation of a QCQP with a ball constraint while it takes around 2100 seconds to solve this relaxation for a QCQP with box constraints.

Table 5.4: Numerical results for minimizing a dense quadratic polynomial on an annulus

- POP size: $m = 2$, $l = 0$; Relaxation order: $k = 2$; SDP size: $\omega = 3$, $a^{\max} = 4$.

POP size	SDP size		Mosek		CGAL	
n	$s^{\max}$	ζ	val	time	val	time
10	66	1343	-3.0295	0.5	-3.0278	1
20	231	16633	-3.6468	69	-3.6458	5
30	496	77873	-3.9108	2546	-3.9079	9
40	861	237063	–	–	-4.7469	28
50	1326	566203	–	–	-6.4170	112
60	1891	1157293	–	–	-5.5841	226
70	2556	2122333	–	–	-7.9325	730
80	3321	3593323	–	–	-7.6164	1355
90	4186	5722263	–	–	-8.1900	3563

Table 5.5: Numerical results for randomly generated dense QCQPs with annulus constraints

- POP size: $m = 2$, $l = \lceil n/4 \rceil$; Relaxation order: $k = 2$; SDP size: $\omega = 3$, $a^{\max} = 4$.

POP size		SDP size		Mosek		CGAL	
n	l	$s^{\max}$	ζ	val	time	val	time
10	3	66	1541	-2.7950	0.5	-2.7934	2
20	5	231	17788	-3.5048	95	-3.5027	10
30	8	496	81841	-3.3964	4237	-3.3937	45
40	10	861	245673	–	–	-4.6573	140
50	13	1326	583441	–	–	-3.8236	437
60	15	1891	1185658	–	–	-4.5246	1076
70	18	2556	2168341	–	–	-6.2924	4783

Table 5.6: Numerical results for minimizing a dense quadratic polynomial on a box

- POP size: $m = n$, $l = 0$; Relaxation order: $k = 2$; SDP size: $\omega = n + 1$, $a^{\max} = 3$.

POP size	SDP size		Mosek		CGAL	
n	$s^{\max}$	ζ	val	time	val	time
10	66	1871	-2.7197	0.5	-2.7189	1
20	231	20791	-3.3560	98	-3.3501	57
30	496	91761	-4.6372	5150	-4.6242	285
40	861	269781	–	–	-4.5788	409
50	1326	629851	–	–	-4.2313	2083
60	1891	1266971	–	–	-4.0135	5525
70	2556	2296141	–	–	-5.4019	15172

Table 5.7: Numerical results for randomly generated dense QCQPs with box constraints

- POP size: $m = n$, $l = \lceil n/7 \rceil$; Relaxation order: $k = 2$; SDP size: $\omega = n + 1$, $a^{\max} = 3$.

POP size		SDP size		Mosek		CGAL	
n	l	$s^{\max}$	ζ	val	time	val	time
10	2	66	2003	-1.8320	0.6	-1.8321	3
20	3	231	21484	-3.1797	175	-3.1781	106
30	5	496	94241	-2.2949	6850	-2.2982	528
40	6	861	274947	–	–	-3.8651	933
50	8	1326	640459	–	–	-3.6267	6159

Table 5.8: Numerical results for minimizing a dense quadratic polynomials on a simplex

- POP size: $m = n + 2$, $l = 0$; Relaxation order: $k = 2$; SDP size: $\omega = n + 3$, $a^{\max} = 5$.

POP size	SDP size		Mosek		CGAL	
n	$s^{\max}$	ζ	val	time	val	time
10	66	2003	-1.9954	0.3	-1.9950	7
20	231	21253	-1.5078	58	-1.5055	116
30	496	92753	-2.0537	2804	-2.0480	377
40	861	271503	–	–	-2.3034	950
50	1326	632503	–	–	-1.8366	9539

Table 5.9: Numerical results for randomly generated dense QCQPs with simplex constraints

- POP size: $m = n + 2$, $l = \lceil n/7 \rceil$; Relaxation order: $k = 2$; SDP size: $\omega = n + 3$, $a^{\max} = 5$.

POP size		SDP size		Mosek		CGAL	
n	l	$s^{\max}$	ζ	val	time	val	time
10	2	66	2135	-1.0605	0.4	-1.0606	176
20	3	231	21946	-1.6629	72	-1.6628	512
30	5	496	95233	-1.0091	6206	-1.0249	1089
40	6	861	276669	–	–	-0.3256	2314
50	8	1326	643111	–	–	-1.4200	10035

5.2.4 Randomly generated dense QCQPs with simplex constraints

Test problems: We construct randomly generated dense QCQPs as in Section 5.2.1, where the ball constraint is now replaced by simplex constraints. Namely, in Step 2 we take g such that (5.1.20) holds with $L = R = 1$. The numerical results are displayed in Table 5.8 and 5.9.

Discussion: Again we observe a behavior of the solvers similar to that in Section 5.2.1. One can also see that solving a QCQP with simplex constraints by CGAL is significantly slower than solving the same one with box constraints. For instance, when $n = 50$, CGAL takes 2100 seconds to solve the second order moment relaxation for a QCQP with box constraints while it takes 9500 seconds with simplex constraints.

5.2.5 Numerical comparison between CGAL and ADMM

In Table 5.10, we make a numerical comparison between CGAL (with our CTP-exploiting method) and `COSMO`, an SDP solver based on ADMM (see Table 1.3), on some randomly generated dense QCQPs with a ball constraint (as in Section 5.2.1).

Discussion: Table 5.10 indicates that both CGAL and `COSMO` provide approximate values with gap within 1% w.r.t. the ones returned by `Mosek` when $n \leq 30$. In addition, `COSMO` is slightly more accurate for $n \in \{20, 30\}$ while CGAL offers an increasing speedup when $n \geq 30$.

5.2.6 Dense POPs with a ball constraint

We construct randomly generated dense POPs as in Section 5.2.1, with input data of degree $d \in \{3, 4\}$.

The numerical results, displayed in Table 5.11, indicate that CGAL returns an optimal value with gap within 1% w.r.t. the one of `Mosek`, and is faster when the largest block size increases.

Table 5.10: Numerical comparison with ADMM (`COSMO`) on randomly generated dense QCQPs with a ball constraint

- POP size: $m = 1$, $l = \lceil n/4 \rceil$; Relaxation order: $k = 2$; SDP size: $\omega = 2$, $a^{\max} = 3$.

POP size		SDP size		`Mosek`		CGAL		`COSMO`	
n	l	$s^{\max}$	ζ	val	time	val	time	val	time
10	3	66	1475	-2.3153	0.7	-2.3134	0.1	-2.3125	0.2
20	5	231	17557	-3.6585	57	-3.6562	7	-3.6582	5
30	8	496	81345	-4.6221	4670	-4.6177	69	-4.6230	91
40	10	861	244812	–	–	-4.9932	173	-4.9989	532
50	13	1326	582115	–	–	-5.0394	524	-5.0418	2468
60	15	1891	1183767	–	–	-5.3537	735	-5.3548	6176

Table 5.11: Numerical results for randomly generated dense POPs with a ball constraint

- POP size: $m = 1$, $l = \lceil n/4 \rceil$; SDP size: $\omega = 2$.

POP size			SDP size				`Mosek`		CGAL	
n	l	d	k	$a^{\max}$	$s^{\max}$	ζ	val	time	val	time
15	4	3	2	3	136	5581	-3.0127	5	-3.0089	1
			3	4	816	288933	–	–	-3.0021	290
10	3	4	2	3	66	1280	-2.0327	0.3	-2.0194	0.6
			3	4	286	35443	-1.9337	41	-1.9310	16

5.3 Appendix

5.3.1 Exploiting CTP for POPs with CS

In this section, we extend the CTP-exploiting framework to POPs with sparsity. For clarity of exposition we only consider correlative sparsity (CS). However, in Appendix 5.3.2 we also treat term sparsity (TS) [212] as well as correlative-term sparsity (CS-TSSOS) [213]. Since the methodology is very similar to that in the dense case described earlier, we omit details and only present the main results.

To make this appendix self-contained, we recall some basic facts already stated in Chapter 3 on exploiting CS for POP (5.0.2) initially proposed in [203] by Waki et al.

For $\boldsymbol{\alpha} \in \mathbb{N}^n$, let $\mathrm{supp}(\boldsymbol{\alpha}) := \{j \in [n] : \alpha_j > 0\}$. For $I \subseteq [n]$, let $\mathbf{x}(I) := \{x_j : j \in I\}$ and $\mathbb{N}^I_d := \{\boldsymbol{\alpha} \in \mathbb{N}^n_d : \mathrm{supp}(\boldsymbol{\alpha}) \subseteq I\}$. Assume $I \subseteq [n]$. Given $\mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in \mathbb{N}^n_{2d}}$, the moment (resp. localizing) submatrix associated with I of order d is defined by $\mathbf{M}_d(\mathbf{y}, I) := (y_{\boldsymbol{\alpha}+\boldsymbol{\beta}})_{\alpha,\beta \in \mathbb{N}^I_d}$ (resp. $\mathbf{M}_d(q\mathbf{y}, I) := (\sum_{\boldsymbol{\gamma}} q_{\boldsymbol{\gamma}} y_{\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\gamma}})_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{N}^I_d}$ for $q \in \mathbb{R}[\mathbf{x}(I)]$). Let $\mathbf{v}^I_d := (\mathbf{x}^{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in \mathbb{N}^I_d}$ with length $b(|I|, d) := \binom{|I|+d}{n}$. For matrices $\mathbf{A}$ and $\mathbf{B}$ of same sizes, the Hadamard product of $\mathbf{A}$ and $\mathbf{B}$, denoted by $\mathbf{A} \circ \mathbf{B}$, is the matrix with entries $[\mathbf{A} \circ \mathbf{B}]_{i,j} = A_{i,j} B_{i,j}$.

POPs with CS

Assume that $\{I_c\}_{c \in [p]}$ (with $n_c := |I_c|$) are the maximal cliques of (a chordal extension of) the correlative sparsity pattern (csp) graph associated with POP (5.0.2), as defined in [203].

Let $\{J_c\}_{c \in [p]}$ (resp. $\{W_c\}_{c \in [p]}$) be a partition of $[m]$ (resp. $[l]$) such that for all $i \in J_c$, $g_i \in \mathbb{R}[\mathbf{x}(I_c)]$ (resp. $i \in W_c$, $h_i \in \mathbb{R}[\mathbf{x}(I_c)]$), $c \in [p]$. For each $c \in [p]$, let $m_c := |J_c|$, $R_c := |W_c|$ and $\mathfrak{g}_{J_c} := \{g_i : i \in J_c\}$, $\mathfrak{h}_{W_c} := \{h_i : i \in W_c\}$. Then $\mathcal{Q}(\mathfrak{g}_{J_c})$ (resp. $\mathcal{I}(\mathfrak{h}_{W_c})$) is a quadratic module (resp. an ideal) in $\mathbb{R}[\mathbf{x}(I_c)]$, for $c \in [p]$.

For each $k \in \mathbb{N}^{\geq k_{\min}}$, consider the following sparse SOS strengthening:

$$\rho^{\mathrm{cs}}_k := \sup \left\{ \xi : f - \xi \in \sum_{c \in [p]} (\mathcal{Q}_k(\mathfrak{g}_{J_c}) + \mathcal{I}_k(\mathfrak{h}_{W_c})) \right\}. \tag{5.3.1}$$

It is equivalent to the SDP:

$$\rho_k^{\rm cs} = \sup_{\xi,\mathbf{G}_i^{(c)},\lceil g_i\rceil^{(c)}} \left\{ \xi \;\middle|\; \begin{array}{l} \mathbf{G}_i^{(c)} \succeq 0\,,\ i\in\{0\}\cup J_c\,,\ c\in[p]\,,\\ f-\xi = \sum_{c\in[p]} \Big((\mathbf{v}_k^{I_c})^\top \mathbf{G}_0^{(c)} \mathbf{v}_k^{I_c} \\ \qquad\qquad + \sum_{i\in J_c} g_i (\mathbf{v}_{k-\lceil g_i\rceil}^{I_c})^\top \mathbf{G}_i^{(c)} \mathbf{v}_{k-\lceil g_i\rceil}^{I_c} \\ \qquad\qquad + \sum_{j\in W_c} h_j (\mathbf{v}_{2(k-\lceil h_j\rceil)}^{I_c})^\top \mathbf{u}_j^{(c)} \Big) \end{array} \right\}. \tag{5.3.2}$$

The dual of (5.3.2) reads

$$\tau_k^{\rm cs} := \inf_{\mathbf{y}\in\mathbb{R}^{b(2k)}} \left\{ L_{\mathbf{y}}(f) \;\middle|\; \begin{array}{l} \mathbf{M}_k(\mathbf{y}, I_c) \succeq 0\,,\ c\in[p]\,,\ y_{\mathbf{0}} = 1\,.\\ \mathbf{M}_{k-\lceil g_i\rceil}(g_i\,\mathbf{y}, I_c) \succeq 0\,,\ i\in J_c\,,\ c\in[p]\,,\\ \mathbf{M}_{k-\lceil h_j\rceil}(h_j\,\mathbf{y}, I_c) = 0\,,\ j\in W_c\,,\ c\in[p] \end{array} \right\}. \tag{5.3.3}$$

It is shown in [103, Theorem 3.6] that convergence of the primal-dual (5.3.2)-(5.3.3) to $f^\star$ is guaranteed if there are additional ball constraints on each clique of variables.

Exploiting CTP for POPs with CS

For every $c\in[p]$, we denote by $\mathcal{S}^{(c,k)}$ the set of real symmetric matrices of size $b(n_c,k) + \sum_{i\in J_c} b(n_c, k-\lceil g_i\rceil)$ in a block diagonal form: $\mathbf{X} = \mathrm{diag}(\mathbf{X}_0, (\mathbf{X}_i)_{i\in J_c})$ such that $\mathbf{X}_0$ is a block of size $b(k,n_c)$ and $\mathbf{X}_i$ is a block of size $b(n_c, k-\lceil g_i\rceil)$ for $i\in J_c$.

Consider POP (5.0.2) with CS described in Section 5.3.1. For every $c\in[p]$ and for every $k\in\mathbb{N}^{\geq k_{\min}}$, letting $\mathbf{D}_k(\mathbf{y}, I_c) := \mathrm{diag}(\mathbf{M}_k(\mathbf{y}, I_c), (\mathbf{M}_{k-\lceil g_i\rceil}(g_i\mathbf{y}, I_c))_{i\in J_c})$ for $\mathbf{y}\in\mathbb{R}^{s(2k)}$, SDP (5.3.3) can be rewritten as

$$\tau_k^{\rm cs} := \inf_{\mathbf{y}\in\mathbb{R}^{b(2k)}} \left\{ L_{\mathbf{y}}(f) \;\middle|\; \begin{array}{l} \mathbf{D}_k(\mathbf{y}, I_c) \succeq 0\,,\ j\in[p]\,,\ y_{\mathbf{0}} = 1\,,\\ \mathbf{M}_{k-\lceil h_j\rceil}(h_j\,\mathbf{y}, I_c) = 0\,,\ j\in W_c\,,\ c\in[p] \end{array} \right\}. \tag{5.3.4}$$

We define CTP for POP with CS as follows.

Definition 5.2. *(CTP for a POP with CS) We say that POP* (5.0.2) *with CS has CTP if for every $k\in\mathbb{N}^{\geq k_{\min}}$ and for every $c\in[p]$, there exists a positive number $a_k^{(c)}$ and a positive definite matrix $\mathbf{P}_k^{(c)}\in\mathcal{S}^{(c,k)}$ such that for all $\mathbf{y}\in\mathbb{R}^{b(2k)}$,*

$$\left.\begin{array}{l} \mathbf{M}_{k-\lceil h_j\rceil}(h_i\mathbf{y}, I_c) = 0\,,\ j\in W_c\,,\\ y_{\mathbf{0}} = 1 \end{array}\right\} \Rightarrow \mathrm{trace}(\mathbf{P}_k^{(c)}\mathbf{D}_k(\mathbf{y}, I_c)\mathbf{P}_k^{(c)}) = a_k^{(c)}\,. \tag{5.3.5}$$

The following result provides a sufficient condition for a POP with CS to have CTP.

Theorem 5.3. *Assume that there is a ball constraint on each clique of variables, i.e.,*

$$\forall\ c\in[p],\ R_c - \|\mathbf{x}(I_c)\|_2^2 \in g \text{ for some } R_c > 0\,. \tag{5.3.6}$$

Then one has $\mathbb{R}^{>0}\subseteq \mathcal{Q}_k^\circ(\mathfrak{g}_{J_c})$, for all $k\in\mathbb{N}^{\geq k_{\min}}$ and for all $c\in[p]$. As a consequence, POP (5.0.2) *has CTP.*

The proof of Theorem 5.3 being very similar to that of Theorem 5.2 by considering each clique of variables, is omitted.

Again by considering each clique of variables, the following result can be obtained from Theorem 5.3 in the same way Corollary 5.1 was obtained.

Corollary 5.3. *If* (5.3.6) *holds then Slater's condition for SDP* (5.3.2) *holds for all $k\in\mathbb{N}^{\geq k_{\min}}$.*

We are now in position to provide a general method to solve POPs with CS which have CTP.

Consider POP (5.0.2) with CS described in Section 5.3.1. Assume that POP (5.0.2) has CTP and let $k\in\mathbb{N}^{\geq k_{\min}}$ be fixed.

Letting

$$\mathbf{X}_c = \mathbf{P}_k^{(c)}\mathbf{D}_k(\mathbf{y}, I_c)\mathbf{P}_k^{(c)}\,,\ c\in[p]\,, \tag{5.3.7}$$

SDP (5.3.4) can be rewritten as

$$\tau_k^{\rm cs} = \inf_{\mathbf{X}_c \in \mathcal{S}_+^{(c,k)}} \left\{ \sum_{c\in[p]} \langle \mathbf{C}_{c,k}, \mathbf{X}_c \rangle \ : \ \sum_{c\in[p]} \mathcal{A}_{c,k}\mathbf{X}_c = \mathbf{b}_k \, , \, c \in [p] \right\} , \tag{5.3.8}$$

where for every $c \in [p]$, $\mathcal{A}_{c,k} : \mathcal{S}^{(c,k)} \to \mathbb{R}^{\zeta_k}$ is a linear operator of the form

$$\mathcal{A}_{c,k}\mathbf{X} = (\langle \mathbf{A}_{c,k,1}, \mathbf{X}\rangle, \ldots, \langle \mathbf{A}_{c,k,\zeta_k}, \mathbf{X}\rangle)$$

with $\mathbf{A}_{c,k,i} \in \mathcal{S}^{(c,k)}$, $i \in [\zeta_k]$, $\mathbf{C}_{c,k} \in \mathcal{S}^{(c,k)}$, $c \in [p]$ and $\mathbf{b}_k \in \mathbb{R}^{\zeta_k}$. See Appendix 5.3.6 for the conversion of SDP (5.3.4) to the form (5.3.8).

The dual of SDP (5.3.8) reads

$$\rho_k^{\rm cs} = \sup_{\mathbf{y}\in\mathbb{R}^{\zeta}} \left\{ \mathbf{b}_k^\top \mathbf{y} \ : \ \mathcal{A}_{c,k}^\top \mathbf{y} - \mathbf{C}_{c,k} \in \mathcal{S}_+^{(c,k)} \, , c \in [p] \right\} , \tag{5.3.9}$$

where $\mathcal{A}_{c,k}^\top : \mathbb{R}^{\zeta} \to \mathcal{S}^{(c,k)}$ is the adjoint operator of $\mathcal{A}_{c,k}$, i.e., $\mathcal{A}_{c,k}^\top \mathbf{z} = \sum_{i\in[\zeta]} z_i \mathbf{A}_{c,k,i}$, $c \in [p]$. By Definition 5.2, it holds that for every $k \in \mathbb{N}^{\geq k_{\min}}$,

$$\left.\begin{array}{l} \forall \ \mathbf{X}_c \in \mathcal{S}^{(c,k)} \, , \, c \in [p] \\ \sum_{c\in[p]} \mathcal{A}_{c,k}\mathbf{X}_c = \mathbf{b}_k \end{array}\right\} \Rightarrow \mathrm{trace}(\mathbf{X}_c) = a_k^{(c)} \, , \, c \in [p] \, . \tag{5.3.10}$$

After replacing $(\mathcal{A}_{c,k}, \mathbf{A}_{c,k,i}, \mathbf{b}_k, \mathbf{C}_{c,k}, \mathcal{S}^{(c,k)}, \zeta_k, \tau_k^{\rm cs}, a_k^{(c)})$ by $(\mathcal{A}_c, \mathbf{A}_{i,c}, \mathbf{b}, \mathbf{C}_c, \mathcal{S}^{(c)}, \zeta, \tau, a_c)$, SDP (5.3.8) then becomes SDP (5.3.28); see Appendix 5.3.4 with $\omega_c = m_c{+}1$ and $s^{\max} = \max_{c\in[p]} b(n_c, k)$.

If there is a ball constraint on each clique of variables then by Corollary 5.3, strong duality holds for the pair (5.3.8)-(5.3.9), for every $k \in \mathbb{N}^{\geq k_{\min}}$.

The two algorithms (CGAL and SM) based on first-order methods are then leveraged to solve the primal-dual (5.3.8)-(5.3.9); see Appendix 5.3.4 and Appendix 5.3.5.

Verifying CTP for POPs with CS via LP

As in the dense case, we can verify CTP for a POP with CS via a series of LPs.

For every $k \in \mathbb{N}^{\geq k_{\min}}$ and for every $c \in [p]$, let $\hat{\mathcal{S}}^{(c,k)}$ be the set of real diagonal matrices of size $b(n_c, k)$ and consider the following LP:

$$\inf_{\xi, \mathbf{G}_i, \mathbf{u}_i} \left\{ \xi \ \left| \ \begin{array}{l} \mathbf{G}_0 - \mathbf{I}_0 \in \hat{\mathcal{S}}_+^{(c,k)} \, , \, \mathbf{G}_i - \mathbf{I}_i \in \hat{\mathcal{S}}_+^{(c,k-\lceil g_i \rceil)} \, , \, i \in J_c \, , \\ \xi = (\mathbf{v}_k^{I_c})^\top \mathbf{G}_0 \mathbf{v}_k^{I_c} + \sum_{i\in J_c} g_i (\mathbf{v}_{k-\lceil g_i \rceil}^{I_c})^\top \mathbf{G}_i \mathbf{v}_{k-\lceil g_i \rceil}^{I_c} \\ \qquad\qquad + \sum_{j\in W_c} h_j (\mathbf{v}_{2(k-\lceil h_j \rceil)}^{I_c})^\top \mathbf{u}_j \end{array} \right. \right\} , \tag{5.3.11}$$

where $\mathbf{I}_i$ is the identity matrix, for every $i \in \{0\} \cup J_c$.

Lemma 5.3. *Let POP* (5.0.2) *with CS be described in Section 5.3.1. If LP* (5.3.11) *has a feasible solution* $(\xi_k^{(c)}, \mathbf{G}_{i,k}^{(c)}, \mathbf{u}_{i,k}^{(c)})$*, for every* $k \in \mathbb{N}^{\geq k_{\min}}$ *and for every* $c \in [p]$*, then POP* (5.0.2) *has CTP with* $\mathbf{P}_k^{(c)} = \mathrm{diag}(\mathbf{G}_{0,k}^{1/2}, (\mathbf{G}_{i,k}^{1/2})_{i\in J_i})$ *and* $a_k^{(c)} = \xi_k^{(c)}$*, for* $k \in \mathbb{N}^{\geq k_{\min}}$ *and for* $c \in [p]$*.*

The proof of Lemma 5.3 is similar to that of Lemma 5.2.

For instance, for POPs with ball or annulus constraints on subsets of each clique of variables, CTP can be verified by LP.

Proposition 5.3. *Let POP* (5.0.2) *with CS be described in Section 5.3.1. Let* $(T_i)_{i\in[r]\cup([m]\setminus[2r])}$ *be as in Assumption 5.1 and further assume that for every* $c \in [p]$*,* $(\cup_{q\in J_c\cap[r]} T_q) \cup (\cup_{q\in J_c\setminus[2r]} T_q) = I_c$*. Then LP* (5.3.11) *has a feasible solution for every* $k \in \mathbb{N}^{\geq k_{\min}}$*, and therefore POP* (5.0.2) *has CTP.*

Proof. To prove that POP (5.0.2) has CTP on each clique of variables, it is sufficient to show that (5.3.11) has a feasible solution, for every $k \in \mathbb{N}^{\geq k_{\min}}$ and for every $c \in [p]$ due to Lemma 5.3.

For every $c \in [p]$, let $\mathbf{u}^{(c)} = (u_i^{(c)})_{i\in I_c} \subseteq \mathbb{N}^{\leq |J_c|}$ be defined by

$$u_i^{(c)} = |\{q \in J_c \cap [r] \ : \ i \in T_q\}| + |\{q \in J_c \setminus [2r] \ : \ i \in T_q\}| \, , \quad i \in I_c \, . \tag{5.3.12}$$

For every $c \in [p]$, one has $u_i^{(c)} \in \mathbb{N}^{\geq 1}$, $i \in I_c$, according to $(\cup_{q\in J_c\cap[r]} T_q) \cup (\cup_{q\in J_c\setminus[2r]} T_q) = I_c$. Moreover,

$$\|\mathbf{u}^{(c)} \circ \mathbf{x}(I_c)\|_2^2 = \sum_{i\in J_c\cap[r]} \|\mathbf{x}(T_i)\|_2^2 + \sum_{i\in J_c\setminus[2r]} \|\mathbf{x}(T_i)\|_2^2\,, \quad \forall c \in [p]\,. \tag{5.3.13}$$

For every $c \in [p]$, with $R^{(c)} := \sum_{i\in J_c\cap[r]}(\underline{R}_i + \overline{R}_i) + \sum_{i\in J_c\setminus[2r]} \overline{R}_i$, by replacing $\mathbf{x}$ (resp. R) by $\mathbf{u}^{(c)} \circ \mathbf{x}(I_c)$ (resp. $R^{(c)}$) in Lemma 5.1, we obtain

$$(R^{(c)} + 1)^k = (1 + \|\mathbf{u}^{(c)} \circ \mathbf{x}(I_c)\|_2^2)^k + \Lambda_{k-1}^{(c)} \sum_{i\in J_c} \delta_i g_i\,,\ \forall c \in [p]\,,\ \forall k \in \mathbb{N}^{\geq k_{\min}}\,, \tag{5.3.14}$$

where $\Lambda_{k-1}^{(c)} := \sum_{r=0}^{k-1}(R^{(c)} + 1)^r(1 + \|\mathbf{u}^{(c)} \circ \mathbf{x}(I_c)\|_2^2)^{k-r-1}$ and

$$\delta_i := \frac{\underline{R}_i}{\overline{R}_i - \underline{R}_i}\,,\ \delta_{i+r} := \frac{\overline{R}_i}{\overline{R}_i - \underline{R}_i}\,,\ i \in J_c \cap [r] \text{ and } \delta_q = 1\,,\ q \in J_c\setminus[2r]. \tag{5.3.15}$$

It is due to the fact that

$$R^{(c)} - \|\mathbf{u}^{(c)} \circ \mathbf{x}(I_c)\|_2 = \sum_{i\in J_c\cap[r]} (\underline{R}_i + \overline{R}_i - \|\mathbf{x}(T_i)\|_2^2) + \sum_{i\in J_c\setminus[2r]} (\overline{R}_i - \|\mathbf{x}(T_i)\|_2^2)\,, \tag{5.3.16}$$

and $\underline{R}_i + \overline{R}_i - \|\mathbf{x}(T_i)\|_2^2 = \delta_i g_i + \delta_{i+r} g_{i+r}$, $i \in J_c \cap [r]$. For every $c \in [p]$, for each $k \in \mathbb{N}^{\geq k_{\min}}$, let $(\theta_{k,\boldsymbol{\alpha}}^{(c)})_{\boldsymbol{\alpha}\in\mathbb{N}_k^{I_c}} \subseteq \mathbb{R}^{>0}$ and $(\eta_{k-1,\boldsymbol{\alpha}}^{(c)})_{\boldsymbol{\alpha}\in\mathbb{N}_{k-1}^{I_c}} \subseteq \mathbb{R}^{>0}$ be such that

$$(1 + \|\mathbf{u}^{(c)} \circ \mathbf{x}(I_c)\|_2^2)^k = \sum_{\boldsymbol{\alpha}\in\mathbb{N}_k^{I_c}} \theta_{k,\boldsymbol{\alpha}}^{(c)} \mathbf{x}^{2\boldsymbol{\alpha}} \quad \text{and} \quad \Lambda_{k-1}^{(c)} = \sum_{\boldsymbol{\alpha}\in\mathbb{N}_{k-1}^{I_c}} \eta_{k-1,\boldsymbol{\alpha}}^{(c)} \mathbf{x}^{2\boldsymbol{\alpha}}\,,$$

and define the diagonal matrices

$$\mathbf{G}_k^{(c,0)} := \operatorname{diag}((\theta_{k,\boldsymbol{\alpha}}^{(c)})_{\boldsymbol{\alpha}\in\mathbb{N}_k^{I_c}}) \quad \text{and} \quad \mathbf{G}_{k-1}^{(c,i)} := \operatorname{diag}((\delta_i \eta_{k-1,\boldsymbol{\alpha}}^{(c)})_{\boldsymbol{\alpha}\in\mathbb{N}_{k-1}^{I_c}})\,,\ i \in J_c\,. \tag{5.3.17}$$

For every $c \in [p]$, (5.3.14) yields that for every $k \in \mathbb{N}^{\geq k_{\min}}$,

$$(R^{(c)} + 1)^k = (\mathbf{v}_k^{I_c})^\top \mathbf{G}_k^{(c,0)} \mathbf{v}_k^{I_c} + \sum_{i\in J_c} g_i (\mathbf{v}_{k-1}^{I_c})^\top \mathbf{G}_{k-1}^{(c,i)} \mathbf{v}_{k-1}^{I_c}\,. \tag{5.3.18}$$

Hence $((R^{(c)} + 1)^k, \mathbf{G}_k^{(c,i)}, \mathbf{0})$ is a feasible solution of (5.3.11), for every $k \in \mathbb{N}^{\geq k_{\min}}$ and for every $c \in [p]$. □

Main algorithm

Algorithm 7 below solves POP (5.0.2) with CS and whose CTP can be verified by LP.

Algorithm 7 Approximating the optimal value of a POP with CS and CTP

Input: POP (5.0.2) with CS and a relaxation order $k \in \mathbb{N}^{\geq k_{\min}}$
Output: The optimal value τ_k^{cs} of SDP (5.3.8)

1: **for** $c \in [p]$ **do**
2: Solve LP (5.3.11) to obtain an optimal solution $(\xi_k^{(c)}, \mathbf{G}_{i,k}^{(c)}, \mathbf{u}_{c,k}^{(c)})$;
3: Let $a_k^{(c)} = \xi_k^{(c)}$ and $\mathbf{P}_k^{(c)} = \operatorname{diag}((\mathbf{G}_{0,k}^{(c)})^{1/2}, \ldots, (\mathbf{G}_{m,k}^{(c)})^{1/2})$;
4: Compute the optimal value τ_k^{cs} of SDP (5.3.8) by running an algorithm based on first-order methods and which exploits CTP.

In Step 4 of Algorithm 7 the two algorithms CGAL (Algorithm 9 in Appendix 5.3.4 or SM (Algorithm 11 in Appendix 5.3.5) are good candidates.

5.3.2 Exploiting CTP for POPs with TS and CS-TS

In this section, we restate the main results of TS in [212] and CS-TSSOS in [213]. Similarly to dense POPs and POPs with CS, one can easily exploit CTP for POPs with TS and CS-TS. The central reason is that the diagonal of each moment/localizing matrix in a given moment relaxation of a dense POP (resp. POP with CS) does not change when TS (resp. CS-TS) is exploited.

Term sparsity (TS)

Fix a relaxation order $k \in \mathbb{N}^{\geq k_{\min}}$ and a sparse order $t \in \mathbb{N}\backslash\{0\}$. We compute as in [212, Section 5] the following block diagonal (up to permutation) $(0,1)$-binary matrices: $\mathbf{G}_{k,t}^{(0)}$ of size $b(k)$; $\mathbf{G}_{k,t}^{(i)}$ of size $b(k-\lceil g_i \rceil)$, $i \in [m]$; $\mathbf{H}_{k,t}^{(j)}$ of size $b(k-\lceil h_j \rceil)$, $j \in [l]$. Then we consider the following sparse moment relaxation of POP (5.0.2):

$$\tau_{k,t}^{\text{ts}} := \inf_{\mathbf{y}\in\mathbb{R}^{b(2k)}} \left\{ L_{\mathbf{y}}(f) \;\middle|\; \begin{array}{l} \mathbf{G}_{k,t}^{(0)} \circ \mathbf{M}_k(\mathbf{y}) \succeq 0\,,\; y_{\mathbf{0}} = 1\,, \\ \mathbf{G}_{k,t}^{(i)} \circ \mathbf{M}_{k-\lceil g_i \rceil}(g_i\,\mathbf{y}) \succeq 0\,,\; i \in [m]\,, \\ \mathbf{H}_{k,t}^{(j)} \circ \mathbf{M}_{k-\lceil h_j \rceil}(h_j\,\mathbf{y}) = 0\,,\; j \in [l] \end{array} \right\}. \tag{5.3.19}$$

One has $\tau_{k,t-1}^{\text{ts}} \leq \tau_{k,t}^{\text{ts}} \leq \tau_k \leq f^\star$, for all (k,t). Moreover, we have the following theorem.

Theorem 5.4. *(Wang et al. [212, Theorem 5.1]) For each $k \in \mathbb{N}^{\geq k_{\min}}$, the sequence $(\tau_{k,t}^{ts})_{t\in\mathbb{N}\backslash\{0\}}$ converges to τ_k (the optimal value of SDP (2.4.5)) in finitely many steps.*

The dual of (5.3.19) reads

$$\rho_{k,t}^{\text{ts}} = \sup_{\xi,\mathbf{Q}_i,\mathbf{U}_i} \left\{ \xi \;\middle|\; \begin{array}{l} \bar{\mathbf{Q}}_i = \mathbf{G}_{k,t}^{(i)} \circ \mathbf{Q}_i \succeq 0\,,\; i \in \{0\} \cup [m]\,, \\ \bar{\mathbf{U}}_i = \mathbf{H}_{k,t}^{(i)} \circ \mathbf{U}_i\,,\; i \in [l]\,, \\ f - \xi = \mathbf{v}_k^\top \bar{\mathbf{Q}}_0 \mathbf{v}_k + \sum_{i\in[m]} g_i \mathbf{v}_{k-\lceil g_i \rceil}^\top \bar{\mathbf{Q}}_i \mathbf{v}_{k-\lceil g_i \rceil} \\ \qquad\qquad + \sum_{j\in[l]} h_j \mathbf{v}_{k-\lceil h_j \rceil}^\top \bar{\mathbf{U}}_j \mathbf{v}_{k-\lceil h_j \rceil} \end{array} \right\}. \tag{5.3.20}$$

Correlative-Term sparsity (CS-TSSOS)

The basic idea of correlative-term sparsity is to exploit term sparsity for each clique. The clique structure of the initial set of variables is derived from correlative sparsity (Section 5.3.1).

Fix a relaxation order $k \in \mathbb{N}^{\geq k_{\min}}$. For every sparse order $t \in \mathbb{N}\backslash\{0\}$ and for every $c \in [p]$, we compute the following block diagonal (up to permutation) $(0,1)$-binary matrices (see [213]): $\mathbf{G}_{k,t,c}^{(0)}$ of size $b(n_c,k)$; $\mathbf{G}_{k,t,c}^{(i)}$ of size $b(n_c,k-\lceil g_i \rceil)$, $i \in J_c$; $\mathbf{H}_{k,t,c}^{(j)}$ of size $b(n_c,k-\lceil h_j \rceil)$, $j \in W_c$. Then let us consider the following CS-TSSOS moment relaxation:

$$\tau_{k,t}^{\text{cs-ts}} := \inf_{\mathbf{y}\in\mathbb{R}^{b(2k)}} \left\{ L_{\mathbf{y}}(f) \;\middle|\; \begin{array}{l} \mathbf{G}_{k,t,c}^{(0)} \circ \mathbf{M}_k(\mathbf{y}, I_c) \succeq 0\,,\; c \in [p]\,,\; y_{\mathbf{0}} = 1\,, \\ \mathbf{G}_{k,t,c}^{(i)} \circ \mathbf{M}_{k-\lceil g_i \rceil}(g_i\,\mathbf{y}, I_c) \succeq 0\,,\; i \in J_c\,,\; c \in [p]\,, \\ \mathbf{H}_{k,t,c}^{(j)} \circ \mathbf{M}_{k-\lceil h_j \rceil}(h_j\,\mathbf{y}, I_c) = 0\,,\; j \in W_c\,,\; c \in [p] \end{array} \right\}. \tag{5.3.21}$$

One has $\tau_{k,t-1}^{\text{cs-ts}} \leq \tau_{k,t}^{\text{cs-ts}} \leq \tau_k^{\text{cs}} \leq \tau_k \leq f^\star$, for all (k,t). Moreover, we have the following theorem.

Theorem 5.5. *(Wang et al. [213]) For each $k \in \mathbb{N}^{\geq k_{\min}}$, the sequence $(\tau_{k,t}^{cs\text{-}ts})_{t\in\mathbb{N}\backslash\{0\}}$ converges to τ_k^{cs} (the optimal value of SDP (5.3.3)) in finitely many steps.*

The dual of (5.3.21) reads

$$\rho_{k,t}^{\text{cs-ts}} = \sup_{\xi,\mathbf{Q}_i^{(c)},\mathbf{U}_i^{(c)}} \left\{ \xi \;\middle|\; \begin{array}{l} \bar{\mathbf{Q}}_i^{(j)} = \mathbf{G}_{k,t,c}^{(i)} \circ \mathbf{Q}_i^{(c)} \succeq 0\,,\; i \in \{0\} \cup J_c\,,\; c \in [p]\,, \\ \bar{\mathbf{U}}_i^{(c)} = \mathbf{H}_{k,t,c}^{(i)} \circ \mathbf{U}_i^{(c)}\,,\; i \in W_c\,,\; c \in [p]\,, \\ f - \xi = \sum_{c\in[p]} \Big((\mathbf{v}_k^{I_c})^\top \bar{\mathbf{Q}}_0^{(c)} \mathbf{v}_k^{I_c} \\ \qquad + \sum_{i\in J_c} g_i (\mathbf{v}_{k-\lceil g_i \rceil}^{I_c})^\top \bar{\mathbf{Q}}_i^{(c)} \mathbf{v}_{k-\lceil g_i \rceil}^{I_c} \\ \qquad + \sum_{j\in W_c} h_j (\mathbf{v}_{k-\lceil h_j \rceil}^{I_c})^\top \bar{\mathbf{U}}_j^{(c)} \mathbf{v}_{k-\lceil h_j \rceil}^{I_c} \Big) \end{array} \right\}. \tag{5.3.22}$$

Table 5.12: Numerical results for minimizing a random quadratic polynomial with TS on the unit ball

- POP size: $m = 1$, $l = 0$; Relaxation order: $k = 2$; Sparse order: $t = 1$; SDP size: $\omega = 4$, $a^{\max} = 3$.

POP size	SDP size		Mosek		CGAL	
n	$s^{\max}$	ζ	val	time	val	time
10	56	937	-1.5681	4	-1.5527	0.7
20	211	13722	-2.4275	36	-2.3996	1
30	466	68357	-3.0748	1930	-3.0577	8
40	821	214842	–	–	-3.6999	20
50	1276	523177	–	–	-4.1603	128
60	1831	1083362	–	–	-4.1914	655
70	2486	2005397	–	–	-4.9578	1461
80	3241	3419282	–	–	-5.6452	7253

Table 5.13: Numerical results for randomly generated QCQPs with TS and a ball constraint

- POP size: $m = 1$, $l = \lceil n/4 \rceil$; Relaxation order: $k = 2$; Sparse order: $t = 1$; SDP size: $\omega = 4$, $a^{\max} = 3$.

POP size		SDP size		Mosek		CGAL	
n	l	$s^{\max}$	ζ	val	time	val	time
10	3	56	1105	-0.60612	0.7	-0.60550	2
20	5	211	14777	-2.3115	47	-2.3097	17
30	8	466	72085	-2.8344	3102	-2.8321	112
40	10	821	223052	–	–	-3.4081	476
50	13	1276	539765	–	–	-3.3552	1845
60	15	1831	1110827	–	–	-3.5620	2992

5.3.3 Numerical experiments for sparse POPs

In this section we report results of numerical experiments for sparse POPs with the same settings and notations as in Section 5.2.

Randomly generated QCQPs with TS and ball constraints

Test problems: We construct randomly generated QCQPs with TS and a ball constraint as follows:

1. Generate a quadratic polynomial objective function f such that for $\boldsymbol{\alpha} \in \mathbb{N}^n_2$ with $|\boldsymbol{\alpha}| \neq 2$, $f_{\boldsymbol{\alpha}} = 0$ and for $\boldsymbol{\alpha} \in \mathbb{N}^n_2$ with $|\boldsymbol{\alpha}| = 2$, the coefficient $f_{\boldsymbol{\alpha}}$ is randomly generated in $(-1, 1)$ w.r.t. the uniform distribution;
2. Take $m = 1$ and $g_1 := 1 - \|\mathbf{x}\|_2^2$;
3. Take a random point $\mathbf{a}$ in $S(\mathfrak{g})$ w.r.t. the uniform distribution;
4. For every $j \in [l]$, generate a quadratic polynomial h_j by
   (i) setting $h_{j,\boldsymbol{\alpha}} = 0$ for each $\boldsymbol{\alpha} \in \mathbb{N}^n_2 \backslash \{\mathbf{0}\}$ with $|\boldsymbol{\alpha}| \neq 2$;
   (ii) for each $\boldsymbol{\alpha} \in \mathbb{N}^n_2 \backslash \{\mathbf{0}\}$ with $|\boldsymbol{\alpha}| = 2$, taking a random coefficient $h_{j,\boldsymbol{\alpha}}$ for h_j in $(-1, 1)$ w.r.t. the uniform distribution;
   (iii) setting $h_{j,\mathbf{0}} := -\sum_{\boldsymbol{\alpha} \in \mathbb{N}^n_2 \backslash \{\mathbf{0}\}} h_{j,\boldsymbol{\alpha}} \mathbf{a}^{\boldsymbol{\alpha}}$.

   Then $\mathbf{a}$ is a feasible solution of POP (5.0.2).

The numerical results are displayed in Table 5.12 and 5.13.

Table 5.14: Numerical results for minimizing a random quadratic polynomial with TS on a box

- POP size: $m = n$, $l = 0$; Relaxation order: $k = 2$; Sparse order: $t = 1$; SDP size: $a^{\max} = 3$.

POP size	SDP size			Mosek		CGAL	
n	ω	$s^{\max}$	ζ	val	time	val	time
10	22	56	1441	-1.0539	3	-1.0519	14
20	42	211	17731	-1.3925	93	-1.3802	161
30	62	466	81871	-2.2301	4392	-2.2128	567
40	82	821	246861	–	–	-2.5209	1602
50	102	1276	585701	–	–	-3.0282	2583
60	122	1831	1191391	–	–	-3.0470	10858

Table 5.15: Numerical results for randomly generated QCQPs with TS and box constraints

- POP size: $m = n$, $l = \lceil n/7 \rceil$; Relaxation order: $k = 2$; Sparse order: $t = 1.$; SDP size: $\omega = n + 1$, $a^{\max} = 3$.

POP size		SDP size			Mosek		CGAL	
n	l	ω	$s^{\max}$	ζ	val	time	val	time
10	2	22	56	1553	-0.77189	0.2	-0.77214	9
20	3	42	211	18364	-1.7962	71	-1.8009	150
30	5	62	466	84201	-1.8529	5814	-1.8625	650
40	6	82	821	251787	–	–	-2.1930	2994
50	8	102	1276	595909	–	–	-2.4655	8397

Discussion: The behavior of solvers is similar to that in the dense case.

Randomly generated QCQPs with TS and box constraints

Test problems: We construct randomly generated QCQPs with TS as in Section 5.3.3, where the ball constraint is now replaced by box constraints. The numerical results are displayed in Table 5.14 and 5.15.

Discussion: Again the behavior of solvers is similar to that in the dense case.

Randomly generated QCQPs with CS and ball constraints on each clique of variables

Test problems: We construct randomly generated QCQPs with CS and ball constraints on each clique of variables as follows:

1. Take a positive integer u, $p := \lfloor n/u \rfloor + 1$ and let

$$I_c = \begin{cases} [u], & \text{if } c = 1\,, \\ \{u(c-1), \ldots, uc\}, & \text{if } c \in \{2, \ldots, p-1\}\,, \\ \{u(p-1), \ldots, n\}, & \text{if } c = p\,; \end{cases} \tag{5.3.23}$$

2. Generate a quadratic polynomial objective function $f = \sum_{c\in[p]} f_c$ such that for each $c \in [p]$, $f_c \in \mathbb{R}[\mathbf{x}(I_c)]_2$, and the coefficient $f_{c,\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \mathbb{N}_2^{I_c}$ of f_c is randomly generated in $(-1, 1)$ w.r.t. the uniform distribution;

3. Take $m = p$ and $g_i := -\|\mathbf{x}(I_i)\|_2^2 + 1$, $i \in [m]$;

4. Take a random point $\mathbf{a}$ in $S(\mathfrak{g})$ w.r.t. the uniform distribution;

Table 5.16: Numerical results for minimizing a random quadratic polynomial with CS and ball constraints on each clique of variables

- POP size: $n = 1000$, $m = p$, $l = 0$, $u^{\max} = u + 1$; Relaxation order: $k = 2$; SDP size: $\omega = 2p$, $a^{\max} = 3$.

POP size		SDP size			Mosek		CGAL	
u	p	ω	$s^{\max}$	ζ	val	time	val	time
11	91	182	91	222712	-240.54	124	-240.37	98
16	63	126	171	550692	-205.45	1389	-205.19	280
21	48	96	276	1107682	–	–	-175.60	321
26	39	78	406	1955879	–	–	-165.65	559
31	33	66	561	3167072	–	–	-149.10	973
36	28	56	741	4758727	–	–	-140.21	1315
41	25	50	946	6839993	–	–	-126.55	1926

Table 5.17: Numerical results for randomly generated QCQPs with CS and ball constraints on each clique of variables

- POP size: $n = 1000$, $m = p$, $l = 143$, $u^{\max} = u + 1$; Relaxation order: $k = 2$; SDP size: $\omega = 2p$, $a^{\max} = 3$.

POP size		SDP size			Mosek		CGAL	
u	p	ω	$s^{\max}$	ζ	val	time	val	time
11	91	182	91	235023	-224.15	163	-224.09	204
16	63	126	171	572905	-192.45	1830	-192.30	335
21	48	96	276	1139460	–	–	-162.79	537
26	39	78	406	2005124	–	–	-148.77	1014
31	33	66	561	3239573	–	–	-142.38	2115
36	28	56	741	4862292	–	–	-124.97	5304

5. Let $r := \lfloor l/p \rfloor$ and

$$W_c := \begin{cases} \{(c-1)r+1, \ldots, cr\}, & \text{if } c \in [p-1]\,, \\ \{(p-1)r+1, \ldots, l\}, & \text{if } c = p\,. \end{cases} \tag{5.3.24}$$

For every $c \in [p]$ and every $i \in W_c$, generate a quadratic polynomial $h_i \in \mathbb{R}[\mathbf{x}(I_c)]_2$ by

(a) for each $\boldsymbol{\alpha} \in \mathbb{N}_2^{I_c} \backslash \{\mathbf{0}\}$, taking a random coefficient $h_{i,\boldsymbol{\alpha}}$ of h_i in $(-1, 1)$ w.r.t. the uniform distribution;

(b) setting $h_{i,\mathbf{0}} := -\sum_{\boldsymbol{\alpha} \in \mathbb{N}_2^{I_c} \backslash \{\mathbf{0}\}} h_{c,\boldsymbol{\alpha}} \mathbf{a}^{\boldsymbol{\alpha}}$.

Then $\mathbf{a}$ is a feasible solution of POP (5.0.2).

The numerical results are displayed in Table 5.16 and 5.17.

Discussion: The number of variables is fixed as $n = 1000$. We increase the clique size u so that the number of variable cliques p decreases accordingly. Again results in Table 5.16 and 5.17 show that CGAL is faster and returns an optimal value of gap within 1% w.r.t. the one returned by `Mosek` (for $u \leq 16$). Moreover `Mosek` runs out of memory when $u \geq 21$.

Randomly generated QCQPs with CS and box constraints on each clique of variables

Test problems: We construct randomly generated QCQPs with CS as in Section 5.3.3, where ball constraints are now replaced by box constraints. Namely, in Step 3 we take $m = n$, $g_j := -x_j^2 + 1/u$, $j \in [n]$.

The numerical results are displayed in Table 5.18 and 5.19.

Table 5.18: Numerical results for minimizing a random quadratic polynomial with CS and box constraints on each clique of variables

- POP size: $n = m = 1000$, $l = 0$, $u^{\max} = u+1$; Relaxation order: $k = 2$; Constant trace: $a^{\max} \in [3, 4]$.

POP size		SDP size			Mosek		CGAL	
u	p	ω	$s^{\max}$	ζ	val	time	val	time
11	91	1181	91	313361	-204.89	443	-204.69	753
16	63	1125	171	720323	-163.11	3082	-162.88	3059
21	48	1095	276	1380918	–	–	-147.92	5655
26	39	1077	406	2357161	–	–	-131.00	8889

Table 5.19: Numerical results for QCQPs with CS and box constraints on each clique of variables

- POP size: $n = m = 1000$, $l = 143$, $u^{\max} = u + 1$; Relaxation order: $k = 2$; Constant trace: $a^{\max} \in [3, 4]$.

POP size		SDP size			Mosek		CGAL	
u	p	ω	$s^{\max}$	ζ	val	time	val	time
11	91	1181	91	325672	-187.01	402	-186.98	1915
16	63	1125	171	742536	-142.16	4323	-142.27	4126
21	48	1095	276	1412696	–	–	-131.14	5334
26	39	1077	406	2406406	–	–	-113.44	8037

Discussion: The number of variables is fixed as $n = 1000$. We increase the clique size u so that the number of variable cliques p decreases accordingly. From results in Table 5.16 and 5.17, one observes that when the largest size of variable cliques is relatively small (say $u \leq 11$), `Mosek` is the fastest solver. However when the largest size of variable cliques is relatively large (say $u \geq 21$), `Mosek` runs out of memory while CGAL still works well.

Randomly generated QCQPs with CS-TSSOS and ball constraints on each clique of variables

Test problems: We construct randomly generated QCQPs with CS-TSSOS and ball constraints on each clique of variables as follows:

1. Take a positive integer u, $p := \lfloor n/u \rfloor + 1$ and let $(I_c)_{c\in[p]}$ be defined as in (5.3.23);
2. Generate a quadratic polynomial objective function $f = \sum_{c\in[p]} f_c$ such that for each $c \in [p]$, $f_c \in \mathbb{R}[\mathbf{x}(I_c)]_2$ and the nonzero coefficient $f_{c,\boldsymbol{\alpha}}$ with $\boldsymbol{\alpha} \in \mathbb{N}_2^{I_c}$ and $|\boldsymbol{\alpha}| = 2$ is randomly generated in $(-1, 1)$ w.r.t. the uniform distribution;
3. Take $m = p$ and $g_i := -\|\mathbf{x}(I_i)\|_2^2 + 1$, $i \in [m]$;
4. Take a random point $\mathbf{a}$ in $S(\mathfrak{g})$ w.r.t. the uniform distribution;
5. Let $r := \lfloor l/p \rfloor$ and $(W_c)_{c\in[p]}$ be as in (5.3.24). For every $c \in [p]$ and every $i \in W_c$, generate a quadratic polynomial $h_i \in \mathbb{R}[\mathbf{x}(I_c)]_2$ by
   (a) for each $\boldsymbol{\alpha} \in \mathbb{N}_2^{I_c}\backslash\{\mathbf{0}\}$ with $|\boldsymbol{\alpha}| \neq 2$, taking $h_{i,\boldsymbol{\alpha}} = 0$;
   (b) for each $\boldsymbol{\alpha} \in \mathbb{N}_2^{I_c}$ with $|\boldsymbol{\alpha}| = 2$, taking a random coefficient $h_{i,\boldsymbol{\alpha}}$ of h_i in $(-1, 1)$ w.r.t. the uniform distribution;
   (c) setting $h_{i,\mathbf{0}} := -\sum_{\boldsymbol{\alpha}\in\mathbb{N}_2^{I_c}\backslash\{\mathbf{0}\}} h_{c,\boldsymbol{\alpha}}\mathbf{a}^{\boldsymbol{\alpha}}$.

   Then $\mathbf{a}$ is a feasible solution of POP (5.0.2).

The numerical results are displayed in Table 5.20 and 5.21.

Table 5.20: Numerical results for minimizing a random quadratic polynomial with CS-TSSOS and ball constraints on each clique of variables

- POP size: $n = 1000$, $m = p$, $l = 0$, $u^{\max} = u + 1$; Relaxation order: $k = 2$; Sparse order: $t = 1$; SDP size: $a^{\max} = 3$.

POP size		SDP size			Mosek		CGAL	
u	p	ω	$s^{\max}$	ζ	val	time	val	time
11	91	364	79	169654	-160.05	163	-160.01	498
16	63	252	154	448354	-135.78	1422	-135.74	768
21	48	192	254	939619	–	–	-117.17	1605
26	39	156	379	1705763	–	–	-106.26	3150

Table 5.21: Numerical results for QCQPs with CS-TSSOS and ball constraints on each clique of variables

- POP size: $n = 1000$, $m = p$, $l = 143$, $u^{\max} = u + 1$; Relaxation order: $k = 2$; Sparse order: $t = 1$; SDP size: $a^{\max} = 3$.

POP size		SDP size			Mosek		CGAL	
u	p	ω	$s^{\max}$	ζ	val	time	val	time
11	91	364	79	180303	-155.91	158	-155.87	604
16	63	252	154	468290	127.42	1707	-127.36	1053
21	48	192	254	939619	–	–	-114.85	2877
26	39	156	379	1751556	–	–	-102.30	6878

Discussion: The behavior of solvers is similar to that in Section 5.3.3. Here, we also emphasize that our framework is less efficient than interior-point methods for most benchmarks presented in [213]. The two underlying reasons are that (1) the block size of the resulting SDP relaxations is small, in which case `Mosek` performs more efficiently, e.g., for the benchmarks from [213, Section 5.2], and (2) it is harder to find the constant trace, e.g., for the benchmarks from [213, Section 5.4]. Thus our proposed method complements that in [213] when the block size of the SDP relaxations is large and/or when CTP can be efficiently verified.

Randomly generated QCQPs with CS-TSSOS and box constraints on each clique of variables

Test problems: We construct randomly generated QCQPs with CS-TSSOS as in Section 5.3.3, where ball constraints are now replaced by box constraints. Namely, in Step 3 we take $m = n$, $g_j := -x_j^2 + 1/u$, $j \in [n]$. The numerical results are displayed in Table 5.22 and 5.23.

Discussion: The behavior of solvers is similar to that in Section 5.3.3.

Table 5.22: Numerical results for minimizing a random quadratic polynomial with CS-TSSOS and box constraints on each clique of variables

- POP size: $n = m = 1000$, $l = 0$, $u^{\max} = u + 1$; Relaxation order: $k = 2$; Sparse order: $t = 1$; Constant trace: $a^{\max} \in [3, 4]$.

POP size		SDP size			Mosek		CGAL	
u	p	ω	$s^{\max}$	ζ	val	time	val	time
11	91	2362	79	248335	-126.15	151	-126.04	1982
16	63	2250	154	601081	-100.75	2225	-100.64	7323
21	48	2190	254	1191001	–	–	-87.804	10734
26	39	2154	379	2080265	–	–	-81.908	20294

Table 5.23: Numerical results for QCQPs with CS-TSSOS and box constraints on each clique of variables

- POP size: $n = m = 1000$, $l = 143$, $u^{\max} = u + 1$; Relaxation order: $k = 2$; Sparse order: $t = 1$; Constant trace: $a^{\max} \in [3, 4]$.

POP size		SDP size			Mosek		CGAL	
u	p	ω	$s^{\max}$	ζ	val	time	val	time
11	91	2362	79	258984	-114.53	325	-114.27	482
16	63	2250	154	621017	-96.199	4450	-96.079	1245
21	48	2190	254	1220027	–	–	-83.013	8204
26	39	2154	379	2126058	–	–	-74.532	27600

5.3.4 Conditional gradient-based augmented Lagrangian

SDP with CTP

Let $s, l, s^{(j)} \in \mathbb{N}^{\geq 1}$, $j \in [\omega]$, be fixed such that $s = \sum_{j=1}^{\omega} s^{(j)}$. Let $\mathcal{S}$ be the set of real symmetric matrices of size s in a block diagonal form: $\mathbf{X} = \operatorname{diag}(\mathbf{X}_1, \ldots, \mathbf{X}_\omega)$, such that $\mathbf{X}_j$ is a block of size $s^{(j)}$, $j \in [\omega]$. Let $s^{\max} := \max_{j\in[\omega]} s^{(j)}$. Let $\mathcal{S}^+$ be the set of all $\mathbf{X} \in \mathcal{S}$ such that $\mathbf{X} \succeq 0$, i.e., $\mathbf{X}$ has only nonnegative eigenvalues. Then $\mathcal{S}$ is a Hilbert space with scalar product $\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{trace}(\mathbf{B}^\top \mathbf{A})$ and $\mathcal{S}_+$ is a self-dual cone.

Let us consider the following SDP:

$$\tau = \inf_{\mathbf{X}\in\mathcal{S}_+} \{ \langle \mathbf{C}, \mathbf{X} \rangle \, : \, \mathcal{A}\mathbf{X} = \mathbf{b} \}, \tag{5.3.25}$$

where $\mathcal{A} : \mathcal{S} \to \mathbb{R}^\zeta$ is a linear operator of the form $\mathcal{A}\mathbf{X} = [\langle \mathbf{A}_1, \mathbf{X} \rangle, \ldots, \langle \mathbf{A}_\zeta, \mathbf{X} \rangle]$, with $\mathbf{A}_i \in \mathcal{S}$, $i \in [\zeta]$, $\mathbf{C} \in \mathcal{S}$ is the cost matrix and $\mathbf{b} \in \mathbb{R}^\zeta$ is a vector.

The dual of SDP (5.3.25) reads

$$\rho = \sup_{\mathbf{y}\in\mathbb{R}^\zeta} \{ \mathbf{b}^\top \mathbf{y} \, : \, \mathcal{A}^\top \mathbf{y} - \mathbf{C} \in \mathcal{S}_+ \}, \tag{5.3.26}$$

where $\mathcal{A}^\top : \mathbb{R}^\zeta \to \mathcal{S}$ is the adjoint operator of $\mathcal{A}$, i.e., $\mathcal{A}^\top \mathbf{y} = \sum_{i\in[\zeta]} y_i \mathbf{A}_i$.

The following assumption will be used later on.

Assumption 5.3. *Consider the following conditions:*

1. *Strong duality of primal-dual* (5.3.25)-(5.3.26) *holds, i.e., $\rho = \tau$ and $\rho \in \mathbb{R}$.*
2. *Constant trace property (CTP): $\exists a > 0 \; : \; \forall \, \mathbf{X} \in \mathcal{S}$, $\mathcal{A}\mathbf{X} = \mathbf{b} \Rightarrow \operatorname{trace}(\mathbf{X}) = a$.*

For $\mathbf{X} \in \mathcal{S}$, the Frobenius norm of $\mathbf{X}$ is defined by $\|\mathbf{X}\|_F := \sqrt{\langle \mathbf{X}, \mathbf{X} \rangle}$. We denote by $\|\mathcal{A}\|$ the operator norm of $\mathcal{A}$, i.e., $\|\mathcal{A}\| := \max_{\mathbf{X}\in\mathcal{S}} \|\mathcal{A}\mathbf{X}\|_2 / \|\mathbf{X}\|_F$. The smallest eigenvalue of a real symmetric matrix $\mathbf{D}$ is denoted by $\lambda_{\min}(\mathbf{D})$.

Algorithm. In [219], Yurtsever et al. stated Algorithm 8 (see below) to solve SDP (5.3.25) with CTP. This procedure is based on the augmented Lagrangian paradigm combined together with the conditional gradient method.

The convergence of the sequence $(\mathbf{X}_t)_{t\in\mathbb{N}}$ in Algorithm 8 to the set of optimal solutions of SDP (5.3.25) is guaranteed as follows:

Theorem 5.6. *[219, Fact 3.1] Consider SDP* (5.3.25) *such that Assumption 5.3 holds. Let $(\mathbf{X}_t)_{t\in\mathbb{N}}$ be in the output of Algorithm 8. Then $\mathbf{X}_t \succeq 0$, for all $t \in \mathbb{N}$ and $\|\mathcal{A}\mathbf{X}_t - \mathbf{b}\|_2 \to 0$, $|\langle \mathbf{C}, \mathbf{X}_t \rangle - \tau| \to 0$ as $t \to \infty$, with the rate of order $\mathcal{O}(1/\sqrt{t})$.*

Remark 5.4. *In order to achieve the best convergence rate for Algorithm 8, we scale the problem's input as follows: $\|\mathbf{C}\|_F = \|\mathcal{A}\| = a = 1$ and $\|\mathbf{A}_1\|_F = \cdots = \|\mathbf{A}_\zeta\|_F$.*

Algorithm 8 CGAL-SDP-CTP

Input: SDP (5.3.25) such that Assumption 5.3 holds; Parameter $K > 0$.
Output: $(\mathbf{X}_t)_{t\in\mathbb{N}}$.

1: Set $\mathbf{X}_0 := \mathbf{0}_{\mathcal{S}}$ and $\mathbf{y}_0 := \mathbf{0}_{\mathbb{R}^\zeta}$.
2: **for** $t \in \mathbb{N}$ **do**
3: Set $\beta_t := \sqrt{t+1}$ and $\eta_t := 2/(t+1)$;
4: Take an eigenvector $\mathbf{u}_t$ corresponding to $\lambda_{\min}(\mathbf{C} + \mathcal{A}^\top(\mathbf{y}_{t-1} + \eta_t(\mathcal{A}\mathbf{X}_{t-1} - \mathbf{b})))$;
5: Set $\mathbf{X}_t := (1-\eta_t)\mathbf{X}_{t-1} + \eta_t a \mathbf{u}_t\mathbf{u}_t^\top$;
6: Select γ_t as the largest $\gamma \in [0,1]$ such that:
7: $\gamma\|\mathcal{A}\mathbf{X}_t - \mathbf{b}\|_2^2 \le \beta_t\eta_t^2 a^2\|\mathcal{A}\|^2$ and $\|\mathbf{y}_{t-1} + \gamma(\mathcal{A}\mathbf{X}_t - \mathbf{b})\|_2 \le K$;
8: Set $\mathbf{y}_t = \mathbf{y}_{t-1} + \gamma_t(\mathcal{A}\mathbf{X}_t - \mathbf{b})$.

Remark 5.5. *Given $\varepsilon > 0$, the for loop in Algorithm 8 terminates when:*

$$\frac{|\langle \mathbf{C}, \mathbf{X}_{t-1}\rangle - (a\lambda_{\min}(\mathbf{C} + \mathcal{A}^\top(\mathbf{y}_{t-1} + \eta_t(\mathcal{A}\mathbf{X}_{t-1} - \mathbf{b}))) - \mathbf{b}^\top\mathbf{y}_{t-1})|}{1 + \max\{|\langle \mathbf{C}, \mathbf{X}_{t-1}\rangle|, |a\lambda_{\min}(\mathbf{C} + \mathcal{A}^\top(\mathbf{y}_{t-1} + \eta_t(\mathcal{A}\mathbf{X}_{t-1} - \mathbf{b}))) - \mathbf{b}^\top\mathbf{y}_{t-1}|\}} \le \varepsilon \tag{5.3.27}$$

and $\|\mathcal{A}\mathbf{X}_{t-1} - \mathbf{b}\|_2 / \max\{1, \|b\|_2\} \le \varepsilon$. In our experiments, we choose $\varepsilon = 10^{-3}$. Note that the left hand side in (5.3.27) *is the relative gap between the primal and dual approximate values obtained at each iteration.*

Remark 5.6. *To save memory at each iteration, we can run Algorithm 8 with an implicit $\mathbf{X}_t$ by setting $\mathbf{w}_t := \mathcal{A}\mathbf{X}_t - \mathbf{b}$. In this case, Step 5 becomes $\mathbf{w}_t := (1-\eta_t)\mathbf{w}_{t-1} + \eta_t[\mathcal{A}(a\mathbf{u}_t\mathbf{u}_t^\top) - b]$. Thus we only obtain an approximate dual solution $\mathbf{y}_t$ of SDP* (5.3.25) *when Algorithm 8 terminates.*

In Appendix 5.3.4, we provide an analogous method to solve an SDP with CTP on each subset of blocks.

SDP with CTP on each subset of blocks

Let $p \in \mathbb{N}^{\ge 1}$, $s_c, \omega_c \in \mathbb{N}$, $c \in [p]$, and $s^{(i,c)} \in \mathbb{N}^{\ge 1}$, $i \in [\omega_p]$, $c \in [p]$, be fixed such that $s_c = \sum_{i\in[\omega_c]} s^{(i,c)}$, $c \in [p]$. For every $c \in [p]$, let $\mathcal{S}^{(c)}$ be the set of real symmetric matrices of size s_c in a block diagonal form: $\mathbf{X}_c = \mathrm{diag}(\mathbf{X}_{1,c}, \ldots, \mathbf{X}_{\omega_c,c})$, such that $\mathbf{X}_{i,c}$ is a block of size $s^{(i,c)}$, $i \in [\omega_c]$. Let $s^{\max} := \max_{i\in[\omega_p], c\in[p]} s^{(i,c)}$. For every $c \in [p]$, let $\mathcal{S}_+^{(c)}$ be the set of all $\mathbf{X}_c \in \mathcal{S}^{(c)}$ such that $\mathbf{X}_c \succeq 0$. Then for every $c \in [p]$, $\mathcal{S}^{(c)}$ is a Hilbert space with scalar product $\langle \mathbf{A}, \mathbf{B}\rangle = \mathrm{trace}(\mathbf{B}^\top\mathbf{A})$ and $\mathcal{S}_+^{(c)}$ is a self-dual cone.

Let us consider the following SDP:

$$\tau = \inf_{\mathbf{X}_c \in \mathcal{S}_+^{(c)}} \left\{ \sum_{c\in[p]} \langle \mathbf{C}_c, \mathbf{X}_c\rangle \,:\, \sum_{c\in[p]} \mathcal{A}_c\mathbf{X}_c = \mathbf{b} \right\}, \tag{5.3.28}$$

where $\mathcal{A}_c : \mathcal{S}^{(c)} \to \mathbb{R}^\zeta$ is a linear operator of the form $\mathcal{A}_c\mathbf{X} = [\langle \mathbf{A}_{1,c}, \mathbf{X}\rangle, \ldots, \langle \mathbf{A}_{\zeta,c}, \mathbf{X}\rangle]$, with $\mathbf{A}_{i,c} \in \mathcal{S}^{(c)}$, $i \in [\zeta]$, $\mathbf{C}_c \in \mathcal{S}^{(c)}$, $c \in [p]$, and $\mathbf{b} \in \mathbb{R}^\zeta$.

The dual of SDP (5.3.28) reads

$$\rho = \sup_{\mathbf{y}\in\mathbb{R}^\zeta} \left\{ \mathbf{b}^\top\mathbf{y} \,:\, \mathcal{A}_c^\top\mathbf{y} - \mathbf{C}_c \in \mathcal{S}_+^{(c)}, c \in [p] \right\}, \tag{5.3.29}$$

where $\mathcal{A}_c^\top : \mathbb{R}^\zeta \to \mathcal{S}^{(c)}$ is the adjoint operator of $\mathcal{A}_c$, i.e., $\mathcal{A}_c^\top\mathbf{z} = \sum_{i\in[\zeta]} z_i\mathbf{A}_{i,c}$, $c \in [p]$.

The following assumption will be used later on:

Assumption 5.4. *Consider the following conditions:*

1. *Strong duality of primal-dual* (5.3.28)-(5.3.29) *holds, i.e., $\rho = \tau$ and $\rho \in \mathbb{R}$.*
2. *Constant trace property (CTP): there exist $a_c > 0$ and $c \in [p]$, such that*

$$\left.\begin{array}{l} \forall\ \mathbf{X}_c \in \mathcal{S}^{(c)}, c \in [p], \\ \sum_{c\in[p]} \mathcal{A}_c\mathbf{X}_c = \mathbf{b} \end{array}\right\} \Rightarrow \mathrm{trace}(\mathbf{X}_c) = a_c, c \in [p]. \tag{5.3.30}$$

Recall that $\lambda_{\min}(\mathbf{D})$ stands for the smallest eigenvalue of a real symmetric matrix $\mathbf{D}$. We denote by $\prod_{c\in[p]}\mathcal{S}^{(c)}$ the set of all $\mathbf{X}=\mathrm{diag}(\mathbf{X}_c)_{c\in[p]}$ such that $\mathbf{X}_c\in\mathcal{S}^{(c)}$, for $c\in[p]$. Let $\mathbf{C}:=\mathrm{diag}(\mathbf{C}_c)_{c\in[p]}$ and let $\mathcal{A}:\prod_{c\in[p]}\mathcal{S}^{(c)}\to\mathbb{R}^\zeta$ be a linear operator of the form: $\mathcal{A}\mathbf{X}=\sum_{c\in[p]}\mathcal{A}_c\mathbf{X}_c$, for all $\mathbf{X}=\mathrm{diag}(\mathbf{X}_c)_{c\in[p]}\in\prod_{c\in[p]}\mathcal{S}^{(c)}$. Then for every $\mathbf{X}=\mathrm{diag}(\mathbf{X}_c)_{c\in[p]}\in\prod_{c\in[p]}\mathcal{S}^{(c)}$, we have $\langle\mathbf{C},\mathbf{X}\rangle=\sum_{c\in[p]}\langle\mathbf{C}_c,\mathbf{X}_c\rangle$ and $\mathcal{A}\mathbf{X}=[\langle\mathbf{A}^{(1)},\mathbf{X}\rangle,\ldots,\langle\mathbf{A}^{(\zeta)},\mathbf{X}\rangle]$, where $\mathbf{A}^{(i)}:=\mathrm{diag}((\mathbf{A}_{i,c})_{c\in[p]})$, for $i\in[\zeta]$.

SDP (5.3.28) can be rewritten as $\tau=\inf_{\mathbf{X}\in\prod_{c\in[p]}\mathcal{S}_+^{(c)}}\{\langle\mathbf{C},\mathbf{X}\rangle:\mathcal{A}\mathbf{X}=\mathbf{b}\}$.

The dual operator $\mathcal{A}^\top:\mathbb{R}^\zeta\to\prod_{c\in[p]}\mathcal{S}^{(c)}$ of $\mathcal{A}$ reads $\mathcal{A}^\top\mathbf{z}=\mathrm{diag}((\mathcal{A}_c^\top\mathbf{z})_{c\in[p]})$. Note $\Delta_c:=\{\mathbf{X}_c\in\mathcal{S}_+^{(c)}:\mathrm{trace}(\mathbf{X}_c)=a_c\}$, for $c\in[p]$.

Algorithm. In order to solve SDP (5.3.28) with CTP on each subset of blocks, we use [218, Algorithm 1] due to Yurtsever et al. to describe Algorithm 9 with the following setting: $\mathcal{X}\leftarrow\Delta:=\prod_{c\in[p]}\Delta_c$, $\mathcal{K}\leftarrow\{\mathbf{b}\}$, $p\leftarrow\zeta$, $Ax\leftarrow\mathcal{A}\mathbf{X}$, $f(x)\leftarrow\langle\mathbf{C},\mathbf{X}\rangle$, $\lambda_0\leftarrow1$, $\lambda_k\leftarrow\beta_k$, $\sigma_k\leftarrow\gamma_k$. $D_{\mathcal{Y}_{k+1}}\leftarrow K$, $L_f\leftarrow0$, $\bar{r}_{k+1}\leftarrow\mathbf{b}$, $D_{\mathcal{X}}^2\leftarrow2\sum_{c\in[p]}a_c^2$, $v_k\leftarrow\mathbf{C}+\mathcal{A}^\top\mathbf{z}_k$, $\arg\min_{x\in\mathcal{X}}\langle v_k,x\rangle\leftarrow\arg\min_{X\in\Delta}\langle\mathbf{C}+\mathcal{A}^\top\mathbf{z}_k,\mathbf{X}\rangle$.

With fixed $\mathbf{z}_k$, we have:

$$\begin{aligned}\min_{\mathbf{X}\in\Delta}\langle\mathbf{C}+\mathcal{A}^\top\mathbf{z}_k,\mathbf{X}\rangle &= \min_{\mathrm{diag}((\mathbf{X}_c)_{c\in[p]})\in\prod_{c\in[p]}\Delta_c}\sum_{c\in[p]}\langle\mathbf{C}_c+\mathcal{A}_c^\top\mathbf{z}_k,\mathbf{X}_c\rangle\\ &=\sum_{c\in[p]}\min_{\mathbf{X}_c\in\Delta_c}\langle\mathbf{C}_c+\mathcal{A}_c^\top\mathbf{z}_k,\mathbf{X}_c\rangle=\sum_{c\in[p]}a_c\lambda_{\min}(\mathbf{C}_c+\mathcal{A}_c^\top\mathbf{z}_k)\,.\end{aligned}$$

Let $\mathbf{u}_k^{(c)}$ be a uniform eigenvector corresponding to $\lambda_{\min}(\mathbf{C}_c+\mathcal{A}_c^\top\mathbf{z}_k)$, for $c\in[p]$. Then one has $\mathrm{diag}((a_c\mathbf{u}_k^{(c)}(\mathbf{u}_k^{(c)})^\top)_{c\in[p]})\in\arg\min_{\mathbf{X}\in\Delta}\langle\mathbf{C}+\mathcal{A}^\top\mathbf{z}_k,\mathbf{X}\rangle$. Thus we can set $s_k\leftarrow\mathrm{diag}((a_c\mathbf{u}_k^{(c)}(\mathbf{u}_k^{(c)})^\top)_{c\in[p]})$ in [218, Algorithm 1].

Algorithm 9 CGAL-SDP-CTP-Blocks

Input: SDP (5.3.28) such that Assumption 5.4 holds; Parameter $K>0$.

Output: $((\mathbf{X}_c^{(t)})_{c\in[p]})_{t\in\mathbb{N}}$.

Set $(\mathbf{X}_c^{(0)})_{c\in[p]}:=(\mathbf{0}_{\mathcal{S}})_{c\in[p]}$ and $\mathbf{y}_0:=\mathbf{0}_{\mathbb{R}^\zeta}$.
for $t\in\mathbb{N}$ **do**
  Set $\beta_t:=\sqrt{t+1}$ and $\eta_t:=2/(t+1)$;
  Set $\mathbf{z}_t:=\mathbf{y}_{t-1}+\eta_t(\sum_{c\in[p]}\mathcal{A}_c\mathbf{X}_c^{(t-1)}-\mathbf{b})$;
  for $c\in[p]$ **do**
    Take a uniform eigenvector $\mathbf{u}_t^{(c)}$ corresponding to $\lambda_{\min}(\mathbf{C}_c+\mathcal{A}_c^\top\mathbf{z}_t)$;
    Set $\mathbf{X}_c^{(t)}:=(1-\eta_t)\mathbf{X}_c^{(t-1)}+\eta_t a_c\mathbf{u}_t^{(c)}(\mathbf{u}_t^{(c)})^\top$;
  Select γ_t as the larges $\gamma\in[0,1]$ such that:
    $\gamma\|\sum_{c\in[p]}\mathcal{A}_c\mathbf{X}_c^{(t)}-\mathbf{b}\|_2^2\le\beta_t\eta_t^2(\sum_{c\in[p]}a_c^2)\|\mathcal{A}\|^2$ and $\|\mathbf{y}_{t-1}+\gamma(\sum_{c\in[p]}\mathcal{A}_c\mathbf{X}_c^{(t)}-\mathbf{b})\|_2\le K$;
  Set $\mathbf{y}_t=\mathbf{y}_{t-1}+\gamma_t(\sum_{c\in[p]}\mathcal{A}_c\mathbf{X}_c^{(t)}-\mathbf{b})$.

Relying on [218, Theorem 3.1], we guarantee the convergence of the sequence $((\mathbf{X}_c^{(t)})_{c\in[p]})_{t\in\mathbb{N}}$ in Algorithm 9 to the set of optimal solutions of SDP (5.3.28) in the following theorem:

Theorem 5.7. *Consider SDP* (5.3.28) *such that Assumption 5.4 holds. Let* $((\mathbf{X}_c^{(t)})_{c\in[p]})_{t\in\mathbb{N}}$ *be the output of Algorithm 9. Then* $\mathbf{X}_c^{(t)}\succeq0$, *for all* $c\in[p]$ *and for all* $t\in\mathbb{N}$ *and* $\left\|\sum_{c\in[p]}\mathcal{A}_c\mathbf{X}_c^{(t)}-\mathbf{b}\right\|_2\to0$ *and* $\left|\sum_{c\in[p]}\langle\mathbf{C}_c,\mathbf{X}_c^{(t)}\rangle-\tau\right|\to0$ *as* $t\to\infty$ *with the rate* $\mathcal{O}(1/\sqrt{t})$.

Remark 5.7. *Before running Algorithm 9, we scale the problem's input as follows:* $\|\mathbf{C}\|_F=\|\mathcal{A}\|=a_1=\cdots=a_p=1$ *and* $\|\mathbf{A}^{(1)}\|_F=\cdots=\|\mathbf{A}^{(\zeta)}\|_F$.

Remark 5.8. *Given $\varepsilon > 0$, the for loop in Algorithm 9 terminates when:*

$$\frac{|\sum_{c\in[p]}\langle \mathbf{C}_c, \mathbf{X}_c^{(t-1)}\rangle - \sum_{c\in[p]}(a_c\lambda_{\min}(\mathbf{C}_c+\mathcal{A}_c^\top \mathbf{z}_t) - \mathbf{b}^\top \mathbf{y}_{t-1})|}{1+\max\{|\sum_{c\in[p]}\langle \mathbf{C}_c, \mathbf{X}_c^{(t-1)}\rangle|, |\sum_{c\in[p]}(a_j\lambda_{\min}(\mathbf{C}_c+\mathcal{A}_c^\top \mathbf{z}_t) - \mathbf{b}^\top \mathbf{y}_{t-1})|\}} \le \varepsilon$$

and $\|\sum_{c\in[p]}\mathcal{A}_c\mathbf{X}_c^{(t-1)} - \mathbf{b}\|_2/\max\{1,\|b\|_2\} \le \varepsilon$. *In our experiments, we choose* $\varepsilon = 10^{-2}$.

Remark 5.9. *To save memory at each iteration, we can run Algorithm 9 with implicit* $\mathbf{X}_c^{(t)}$, $c \in [p]$, *by setting* $\mathbf{w}_t := \sum_{c\in[p]}\mathcal{A}_c\mathbf{X}_c^{(t)} - \mathbf{b}$. *In this case, Step 7 becomes* $\mathbf{w}_t := (1-\eta_t)\mathbf{w}_{t-1} + \eta_t[\sum_{c\in[p]}\mathcal{A}_c(a_c\mathbf{u}_t^{(c)}(\mathbf{u}_t^{(c)})^\top) - \mathbf{b}]$. *Thus we only obtain an approximate dual solution* $\mathbf{y}_t$ *of SDP* (5.3.28) *when Algorithm 9 terminates.*

5.3.5 Spectral method

SDP with CTP

Consider SDP with CTP described in Appendix 5.3.4. The following assumption will be used later on.

Assumption 5.5. *Dual attainability: SDP* (5.3.26) *has an optimal solution.*

Lemma 5.4. *Let Assumption 5.3 hold and let* $\varphi : \mathbb{R}^\zeta \to \mathbb{R}$ *be a function defined by* $\mathbf{y} \mapsto \varphi(\mathbf{y}) := a\lambda_{\min}(\mathbf{C} - \mathcal{A}^\top\mathbf{y}) + \mathbf{b}^\top\mathbf{y}$. *Then,*

$$\tau = \sup_{\mathbf{y}\in\mathbb{R}^\zeta} \varphi(\mathbf{y})\,. \tag{5.3.31}$$

Moreover, if Assumption 5.5 holds, then problem (5.3.31) *has an optimal solution.*

Notice that φ in Lemma 5.4 is concave and continuous but not differentiable in general. The subdifferential of φ at $\mathbf{y}$ reads: $\partial\varphi(\mathbf{y}) = \{\mathbf{b} - a\mathcal{A}\mathbf{U} : \mathbf{U} \in \mathrm{conv}(\Gamma(\mathbf{C} - \mathcal{A}^\top\mathbf{y}))\}$, where for each $\mathbf{A} \in \mathcal{S}$, $\Gamma(\mathbf{A}) := \{\mathbf{u}\mathbf{u}^\top \ : \ \mathbf{A}\mathbf{u} = \lambda_{\min}(\mathbf{A})\mathbf{u}\ ,\ \|\mathbf{u}\|_2 = 1\}$.

Next, we describe Algorithm 10 to solve SDP (5.3.25), which is based on nonsmooth first-order optimization methods (e.g., LMBM [71, Algorithm 1]).

Algorithm 10 Spectral-SDP-CTP

Input: SDP (5.3.25) with unknown optimal value and optimal solution;
method (T) for solving convex nonsmooth unconstrained optimization problems (NSOP).
Output: the optimal value τ of SDP (5.3.25).

1: Compute the optimal value τ and an optimal solution $\bar{\mathbf{y}}$ of the NSOP (5.3.31) by using method (T).

Corollary 5.4. *Let Assumption 5.3 hold. Assume that the method (T) is globally convergent for NSOP* (5.3.31) *(e.g., (T) is LMBM). Then output* τ *of Algorithm 10 is well-defined. Moreover, if Assumption 5.5 holds, the vector* $\bar{\mathbf{y}}$ *mentioned at Step 1 of Algorithm 10 exists.*

SDP with CTP on each subset of blocks

Consider SDP with CTP on each subset of blocks described in Appendix 5.3.4.
The following assumption will be used later on.

Assumption 5.6. *Dual attainability: SDP* (5.3.29) *has an optimal solution.*

Lemma 5.5. *Let Assumption 5.4 hold and let* $\psi : \mathbb{R}^\zeta \to \mathbb{R}$ *be a function defined by* $\mathbf{y} \mapsto \psi(\mathbf{y}) := \mathbf{b}^\top\mathbf{y} + \sum_{c\in[p]} a_j\lambda_{\min}(\mathbf{C}_c - \mathcal{A}_c^\top\mathbf{y})$. *Then,*

$$\tau = \sup_{\mathbf{y}\in\mathbb{R}^\zeta} \psi(\mathbf{y})\,. \tag{5.3.32}$$

Moreover, if Assumption 5.6 holds, then problem (5.3.32) *has an optimal solution.*

Proof. From (5.3.28) and Condition 4 of Assumption 5.4,

$$\tau = \inf_{\mathbf{X}_c \in \mathcal{S}_+^{(c)}} \left\{ \sum_{c\in[p]} \langle \mathbf{C}_c, \mathbf{X}_c \rangle \;\middle|\; \begin{array}{l} \sum_{c\in[p]} \mathcal{A}_c \mathbf{X}_c = \mathbf{b}\,, \\ \langle \mathbf{I}_c, \mathbf{X}_c \rangle = a_c\,,\ c \in [p] \end{array} \right\}, \tag{5.3.33}$$

where $\mathbf{I}_c \in \mathcal{S}^{(c)}$ is the identity matrix, for $c \in [p]$. Note that $\langle \mathbf{I}_c, \mathbf{X}_c \rangle = \mathrm{trace}(\mathbf{X}_c)$, for $\mathbf{X}_c \in \mathcal{S}^{(c)}$, $c \in [p]$. The dual of this SDP reads

$$\rho = \sup_{(\xi,\mathbf{y})\in\mathbb{R}^{p+\zeta}} \left\{ \sum_{c\in[p]} a_c \xi_c + \mathbf{b}^\top \mathbf{y} \,:\, \mathbf{C}_c - \mathcal{A}_c^\top \mathbf{y} - \xi_c \mathbf{I}_c \in \mathcal{S}_+^{(c)}\,, c \in [p] \right\}. \tag{5.3.34}$$

It implies that $\rho = \sup_{\xi,\mathbf{y}} \{ \sum_{c\in[p]} a_c \xi_c + \mathbf{b}^\top \mathbf{y} \,:\, \xi_c \le \lambda_{\min}(\mathbf{C}_c - \mathcal{A}_c^\top \mathbf{y})\,, c \in [p] \}$. From this, the result follows since $\rho = \tau$. □

Proposition 5.4. *The function ψ in Lemma 5.5 has the following properties:*

1. *ψ is concave and continuous but not differentiable in general.*
2. *The subdifferential of ψ at $\mathbf{y}$ satisfies: $\partial\psi(\mathbf{y}) = \mathbf{b} + \sum_{c\in[p]} a_j \partial\psi_j(\mathbf{y})$, where for every $c \in [p]$, $\psi_c : \mathbb{R}^\zeta \to \mathbb{R}$ is a function defined by $\psi_c(\mathbf{y}) = \lambda_{\min}(\mathbf{C}_c - \mathcal{A}_c^\top \mathbf{y})$ and $\partial\psi_c(\mathbf{y}) = \{-\mathcal{A}_c \mathbf{U} \,:\, \mathbf{U} \in \mathrm{conv}(\Gamma(\mathbf{C}_c - \mathcal{A}_c^\top \mathbf{y}))\}$.*

Proof. It is not hard to prove the first statement. Indeed, ψ is a positive combination of $\mathbf{z} \mapsto \mathbf{b}^\top \mathbf{z}$, ψ_c, $c \in [p]$, which are convex, continuous functions. The second statement follows by applying the subdifferential sum rule and notice that the domains of $\mathbf{z} \mapsto \mathbf{b}^\top \mathbf{z}$, ψ_c, $c \in [p]$, are both $\mathbb{R}^n$. □

Next, we describe Algorithm 11 to solve SDP (5.3.28), which is based on nonsmooth first-order optimization methods (e.g., LMBM [71, Algorithm 1]).

Algorithm 11 Spectral-SDP-CTP-Blocks

Input: SDP (5.3.28) with unknown optimal value and optimal solution;
method (T) for solving NSOP.
Output: the optimal value ρ of SDP (5.3.28).

1: Compute the optimal value τ and an optimal solution $\bar{\mathbf{y}}$ of the NSOP (5.3.32) by using method (T).

The fact that Algorithm 11 is well-defined under certain conditions is a corollary of Lemma 5.5 and [134, Lemma A.2].

Corollary 5.5. *Let Assumption 5.4 hold. Assume that the method (T) is globally convergent for NSOP* (5.3.32) *(e.g., (T) is LMBM). Then output τ of Algorithm 11 is well-defined. Moreover, if Assumption 5.6 holds, the vector $\bar{\mathbf{y}}$ involved at Step 1 of Algorithm 11 exists.*

5.3.6 Converting the moment relaxation to the standard SDP

The dense case

Let $k \in \mathbb{N}^{\ge k_{\min}}$ be fixed. We will present a way to transform SDP (5.1.1) to the form (5.1.3). By adding slack variables $\mathbf{y}^{(i)} \in \mathbb{R}^{s(2(k-\lceil g_i \rceil))}$, $i \in [m]$, SDP (5.1.1) is equivalent to

$$\tau_k := \inf_{\mathbf{y},\mathbf{y}^{(i)}} \left\{ L_{\mathbf{y}}(f) \;\middle|\; \begin{array}{l} \mathbf{W}_k(\mathbf{y}, \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(m)}) \in \mathcal{S}_+^{(k)}\,, \\ \mathbf{M}_{k-\lceil g_i \rceil}(\mathbf{y}^{(i)}) = \mathbf{M}_{k-\lceil g_i \rceil}(g_i\, \mathbf{y})\,,\ i \in [m]\,, \\ \mathbf{M}_{k-\lceil h_j \rceil}(h_j\, \mathbf{y}) = 0\,,\ j \in [l] \end{array} \right\}, \tag{5.3.35}$$

where $\mathbf{W}_k(\mathbf{y}, \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(m)}) := \mathrm{diag}(\mathbf{M}_k(\mathbf{y}), \mathbf{M}_{k-\lceil g_1 \rceil}(\mathbf{y}^{(1)}), \ldots, \mathbf{M}_{k-\lceil g_m \rceil}(\mathbf{y}^{(m)}))$.

Let $\mathcal{V} = \{\mathbf{M}_k(\mathbf{z}) : \mathbf{z} \in \mathbb{R}^{s(2k)}\}$ and $\mathcal{V}_i = \{\mathbf{M}_{k-\lceil g_i \rceil}(\mathbf{z}) : \mathbf{z} \in \mathbb{R}^{s(2(k-\lceil g_i \rceil))}\}$, $i \in [m]$. Then $\mathcal{V}$ and $\mathcal{V}_i$, $i \in [m]$, are the linear subspaces of the spaces of real symmetric matrices of size $s(k)$ and $s(k - \lceil g_i \rceil)$, $i \in [m]$, respectively.

Denote by $\mathcal{V}^\perp$, $\mathcal{V}_i^\perp$, $i \in [m]$, the orthogonal complements of $\mathcal{V}$, $\mathcal{V}_i$, $i \in [m]$, respectively. In [134, Appendix A.2], we show how to take a basis $\{\hat{\mathbf{A}}_j\}_{j\in[r]}$ of $\mathcal{V}^\perp$. Similarly we can take a basis $\{\hat{\mathbf{A}}_j^{(i)}\}_{j\in[r_i]}$ of $\mathcal{V}_i^\perp$, $i \in [m]$. Here $r = \dim(\mathcal{V}^\perp)$ and $r_i = \dim(\mathcal{V}_i^\perp)$, $i \in [m]$.

Notice that if $\mathbf{X}_0$ is a real symmetric matrix of size $s(k)$, then $\mathbf{X}_0 = \mathbf{M}_k(\mathbf{y})$ for some $\mathbf{y} \in \mathbb{R}^{s(2k)}$ if and only if $\langle \hat{\mathbf{A}}_j, \mathbf{X}_0 \rangle = 0$, $j \in [r]$. It implies that if $\mathbf{X} = \operatorname{diag}(\mathbf{X}_0, \ldots, \mathbf{X}_m) \in \mathcal{S}^{(k)}$, then there exist $\mathbf{y}$ and $\mathbf{y}^{(i)}$, $i \in [m]$, such that $\mathbf{X} = \mathbf{W}_k(\mathbf{y}, \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(m)}) \Leftrightarrow \langle \bar{\mathbf{A}}, \mathbf{X} \rangle = 0$, $\bar{\mathbf{A}} \in \mathcal{B}_1$, where $\mathcal{B}_1$ involves matrices $\bar{\mathbf{A}}$ defined as

- $\bar{\mathbf{A}} = \operatorname{diag}(\hat{\mathbf{A}}_j, \mathbf{0}, \ldots, \mathbf{0})$ for some $j \in [r]$;
- $\bar{\mathbf{A}} = \operatorname{diag}(\mathbf{0}, \hat{\mathbf{A}}_j^{(1)}, \ldots, \mathbf{0})$ for some $j \in [r_1]$;
- ...
- $\bar{\mathbf{A}} = \operatorname{diag}(\mathbf{0}, \mathbf{0}, \ldots, \hat{\mathbf{A}}_j^{(m)})$ for some $j \in [r_m]$.

Notice that

$$|\mathcal{B}_1| = r + \textstyle\sum_{i\in[m]} r_i = \begin{array}{l} \dfrac{b(k)(b(k)+1)}{2} - b(2k) \\ \quad + \displaystyle\sum_{i\in[m]} \left(\frac{b(k-\lceil g_i \rceil)(b(k-\lceil g_i \rceil)+1)}{2} - b(2(k-\lceil g_i \rceil)) \right) . \end{array} \tag{5.3.36}$$

The constraints $\mathbf{M}_{k-\lceil g_i \rceil}(\mathbf{y}^{(i)}) = \mathbf{M}_{k-\lceil g_i \rceil}(g_i\, \mathbf{y})$, $i \in [m]$, of SDP (5.3.35) are equivalent to $\mathbf{y}_{\boldsymbol{\alpha}}^{(i)} = \sum_{\boldsymbol{\gamma}\in\mathbb{N}^n_{2\lceil g_i \rceil}} g_i \mathbf{y}_{\boldsymbol{\alpha}+\boldsymbol{\gamma}}$, $\boldsymbol{\alpha} \in \mathbb{N}^n_{2(k-\lceil g_i \rceil)}$, $i \in [m]$. They can be written as $\langle \bar{\mathbf{A}}, \mathbf{W}_k(\mathbf{y}, \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(m)}) \rangle = 0$, for $\bar{\mathbf{A}} \in \mathcal{B}_2$, where $\mathcal{B}_2$ involves matrices $\bar{\mathbf{A}}$ defined by $\bar{\mathbf{A}} = \operatorname{diag}(\tilde{\mathbf{A}}, \mathbf{0}, \ldots, \mathbf{0}, \tilde{\mathbf{A}}^{(i)}, \mathbf{0}, \ldots, \mathbf{0})$, with $\tilde{\mathbf{A}} = (\tilde{A}_{\boldsymbol{\mu},\boldsymbol{\nu}})_{\boldsymbol{\mu},\boldsymbol{\nu}\in\mathbb{N}^n_k}$ being defined as follows:

$$\tilde{A}_{\boldsymbol{\mu},\boldsymbol{\nu}} = \begin{cases} g_{i,\boldsymbol{\gamma}} & \text{if } \boldsymbol{\mu} = \boldsymbol{\nu}\,,\ \boldsymbol{\mu}+\boldsymbol{\nu} = \boldsymbol{\alpha}+\boldsymbol{\gamma}\,, \\ \frac{1}{2} g_{i,\boldsymbol{\gamma}} & \text{if } \boldsymbol{\mu} \neq \boldsymbol{\nu}\,,\ (\boldsymbol{\mu},\boldsymbol{\nu}) \in \{(\boldsymbol{\mu}_1,\boldsymbol{\nu}_1),(\boldsymbol{\nu}_1,\boldsymbol{\mu}_1)\} \\ & \qquad \text{with } (\boldsymbol{\mu}_1,\boldsymbol{\nu}_1) = \text{minimal}(\{(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\nu}}) \in (\mathbb{N}^n_k)^2 \,:\, \bar{\boldsymbol{\mu}}+\bar{\boldsymbol{\nu}} = \boldsymbol{\alpha}+\boldsymbol{\gamma}\})\,, \\ 0 & \text{otherwise,} \end{cases} \tag{5.3.37}$$

and $\tilde{\mathbf{A}}^{(i)} = (\tilde{A}^{(i)}_{\boldsymbol{\mu},\boldsymbol{\nu}})_{\boldsymbol{\mu},\boldsymbol{\nu}\in\mathbb{N}^n_{k-\lceil g_i \rceil}}$ being defined as follows:

$$\tilde{A}^{(i)}_{\boldsymbol{\mu},\boldsymbol{\nu}} = \begin{cases} -1 & \text{if } \boldsymbol{\mu} = \boldsymbol{\nu}\,,\ \boldsymbol{\mu}+\boldsymbol{\nu} = \boldsymbol{\alpha}\,, \\ -\frac{1}{2} & \text{if } \boldsymbol{\mu} \neq \boldsymbol{\nu}\,,\ (\boldsymbol{\mu},\boldsymbol{\nu}) \in \{(\boldsymbol{\mu}_1,\boldsymbol{\nu}_1),(\boldsymbol{\nu}_1,\boldsymbol{\mu}_1)\} \\ & \qquad \text{with } (\boldsymbol{\mu}_1,\boldsymbol{\nu}_1) = \text{minimal}(\{(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\nu}}) \in (\mathbb{N}^n_k)^2 \,:\, \bar{\boldsymbol{\mu}}+\bar{\boldsymbol{\nu}} = \boldsymbol{\alpha}\})\,, \\ 0 & \text{otherwise,} \end{cases} \tag{5.3.38}$$

for some $\boldsymbol{\alpha} \in \mathbb{N}^n_{2(k-\lceil g_i \rceil)}$ and $i \in [m]$. Notice that $|\mathcal{B}_2| = \sum_{i\in[m]} b(2(k - \lceil g_i \rceil))$. Here minimal$(T)$ is the minimal element of T, for every $T \subseteq \mathbb{N}^{2n}$ with respect to the graded lexicographic order.

The constraints $\mathbf{M}_{k-\lceil h_j \rceil}(h_j\ \mathbf{y}) = 0$, $j \in [l]$, can be simplified as $\sum_{\boldsymbol{\gamma}\in\mathbb{N}^n_{2\lceil h_j \rceil}} h_{j,\boldsymbol{\gamma}} y_{\boldsymbol{\alpha}+\boldsymbol{\gamma}} = 0$, $\boldsymbol{\alpha} \in \mathbb{N}^n_{2(k-\lceil h_j \rceil)}$, $j \in [l]$. They are equivalent to the following trace equality constraints: $\langle \bar{\mathbf{A}}, \mathbf{W}_k(\mathbf{y}, \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(m)}) \rangle = 0$, $\bar{\mathbf{A}} \in \mathcal{B}_3$, where $\mathcal{B}_3$ involves matrices $\bar{\mathbf{A}} = \operatorname{diag}(\tilde{\mathbf{A}}, \mathbf{0}, \ldots, \mathbf{0})$, with $\tilde{\mathbf{A}} = (\tilde{A}_{\boldsymbol{\mu},\boldsymbol{\nu}})_{\boldsymbol{\mu},\boldsymbol{\nu}\in\mathbb{N}^n_k}$ being defined as follows:

$$\tilde{A}_{\boldsymbol{\mu},\boldsymbol{\nu}} = \begin{cases} h_{j,\boldsymbol{\gamma}} & \text{if } \boldsymbol{\mu} = \boldsymbol{\nu}\,,\ \boldsymbol{\mu}+\boldsymbol{\nu} = \boldsymbol{\alpha}+\boldsymbol{\gamma}\,, \\ \frac{1}{2} h_{j,\boldsymbol{\gamma}} & \text{if } \boldsymbol{\mu} \neq \boldsymbol{\nu}\,,\ (\boldsymbol{\mu},\boldsymbol{\nu}) \in \{(\boldsymbol{\mu}_1,\boldsymbol{\nu}_1),(\boldsymbol{\nu}_1,\boldsymbol{\mu}_1)\} \\ & \qquad \text{with } (\boldsymbol{\mu}_1,\boldsymbol{\nu}_1) = \text{minimal}(\{(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\nu}}) \in (\mathbb{N}^n_k)^2 \,:\, \bar{\boldsymbol{\mu}}+\bar{\boldsymbol{\nu}} = \boldsymbol{\alpha}+\boldsymbol{\gamma}\})\,, \\ 0 & \text{otherwise.} \end{cases}$$

Notice that $|\mathcal{B}_3| = \sum_{j\in[l]} b(2(k-\lceil h_j\rceil))$.

Let $\cup_{j\in[3]}\mathcal{B}_j = (\bar{\mathbf{A}}_i)_{i\in[\zeta_k-1]}$, where

$$\zeta_k = 1+\sum_{j\in[3]}|\mathcal{B}_j| = \; 1+\frac{b(k)(b(k)+1)}{2}-b(2k) + \sum_{i\in[m]}\frac{b(k-\lceil g_i\rceil)(b(k-\lceil g_i\rceil)+1)}{2}+\sum_{j\in[l]} b(2(k-\lceil h_j\rceil))\,.$$

The final constraint $y_{\mathbf{0}}=1$ can be rewritten as $\langle \bar{\mathbf{A}}_{\zeta_k}, \mathbf{W}_k(\mathbf{y},\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(m)})\rangle = 1$ with $\bar{\mathbf{A}}_{\zeta_k}\in\mathcal{S}^{(k)}$ having zero entries except the top left one $[\bar{A}_{\zeta_k}]_{\mathbf{0},\mathbf{0}}=1$. Thus we select real vector $\mathbf{b}_k$ of length t_k such that all entries of $\mathbf{b}_k$ are zeros except the final one $b_{k,\zeta_k}=1$.

The function $L_{\mathbf{y}}(f)=\sum_{\boldsymbol{\gamma}} f_{\boldsymbol{\gamma}} y_{\boldsymbol{\gamma}}$ is equal to $\langle \bar{\mathbf{C}}, \mathbf{W}_k(\mathbf{y},\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(m)})\rangle$ with $\bar{\mathbf{C}} := \operatorname{diag}(\tilde{\mathbf{C}},\mathbf{0},\ldots,\mathbf{0})$, where $\tilde{\mathbf{C}} = (\tilde{C}_{\boldsymbol{\mu},\boldsymbol{\nu}})_{\boldsymbol{\mu},\boldsymbol{\nu}\in\mathbb{N}^n_k}$ is defined by

$$\tilde{C}_{\boldsymbol{\mu},\boldsymbol{\nu}} = \begin{cases} f_{\boldsymbol{\gamma}} & \text{if } \boldsymbol{\mu}=\boldsymbol{\nu}\,,\ \boldsymbol{\mu}+\boldsymbol{\nu}=\boldsymbol{\gamma}\,,\\ \frac{1}{2}f_{\boldsymbol{\gamma}} & \text{if } \boldsymbol{\mu}\neq\boldsymbol{\nu}\,,\ (\boldsymbol{\mu},\boldsymbol{\nu})\in\{(\boldsymbol{\mu}_1,\boldsymbol{\nu}_1),(\boldsymbol{\nu}_1,\boldsymbol{\mu}_1)\}\\ & \qquad \text{with } (\boldsymbol{\mu}_1,\boldsymbol{\nu}_1)=\text{minimal}(\{(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\nu}})\in(\mathbb{N}^n_k)^2 \,:\, \bar{\boldsymbol{\mu}}+\bar{\boldsymbol{\nu}}=\boldsymbol{\gamma}\})\,,\\ 0 & \text{otherwise.}\end{cases}$$

By writing $\bar{\mathbf{X}} = \mathbf{W}_k(\mathbf{y},\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(m)})$, SDP (5.3.35) has the standard form

$$\tau_k = \inf_{\bar{\mathbf{X}}\in\mathcal{S}^{(k)}_+}\{\,\langle\bar{\mathbf{C}},\bar{\mathbf{X}}\rangle \,:\, \bar{\mathcal{A}}\bar{\mathbf{X}}=\mathbf{b}_k\}\,, \tag{5.3.39}$$

where $\bar{\mathcal{A}}:\mathcal{S}^{(k)}\to\mathbb{R}^{\zeta_k}$ is a linear operator of the form $\bar{\mathcal{A}}\mathbf{X} = \big[\langle\bar{\mathbf{A}}_1,\mathbf{X}\rangle,\ldots,\langle\bar{\mathbf{A}}_{\zeta_k},\mathbf{X}\rangle\big]$. Since $\langle\mathbf{U},\mathbf{V}\rangle = \langle\mathbf{P}_k^{-1}\mathbf{U}\mathbf{P}_k^{-1},\mathbf{P}_k\mathbf{V}\mathbf{P}_k\rangle$, for all $\mathbf{U},\mathbf{V}\in\mathcal{S}^{(k)}$, by noting $\mathbf{X}=\mathbf{P}_k\bar{\mathbf{X}}\mathbf{P}_k$, SDP (5.3.39) can be written as (5.1.3) with $\mathbf{A}_{k,i}=\mathbf{P}_k^{-1}\bar{\mathbf{A}}_i\mathbf{P}_k^{-1}$, $i\in[\zeta_k]$, and $\mathbf{C}_k=\mathbf{P}_k^{-1}\bar{\mathbf{C}}\mathbf{P}_k^{-1}$.

The sparse case

Let $k\in\mathbb{N}^{\geq k_{\min}}$ be fixed. We will present a way to transform SDP (5.3.4) to the form (5.3.8). Doing a similar process as in Appendix 5.3.6 on every clique, by noting (5.3.7), for every $c\in[p]$, the constraints

$$\begin{cases}\mathbf{D}_k(\mathbf{y},I_c)\succeq 0\,,\ y_{\mathbf{0}}=1\,,\\ \mathbf{M}_{k-\lceil h_j\rceil}(h_j\,\mathbf{y},I_c)=0\,,\ j\in W_c\,,\end{cases} \tag{5.3.40}$$

become $\hat{\mathcal{A}}_c\mathbf{X}_c=\hat{\mathbf{b}}_c$ for some linear operator $\hat{\mathcal{A}}_c:\mathcal{S}^{(c,k)}\to\mathbb{R}^{\hat{\zeta}_c}$ and vector $\hat{\mathbf{b}}_c\in\mathbb{R}^{\hat{\zeta}_c}$. Moreover, $L_{\mathbf{y}}(f_c)=\langle\mathbf{C}_c,\mathbf{X}_c\rangle$ for some matrix $\mathbf{C}_c\in\mathcal{S}^{(c,k)}$ since $f_c\in\mathbb{R}[\mathbf{x}(I_c)]$, for every $c\in[p]$. Then from (5.3.7), the objective function of SDP (5.3.4) is $L_{\mathbf{y}}(f)=\sum_{c\in[p]}\langle\mathbf{C}_c,\mathbf{X}_c\rangle$.

Next we describe the constraints depending on common moments on cliques. For every $\boldsymbol{\alpha}\in\cup_{c\in[p]}\mathbb{N}^{I_c}_k$, note $T(\boldsymbol{\alpha}):=\{c\in[p]\,:\,\boldsymbol{\alpha}\in\mathbb{N}^{I_c}_k\}$. In other words, $T(\boldsymbol{\alpha})$ indices the cliques sharing the same moment $y_{\boldsymbol{\alpha}}$. For $\boldsymbol{\alpha}\in\cup_{c\in[p]}\mathbb{N}^{I_c}_k$ such that $|T(\boldsymbol{\alpha})|\geq 2$, for every $c\in T(\boldsymbol{\alpha})$, let $\hat{\mathbf{A}}^{(\boldsymbol{\alpha})}_c\in\mathcal{S}^{(c,k)}$ be such that $\langle\hat{\mathbf{A}}^{(\boldsymbol{\alpha})}_c,\mathbf{X}_c\rangle = y_{\boldsymbol{\alpha}}$ It implies the constraints $\langle\hat{\mathbf{A}}^{(\boldsymbol{\alpha})}_{j_0},\mathbf{X}_{j_0}\rangle - \langle\hat{\mathbf{A}}^{(\boldsymbol{\alpha})}_i,\mathbf{X}_i\rangle = 0$, $i\in T(\boldsymbol{\alpha})\backslash\{j_0\}$, for every $\boldsymbol{\alpha}\in\cup_{c\in[p]}\mathbb{N}^{I_c}_k$ such that $|T(\boldsymbol{\alpha})|\geq 2$, for some $j_0\in T(\boldsymbol{\alpha})$. We denote by $\tilde{\mathcal{A}}\mathbf{X}=\mathbf{0}_{\mathbb{R}^{\tilde{\zeta}}}$ all these constraints with $\mathbf{X}=\operatorname{diag}((\mathbf{X}_c)_{c\in[p]})$.

Set $\zeta:=\sum_{c\in[p]}\hat{\zeta}_c+\tilde{\zeta}$ and $\mathbf{b}=[(\hat{\mathbf{b}}_c)_{c\in[p]},\mathbf{0}_{\mathbb{R}^{\tilde{\zeta}}}]\in\mathbb{R}^{\zeta}$. Define the linear operator $\mathcal{A}:\prod_{c\in[p]}\mathcal{S}^{(c,k)}\to\mathbb{R}^{\zeta}$ such that $\mathcal{A}\mathbf{X}=[(\hat{\mathcal{A}}_c\mathbf{X}_c)_{c\in[p]},\tilde{\mathcal{A}}\mathbf{X}]$, for all $\mathbf{X}=\operatorname{diag}(\mathbf{X}_c)_{c\in[p]}\in\prod_{c\in[p]}\mathcal{S}^{(c)}$. From (5.3.7), the affine constraints of SDP (5.3.4) are now equivalent to $\mathcal{A}\mathbf{X}=\mathbf{b}$.

Let $\mathbf{A}^{(i)}:=\operatorname{diag}((\mathbf{A}_{i,c})_{c\in[p]})\in\prod_{c\in[p]}\mathcal{S}^{(c)}$, $i\in[\zeta]$, be such that

$$\mathcal{A}\mathbf{X}=[\langle\mathbf{A}^{(1)},\mathbf{X}\rangle,\ldots,\langle\mathbf{A}^{(\zeta)},\mathbf{X}\rangle]\,,$$

for all $\mathbf{X}=\operatorname{diag}(\mathbf{X}_c)_{c\in[p]}\in\prod_{c\in[p]}\mathcal{S}^{(c)}$. For every $c\in[p]$, define $\mathcal{A}_c:\mathcal{S}^{(c)}\to\mathbb{R}^{\zeta}$ as a linear operator of the form $\mathcal{A}_c\mathbf{X}:=[\langle\mathbf{A}_{1,c},\mathbf{X}\rangle,\ldots,\langle\mathbf{A}_{\zeta,c},\mathbf{X}\rangle]$. Then $\mathcal{A}\mathbf{X}=\sum_{c\in[p]}\mathcal{A}_c\mathbf{X}_c$, for all $\mathbf{X}=\operatorname{diag}(\mathbf{X}_c)_{c\in[p]}\in\prod_{c\in[p]}\mathcal{S}^{(c)}$. Hence we obtain the data $(\mathbf{C}_{c,k},\mathcal{A}_{c,k},\mathbf{b}_k,\zeta_k)=(\mathbf{C}_c,\mathcal{A}_c,\mathbf{b},\zeta)$ of the standard form (5.3.8) by plugging k.

Part II

Positivity certificates with denominators

Chapter 6

Polynomial optimization over noncompact semialgebraic sets

Most of the content of this chapter is from [130].

In this chapter, we focus on polynomial optimization on noncompact semialgebraic sets by relying on positivity certificates involving denominators. The main motivation is to voluntarily avoid the *big-ball trick* which reduces the problem to the compact case. The big-ball "trick" is to simply assume that the global minimum is attained in some *a priori known* ball $B(\mathbf{0}, R)$ centered at zero of radius $R > 0$ potentially large. Therefore, by adding this additional constraint to the definition of the feasible set, one is back to the compact case. Why? This "trick" has definitely some merit since in some practical applications such an R can be sometimes determined with *ad-hoc* arguments. However, it is not satisfactory from a mathematical point of view. Indeed after one has found a minimizer $\mathbf{x}^\star \in B(\mathbf{0}, R)$, one is still left with the question: *Is really* $\mathbf{x}^\star$ *a global minimizer? Was* R *chosen sufficiently large?* In other words, in doing so one does not obtain an certificate that $\mathbf{x}^\star$ is a global minimizer. As we will see in this chapter, we deal with the challenge to *adapt* some certificates of positivity on noncompact sets already available in the literature, to turn them into a practical algorithm.

Let us recall the Positivstellensatz [170] of Putinar and Vasilescu in the following theorem:

Theorem 6.1. *(Putinar–Vasilescu[170, Corollary 4.3 and 4.4]) Let* $\theta \in \mathbb{R}[\mathbf{x}]$ *be the quadratic polynomial* $x \mapsto \theta(\mathbf{x}) := 1 + \|\mathbf{x}\|_2^2$*, and denote by* $\tilde{p} \in \mathbb{R}[\mathbf{x}, x_{n+1}]$ *the homogeneous polynomial associated with* $p \in \mathbb{R}[\mathbf{x}]$*, defined by* $x \mapsto \tilde{p}(\mathbf{x}) := x_{n+1}^{\deg(p)} p(x/x_{n+1})$*.*

1. *Let* $f \in \mathbb{R}[\mathbf{x}]$ *such that* $\tilde{f} > 0$ *on* $\mathbb{R}^{n+1}\backslash\{\mathbf{0}\}$*. Then* $\theta^k f \in \Sigma[\mathbf{x}]$ *for some* $k \in \mathbb{N}$*.*
2. *Let* $f, g_1, \ldots, g_m \in \mathbb{R}[\mathbf{x}]$ *satisfy the following two conditions:*
   (a) $f = f_0 + f_1$ *such that* $\deg(f_0) < \deg(f_1)$ *and* $\tilde{f}_1 > 0$ *on* $\mathbb{R}^{n+1}\backslash\{\mathbf{0}\}$*;*
   (b) $f > 0$ *on* $S(\mathfrak{g})$*.*

   Then $\theta^{2k} f \in \mathcal{Q}(\mathfrak{g})$ *for some* $k \in \mathbb{N}$*.*

As a consequence, they also obtain:

Corollary 6.1. *(Putinar–Vasilescu [170, Final remark 2]) Let* $\theta := 1 + \|\mathbf{x}\|_2^2$*.*

1. *Let* $f \in \mathbb{R}[\mathbf{x}]_{2d}$ *be such that* $f \geq 0$ *on* $\mathbb{R}^n$*. Then for all* $\varepsilon > 0$*, there exists* $k \in \mathbb{N}$ *such that* $\theta^k(f + \varepsilon\theta^d) \in \Sigma[\mathbf{x}]$*.*
2. *Let* $f \in \mathbb{R}[\mathbf{x}]$ *such that* $f \geq 0$ *on* $S(\mathfrak{g})$*. Let* $d \in \mathbb{N}$ *such that* $2d > \deg(f)$*. Then for all* $\varepsilon > 0$*, there exists* $k \in \mathbb{N}$ *such that* $\theta^{2k}(f + \varepsilon\theta^d) \in \mathcal{Q}(\mathfrak{g})$*.*

Marshall [138, Corollary 4.3] states a slightly more general result but with no explicit d, and Schweighofer [187, Corollary 6.3] provides a new algebraic proof of Marshall's result. To summarize, for every polynomial f nonnegative on a general basic semialgebraic set $S(\mathfrak{g})$, one obtains the

following representation result: for a given $\varepsilon > 0$, there exist a nonnegative integer k and SOS polynomials $\sigma_0, \sigma_1, \ldots, \sigma_m$, such that

$$f + \varepsilon\theta^d = \frac{\sigma_0 + \sigma_1 g_1 + \cdots + \sigma_m g_m}{\theta^k}\,. \tag{6.0.1}$$

Although this representation is theoretically attractive, their previous proofs are not constructive and do not provide any explicit algorithm, especially in polynomial optimization.

Contribution. As already mentioned, our approach is to treat the noncompact case frontally and avoid the big-ball trick. Our contribution is threefold:

I. In Section 6.1 we first provide an alternative proof of (6.0.1), *with an explicit degree bound* on the SOS weights, by relying on Jacobi's technique in the proof of [85, Theorem 7]; this is crucial as it has immediate implications on the algorithmic side. More precisely, the degrees of SOS weights σ_i are bounded above by $k + d - \lceil g_i \rceil$. First, one transforms the initial polynomials to homogeneous forms, then one relies on Putinar's Positivstellensatz for the compact case, and finally one transforms back the obtained forms to dehomogenized polynomials. As a consequence, with $\varepsilon > 0$ fixed, arbitrary, this degree bound allows us to provide hierarchies $(\rho_k^i(\varepsilon))_{k\in\mathbb{N}}$, $i = 1, 2, 3$ for unconstrained polynomial optimization ($m = 0$ and $i = 1$, see Section 6.2.1) as well as for constrained polynomial optimization ($m \geq 1$ and $i = 2, 3$, see Section 6.2.2). Computing each $\rho_k^i(\varepsilon)$ boils down to solving a single SDP, with strong duality property. For k sufficiently large, $\rho_k^i(\varepsilon)$ becomes an upper bound for the optimal value $f^\star$ of the corresponding polynomial optimization problem (POP) $\min_{\mathbf{x}\in S(\mathfrak{g})} f(\mathbf{x})$. If this problem has an optimal solution $\mathbf{x}^\star$, the gap between $\rho_k^i(\varepsilon)$ and $f^\star$ is at most $\varepsilon\theta(\mathbf{x}^\star)^d$. The related convergence rates are also analyzed in these sections.

II. In Section 6.2.3, we provide a new algorithm to find a feasible solution in the set $S(\mathfrak{g}, \mathfrak{h})$ defined in (1.1.1). The idea is to include appropriate additional spherical equality constraints $\varphi_t := \xi_t - \|\mathbf{x} - \mathbf{a_t}\|_2^2$, $t = 0, \ldots, n$, in $S(\mathfrak{g}, \mathfrak{h})$ so that the system $S(\mathfrak{g}, \mathfrak{h} \cup \{\varphi_0, \ldots, \varphi_n\})$ has a unique real solution. The nonnegative reals $(\xi_t)_{t=0,\ldots,n}$ are computed with an adequate Moment-SOS hierarchy. Moreover, this solution might be extracted in certain cases by checking whether some (moment) matrix satisfies a flat extension condition.

III. Finally we use this method to approximate a global minimizer of f on $S(\mathfrak{g}, \mathfrak{h})$. Namely, we fix $\varepsilon > 0$ small and find a point in $S(\mathfrak{g} \cup \{\rho_k^i(\varepsilon) - f\}, \mathfrak{h})$. This procedure works in certain cases, even if the set of minimizers is infinite. This is in deep contrast with the extraction procedure of [77] (via some flat extension condition) which works only for finite solution sets. Assuming that the set of solutions is finite, one may compare our algorithm with the procedure from [77] as follows. On the one hand, the latter extraction procedure provides global optimizers, provided that one has solved an SDP-relaxation with sufficiently large "k" (so as to get an appropriate rank condition). On the other hand, our algorithm that adds spherical equality constraints "divides" the problem into $n + 1$ SDP relaxations with additional constraints but with smaller order "k" (which is *the* crucial parameter for the SDP solvers). Numerical examples are provided in Section 6.3 to illustrate the difference between these two strategies.

For clarity of exposition, most proofs are omitted, they are available in [130, Appendix].

6.1 Representation theorems

In this section we provide two exact representations of globally nonnegative polynomials and polynomials nonnegative on basic semialgebraic sets (not necessarily compact). The representations are obtained thanks to a perturbation argument as well as existing representations for positive definite forms. Let $\theta := 1 + \|\mathbf{x}\|_2^2$.

6.1.1 Globally nonnegative polynomials

Let us note $\|q\|_1 := \sum_{\boldsymbol{\alpha}} |q_{\boldsymbol{\alpha}}|$ for a given $q \in \mathbb{R}[\mathbf{x}]$. The following result provides a representation of globally nonnegative polynomials.

Theorem 6.2. *Let $f \in \mathbb{R}[\mathbf{x}]_{2d}$ be nonnegative on $\mathbb{R}^n$. Then for every $\varepsilon > 0$, for $k_\varepsilon \in \mathbb{N}$ and*

$$k_\varepsilon \geq \frac{2(n+1)d(2d-1)}{4\log 2}(\varepsilon^{-1}\|f\|_1 + 1) - \frac{n+1+2d}{2}\,,$$

one has

$$\theta^{k_\varepsilon}\left(f + \varepsilon\theta^d\right) \in \Sigma[\mathbf{x}]_{k_\varepsilon+d}\,. \tag{6.1.1}$$

Proof. The proof consists of three steps:

1. Associate a positive definite form to the globally nonnegative polynomial f.
2. Use Reznick's representation from Theorem 1.2 to get a representation of this homogeneous form.
3. Transform back the homogeneous polynomial together with its representation to the original polynomial.

Let $\tilde{f} = x_{n+1}^{2d} f(\mathbf{x}/x_{n+1})$ be the degree $2d$ homogenization of f. Since f is globally nonnegative, $\tilde{f}$ is nonnegative on $\mathbb{R}^{n+1}$. Let $\varepsilon > 0$ be fixed. We claim that

$$\tilde{f} + \varepsilon\|(\mathbf{x}, x_{n+1})\|_2^{2d} \in \mathbb{R}[\mathbf{x}, x_{n+1}]$$

is positive definite, i.e., is homogeneous and positive on $\mathbb{R}^{n+1}\backslash\{\mathbf{0}\}$. Since

$$\|(\mathbf{x}, x_{n+1})\|_2^{2d} = (x_1^2 + \cdots + x_n^2 + x_{n+1}^2)^d\,,$$

the polynomial $\|(\mathbf{x}, x_{n+1})\|_2^{2d}$ is homogeneous of degree $2d$ on $\mathbb{R}^{n+1}$. From this and since $\tilde{f}$ is homogeneous of degree $2d$, $\tilde{f} + \varepsilon\|(\mathbf{x}, x_{n+1})\|_2^{2d}$ is homogeneous of degree $2d$. For every $(\mathbf{x}, x_{n+1}) \in \mathbb{R}^{n+1}\backslash\{\mathbf{0}\}$, $\|(\mathbf{x}, x_{n+1})\|_2 > 0$. From this and since $\tilde{f}$ is nonnegative on $\mathbb{R}^{n+1}$,

$$\tilde{f}(\mathbf{x}, x_{n+1}) + \varepsilon\|(\mathbf{x}, x_{n+1})\|_2^{2d} > 0\,,$$

for all $(\mathbf{x}, x_{n+1}) \in \mathbb{R}^{n+1}\backslash\{\mathbf{0}\}$. In addition, it is not hard to show that

$$\inf\{\tilde{f}(\mathbf{x}, x_{n+1}) + \varepsilon\|(\mathbf{x}, x_{n+1})\|_2^{2d} : (\mathbf{x}, x_{n+1}) \in \mathbb{S}^n\} \geq \varepsilon$$

and

$$\begin{array}{ll} & \sup\{\tilde{f}(\mathbf{x}, x_{n+1}) + \varepsilon\|(\mathbf{x}, x_{n+1})\|_2^{2d} : (\mathbf{x}, x_{n+1}) \in \mathbb{S}^n\} \\ \leq & \sup\{\tilde{f}(\mathbf{x}, x_{n+1}) : (\mathbf{x}, x_{n+1}) \in \mathbb{S}^n\} + \varepsilon \\ \leq & \|\tilde{f}\|_1 + \varepsilon\,. \end{array}$$

Thus, $\delta(\tilde{f} + \|.\|_2^{2d}) \leq (\|\tilde{f}\|_1 + \varepsilon)/\varepsilon = (\|f\|_1 + \varepsilon)/\varepsilon$. From this and by applying Theorem 1.2 with $p = \tilde{f} + \varepsilon\|(\mathbf{x}, x_{n+1})\|_2^{2d}$, for $k_\varepsilon \in \mathbb{N}$ and

$$k_\varepsilon \geq \frac{2(n+1)d(2d-1)}{(4\log 2)}(\varepsilon^{-1}\|f\|_1 + 1) - \frac{n+1+2d}{2}\,,$$

there exists $\tilde{\sigma}_\varepsilon \in \Sigma[\mathbf{x}, x_{n+1}]_{k_\varepsilon+d}$ such that

$$\|(\mathbf{x}, x_{n+1})\|_2^{2k_\varepsilon}(\tilde{f} + \varepsilon\|(\mathbf{x}, x_{n+1})\|_2^{2d}) = \tilde{\sigma}_\varepsilon\,.$$

By replacing x_{n+1} by 1, one has

$$\theta^{k_\varepsilon}(f + \varepsilon\theta^d) = \tilde{\sigma}_\varepsilon(\mathbf{x}, 1)\,.$$

Let us note $\sigma_\varepsilon(\mathbf{x}) := \tilde{\sigma}_\varepsilon(\mathbf{x}, 1)$, for every $\mathbf{x} \in \mathbb{R}^n$. Since $\tilde{\sigma}_\varepsilon \in \Sigma[\mathbf{x}, x_{n+1}]_{k_\varepsilon+d}$, it follows that $\sigma_\varepsilon \in \Sigma[\mathbf{x}]_{k_\varepsilon+d}$, yielding the desired result. □

6.1.2 Polynomials nonnegative on a basic semialgebraic set

We recall the definition of the truncated quadratic module of order d associated with $S(\mathfrak{g})$:

$$\mathcal{Q}_d(\mathfrak{g}) := \left\{ \sigma_0 + \sum_{j=1}^{m} \sigma_i g_i : \ \sigma_0 \in \Sigma[\mathbf{x}]_d \, , \, \sigma_i \in \Sigma[\mathbf{x}]_{d-\lceil g_i \rceil} \right\} .$$

For every $q \in \mathbb{R}[\mathbf{x}]$, let us define

$$d_1(q) := 1 + \lfloor \deg(q)/2 \rfloor \qquad \text{and} \qquad d_2(q) := \lceil \deg(q)/2 \rceil = \lceil q \rceil .$$

The following result provides a degree bound for the SOS weights of [169, Theorem 1].

Theorem 6.3. *Let $\mathfrak{g} = \{g_1, \ldots, g_m\} \subset \mathbb{R}[\mathbf{x}]$ and $f \in \mathbb{R}[\mathbf{x}]$ such that f is nonnegative on $S(\mathfrak{g})$. Let $\varepsilon > 0$ and $d \in \mathbb{N}$ be such that at least one of the following two conditions is satisfied:*
(i) $d \geq d_1(f)$;
(ii) $d \geq d_2(f)$ and $g_m := f + \lambda$ for some real $\lambda \geq 0$.
Then there exist $k_\varepsilon \in \mathbb{N}$ such that

$$\theta^{k_\varepsilon} \left(f + \varepsilon \, \theta^d \right) \in \mathcal{Q}_{k_\varepsilon + d}(\mathfrak{g}) \, . \tag{6.1.2}$$

The detailed proof of Theorem 6.3 relies on Jacobi's technique in his proof of [85, Theorem 7] and is available in [130, Appendix]. This proof consists of three steps:

1. Associate a homogeneous polynomial $\tilde{f}$ to the polynomial f.
2. Use Putinar's Positivstellensatz (Theorem 1.1 (i)) to obtain a representation of $\tilde{f}$.
3. Transform back the representation of $\tilde{f}$ to obtain a representation of f.

Remark 6.1. *Theorem 6.3 is an extension of Putinar's Positivstellensatz to (possibly) noncompact sets $S(\mathfrak{g})$, and so does not require the Archimedean condition. The price to pay for such an extension is the presence of the multiplier θ^{k_ε} in front of f and the perturbation term $\varepsilon \, \theta^d$. Note that (ii) involves a tighter bound for d, compared to (i), since $d_1(f) \geq d_2(f)$. The counterpart is that (ii) requires to include the additional constraint $f + \lambda \geq 0$, for some $\lambda \geq 0$.*

Complexity of Putinar–Vasilescu's Positivstellensatz. For each $q \in \mathbb{R}[\mathbf{x}]_{2r}$, we denote $\|q\|_{\max,r} := \max_{\boldsymbol{\alpha}} \{ |q_{\boldsymbol{\alpha}}| / c_{n+1}(\boldsymbol{\alpha}, 2r - |\boldsymbol{\alpha}|) \}$. Let us recall $c_n(\boldsymbol{\alpha}) := \frac{|\boldsymbol{\alpha}|!}{\alpha_1! \ldots \alpha_n!}$ for each $\boldsymbol{\alpha} \in \mathbb{N}^n$. By relying on Baldi–Mourrain's result (Theorem 1.3) after the homogenization trick, it is straightforward to analyze the complexity of Putinar–Vasilescu's Positivstellensatz:

Proposition 6.1. *Assume that all assumptions of Theorem 6.3 hold and $\mathbf{0}_{\mathbb{R}^n} \in S(\mathfrak{g})$. Then there exist real numbers $\mathfrak{c}_j > 0$ depending on $\mathfrak{g}$ such that for all $k_\varepsilon \in \mathbb{N}$ satisfying*

$$k_\varepsilon \geq \mathfrak{c}_1 d^{\mathfrak{c}_2} \left(\varepsilon^{-1} \|f\| \right)^{\mathfrak{c}_3} - d \, , \tag{6.1.3}$$

one has $\theta^{k_\varepsilon} \left(f + \varepsilon \, \theta^d \right) \in \mathcal{Q}_{k_\varepsilon + d}(\mathfrak{g})$.

The proof of Proposition 6.1 is similar to [130, Appendix].

Discussion about the ε parameter. The (arbitrary small) positive parameter ε in Theorem 6.2 and Theorem 6.3 ensures the positivity of polynomials over the respective considered domain $\mathbb{R}^n$ or $S(\mathfrak{g})$, excluding the origin in the homogenized representations. However these representations can still hold, even when $\varepsilon = 0$, as illustrated in the following two examples:

Example 6.1. *(i) Motzkin's polynomial $f = x_1^4 x_2^2 + x_1^2 x_2^4 + 1 - 3 x_1^2 x_2^2$ is globally nonnegative but not SOS. However, θf is SOS since*

$$\begin{aligned} \theta f = \ & 2(\tfrac{1}{2} x_1^3 x_2 + \tfrac{1}{2} x_1 x_2^3 - x_1 x_2)^2 + (x_1^2 x_2 - x_2)^2 + (x_1 x_2^2 - x_1)^2 \\ & + \tfrac{1}{2} (x_1^3 x_2 - x_1 x_2)^2 + \tfrac{1}{2} (x_1 x_2^3 - x_1 x_2)^2 + (x_1^2 x_2^2 - 1)^2 \, . \end{aligned}$$

(ii) Let $f = \left(x_1^2 + x_2^2 \right) x_1^2 x_2^2 - 3 x_1^2 x_2^2$ and $g = x_1^2 + x_2^2 - 4$. It is not hard to show that f is nonnegative on the noncompact set $S(\mathfrak{g})$. Moreover, $f = \frac{1}{4} x_1^2 x_2^2 \left(x_1^2 + x_2^2 \right) + \frac{3}{4} x_1^2 x_2^2 g$. Thus, $\theta^0 f \in \mathcal{Q}_3(\mathfrak{g})$.

However, the certificate (6.1.1) for global nonnegativity with $\varepsilon = 0$ is not true in general, as shown in the following lemma. This is due to the fact that any SOS multiplier for Delzell's polynomial must have a zero at the "bad point", so one can not find any globally positive multiplier.

Lemma 6.1. *The nonnegative dehomogenized Delzell's polynomial [43]:*

$$f = x_1^4x_2^2 + x_2^4x_3^2 + x_1^2x_3^4 - 3x_1^2x_2^2x_3^2 + x_3^8$$

satisfies that $\theta^k f \notin \Sigma[\mathbf{x}]$ for all $k \in \mathbb{N}$.

Proof. Assume by contradiction that $\theta^K f \in \Sigma[\mathbf{x}]$ for some $K \in \mathbb{N}$. Note that $n = 3$ here. We denote by $\tilde{f}$ the degree 8 homogenization of f, i.e.,

$$\tilde{f} = x_4^2(x_1^4x_2^2 + x_2^4x_3^2 + x_1^2x_3^4 - 3x_1^2x_2^2x_3^2) + x_3^8\,.$$

Then $\|(\mathbf{x}, x_{n+1})\|_2^{2K}\tilde{f} \in \Sigma[\mathbf{x}, x_{n+1}]$. As shown in [176, Section 6], it is impossible. This contradiction yields the conclusion. □

The certificate (6.1.2) for global nonnegativity on basic semialgebraic sets with $\varepsilon = 0$ is also not true in general, as shown in the following lemma:

Lemma 6.2. *With $n = 1$, let $f = x$ and $\mathfrak{g} = \{x^3, -x^3\}$. Then $f = 0$ on $S(\mathfrak{g}) = \{0\}$. It follows that f is nonnegative on $S(g)$, but:*
(i) $\theta^k f \notin \mathcal{Q}(\mathfrak{g})$ for all $k \in \mathbb{N}$ and;
(ii) for every $\varepsilon > 0$, $\theta^k(f + \varepsilon\theta) \in \mathcal{Q}_{k+1}(\mathfrak{g})$ for all $k \in \mathbb{N}$ with $k \geq \max\{2\,, \varepsilon^{-2}/4 - 1\}$.

Proof. We will show statement (i). Assume by contradiction that there exists $k \in \mathbb{N}$ such that $\theta^k f \in \mathcal{Q}(\mathfrak{g})$. Then there exists $q_j(x) \in \mathbb{R}[x]$, $j = 0, \ldots, r$ such that

$$\theta^k f = \sum_{j=1}^{m} q_j(x)^2 + q_0(x)x^3.$$

Assume that $q_j(x) = a_j + b_jx + x^2d_j(x)$, where $a_j, b_j \in \mathbb{R}$ and $d_j \in \mathbb{R}[x]$, $j \in [r]$. From this and since $\theta^k = 1 + x^2e(x)$ for some $e \in \mathbb{R}[\mathbf{x}]$, one has

$$(1 + x^2e(x))x = \sum_{j=1}^{r} a_j^2 + 2\sum_{j=1}^{r} a_jb_jx + x^2p(x)\,,$$

for some $q \in \mathbb{R}[x]$. By comparing coefficients of monomials 1 and x in the two sides of the above equality, $\sum\limits_{j=1}^{r} a_j^2 = 0$ and $2\sum\limits_{j=1}^{r} a_jb_j = 1$. It implies that $a_j = 0$, $j \in [r]$, and $2\sum\limits_{j=1}^{r} a_jb_j = 1$. It follows that $0 = 1$. It is impossible.

Let us prove the statement (ii). Let $\varepsilon > 0$ and $k \in \mathbb{N}$, $k \geq 2$. Since $\theta^k = 1 + kx^2 + x^4e(x)$ for some $e \in \mathbb{R}[x]_{2k-4}$, one has

$$\theta^k(f + \varepsilon\theta) = (1 + kx^2 + x^4e(x))(\varepsilon + x + \varepsilon x^2) = \varepsilon + x + \varepsilon(k+1)x^2 + x^3q(x)\,,$$

for some $q \in \mathbb{R}[x]_{2k-2}$. Assume that $k \geq \varepsilon^{-2}/4 - 1$. Then

$$\theta^k(f + \varepsilon\theta) = \varepsilon - \frac{1}{4\varepsilon(k+1)} + \left(x\sqrt{\varepsilon(k+1)} + \frac{1}{2\sqrt{\varepsilon(k+1)}}\right)^2 + x^3q(x) \in \mathcal{Q}_{k+1}(\mathfrak{g})\,.$$

□

From Lemma 6.1 and Lemma 6.2, we conclude that the strict positivity of the ε parameter is necessary in general although the certification with $\varepsilon = 0$ may happen in many cases.

When the certificate (6.1.1) with $\varepsilon = 0$ occurs, one has the following remark about the exponent of θ in (6.1.1).

Remark 6.2. *If $n = 2$, there does not exist a fixed $K \in \mathbb{N}$ such that for all nonnegative $f \in \mathbb{R}[\mathbf{x}]_6$, $\theta^K f$ is SOS. Indeed, assume by contradiction that there exists such a K. Then, the degree 6 homogenization $\tilde{f}$ of f would be a positive ternary sextic such that $\|(\mathbf{x}, x_{n+1})\|_2^{2K} \tilde{f}$ is SOS. By using [177, Theorem 1] and the fact that the homogeneous Motzkin's polynomial is a positive ternary sextic which is not SOS, we obtain a contradiction.*

In (6.1.1) with $\varepsilon = 0$, the multiplier θ^{k_ε} can be replaced with other SOS multipliers; see, e.g., Leep-Starr's polynomial [114, Example 2]. With the multiplier $\theta := x_1^2 + x_2^2 + x_3^2$ and the homogenized Delzell's polynomial D, Schabert provides in [182, Example 4.4] the exact SOS decomposition of the product $\Theta(x_1, x_2, x_3)D(x_1, x_2, x_3, x_4)$.

6.2 Polynomial optimization

In this section, we exploit the two representations from Theorem 6.2 and Theorem 6.3 to construct new hierarchies of semidefinite programs for POPs of the form $f^\star = \inf\{f(\mathbf{x}) : \mathbf{x} \in \Omega\}$ where $\Omega = \mathbb{R}^n$ for the unconstrained case and $\Omega = S(g)$ for the constrained case (with no compactness assumption), respectively. Instead of solving the original problem, we are rather interested in the perturbed problem:

$$f_\varepsilon^\star := \inf\{f(\mathbf{x}) + \varepsilon\,\theta(\mathbf{x})^d : \mathbf{x} \in \Omega\}, \tag{6.2.1}$$

where $\varepsilon > 0$ is fixed, $\theta(\mathbf{x}) := 1 + \|\mathbf{x}\|_2^2$, and $2d \geq \deg(f)$. Now, assume that the optimal value $f^\star$ of the original problem is attained at some $\mathbf{x}^\star \in \Omega$. It is not difficult to show that if Ω is unbounded, the polynomial $f + \varepsilon\,\theta^d$ is coercive on Ω, i.e.

$$\lim_{\mathbf{x}\in\Omega\,,\,\|\mathbf{x}\|_2\to\infty} \left(f(\mathbf{x}) + \varepsilon\theta(\mathbf{x})^d\right) = \infty,$$

(see more in [12]). Indeed, it is due to the fact that f is bounded from below by $f^\star$ on Ω and $\theta(\mathbf{x})^d \to \infty$ as $\|\mathbf{x}\|_2 \to \infty$. Thus, the optimal value $f_\varepsilon^\star$ of the perturbed problem (6.2.1) is always attained at some global minimizer $\mathbf{x}_\varepsilon^\star$ even if Ω is noncompact. Then:

$$\begin{aligned} f^\star + \varepsilon\,\theta(\mathbf{x}^\star)^d &= f(\mathbf{x}^\star) + \varepsilon\,\theta(\mathbf{x}^\star)^d \\ &\geq f_\varepsilon^\star = f(\mathbf{x}_\varepsilon^\star) + \varepsilon\,\theta(\mathbf{x}_\varepsilon^\star)^d \geq f(\mathbf{x}_\varepsilon^\star) \geq f^\star. \end{aligned}$$

Thus, $f^\star \in [f_\varepsilon^\star - \varepsilon\,\theta(\mathbf{x}^\star)^d, f_\varepsilon^\star]$, i.e., $f_\varepsilon^\star$ is a perturbation of $f^\star$ and the gap between both of them is at most $\varepsilon\,\theta(\mathbf{x}^\star)^d$. Next, observe that:

$$\begin{aligned} f_\varepsilon^\star &= \sup\{\lambda \in \mathbb{R} : f + \varepsilon\,\theta^d - \lambda \geq 0 \text{ on } \Omega\} \\ &= \sup\{\lambda \in \mathbb{R} : \theta^k(f + \varepsilon\,\theta^d - \lambda) \geq 0 \text{ on } \Omega\},\ k \in \mathbb{N}. \end{aligned}$$

The following hierarchies are based on the simple idea of replacing constraint "$\theta^k(f + \varepsilon\theta^d - \lambda) \geq 0$ on Ω" by relaxed constraint"$\theta^k(f + \varepsilon\theta^d - \lambda)$ is in the truncated quadratic module associated with Ω".

6.2.1 Unconstrained case

Given $f \in \mathbb{R}[\mathbf{x}]_{2d}$, let us consider the following problem:

$$f^\star := \inf_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}). \tag{6.2.2}$$

In the sequel, we assume that $f^\star > -\infty$ and let $\varepsilon > 0$ be fixed. Consider the hierarchy of semidefinite programs indexed by $k \in \mathbb{N}$:

$$\begin{aligned} \tau_k^1(\varepsilon) := \inf\ & L_{\mathbf{y}}(\theta^k(f + \varepsilon\theta^d)) \\ \text{s.t.}\ & \mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{2(d+k)}} \subset \mathbb{R}, \\ & \mathbf{M}_{k+d}(\mathbf{y}) \succeq 0, \\ & L_{\mathbf{y}}(\theta^k) = 1. \end{aligned} \tag{6.2.3}$$

Theorem 6.4. *For every $k \in \mathbb{N}$, the dual of* (6.2.3) *reads:*

$$\rho_k^1(\varepsilon) := \sup\{\lambda \in \mathbb{R} : \ \theta^k(f - \lambda + \varepsilon\theta^d) \in \Sigma[\mathbf{x}]_{k+d}\}\,. \tag{6.2.4}$$

The following statements hold:

1. *The sequence $(\rho_k^1(\varepsilon))_{k\in\mathbb{N}}$ is monotone non-decreasing.*
2. *Assume that $f^\star$ in* (6.2.2) *is attained at $\mathbf{x}^\star \in \mathbb{R}^n$. Then there exists $K \in \mathbb{N}$ such that $f^\star \le \rho_k^1(\varepsilon) \le f^\star + \varepsilon\,\theta(\mathbf{x}^\star)^d$ for all $k \ge K$. In particular, K is upper bounded by $\mathcal{O}(\varepsilon^{-1})$ as $\varepsilon \downarrow 0$.*

Proof. 1. Let $k \in \mathbb{N}$ and fix $\bar\varepsilon > 0$, arbitrary. By (6.2.4), there exists a real $\bar\lambda$ such that

$$\rho_k^1(\varepsilon) - \bar\varepsilon \le \bar\lambda \text{ and } \theta^k(f - \bar\lambda + \varepsilon\theta^d) \in \Sigma[\mathbf{x}]_{k+d}\,.$$

Since $\theta \in \Sigma[\mathbf{x}]_1$, $\theta^{k+1}(f - \bar\lambda + \varepsilon\theta^d) \in \Sigma[\mathbf{x}]_{k+d+1}$. By (6.2.4), $\rho_{k+1}^1(\varepsilon) \ge \bar\lambda \ge \rho_k^1(\varepsilon) - \bar\varepsilon$. This implies that $\rho_{k+1}^1(\varepsilon) \ge \rho_k^1(\varepsilon)$.

2. By (6.2.2), $f - f^\star$ is nonnegative. By Theorem 6.2, there exists $K \in \mathbb{N}$ and $K = O(\varepsilon^{-1})$ as $\varepsilon \downarrow 0$ such that

$$\theta^K(f - f^\star + \varepsilon\theta^d) \in \Sigma[\mathbf{x}]_{K+d}\,.$$

Let $k \ge K$ be fixed. Since $\theta \in \Sigma[\mathbf{x}]_1$, one has

$$\theta^k(f - f^\star + \varepsilon\theta^d) = \theta^{K+(k-K)}(f - f^\star + \varepsilon\theta^d) \in \Sigma[\mathbf{x}]_{k+d}\,.$$

By (6.2.4), $f^\star \le \rho_k^1(\varepsilon)$. Thus, $f^\star \le \rho_k^1(\varepsilon)$ for all $k \ge K$. Let $k \in \mathbb{N}$ and fix $\bar\varepsilon > 0$, arbitrary. By (6.2.4), there exists a real $\bar\lambda$ such that

$$\rho_k^1(\varepsilon) - \bar\varepsilon \le \bar\lambda \text{ and } \theta^k(f - \bar\lambda + \varepsilon\theta^d) \in \Sigma[\mathbf{x}]_{k+d}\,.$$

It follows that $f - \bar\lambda + \varepsilon\theta^d \ge 0$ on $\mathbb{R}^n$. From this,

$$f^\star + \varepsilon\theta(\mathbf{x}^\star)^d = f(\mathbf{x}^\star) + \varepsilon\theta(\mathbf{x}^\star)^d \ge \bar\lambda \ge \rho_k^1 - \bar\varepsilon\,.$$

This implies $f^\star + \varepsilon\theta(\mathbf{x}^\star)^d \ge \rho_k^1(\varepsilon)$, the desired result.

□

We guarantee strong duality for previous primal-dual problems:

Proposition 6.2. *Let $k \in \mathbb{N}$. Then $\tau_k^1(\varepsilon) = \rho_k^1(\varepsilon)$. Moreover, if $\tau_k^1(\varepsilon) > -\infty$ then the optimal value $\rho_k^1(\varepsilon)$ is attained.*

Proof. By Slater's constraint qualification [25, Section 5.2.3], it suffices to show that (6.2.3) admits a strictly feasible solution. Let us denote by μ the measure with density $\chi_{[0,1]^n}\theta^{-k}$ with respect to the Lebesgue measure, where χ_A is the characteristic function of a given set $A \subset \mathbb{R}^n$. Set $y_{\boldsymbol{\alpha}} := \int \mathbf{x}^{\boldsymbol{\alpha}} d\mu$ for all $\boldsymbol{\alpha} \in \mathbb{N}^n$. We claim that $y_{\boldsymbol{\alpha}} \in \mathbb{R}$ for all $\boldsymbol{\alpha} \in \mathbb{N}^n$, $L_{\mathbf{y}}(\theta^k) = 1$ and $\mathbf{M}_{k+d}(\mathbf{y}) \succ 0$. Indeed, for all $\boldsymbol{\alpha} \in \mathbb{N}^n$

$$\begin{aligned} |y_{\boldsymbol{\alpha}}| &= |\textstyle\int \mathbf{x}^{\boldsymbol{\alpha}}\chi_{[0,1]^n}\theta^{-k}d\mathbf{x}| = |\int_{[0,1]^n} \mathbf{x}^{\boldsymbol{\alpha}}\theta^{-k}d\mathbf{x}| \\ &\le \textstyle\int_{[0,1]^n} |x_1|^{\alpha_1}\ldots|x_n|^{\alpha_n}\theta^{-k}d\mathbf{x} \le 1\,, \end{aligned}$$

since $\theta^{-k} \le 1$. Thus, $y_{\boldsymbol{\alpha}} \in \mathbb{R}$ for all $\boldsymbol{\alpha} \in \mathbb{N}^n$. In addition,

$$L_{\mathbf{y}}(\theta^k) = \int \theta^k\chi_{[0,1]^n}\theta^{-k}d\mathbf{x} = \int_{[0,1]^n} d\mathbf{x} = 1\,.$$

Let $\mathbf{p} \in \mathbb{R}^{s(d+k)}\backslash\{\mathbf{0}\}$ be fixed. We state that $\mathbf{p}^\top\mathbf{M}_{d+k}(\mathbf{y})\mathbf{p} > 0$. Assume by contradiction that $\mathbf{p}^\top\mathbf{M}_{d+k}(\mathbf{y})\mathbf{p} \le 0$. One has

$$\begin{aligned} 0 &\ge \mathbf{p}^\top\mathbf{M}_{d+k}(\mathbf{y})\mathbf{p} = \textstyle\int \mathbf{p}^\top\mathbf{v}_{d+k}\mathbf{v}_{d+k}^\top\mathbf{p}d\mu \\ &= \textstyle\int \mathbf{p}^\top\mathbf{v}_{d+k}\mathbf{v}_{d+k}^\top\mathbf{p}\chi_{[0,1]^n}\theta^{-k}d\mathbf{x} = \int_{[0,1]^n} \left(\mathbf{p}^\top\mathbf{v}_{d+k}\right)^2\theta^{-k}d\mathbf{x} \ge 0\,. \end{aligned}$$

It follows that $\mathbf{p}^\top\mathbf{v}_{d+k} = 0$ on $[0,1]^n$, thus $\mathbf{p} = \mathbf{0}$ yielding a contradiction. From this, $(y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{d+k}}$ is a feasible solution of (6.2.3) with $\mathbf{M}_{k+d}(\mathbf{y}) \succ 0$. By strong duality, the conclusion follows. □

6.2.2 Constrained case

Consider the following problem:

$$f^\star := \inf_{\mathbf{x}\in S(\mathfrak{g})} f(\mathbf{x})\,, \tag{6.2.5}$$

where $f \in \mathbb{R}[\mathbf{x}]$, $\mathfrak{g} = \{g_1, \ldots, g_m\} \subset \mathbb{R}[\mathbf{x}]$. Assume that $S(\mathfrak{g}) \neq \emptyset$ and $f^\star > -\infty$. Let $\varepsilon > 0$ be fixed.

Unknown lower bound

Let $d := \lfloor \deg(f)/2 \rfloor + 1$ and consider the hierarchy of semidefinite programs indexed by $k \in \mathbb{N}$:

$$\begin{array}{rll} \tau_k^2(\varepsilon) := \inf & L_{\mathbf{y}}(\theta^k(f+\varepsilon\theta^d)) & \\ \text{s.t.} & \mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{2(d+k)}} \subset \mathbb{R}\,, & \\ & \mathbf{M}_{k+d}(\mathbf{y}) \succeq 0\,, & \\ & \mathbf{M}_{k+d-\lceil g_i\rceil}(g_i\mathbf{y}) \succeq 0,\ i \in [m]\,, & \\ & L_{\mathbf{y}}(\theta^k) = 1\,. & \end{array} \tag{6.2.6}$$

Theorem 6.5. *For every $k \in \mathbb{N}$, the dual of* (6.2.6) *reads:*

$$\rho_k^2(\varepsilon) := \sup\{\lambda \in \mathbb{R} :\ \theta^k\,(f - \lambda + \varepsilon\,\theta^d) \in \mathcal{Q}_{k+d}(\mathfrak{g})\}\,. \tag{6.2.7}$$

The following statements hold:

1. *The sequence $(\rho_k^2(\varepsilon))_{k\in\mathbb{N}}$ is monotone non-decreasing.*
2. *Assume that problem* (6.2.5) *has an optimal solution $\mathbf{x}^\star$. Then there exists $K \in \mathbb{N}$ such that $f^\star \leq \rho_k^2(\varepsilon) \leq f^\star + \varepsilon\,\theta(\mathbf{x}^\star)^d$ for all $k \geq K$. In particular, K is upper bounded by $\mathcal{O}(\varepsilon^{-\mathfrak{c}}) - d$ as $\varepsilon \downarrow 0$, for some $\mathfrak{c} > 0$ depending on $\mathfrak{g}$.*

The proof of Theorem 6.5 relies on Theorem 6.3 (i) and is similar to Theorem 6.4. The upper bound on K is based on Proposition 6.1.

We guarantee strong duality for previous primal-dual problems:

Proposition 6.3. *There exists $K \in \mathbb{N}$ such that $\tau_k^2(\varepsilon) = \rho_k^2(\varepsilon)$ for all $k \geq K$. Moreover, if $\tau_k^2(\varepsilon) > -\infty$, the optimal value $\rho_k^2(\varepsilon)$ is attained.*

The proof of Proposition 6.3 can be found in [130, Appendix].

Remark 6.3. *If $S(\mathfrak{g})$ has nonempty interior then strong duality holds for all orders k of the primal-dual problems* (6.2.6)-(6.2.7). *Indeed, by constructing a sequence of moments from the Lebesgue measure on an open ball contained in $S(\mathfrak{g})$, one can find a strictly feasible solution of* (6.2.6) *and then apply Slater's constraint qualification [25, Section 5.2.3].*

Known lower bound

Assume that $g_m := f - \underline{f}$ for some real $\underline{f} \leq f^\star$ and let $d := \lceil \deg(f)/2 \rceil$. We then obtain the same conclusion as Theorem 6.5 with replacing here $\tau_k^2(\varepsilon)$ and $\rho_k^2(\varepsilon)$ by $\tau_k^3(\varepsilon)$ and $\rho_k^3(\varepsilon)$, respectively. The proof relies on Theorem 6.3 (ii) and is similar to Theorem 6.4. Note that here $g_m = (f - f^\star) + (f^\star - \underline{f})$ with $f - f^\star \geq 0$ on $S(\mathfrak{g})$ and $f^\star - \underline{f} \geq 0$. The upper bound on K is also based on Proposition 6.1.

The next proposition states that strong duality is guaranteed for each relaxation order k.

Proposition 6.4. *Let $k \in \mathbb{N}$. Then $\tau_k^3(\varepsilon) = \rho_k^3(\varepsilon)$. Moreover, if $\tau_k^3(\varepsilon) > -\infty$ then the optimal value $\rho_k^3(\varepsilon)$ is attained.*

The proof of Proposition 6.4 can be found in [130, Appendix].

Remark 6.4. *A lower bound $\underline{f}$ of problem* (6.2.5) *can be obtained by solving the following SDP:*

$$\sup\{\lambda \in \mathbb{R} : \ \theta^k\,(f-\lambda) \in \mathcal{Q}_{k+d}(\mathfrak{g})\}\,,\ k \in \mathbb{N}\,.$$

Assume that we know a lower bound $\underline{f}$ of problem (6.2.5)*. By adding the inequality constraint $f - \underline{f} \geq 0$ in $S(\mathfrak{g})$, we obtain the same semialgebraic set, i.e. $S(\mathfrak{g} \cup \{f - \underline{f}\}) = S(\mathfrak{g})$. Thus, the two problems $\inf\{f(\mathbf{x}) : x \in S(\mathfrak{g})\}$ and $\inf\{f(\mathbf{x}) : x \in S(\mathfrak{g} \cup \{f - \underline{f}\})\}$ are identical and the primal-dual SDP relaxations with values $\rho_k^3(\varepsilon)$ and $\tau_k^3(\varepsilon)$ of the latter one satisfy strong duality for each relaxation order k.*

General case. Since $g_i \geq 0$ on $S(\mathfrak{g})$ and $-g_i \geq 0$ on $S(\mathfrak{g})$ is equivalent to $g_i = 0$ on $S(\mathfrak{g})$, $S(\mathfrak{g})$ can be rewritten as $S(\mathfrak{g}, \mathfrak{h})$ with $\mathfrak{g} = \{g_1, \ldots, g_m\}$ is the set of polynomials involved in the inequality constraints and $\mathfrak{h} = \{h_1, \ldots, h_l\}$ is the set of polynomials involved in the equality constraints; in addition, $\mathbb{R}^n = S(\{\mathbf{0}\}, \emptyset)$. Consider the general POP:

$$f^\star := \inf_{\mathbf{x} \in S(\mathfrak{g},\mathfrak{h})} f(\mathbf{x})\,, \tag{6.2.8}$$

with $f^\star \in \mathbb{R}$ and define

$$(d,i) = \begin{cases} (\lceil \deg(f)/2 \rceil, 1) & \text{if } S(\mathfrak{g},\mathfrak{h}) = \mathbb{R}^n\,, \\ (1 + \lfloor \deg(f)/2 \rfloor, 2) & \text{if } S(\mathfrak{g},\mathfrak{h}) \neq \mathbb{R}^n \text{ and lower bound } \underline{f} \text{ is unknown}\,, \\ (\lceil \deg(f)/2 \rceil, 3) & \text{otherwise and set } g_m := f - \underline{f}\,. \end{cases}$$

For fixed $\varepsilon > 0$, one considers the following SDP relaxation of POP (6.2.8)

$$\begin{array}{rll} \tau_k^i(\varepsilon) := \inf & L_{\mathbf{y}}(\theta^k(f + \varepsilon\theta^d)) & \\ \text{s.t.} & \mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in \mathbb{N}^n_{2(d+k)}} \subset \mathbb{R}\,, & \\ & \mathbf{M}_{k+d-\lceil g_i \rceil}(g_i \mathbf{y}) \succeq 0\,,\ i = 0, \ldots, m\,, & \\ & \mathbf{M}_{k+d-\lceil h_j \rceil}\,(h_j \mathbf{y}) = 0\,,\ j \in [l]\,, & \\ & L_{\mathbf{y}}(\theta^k) = 1\,, & \end{array} \tag{6.2.9}$$

where $g_0 := 1$. Its dual is the semidefinite program:

$$\rho_k^i(\varepsilon) := \sup\{\lambda \in \mathbb{R} : \ \theta^k(f - \lambda + \varepsilon\theta^d) \in \mathcal{Q}_{k+d}(\mathfrak{g}, \mathfrak{h})\} \tag{6.2.10}$$

The zero-duality gap between SDP (6.2.9) and SDP (6.2.10) is guaranteed for large enough k.

Remark 6.5. *The condition $\tau_k^i(\varepsilon) > -\infty$ is always satisfied whenever k is sufficently large. Indeed by weak duality, when ε is fixed and k is sufficiently large then $\tau_k^i(\varepsilon) \geq \rho_k^i(\varepsilon) \geq f^\star > -\infty$. However, when k is small, $\tau_k^i(\varepsilon) = -\infty$ may happen.*

Let us now assume that the POP (6.2.8) has an optimal solution $\mathbf{x}^\star$. Then $\rho_k^i(\varepsilon) \in [f^\star, f^\star + \varepsilon\theta(\mathbf{x}^\star)^d]$ when $\varepsilon > 0$ is fixed and k is sufficiently large. Moreover, the gap between $\rho_k^i(\varepsilon)$, and $f^\star$ is at most $\varepsilon\,\theta(\mathbf{x}^\star)^d$. Therefore, $\rho_k^i(\varepsilon)$ is indeed an approximation of $f^\star$. In practice, $(\rho_k^i(\varepsilon))_{k \in \mathbb{N}}$ often converges to the optimal value $f_\varepsilon^\star := \min\{f(\mathbf{x}) + \varepsilon\,\theta(\mathbf{x})^d : \mathbf{x} \in S(\mathfrak{g}, \mathfrak{h})\}$ after finitely many steps (see Section 6.3).

Remark 6.6. *(i) The term θ^d in both* (6.2.9) *and* (6.2.10) *can be replaced by $\varphi_d(\mathbf{x}, 1)$ where $\varphi_d : \mathbb{R}^{n+1} \to \mathbb{R}$ is a positive form of degree $2d$. For instant, one can select $\varphi_d(\mathbf{x}, 1) = x_1^{2d} + \cdots + x_n^{2d} + 1$. (ii) Let $r \in \mathbb{N}$ be fixed. For every k divisible by $2r$, the term θ^k appearing in both* (6.2.9) *and* (6.2.10) *can be replaced by $\psi_r(\mathbf{x}, 1)^{k/(2r)}$ where $\psi_r : \mathbb{R}^{n+1} \to \mathbb{R}$ is a coercive positive form of degree $2r$. For instant, one can select $\psi_r(\mathbf{x}, 1) = x_1^{2r} + \cdots + x_n^{2r} + 1$.*

Relation between classical optimality conditions and nonnegativity certificates. When $\rho_k^i(0) = f^\star$, the constraint qualification conditions hold at $\mathbf{x}^\star$.

Proposition 6.5. *Assume that $\rho_k^i(0) = f^\star$ for some $k \in \mathbb{N}$, i.e., there exists $\sigma_i \in \Sigma[\mathbf{x}]$, $i = 0, \ldots, m$ and $\phi_j \in \mathbb{R}[\mathbf{x}]$, $j \in [l]$ such that $\theta^k(f - f^\star) = \sigma_0 + \sum_{i=1}^m \sigma_i g_i + \sum_{j=1}^l \phi_j h_j$. Then the constraint qualification conditions hold at $\mathbf{x}^\star$:*

1. $\sigma_i(\mathbf{x}^\star) \geq 0$ *and* $g_i(\mathbf{x}^\star) \geq 0$, *for all* $i \in [m]$;

2. $\sigma_i(\mathbf{x}^\star) g_i(\mathbf{x}^\star) = 0$, *for all* $i \in [m]$;

3. $\theta(\mathbf{x}^\star)^k \nabla f(\mathbf{x}^\star) = \sum_{i=1}^{m} \sigma_i(\mathbf{x}^\star) \nabla g_i(\mathbf{x}^\star) + \sum_{j=1}^{l} \phi_j(\mathbf{x}^\star) \nabla h_j(\mathbf{x}^\star)$.

The proof of Proposition 6.5 is similar to [105, Theorem 7.4].

If we take an arbitrary small $\varepsilon > 0$ then $\rho_k^i(\varepsilon)$ is arbitrary close to $f^\star$ for large enough k. However, if one sets $\varepsilon = 0$, the statement "$\rho_K^i(0) = f^\star$ for some $K \in \mathbb{N}$" is not true in general as stated in the following proposition:

Proposition 6.6. *If the first order optimality condition fails at a global minimizer of problem* (6.2.8)*, then* $\rho_k^i(0) < f^\star$ *for all* $k \in \mathbb{N}$.

The proof of Proposition 6.6 is similar to [147, Proposition 3.4].

By applying Proposition 6.6 to POP $\min\{x : x^3 = 0\}$, we obtain the statement (i) of Lemma 6.2. Indeed, the first order optimality condition fails at the global minimizer 0 of this problem. Therefore, the positivity of ε ensures convergence of $(\rho_k^i(\varepsilon))_{k\in\mathbb{N}}$ to the neighborhood $[f^\star, f^\star + \varepsilon\theta(\mathbf{x}^\star)^d]$ of the optimal value $f^\star$. As proved by Huang, Nie and Yuan in [83], $\rho_K^i(0) = f^\star$ for some $K \in \mathbb{N}$ when some optimality conditions hold at every global minimizer of (6.2.8). In many cases, $\rho_K^i(0) = f^\star$ with $K = 0, 1$ when the KKT conditions hold (see Example 6.1 and [146, Example 4.4] with $x_n = 1$). However, the KKT conditions are not enough for this conjecture due to the fact that the minimizer $\mathbf{x}^\star = (0, 0, 0)$ of dehomogenized Delzell's polynomial in Lemma 6.1 satisfies the KKT conditions and $\rho_k^i(0) < f^\star$ for all $k \in \mathbb{N}$ in this case.

Reducing the noncompact case to a compact case. Consider the POP: $f^\star := \inf\{f(\mathbf{x}) : \mathbf{x} \in S(\mathfrak{g}, \mathfrak{h})\}$ where the feasible set $S(\mathfrak{g}, \mathfrak{h})$ is possibly noncompact, and the associated perturbed POP: $f_\varepsilon^\star := \inf\{f(\mathbf{x}) + \varepsilon\theta(\mathbf{x})^d : \mathbf{x} \in S(\mathfrak{g}, \mathfrak{h})\}$ with fixed $\varepsilon > 0$. Here one assumes that $f^\star$ is attained at $\mathbf{x}^\star$ and $2d > \deg(f)$. As in Section 6.2, $f^\star \in [f_\varepsilon^\star - \varepsilon\,\theta(\mathbf{x}^\star)^d, f_\varepsilon^\star]$. Suppose that a point $\bar{\mathbf{x}}$ in $S(\mathfrak{g}, \mathfrak{h})$ is known. It is not hard to show that $f + \varepsilon\theta^d$ is coercive and therefore with $C := f(\bar{\mathbf{x}}) + \varepsilon\theta(\bar{\mathbf{x}})^d$, the set $S(\{C - f - \varepsilon\theta^d\})$ is compact. Moreover,

$$f_\varepsilon^\star = \inf\{f(\mathbf{x}) + \varepsilon\theta(\mathbf{x})^d : x \in S(\mathfrak{g} \cup \{C - f - \varepsilon\theta^d\}, \mathfrak{h})\}\,. \tag{6.2.11}$$

Note that the quadratic module associated with the constraint set of POP (6.2.11) is Archimedean and so $f_\varepsilon^\star$ can be approximated as close as desired by the Moment-SOS hierarchy. This approach is similar in spirit to that of [88]. However, determining a point $\bar{\mathbf{x}}$ in $S(\mathfrak{g}, \mathfrak{h})$ is not easy in general. The hierarchy (6.2.10) relying on Putinar–Vasilescu's Positivstellensatz goes beyond this restriction.

In a different way, if one relies on the big ball trick, we consider the following POP, for all $\varepsilon > 0$:

$$\hat{f}_\varepsilon^* := \inf\{f(\mathbf{x}) : x \in S(\mathfrak{g} \cup \{1 - \varepsilon\|\mathbf{x}\|_2^2\}, \mathfrak{h})\}\,. \tag{6.2.12}$$

Obviously, one has $\hat{f}_\varepsilon^* \downarrow f^\star$ as $\varepsilon \downarrow 0$. The quadratic module associated with the constraint set of POP (6.2.12) is Archimedean and $f_\varepsilon^\star$ can be approximated by the Moment-SOS hierarchy:

$$\rho_k(\varepsilon) := \sup\{\lambda \in \mathbb{R} : f - \lambda \in \mathcal{Q}(\mathfrak{g} \cup \{1 - \varepsilon\|\mathbf{x}\|_2^2\}, \mathfrak{h})\}\,, k \in \mathbb{N}\,, \varepsilon > 0\,. \tag{6.2.13}$$

Thus the SOS weight of $1-\varepsilon\|\mathbf{x}\|_2^2$ appears in the SOS decomposition of $f-\lambda$. However, the hierarchy (6.2.10) relying on Putinar–Vasilescu's Positivstellensatz does not require such SOS weight in the decomposition of $f - \lambda$, but requires a perturbation $\varepsilon\theta^d$ and a multiplier θ^k.

6.2.3 Global optimizers

In this section we introduce a new method to find an approximation of a feasible point of a basic semialgebraic set $S(\mathfrak{g}, \mathfrak{h})$ as defined in (1.1.1). We then apply this method to obtain an approximation of a global minimizer $\mathbf{x}^\star$ associated to $f^\star = \min\{f(\mathbf{x}) : \mathbf{x} \in S(\mathfrak{g}, \mathfrak{h})\}$ via finding a feasible solution of $S(\mathfrak{g} \cup \{\rho_k^i(\varepsilon) - f\}, \mathfrak{h})$.

Remark 6.7. *Let $\varepsilon > 0$ be fixed and $k \in \mathbb{N}$ be sufficiently large such that $\rho_k^i(\varepsilon)$is an upper bound of $f^\star$. Let $\mathbf{x}^\star$ be a global minimizer of f on $S(\mathfrak{g}, \mathfrak{h})$ and let $\bar{\mathbf{x}} \in S(\mathfrak{g} \cup \{\rho_k^i(\varepsilon) - f\}, \mathfrak{h})$. Then $\bar{\mathbf{x}} \in S(\mathfrak{g}, \mathfrak{h})$ and $f^\star \leq f(\bar{\mathbf{x}}) \leq \rho_k^i(\varepsilon) \leq f^\star + \varepsilon\theta(\mathbf{x}^\star)^d$.*

Let us consider an arbitrary small $\varepsilon > 0$. The difference between $\rho_k^i(\varepsilon)$ and $f^\star$ will be as close as desired to $\varepsilon\theta(\mathbf{x}^\star)^d$ for large enough k. Assume that the solution set $S(\mathfrak{g} \cup \{f^\star - f\}, \mathfrak{h})$ is finite and denote by $\mathbf{y}_\varepsilon^\star$ an optimal solution of SDP (6.2.9). In practice, when k is sufficiently large, $\mathbf{y}_\varepsilon^\star$ satisfies numerically the flat extension condition defined in Section 2.5. One may then use the algorithm of Henrion and Lasserre [77] to extract numerically the support of a representing measure for $\mathbf{y}_\varepsilon^\star$ which may include global minimizers of $f^\star = \min\{f(\mathbf{x}) : \mathbf{x} \in S(\mathfrak{g}, \mathfrak{h})\}$ (see the same extraction in [89, Section 3.2]). However we cannot guarantee the success of this extraction procedure in theory because the set $S(\mathfrak{g} \cup \{\rho_k^i(\varepsilon) - f\}, \mathfrak{h})$ may not be zero dimensional when $\rho_k^i(\varepsilon) > f^\star$. For example, if $f = \|\mathbf{x}\|_2^2$ and $S(\mathfrak{g}, \mathfrak{h}) = \mathbb{R}^n$, $S(\mathfrak{g} \cup \{\rho_k^i(\varepsilon) - f\}, \mathfrak{h})$ is a closed ball centered at the origin with radius $\rho_k^i(\varepsilon)^{1/2}$. The following method aims at overcoming this issue from both theoretical and algorithmic sides. For further application cases of the flat moment criterion, we refer the interested reader to the framework from [171], which is based on altering the bottom right part of the moment matrix.

The Adding-Spherical-Constraints method (ASC): For $\mathbf{a} \in \mathbb{R}^n$ and $r \geq 0$, let $\overline{B(\mathbf{a}, r)}$ (resp. $\partial B(\mathbf{a}, r)$) be the closed ball (resp. sphere) centered at $\mathbf{a}$ with radius r, i.e.,

$$\overline{B(\mathbf{a}, r)} = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{a}\|_2 \leq r\} \quad (\text{resp.} \quad \partial B(\mathbf{a}, r) = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{a}\|_2 = r\}) .$$

The following result provides an efficient way to find a sequence of additional spherical equality constraints for a given semialgebraic set such that (i) the resulting set is a singleton (i.e. it contains a *single* real point), and (ii) this point is solution of a non-singular system of *linear equations*.

Lemma 6.3. *Assume that $S(\mathfrak{g}, \mathfrak{h}) \neq \emptyset$. Let $(\mathbf{a}_t)_{t=0,1,\ldots,n} \subset \mathbb{R}^n$ such that $\mathbf{a}_t - \mathbf{a}_0$, $t \in [n]$ are linearly independent in $\mathbb{R}^n$. Let us define the sequence $(\xi_t)_{t=0,1,\ldots,n} \subset \mathbb{R}_+$ as follows:*

$$\begin{cases} \xi_0 := \min\{\|\mathbf{x} - \mathbf{a_0}\|_2^2 : \mathbf{x} \in S(\mathfrak{g}, \mathfrak{h})\} , \\ \xi_t := \min\{\|\mathbf{x} - \mathbf{a_t}\|_2^2 : \mathbf{x} \in S(\mathfrak{g}, \mathfrak{h} \cup \{\xi_j - \|\mathbf{x} - \mathbf{a_j}\|_2^2 : j = 0, \ldots, t-1\})\} , \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad t \in [n] . \end{cases} \tag{6.2.14}$$

Then there exists a unique real point $\mathbf{x}^\star$ in $S(\mathfrak{g}, \mathfrak{h} \cup \{\xi_t - \|\mathbf{x} - \mathbf{a}_t\|_2^2 : t = 0, \ldots, n\})$ which satisfies the non-singular linear system of equations

$$(\mathbf{a}_t - \mathbf{a}_0)^\top \mathbf{x}^\star = -\frac{1}{2}(\xi_t - \xi_0 - \|\mathbf{a}_t\|_2^2 + \|\mathbf{a}_0\|_2^2), \ t \in [n] . \tag{6.2.15}$$

The proof of Lemma 6.3 is available in [130, Appendix]. Geometrically speaking, we find a

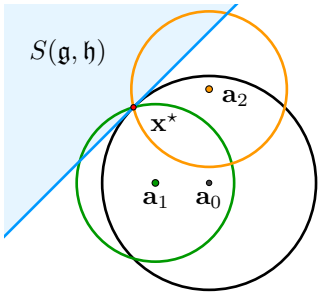


Figure 6.1: Illustration of Lemma 6.3.

sequence of spheres $\partial B(\mathbf{a}_t, \xi_t^{1/2})$, $t = 0, \ldots, n$, such that the intersection between these spheres and $S(\mathfrak{g}, \mathfrak{h})$ is the singleton $\{\mathbf{x}^\star\}$ (see Figure 6.1). Next, we use Lasserre's hierarchy to compute the optimal values ξ_t, $t = 0, \ldots, n$ of problem (6.2.14).

Theorem 6.6. *Assume that $S(\mathfrak{g},\mathfrak{h})\cap\overline{B(\mathbf{0},R^{1/2})}\neq\emptyset$ for some $R>0$. Let $(\mathbf{a}_t)_{t=0,1,\ldots,n}\subset\mathbb{R}^n$ such that $\mathbf{a}_t-\mathbf{a}_0,\ t\in[n]$ are linearly independent in $\mathbb{R}^n$. Assume that the Moment-SOS hierarchies associated with the following POPs:*

$$\begin{cases}\xi_0:=\min\{\|\mathbf{x}-\mathbf{a}_0\|_2^2:\ \mathbf{x}\in S(\mathfrak{g}\cup\{R-\|\mathbf{x}\|_2^2\},\mathfrak{h})\},\\ \xi_t:=\min\{\|\mathbf{x}-\mathbf{a}_t\|_2^2:\ \mathbf{x}\in S(\mathfrak{g},\mathfrak{h}\cup\{\xi_j-\|\mathbf{x}-\mathbf{a}_j\|_2^2:\ j=0,\ldots,t-1\})\},\\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad t\in[n],\end{cases}\tag{6.2.16}$$

have finite convergence, and let $w:=\max\{\lceil g_i\rceil,\lceil h_j\rceil,1\}$. For every $k\in\mathbb{N}$, consider the following semidefinite programs:

$$\begin{cases}\eta_k^0:=\inf\limits_{\mathbf{y}\in\mathbb{R}^{s(2(k+w))}} & L_{\mathbf{y}}(\|\mathbf{x}-\mathbf{a}_0\|_2^2)\\ \quad s.t. & \mathbf{M}_{k+w-\lceil g_i\rceil}(g_i\mathbf{y})\succeq 0,\\ & \mathbf{M}_{k+w-1}((R-\|\mathbf{x}\|_2^2)\mathbf{y})\succeq 0,\\ & \mathbf{M}_{k+w-\lceil h_j\rceil}(h_j\mathbf{y})=0,\\ & y_{\mathbf{0}}=1,\\ \eta_k^t:=\inf\limits_{\mathbf{y}\in\mathbb{R}^{s(2(k+w))}} & L_{\mathbf{y}}(\|\mathbf{x}-\mathbf{a}_t\|_2^2)\\ \quad s.t. & \mathbf{M}_{k+w-\lceil g_i\rceil}(g_i\mathbf{y})\succeq 0,\\ & \mathbf{M}_{k+w-\lceil h_j\rceil}(h_j\mathbf{y})=0,\\ & \mathbf{M}_{k+w-1}((\eta_k^j-\|\mathbf{x}-\mathbf{a}_j\|_2^2)\mathbf{y})=0,\ j=0,\ldots,t-1,\\ & y_{\mathbf{0}}=1,\\ & \qquad\qquad t\in[n].\end{cases}\tag{6.2.17}$$

Then there exists $K\in\mathbb{N}$ such that for all $k\geq K$, $\eta_k^t=\xi_t$, $t=0,\ldots,n$. Moreover, there exist $t\in\{0,\ldots,n\}$ and $\tilde{K}\in\mathbb{N}$ such that for all $k\geq\tilde{K}$, the solution $\mathbf{y}$ of SDP (6.2.17) with value η_k^t satisfies the flat extension condition, i.e., rank$(\mathbf{M}_{k+w}(\mathbf{y}))=$ rank$(\mathbf{M}_k(\mathbf{y}))$. In addition, $\mathbf{y}$ has a representing rank$(\mathbf{M}_k(\mathbf{y}))$-atomic measure μ and supp$(\mu)\subset S(\mathfrak{g},\mathfrak{h})$.

The proof of Theorem 6.6 is available in [130, Appendix].

Remark 6.8. *The Moment-SOS hierarchy of each POP* (6.2.16) *has finite convergence when one of the following conditions is satisfied:*

1. *(Lasserre [105, Theorem 7.5]) The ideal $\mathcal{I}(\mathfrak{h})$ is real radical, and the second-order sufficient condition holds at every global minimizer of each POP in* (6.2.16).
2. *(Lasserre et al. [108, Proposition 1.1] and [105, Theorem 6.13]) The real variety $V(\mathfrak{h})$ $(=S(\emptyset,\mathfrak{h}))$ is finite.*

Remark 6.9. *In the final conclusion of Theorem 6.6, when $\mathbf{y}$ has a representing* rank$(\mathbf{M}_k(\mathbf{y}))$*-atomic measure μ, we may use the extraction algorithm from [77] to obtain the atomic support of μ.*

Based on Theorem 6.6, Algorithm 6.2.3 below finds a feasible point in a nonempty (possibly noncompact) semialgebaric set $S(\mathfrak{g},\mathfrak{h})$.

Proposition 6.7. *Let the assumptions of Theorem 6.6 hold. For k sufficiently large, Algorithm 6.2.3 terminates and $\bar{\mathbf{x}}\in S(\mathfrak{g},\mathfrak{h})$.*

Proof. The proof follows from Theorem 6.6 and Remark 6.9. □

In Algorithm 6.2.3, step 1 computes the radius $L^{1/2}$ of the ball $\overline{B(\mathbf{0},L^{1/2})}$ which has non-empty intersection with $S(\mathfrak{g},\mathfrak{h})$. Then step 2 checks the flat extension condition and extracts the solution $\bar{\mathbf{x}}$.

Remark 6.10. *At step 2 in Algorithm 6.2.3, for k sufficiently large, the rank of the moment matrix* rank$(\mathbf{M}_{k+w}(\mathbf{y}))$ *decreases to one when t goes from 0 to n. Indeed, for each t between 0 and n, we replace the semialgebraic set $S(\mathfrak{g},\mathfrak{h})$ by its intersection with the t spheres $\partial B(\mathbf{a}_j,\xi_j^{1/2})$, $j=0,\ldots,t-1$. This intersection includes the support of the measure with moments $\mathbf{y}$. Since $S(\mathfrak{g},\mathfrak{h})\cap\bigcap_{j=0}^n\partial B(\mathbf{a}_j,\xi_j^{1/2})=\{\mathbf{x}^\star\}$, this support converges to $\{\mathbf{x}^\star\}$ when t goes from 0 to n. Thus for large enough k, the solution $\mathbf{y}$ of SDP* (6.2.17) *with value η_k^n has a representing measure supported on $\mathbf{x}^\star=(y_{\mathbf{e}_1},\ldots,y_{\mathbf{e}_n})$. Here $\mathbf{e}_i$, $i\in[n]$ is canonical basis of $\mathbb{R}^n$.*

Algorithm 12 PolySys

Input: $S(\mathfrak{g},\mathfrak{h}) \neq \emptyset$, $(\mathbf{a}_t)_{t=0,1,\ldots,n} \subset \mathbb{R}^n$ such that $\mathbf{a}_t - \mathbf{a}_0$, $t \in [n]$ are linearly independent, $\varepsilon > 0$ and $k \in \mathbb{N}$.
Output: $\bar{\mathbf{x}}$.
Begin with $t := 0$ and do:

1: Solve SDP (6.2.10) with $f = \|\mathbf{x}\|_2^2$ to obtain $\rho_k^i(\varepsilon)$. Set $L := \rho_k^i(\varepsilon)$ and go to step 2.
2: Solve SDP (6.2.17) to obtain η_k^t and an associated solution $\mathbf{y}$.
a: If $t \leq n$ and $\operatorname{rank}(\mathbf{M}_{k+w}(\mathbf{y})) = \operatorname{rank}(\mathbf{M}_k(\mathbf{y}))$, i.e., $\mathbf{y}$ has a representing measure μ, extract $\operatorname{supp}(\mu)$ from $\mathbf{y}$ by using the algorithm from [77]. Take $\bar{\mathbf{x}} \in \operatorname{supp}(\mu)$ and stop.
b: If $t \leq n$ and $\operatorname{rank}(\mathbf{M}_{k+w}(\mathbf{y})) \neq \operatorname{rank}(\mathbf{M}_k(\mathbf{y}))$, set $t := t+1$ and do again step 2.
c: If $t = n+1$, stop.

Table 6.1: Decrease of the moment matrix rank in Algorithm 6.2.3.

t	$\mathbf{a}_t$	η_0^t	$\operatorname{rank}(\mathbf{M}_1(\mathbf{y}))$	$\operatorname{rank}(\mathbf{M}_0(\mathbf{y}))$	$\bar{\mathbf{x}}$
0	$\mathbf{0}_{\mathbb{R}^4}$	2.0000	5	1	–
1	$\mathbf{e}_1$	1.0000	3	1	–
2	$\mathbf{e}_2$	2.9997	3	1	–
3	$\mathbf{e}_3$	1.9998	1	1	(0.9999, 0.0001, 0.5028, -0.8611)
4	$\mathbf{e}_4$	1.3089	1	1	(1.0000, 0.0002, 0.4968, 0.8329)

The decrease of the moment matrix rank in Algorithm 6.2.3 for the kissing number problem with $g_1 = x_1^2+x_2^2+x_3^2+x_4^2-2x_1x_3-2x_2x_4-1$, $h_1 = x_1^2+x_2^2-1$ and $h_2 = x_3^2+x_4^2-1$ is illustrated in Table 6.1. Here e_i, $i \in [4]$ is the canonical basis of $\mathbb{R}^4$. In this example, $\operatorname{rank}(\mathbf{M}_1(\mathbf{y}))$ decreases from 5 to 1 when t goes from 0 to 4 and $\mathbf{M}_1(\mathbf{y})$ fulfills the flat extension condition at from $t = 3$.

Remark 6.11. *ASC can be used to find an approximation of a real point in $S(\mathfrak{g},\mathfrak{h})$ even if $S(\mathfrak{g},\mathfrak{h})$ is positive dimensional. This is illustrated later on by our numerical experiments from Section 6.3 (see the polynomial systems corresponding to Id 6, 7, 8 and 13).*

Obtaining a minimizer by using the ASC method: We rely on the following algorithm to find the value $\rho_k^i(\varepsilon)$ of SDP (6.2.10), which approximates $f^\star = \min\{f(\mathbf{x}) : \mathbf{x} \in S(\mathfrak{g},\mathfrak{h})\}$, together with an approximation $\bar{\mathbf{x}}$ of a minimizer $\mathbf{x}^\star$ for this problem.

Algorithm 13 PolyOpt

Input: f, $S(\mathfrak{g},\mathfrak{h}) \neq \emptyset$, $\varepsilon > 0$ and $k \in \mathbb{N}$.
Output: $\rho_k^i(\varepsilon)$ and $\bar{\mathbf{x}}$.

1: Solve SDP (6.2.10) to obtain $\rho_k^i(\varepsilon)$.
2: Compute $\bar{\mathbf{x}}$ in $S(\mathfrak{g} \cup \{\rho_k^i(\varepsilon) - f\},\mathfrak{h})$ by using Algorithm 6.2.3 and stop.

Proposition 6.8. *If POP $f^\star := \inf\{f(\mathbf{x}) : \mathbf{x} \in S(\mathfrak{g},\mathfrak{h})\}$ admits an optimal solution at $\mathbf{x}^\star$, then for k large enough, Algorithm 6.2.3 terminates and $f^\star \leq \rho_k^i(\varepsilon) \leq f^\star + \varepsilon\theta(\mathbf{x}^\star)^d$. Moreover, for k large enough, $\bar{\mathbf{x}} \in S(\mathfrak{g},\mathfrak{h})$ and $f^\star \leq f(\bar{\mathbf{x}}) \leq f^\star + \varepsilon\theta(\mathbf{x}^\star)^d$ if the assumption of Theorem 6.6 holds for $S(\mathfrak{g} \cup \{\rho_k^i(\varepsilon) - f\},\mathfrak{h})$.*

In practice, one performs Algorithm 6.2.3 several times by updating $k := k+1$ until one obtains $\bar{\mathbf{x}}$ in $S(\mathfrak{g} \cup \{\rho_k^i(\varepsilon) - f\},\mathfrak{h})$. Obviously, one has $f^\star + \varepsilon\theta(\mathbf{x}^\star)^d \geq \rho_k^i(\varepsilon) \geq f(\bar{\mathbf{x}}) \geq f^\star$.

6.3 Examples

In this section, we report results obtained after solving some instances of POP (6.2.8) with Algorithm 6.2.3. As before, let us note $\mathfrak{g} = \{g_1,\ldots,g_m\}$ and $\mathfrak{h} = \{h_1,\ldots,h_l\}$ the sets of polynomials involved in the inequality constraints and the equality constraints, respectively. In particular, the resulting set $S(\mathfrak{g},\mathfrak{h})$ is unbounded for all examples. The experiments are performed with both

MATLAB R2018a/Yalmip and Julia 1.1.1/JuMP to model the semidefinite optimization problems and `Mosek` 8.0 to solve these problems. The codes for Algorithm 6.2.3 (PolySys) and Algorithm 6.2.3 (PolyOpt) can be downloaded from the link: https://github.com/maihoanganh. In these codes, we always set $\mathbf{a}_0 := 0_{\mathbb{R}^n}$ and $\mathbf{a}_1, \ldots, \mathbf{a}_n$ as the canonical basis of $\mathbb{R}^n$. We use a desktop computer with an Intel(R) Pentium(R) CPU N4200 @ 1.10GHz and 4.00 GB of RAM. The input data given in Table 6.2 include examples of unconstrained and constrained POPs. The corresponding output data, the exact results and timings are given in Table 6.3. In these tables, the SOS hierarchy (6.2.10) is solved by optimization models in Yalmip (Y) and JuMP (J). The symbol "$-$" in a column entry indicates that the calculation did not finish in a couple of hours.

Id 1-5 are unconstrained POPs. Id 6-12 are POPs with inequality constrains, Id 13-18 are POPs with equality constraints and Id 19-25 are POPs with both inequality and equality constraints. Id 8, 11 and 12 correspond to examples from Jeyakumar et al. [88, 87]. Id 9 and 10 are selected from Demmel et al. [44]. Id 13-17 come from Greuet et al. [67]. Id 23, 24 and 25 are POPs constructed from some inequalities issued from Mathematics competitions mentioned in [190, 50], yielding noncompact POPs with known optimal values and optimizers.

Even though the sets of minimizers associated to Id 6, 7, 8 and 13 are positive dimensional, we can still extract a single approximate optimal solution by using our ASC algorithm. Note that ASC computes a real point $\bar{\mathbf{x}}$ in $S(\mathfrak{g} \cup \{\rho_k^i(\varepsilon) - f\}, \mathfrak{h})$ which is an outer approximation of $\{\mathbf{x}^\star \in S(\mathfrak{g}, \mathfrak{h}) : f(\mathbf{x}^\star) = f^\star\}$ for k sufficiently large.

In Table 6.3, Algorithm 6.2.3 terminates at some order $k \leq 5$ for all POPs except Id 16. Note that for Id 16, the global minimum does not satisfy the KKT conditions. Thus the method of Demmel et al. [44, 149] and Nie's [148] cannot be used to solve this POP. Moreover, the convergence rate of $(\rho_k^i(\varepsilon))_{k\in\mathbb{N}}$ in Id 16 is very poor when $\varepsilon \leq 10^{-5}$. We overcome this issue by fixing k, multiplying ε by 10, and solving again the relaxations. The computational cost that we must pay here is due to the largest gap $\varepsilon\,\theta(\mathbf{x}^\star)^d$ between $\rho_k^i(\varepsilon)$ and $f^\star$. This behavior is illustrated in Table 6.4.

In Id 18, even if the ideal $\mathcal{I}(\mathfrak{h})$ is not radical and $V(\mathfrak{h})$ is not equidimensional (the assumptions required to apply the framework in [67] are not guaranteed) our ASC method can still extract one solution of the problem.

For Id 21, we can improve the quality of the approximation $\rho_k^i(\varepsilon)$ of the optimal value $f^\star$ by fixing $k = 1$, dividing ε by 10, and solving again the relaxations. This is illustrated in Table 6.5.

We emphasize that we can customize the ε parameter for different purposes. On the one hand, one increases ε to improve the convergence speed of the sequence $(\rho_k^i(\varepsilon))_{k\in\mathbb{N}}$ to the neighborhood $[f^\star, f^\star + \varepsilon\theta(\mathbf{x}^\star)^d]$ of $f^\star$ (see Table 6.4). On the other hand, one decreases ε to improve the accuracy of the approximate optimal value $\rho_k^i(\varepsilon)$ and the approximate optimal solution $\bar{\mathbf{x}}$ (see Table 6.5).

Our numerical benchmarks also show that modeling in JuMP is faster and provides more accurate outputs than modeling in Yalmip. In particular, the JuMP implementation is the only one which provides solutions for Id 11, 12, 17 and 23.

Let us now denote by k_ε the smallest nonnegative integer such that $\rho_{k_\varepsilon}^i(\varepsilon) \geq f^\star$, for each $\varepsilon > 0$. The graph of the function $\varepsilon^{-1} \mapsto k_\varepsilon$ on $(0, 100]$ for Id 9 and Id 16 is illustrated in Figure 6.2. Here

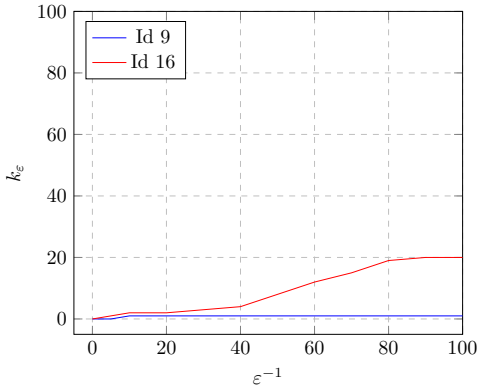


Figure 6.2: Plot of the complexity.

Table 6.2: Examples of POPs.

Id	reference	input data
1	Motzkin	$f = x_1^2x_2^2(x_1^2+x_2^2-1)$ $\mathfrak{g}=\emptyset$ $\mathfrak{h}=\emptyset$
2	Robinson	$f = x_1^6+x_2^6-x_1^4x_2^2-x_1^2x_2^4-x_1^4-x_2^4-x_1^2-x_2^2+3x_1^2x_2^2$ $\mathfrak{g}=\emptyset$ $\mathfrak{h}=\emptyset$
3	Choi-Lam	$f = x_1^4x_2^2+x_2^4+x_1^2-3x_1^2x_2^2$ $\mathfrak{g}=\emptyset$ $\mathfrak{h}=\emptyset$
4	Lax-Lax	$f = x_1x_2x_3-x_1(x_2-x_1)(x_3-x_1)(1-x_1)-x_2(x_1-x_2)(x_3-x_2)(1-x_2)-x_3(x_1-x_3)(x_2-x_3)(1-x_3)-(x_1-1)(x_2-1)(x_3-1)$ $\mathfrak{g}=\emptyset$ $\mathfrak{h}=\emptyset$
5	Delzell	$f = x_1^4x_2^2+x_2^4x_3^2+x_1^2x_3^4-3x_1^2x_2^2x_3^2+x_3^8$ $\mathfrak{g}=\emptyset$ $\mathfrak{h}=\emptyset$
6	Modified Motzkin	$f = (x_1^2+x_2^2-3)x_1^2x_2^2$ $\mathfrak{g}=\{x_1^2+x_2^2-4\}$ $\mathfrak{h}=\emptyset$
7	[82, Example 4.3]	$f = x_1^4+x_2^4+x_3^4-4x_1x_3^3$ $\mathfrak{g}=\{1-x_1^4+\frac{1}{2}x_2^4-x_3^4\}$ $\mathfrak{h}=\emptyset$
8	[88, Example 3.1]	$f = x_1^2+1$ $\mathfrak{g}=\{1-x_2^2, x_2^2-1/4\}$ $\mathfrak{h}=\emptyset$
9	[44, Example 4.5]	$f = x_1^2+x_2^2$ $\mathfrak{g}=\{x_1^2-x_1x_2-1, x_1^2+x_1x_2-1, x_2^2-1\}$ $\mathfrak{h}=\emptyset$
10	[44, Example 4.4]	$f = -\frac{4}{3}x_1^2+\frac{2}{3}x_2^2-2x_1x_2$ $\mathfrak{g}=\{x_2^2-x_1^2, -x_1x_2\}$ $\mathfrak{h}=\emptyset$
11	[87, Section 5.2]	$f = 1+\sum_{j=2}^{8}((x_j-x_{j-1}^2)^2+(1-x_j^2))$ $\mathfrak{g}=\{x_1,\ldots,x_8\}$ $\mathfrak{h}=\emptyset$
12	[87, Section 5.3]	$f = 1+\sum_{l=1}^{3}((x_{2l}-x_{2l-1}^2)^2+(1-x_{2l-1})^2+90(x_{2l+2}^2-x_{2l+1})^2+(x_{2l+1}-1)^2+10(x_{2l}+x_{2l+2}-2)^2+\frac{1}{10}(x_{2l}-x_{2l+2})^2)$ $\mathfrak{g}=\{x_1,\ldots,x_8\}$ $\mathfrak{h}=\emptyset$
13	[67, Example A.2]	$f = (x_1^2+x_2^2-2)(x_1^2+x_2^2)$ $\mathfrak{g}=\emptyset$ $\mathfrak{h}=\{(x_1^2+x_2^2-1)(x_1-3)\}$
14	[67, Example A.5]	$f = x_1^6+x_2^6+x_3^6+3x_1^2x_2^2x_3^2-x_1^2(x_2^4+x_3^4)-x_2^2(x_3^4+x_1^4)-x_3^2(x_1^4+x_2^4)$ $\mathfrak{g}=\emptyset$ $\mathfrak{h}=\{x_1+x_2+x_3-1\}$
15	[67, Example A.6]	$f = x_1x_2x_3x_4-x_1(x_2-x_1)(x_3-x_1)(x_4-x_1)-x_2(x_1-x_2)(x_3-x2)(x_4-x_2)-x_3(x_2-x_3)(x_1-x_3)(x_4-x_3)-x_4(x_2-x_4)(x_3-x_4)(x_1-x_4)$ $\mathfrak{g}=\emptyset$ $\mathfrak{h}=\{x_1, x_2-x_3, x_3-x_4\}$
16	[67, Example A.4]	$f = (x_1+1)^2+x_2^2$ $\mathfrak{g}=\emptyset$ $\mathfrak{h}=\{x_1^3-x_2^2\}$
17	[67, Example A.8]	$f = \frac{1}{6}\sum_{j=1}^{5}(x_j^2+x_{j+5}^2)$ $\mathfrak{g}=\emptyset$ $\mathfrak{h}=\{x_6-1, x_{j+6}-x_{j+5}-\frac{1}{6}(x_{j+5}^2-x_j) : j\in[4]\}$
18	self made	$f = x_1^6+x_2^2$ $\mathfrak{g}=\emptyset$ $\mathfrak{h}=\{(x_1^2+x_2^2)(1-x_1x_2)^2\}$
19	[62, Example 2]	$f = x_1^2+x_2^2+x_3^2+x_4^2$ $\mathfrak{g}=\{\frac{1}{8}-x_4\}$ $\mathfrak{h}=\{x_1+x_2+x_3+x_4-1\}$
20	self made	$f = x_1^3-x_2^2$ $\mathfrak{g}=\{x_1,x_2\}$ $\mathfrak{h}=\{(x_1x_2+1)(x_1-x_2)^2\}$
21	self made	$f = x_1^4-3x_2$ $\mathfrak{g}=\{x_1,x_2\}$ $\mathfrak{h}=\{(x_2-x_1^2)(2x_1^2-x_2)\}$
22	AM-GM inequality	$f = x_1+x_2+x_3$ $\mathfrak{g}=\{x_1,x_2,x_3\}$ $\mathfrak{h}=\{x_1x_2x_3-1\}$
23	[190, USSR Olimpiad 1989]	$f = (x_1+x_2)(x_2+x_3)$ $\mathfrak{g}=\{x_1,x_2,x_3\}$ $\mathfrak{h}=\{x_1x_2x_3(x_1+x_2+x_3)-1\}$
24	[50, IMO 1990]	$f = x_1(x_1+x_2+x_3)(x_1+x_3+x_4)(x_1+x_2+x_4)+x_2(x_1+x_2+x_3)(x_2+x_3+x_4)(x_1+x_2+x_4)+x_3(x_1+x_2+x_3)(x_1+x_3+x_4)(x_3+x_2+x_4)+x_4(x_4+x_2+x_3)(x_1+x_3+x_4)(x_1+x_2+x_4)-\frac{1}{3}(x_1+x_2+x_3)(x_1+x_3+x_4)(x_1+x_2+x_4)(x_2+x_3+x_4)$ $\mathfrak{g}=\{x_1,x_2,x_3,x_4\}$ $\mathfrak{h}=\{x_1x_2+x_2x_3+x_3x_4+x_4x_1-1\}$
25	[50, IMO 2000]	$f = -(x_1x_2-x_2+1)(x_2x_3-x_3+1)(x_3x_1-x_1+1)$ $\mathfrak{g}=\{x_1,x_2,x_3\}$ $\mathfrak{h}=\{x_1x_2x_3-1\}$

Table 6.3: Numerical experiments with $\varepsilon = 10^{-5}$.

Id	$f^\star$	$\{\mathbf{x}^\star \in S(\mathfrak{g},\mathfrak{h}) : f(\mathbf{x}^\star) = f^\star\}$	k	$\rho_k^i(\varepsilon)$	$\bar{\mathbf{x}}$	time (s)
1	$-\frac{1}{27}$	$\{\frac{1}{\sqrt{3}}(\pm 1, \pm 1)\}$	2	-0.0369	(0.5713, 0.5713)	Y: 2.65 J: 2.58
2	-1	$\{(\pm 1, \pm 1), (0, \pm 1), (\pm 1, 0)\}$	2	-0.9999	(-0.0000, -0.9967)	Y: 2.90 J: 2.91
3	0	$\{(\pm 1, \pm 1), (0, 0)\}$	1	0.0000	(-0.0000, -0.0000)	Y: 3.26 J: 0.22
4	0	$\{(1,0,0), (0,1,0), (0,0,1),$ $(1,1,0), (0,1,1), (1,0,1)\}$	1	0.0000	(0.00, 0.00, 0.99)	Y: 4.22 J: 0.30
5	0	$\{(0,0,0)\}$	2	0.0000	(0.00, 0.00, 0.00)	Y: 65.58 J: 31.84
6	0	$\{(0,t),\ (t,0)\ :\ \lvert t\rvert \geq 2\}$	1	0.0062	(2.0000, 0.0000)	Y: 4.47 J: 3.10
7	$1 - \sqrt[4]{27}$	$\{(0,t,0) : t \in \mathbb{R}\} \cup$ $\{t(1,0,\sqrt[4]{3}) : \lvert t\rvert \leq 1/\sqrt[4]{4}\}$	1	-1.2793	(0.70, -0.00, 0.93)	Y: 3.96 J: 1.94
8	1	$\{0\} \times (\left[\frac{1}{2}, 1\right] \cup \left[-1, -\frac{1}{2}\right])$	1	1.0000	(0.0000, -0.5081)	Y: 3.41 J: 2.58
9	$\frac{5+\sqrt{5}}{2}$	$\{(\pm\frac{1+\sqrt{5}}{2}, \pm 1)\}$	2	3.6182	(1.6181, 1.0000)	Y: 4.72 J: 2.53
10	0	$\{(0,0)\}$	5	Y: -0.0005 J: -0.0002	(0.0000, 0.0000)	Y: 10.08 J: 3.82
11	1	$\{(1,\ldots,1)\}$	0	Y: – J: 1.0072	Y: – J: (0.95, 0.96, 0.96, 0.96, 0.97, 0.97, 0.97, 0.97)	Y: – J: 2265.71
12	1	$\{(1,\ldots,1)\}$	0	Y: – J: 1.0072	Y: – J: (0.98, 0.99, 0.99, 0.99, 0.99, 0.99, 0.96, 0.98)	Y: – J: 2642.23
13	-1	$\{(t, \pm\sqrt{1-t^2}) : t \in [-1,1]\}$	2	-0.9999	Y: (1.0000, 0.0000) J: (0.3663, -0.9304)	Y: 8.46 J: 3.99
14	0	$\{\frac{1}{3}(1,1,1)\}$	2	0.0000	(0.33, 0.33, 0.33)	Y: 39.25 J: 6.92
15	0	$\{(0,0,0,0)\}$	1	0.0000	$(0.00, \ldots, 0.00)$	Y: 11.06 J: 4.19
16	1	$\{(0,0)\}$	5 10 15 20	0.9771 0.9802 0.9857 0.9918	J: – J: – J: – J: –	J: 1.83 J: 4.49 J: 15.38 J: 154.60
17			1	Y: 1.3216 J: 1.4883	Y: – J: (1.19, 0.78, 0.46, 0.20, 0.00, 1.0, 0.96, 0.99, 1.08, 1.24)	Y: 1846.74 J: 639.97
18	0	$\{(0,0)\}$	1	0.0000	(0.0000, 0.0000)	Y: 2.55 J: 0.34
19	$\frac{13}{48}$	$\{(\frac{7}{24}, \frac{7}{24}, \frac{7}{24}, \frac{1}{8})\}$	0	0.2708	$(0.29, 0.29, 0.29, 0.12)$	Y: 24.44 J: 17.02
20	$-\frac{4}{27}$	$\{(\frac{2}{3}, \frac{2}{3})\}$	1	-0.1487	(0.6518, 0.6526)	Y: 20.52 J: 2.03
21	-9	$\{(\sqrt{3}, 6)\}$	1	Y: -8.0171 J: -8.5578	Y: (1.4172, 4.0172) J: (1.5280, 4.6701)	Y: 7.60 J: 2.52
22	3	$\{(1,1,1)\}$	2	3.0000	(1.00, 1.00, 1.00)	Y: 8.57 J: 3.89
23	2	$\{(1, \sqrt{2}-1, 1)\}$	5	Y: – J: 2.0000	Y: – J: (0.99, 0.41, 0.99)	Y: – J: 201.29
24	$\frac{81}{16}$	$\{\frac{1}{2}(1,1,1,1)\}$	0	5.0625	(0.49, 0.49, 0.49, 0.49)	Y: 11.37 J: 2.18
25	-1	$\{(1,1,1)\}$	1	-0.9974	(1.0, 1.0, 1.0)	Y: 35.33 J: 7.45

Table 6.4: Numerical experiments for Id 16 with various values of ε.

ε	$\rho_5^i(\varepsilon)$	$\bar{\mathbf{x}}$	time (s)
10^{-4}	Y: 0.9492 J: 0.9778	J: – Y: –	Y: 4.47 J: 2.02
10^{-3}	Y: 0.9528 J: 0.9783	J: – Y: –	Y:4.62 J: 1.59
10^{-2}	Y: 0.9550 J: 0.9884	J: – Y: –	Y: 4.45 J: 1.29
10^{-1}	Y: 1.0479 J: 1.0774	(0.0000, 0.0000)	Y: 17.64 J: 3.76

Table 6.5: Numerical experiments for Id 21 with various values of ε.

ε	$\rho_1^i(\varepsilon)$	$\bar{\mathbf{x}}$	time (s)
10^{-6}	Y: -8.2078 J: -8.9392	Y: $(1.4525, 4.2199)$ J: $(1.6593, 5.5069)$	Y: 8.46 J: 3.10
10^{-7}	Y: -8.4889 J: -8.9935	Y: $(1.5116, 4.5701)$ J: $(1.7086, 5.8387)$	Y: 8.34 J: 2.61
10^{-8}	Y: -8.4915 J: -8.9968	Y: $(1.5122, 4.5739)$ J: $(1.7158, 5.8873)$	Y: 8.37 J: 2.70
10^{-9}	Y: -7.9335 J: -8.9949	Y: $(1.4026, 3.9346)$ J: $(1.7113, 5.8572)$	Y: 8.30 J: 2.54

Id 9 (resp. Id 16) is an example of POP such that the global minimums satisfy the KKT condition (resp. do not satisfy the KKT condition). We can experimentally compare the complexity of Algorithm 6.2.3 in both cases. For Id 9, the function seems to increase as slowly as a constant function, which is in deep contrast with Id 16, where the function increases more quickly and seems to have a step-wise linear growth. Theorem 6.5 states that $k_\varepsilon \leq \mathcal{O}(\varepsilon^{-\mathfrak{c}})$ as $\varepsilon \downarrow 0$ for some $\mathfrak{c} > 0$ independent from ε.

Chapter 7

On the complexity of Putinar–Vasilescu's Positivstellensatz

Most of the content of this chapter is from [132].

In the previous chapter, we have applied Putinar–Vasilescu's Positivstellensatz to design a hierarchy of SDP relaxations that solves POPs over noncompact semialgebraic sets. In this chapter, we provide a constructive proof of the degree bound for Putinar–Vasilescu's Positivstellensatz. It allows us to analyze the convergence rate of the corresponding hierarchy of SDP relaxations.

For a positive $m \in \mathbb{N}$, let us consider the polynomial optimization problem (POP):

$$f^\star := \inf_{\mathbf{x} \in S(\mathfrak{g})} f(\mathbf{x}) \,, \tag{7.0.1}$$

where $f \in \mathbb{R}[\mathbf{x}]$ and

$$S(\mathfrak{g}) := \{\mathbf{x} \in \mathbb{R}^n \,:\, g_i(\mathbf{x}) \geq 0 \,,\, i \in [m]\} \,, \tag{7.0.2}$$

for some $\mathfrak{g} = \{g_i\}_{i \in [m]} \subset \mathbb{R}[\mathbf{x}]$. Here $\mathbb{R}[\mathbf{x}]$ denotes the ring of real polynomials in vector of variables $\mathbf{x} = (x_1, \ldots, x_n)$ and $[m]$ stands for the set $\{1, \ldots, m\}$. Assume that f has degree at most $2d_f$ for some positive $d_f \in \mathbb{N}$. Recall that the set $S(\mathfrak{g})$ is a conjunction of finitely many polynomial inequalities, and therefore is called a *basic semialgebraic set*.

Problem (7.0.1) can be written as

$$f^\star = \sup_{\lambda \in \mathbb{R}} \{\lambda \,:\, f - \lambda > 0 \text{ on } S(\mathfrak{g})\} \,. \tag{7.0.3}$$

We can replace the inequality constraint of problem (7.0.3) by an equality constraint, if one can represent positive polynomials on $S(\mathfrak{g})$. Assume that S has nonempty interior and a ball constraint is present, i.e., $g_1 = R - \|\mathbf{x}\|_2^2$ for some $R > 0$. Our first (minor) contribution is to rely on the representation of polynomials positive on S stated by Putinar and Vasilescu [169], to obtain

$$f - \lambda = \frac{\sigma_0 + \sum_{i \in [m]} \sigma_i g_i}{(1 + \|\mathbf{x}\|_2^2)^k} \,, \tag{7.0.4}$$

for some $k \in \mathbb{N}$, $\sigma_i \in \Sigma[\mathbf{x}]$, $i \in [m]$, being such that $\deg(\sigma_0) \leq 2(k + d_f)$ and $\deg(\sigma_i g_i) \leq 2(k + d_f)$. Here $\Sigma[\mathbf{x}]$ denotes the set of sum-of-squares (SOS) polynomials and $\deg(\cdot)$ stands for the degree of a polynomial. Such a representation of positive polynomials is called a *Positivstellensatz*.

After bounding the degrees of the SOS polynomials involved in (7.0.4), we obtain the following hierarchy of relaxations indexed by $k \in \mathbb{N}$:

$$\begin{aligned} \rho_k^{(0)} := \sup_{\lambda, \sigma_i} \quad & \lambda \\ \text{s.t.} \quad & \lambda \in \mathbb{R} \,,\, \sigma_i \in \Sigma[\mathbf{x}] \,, \\ & (1 + \|\mathbf{x}\|_2^2)^k (f - \lambda) = \sigma_0 + \textstyle\sum_{i \in [m]} \sigma_i g_i \,, \\ & \deg(\sigma_0) \leq 2(k + d_f) \,,\, \deg(\sigma_i g_i) \leq 2(k + d_f) \,. \end{aligned} \tag{7.0.5}$$

Problem (7.0.5) can be solved numerically using semidefinite programming [16]. As usual, one can show that for each $k \in \mathbb{N}$, $\rho_k^{(0)}$ is a lower bound of $f^\star$, that the sequence $(\rho_k^{(0)})_{k\in\mathbb{N}}$ is monotone nondecreasing, and converges to $f^\star$.

In the present chapter, we answer the following two interesting questions:

1. How fast does $(\rho_k^{(0)})_{k\in\mathbb{N}}$ converge to $f^\star$? We show the convergence rate $\mathcal{O}(k^{-1/\mathfrak{c}})$ for some constant $\mathfrak{c} > 0$ depending on f and g_i.

2. Is there any explicit example to illustrate this rate of convergence? If $S(\mathfrak{g})$ is the unit ball, i.e., $m = 1$ and $g_1 = 1 - \|\mathbf{x}\|_2^2$, the sequence $(\rho_k^{(0)})_{k\in\mathbb{N}}$ converges to $f^\star$ with the rate $\mathcal{O}(k^{-1/65})$.

Our contribution is concerned with the case of basic semialgebraic sets having nonempty interiors. Basically one obtains a convergence rate similar in spirit and magnitude of Schweighofer's bound $\bar{\mathfrak{c}}\varepsilon^{-\mathfrak{c}}$, but still based on the quadratic module $\mathcal{Q}_k(\mathfrak{g})$ defined as in (1.3.8) (not the preordering $\mathcal{P}_k(\mathfrak{g})$ defined as in (1.3.6)) thanks to the prescribed denominator $(1 + \|\mathbf{x}\|_2^2)^k$ involved in Putinar–Vasilescu's Positivstellensatz.

Before showing explicitly our contribution, we restate the original result of Putinar and Vasilescu (without degree bound) in (6.1).

Contribution. The construction of the hierarchy of semidefinite relaxations (7.0.5) is based on the Positivstellensatz stated in Corollary 7.2. More explicitly, if $S(\mathfrak{g})$ has nonempty interior such that $g_1 = R - \|\mathbf{x}\|_2^2$ for some $R > 0$ and f is of degree at most $2d_f$ such that f is nonnegative on S, then there exist positive constants $\bar{\mathfrak{c}}$ and $\mathfrak{c}$ depending on f, g_i such that for all $\varepsilon > 0$, for all $k \geq \bar{\mathfrak{c}}\varepsilon^{-\mathfrak{c}}$,

$$(1 + \|\mathbf{x}\|_2^2)^k (f + \varepsilon) = \sigma_0 + \textstyle\sum_{i\in[m]} \sigma_i g_i \,, \tag{7.0.6}$$

for some $\sigma_i \in \Sigma[\mathbf{x}]$ being such that $\deg(\sigma_0) \leq 2(k + d_f)$ and $\deg(\sigma_i g_i) \leq 2(k + d_f)$.

In order to prove (7.0.6), we provide a degree bound on the weighted SOS polynomials for the homogenized Putinar–Vasilescu's Positivstellensatz [169]. This is stated in Theorem 7.1 as follows: If $f, g_1, \ldots, g_m$ are homogeneous polynomials of even degrees such that $S(\mathfrak{g})$ has nonempty interior and f is nonnegative on $S(\mathfrak{g})$, then there exist positive constants $\bar{\mathfrak{c}}$ and $\mathfrak{c}$ depending on f, g_i such that for all $\varepsilon > 0$, for all $k \geq \bar{\mathfrak{c}}\varepsilon^{-\mathfrak{c}}$,

$$\|\mathbf{x}\|_2^{2k} (f + \varepsilon \|\mathbf{x}\|_2^{2d_f}) = \sigma_0 + \textstyle\sum_{i\in[m]} \sigma_i g_i \,, \tag{7.0.7}$$

for some homogeneous SOS polynomials σ_i being such that $\deg(\sigma_0) = \deg(\sigma_i g_i) = 2(k + d_f)$. Here a polynomial q is homogeneous of degree $2t$ if $q(\lambda x) = \lambda^{2t} q(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$ and each $\lambda \in \mathbb{R}$. Remark that the original version of Putinar–Vasilescu's Positivstellensatz in [169] does not include any degree bound on the weighted SOS polynomials σ_i involved in the representation (7.0.7). Our proof of Theorem 7.1 consists of three main steps:

1. Construct iteratively some positive "weight" functions ψ_i such that $f + \varepsilon - \sum_{i\in[m]} \psi_i g_i$ is positive on $[-1, 1]^n$. The idea of this step is similar in spirit to the proof of the inductive property in [186, Proposition 3.1] and relies on the Lojasiewicz inequality.

2. Approximate $\sqrt{\psi_i}$ with the multivariate Bernstein polynomial q_i on $[-1, 1]^n$ such that the polynomial $H = f + \varepsilon - \sum_{i\in[m]} q_i^2 g_i$ is positive on the unit sphere $\mathbb{S}^{n-1}$.

3. Apply Reznick's Positivstellensatz [175] to the homogenization of H.

The complexity analysis of every step is derived to get the final degree bound $\bar{\mathfrak{c}}\varepsilon^{-\mathfrak{c}}$.

Afterwards, we obtain in Corollary 7.1 the same degree bound for the dehomogenized Putinar–Vasilescu's Positivstellensatz. This improves the bound $\mathcal{O}(\exp(\varepsilon^{-\mathfrak{c}}))$ obtained in the previous chapter, based on Nie–Schweighofer's complexity result [150] for Putinar's Positivstellensatz [168]. Corollary 7.1 yields the convergence rate $\mathcal{O}(\varepsilon^{-\mathfrak{c}})$ for the corresponding hierarchy of relaxations for polynomial optimization on general (not necessarily compact) basic semialgebraic sets.

Remark 7.1. *Concerning the assumption that the semialgebraic set $S(\mathfrak{g})$ has nonempty interior, a technique from [1] maybe helps us remove this assumption*[1].

[1] Tien-Son Pham, Personal communication, 2021.

Technical insights. We start to recall the two main steps in the proof of Nie and Schweighofer [150] for the degree bound of SOS polynomials involved in Putinar's Positivstellensatz:

1. Find a large enough $k \in \mathbb{N}$ such that the polynomial

$$F = f + \varepsilon - \lambda \textstyle\sum_{i\in[m]} (g_i - 1)^{2k} g_i \tag{7.0.8}$$

   is positive on $[-1,1]^n$. An estimate $k \geq \mathcal{O}(\varepsilon^{-\mathfrak{c}})$ is obtained. Here $\varepsilon > 0$ measures how close the polynomial f (assumed to be nonnegative on $S(\mathfrak{g})$) is to have a zero on $S(\mathfrak{g})$.

2. Apply Schmüdgen's Positivstellensatz to F on $[-1,1]^n$.

Notice that Schweighofer's degree bound of Schmüdgen's Positivstellensatz is exponential in the degree of the given positive polynomial (n^{d_f} in (1.3.5)). Accordingly, Nie and Schweighofer obtain an exponential bound $n^{\mathcal{O}(\varepsilon^{-\mathfrak{c}})}$ in the second step since $\deg(F) \sim Ck$ as $k \to \infty$ for some positive constant C.

One notable difference in our proof is that the weight $\lambda(g_i - 1)^k$ in (7.0.8) is replaced by a non-differentiable positive function ψ_i. Surprisingly, we can prove that the square root $\sqrt{\psi_i}$ is a Lipschitz continuous function. Thus each $\sqrt{\psi_i}$ can be approximated with a Bernstein polynomials q_i on $[-1,1]$. Here, the advantage of using Bernstein polynomials is that the approximation error between $\sqrt{\psi_i}$ and q_i decreases with a rate which only depends on a Lipschitz constant of $\sqrt{\psi_i}$, and $|q_i|$ is upper bounded by the supremum of $\sqrt{\psi_i}$ on $[-1,1]^n$.

Next, we apply Reznick's Positivstellensatz to the homogeneous polynomial $\tilde{H}$ obtained from the homogenization of

$$H := f + \varepsilon - \textstyle\sum_{i\in[m]} q_i^2 g_i \,, \tag{7.0.9}$$

being such that the bounds of $\tilde{H}$ and H on the unit sphere are the same. The important point to note here is that the degree bound of Reznick's Positivstellensatz is quadratic in the degree of $\tilde{H}$ and linear in the ratio $\delta(\tilde{H})$ (see (1.3.1)). This is in deep contrast with Schmüdgen's Positivstellensatz, as there is no exponential dependency in these two quantities. This leads to the difference between our convergence rate $\mathcal{O}(\varepsilon^{-\mathfrak{c}})$ and Nie–Schweighofer's rate $\mathcal{O}(\exp(\varepsilon^{-\mathfrak{c}}))$.

One may ask whether with the same techniques from our proof, one could improve the existing degree bound for Putinar's Positivstellensatz. We have tried to apply the degree bound (1.3.3) of Pólya's Positivstellensatz to H after a change of coordinate, but unfortunately this leads to the same bound as Nie and Schweighofer. The underlying reason is that the norm $\|q\|$ in (1.3.3) depends on the coefficients of q. In our situation, q coincides with H and the coefficients of H are bounded by a value involving the coefficients of the Bernstein polynomials. The bound on the largest coefficient, even for a univariate Bernstein polynomial, seems to be exponential in the approximation order t, namely, $\binom{2t}{t} \sim \frac{4^t}{\sqrt{\pi t}}$ as $t \to \infty$. The same issue occurs when we apply the degree bound of Schmüdgen's Positivstellensatz instead of the one of Pólya's Positivstellensatz.

7.1 Representation theorems and degree bounds

In this section, we derive representations of polynomials nonnegative on semialgebraic sets together with degree bounds. We extend these representations to the set of continuous functions being nonnegative on compact domains.

7.1.1 Polynomials nonnegative on general semialgebraic sets

We analyze the complexity of Putinar–Vasilescu's Positivstellensatz [169] in the following theorem:

Theorem 7.1. *(Homogenized representation) Let $g_1, \ldots, g_m$ be homogeneous polynomials of even degrees such that the semialgebraic set*

$$S := \{\mathbf{x} \in \mathbb{R}^n \,:\, g_1(\mathbf{x}) \geq 0 \ldots, g_m(\mathbf{x}) \geq 0\} \tag{7.1.1}$$

has nonempty interior. Let f be a homogeneous polynomial of degree $2d_f$ for some $d_f \in \mathbb{N}$ such that f is nonnegative on S. Then there exist positive constants $\bar{\mathfrak{c}}$ and $\mathfrak{c}$ depending on $f, g_1, \ldots, g_m$ such that for all $\varepsilon > 0$, for all $k \in \mathbb{N}$ satisfying

$$k \geq \bar{\mathfrak{c}}\varepsilon^{-\mathfrak{c}}\,, \tag{7.1.2}$$

there exist homogeneous SOS polynomials $\sigma_0, \ldots, \sigma_m$ such that

$$\deg(\sigma_0) = \deg(\sigma_1 g_1) = \cdots = \deg(\sigma_m g_m) = 2(k + d_f) \tag{7.1.3}$$

and

$$\|\mathbf{x}\|_2^{2k}(f + \varepsilon\|\mathbf{x}\|_2^{2d_f}) = \sigma_0 + \sigma_1 g_1 + \cdots + \sigma_m g_m\,. \tag{7.1.4}$$

In particular, if $m = 1$ and $g_1 = x_n^2 - \|\mathbf{x}'\|_2^2$ with $\mathbf{x}' := (x_1, \ldots, x_{n-1})$, then $\mathfrak{c} = 65$. Moreover, each SOS polynomial σ_i involved in (7.1.4) can be chosen as the (single) square of a homogeneous polynomial, for $i \in [m]$.

The proof of Theorem 7.1 is postponed to Subsection 7.1.4. In Theorem 7.1, all polynomials are assumed to have even degrees. Next, we provide two corollaries where each polynomial can have odd degree.

The following corollary is a direct consequence of Theorem 7.1.

Corollary 7.1. *(Dehomogenized representation) Let $g_1, \ldots, g_m$ be polynomials such that the semi-algebraic set*

$$S := \{\mathbf{x} \in \mathbb{R}^n \,:\, g_1(\mathbf{x}) \geq 0 \ldots, g_m(\mathbf{x}) \geq 0\} \tag{7.1.5}$$

has nonempty interior. Let f be a polynomial nonnegative on S. Denote $d_f := \lfloor \deg(f)/2 \rfloor + 1$. Then there exist positive constants $\bar{\mathfrak{c}}$ and $\mathfrak{c}$ depending on $f, g_1, \ldots, g_m$ such that for all $\varepsilon > 0$, for all $k \in \mathbb{N}$ satisfying

$$k \geq \bar{\mathfrak{c}}\varepsilon^{-\mathfrak{c}}\,, \tag{7.1.6}$$

there exist SOS polynomials $\sigma_0, \ldots, \sigma_m$ such that

$$\deg(\sigma_0) \leq 2(k + d_f) \quad \textit{and} \quad \deg(\sigma_i g_i) \leq 2(k + d_f)\,,\ i \in [m]\,, \tag{7.1.7}$$

and

$$\theta^k(f + \varepsilon\theta^{d_f}) = \sigma_0 + \sigma_1 g_1 + \cdots + \sigma_m g_m\,, \tag{7.1.8}$$

where $\theta := 1 + \|\mathbf{x}\|_2^2$. Moreover, if $m = 1$ and $g_1 = 1 - \|\mathbf{x}\|_2^2$, then $\mathfrak{c} = 65$.

Proof. The proof of Corollary 7.1 is similar to the proof of [130, Theorems 4 and 5]. We recall the basic ingredients. Let $\tilde{S}$ be a homogenized version of S, defined by

$$\tilde{S} := \{(\mathbf{x}, x_{n+1}) \in \mathbb{R}^{n+1} \,:\, \tilde{g}_i(\mathbf{x}, x_{n+1}) \geq 0\,,\ i \in [m]\}\,, \tag{7.1.9}$$

with $\tilde{g}_i(\mathbf{x}, x_{n+1}) := x_{n+1}^{2\lceil g_i \rceil} g_i(\mathbf{x}/x_{n+1})$ being the degree-$2\lceil g_i \rceil$ homogenization of g_i, for $i \in [m]$. Then the proof consists of three steps:

1. Prove that the degree-$2d_f$ homogenization of f, denoted by $\tilde{f}$, is nonnegative on $\tilde{S}$.
2. Use Theorem 7.1 to obtain a representation of $\tilde{f}$ together with the degree bound on SOS polynomials.
3. Obtain a representation of f by evaluating the representation of $\tilde{f}$ at $x_{n+1} = 1$.

To apply Theorem 7.1, we need to show that if S has nonempty interior, then $\tilde{S}$ has nonempty interior. This statement holds since when a belongs to the interior of S, one has $\tilde{g}_i(\mathbf{a}, 1) = g_i(\mathbf{a}) > 0$, implying that $(\mathbf{a}, 1)$ belongs to the interior of $\tilde{S}$. □

Note that the ice cream constraint $x_{n+1}^2 - \|\mathbf{x}\|_2^2$ is the degree-2 homogenization associated to the ball constraint $1 - \|\mathbf{x}\|_2^2$.

7.1.2 Polynomials nonnegative on compact semialgebraic sets

The following corollary is deduced from Corollary 7.1.

Corollary 7.2. *Let $g_1, \dots, g_m$ be polynomials such that $g_1 := R - \|\mathbf{x}\|_2^2$ for some $R > 0$ and the semialgebraic set*

$$S := \{\mathbf{x} \in \mathbb{R}^n \,:\, g_1(\mathbf{x}) \geq 0 \dots, g_m(\mathbf{x}) \geq 0\} \tag{7.1.10}$$

has nonempty interior. Let f be a polynomial nonnegative on S. Denote $d_f := \lfloor \deg(f)/2 \rfloor + 1$. Then there exist positive constants $\bar{\mathfrak{c}}$ and $\mathfrak{c}$ depending on $f, g_1, \dots, g_m$ such that for all $\varepsilon > 0$, for all $k \in \mathbb{N}$ satisfying

$$k \geq \bar{\mathfrak{c}} \varepsilon^{-\mathfrak{c}}\,, \tag{7.1.11}$$

there exist SOS polynomials $\sigma_0, \dots, \sigma_m$ such that

$$\deg(\sigma_0) \leq 2(k + d_f) \quad \textit{and} \quad \deg(\sigma_i g_i) \leq 2(k + d_f)\,,\ i \in [m]\,, \tag{7.1.12}$$

and

$$(1 + \|\mathbf{x}\|_2^2)^k (f + \varepsilon) = \sigma_0 + \sigma_1 g_1 + \cdots + \sigma_m g_m\,. \tag{7.1.13}$$

Moreover, if $m = 1$ and $L = 1$, then $\mathfrak{c} = 65$.

Proof. By using Corollary 7.1, there exist positive constants $\bar{\mathfrak{c}}$ and $\mathfrak{c}$ depending on $f, g_1, \dots, g_m$ such that for all $\varepsilon > 0$, for all $k \in \mathbb{N}$ satisfying $k \geq \bar{\mathfrak{c}} \varepsilon^{-\mathfrak{c}}$, there exist SOS polynomials $\sigma_0, \dots, \sigma_m$ such that

$$\deg(\sigma_0) \leq 2(k + d_f) \quad \text{and} \quad \deg(\sigma_i g_i) \leq 2(k + d_f)\,,\ i \in [m]\,. \tag{7.1.14}$$

and

$$\theta^k (f + \varepsilon \theta^{d_f}) = \sigma_0 + \sigma_1 g_1 + \cdots + \sigma_m g_m\,, \tag{7.1.15}$$

where $\theta := 1 + \|\mathbf{x}\|_2^2$. In addition,

$$(R+1)^{d_f} - \theta^{d_f} = (R - \|\mathbf{x}\|_2^2) \sum_{j=0}^{d_f - 1} (R+1)^{d_f - 1 - j} \theta^j = s_1 g_1\,, \tag{7.1.16}$$

where $s_1 = \sum_{j=0}^{d_f-1} (R+1)^{d_f-1-j} \theta^j$ is an SOS polynomial of degree $2(d_f - 1)$. From this,

$$\begin{array}{ll} \theta^k [f + \varepsilon (R+1)^{d_f}] & = \theta^k (f + \varepsilon \theta^{d_f}) + \varepsilon \theta^k [(R+1)^{d_f} - \theta^{d_f}] \\ & = \sigma_0 + (\varepsilon s_1 \theta^k + \sigma_1) g_1 + \sum_{i=2}^m \sigma_i g_i\,, \end{array} \tag{7.1.17}$$

which yields the desired result. □

Remark 7.2. *In Corollary 7.2, if $m = 1$ and $L > 0$, we still obtain $\mathfrak{c} = 65$. Indeed, up to a scaling, one can always assume $L = 1$.*

Remark 7.3. *We can apply the technique used in the proof of Corollary 7.2, which consists of replacing the perturbation $\varepsilon\theta^{d_f}$ by ε, to represent polynomials nonnegative on $\mathbb{R}^n$. Let us consider an arbitrary large positive constant L and a polynomial f of degree $2d_f$ which is nonnegative on $\mathbb{R}^n$. Then, thanks to [130, Theorem 3.2], for any $\varepsilon > 0$, for all $k \in \mathbb{N}$ such that $k \geq \mathcal{O}(\varepsilon^{-1})$, $\theta^k(f + \varepsilon\theta^{d_f})$ is an SOS polynomial, so that $\theta^k(f + \varepsilon) = \sigma_0 + \sigma_1(R - \|\mathbf{x}\|_2^2)$ for some SOS polynomials σ_i, $i = 0, 1$. This is the so-called "big ball trick". This representation yields a linear convergence rate $\mathcal{O}(\varepsilon^{-1})$ for the minimization of polynomials on $\mathbb{R}^n$.*

Remark 7.4. *(Removing the denominator in Putinar–Vasilescu's Positivstellensatz) Let $f, g_i \in \mathbb{R}[\mathbf{x}]$ with $g_1 := 1$. Set $d = \lfloor \deg(f)/2 \rfloor + 1$. Assume that $f \geq 0$ on $S(g) \subset \{\mathbf{x} \in \mathbb{R}^n \,:\, 1 \geq \|\mathbf{x}\|_2^2\}$. Then $f \in \mathbb{R}[\mathbf{x}, x_{n+1}]$ is nonnegative on $S(\mathfrak{g}) \times \mathbb{R}$. Let $\psi(\mathbf{x}, x_{n+1}) := 1 + \|(\mathbf{x}, x_{n+1})\|_2^2$. Putinar–Vasilescu's Positivstellensatz yields that there is $k \in \mathbb{N}$ such that*

$$\psi^k (f + \varepsilon \psi^d) = \sum_{i=1}^m \sigma_i g_i$$

for some SOS σ_i *in* $\mathbb{R}[\mathbf{x}, x_{n+1}]$. *Setting* $x_{n+1} = \sqrt{1 - \|\mathbf{x}\|_2^2}$, *we get* $\psi = 2$ *so*

$$f + \varepsilon 2^d = \frac{1}{2^k} \sum_{i=1}^{m} \sigma_i(x, \sqrt{1 - \|\mathbf{x}\|_2^2}) g_i \,.$$

By removing all terms which are not polynomial, we obtain

$$f + \varepsilon 2^d = \frac{1}{2^k} \sum_{i=1}^{m} [a_i + b_i(1 - \|\mathbf{x}\|_2^2)] g_i \,. \tag{7.1.18}$$

for some SOS a_i, b_i *in* $\mathbb{R}[\mathbf{x}]$. *It is due to the fact that each* $\sigma_i(\mathbf{x}, \sqrt{1 - \|\mathbf{x}\|_2^2})$ *is a sum of the following squares:*

$$(u + v\sqrt{1 - \|\mathbf{x}\|_2^2})^2 = u^2 + v^2(1 - \|\mathbf{x}\|_2^2) + 2uv\sqrt{1 - \|\mathbf{x}\|_2^2} \,,$$

for some $u, v \in \mathbb{R}[\mathbf{x}]$. *Based on the degree bound for Putinar–Vasilescu's Positivstellensatz in Corollary 7.1, we obtain a similar one for representation* (7.1.18).

7.1.3 Preliminary material

This subsection presents some important lemmas that we use to prove the main results.

Given $\Omega \subset \mathbb{R}^n$, the distance of $\mathbf{a} \in \mathbb{R}^n$ to Ω is denoted by $\operatorname{dist}(\mathbf{a}, \Omega)$. Denote by $B(\mathbf{a}, r)$ (resp. $B^\circ(\mathbf{a}, r)$) the closed (resp. open) ball centered at $\mathbf{a} \in \mathbb{R}^n$ with radius $r > 0$.

We recall the Łojasiewicz inequality in the following lemma:

Lemma 7.1. *Let* $g : U \to \mathbb{R}$ *be an analytic function on an open set* $U \subset \mathbb{R}^n$ *and* $Z := \{\mathbf{x} \in K : g(\mathbf{x}) = 0\}$ *for some compact set* $K \subset U$. *Then there exists* $\alpha > 0$ *and* $C > 0$ *such that*

$$d(\mathbf{x}, Z)^\alpha \leq C|g(\mathbf{x})| \,, \forall x \in K \,.$$

As a consequence of Lemma 7.1, we obtain the following result:

Lemma 7.2. *([22, Corollary 2.6.7]) Let* $r > 0$ *and the semialgebraic set* $S := \{\mathbf{x} \in \mathbb{R}^n : g_i(\mathbf{x}) \geq 0 \,, i \in [m]\}$, *where* $g_1, \ldots, g_m$ *are polynomials. Then there exist positive constants* α *and* C *such that, for all* $\mathbf{x}$ *in* $B(\mathbf{0}, r)$,

$$\operatorname{dist}(\mathbf{x}, S)^\alpha \leq -C \min\{g_1(\mathbf{x}), \ldots, g_m(\mathbf{x}), 0\} \,. \tag{7.1.19}$$

Given an open set $U \subset \mathbb{R}^n$ and a differentiable function $\varphi : U \to \mathbb{R}$, denote by $\nabla\varphi(\mathbf{x}) = [\partial_{x_1}\varphi(\mathbf{x}), \ldots, \partial_{x_n}\varphi(\mathbf{x})]$ the gradient of φ at $\mathbf{x} \in U$. Given $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, let $\mathbf{x}' := (x_1, \ldots, x_{n-1})$.

As mentioned in [164, Theorem 3.8], one can prove that $\alpha \in \{1, 2\}$ in Lemma 7.1 if g is a quadratic polynomial. The following lemma states an instance of this result:

Lemma 7.3. *(Łojasiewicz inequality with ice cream constraint) Let* $g := x_n^2 - \|\mathbf{x}'\|_2^2$ *and* $Z := \{\mathbf{x} \in \mathbb{R}^n : g(\mathbf{x}) = 0\}$. *Then for all* $\mathbf{x} \in \mathbb{R}^n$,

$$\operatorname{dist}(\mathbf{x}, Z)^2 \leq \frac{|g(\mathbf{x})|}{2} \,. \tag{7.1.20}$$

Proof. If $\mathbf{x} \in Z$, both sides of (7.1.20) are zeros. Let $\mathbf{x} \in \mathbb{R}^n \backslash Z$ be fixed. Then $d(\mathbf{x}, Z)^2 = \min_{\mathbf{y}}\{\|\mathbf{x} - \mathbf{y}\|_2^2 : g(\mathbf{y}) = 0\}$. Assume that $(y, \mu) \in \mathbb{R}^n \times \mathbb{R}$ satisfies the Karush–Kuhn–Tucker conditions:

$$\begin{cases} \nabla_{\mathbf{y}} \|\mathbf{x} - \mathbf{y}\|_2^2 = \mu \nabla_{\mathbf{y}} g(\mathbf{y}) \,, \\ g(\mathbf{y}) = 0 \,. \end{cases} \tag{7.1.21}$$

The first condition of (7.1.21) implies that $2(\mathbf{x} - \mathbf{y}) = \mu \begin{bmatrix} -2\mathbf{y}' \\ 2y_n \end{bmatrix}$, so $\mathbf{x}' - \mathbf{y}' = -\mu\mathbf{y}'$ and $x_n - y_n = \mu y_n$. Assume that $\mu \notin \{1, -1\}$. Then $\mathbf{y}' = \frac{\mathbf{x}'}{1-\mu}$ and $y_n = \frac{x_n}{1+\mu}$. Since $g(\mathbf{y}) = y_n^2 - \|\mathbf{y}'\|_2^2 = 0$, $y_n = \pm\|\mathbf{y}'\|_2$.

Let us consider the first case $y_n = \|\mathbf{y}'\|_2$. Then $\frac{x_n}{1+\mu} = \frac{\|\mathbf{x}'\|_2}{1-\mu}$. It implies that $\mu = \frac{x_n - \|\mathbf{x}'\|_2}{x_n + \|\mathbf{x}'\|_2}$. Note that $x_n \neq -\|\mathbf{x}'\|_2$ since $g(\mathbf{x}) \neq 0$. From this, $\mathbf{y}' = \frac{(x_n + \|\mathbf{x}'\|_2)\mathbf{x}'}{2\|\mathbf{x}'\|_2}$ and $y_n = \frac{x_n + \|\mathbf{x}'\|_2}{2}$. Thus, $\|\mathbf{x} - \mathbf{y}\|_2^2 = \frac{(x_n - \|\mathbf{x}'\|_2)^2}{2}$.

Similarly, if we consider the case $y_n = -\|\mathbf{y}'\|_2$, then $\|\mathbf{x} - \mathbf{y}\|_2^2 = \frac{(x_n + \|\mathbf{x}'\|_2)^2}{2}$.

Let us consider the case of $\mu \in \{1, -1\}$. Assume that $\mu = 1$. Then $\mathbf{x}' = 0$ and $y_n = \frac{x_n}{2}$. From this and the fact that $0 = g(\mathbf{y}) = y_n^2 - \|\mathbf{y}'\|_2^2$, we obtain $\|\mathbf{y}'\|_2^2 = \frac{x_n^2}{4}$. It implies that $\|\mathbf{x} - \mathbf{y}\|_2^2 = \|\mathbf{y}'\|_2^2 + (x_n - y_n)^2 = \frac{x_n^2}{4} + \frac{x_n^2}{4} = \frac{x_n^2}{2} = \frac{(x_n - \|\mathbf{x}'\|_2)^2}{2}$. Thus, $\|\mathbf{x} - \mathbf{y}\|_2^2 = \frac{(x_n - \|\mathbf{x}'\|_2)^2}{2}$.

Similarly, if we consider the case $\mu = -1$, then $\|\mathbf{x} - \mathbf{y}\|_2^2 = \frac{(x_n + \|\mathbf{x}'\|_2)^2}{2}$.

Thus,

$$\begin{aligned} d(\mathbf{x}, Z)^2 &\leq \tfrac{1}{2}\min\{(x_n - \|\mathbf{x}'\|_2)^2, (x_n + \|\mathbf{x}'\|_2)^2\} \\ &\leq \tfrac{1}{2}\sqrt{(x_n - \|\mathbf{x}'\|_2)^2 (x_n + \|\mathbf{x}'\|_2)^2} \\ &= \tfrac{1}{2}|x_n^2 - \|\mathbf{x}'\|_2^2| = \tfrac{|g(\mathbf{x})|}{2}\,, \end{aligned} \tag{7.1.22}$$

yielding (7.1.20). □

A real-valued function $f : U \to \mathbb{R}$ for some $U \subset \mathbb{R}^n$ is called L-Lipschitz (or Lipschitz) continuous on $K \subset U$ if there exits a real $L > 0$ such that $|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|_2$, for all $x, y \in K$. In this case, L is called the Lipschitz constant of f on K. Given an open set $U \subset \mathbb{R}^n$, a function $f : U \to \mathbb{R}$ is called locally Lipschitz continuous on $K \subset U$ if for every $\mathbf{x} \in K$ there exists a neighborhood $W \subset U$ of x such that f is Lipschitz continuous on $W \cap K$.

The following lemma is similar in spirit to [163, Section 2.4, Lemma 2]:

Lemma 7.4. *Given an open set $U \subset \mathbb{R}^n$, if the function $f : U \to \mathbb{R}$ is locally Lipschitz on a compact set $K \subset U$, then f is Lipschitz on K.*

Proof. Since f is locally Lipschitz on K, for each $\mathbf{x} \in K$ there is some $r_{\mathbf{x}} > 0$ and $L_{\mathbf{x}} > 0$ such that $B(\mathbf{x}, r_{\mathbf{x}}) \subset U$ and f is $L_{\mathbf{x}}$-Lipschitz on $B(\mathbf{x}, r_{\mathbf{x}}) \cap K$. Then the sets $B(\mathbf{x}, \frac{1}{2}r_{\mathbf{x}})$, $\mathbf{x} \in K$ form an open cover of K. Due to the compactness of K, there exists a finite subsequence of $B(\mathbf{x}, \frac{1}{2}r_{\mathbf{x}})$, $\mathbf{x} \in K$ covering K. For convenience, denote these by $B(\mathbf{x}_k, \frac{1}{2}r_k)$ and $L_k := L_{\mathbf{x}_k}$, $k \in [l]$. Let $M := \sup_{\mathbf{x}\in K} |f(\mathbf{x})|$, $r := \frac{1}{2}\min_{k\in[l]} r_k$, $L_0 := \frac{2M}{r}$ and $L := \max\{L_0, L_k : k \in [l]\}$. Then L is a Lipschitz constant of f on K. To see this, pick $\mathbf{x}, \mathbf{y} \in K$. If $\|\mathbf{x} - \mathbf{y}\|_2 \geq r$ then we see that $\frac{|f(\mathbf{x}) - f(\mathbf{y})|}{\|\mathbf{x} - \mathbf{y}\|_2} \leq \frac{2M}{r} = L_0 \leq L$. If $\|\mathbf{x} - \mathbf{y}\|_2 < r$, then for some $\mathbf{x}_k$ we have $\mathbf{x} \in B(\mathbf{x}_k, \frac{1}{2}r_k)$. Then $\mathbf{y} \in B(\mathbf{x}_k, r_k)$ and so $|f(\mathbf{x}) - f(\mathbf{y})| \leq L_k\|\mathbf{x} - \mathbf{y}\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2$. □

Lemma 7.5. *(Kirszbraun's theorem [96]) If U is a subset of $\mathbb{R}^n$ and $f : U \to \mathbb{R}$ is a Lipschitz continuous function, then there is a Lipschitz continuous function $F : \mathbb{R}^n \to \mathbb{R}$ that extends f and has the same Lipschitz constant as f. Moreover the extension is provided by*

$$F(\mathbf{x}) := \inf_{\mathbf{u}\in U}\{f(\mathbf{u}) + L_f\|\mathbf{x} - \mathbf{u}\|_2\}\,, \tag{7.1.23}$$

where L_f is the Lipschitz constant of f on U.

Given $\Omega \subset \mathbb{R}^n$, denote by $C(\Omega)$ the set of continuous functions from Ω to $\mathbb{R}$. We recall basic properties of the multivariate Bernstein polynomials described, e.g., in [80, 73].

Definition 7.1. *(Multivariate Bernstein polynomials) Let $d \in \mathbb{N}^n$ and $f \in C([0,1]^n)$. The polynomials*

$$B_{f,d}(\mathbf{x}) := \textstyle\sum_{k_1=0}^{d_1} \cdots \sum_{k_n=0}^{d_n} f\left(\frac{k_1}{d_1}, \ldots, \frac{k_n}{d_n}\right) \prod_{j=1}^n \left[\binom{d_j}{k_j} x_j^{k_j}(1 - x_j)^{d_j - k_j}\right] \tag{7.1.24}$$

are called the multivariate Bernstein polynomials of f.

Note that $\deg(B_{f,d}) = \sum_{j\in[n]} d_j$ and the binomial identity implies

$$\sup_{\mathbf{x}\in[0,1]^n} |B_{f,d}(\mathbf{x})| \leq \sup_{\mathbf{x}\in[0,1]^n} |f(\mathbf{x})|\,. \tag{7.1.25}$$

Lemma 7.6. *(Error bound) If $f \in C([0,1]^n)$ is L-Lipschitz, namely $|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x}-\mathbf{y}\|_2$ on $[0,1]^n$, then for all $d \in \mathbb{N}^n$, the inequality*

$$|B_{f,d}(\mathbf{x}) - f(\mathbf{x})| \leq \tfrac{L}{2}(\textstyle\sum_{j=1}^n \frac{1}{d_j})^{\frac{1}{2}} \tag{7.1.26}$$

holds for all $\mathbf{x} \in [0,1]^n$.

Proof. Let $\mathbf{x} \in [0,1]^n$ be fixed. To simplify the notation, we define $b_j := \binom{d_j}{k_j} x_j^{k_j}(1-x_j)^{d_j-k_j}$, for $j \in [n]$, and $\mathbf{k} = (\frac{k_1}{d_1}, \ldots, \frac{k_n}{d_n})$. Then, we have:

$$\begin{array}{rl} |B_{f,d}(\mathbf{x}) - f(\mathbf{x})|^2 & \leq (\sum |f(\mathbf{k}) - f(\mathbf{x})| b_1 \ldots b_n)^2 \\ & \leq L^2 \left(\sum \|\mathbf{k}-\mathbf{x}\|_2 b_1 \cdots b_n\right)^2 \\ & \leq L^2 \left(\sum \|\mathbf{k}-\mathbf{x}\|_2^2 b_1 \cdots b_n\right)\left(\sum b_1 \cdots b_n\right) \\ & = L^2 \sum (\sum_r (\frac{k_r}{d_r} - x_r)^2) b_1 \cdots b_n \\ & = L^2 \sum_r (\sum (\frac{k_r}{d_r} - x_r)^2 b_1 \cdots b_n) \\ & = L^2 \sum_r \frac{x_r(1-x_r)}{d_r} \\ & \leq L^2 \sum_r \frac{1}{4d_r}\,. \end{array} \tag{7.1.27}$$

For the first inequality, we have used the multinomial identity $1 = \prod_j [x_j + (1-x_j)]^{d_j} = \sum b_1 \ldots b_n$ and the triangle inequality. For the second one, we have used the fact that f is L-Lipschitz. For the third one, we use that for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ and all $\mathbf{c} \in \mathbb{R}^n_+$, one has $\left(\sum a_k b_k c_k\right)^2 \leq \left(\sum a_k^2 c_k\right)\left(\sum b_k^2 c_k\right)$. The last equality comes from the identities $\sum_{k_r=0}^{d_r}(k_r - d_r x_r)^2 b_r = d_r x_r (1-x_r)$ and $\prod_{j \neq r}(\sum_{k_j=0}^{d_j} b_j) = 1$. The last inequality is obtained by noticing that we have $x_r(1-x_r) \leq 1/4$. □

Let $\mathbf{e} := (1, \ldots, 1) \in \mathbb{R}^n$. As a consequence of Lemma 7.6, we obtain the following result after a change of coordinates.

Lemma 7.7. *If $f \in C([0,1]^n)$ is L-Lipschitz, namely $|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x}-\mathbf{y}\|_2$ on $[-1,1]^n$, then for all $k \in \mathbb{N}^{\geq 1}$, the inequality*

$$\left| B_{y \mapsto f(2\mathbf{y}-\mathbf{e}),k\mathbf{e}}\left(\frac{\mathbf{x}+\mathbf{e}}{2}\right) - f(\mathbf{x})\right| \leq L\left(\frac{n}{k}\right)^{\frac{1}{2}} \tag{7.1.28}$$

holds for all $\mathbf{x} \in [-1,1]^n$. Moreover, we have

$$\sup_{\mathbf{x}\in[-1,1]^n} |B_{\mathbf{y}\mapsto f(2\mathbf{y}-\mathbf{e}),k\mathbf{e}}\left(\tfrac{\mathbf{x}+\mathbf{e}}{2}\right)| \leq \sup_{\mathbf{x}\in[-1,1]^n} |f(\mathbf{x})|\,. \tag{7.1.29}$$

Proof. Define $g : [0,1]^n \to \mathbb{R}$ by $g(\mathbf{x}) := f(2\mathbf{x} - \mathbf{e})$. Let us compute a Lipschitz constant of g. With $\mathbf{x}, \mathbf{y} \in [0,1]^n$, by the Lipschitz continuity of f, we have

$$\begin{array}{rcl} |g(\mathbf{x}) - g(\mathbf{y})| = & |f(2\mathbf{x}-\mathbf{e}) - f(2\mathbf{y}-\mathbf{e})| & \\ \leq & L\|2\mathbf{x}-\mathbf{e}-2\mathbf{y}+\mathbf{e}\|_2 & \\ = & 2L\|\mathbf{x}-\mathbf{y}\|_2\,. & \end{array} \tag{7.1.30}$$

Then $2L$ is a Lipschitz constant of g. Let $k \in \mathbb{N}^{\geq 1}$. Using Lemma 7.6, we get that for all $\mathbf{x} \in [0,1]^n$,

$$|B_{g,k\mathbf{e}}(\mathbf{x}) - g(\mathbf{x})| \leq \tfrac{2L}{2}(\textstyle\sum_{j=1}^n \frac{1}{k})^{\frac{1}{2}} = L(\frac{n}{k})^{\frac{1}{2}}\,. \tag{7.1.31}$$

Let $\mathbf{y} \in [-1,1]^n$. Then $\frac{\mathbf{y}+\mathbf{e}}{2} \in [0,1]$ implies that

$$|B_{g,k\mathbf{e}}(\tfrac{\mathbf{y}+\mathbf{e}}{2}) - f(\mathbf{y})| = |B_{g,k\mathbf{e}}(\tfrac{\mathbf{y}+\mathbf{e}}{2}) - g(\tfrac{\mathbf{y}+\mathbf{e}}{2})| \leq L(\tfrac{n}{k})^{\frac{1}{2}}\,. \tag{7.1.32}$$

yielding (7.1.28).

In addition, from (7.1.25),

$$\begin{array}{rl} \sup_{\mathbf{y}\in[-1,1]} |B_{g,k\mathbf{e}}(\frac{\mathbf{y}+\mathbf{e}}{2})| & = \sup_{\mathbf{x}\in[0,1]^n} |B_{g,k\mathbf{e}}(\mathbf{x})| \\ & \leq \sup_{\mathbf{x}\in[0,1]^n} |g(\mathbf{x})| \\ & = \sup_{\mathbf{y}\in[-1,1]} |g(\frac{\mathbf{y}+\mathbf{e}}{2})| = \sup_{\mathbf{y}\in[-1,1]} |f(\mathbf{y})|\,, \end{array} \tag{7.1.33}$$

which yields (7.1.29). □

7.1.4 The proof of Theorem 7.1

Recall that $[l] := \{1, \ldots, l\}$ for $l \in \mathbb{N}^{\geq 1}$. Given real value functions g, h, we use the notation $\{g * h\} = \{\mathbf{x} \in \mathbb{R}^n \,:\, g(\mathbf{x}) * h(\mathbf{x})\}$, where $* \in \{=, \geq, \leq, >, <\}$. Given a real value function q on $\Omega \subset \mathbb{R}^n$, note $\|q\|_\Omega := \sup_{\mathbf{x}\in\Omega} |q(\mathbf{x})|$. With $\Omega \subset \mathbb{R}^n$, denote by $\mathrm{int}(\Omega)$ the interior of Ω.

Given $U, V \subseteq \mathbb{R}^n$ and $r \in \mathbb{R}$, note $U + V = \{\mathbf{u} + \mathbf{v} \,:\, \mathbf{u} \in U\,,\, \mathbf{v} \in V\}$ and $rU = \{r\mathbf{u} \,:\, \mathbf{u} \in U\}$. Given a function $f : U \to \mathbb{R}$ and $A \subset U \subset \mathbb{R}^n$ such that $A = -A$, f is called even on A if $f(-\mathbf{x}) = f(\mathbf{x})$ for all $\mathbf{x} \in A$.

To begin the proof, let us fix $\varepsilon > 0$. By assumption, $\deg(f) = 2d_f$, $\deg(g_i) = 2d_{g_i}$ for some $d, d_{g_i} \in \mathbb{N}$, for $i \in [m]$.

Construction of the positive weight functions

For $j \in [m]$, define

$$S_j := \{\mathbf{x} \in \mathbb{R}^n \,:\, g_i(\mathbf{x}) \geq 0\,,\, i \in [j]\}\,. \tag{7.1.34}$$

Obviously, we have $S_m = S$. Note $S_0 := \mathbb{R}^n$ and $f_m := f$.

We will prove that there exist functions $\bar{\varphi}_m : \mathbb{R}^n \to \mathbb{R}$ such that the following conditions hold:

1. $\bar{\varphi}_m$ is positive, even and bounded from above by $C_{\bar{\varphi}_m} = \bar{r}_m \varepsilon^{-r_m}$ on $B(\mathbf{0}, \sqrt{n} + m)$ for some positive constants $\bar{r}_m$ and r_m independent of ε.
2. $\bar{\varphi}_m$ is Lipschitz with Lipschitz constant $L_{\bar{\varphi}_m} = \bar{t}_j \varepsilon^{-t_m}$ for some positive constants $\bar{t}_m$ and t_m independent of ε.
3. $f_{m-1} := f_m + \frac{\varepsilon}{2} - \bar{\varphi}_m^2 g_m$ satisfies:
   (a) $f_{m-1} \geq 0$ on $S_{m-1} \cap B(\mathbf{0}, \sqrt{n} + m - 1)$;
   (b) $f_{m-1} \leq C_{f_{m-1}}$ on $B(\mathbf{0}, \sqrt{n}+m)$, where $C_{f_{m-1}} = \bar{c}_{m-1}\varepsilon^{-c_{m-1}}$ for some positive constants $\bar{c}_{m-1}$ and c_{m-1} independent of ε;
   (c) f_{m-1} is Lipschitz on $B(\mathbf{0}, \sqrt{n} + m)$ with Lipschitz constant $L_{f_{m-1}} = \bar{l}_{m-1}\varepsilon^{-l_{m-1}}$ for some positive constants $\bar{l}_{m-1}$ and l_{m-1} independent of ε.

Let

$$M_m := \inf_{\mathbf{x}\in S_m \cap B(\mathbf{0}, \sqrt{n}+m)} \frac{f(\mathbf{x}) + \frac{\varepsilon}{2}}{g_m(\mathbf{x})}\,. \tag{7.1.35}$$

The constant M_m is a positive real number. Let $C_{g_m} = \|g_m\|_{B(\mathbf{0}, \sqrt{n}+m)}$. We claim that $\frac{\varepsilon}{2C_{g_m}} < M_m < \infty$. Indeed, if $\mathbf{z}$ is a feasible solution of (7.1.35), $\mathbf{z} \in S$ yielding $f(\mathbf{z}) \geq 0$ so that

$$\frac{f(\mathbf{z}) + \frac{\varepsilon}{2}}{g_m(\mathbf{z})} \geq \frac{\varepsilon}{2g_m(\mathbf{z})} \geq \frac{\varepsilon}{2C_{g_m}}\,. \tag{7.1.36}$$

From this, we have $M_m > \frac{\varepsilon}{2C_{g_m}}$. On the other hand, there exists $\mathbf{a} \in \mathbb{R}^n$ such that $g_i(\mathbf{a}) > 0$ for $i \in [m]$ since S has nonempty interior. For $i \in [m]$, since g_i is homogeneous, $\mathbf{a} = \mathbf{0}$ yields $g_i(\mathbf{a}) = 0$. It implies that $\mathbf{a} \neq 0$. With $\bar{\mathbf{a}} = \frac{\mathbf{a}}{\|\mathbf{a}\|_2} \in \mathbb{B}^n \subset B(\mathbf{0}, \sqrt{n} + m)$, we obtain $g_i(\bar{\mathbf{a}}) > 0$ for $i \in [m]$ since

$$g_i(\bar{\mathbf{a}}) = g_i\left(\frac{\mathbf{a}}{\|\mathbf{a}\|_2}\right) = \frac{g_i(\mathbf{a})}{\|\mathbf{a}\|_2^{2d_{g_i}}} > 0\,,\, \forall i \in [m]\,. \tag{7.1.37}$$

Thus, $\bar{\mathbf{a}}$ is a feasible solution of (7.1.35) which yields

$$\frac{\varepsilon}{2C_{g_m}} \leq M_m \leq \frac{f(\bar{\mathbf{a}}) + \frac{\varepsilon}{2}}{g_m(\bar{a})} \leq \frac{C_f + \frac{\varepsilon}{2}}{g_m(\bar{\mathbf{a}})} < \infty\,, \tag{7.1.38}$$

where $C_f := \|f\|_{B(\mathbf{0}, \sqrt{n}+m)}$.

Let $\psi_m : \mathbb{R}^n \to \mathbb{R}$ be the function defined by

$$\psi_m(\mathbf{x}) := \begin{cases} \max\{M_m, \frac{f(\mathbf{x})+\frac{\varepsilon}{2}}{g_m(\mathbf{x})}\} & \text{if } g_m(\mathbf{x}) < 0\,, \\ M_m & \text{otherwise.} \end{cases} \tag{7.1.39}$$

The function $f+\frac{\varepsilon}{2}-\psi_m g_m$ is nonnegative on $S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m)$. Namely, we claim that

$$f+\frac{\varepsilon}{2}-\psi_m g_m \geq 0 \text{ on } S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m)\,. \tag{7.1.40}$$

Let $\mathbf{y}\in S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m)$. If $g_m(\mathbf{y})<0$, then

$$\begin{array}{rcl} f(\mathbf{y})+\frac{\varepsilon}{2}-\psi_m(\mathbf{y})g_m(\mathbf{y}) = & f(\mathbf{y})+\frac{\varepsilon}{2}-g_m(\mathbf{y})\max\{M_m,\frac{f(\mathbf{y})+\frac{\varepsilon}{2}}{g_m(\mathbf{y})}\} \\ \geq & f(\mathbf{y})+\frac{\varepsilon}{2}-g_m(\mathbf{y})\frac{f(\mathbf{y})+\frac{\varepsilon}{2}}{g_m(\mathbf{y})}=0\,. \end{array} \tag{7.1.41}$$

Otherwise, $g_m(\mathbf{y})\geq 0$ gives

$$\begin{array}{ll} f(\mathbf{y})+\frac{\varepsilon}{2}-\psi_m(\mathbf{y})g_m(\mathbf{y}) & = f(\mathbf{y})+\frac{\varepsilon}{2}-g_m(\mathbf{y})M_m \\ & \begin{cases} \geq f(\mathbf{y})+\frac{\varepsilon}{2}-g_m(\mathbf{y})\frac{f(\mathbf{y})+\frac{\varepsilon}{2}}{g_m(\mathbf{y})}=0 & \text{if } g_m(\mathbf{y})>0\,, \\ = f(\mathbf{y})+\frac{\varepsilon}{2}\geq 0 & \text{if } g_m(\mathbf{y})=0\,, \end{cases} \end{array} \tag{7.1.42}$$

since $\mathbf{y}\in S$ is a feasible solution of (7.1.35).

The function ψ_m is positive, even on $B(\mathbf{0},\sqrt{n}+m)$ and continuous on $S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m)$. It is easy to see that ψ_m is bounded from below by M_m and continuous on $B(\mathbf{0},\sqrt{n}+m)\backslash\{g_m=0\}$ since the max function $(t_1,t_2)\mapsto\max\{t_1,t_2\}$ is continuous.

We claim that ψ_m is continuous on $S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m)\cap\{g_m=0\}$. Indeed, let us consider a sequence $(y_l)_l\subset S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m)\cap\{g_m<0\}$ such that $\mathbf{y}_l\to\bar{\mathbf{y}}\in S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m)\cap\{g_m=0\}$. Then $g_m(\mathbf{y}_l)\to 0^-$ and $f(\mathbf{y}_l)\to f(\bar{\mathbf{y}})\geq 0$ (since $\bar{\mathbf{y}}\in S$) yielding that $\frac{f(\mathbf{y}_l)+\frac{\varepsilon}{2}}{g_m(\mathbf{y}_l)}\to-\infty$. It implies that $\max\{M_m,\frac{f(\mathbf{y}_l)+\frac{\varepsilon}{2}}{g_m(\mathbf{y}_l)}\}\to M_m$. Thus, $\psi_m=M_m$ on a sufficiently small neighborhood of any point in $S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m)\cap\{g_m=0\}$. On the other hand, ψ_m is even, i.e., $\psi_m(\mathbf{x})=\psi_m(-\mathbf{x})$ due to the fact that $f,g_1,\ldots,g_m$ are even and $B(\mathbf{0},\sqrt{n}+m)=-B(\mathbf{0},\sqrt{n}+m)$.

The upper bound of ψ_m depends on ε. It follows from (7.1.38) that $\psi_m=M_m$ on $B(\mathbf{0},\sqrt{n}+m)\cap\{g_m\geq 0\}$ and so is bounded from above by $\frac{f(\bar{\mathbf{a}})+\frac{\varepsilon}{2}}{g_m(\bar{\mathbf{a}})}$.

Let us compute an upper bound of ψ_m on $S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m)\cap\{g_m<0\}$. Let $\mathbf{y}\in S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m)$ be such that $g_m(\mathbf{y})<0$ and $\frac{f(\mathbf{y})+\frac{\varepsilon}{2}}{g_m(\mathbf{y})}>M_m$. Then $\psi_m(\mathbf{y})=\frac{f(\mathbf{y})+\frac{\varepsilon}{2}}{g_m(\mathbf{y})}$. By using the Łojasiewicz inequality (see Lemma 7.2), there exist $C_m>0$ and $\alpha_m>0$ depending on $g_1,\ldots,g_m$ such that for all $\mathbf{x}\in S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m)\cap\{g_m<0\}$,

$$\operatorname{dist}(\mathbf{x},S)^{\alpha_m}\leq -C_m\min\{g_1(\mathbf{x}),\ldots,g_m(\mathbf{x}),0\}=-C_m g_m(\mathbf{x})\,. \tag{7.1.43}$$

Let $\delta_m=\frac{1}{C_m}(\frac{\varepsilon}{2L_f})^{\alpha_m}$, where L_f is a Lipschitz constant of f on $B(\mathbf{0},\sqrt{n}+m)$. Consider the following two cases:

- Case 1: $g_m(\mathbf{y})\leq-\delta_m<0$. Then

$$\psi_m(\mathbf{y})=\frac{f(\mathbf{y})+\frac{\varepsilon}{2}}{g_m(\mathbf{y})}=\frac{-f(\mathbf{y})-\frac{\varepsilon}{2}}{-g_m(\mathbf{y})}\leq\frac{C_f}{-g_m(\mathbf{y})}\leq\frac{C_f}{\delta_m}\leq C_m C_f\left(\frac{2L_f}{\varepsilon}\right)^{\alpha_m}. \tag{7.1.44}$$

- Case 2: $-\delta_m\leq g_m(\mathbf{y})<0$. Let $\mathbf{z}\in S$ such that $\operatorname{dist}(\mathbf{y},S)=\|\mathbf{y}-\mathbf{z}\|_2$. Then (7.1.43) turns to $-f(\mathbf{y})\leq\frac{\varepsilon}{2}$ according to

$$\begin{array}{rl} -f(\mathbf{y}) & \leq -f(\mathbf{z})+L_f\|\mathbf{y}-\mathbf{z}\|_2\leq L_f\operatorname{dist}(\mathbf{y},S) \\ & \leq L_f(-C_m g_m(\mathbf{y}))^{\frac{1}{\alpha_m}}\leq L_f(C_m\delta_m)^{\frac{1}{\alpha_m}}=\frac{\varepsilon}{2}\,. \end{array} \tag{7.1.45}$$

From this, we obtain

$$M_m<\frac{f(\mathbf{y})+\frac{\varepsilon}{2}}{g_m(\mathbf{y})}=\frac{-f(\mathbf{y})-\frac{\varepsilon}{2}}{-g_m(\mathbf{y})}\leq\frac{\frac{\varepsilon}{2}-\frac{\varepsilon}{2}}{-g_m(\mathbf{y})}=0<M_m\,. \tag{7.1.46}$$

The contradiction indicates that this case does not occur.

Thus, the bound is given as follows

$$\sup_{\mathbf{x}\in S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m)} \psi_m(\mathbf{x}) \leq \max\left\{\frac{f(\bar{\mathbf{a}})+\frac{\varepsilon}{2}}{g_m(\bar{\mathbf{a}})}, C_m C_f \left(\frac{2L_f}{\varepsilon}\right)^{\alpha_m}\right\} =: C_{\psi_m}\,. \tag{7.1.47}$$

Moreover, we obtain the inclusion

$$S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m)\cap\{\xi_m \geq M_m\}\cap\{g_m\leq 0\}\subset\{g_m\leq -\delta_m\}\,, \tag{7.1.48}$$

where $\xi_m(\mathbf{x}) = \frac{f(\mathbf{x})+\frac{\varepsilon}{2}}{g_m(\mathbf{x})}$. Let φ_m be the square root of ψ_m, i.e., $\varphi_m(\mathbf{x}) := \sqrt{\psi_m(\mathbf{x})}$. Then φ_m is well-defined on $B(\mathbf{0},\sqrt{n}+m)$ since ψ_m is positive. Moreover, φ_m is finitely bounded from above on $S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m)$ by $C_{\varphi_m} := \sqrt{C_{\psi_m}}$ and φ_m is continuous on $S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m)$ since ξ_m is continuous on $S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m)$.

The function φ_m is Lipschitz continuous on $S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m-1)$. Keep in mind that ψ_m is defined by the constant function M_m and the function ξ_m. Since φ_m takes the constant value $\sqrt{M_m}$ on $B(\mathbf{0},\sqrt{n}+m)\backslash(\{\xi_m\geq M_m\}\cap\{g_m\leq 0\})$, φ_m is Lipschitz continuous on $B(\mathbf{0},\sqrt{n}+m)\backslash(\{\varphi_m\geq M_m\}\cap\{g_m\leq 0\})$ with zero Lipschitz constant.

On the other hand, $\varphi_m = \sqrt{\xi_m}$ on $B(\mathbf{0},\sqrt{n}+m)\cap\{\xi_m\geq M_m\}\cap\{g_m\leq 0\}$. As a consequence of (7.1.48), we have

$$\begin{array}{rl} & S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m)\cap\{\xi_m\geq M_m\}\cap\{g_m\leq 0\} \\ = & S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m)\cap\{\xi_m\geq M_m\}\cap\{g_m\leq -\delta_m\}\,. \end{array} \tag{7.1.49}$$

It implies that

$$\varphi_m(\mathbf{x}) = \begin{cases} \sqrt{\xi_m(\mathbf{x})} & \text{if } x\in S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m)\cap\{\xi_m\geq M_m\}\cap\{g_m\leq -\delta_m\}\,, \\ \sqrt{M_m} & \text{if } x\in (S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m))\backslash(\{\xi_m\geq M_m\}\cap\{g_m\leq -\delta_m\})\,. \end{cases} \tag{7.1.50}$$

The second equality is due to the fact that $\varphi_m = \sqrt{M_m}$ on $(S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m))\backslash(\{\xi_m\geq M_m\}\cap\{g_m\leq 0\})$ and

$$\begin{array}{rl} & (S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m))\backslash(\{\xi_m\geq M_m\}\cap\{g_m\leq 0\}) \\ = & (S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m))\backslash[S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m)\cap\{\xi_m\geq M_m\}\cap\{g_m\leq 0\}] \\ = & (S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m))\backslash[S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m)\cap\{\xi_m\geq M_m\}\cap\{g_m\leq -\delta_m\}] \\ = & (S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m))\backslash(\{\xi_m\geq M_m\}\cap\{g_m\leq -\delta_m\})\,. \end{array} \tag{7.1.51}$$

Let L_{g_m} be a Lipschitz constant of g_m on $B(\mathbf{0},\sqrt{n}+m)$. Set

$$w_m := \min\left\{1, \frac{\delta_m}{2L_{g_m}}, \frac{\varepsilon\delta_m^2}{8C_{g_m}[L_f C_{g_m} + (C_f+\frac{\varepsilon}{2})L_{g_m}]}\right\}\,. \tag{7.1.52}$$

and

$$W_m := \big(B(\mathbf{0},\sqrt{n}+m-1)\cap\{\xi_m\geq M_m\}\cap\{g_m\leq -\delta_m\}\big) + w_m\mathbb{B}^n\,. \tag{7.1.53}$$

Then $B(\mathbf{0},\sqrt{n}+m-1)\cap\{\xi_m\geq M_m\}\cap\{g_m\leq -\delta_m\}\subset W_m$. Next, we prove that

$$W_m\subset B(\mathbf{0},\sqrt{n}+m)\cap\{\xi_m\geq\frac{M_m}{2}\}\cap\{g_m\leq -\frac{\delta_m}{2}\}\,. \tag{7.1.54}$$

Let $\mathbf{y}\in W_m$. Then $\mathbf{y}=\mathbf{z}+w_m\mathbf{u}$ for some $\mathbf{z}\in S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m-1)\cap\{\xi_m\geq M_m\}\cap\{g_m\leq -\delta_m\}$ and for some $\mathbf{u}\in\mathbb{B}^n$. Combining $\|\mathbf{z}\|_2\leq\sqrt{n}+m-1$, $0<w_m<1$ and $\|\mathbf{u}\|_2\leq 1$, one has $\|\mathbf{y}\|_2\leq\|\mathbf{z}\|_2+w_m\|\mathbf{u}\|_2\leq\sqrt{n}+m$, yielding $\mathbf{y}\in B(\mathbf{0},\sqrt{n}+m)$. Since $g_m(\mathbf{z})\leq -\delta_m$, we have

$$g_m(\mathbf{y})\leq g_m(\mathbf{z})+L_{g_m}\|\mathbf{y}-\mathbf{z}\|_2\leq -\delta_m + L_{g_m}w_m\|\mathbf{u}\|_2\leq -\delta_m + L_{g_m}\frac{\delta_m}{2L_{g_m}}\leq -\frac{\delta_m}{2}\,. \tag{7.1.55}$$

Thus $y \in \{g_m \leq -\frac{\delta_m}{2}\}$. This in turn implies

$$\begin{array}{rl} & |\xi_m(\mathbf{y}) - \xi_m(\mathbf{z})| \\ = & \left| \frac{f(\mathbf{y})+\frac{\varepsilon}{2}}{g_m(\mathbf{y})} - \frac{f(\mathbf{z})+\frac{\varepsilon}{2}}{g_m(\mathbf{z})} \right| \\ = & \frac{|(f(\mathbf{y})+\frac{\varepsilon}{2})g_m(\mathbf{z}) - (f(\mathbf{z})+\frac{\varepsilon}{2})g_m(\mathbf{y})|}{|g_m(\mathbf{y})||g_m(\mathbf{z})|} \\ \leq & \frac{2}{\delta_m^2}|(f(\mathbf{y})+\frac{\varepsilon}{2} - f(\mathbf{z}) - \frac{\varepsilon}{2})g_m(\mathbf{z}) + (f(\mathbf{z})+\frac{\varepsilon}{2})(g_m(\mathbf{z}) - g_m(\mathbf{y}))| \\ \leq & \frac{2}{\delta_m^2}[|f(\mathbf{y}) - f(\mathbf{z})||g_m(\mathbf{z})| + (|f(\mathbf{z})| + \frac{\varepsilon}{2})|g_m(\mathbf{z}) - g_m(\mathbf{y})|] \\ \leq & \frac{2}{\delta_m^2}[L_f\|\mathbf{y}-\mathbf{z}\|_2 C_{g_m} + (C_f + \frac{\varepsilon}{2})L_{g_m}\|\mathbf{z}-\mathbf{y}\|_2] \\ \leq & \frac{2}{\delta_m^2}[L_f C_{g_m} + (C_f + \frac{\varepsilon}{2})L_{g_m}]w_m\|u\|_2 \leq \frac{\varepsilon}{4C_{g_m}} \leq \frac{M_m}{2}\,. \end{array} \tag{7.1.56}$$

Since $\xi_m(\mathbf{z}) \geq M_m$, we obtain $\xi_m(\mathbf{y}) \geq \xi_m(\mathbf{z}) - |\xi_m(\mathbf{y}) - \xi_m(\mathbf{z})| \geq M_m - \frac{M_m}{2} = \frac{M_m}{2}$, yielding $\mathbf{y} \in \{\xi_m \geq \frac{M_m}{2}\}$, which concludes the proof of (7.1.54) and ensures that $\sqrt{\xi_m}$ is well-defined on W_m.

Let us prove that $\sqrt{\xi_m}$ is Lipschitz on W_m. Let $\mathbf{y}, \mathbf{z} \in W_m$ such that $\mathbf{y} \neq \mathbf{z}$. Then

$$\begin{array}{rl} & \frac{|\sqrt{\xi_m(\mathbf{y})} - \sqrt{\xi_m(\mathbf{z})}|}{\|\mathbf{y}-\mathbf{z}\|_2} \\ = & \frac{|\xi_m(\mathbf{y}) - \xi_m(\mathbf{z})|}{\|\mathbf{y}-\mathbf{z}\|_2(\sqrt{\xi_m(\mathbf{y})} + \sqrt{\xi_m(\mathbf{z})})} \\ \leq & \frac{\left| \frac{f(\mathbf{y})+\frac{\varepsilon}{2}}{g_m(\mathbf{y})} - \frac{f(\mathbf{z})+\frac{\varepsilon}{2}}{g_m(\mathbf{z})} \right|}{2\sqrt{\frac{M_m}{2}}\|\mathbf{y}-\mathbf{z}\|_2} \\ \leq & \frac{|(f(\mathbf{y})+\frac{\varepsilon}{2})g_m(\mathbf{z}) - (f_m(\mathbf{z})+\frac{\varepsilon}{2})g_m(\mathbf{y})|}{2g_m(\mathbf{y})g_m(\mathbf{z})\sqrt{\frac{\varepsilon}{4C_{g_m}}}\|\mathbf{y}-\mathbf{z}\|_2} \\ \leq & \frac{2|(f(\mathbf{y})+\frac{\varepsilon}{2})g_m(\mathbf{z}) - (f_m(\mathbf{z})+\frac{\varepsilon}{2})g_m(\mathbf{y})|}{\delta_m^2\sqrt{\frac{\varepsilon}{4C_{g_m}}}\|\mathbf{y}-\mathbf{z}\|_2} \\ = & \frac{2|(f(\mathbf{y})+\frac{\varepsilon}{2} - f(\mathbf{z}) - \frac{\varepsilon}{2})g_m(\mathbf{z}) + (f(\mathbf{z})+\frac{\varepsilon}{2})(g_m(\mathbf{z}) - g_m(\mathbf{y}))|}{\delta_m^2\sqrt{\frac{\varepsilon}{4C_{g_m}}}\|\mathbf{y}-\mathbf{z}\|_2} \\ \leq & \frac{2[|f(\mathbf{y}) - f(\mathbf{z})||g_m(\mathbf{z})| + (|f(\mathbf{z})| + \frac{\varepsilon}{2})|g_m(\mathbf{z}) - g_m(\mathbf{y})|]}{\delta_m^2\sqrt{\frac{\varepsilon}{4C_{g_m}}}\|\mathbf{y}-\mathbf{z}\|_2} \\ \leq & \frac{2[L_f\|\mathbf{y}-\mathbf{z}\|_2 C_{g_m} + (C_f + \frac{\varepsilon}{2})L_{g_m}\|\mathbf{z}-\mathbf{y}\|_2]}{\delta_m^2\sqrt{\frac{\varepsilon}{4C_{g_m}}}\|\mathbf{y}-\mathbf{z}\|_2} \\ \leq & \frac{2[L_f C_{g_m} + (C_f + \frac{\varepsilon}{2})L_{g_m}]}{\delta_m^2\sqrt{\frac{\varepsilon}{4C_{g_m}}}} =: L_{\sqrt{\xi_m}}\,. \end{array} \tag{7.1.57}$$

Thus, $L_{\sqrt{\xi_m}}$ is a Lipschitz constant of $\sqrt{\xi_m}$ on W_m.

Set $K := S_{m-1} \cap B(\mathbf{0}, \sqrt{n}+m-1)$, $K_1 := K \cap \{\xi_m \geq M_m\} \cap \{g_m \leq -\delta_m\}$ and $K_2 := K \backslash (\{\xi_m \geq M_m\} \cap \{g_m \leq -\delta_m\})$. Note that $K = K_1 \cup K_2$ and $K_1 \cap K_2 = \emptyset$. From (7.1.50), $\varphi_m = \sqrt{\xi_m}$ on K_1 and $\varphi_m = \sqrt{M_m}$ on K_2.

To conclude that φ_m is Lipschitz on K according to Lemma 7.4 (see Figure 7.1), it is sufficient to prove that φ_m is locally Lipschitz on K.

Explicitly, we will show that for all $\mathbf{z} \in K$, φ_m is Lipschitz on $B(\mathbf{z}, \frac{w_m}{2}) \cap K$ with Lipschitz constant $L_{\sqrt{\xi_m}}$. Let $\mathbf{z} \in K$. Let $u, v \in B(\mathbf{z}, \frac{w_m}{2}) \cap K$ and consider the following cases:

- Case 1: $\mathbf{u}, \mathbf{v} \in K_1$. Then $\mathbf{u}, \mathbf{v} \in W_m$ by definition of W_m. Moreover, $\varphi_m(\mathbf{u}) = \sqrt{\xi_m(\mathbf{u})}$ and $\varphi_m(\mathbf{v}) = \sqrt{\xi_m(\mathbf{v})}$. In this case, by the Lipschitz continuity of $\sqrt{\xi_m}$ on W_m,

$$|\varphi_m(\mathbf{u}) - \varphi_m(\mathbf{v})| = |\sqrt{\xi_m(\mathbf{u})} - \sqrt{\xi_m(\mathbf{v})}| \leq L_{\sqrt{\xi_m}}\|\mathbf{u}-\mathbf{v}\|_2\,. \tag{7.1.58}$$

- Case 2: $\mathbf{u}, \mathbf{v} \in K_2$. In this case, $\varphi_m(\mathbf{u}) = \varphi_m(\mathbf{v}) = \sqrt{M_m}$, so that

$$|\varphi_m(\mathbf{u}) - \varphi_m(\mathbf{v})| = 0 \leq L_{\sqrt{\xi_m}}\|\mathbf{u}-\mathbf{v}\|_2\,. \tag{7.1.59}$$

- Case 3: $\mathbf{u} \in K_1$ and $\mathbf{v} \in K_2$. We claim that $B(\mathbf{z}, \frac{w_m}{2}) \subset W_m$. Let $\mathbf{q} \in B(\mathbf{z}, \frac{w_m}{2})$. Then $\|\mathbf{q}-\mathbf{u}\|_2 \leq \|\mathbf{q}-\mathbf{z}\|_2 + \|\mathbf{z}-\mathbf{u}\|_2 \leq w_m$ yielding $\mathbf{q} \in \mathbf{u} + w_m\mathbb{B}^n \subset K_1 + w_m\mathbb{B}^n \subset W_m$. Then $\mathbf{u}, \mathbf{v} \in B(\mathbf{z}, \frac{w_m}{2}) \subset W_m$. Moreover, $\varphi_m(\mathbf{u}) = \sqrt{\xi_m(\mathbf{u})}$ and $\varphi_m(\mathbf{v}) = \sqrt{M_m}$. According to the continuity of ξ_m on $B(\mathbf{z}, \frac{w_m}{2}) \subset W_m$ and the convexity of $B(\mathbf{z}, \frac{w_m}{2})$, there exists

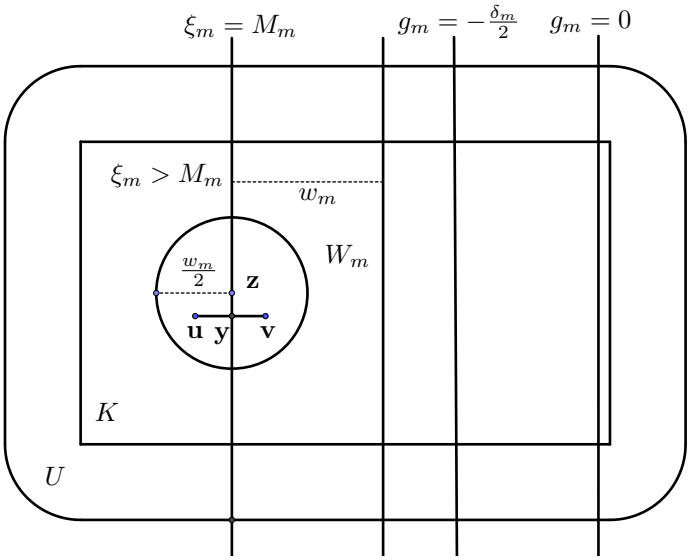


Figure 7.1: Illustration for the proof of the Lipschitz continuity of φ_m on K (rectangle). Here $K = S_{m-1} \cap B(\mathbf{0}, \sqrt{n} + m - 1)$ and $U = K + \frac{w_m}{2} B^\circ(\mathbf{0}, 1)$ with the notation of Lemma 7.4.

$\mathbf{y} \in B(\mathbf{z}, \frac{w_m}{2}) \cap \{\xi_m = M_m\} \cap \{t\mathbf{u} + (1-t)\mathbf{v} \,:\, t \in [0,1]\}$. Then with $\mathbf{y} = \lambda\mathbf{u} + (1-\lambda)\mathbf{v}$ for some $\lambda \in [0,1]$, we have

$$\begin{aligned} |\varphi_m(\mathbf{u}) - \varphi_m(\mathbf{v})| \leq\ & |\varphi_m(\mathbf{u}) - \varphi_m(\mathbf{y})| + |\varphi_m(\mathbf{y}) - \varphi_m(\mathbf{v})| \\ \leq\ & |\sqrt{\xi_m(\mathbf{u})} - \sqrt{\xi_m(\mathbf{y})}| + |\sqrt{M_m} - \sqrt{M_m}| \\ \leq\ & L_{\sqrt{\xi_m}} \|\mathbf{u} - \mathbf{y}\|_2 \\ \leq\ & L_{\sqrt{\xi_m}} \|\mathbf{u} - \lambda\mathbf{u} - (1-\lambda)\mathbf{v}\|_2 \\ \leq\ & L_{\sqrt{\xi_m}} (1-\lambda) \|\mathbf{u} - \mathbf{v}\|_2 \leq L_{\sqrt{\xi_m}} \|\mathbf{u} - \mathbf{v}\|_2 \,. \end{aligned} \tag{7.1.60}$$

From the proof of Lemma 7.4, the Lipschitz constant of φ_m on K is given by

$$L_{\bar{\varphi}_m} := \max \left\{ \frac{4C_{\varphi_m}}{w_m}, L_{\sqrt{\xi_m}} \right\}, \tag{7.1.61}$$

Here we have covered K by a finite sequence of balls with radii $\frac{w_m}{2}$ and centers lying on K.

The function φ_m has a Lipschitz continuous extension $\bar{\varphi}_m$. Let $\bar{\varphi}_m : \mathbb{R}^n \to \mathbb{R}$ be the function defined by

$$\bar{\varphi}_m(\mathbf{x}) := \inf_{\mathbf{y}} \{\varphi_m(\mathbf{y}) + L_{\bar{\varphi}_m} \|\mathbf{x} - \mathbf{y}\|_2 \,:\, y \in S_{m-1} \cap B(\mathbf{0}, \sqrt{n} + m - 1)\} \,. \tag{7.1.62}$$

By Kirszbraun's theorem (stated in Lemma 7.5), $\bar{\varphi}_m$ is Lipschitz continuous with Lipschitz constant $L_{\bar{\varphi}_m}$ and $\bar{\varphi}_m = \varphi_m$ on $S_{m-1} \cap B(\mathbf{0}, \sqrt{n} + m - 1)$.

The function $\bar{\varphi}_m$ is even, positive and has a finite upper bound on $B(\mathbf{0}, \sqrt{n} + m)$ depending on ε. Let us prove that $\bar{\varphi}_m$ is even. Consider

$$\bar{\varphi}_m(-\mathbf{x}) = \inf_{\mathbf{y}} \{\varphi_m(\mathbf{y}) + L_{\bar{\varphi}_m} \| - \mathbf{x} - \mathbf{y}\|_2 \,:\, \mathbf{y} \in S_{m-1} \cap B(\mathbf{0}, \sqrt{n} + m - 1)\} \,. \tag{7.1.63}$$

Let $\mathbf{y}$ be any feasible solution of (7.1.63). Since $g_1, \ldots, g_{m-1}$ are even, $S_{m-1} \cap B(\mathbf{0}, \sqrt{n} + m)$ is symmetric, i.e., $S_{m-1} \cap B(\mathbf{0}, \sqrt{n} + m) = -S_{m-1} \cap B(\mathbf{0}, \sqrt{n} + m)$, it turns out that $-\mathbf{y}$ is a feasible

solution of (7.1.63). Thus,

$$\begin{array}{rcl}\bar{\varphi}_m(-\mathbf{x}) = & \inf_{-\mathbf{y}}\{\varphi_m(-\mathbf{y}) + L_{\bar{\varphi}_m}\|-\mathbf{x}+\mathbf{y}\|_2 \,:\, -\mathbf{y}\in S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m-1)\} \\ = & \inf_{\mathbf{y}}\{\varphi_m(\mathbf{y}) + L_{\bar{\varphi}_m}\|\mathbf{y}-\mathbf{x}\|_2 \,:\, y\in S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m-1)\} = \bar{\varphi}_m(\mathbf{x})\,,\end{array} \tag{7.1.64}$$

where the latter inequality is due to the fact that φ_m is even (since ξ_m, g_m are even). From this, $\bar{\varphi}_m$ is even. It is not hard to show that $\bar{\varphi}_m \geq \sqrt{M_m}$ since $\varphi_m \geq \sqrt{M_m}$.

Let us estimate the upper bound of $\bar{\varphi}_m$ on $B(\mathbf{0},\sqrt{n}+m)$. Let $\mathbf{x}\in B(\mathbf{0},\sqrt{n}+m)$ and $\mathbf{y}\in S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m-1)$. From (7.1.62), we get

$$\bar{\varphi}_m(\mathbf{x}) \leq \varphi_m(\mathbf{y}) + L_{\bar{\varphi}_m}\|\mathbf{x}-\mathbf{y}\|_2 \leq C_{\varphi_m} + (2(\sqrt{n}+m)-1)L_{\bar{\varphi}_m} =: C_{\bar{\varphi}_m}\,. \tag{7.1.65}$$

Thus,

$$\sup_{\mathbf{x}\in B(\mathbf{0},\sqrt{n}+m)} \bar{\varphi}_m(\mathbf{x}) \leq C_{\bar{\varphi}_m}\,. \tag{7.1.66}$$

Set $f_{m-1} := f + \frac{\varepsilon}{2} - \bar{\varphi}_m^2 g_m$.

From (7.1.40) and since $\bar{\varphi}_m = \varphi_m = \sqrt{\psi_m}$ on $S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m-1)$, $f_{m-1}\geq 0$ on $S_{m-1}\cap B(\mathbf{0},\sqrt{n}+m-1)$. Since $\bar{\varphi}_m$ is Lipschitz continuous, f_{m-1} is Lipschitz continuous on $B(\mathbf{0},\sqrt{n}+m)$.

A bound and a Lipschitz constant of f_{m-1} on $B(\mathbf{0},\sqrt{n}+m)$ both depend on ε. Let us compute an upper bound of $|f_{m-1}|$ on $B(\mathbf{0},\sqrt{n}+m)$. Let $y\in B(\mathbf{0},\sqrt{n}+m)$. Then

$$|f_{m-1}(\mathbf{y})| \leq |f(\mathbf{y})| + \tfrac{\varepsilon}{2} + \bar{\varphi}_m(\mathbf{y})^2|g_m(\mathbf{y})| \leq C_f + \tfrac{\varepsilon}{2} + C_{g_m}C_{\bar{\varphi}_m}^2 =: C_{f_{m-1}}\,. \tag{7.1.67}$$

Thus,

$$\|f_{m-1}\|_{B(\mathbf{0},\sqrt{n}+m)} \leq C_{f_{m-1}}\,. \tag{7.1.68}$$

We now estimate the Lipschitz constant of f_{m-1} on $B(\mathbf{0},\sqrt{n}+m)$. Let $\mathbf{y},\mathbf{z}\in B(\mathbf{0},\sqrt{n}+m)$ such that $\mathbf{y}\neq\mathbf{z}$. Then

$$\begin{array}{rl} & \frac{|f_{m-1}(\mathbf{y})-f_{m-1}(\mathbf{z})|}{\|\mathbf{y}-\mathbf{z}\|_2} \\ \leq & \frac{|f(\mathbf{y})-f(\mathbf{z})|+|\bar{\varphi}_m(\mathbf{y})^2 g_m(\mathbf{y})-\bar{\varphi}_m(\mathbf{z})^2 g_m(\mathbf{z})|}{\|\mathbf{y}-\mathbf{z}\|_2} \\ \leq & L_f + \frac{|\bar{\varphi}_m(\mathbf{y})^2 g_m(\mathbf{y})-\bar{\varphi}_m(\mathbf{z})^2 g_m(\mathbf{y})|}{\|\mathbf{y}-\mathbf{z}\|_2} \\ & + \frac{|\bar{\varphi}_m(\mathbf{z})^2 g_m(\mathbf{y})-\bar{\varphi}_m(\mathbf{z})^2 g_m(\mathbf{z})|}{\|\mathbf{y}-\mathbf{z}\|_2} \\ = & L_f + \frac{|g_m(\mathbf{y})||\bar{\varphi}_m(\mathbf{y})+\bar{\varphi}_m(\mathbf{z})||\bar{\varphi}_m(\mathbf{y})-\bar{\varphi}_m(\mathbf{z})|+\bar{\varphi}_m(\mathbf{z})^2|g_m(\mathbf{y})-g_m(\mathbf{z})|}{\|\mathbf{y}-\mathbf{z}\|_2} \\ \leq & L_f + \frac{2C_{g_m}C_{\bar{\varphi}_m}L_{\bar{\varphi}_m}\|\mathbf{y}-\mathbf{z}\|_2 + C_{\bar{\varphi}_m}^2 L_{g_m}\|\mathbf{y}-\mathbf{z}\|_2}{\|\mathbf{y}-\mathbf{z}\|_2} \\ = & L_f + 2C_{g_m}L_{\bar{\varphi}_m}C_{\bar{\varphi}_m} + L_{g_m}C_{\bar{\varphi}_m}^2 =: L_{f_{m-1}}\,. \end{array} \tag{7.1.69}$$

Then, $L_{f_{m-1}}$ is a Lipschitz constant of f_{m-1} on $B(\mathbf{0},\sqrt{n}+m)$.

Notice that $C_{\bar{\varphi}_m}, L_{\bar{\varphi}_m}, C_{f_{m-1}}, L_{f_{m-1}}$ are obtained by composing finitely many times the following operators: "$+$", "$-$", "$\times$", "$\div$", "$|\cdot|$","$(x_1,x_2)\mapsto\max\{x_1,x_2\}$", "$(x_1,x_2)\mapsto\min\{x_1,x_2\}$", "$(\cdot)^{\alpha_m}$" and "$\sqrt{\cdot}$", where all arguments possibly depend on ε. Without loss of generality we can assume $C_{\bar{\varphi}_m} = \bar{r}_m\varepsilon^{-r_m}$, $L_{\bar{\varphi}_m} = \bar{t}_m\varepsilon^{-t_m}$, $C_{f_{m-1}} = \bar{c}_{m-1}\varepsilon^{-c_{m-1}}$, $L_{f_{m-1}} = \bar{l}_{m-1}\varepsilon^{-l_{m-1}}$ for some $\bar{r}_m$, r_m, $\bar{t}_m$, t_m, $\bar{c}_{m-1}$, c_{m-1}, $\bar{l}_{m-1}$, l_{m-1} large enough and independent of ε.

Backward induction. Repeating the above process (after replacing f_j by f_{j-1}) several times, we obtain functions $\bar{\varphi}_j:\mathbb{R}^n\to\mathbb{R}$, $j=m,m-1,\ldots,1$, such that,

1. $\bar{\varphi}_j$ is positive, even and bounded from above by $C_{\bar{\varphi}_j} = \bar{r}_j\varepsilon^{-r_j}$ on $B(\mathbf{0},\sqrt{n}+j)$ for some positive constants $\bar{r}_j$ and r_j independent of ε.
2. $\bar{\varphi}_j$ is Lipschitz with Lipschitz constant $L_{\bar{\varphi}_j} = \bar{t}_j\varepsilon^{-t_j}$ for some positive constants $\bar{t}_j$ and t_j independent of ε.
3. $f_{j-1} := f_j + \frac{\varepsilon}{2^{m-j+1}} - \bar{\varphi}_j^2 g_j$ satisfies:
   (a) $f_{j-1}\geq 0$ on $S_{j-1}\cap B(\mathbf{0},\sqrt{n}+j-1)$;

(b) $f_{j-1} \le C_{f_{j-1}}$ on $B(\mathbf{0}, \sqrt{n}+j)$, where $C_{f_{j-1}} = \bar{c}_{j-1}\varepsilon^{-c_{j-1}}$ for some positive constants $\bar{c}_{j-1}$ and c_{j-1} independent of ε;

(c) f_{j-1} is Lipschitz on $B(\mathbf{0}, \sqrt{n}+j)$ with Lipschitz constant $L_{f_{j-1}} = \bar{l}_{j-1}\varepsilon^{-l_{j-1}}$ for some positive constants $\bar{l}_{j-1}$ and l_{j-1} independent of ε.

Then

$$\begin{aligned} f_0 &= f_1 + \tfrac{\varepsilon}{2^m} - \bar{\varphi}_1^2 g_1 \\ &= \big(f_2 + \tfrac{\varepsilon}{2^{m-1}} - \bar{\varphi}_2^2 g_2\big) + \tfrac{\varepsilon}{2^m} - \bar{\varphi}_1^2 g_1 \\ &= f_2 + \big(\tfrac{\varepsilon}{2^{m-1}} + \tfrac{\varepsilon}{2^m}\big) - \bar{\varphi}_2^2 g_2 - \bar{\varphi}_1^2 g_1 \\ &= \cdots = f_m + \varepsilon \textstyle\sum_{i=1}^m \frac{1}{2^i} - \sum_{i=1}^m \bar{\varphi}_i^2 g_i \\ &= f + \tfrac{\varepsilon}{2}\tfrac{1-\frac{1}{2^m}}{1-\frac{1}{2}} - \textstyle\sum_{i=1}^m \bar{\varphi}_i^2 g_i \\ &= f + \varepsilon(1 - \tfrac{1}{2^m}) - \textstyle\sum_{i=1}^m \bar{\varphi}_i^2 g_i \,. \end{aligned} \tag{7.1.70}$$

From this and since $f_0 \ge 0$ on $S_0 \cap B(\mathbf{0}, \sqrt{n}) = B(\mathbf{0}, \sqrt{n}) \supset [-1,1]^n$, we obtain

$$f + \varepsilon - \sum_{i=1}^m \bar{\varphi}_i^2 g_i \ge \frac{\varepsilon}{2^m} \text{ on } [-1,1]^n \,. \tag{7.1.71}$$

Polynomial approximations for the weight functions

Approximating with Bernstein polynomials. For each $i \in [m]$, we now approximate $\bar{\varphi}_i$ on $[-1,1]^n$ with the following Bernstein polynomials:

$$B_i^{(d)}(\mathbf{x}) = B_{\mathbf{y}\mapsto\bar{\varphi}_i(2\mathbf{y}-\mathbf{e}),d\mathbf{e}}\left(\frac{\mathbf{x}+\mathbf{e}}{2}\right), \quad d \in \mathbb{N}\,, \tag{7.1.72}$$

with $\mathbf{e} = (1,\ldots,1) \in \mathbb{R}^n$. By using Lemma 7.7, for all $\mathbf{x} \in [-1,1]^n$, for $i \in [m]$,

$$|B_i^{(d)}(\mathbf{x}) - \bar{\varphi}_i(\mathbf{x})| \le L_{\bar{\varphi}_i}\left(\frac{n}{d}\right)^{\frac{1}{2}}, \quad d \in \mathbb{N}\,, \tag{7.1.73}$$

and the following inequality holds for all $\mathbf{x} \in [-1,1]^n$, for $i \in [m]$:

$$|B_i^{(d)}(\mathbf{x})| \le \sup_{\mathbf{x}\in[-1,1]^n} |\bar{\varphi}_i(\mathbf{x})| \le C_{\bar{\varphi}_i}\,. \tag{7.1.74}$$

For $i \in [m]$, let

$$d_i := 2u_i \quad \text{with} \quad u_i = \left\lceil \frac{2C_{g_i}^2 C_{\bar{\varphi}_i}^2 n L_{\bar{\varphi}_i}^2 (m+1)^2 2^{2m}}{\varepsilon^2} \right\rceil, \tag{7.1.75}$$

where $C_{g_i} := \|g_i\|_{B(\mathbf{0},\sqrt{n}+i)}$, for $i \in [m]$. Then for all $\mathbf{x} \in [-1,1]^n$,

$$\begin{aligned} |B_i^{(d_i)}(\mathbf{x}) - \bar{\varphi}_i(\mathbf{x})| &\le L_{\bar{\varphi}_i}\left(\tfrac{n}{d_i}\right)^{\frac{1}{2}} \\ &\le L_{\bar{\varphi}_i}\left(\frac{n}{\frac{4C_{g_i}^2 C_{\bar{\varphi}_i}^2 n L_{\bar{\varphi}_i}^2 (m+1)^2 2^{2m}}{\varepsilon^2}}\right)^{\frac{1}{2}} \\ &= \tfrac{\varepsilon}{2C_{g_i} C_{\bar{\varphi}_i}(m+1)2^m}\,. \end{aligned} \tag{7.1.76}$$

Converting to homogeneous approximations. For $i \in [m]$, we write $B_i^{(d_i)} = \sum_{j=0}^{nd_i} h_i^{(j)}$ such that $h_i^{(j)}$ is a homogeneous polynomial with $\deg(h_i^{(j)}) = j$. Set $p_i := \frac{1}{2}[B_i^{(d_i)}(\mathbf{x}) + B_i^{(d_i)}(-\mathbf{x})]$, for $i \in [m]$. Then $p_i = \sum_{t=0}^{nu_i} h_i^{(2t)}$, for $i \in [m]$, since $h_i^{(j)}(\mathbf{x}) = h_i^{(j)}(-\mathbf{x})$ if j is even and $h_i^{(j)}(\mathbf{x}) = -h_i^{(j)}(-\mathbf{x})$ otherwise. Since $\bar{\varphi}_i$ is even, $\bar{\varphi}_i(\mathbf{x}) = \frac{1}{2}[\bar{\varphi}_i(\mathbf{x}) + \bar{\varphi}_i(-\mathbf{x})]$. It implies that for $\mathbf{x} \in [-1,1]^n$, for $i \in [m]$,

$$\begin{aligned} |p_i(\mathbf{x}) - \bar{\varphi}_i(\mathbf{x})| &= |\tfrac{1}{2}[B_i^{(d_i)}(\mathbf{x}) + B_i^{(d_i)}(-\mathbf{x})] - \tfrac{1}{2}[\bar{\varphi}_i(\mathbf{x}) + \bar{\varphi}_i(-\mathbf{x})]| \\ &\le \tfrac{1}{2}|B_i^{(d_i)}(\mathbf{x}) - \bar{\varphi}_i(\mathbf{x})| + \tfrac{1}{2}|B_i^{(d_i)}(-\mathbf{x}) - \bar{\varphi}_i(-\mathbf{x})| \\ &\le \tfrac{\varepsilon}{4C_{g_i}C_{\bar{\varphi}_i}(m+1)2^m} + \tfrac{\varepsilon}{4C_{g_i}C_{\bar{\varphi}_i}(m+1)2^m} = \tfrac{\varepsilon}{2C_{g_i}C_{\bar{\varphi}_i}(m+1)2^m}\,. \end{aligned} \tag{7.1.77}$$

and

$$|p_i(\mathbf{x})| \le \frac{1}{2}(|B_i^{(d_i)}(\mathbf{x})| + |B_i^{(d_i)}(\mathbf{x})|) \le \frac{1}{2}(C_{\bar{\varphi}_i} + C_{\bar{\varphi}_i}) = C_{\bar{\varphi}_i}\,. \tag{7.1.78}$$

Set $q_i := \sum_{t=0}^{nu_i} h_i^{(2t)} \|\mathbf{x}\|_2^{2(nu_i-t)}$. Then q_i is a homogeneous polynomial of degree $2nu_i$ and $q_i = p_i$ on $\mathbb{S}^{n-1}$, for $i \in [m]$. Thus for $i \in [m]$, $|q_i(\mathbf{x}) - \bar{\varphi}_i(\mathbf{x})| \le \frac{\varepsilon}{2C_{g_i}C_{\bar{\varphi}_i}(m+1)2^m}$ and $|q_i(\mathbf{x})| \le C_{\bar{\varphi}_i}$, for all $\mathbf{x} \in \mathbb{S}^{n-1}$. From these and (7.1.71), for all $\mathbf{x} \in \mathbb{S}^{n-1}$,

$$\begin{array}{rl} & f(\mathbf{x}) + \varepsilon - \sum_{i=1}^m q_i(\mathbf{x})^2 g_i(\mathbf{x}) \\ = & f(\mathbf{x}) + \varepsilon - \sum_{i=1}^m \bar{\varphi}_i(\mathbf{x})^2 g_i(\mathbf{x}) + \sum_{i=1}^m g_i(\mathbf{x})[\bar{\varphi}_i(\mathbf{x})^2 - q_i(\mathbf{x})^2] \\ \ge & \frac{\varepsilon}{2^m} - \sum_{i=1}^m |g_i(\mathbf{x})||\bar{\varphi}_i(\mathbf{x}) + q_i(\mathbf{x})||\bar{\varphi}_i(\mathbf{x}) - q_i(\mathbf{x})| \\ \ge & \frac{\varepsilon}{2^m} - \sum_{i=1}^m C_{g_i}(|\bar{\varphi}_i(\mathbf{x})| + |q_i(\mathbf{x})|)\frac{\varepsilon}{2C_{g_i}C_{\bar{\varphi}_i}(m+1)2^m} \\ \ge & \frac{\varepsilon}{2^m} - \sum_{i=1}^m 2C_{g_i}C_{\bar{\varphi}_i}\frac{\varepsilon}{2C_{g_i}C_{\bar{\varphi}_i}(m+1)2^m} \\ = & \frac{\varepsilon}{2^m} - \frac{m\varepsilon}{(m+1)2^m} = \frac{\varepsilon}{(m+1)2^m}\,. \end{array} \tag{7.1.79}$$

Moreover, for all $\mathbf{x} \in \mathbb{S}^{n-1}$,

$$f(\mathbf{x}) + \varepsilon - \sum_{i=1}^m q_i(\mathbf{x})^2 g_i(\mathbf{x}) \le C_f + \varepsilon + \sum_{i=1}^m C_{\bar{\varphi}_i}^2 C_{g_i} =: C_F\,. \tag{7.1.80}$$

Applying the global positivity certificate

Set $D := \max_{i\in[m]}\{2nu_i + d_{g_i}, d_f\}$ and

$$F = \|\mathbf{x}\|_2^{2(D-d_f)}(f + \varepsilon\|\mathbf{x}\|_2^{2d_f}) - \sum_{i=1}^m g_i q_i^2 \|\mathbf{x}\|_2^{2(D-2nu_i-d_{g_i})}\,. \tag{7.1.81}$$

Then F is a homogeneous polynomial of degree $2D$ and for all $\mathbf{x} \in \mathbb{S}^{n-1}$,

$$C_F \ge F(\mathbf{x}) = f(\mathbf{x}) + \varepsilon - \sum_{i=1}^m q_i(\mathbf{x})^2 g_i(\mathbf{x}) \ge \frac{\varepsilon}{(m+1)2^m}\,. \tag{7.1.82}$$

It implies that F is a positive definite form of degree $2D$ with $\inf_{\mathbf{x}\in\mathbb{S}^{n-1}} F(\mathbf{x}) \ge \frac{\varepsilon}{(m+1)2^m}$ and $\sup_{\mathbf{x}\in\mathbb{S}^{n-1}} F(\mathbf{x}) \le C_F$. There is no loss of generality in assuming $C_F = b\varepsilon^{-b}$ for some large enough $b > 0$ independent of ε. Similarly assume that $D \ge d\varepsilon^{-d}$ for some large enough $d > 0$ independent of ε. From this,

$$\delta(F) \le \frac{b\varepsilon^{-b}}{\frac{\varepsilon}{(m+1)2^m}} = b(m+1)2^m\varepsilon^{-b-1}\,. \tag{7.1.83}$$

Set

$$\bar{K} := \frac{2nd\varepsilon^{-d}(2d\varepsilon^{-d}-1)}{4\log 2} b(m+1)2^m\varepsilon^{-b-1}\,. \tag{7.1.84}$$

Then

$$\bar{K} \ge \frac{2nD(2D-1)}{4\log 2}\delta(F) - \frac{n+2D}{2}\,. \tag{7.1.85}$$

Clearly there exist positive constants $\bar{\mathfrak{c}}$ and $\mathfrak{c}$ independent of ε such that $\bar{\mathfrak{c}}\varepsilon^{-\mathfrak{c}} \ge \bar{K}$. Let $K \in \mathbb{N}$ and $K \ge \bar{\mathfrak{c}}\varepsilon^{-\mathfrak{c}} \ge \bar{K}$. According to Lemma 1.3.1, there exists a homogeneous SOS polynomial s_0 of degree $2(D+K)$ such that $\|\mathbf{x}\|_2^{2K}F = s_0$. It implies that

$$\begin{array}{rl} \|\mathbf{x}\|_2^{2(D-d_f+K)}(f + \varepsilon\|\mathbf{x}\|_2^{2d_f}) & = s_0 + \sum_{i=1}^m g_i q_i^2 \|\mathbf{x}\|_2^{2(D-2nu_i-d_{g_i}+K)} \\ & = s_0 + \sum_{i=1}^m g_i s_i\,, \end{array} \tag{7.1.86}$$

where $s_i := q_i^2\|\mathbf{x}\|_2^{2(D-2nu_i-d_{g_i}+K)}$ is a homogeneous SOS polynomial such that $\deg(g_is_i) = 2(K+D)$, for $i \in [m]$. Set $k = D - d_f + K$. Then $\|\mathbf{x}\|_2^{2k}(f + \varepsilon\|\mathbf{x}\|_2^{2d_f}) = s_0 + \sum_{i=1}^m g_i s_i$ with $\deg(s_0) = \deg(g_is_i) = 2(k+d_f)$, for $i \in [m]$.

The case of the ice cream constraint. Assume that $m = 1$ and $g_1 = x_n^2 - \|\mathbf{x}'\|_2^2$ with $\mathbf{x}' := (x_1, \ldots, x_{n-1})$. We shall show that $c = 65$. Using Lemma 7.3, we can take $\alpha_m = 2$ in (7.1.43). We then obtain the following asymptotic equivalences as $\varepsilon \to 0^+$:

$$\begin{array}{l} \delta_m \sim R_1\varepsilon^2 \Rightarrow C_{\psi_m} \sim R_2\varepsilon^{-2} \Rightarrow C_{\varphi_m} \sim R_3\varepsilon^{-1} \Rightarrow w_m \sim R_4\varepsilon^5 \Rightarrow L_{\sqrt{\xi_m}} \sim R_5\varepsilon^{-\frac{9}{2}} \\ \Rightarrow L_{\bar{\varphi}_m} \sim R_6\varepsilon^{-6} \Rightarrow C_{\bar{\varphi}_m} \sim R_7\varepsilon^{-6} \Rightarrow u_m \sim R_8\varepsilon^{-26} \Rightarrow d_m \sim R_9\varepsilon^{-26} \\ \Rightarrow C_F \sim R_{10}\varepsilon^{-12} \Rightarrow D \sim R_{11}\varepsilon^{-26} \Rightarrow b = 12 \Rightarrow d = 26 \Rightarrow \bar{K} \sim R_{12}\varepsilon^{-65} \\ \Rightarrow \mathfrak{c} = 65. \end{array} \tag{7.1.87}$$

for some $R_j > 0$ independent of ε, $j \in [12]$. This completes the proof of Theorem 7.1.

7.2 Polynomial optimization

This section is concerned with some applications to polynomial optimization.

Consider the following POP:

$$f^\star := \inf_{\mathbf{x} \in S(\mathfrak{g})} f(\mathbf{x}) \,, \tag{7.2.1}$$

where $f \in \mathbb{R}[\mathbf{x}]$ and

$$S(\mathfrak{g}) = \{\mathbf{x} \in \mathbb{R}^n \,:\, g_i(\mathbf{x}) \geq 0 \,,\, i \in [m]\} \,, \tag{7.2.2}$$

for some $\mathfrak{g} = \{g_i\}_{i \in [m]} \in \mathbb{R}[\mathbf{x}]$. Recall that $\theta = 1 + \|\mathbf{x}\|_2^2$.

7.2.1 General case

In this subsection, we improve the convergence rate of the Moment-SOS hierarchy described in [130, Theorem 4.3], based on Putinar–Vasilescu's Positivstellensatz [169].

Theorem 7.2. *Let $f, g_1, \ldots, g_m$ be polynomials such that $f^\star$ defined as in (7.2.1) and $S(\mathfrak{g})$ defined as in (7.2.2) satisfy that $S(\mathfrak{g})$ has nonempty interior and $f^\star > -\infty$. Let $\varepsilon > 0$ and denote $g_0 := 1$. Let $d := \lfloor \deg(f)/2 \rfloor + 1$. Consider the hierarchy of semidefinite programs indexed by $k \in \mathbb{N}$:*

$$\begin{array}{rll} \tau_k^{(\varepsilon)} := \inf & L_{\mathbf{y}}(\theta^k(f + \varepsilon\theta^d)) & \\ s.t. & \mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in \mathbb{N}^n_{2(d+k)}} \subset \mathbb{R} \,, & \\ & \mathbf{M}_{k+d}(\mathbf{y}) \succeq 0 \,, & \\ & \mathbf{M}_{k+d-\lceil g_i \rceil}(g_i\mathbf{y}) \succeq 0 \,,\; i \in [m] \,, & \\ & L_{\mathbf{y}}(\theta^k) = 1 \,. & \end{array} \tag{7.2.3}$$

For every $k \in \mathbb{N}$, the dual of (7.2.3) reads as:

$$\rho_k^{(\varepsilon)} := \sup_{\lambda \in \mathbb{R}} \{\lambda \,:\; \theta^k \,(f - \lambda + \varepsilon\,\theta^d) \in \mathcal{Q}_{k+d}(\mathfrak{g})\} \,, \tag{7.2.4}$$

where $\mathcal{Q}_r(\mathfrak{g})$ is defined as in (1.3.8). The following statements hold:

1. *For all $k \in \mathbb{N}$,*

$$\rho_k^{(\varepsilon)} \leq \rho_{k+1}^{(\varepsilon)} \leq f^\star \,. \tag{7.2.5}$$

2. *Assume that problem (7.2.1) has an optimal solution $\mathbf{x}^\star$. Then there exist positive constants $\bar{\mathfrak{c}}$ and $\mathfrak{c}$ depending on $f, g_1, \ldots, g_m$ such that for all $k \geq \bar{\mathfrak{c}}\varepsilon^{-\mathfrak{c}}$,*

$$0 \leq \rho_k^{(\varepsilon)} - f^\star \leq \varepsilon\theta(\mathbf{x}^\star)^d \,. \tag{7.2.6}$$

3. *Strong duality holds for all orders k of the primal-dual problems (7.2.3)-(7.2.4).*

The proof of Theorem 7.2 is exactly the same as the proof of [130, Theorem 7]. The second statement relies on Corollary 7.1. The third statement is due to the Slater condition [201, Theorem 3.1] since $S(\mathfrak{g})$ has nonempty interior (see in detail [130, Proposition 2 and Remark 3]).

7.2.2 Compact case

In this subsection, we consider the case when $S(\mathfrak{g})$ is compact by assuming that a ball constraint is present. We can then remove the perturbation term $\varepsilon\theta^d$ in the hierarchy based on Putinar–Vasilescu's Positivstellensatz, described in the previous subsection.

Theorem 7.3. *Let $f, g_1, \ldots, g_m$ be polynomials such that $f^\star$ defined as in* (7.2.1) *and $S(\mathfrak{g})$ defined as in* (7.2.2) *satisfy that $S(\mathfrak{g})$ has nonempty interior and $f^\star > -\infty$. Denote $g_0 := 1$. Let $d := \lfloor \deg(f)/2 \rfloor + 1$. Assume that $g_1 = R - \|\mathbf{x}\|_2^2$ for some $R > 0$. Consider the hierarchy of semidefinite programs indexed by $k \in \mathbb{N}$:*

$$\begin{aligned} \tau_k^{(0)} := \inf \quad & L_{\mathbf{y}}(\theta^k f) \\ s.t. \quad & \mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in \mathbb{N}^n_{2(d+k)}} \subset \mathbb{R}\,, \\ & \mathbf{M}_{k+d}(\mathbf{y}) \succeq 0\,, \\ & \mathbf{M}_{k+d-\lceil g_i \rceil}(g_i \mathbf{y}) \succeq 0,\ i \in [m]\,, \\ & L_{\mathbf{y}}(\theta^k) = 1\,. \end{aligned} \tag{7.2.7}$$

For every $k \in \mathbb{N}$, the dual of (7.2.7) *reads as:*

$$\rho_k^{(0)} := \sup_{\lambda \in \mathbb{R}} \{\lambda \,:\, \theta^k\,(f - \lambda) \in \mathcal{Q}_{k+d}(\mathfrak{g})\}\,, \tag{7.2.8}$$

where $\mathcal{Q}_r(\mathfrak{g})$ is defined as in (1.3.8)*. The following statements hold:*

1. *For all $k \in \mathbb{N}$,*
$$\rho_k^{(0)} \leq \rho_{k+1}^{(0)} \leq f^\star\,. \tag{7.2.9}$$

2. *There exist positive constants $\bar{\mathfrak{c}}$ and $\mathfrak{c}$ depending on $f, g_1, \ldots, g_m$ such that*
$$0 \leq f^\star - \rho_k^{(0)} \leq \left(\frac{\bar{\mathfrak{c}}}{k}\right)^{\frac{1}{\mathfrak{c}}} \tag{7.2.10}$$

3. *Strong duality holds for all orders k of the primal-dual problems* (7.2.7)-(7.2.8)*.*

Proof. The first and third statements of Theorem 7.3 can be proved similarly to the ones of Theorem 7.2. Let us prove the second statement. By using Corollary 7.2, there exist positive constants $\bar{\mathfrak{c}}$ and $\mathfrak{c}$ depending on $f, g_1, \ldots, g_m$ such that for any $\varepsilon > 0$, for all $k \geq \bar{\mathfrak{c}}\varepsilon^{-\mathfrak{c}}$,

$$\theta^k(f - f^\star + \varepsilon) \in \mathcal{Q}_{k+d}(\mathfrak{g})\,. \tag{7.2.11}$$

Let $K \in \mathbb{N}$. Set $\epsilon = (\frac{\bar{\mathfrak{c}}}{K})^{\frac{1}{\mathfrak{c}}}$. Then $\epsilon > 0$ and $K = \bar{\mathfrak{c}}\epsilon^{-\mathfrak{c}}$, so that

$$\theta^K(f - f^\star + \epsilon) \in \mathcal{Q}_{K+d}(\mathfrak{g})\,. \tag{7.2.12}$$

It implies that $f^\star - \epsilon$ is a feasible solution of (7.2.8) with relaxation order K, yielding

$$0 \leq f^\star - \rho_K^{(0)} \leq f^\star - (f^\star - \epsilon) = \epsilon = \left(\frac{\bar{\mathfrak{c}}}{K}\right)^{\frac{1}{\mathfrak{c}}}\,. \tag{7.2.13}$$

Hence the desired result follows. □

Remark 7.5. *The authors' bounds in Theorems 7.2 and 7.3 are only for worst cases. In fact, for generic cases of polynomials, the Moment-SOS hierarchy based on Putinar's Positivstellensatz (in [102]) has finite convergence [147].*

Chapter 8

A sparse version of Reznick's Positivstellensatz

Most of the content of this chapter is from [133].

Inspired by correlative sparsity due to Waki et al. [203, 103] and Putinar–Vasilescu's Positivstellensatz that is applied for polynomial optimization in the previous two chapters, we prove a sparse version of Reznick's Positivstellensatz. Accordingly we obtain some representations that involve uniform denominators and quadratic modules in the sparse setting.

Exploiting sparsity pattern. Let $n, m \in \mathbb{N}^{>0}$. For $T \subset [n]$, denote by $\mathbb{R}[\mathbf{x}(T)]$ (resp. $\Sigma[\mathbf{x}(T)]$) the ring of polynomials (resp. the subset of SOS polynomials) in the variables $\mathbf{x}(T) := \{x_j : j \in T\}$. Also denote by $\mathbb{R}[\mathbf{x}(T)]_t$ (resp. $\Sigma[\mathbf{x}(T)]_t$) the restriction of $\mathbb{R}[\mathbf{x}(T)]$ (resp. $\Sigma[\mathbf{x}(T)]$) to polynomials of degree at most t (resp. $2t$). For $W \subset [m]$, we note $\mathfrak{g}_W := \{g_i : i \in W\}$.

Designing alternative hierarchies for solving $f^\star := \inf\{f(\mathbf{x}) : x \in S(\mathfrak{g})\}$, significantly (computationally) cheaper than their dense version (1.2.4)*, while maintaining convergence to the optimal value $f^\star$ is a real challenge with important implications.*

One first such successful contribution is due to Waki et al. [203] when the input polynomial data f, g_i are sparse, where by sparse we mean the following:

Assumption 8.1. *The following conditions hold:*

(i) Running intersection property (RIP): $[n] = \bigcup_{c=1}^{p} I_c$ *with* $p \in \mathbb{N}^{\geq 2}$, $I_c \neq \emptyset$, $c \in [p]$, *and for every* $c \in \{2, \ldots, p\}$, *there exists* $s_c \in [c-1]$, *such that* $\hat{I}_c \subset I_{s_c}$, *where* $\hat{I}_c := I_c \cap \left(\bigcup_{j=1}^{c-1} I_j\right)$. *Note that* $s_2 = 1$ *and w.l.o.g, set* $\hat{I}_1 := \emptyset$. *Denote* $n_c := |I_c|$ *and* $\hat{n}_c := |\hat{I}_c|$, $c \in [p]$.

(ii) Structured sparsity pattern for the objective function[1]*:* $f = \sum_{c=1}^{p} f_c$ *where* $f_c \in \mathbb{R}[\mathbf{x}(I_c)]_{\deg(f)}$, $c \in [p]$.

(iii) Structured sparsity pattern for the constraints: $[m] = \bigcup_{c=1}^{p} J_c$ *and for every* $i \in J_c$, $g_i \in \mathbb{R}[\mathbf{x}(I_c)]$, $c \in [p]$.

(iv) Additional redundant quadratic constraints: There exists $R > 0$ *such that* $\|\mathbf{x}\|_2^2 \leq R$ *for all* $\mathbf{x} \in S(\mathfrak{g})$ *and* $R - \|\mathbf{x}(I_c)\|_2^2 \in \mathfrak{g}_{J_c}$, $c \in [p]$.

With τ ($\leq n$) being the maximum number of variables appearing in each index subset I_c, i.e., $\tau := \max\{n_c : c \in [p]\}$, Table 8.1 displays the respective computational complexity of the sparse hierarchy of Waki et al. [203] and the dense hierarchy of Lasserre [102] for SDPs with same relaxation order $k \in \mathbb{N}$. Obviously the sparse hierarchy provides a potentially high computational saving when compared to the dense one. In addition, convergence of the hierarchy of Waki et al. to the optimal value of the original POP was proved in [103], resulting in the following sparse version of Putinar's Positivstellensatz:

[1] If there are f_c in the sum f such that $\deg(f_c) > \deg(f)$, we can always remove the high degree redundant term in f_c which cancel with each other to make degree of f_c at most $\deg(f)$.

Table 8.1: Comparing the computational complexity of the sparse and dense hierarchies.

SDP of order k	sparse hierarchy	dense hierarchy
number of variables	$O(\tau^{2k})$	$O(n^{2k})$
largest size of SDP matrix	$O(\tau^{k})$	$O(n^{k})$

Theorem 8.1. *(Lasserre, Waki et al.) Let Assumption 8.1 hold. If a polynomial f is positive on $S(\mathfrak{g})$, then there exist $\sigma_{0,c} \in \Sigma[\mathbf{x}(I_c)]_k$, $\sigma_{i,c} \in \Sigma[\mathbf{x}(I_c)]_{k-\lceil g_i \rceil}$, $i \in J_c$, $c \in [p]$ such that*

$$f = \sum_{c=1}^{p} \left(\sigma_{0,c} + \sum_{i \in J_c} \sigma_{i,c} g_i \right). \tag{8.0.1}$$

Compactness of the feasible set $S(\mathfrak{g})$ is a crucial ingredient of the proof in [103]; shortly after, Grimm et al. [68] provided another (simpler) proof where $\mathrm{int}(S(\mathfrak{g})) \neq \emptyset$ is not needed, but where compactness of $S(\mathfrak{g})$ is still a crucial assumption.

Motivation for sparse representations on noncompact sets. We remark that Theorem 8.1 requires the additional redundant quadratic constraints (Assumption 8.1 (iv)), which is slightly stronger than just assuming the compactness of $S(\mathfrak{g})$. When $S(\mathfrak{g})$ is compact, we can always add these constraints but we need to know the radius $R > 0$ of a ball centered at the origin and containing $S(\mathfrak{g})$. In this case, adding such constraints increases the number of positive semidefinite matrices from m to $m + p$ in each SDP. In addition, it may be hard to verify compactness of $S(\mathfrak{g})$ and obtain such a radius R.

To the best of our knowledge, in the noncompact case there is still no Positivstellensatz allowing one to build hierarchies for POPs satisfying:

- the RIP and the structured sparsity pattern from Assumption 8.1 (i)-(iii),
- and a guarantee of convergence to the global optimum.

In fact we provide Examples 6.1, 8.3, and 8.4, which show that in both unconstrained and constrained cases, there exist sparse nonnegative polynomials which do *not* have a sparse SOS-based decomposition (8.0.1) à la Putinar. Such examples have been our motivation to investigate existence of sparse representations in the noncompact case, as well as to construct converging SDP-hierarchies for sparse polynomial optimization in general.

Contribution. Our contribution is threefold:

I. We first provide a rational SOS representation for a positive definite rational form which is a sum of sparse rational functions with uniform denominators, satisfying the structured sparsity pattern and the RIP stated in Assumption 8.1 (i). This representation is provided in Theorem 8.2. As a direct consequence, we obtain a sparse version of Reznick's Positivstellensatz in Corollary 8.1.

II. Then, we provide two positivity certificates for arbitrary small perturbations of – globally nonnegative polynomials in Corollary 8.2 – and polynomials nonnegative on a (possibly noncompact) basic semialgebraic set in Corollary 8.3, when the input data satisfy a similar sparsity pattern. These two certificates are obtained via a sparse version of Putinar–Vasilescu's Positivstellensatz and do not require the additional constraints from Assumption 8.1 (iv).

III. In Section 8.2, we build up a hierarchy of semidefinite relaxations for polynomial optimization based on the sparse version of Putinar–Vasilescu's Positivstellensatz. Convergence of the hierarchy is guaranteed and illustrated on minimization of random quadratic forms on the nonnegative orthant. However a naive implementation of this hierarchy leads to a heavy computational burden when the number of variables is larger than 10.

Illustrations of such positivity certificates for polynomials nonnegative on noncompact basic semialgebraic sets are provided in Example 8.1, 6.1, 8.3 and 8.4, for which positivity certificates (1.2.4) do not exist.

The existence of such sparse SOS-representations is proved by combining different tools:

(a) First, we use an idea similar to that developed in Grimm et al. [68] (in the compact case) to prove that a sparse positive definite form can be decomposed as SOS of sparse positive definite rational forms; as expected the noncompact case is technically more involved. This yields a *sparse version* of Hilbert-Artin's representation theorem in the case of positive definite forms.

(b) Next, we use generalizations of Schmüdgen's Positivstellensatz presented by Schweighofer [187], Berr-Wörmann [17], Jacobi [85], and Marshall [136, 137], for a finitely generated $\mathbb{R}$-algebra in each term of the sum, to obtain again a *sparse version*, this time of Reznick's Positivstellensatz for positive definite forms.

(c) Finally we combine the homogenization/dehomogenization method that we already used in [130] together with limit tools, to provide the two sparse versions of Putinar–Vasilescu's Positivstellensatz.

We acknowledge that the computational benefits are so far rather limited and that the present contribution is essentially theoretical, as it provides a sparse analogue of Reznick's Positivstellenstaz. In our opinion, the sparse analogue of Putinar's Positivstellenstaz is still the "champion" algorithm to beat. Indeed the versatility of its power, which applies to both dense (but of modest size) problems and large size sparse problems, is yet to be surpassed. So the practical benefits of sparse Reznick Positivstellensatz are not immediately available with its obvious (but naive) implementation. In order to address this computational issue, we propose to use a sampling technique in the spirit of that advocated by Parrilo and Löfberg [118] for polynomial optimization, and briefly described in Section 8.2.3. In our context it allows us to avoid clearing denominators and its efficiency is illustrated on some numerical experiments presented in Section 8.2.4. Its complete validation for rational functions is beyond the scope of the present chapter and we believe that in view of the appealing form of our sparse version of Reznick's Positivstellensatz, additional investigation of powerful algorithmic implementations are worth pursuing.

8.1 Representation theorems

8.1.1 Notation and definitions

A function h is *homogeneous* of degree t if $h(\lambda\mathbf{x}) = \lambda^t h(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$ and each $\lambda \in \mathbb{R}$. Therefore a homogeneous polynomial can be written as $h = \sum_{|\alpha|=t} h_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}}$. A function $f : \mathbb{R}^n \to \mathbb{R}$ is even if $f(\mathbf{x}) = f(-\mathbf{x})$ for all $\mathbf{x}$. A rational function h is the ratio of two polynomials and denote by $\mathbb{R}(\mathbf{x})$ the space of all rational functions. A homogeneous rational function (also called be a *rational form*, or form in short) can be written as the ratio of two homogeneous polynomials.

The degree-d *homogenization* $\tilde{h}$ of $h \in \mathbb{R}(x_1, \ldots, x_n)$ is a homogeneous rational function in $\mathbb{R}(x_1, \ldots, x_{n+1})$ of degree d defined by $\tilde{h}(\mathbf{x}, x_{n+1}) = x_{n+1}^d h(\mathbf{x}/x_{n+1})$. A *rational positive definite form* of degree t is a homogeneous rational function of degree t which is positive everywhere except at the origin. Equivalently, a homogeneous rational function h of degree t is a rational positive definite form of degree t if and only if there exists $\varepsilon > 0$ such that $h \geq \varepsilon \|\mathbf{x}\|_2^{2t}$.

For $(i, j) \in \mathbb{N}^2$, we denote the Kronecker delta function by

$$\delta_{i,j} := \begin{cases} 1 & \text{if } i = j\,, \\ 0 & \text{if } i \neq j\,. \end{cases}$$

When Assumption 8.1 (i) holds, define

$$\Phi_c := \begin{cases} \|\mathbf{x}(\hat{I}_c)\|_2^{2(1-\delta_{c,1})} \prod_{j=c+1}^{p} \|\mathbf{x}(\hat{I}_j)\|_2^{2\delta_{c,s_j}} & \text{if } c \in [p-1]\,, \\ \|\mathbf{x}(\hat{I}_c)\|_2^{2(1-\delta_{c,1})} & \text{if } c = p\,. \end{cases}$$

Obviously, one has $\Phi_c \in \mathbb{R}[\mathbf{x}(I_c)]$, for each $c \in [p]$.

8.1.2 A key result

Let us state the first main result of this chapter which is of independent interest and very useful in proving our representation results. In particular it yields a sparse version of Reznick's Positivstellensatz as a particular case.

Theorem 8.2. *Let Assumption 8.1 (i) hold. Let $f \in \mathbb{R}(\mathbf{x})$ be a positive definite rational form of degree $2d$ with $d \in \mathbb{N}^{>0}$ such that*

$$f = \sum_{c=1}^{p} \frac{q_c}{\|\mathbf{x}(I_c)\|_2^{2k_c}},$$

where $q_c \in \mathbb{R}[\mathbf{x}(I_c)]$ is homogeneous of degree $2(d+k_c)$ for some $k_c \in \mathbb{N}$, $c \in [p]$. Then there exist $k \in \mathbb{N}$ and $\sigma_c \in \Sigma[\mathbf{x}(I_c)]_{d+k(1+\deg(\Phi_c)/2)}$, $c \in [p]$, such that

$$f = \sum_{c=1}^{p} \frac{\sigma_c}{\|\mathbf{x}(I_c)\|_2^{2k}\Phi_c^k}. \tag{8.1.1}$$

The proof of Theorem 8.2 can be found in [133, Section 4].

As a consequence, we obtain the following sparse version of Reznick's Positivstellensatz.

Corollary 8.1. *Let Assumption 8.1 (i) hold. Assume that f is a positive definite form of degree $2d$ with $d \in \mathbb{N}^{>0}$ and $f = \sum_{c=1}^{p} f_c$, where $f_c \in \mathbb{R}[\mathbf{x}(I_c)]$ is homogeneous of degree $2d$, $c \in [p]$. Then there exist $k \in \mathbb{N}$ and $\sigma_c \in \Sigma[\mathbf{x}(I_c)]_{d+k(1+\deg(\Phi_c)/2)}$, $c \in [p]$, such that*

$$f = \sum_{c=1}^{p} \frac{\sigma_l}{H_c^k}, \tag{8.1.2}$$

where $H_c := \|\mathbf{x}(I_c)\|_2^2\,\Phi_c$, $c \in [p]$.

To prove Corollary 8.1, we apply Theorem 8.2 with $k_c = 0$, $c \in [p]$. The representation (8.1.2) can still hold even when f is not a positive definite form, as illustrated in the following example.

Example 8.1. *Let $f = f_1 + f_2$, where*

$$f_1 := x_4^2(x_1^4x_2^2 + x_2^4x_3^2 + x_1^2x_3^4 - 3x_1^2x_2^2x_3^2) + x_3^8$$

is the so-called Delzell's polynomial and $f_2 := x_1^2x_2^2x_3^2x_5^2$. The polynomial f_1 is nonnegative, but not SOS as shown in [176, Section 6]. Let $I_1 := \{1,2,3,4\}$ and $I_2 := \{1,2,3,5\}$. Then $f_1 \in \mathbb{R}[\mathbf{x}(I_1)]$ and $f_2 \in \mathbb{R}[\mathbf{x}(I_2)]$ are nonnegative and homogeneous of degree 8. Since f_1 is nonnegative then f is nonnegative. The following statements hold:

1. *f is a nonnegative form, but is not positive definite;*
2. *$f \notin \Sigma[\mathbf{x}(I_1)] + \Sigma[\mathbf{x}(I_2)]$, but $f \in \frac{\Sigma[\mathbf{x}(I_1)]_6}{\|\mathbf{x}(I_1)\|_2^2\Phi_1} + \frac{\Sigma[\mathbf{x}(I_2)]_6}{\|\mathbf{x}(I_2)\|_2^2\Phi_2}$.*

The first statement follows from the fact that $f(0,0,0,1,1) = 0$, ensuring that f is not a positive definite form.
Proof of the second statement: Assume by contradiction that $f = \sigma_1 + \sigma_2$ for some $\sigma_c \in \Sigma[\mathbf{x}(I_c)]$, $c = 1,2$. Evaluation at $x_5 = 0$ yields $f_1 = \sigma_1 + \sigma_2(x_1,x_2,x_3,0)$, so that f_1 is an SOS, which is impossible. Thus, $f \notin \Sigma[\mathbf{x}(I_1)] + \Sigma[\mathbf{x}(I_2)]$. However, $(x_1^2+x_2^2+x_3^2)f_1$ is SOS according to [182, Example 4.4], so $(x_1^2+x_2^2+x_3^2)f \in \Sigma[\mathbf{x}(I_1)]_5 + \Sigma[\mathbf{x}(I_2)]_5$. Note that $\Phi_1 = \Phi_2 = x_1^2+x_2^2+x_3^2$. Therefore

$$f \in \frac{\Sigma[\mathbf{x}(I_1)]_5}{\Phi_1} + \frac{\Sigma[\mathbf{x}(I_2)]_5}{\Phi_2} \subset \frac{\Sigma[\mathbf{x}(I_1)]_6}{H_1} + \frac{\Sigma[\mathbf{x}(I_2)]_6}{H_2}.$$

Remark 8.1. ***A possibly simpler proof of Corollary 8.1 based on Carathéodory's theorem.***
In [14, Chapter I, Section 3], Barvinok provides a simple proof of Reznick's theorem, based on Carathéodory's theorem. The basic idea is that a large enough power of the linear form $\langle \mathbf{v}, \mathbf{x} \rangle^{2k}$ is a Dirac function on the unit sphere, centered at $\mathbf{v}$. Therefore $\|\mathbf{x}\|_2^{2k} f$ (which is the same as f on the unit sphere, and strictly positive) can naturally be written as a positive combination of Dirac functions, which can be in turn decomposed as a positive weighted sum of powers of linear forms. Compared to Reznick's proof, this simple proof has a less constructive flavor. It is due to the fact that we need to know the power of the linear form $\langle \mathbf{v}, \mathbf{x} \rangle^{2k}$ to construct the convex hull of such

Dirac functions in order to apply Carathéodory's theorem. If the shape of the uniform denominators involved in Reznick's result is not available, finding such convex hull is hard and unnatural. To prove our sparse version of Reznick's result via the application of Carathéodory's theorem, we would need to construct the Dirac functions of the form σ_c/H_c^k. However these Dirac unctions are complicated because each denominator H_c^k depends on the RIP. In our proof, we obtain the general form of H_c via inductions. In case that such denominators H_c^k are known in advance, it might be possible to prove our result in a simpler way.

8.1.3 Global nonnegativity

When Assumption 8.1 (i) holds, define the following polynomials, for each $c \in [p]$:

- $\theta_c := \|\mathbf{x}(I_c)\|_2^2 + 1$ and $\hat{\theta}_c := \|\mathbf{x}(\hat{I}_c)\|_2^2 + 1$;
- $D_c := \begin{cases} \hat{\theta}_c^{1-\delta_{c,1}} \prod_{j=c+1}^{p} \hat{\theta}_j^{\delta_{c,s_j}} & \text{if } c < p\,, \\ \hat{\theta}_c^{1-\delta_{c,1}} & \text{if } c = p\,; \end{cases}$
- $\Theta_c := \theta_c D_c$ and $\omega_c := \deg(\Theta_c)/2$.

Note that $\Theta_c \in \Sigma[\mathbf{x}(I_c)]_{\omega_c}$, for each $c \in [p]$. We next state the following *sparse version* of Putinar–Vasilescu's Positivstellensatz for polynomials nonnegative on $\mathbb{R}^n$.

Corollary 8.2. *Let f be a nonnegative polynomial such that the conditions (i) and (ii) of Assumption 8.1 hold. Let $\varepsilon > 0$ and $d \geq \deg(f)/2$. Then there exist $k \in \mathbb{N}$ and $\sigma_c \in \Sigma[\mathbf{x}(I_c)]_{d+k\omega_c}$, $c \in [p]$, such that*

$$f + \varepsilon \sum_{c=1}^{p} \theta_c^d = \sum_{c=1}^{p} \frac{\sigma_c}{\Theta_c^k}\,. \tag{8.1.3}$$

The proof of Corollary 8.2 can be found in [133, Section 4].

The representation (8.1.3) can still hold even if $\varepsilon = 0$, as illustrated in the following examples.

Example 8.2. *Let $f = f_1 + f_2$, where*

$$f_1 := 8 + \frac{1}{2}x_1^2x_2^4 + (x_1^2 - 2x_1^3)x_2^3 + (2x_1 + 10x_1^2 + 4x_1^3 + 3x_1^4)x_2^2 + 4(x_1 - 2x_1^2)x_2$$

is the so-called Leep-Starr's polynomial and $f_2 := x_1^2x_3^2$. Let $I_1 := \{1,2\}$ and $I_2 := \{1,3\}$, so that $f_1 \in \mathbb{R}[\mathbf{x}(I_1)]$ and $f_2 \in \mathbb{R}[\mathbf{x}(I_2)]$. As shown in [114, Example 2], f_1 is nonnegative but not an SOS. In addition, f_2 is an SOS, so that f is nonnegative.

We claim that $f \notin \Sigma[\mathbf{x}(I_1)] + \Sigma[\mathbf{x}(I_2)]$. Indeed, assume by contradiction that $f = \sigma_1 + \sigma_2$ for some $\sigma_c \in \Sigma[\mathbf{x}(I_c)]$, $c = 1, 2$. Evaluation at $x_3 = 0$, yields $f_1 = \sigma_1 + \sigma_2(x_2, 0)$, so that f_1 is an SOS, which is impossible.

However, $(x_1^2+1)^2 f_1$ is a sum of three squares of polynomials according to [114, Example 2], so $(x_1^2+1)^2 f \in \Sigma[\mathbf{x}(I_1)]_5 + \Sigma[\mathbf{x}(I_2)]_5$. Note that $D_1 = D_2 = x_1^2 + 1$. Thus,

$$f \in \frac{\Sigma[\mathbf{x}(I_1)]_5}{D_1^2} + \frac{\Sigma[\mathbf{x}(I_2)]_5}{D_2^2} \subset \frac{\Sigma[\mathbf{x}(I_1)]_7}{\Theta_1^2} + \frac{\Sigma[\mathbf{x}(I_2)]_7}{\Theta_2^2}\,.$$

Example 8.3. *As shown in [100, Example 5.2], the nonnegative polynomial*

$$f = x_1^2 - 2x_1x_2 + 3x_2^2 - 2x_1^2x_2 + 2x_1^2x_2^2 - 2x_2x_3 + 6x_3^2 + 18x_2^2x_3 - 54x_2x_3^2 + 142x_2^2x_3^2$$

satisfies $f \in \mathbb{R}[\mathbf{x}(I_1)] + \mathbb{R}[\mathbf{x}(I_2)]$ and $f \notin \Sigma[\mathbf{x}(I_1)] + \Sigma[\mathbf{x}(I_2)]$, with $I_1 = \{1,2\}$ and $I_2 = \{2,3\}$. However, $f \in \frac{\Sigma[\mathbf{x}(I_1)]_4}{\Theta_1} + \frac{\Sigma[\mathbf{x}(I_2)]_4}{\Theta_2}$, where $\Theta_1 = (x_2^2+1)(x_1^2+x_2^2+1)$ and $\Theta_2 = (x_2^2+1)(x_2^2+x_3^2+1)$. It is due to the fact that $f = \frac{\sigma_1}{D_1} + \frac{\sigma_2}{D_2}$, where $D_1 = D_2 = x_2^2 + 1$ and σ_1, σ_2 are SOS polynomials given in [133, Appendix].

8.1.4 Positivity on a semialgebraic set

We next state our second main result, namely a *sparse version* of Putinar–Vasilescu's Positivstellensatz for polynomials nonnegative on (possibly noncompact) basic semialgebraic sets.

Corollary 8.3. *Let $f \in \mathbb{R}[\mathbf{x}]$ be nonnegative on $S(\mathfrak{g})$ such that the conditions (i), (ii) and (iii) of Assumption 8.1 hold. Let $\varepsilon > 0$ and $d \geq 1 + \lfloor \deg(f)/2 \rfloor$. Then there exist $k \in \mathbb{N}$, $\sigma_{0,c} \in \Sigma[\mathbf{x}(I_c)]_{d+k\omega_c}$ and $\sigma_{i,c} \in \Sigma[\mathbf{x}(I_c)]_{d+k\omega_c - \lceil g_i \rceil}$, $i \in J_c$, $c \in [p]$, such that*

$$f + \varepsilon \sum_{c=1}^{p} \theta_c^d = \sum_{c=1}^{p} \frac{\sigma_{0,c} + \sum_{i \in J_c} \sigma_{i,c} g_i}{\Theta_c^k} \,. \tag{8.1.4}$$

The proof of Corollary 8.3 can be found in [133, Section 4].

Example 8.4. *Let $f = f_1 + f_2$, where $f_1 = x_1 x_2$ and $f_2 = x_2^2 x_3$. Let $\mathfrak{g} = \{g_1, g_2, g_3\}$, where $g_1 = x_2^3$, $g_2 = -g_1$ and $g_3 = x_3$. It is not hard to show that $f = 0$ on $S(\mathfrak{g})$, so that $f \geq 0$ on $S(\mathfrak{g})$. By noting $I_1 := \{1, 2\}$ and $I_2 := \{2, 3\}$, one has $\{f_1, g_1, g_2\} \subset \mathbb{R}[\mathbf{x}(I_1)]$ and $\{f_2, g_3\} \subset \mathbb{R}[\mathbf{x}(I_2)]$. We claim the following statements:*

1. *$f \notin \Sigma[\mathbf{x}(I_1)] + g_1 \mathbb{R}[\mathbf{x}(I_1)] + \Sigma[\mathbf{x}(I_2)] + g_3 \Sigma[\mathbf{x}(I_2)]$;*

2. *for every $\varepsilon > 0$,*

$$f + \varepsilon(\theta_1^2 + \theta_2^2) \in \frac{\Sigma[\mathbf{x}(I_1)]_{2k+2} + g_1 \mathbb{R}[\mathbf{x}(I_1)]_{4k+1}}{\Theta_1^k} + \frac{\Sigma[\mathbf{x}(I_2)]_{2k+2} + g_3 \Sigma[\mathbf{x}(I_2)]_{4k+3}}{\Theta_2^k} \,,$$

for some $k \in \mathbb{N}$ depending on ε.

Proof of the first statement: Assume by contradiction that there exist $\sigma_1 \in \Sigma[\mathbf{x}(I_1)]$, $\psi_1 \in \mathbb{R}[\mathbf{x}(I_1)]$ and $\sigma_2, \sigma_3 \in \Sigma[\mathbf{x}(I_2)]$ such that $f = \sigma_1 + \psi_1 g_1 + \sigma_2 + \sigma_3 g_3$. Evaluation at $x_1 = 1$ and $x_3 = 0$ yields

$$x_2 = \sigma_1(1, x_2) + \psi_1(1, x_2) x_2^3 + \sigma_2(x_2, 0) \in \Sigma[x_2] + x_2^3 \mathbb{R}[x_2] \,,$$

which is impossible due to [130, Lemma 3.3 (i)].
Proof of the second statement: With $\varepsilon > 0$ fixed,

$$f_1 + \varepsilon \theta_1^2 = x_1 x_2 + \varepsilon(1 + x_1^2 + x_2^2)^2 = x_1 x_2 + \varepsilon + \varepsilon x_1^2 + \sigma_4 \,,$$

for some $\sigma_4 \in \Sigma[\mathbf{x}(I_1)]_2$. Let $k \in \mathbb{N}^{\geq 2}$ be fixed. Then $D_1^k = (1 + x_2^2)^k = 1 + k x_2^2 + x_2^4 \sigma_5$ for some $\sigma_5 \in \Sigma[x_2]_{k-2}$, which implies

$$D_1^k (f_1 + \varepsilon \theta_1^2) = x_1 x_2 + \varepsilon x_1^2 + \varepsilon k x_2^2 + \sigma_6 + \psi_2 x_2^3 \,,$$

for some $\sigma_6 \in \Sigma[\mathbf{x}(I_1)]_{k+2}$ and $\psi_2 \in \mathbb{R}[\mathbf{x}(I_1)]_{2k+1}$. Assume that $k \geq \varepsilon^{-2}/4$. Then

$$\begin{aligned} D_1^k (f_1 + \varepsilon \theta_1^2) \;&= x_1^2 \left(\varepsilon - \tfrac{1}{4\varepsilon k}\right) + \left(x_2 \sqrt{\varepsilon k} + \tfrac{x_1}{2\sqrt{\varepsilon k}}\right)^2 + \sigma_6 + \psi_2 x_2^3 \\ &\in \Sigma[\mathbf{x}(I_1)]_{k+2} + g_1 \mathbb{R}[\mathbf{x}(I_1)]_{2k+1} \,, \end{aligned}$$

which implies $f_1 + \varepsilon \theta_1^2 \in \frac{\Sigma[\mathbf{x}(I_1)]_{2k+2} + g_1 \mathbb{R}[\mathbf{x}(I_1)]_{4k+1}}{\Theta_1^k}$. We also have

$$f_2 + \varepsilon \theta_2^2 \in \frac{\Sigma[\mathbf{x}(I_2)]_{2k+2} + g_3 \Sigma[\mathbf{x}(I_2)]_{4k+3}}{\Theta_2^k}$$

since $f_2 \in g_3 \Sigma[\mathbf{x}(I_2)]_1$, proving the second statement.

8.1.5 General case

For $f \in \mathbb{R}[\mathbf{x}]$, define

$$d := \begin{cases} \lceil \deg(f)/2 \rceil & \text{if } S(\mathfrak{g}) = \mathbb{R}^n\,, \\ 1 + \lfloor \deg(f)/2 \rfloor & \text{if } S(\mathfrak{g}) \neq \mathbb{R}^n\,. \end{cases} \tag{8.1.5}$$

For $l \in [p]$, let us note

$$\mathcal{Q}(\mathfrak{g}_{J_c})[\mathbf{x}(I_c)]_k = \left\{ \sigma_{0,c} + \sum_{i \in J_c} \sigma_{i,c} g_i \;\middle|\; \begin{array}{l} \sigma_{0,c} \in \Sigma[\mathbf{x}(I_c)]_k, \\ \sigma_{i,c} \in \Sigma[\mathbf{x}(I_c)]_{k - \lceil g_i \rceil}\,,\ i \in J_c \end{array} \right\} \tag{8.1.6}$$

the truncated quadratic module of order k associated with the polynomials $\mathfrak{g}_{J_c} = \{g_i \,:\, i \in J_c\}$.

Gathering the two situations from Corollary 8.2 and Corollary 8.3, we obtain the following general statement.

Theorem 8.3. *Let $f \in \mathbb{R}[\mathbf{x}]$ be nonnegative on $S(\mathfrak{g})$. Let us suppose that the conditions (i) and (ii) of Assumption 8.1 hold if $S(\mathfrak{g}) = \mathbb{R}^n$, otherwise let us suppose that the conditions (i), (ii) and (iii) of Assumption 8.1 hold. Let us fix $\varepsilon > 0$ and let d be defined as in* (8.1.5)*. Then there exists $k \in \mathbb{N}$ such that*

$$f + \varepsilon \sum_{c=1}^{p} \theta_c^d \in \sum_{c=1}^{p} \frac{\mathcal{Q}(\mathfrak{g}_{J_c})[\mathbf{x}(I_c)]_{d+k\omega_c}}{\Theta_c^k}\,. \tag{8.1.7}$$

8.2 Application to polynomial optimization

Based on Theorem 8.3, we build up two hierarchies of SDP relaxations for unconstrained sparse POPs and sparse POPs on possibly noncompact basic semialgebraic sets.

Consider the general POP:

$$f^\star := \inf_{\mathbf{x} \in S(\mathfrak{g})} f(\mathbf{x})\,. \tag{8.2.1}$$

When the condition (i) of Assumption 8.1 holds, with d defined as in (8.1.5) one notes:

- $\psi_d := \sum_{c=1}^{p} \theta_c^d$;
- $\phi := \prod_{c=1}^{p} \Theta_c$;
- $\phi_c := \phi/\Theta_c = \prod_{r=1, r\neq c}^{p} \Theta_r$, $c \in [p]$.

Let $\varepsilon > 0$ be fixed. Let us suppose that the conditions (i) and (ii) of Assumption 8.1 hold if $S(\mathfrak{g}) = \mathbb{R}^n$, otherwise let us suppose that the conditions (i), (ii) and (iii) of Assumption 8.1 hold.

8.2.1 Semidefinite relaxations

Consider the hierarchy of semidefinite programs indexed by $k \in \mathbb{N}$:

$$\rho_k^{\text{sparse}}(\varepsilon) := \sup \left\{ \lambda \in \mathbb{R} \,:\, \phi^k (f - \lambda + \varepsilon \psi_d) \in \sum_{c=1}^{p} \phi_c^k \mathcal{Q}(\mathfrak{g}_{J_c})[\mathbf{x}(I_c)]_{d+k\omega_c} \right\}. \tag{8.2.2}$$

Theorem 8.4. *The following statements hold:*

1. *The sequence $(\rho_k^{\text{sparse}}(\varepsilon))_{k\in\mathbb{N}}$ is monotone non-decreasing.*
2. *Assume that problem* (8.2.1) *has an optimal solution $\mathbf{x}^\star$. Then there exists $K \in \mathbb{N}$ such that $f^\star \leq \rho_k^{\text{sparse}}(\varepsilon) \leq f^\star + \varepsilon\, \psi_d(\mathbf{x}^\star)$ for all $k \geq K$.*

The proof of Theorem 8.4 can be found in [133, Section 4].

8.2.2 Duality

For every $k \in \mathbb{N}$, the dual of (8.2.2) reads:

$$\begin{aligned} \tau_k^{\text{sparse}}(\varepsilon) := \inf_{\mathbf{y}} \quad & L_{\mathbf{y}}(\phi^k(f + \varepsilon\psi_d)) \\ \text{s.t.} \quad & \mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in \mathbb{N}^n_{2(d+pk\max_c \omega_c)}} \subset \mathbb{R}\,, \\ & \mathbf{M}_{d+k\omega_c}(\phi_c^k \mathbf{y}, I_c) \succeq 0\,,\ c \in [p]\,, \\ & \mathbf{M}_{d+k\omega_c-u_i}(\phi_c^k g_i \mathbf{y}, I_c) \succeq 0\,,\ i \in J_c\,,\ c \in [p]\,, \\ & L_{\mathbf{y}}(\phi^k) = 1\,. \end{aligned} \tag{8.2.3}$$

We guarantee strong duality for the pair of primal-dual problems (8.2.3)-(8.2.2).

Proposition 8.1. *Assume that $S(\mathfrak{g})$ has nonempty interior. Let $k \in \mathbb{N}$. Then $\tau_k^{\text{sparse}}(\varepsilon) = \rho_k^{\text{sparse}}(\varepsilon)$. Moreover, if $\tau_k^{\text{sparse}}(\varepsilon) > -\infty$, the optimal value $\rho_k^{\text{sparse}}(\varepsilon)$ is attained.*

The proof of Proposition 8.1 can be found in [133, Section 4].

Remark 8.2. *The condition $\tau_k^{\text{sparse}}(\varepsilon) > -\infty$ is always satisfied whenever k is sufficently large. Indeed by weak duality, when ε is fixed and k is sufficiently large then $\tau_k^{\text{sparse}}(\varepsilon) \geq \rho_k^{\text{sparse}}(\varepsilon) \geq f^\star > -\infty$. However, when k is small, $\tau_k^{\text{sparse}}(\varepsilon) = -\infty$ may happen.*

8.2.3 Sampling technique

Even though our sparse variant of Reznick's Positivstellensatz has an appealing shape, the computational benefits can be limited in the context of polynomial optimization. It is due to the very large number of constraints involved in the SDP relaxations when we clear the denominators in these certificates. To avoid this explosion of SDP constraints, one possible route is to use the sampling technique suggested in Parrilo and Löfberg [118] for semidefinite relaxations in polynomial optimization. To state that two polynomials are identical, instead of equating their coefficients one rather states that their respective values at sufficiently many points are equal.

More explicitly, in our context we take a sample $(\mathbf{a}_i)_{i=1}^N \subset \mathbb{R}^n$ and consider the following SDP relaxations indexed by $k \in \mathbb{N}$:

$$\begin{aligned} \rho_{k,N}^{\text{sample}}(\varepsilon) := \sup \quad & \lambda \\ \text{s.t} \quad & f(\mathbf{a}_i) - \lambda + \varepsilon \sum_{c=1}^p \theta_c(\mathbf{a}_i)^d = \sum_{c=1}^p \frac{\sigma_{0,c}(\mathbf{a}_i) + \sum_{j \in J_c} \sigma_{j,c}(\mathbf{a}_i) g_j(\mathbf{a}_i)}{\Theta_c(\mathbf{a}_i)^k}\,, \\ & \qquad\qquad i \in [N]\,, \\ & \sigma_{0,c} \in \Sigma[\mathbf{x}(I_c)]_{d+k\omega_c}\,, \\ & \sigma_{j,c} \in \Sigma[\mathbf{x}(I_c)]_{d+k\omega_c - \lceil g_j \rceil}\,,\ j \in J_c\,,\ c \in [p]\,. \end{aligned} \tag{8.2.4}$$

It is not hard to see that

$$\rho_k^{\text{sparse}}(\varepsilon) \leq \rho_{k,N}^{\text{sample}}(\varepsilon) \leq \rho_{k,N-1}^{\text{sample}}(\varepsilon)\,. \tag{8.2.5}$$

Since the denominators Θ_c^k have fixed forms, their evaluations $\Theta_c(\mathbf{a}_i)^k$ become constants in SDP (8.2.4). Thus, we do not need to clear the denominators as in the SDP (8.2.2) with value $\rho_k^{\text{sparse}}(\varepsilon)$. The constraints

$$f(\mathbf{a}_i) - \lambda + \varepsilon \sum_{c=1}^p \theta_c(\mathbf{a}_i)^d = \sum_{c=1}^p \frac{\sigma_{0,c}(\mathbf{a}_i) + \sum_{j \in J_c} \sigma_{j,c}(\mathbf{a}_i) g_j(\mathbf{a}_i)}{\Theta_c(\mathbf{a}_i)^k}\,, \quad i \in [N]\,,$$

directly provide linear constraints on the coefficients of the polynomial weights $\sigma_{j,c}$.

The underlying rationale behind (8.2.4) comes from the following observation in the case of sparse polynomial optimization, i.e., without denominators. Let us consider the simplified equality constraint:

$$\sum_{c=1}^p f_c = \sum_{c=1}^p (\sigma_{0,c} + \sum_{i \in J_c} \sigma_{i,c} g_i)\,, \tag{8.2.6}$$

with $\sigma_{0,c} \in \Sigma[\mathbf{x}(I_c)]_k$ and $\sigma_{i,c} \in \Sigma[\mathbf{x}(I_c)]_{k - \lceil g_i \rceil}$.

The polynomials involved in both left-hand side and right-hand side of (8.2.6) have degree less than $2k$.

If we do not exploit sparsity, we need to sample $N = \binom{2k+n}{n}$ points $(\mathbf{a}_i)_{i=1}^N \subset \mathbb{R}^n$, in order to potentially obtain a set of equality constraints equivalent to (8.2.6). Let us consider the multi-dimensional Vandermonde matrix associated to $(\mathbf{a}_i)_{i=1}^N$, with rows indexed by all monomials of degree $\leq 2k$. If this Vandermonde matrix is invertible, then the set of equations obtained after sampling is equivalent to (8.2.6) as a consequence of Theorem 4.1 from [153], namely the N points are in this case so-called "polynomially poised" if and only if they do not belong to a common algebraic hypersurface of degree $\leq 2k$.

If we exploit sparsity, we can do a similar reasoning with the multi-dimensional sparse Vandermonde matrix, with rows indexed by all sparse monomials of degree $\leq 2k$. In this case, the number of rows is upper bounded by $\sum_{c=1}^p \binom{2k+c}{c}$, and so is the number of required sample points. By using Theorem 4.1 from [153], the points are poised if and only if the Vandermonde matrix is non-singular, which guarantees a unique sparse interpolating polynomial of degree $\leq 2k$ passing through the data points.

Obtaining similar equivalent conditions in the case of rational function evaluation is left for future work. Our numerical experiments presented in the following section suggest that we obtain the same value as $\rho_k^{\text{sparse}}(\varepsilon)$ with (8.2.4), after selecting a large enough number N of samples.

8.2.4 Numerical experiments

The main goal of this section is to illustrate the correctness of our representation results, on a sample of nontrivial polynomials involving up to 10 variables. Let us report the numerical results obtained with SDP (8.2.2) to approximate the minimum of quadratic polynomials on the nonnegative orthant. The quadratic polynomials are generated randomly as follows:

1. Take $u \in \mathbb{N}$, $p := \lfloor n/u \rfloor$ and
$$I_c = \begin{cases} \{1, \ldots, u\} & \text{if } c = 1\,, \\ \{u(c-1), \ldots, uc\} & \text{if } c \in \{2, \ldots, p-1\}\,, \\ \{u(p-1), \ldots, n\} & \text{if } c = p\,. \end{cases} \tag{8.2.7}$$

2. Let $f = \sum_{c=1}^p f_c$ such that $f_c \in \mathbb{R}[\mathbf{x}(I_c)]_2$, $c \in [p]$. For each $c \in [p]$, the coefficient $f_{c,\boldsymbol{\alpha}}$ of f_c is generated randomly in $(0,1)$ with respect to the uniform distribution, for every $\boldsymbol{\alpha} \in \mathbb{N}_2^{I_c} \backslash \{\mathbf{0}\}$ and $f_{c,\mathbf{0}} = 0$.

3. Take $m = n$ and $g_j := x_j$, $j \in [n]$. Then $S(\mathfrak{g})$ is the nonnegative orthant which is a noncompact set. Set $J_1 = I_1$ and $J_c = I_c \backslash \cup_{i=1}^{c-1} I_i$, for $c = 2, \ldots, p$.

4. Take a sequence $(\mathbf{a}_i)_{i=1}^N \subset [-1,1]^n$ of uniformly random sample points as mentioned in Section 8.2.3.

With the above setting, the conditions (i), (ii) and (iii) of Assumption 8.1 hold and the optimal value of the corresponding POP (8.2.1) should be $f^\star = 0$.
The experiments are performed with the `JuMP` Julia package [53], relying on the `Mosek` solver [41] to solve the SDP relaxation (8.2.2). We use a desktop computer with an Intel(R) Core(TM) i7-8665UCPU @ 1.9GHz×8 and 31.2 GB of RAM. The numerical results are displayed in Table 5.16 with the following information:

- nummat: number of semidefinite matrices involved in SDP (8.2.2);
- numcons: number of constraints involved in SDP (8.2.2);
- val: the optimal value of SDP (8.2.2);
- time: the total time in seconds;
- the symbol "−" means that the SDP solver runs out of memory.

Table 8.2: Numerical results obtained when solving the SDP relaxations (8.2.2), (8.2.4) associated to the minimization of quadratic forms on the nonnegative orthant with $\varepsilon = 10^{-5}$ and $k = 1$.

POP size		SDP size		$\rho_k^{\text{sparse}}(\varepsilon)$ (JuMP+`Mosek`)		$\rho_{k,N}^{\text{sample}}(\varepsilon)$ (JuMP+`Mosek`)		
n	u	nummat	numcons	val	time	N	val	time
5	2	10	432	3.0×10^{-5}	3	100	3.0×10^{-5}	3
7	2	14	2368	6.0×10^{-5}	3	140	6.0×10^{-5}	4
10	3	20	22528	9.0×10^{-5}	109	200	9.0×10^{-5}	17
12	4	22	229520	–	–	240	1.0×10^{-4}	65
15	5	29	1157120	–	–	300	1.3×10^{-4}	2601

For $n \leq 10$, the optimal value obtained at the first order SDP relaxations (8.2.2) and (8.2.4) is very close to the exact optimal value of POP (8.2.1). Besides the solver runs out of memory for $n \geq 12$ when we compute $\rho_k^{\text{sparse}}(\varepsilon)$ as the number of equality constraints in SDP (8.2.2) is already very large even for $k = 1$. This large number of constraints arises while computing the common denominator within the sparse representation. Explicitly, it is the number of terms of the polynomial $\phi^k(f - \lambda + \varepsilon\psi_d)$, according to (8.2.2). This number is upper bounded by $\binom{n+2d+k\deg(\phi)}{n}$, where d is the smallest positive integer such that $2d > \deg(f)$. Note that k needs to be sufficiently large to ensure that the cone $\sum_{c=1}^{p} \phi_c^k \mathcal{Q}(\mathfrak{g}_{J_c})[\mathbf{x}(I_c)]_{d+k\omega_c}$ in (8.2.2) is well-defined, namely $2(d + k \min_{c\in[p]} \omega_c) \geq \max_{i\in[m]} \deg(g_i)$. Nevertheless, we still obtain the optimal value of SDP relaxation (8.2.4) for $n \geq 12$ by using the sampling technique from Section 8.2.3. The value is also close to the exact optimal value of POP (8.2.1). The number of constraints involved in SDP (8.2.4) is equal to the sample size N which is much smaller than the one involved in SDP (8.2.2).

Chapter 9

Exploiting nonnegativity of variables

In the previous three chapters, we have relied on uniform denominators to solve POPs over possibly noncompact semialgebraic sets. In this chapter, we present another advantage related to uniform denominators. More explicitly, such denominators enable us to decompose the Gram matrices, arising in the SOS strengthenings of POPs over the nonnegative orthant, into finitely many smaller ones with prescribed sizes.

For each $\mathcal{A} \subset \mathbb{N}^n$, denote $\mathbf{v}_{\mathcal{A}}(\mathbf{x}) = (\mathbf{x}^{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathcal{A}}$. We say that a polynomial q is *even in each variable* if for every $j \in [n]$, $q(x_1, \ldots, x_{j-1}, -x_j, x_{j+1}, \ldots, x_n) = q(x_1, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_n)$. A polynomial q is called a *SOS of monomials* if $q = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^n} \lambda_{\boldsymbol{\alpha}} \mathbf{x}^{2\boldsymbol{\alpha}}$ for some $\lambda_{\boldsymbol{\alpha}} \geq 0$. Accordingly, if q is an SOS of monomials, then $q = \mathbf{v}_d^\top \operatorname{diag}(\mathbf{u})\mathbf{v}_d$ for some $d \in \mathbb{N}$ and $\mathbf{u} \in \mathbb{R}_+^{b(n,d)}$.

Factor width: Originally defined in [23], the factor width of a real positive semidefinite matrix $\mathbf{G}$ is the smallest integer s for which there exists a real matrix $\mathbf{P}$ such that $\mathbf{G}$ can be decomposed as $\mathbf{G} = \mathbf{P}\mathbf{P}^\top$ and each column of $\mathbf{P}$ contains at most s nonzeros. In this case, if $\mathbf{u}$ is a vector of several monomials in $\mathbf{x}$, the SOS polynomial $\mathbf{u}^\top\mathbf{G}\mathbf{u}$ can be written as $\mathbf{u}(\mathbf{x})^\top\mathbf{G}\mathbf{u}(\mathbf{x}) = \sum_i (\mathbf{q}_i^\top \mathbf{u}(\mathbf{x}))^2$, where $\mathbf{q}_i$ is the i-th column of $\mathbf{P}$. It is not hard to prove that the Gram matrix of each square $(\mathbf{q}_i^\top \mathbf{u}(\mathbf{x}))^2$ has size at most s since $\mathbf{q}_i$ has at most s nonzeros. Thus, if an SOS polynomial has Gram matrix of factor width at most s, it can be written as a sum of SOS polynomials with Gram matrix sizes at most s. The inverse also holds true thanks to eigen decomposition. The applications of factor width for polynomial optimization can be found in, e.g., [3, 141].

POP with nonnegative variables: In the present chapter, we focus on the following POP on the nonnegative orthant:

$$f^\star := \inf_{\mathbf{x}\in S} f(\mathbf{x}) \,, \tag{9.0.1}$$

where f is a polynomial and S is a semialgebraic set defined by

$$S := \{\mathbf{x} \in \mathbb{R}^n \,:\, x_j \geq 0 \,,\, j \in [n] \,,\, g_i(\mathbf{x}) \geq 0 \,,\, i \in [m]\} \,, \tag{9.0.2}$$

for some $g_i \in \mathbb{R}[\mathbf{x}]$, $i \in [m]$ with $g_m := 1$. Letting $\check{q}(\mathbf{x}) := q(\mathbf{x}^2)$ (with $\mathbf{x}^2 := (x_1^2, \ldots, x_n^2)$) whenever $q \in \mathbb{R}[\mathbf{x}]$, it follows immediately that problem (9.0.1) is equivalent to solving

$$f^\star = \inf_{\mathbf{x}\in\check{S}} \check{f}(\mathbf{x}) \,, \tag{9.0.3}$$

where $\check{S}$ is a subset of $\mathbb{R}^n$ defined by

$$\check{S} := \{\mathbf{x} \in \mathbb{R}^n \,:\, \check{g}_i(\mathbf{x}) \geq 0 \,,\, i \in [m]\} \,. \tag{9.0.4}$$

Contribution. Our contribution is twofold:

I. In our first contribution, we provide in Corollary 9.2 a degree bound for the extension of Pólya's Positivstellensatz originally stated in [45]. Explicitly, if

- $\check{f}, \check{g}_1, \ldots, \check{g}_m$ are polynomials even in each variable,
- $\check{S}$ defined as in (9.0.4) has nonempty interior, $\check{g}_1 = R - \|\mathbf{x}\|_2^2$ for some $R > 0$,
- $\check{f}$ is of degree at most $2d_f$, each $\check{g}_i$ is of degree at most $2d_{g_i}$, and $\check{f} - f^\star$ is nonnegative on $\check{S}$,

then there exist positive constants $\bar{\mathfrak{c}}$ and $\mathfrak{c}$ depending on $\check{f}, \check{g}_i$ such that for all $\varepsilon > 0$, for all $k \geq \bar{\mathfrak{c}}\varepsilon^{-\mathfrak{c}}$,

$$(1 + \|\mathbf{x}\|_2^2)^k(\check{f} - f^\star + \varepsilon) = \textstyle\sum_{i\in[m]} \sigma_i \check{g}_i \,, \tag{9.0.5}$$

for some σ_i being SOS of monomials such that $\deg(\sigma_i \check{g}_i) \leq 2(k + d_f)$. (Here $\check{g}_m := 1$.)

Consequently, the resulting LP-hierarchy of lower bounds $(\rho_k^{\text{Pól}})_{k\in\mathbb{N}}$ for POP (9.0.3):

$$\begin{aligned} \rho_k^{\text{Pól}} := \sup_{\lambda, \mathbf{u}_i} \quad & \lambda \\ \text{s.t.} \quad & \lambda \in \mathbb{R}\,,\, \mathbf{u}_i \in \mathbb{R}_+^{b(n,k_i)}\,,\, i \in [m]\,, \\ & \theta^k(\check{f} - \lambda) = \textstyle\sum_{i\in[m]} \check{g}_i \mathbf{v}_{k_i}^\top \operatorname{diag}(\mathbf{u}_i)\mathbf{v}_{k_i}\,. \end{aligned} \tag{9.0.6}$$

where $k_i := k + d_f - d_{g_i}$, for $i \in [m]$. converges to $f^\star$ with a rate at least $\mathcal{O}(\varepsilon^{-\mathfrak{c}})$. This linear hierarchy is originally stated by Dickinson and Povh in [46] without convergence rate.

Unfortunately, for large relaxation order k this LP is potentially ill-conditioned (see for instance Example 9.2). In order to address this issue, we replace each diagonal Gram matrix $\operatorname{diag}(u_j)$ in LP (9.0.6) by a Gram matrix of factor width at most $s \in \mathbb{N}_{>0}$ to obtain a semidefinite relaxation, which is tighter than LP (9.0.6). Namely, consider the following SDP indexed by $k \in \mathbb{N}$ and $s \in \mathbb{N}_{>0}$:

$$\begin{aligned} \rho_{k,s}^{\text{Pól}} := \sup_{\lambda, \mathbf{G}_{ij}} \quad & \lambda \\ \text{s.t.} \quad & \lambda \in \mathbb{R}\,,\, \mathbf{G}_{ij} \succeq 0\,,\, j \in [b(n,k_i)]\,,\, i \in [m]\,, \\ & \theta^k(\check{f} - \lambda) = \textstyle\sum_{i\in[m]} \check{g}_i \big(\sum_{j\in[b(n,k_i)]} \mathbf{v}_{\mathcal{A}_j^{(s,k_i)}}^\top \mathbf{G}_{ij} \mathbf{v}_{\mathcal{A}_j^{(s,k_i)}} \big)\,. \end{aligned} \tag{9.0.7}$$

where each $\mathcal{A}_r^{(s,d)} \subset \mathbb{N}_d^n$, chosen as in Section 9.2.2, is such that $(\mathcal{A}_r^{(s,d)})_{r\in[b(n,d)]}$ covers $\mathbb{N}_d^n$, i.e.,

$$\cup_{r=1}^{b(n,d)} \mathcal{A}_r^{(s,d)} = \mathbb{N}_d^n\,, \tag{9.0.8}$$

and the cardinal number of $\mathcal{A}_r^{(s,d)}$ is at most s. Here $\check{g}_m := 1$. We call s the factor width upper bound associated with the semidefinite relaxation (9.0.7). It is easy to see that the size of each Gram matrix $\mathbf{G}_{ij}$ in (9.0.7) is at most s. In addition, due to (9.0.8), we obtain the following estimate for every $s \in [b(n,k)]$:

$$\rho_k^{\text{Pól}} = \rho_{k,1}^{\text{Pól}} \leq \rho_{k,s}^{\text{Pól}} \leq f^\star\,, \tag{9.0.9}$$

so that for every fixed $s \in \mathbb{N}_{>0}$, $\tau_{k,s}^{\text{Pól}} \to f^\star$ as k increases, with a rate at least $\mathcal{O}(\varepsilon^{-\mathfrak{c}})$. Notice that when $s = 2$, (9.0.7) becomes an SOCP thanks to [123, Lemma 15].

We emphasize that in our semidefinite relaxation (9.0.7), for fixed k the size of Gram matrices $\mathbf{G}_{ij}$ can be bounded from above by any $s \in \mathbb{N}_{>0}$ while the maximal matrix size of the standard semidefinite relaxation for POP (9.0.1) (defined as in Section 2.4) is fixed for each relaxation order k. Nevertheless, since we convert (9.0.1) to the form (9.0.3) (so as to use Corollary 9.2), the degrees of the resulting objective and constraint polynomials are doubled, i.e., $\deg(\check{f}) = 2\deg(f)$ and $\deg(\check{g}_i) = 2\deg(g_i)$.

However, numerical experiments in Sections 9.3 and 9.4.7 suggest that our method works better than existing methods on examples of POPs with nonnegative variables. For instance, for 20-variable dense POPs on the nonnegative orthant, the standard SOS-relaxations based on Putinar's Positivstellensatz provide a lower bound for $f^\star$ in 356 seconds while we can provide a better lower bound in 5 seconds.

Next, in Sections 9.4.6 and 9.4.6 we provide two convergent hierarchies of linear and semidefinite relaxations for POPs on the nonegative orthant, that exploit *correlative sparsity*, and with properties similar to those of (9.0.6) and (9.0.7). Accordingly, for POPs on the nonnegative orthant with up to 1000 variables, we can provide lower bounds in 19 seconds which are better than those

obtained in 56360 seconds with the sparsity-adapted version of the standard SOS-relaxations of Waki et al. [203].

II. In our second contribution, we provide a degree bound for an extended version of Handelman's Positivstellensatz to arbitrary compact basic semialgebraic sets. More explicitly, Corollary 9.3 states the following result. If

- $\check{f}, \check{g}_1, \ldots, \check{g}_m$ are polynomials even in each variable,
- $\check{S}$ defined as in (9.0.4) has nonempty interior, $\check{g}_1 = R - \|\mathbf{x}\|_2^2$ for some $R > 0$,
- $\check{g}_i$ is of degree at most $2d_{g_i}$ and $\check{f} - f^\star$ is nonnegative on $\check{S}$,

then there exist positive constants $\bar{\mathfrak{c}}$ and $\mathfrak{c}$ depending on $\check{f}, \check{g}_i$ such that for all $\varepsilon > 0$, for all $k \geq \bar{\mathfrak{c}}\varepsilon^{-\mathfrak{c}}$,

$$(\check{f} - f^\star) + \varepsilon = \sum_{i\in[m]} \sum_{j=0}^{k-d_{g_i}} \sigma_{ij}\check{g}_i\check{g}_1^j\,, \tag{9.0.10}$$

for some σ_{ij} being SOS of monomials such that $\deg(\sigma_{ij}\check{g}_i\check{g}_1^j) \leq 2k$. (Here $\check{g}_m := 1$.)

When compared with the extension of Pólya's Positivstellensatz in (9.0.5), the one of Handelman's Positivstellensatz in (9.0.10) does not have multiplier $(1 + \|\mathbf{x}\|_2^2)^k$ but its number of SOS of monomials is $\sum_{i=1}^m (k - d_{g_i} + 1)$ becomes larger when k increases. In contrast, the extension of Pólya's Positivstellensatz involves the same multiplier and its number of SOS of monomials is $m + 1$, so does not depend on k.

As a consequence, we obtain in Section 9.2.2 the rate of convergence for the hierarchy of linear relaxations (9.2.8) based on the extension of Handelman's Positivstellensatz. In addition, we also propose a new hierarchy of semidefinite relaxations in (9.2.22) based on even symmetry and the concept of factor width similarly to the one relying on Pólya's Positivstellensatz. A sparse version of this semidefinite hierarchy is also obtained in Section 9.4.6.

As shown in Sections 9.3 and 9.4.7, these hierarchies of semidefinite relaxations have the same numerical behavior as the ones based on Pólya's Positivstellensatz. In almost all cases, the ones based on the extension of Handelman's Positivstellensatz are several times slower but provide slightly better accurate bounds, compared to the ones based on the extension of Pólya's Positivstellensatz.

Related works

Exploiting sparsity: Structure exploitation in (9.0.7), is comparable to term sparsity and correlative sparsity (see Chapter 3) but here we can deal with dense POPs of the form (9.0.1). Moreover, the maximal block sizes involved in the sparsity-exploiting SDP relaxations mainly depend on the POP itself as well as on the relaxation order. By comparison, the maximal block size of our SDP relaxations is controllable. Under mild conditions, the rate of convergence $\rho_{k,s}^{\mathrm{Han}} \to f^\star$ as k increases, is at least $\mathcal{O}(\varepsilon^{-\mathfrak{c}})$.

Dickinson–Povh's hierarchy of linear relaxations: Dickinson and Povh state in [45] a specific constrained version of Pólya's Positivstellensatz. Explicitly, if $f, g_1, \ldots, g_m$ are homogeneous polynomials, S is defined as in (9.0.2), and f is positive on $S\backslash\{\mathbf{0}\}$, then

$$(\sum_{j\in[n]} x_j)^k f = \sum_{i\in[m]} \sigma_i g_i\,, \tag{9.0.11}$$

for some homogeneous polynomials σ_i with positive coefficients. (Here $g_m := 1$.) They also construct a hierarchy of linear relaxations associated with (9.0.11).

The extension of Pólya's Positivstellensatz restated in Corollary 9.2 is indeed analogous to (9.0.11). However the approach is different and importantly, the result is more convenient as we provide *degree bounds* for the SOS of monomials involved in the representation. Similarly, our corresponding linear relaxations (9.2.6) are the analogues to those of Dickinson and Povh [46]. As shown in Example 9.2 and other examples in Sections 9.3 and 9.4.7, this hierarchy of linear relaxations usually have a poor numerical behavior in practice when k is large. Our new hierarchy of semidefinite relaxations (9.2.19) is used to improve this issue.

DSOS and SDSOS: Recent work of Ahmadi and Majumdar [3] presents two alternative cones for SOS cones, namely, DSOS and SDSOS, which involve factor widths at most 2 and are more tractable than SOS cones. In the unconstrained case of POP (9.0.3), our semidefinite hierarchy based on the extension of Pólya's Positivstellensatz can be seen as a generalization of DSOS and SDSOS while using the notion of factor width, see Remark 9.17. In fact, to obtain our semidefinite relaxations for the constrained case (9.0.3), we replace each SOS of monomials involved in the certificate (9.0.5) by an SOS polynomial whose Gram matrix has factor width at most s; see Remark 9.10.

9.1 Representation theorems

In this section, we derive representations of polynomials nonnegative on semialgebraic sets together with degree bounds.

9.1.1 Polynomials nonnegative on general semialgebraic sets

Extension of Pólya's Positivstellensatz: We analyze the complexity of the extension of Pólya's Positivstellensatz in the following theorem:

Theorem 9.1. *(Homogenized representation) Let $g_1, \dots, g_m$ be homogeneous polynomials such that $g_1, \dots, g_m$ are even in each variable. Let S be the semialgebraic set defined by*

$$S := \{\mathbf{x} \in \mathbb{R}^n \,:\, g_1(\mathbf{x}) \geq 0 \dots, g_m(\mathbf{x}) \geq 0\}\,. \tag{9.1.1}$$

Let f be a homogeneous polynomial of degree $2d_f$ for some $d_f \in \mathbb{N}$ such that f is even in each variable and nonnegative on S. Then the following statements hold:

1. *For all $\varepsilon > 0$, there exists $K_\varepsilon \in \mathbb{N}$ such that for all $k \geq K_\varepsilon$, there exist homogeneous SOS of monomials σ_i satisfying*

$$\deg(\sigma_0) = \deg(\sigma_1 g_1) = \cdots = \deg(\sigma_m g_m) = 2(k + d_f) \tag{9.1.2}$$

and

$$\|\mathbf{x}\|_2^{2k}(f + \varepsilon\|\mathbf{x}\|_2^{2d_f}) = \sigma_0 + \sigma_1 g_1 + \cdots + \sigma_m g_m\,. \tag{9.1.3}$$

2. *If S has nonempty interior, there exist positive constants $\bar{\mathfrak{c}}$ and $\mathfrak{c}$ depending on f, g_i such that for all $\varepsilon > 0$, one can take $K_\varepsilon = \bar{\mathfrak{c}}\varepsilon^{-\mathfrak{c}}$.*

The proof of Theorem 9.1 is postponed to Section 9.4.2.

Note that some other homogeneous representations for globally nonnegative polynomials even in each variable have been studied in [64, 72, 37].

Remark 9.1. *The Gram matrix associated with each SOS of monomials is diagonal. In other word, it is a block-diagonal matrix with maximal block size one. It would be interesting to know for which types of input polynomials we could obtain other representations involving SOS with block-diagonal Gram matrices of very small maximal block size, similarly to Theorem 9.1. Some of them have been discussed in [66, 123] that includes SOS of binomials, trinomials, tetranomials and SOS of any s-nomials. We emphasize that such representations allow one to build up SDP relaxations of small maximal matrix size that can be solved efficiently by using interior-point methods as shown later in Section 9.3.*

The following corollary is a direct consequence of Theorem 9.1.

Corollary 9.1. *(Dehomogenized representation) Let $g_1, \dots, g_m$ be polynomials even in each variable. Let S be the semialgebraic set defined by* (9.1.1)*. Let f be a polynomial even in each variable and nonnegative on S. Denote $d_f := \lfloor \deg(f)/2 \rfloor + 1$. Then the following statements hold:*

1. *For all $\varepsilon > 0$, there exists $K_\varepsilon \in \mathbb{N}$ such that for all $k \geq K_\varepsilon$, there exist SOS of monomials σ_i satisfying*

$$\deg(\sigma_0) \leq 2(k + d_f) \quad \text{and} \quad \deg(\sigma_i g_i) \leq 2(k + d_f)\,,\ i \in [m]\,, \tag{9.1.4}$$

and

$$\theta^k(f + \varepsilon\theta^{d_f}) = \sigma_0 + \sigma_1 g_1 + \cdots + \sigma_m g_m\,, \tag{9.1.5}$$

where $\theta := 1 + \|\mathbf{x}\|_2^2$.

2. *If S has nonempty interior, there exist positive constants $\bar{\mathfrak{c}}$ and $\mathfrak{c}$ depending on f, g_i such that for all $\varepsilon > 0$, one can take $K_\varepsilon = \bar{\mathfrak{c}}\varepsilon^{-\mathfrak{c}}$.*

The proof of Corollary 9.1 is similar to the proof of [132, Corollary 1].

9.1.2 Polynomials nonnegative on compact semialgebraic sets

In this section, we provide a representation of polynomials nonnegative on semialgebraic sets when the input polynomials are even in each variable. We also derive in Section 9.4.5 some sparse representations when the input polynomials have correlative sparsity.

Extension of Pólya's Positivstellensatz

The following corollary is deduced from Corollary 9.1.

Corollary 9.2. *Let f, g_i, S, d_f be as in Corollary 9.1 such that $g_1 := R - \|\mathbf{x}\|_2^2$ for some $R > 0$. Then the following statements hold:*

1. *For all $\varepsilon > 0$, there exists $K_\varepsilon \in \mathbb{N}$ such that for all $k \geq K_\varepsilon$, there exist SOS of monomials σ_i satisfying* (9.1.4) *and*

$$(1 + \|\mathbf{x}\|_2^2)^k(f + \varepsilon) = \sigma_0 + \sigma_1 g_1 + \cdots + \sigma_m g_m\,. \tag{9.1.6}$$

2. *If S has nonempty interior, there exist positive constants $\bar{\mathfrak{c}}$ and $\mathfrak{c}$ depending on f, g_i such that for all $\varepsilon > 0$, one can take $K_\varepsilon = \bar{\mathfrak{c}}\varepsilon^{-\mathfrak{c}}$.*

Corollary 9.2 can be proved in the same way as [132, Corollary 2].

Remark 9.2. *If we remove the multiplier $(1 + \|\mathbf{x}\|_2^2)^k$ in* (9.1.6)*, Corollary 9.2 is no longer true. Indeed, let $n = 1$, $f := (x^2 - \frac{3}{2})^2$ and assume that $f = \sigma_0 + \sigma_1(1 - x^2)$ for some SOS of monomials σ_i, $i = 0, 1$. Note that f is even and positive on $[-1, 1]$. We write $\sigma_i := a_i + b_i x^2 + x^4 r_i(x)$ for some $a_i, b_i \in \mathbb{R}_+$ and $r_i \in \mathbb{R}[x]$. It implies that*

$$x^4 - 3x^2 + \tfrac{9}{4} = (a_0 + b_0 x^2 + x^4 r_0(x)) + (a_1 + b_1 x^2 + x^4 r_1(x))(1 - x^2)\,. \tag{9.1.7}$$

Then we obtain the system of linear equations: $\frac{9}{4} = a_0 + a_1$ and $-3 = b_0 - a_1 + b_1$. Summing gives $-\frac{3}{4} = a_0 + b_0 + b_1$. However, $a_0 + b_0 + b_1 \geq 0$ since $a_i, b_i \in \mathbb{R}_+$. This contradiction yields the conclusion. Thus, Putinar's Positivstellensatz with SOS of monomials does not exist when the input polynomials are even in each variable. However, we are still able to exploit term sparsity/even symmetry for Putinar's Positivstellensatz in this case as shown later in Proposition 9.1.

It is not hard to see that with the multiplier $(1 + x^2)^2$, we obtain the Pólya's Positivstellensatz as follows:

$$(1 + x^2)^2 f = \bar{\sigma}_0 + \bar{\sigma}_1(1 - x^2)\,, \tag{9.1.8}$$

where $\bar{\sigma}_0 := x^8$ and $\bar{\sigma}_1 := x^4 + \frac{15}{4}x^2 + \frac{9}{4}$ are SOS of monomials.

We prove in the following proposition the existence of block-diagonal Gram matrices in Putinar's Positivstellensatz when the input polynomials are even in each variable:

Proposition 9.1. *Let $f, g_1, \ldots, g_m$ be polynomials in $\mathbb{R}[\mathbf{x}]$ such that f, g_i are even in each variable. Assume that there exists a decomposition:*

$$f = \sum_{i=1}^{m} g_i \mathbf{v}_{d_i}^\top \mathbf{G}^{(i)} \mathbf{v}_{d_i} \,, \tag{9.1.9}$$

for some $d_i \in \mathbb{N}$ and real symmetric matrices $\mathbf{G}^{(i)} = (G^{(i)}_{\boldsymbol{\alpha},\boldsymbol{\beta}})_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{N}^n_{d_i}}$. For every $i \in [m]$, define $\bar{\mathbf{G}^{(i)}} := (\bar{G}^{(i)}_{\boldsymbol{\alpha},\boldsymbol{\beta}})_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{N}^n_{d_i}}$, where:

$$\bar{G}^{(i)}_{\boldsymbol{\alpha},\boldsymbol{\beta}} := \begin{cases} G^{(i)}_{\boldsymbol{\alpha},\boldsymbol{\beta}} & \text{if } \boldsymbol{\alpha}+\boldsymbol{\beta} \in 2\mathbb{N}^n \,, \\ 0 & \text{otherwise} \,. \end{cases} \tag{9.1.10}$$

Then $\bar{\mathbf{G}}^{(i)}$ are block-diagonal up to permutation and

$$f = \sum_{i=1}^{m} g_i \mathbf{v}_{d_i}^\top \bar{\mathbf{G}}^{(i)} \mathbf{v}_{d_i} \,. \tag{9.1.11}$$

Moreover, if $\mathbf{G}^{(i)} \succeq 0$, then $\bar{\mathbf{G}}^{(i)} \succeq 0$.

Proof. The proof is inspired by [61, Section 8.1]. Removing all terms in (9.1.9) except the terms of monomials $\mathbf{x}^{2\boldsymbol{\alpha}}$, $\boldsymbol{\alpha} \in \mathbb{N}^n$, we obtain (9.1.11). It is due to the fact that f, g_i only have terms of the form $\mathbf{x}^{2\boldsymbol{\alpha}}$, $\boldsymbol{\alpha} \in \mathbb{N}^n$ and

$$\mathbf{v}_{d_i}^\top \mathbf{G}^{(i)} \mathbf{v}_{d_i} = \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{N}^n_{d_i}} G^{(i)}_{\boldsymbol{\alpha},\boldsymbol{\beta}} \mathbf{x}^{\boldsymbol{\alpha}+\boldsymbol{\beta}} \,. \tag{9.1.12}$$

Next, we show the block-diagonal structure of $\bar{\mathbf{G}}^{(i)}$. For every $\boldsymbol{\gamma} \in \{0,1\}^n$, define

$$\Lambda^{(i)}_{\boldsymbol{\gamma}} := \{\boldsymbol{\alpha} \in \mathbb{N}^n_{d_i} \,:\, , \boldsymbol{\alpha} - \boldsymbol{\gamma} \in 2\mathbb{N}^n\} \,. \tag{9.1.13}$$

Then $\Lambda^{(i)}_{\boldsymbol{\gamma}} \cap \Lambda^{(i)}_{\boldsymbol{\eta}} = \emptyset$ if $\boldsymbol{\gamma} \neq \boldsymbol{\eta}$ and $\mathbb{N}^n_{d_i} := \cup_{\boldsymbol{\gamma}\in\{0,1\}^n} \Lambda^{(i)}_{\boldsymbol{\gamma}}$. In addition, for all $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \Lambda^{(i)}_{\boldsymbol{\gamma}}$, $\boldsymbol{\alpha}+\boldsymbol{\beta} \in 2\mathbb{N}^n$. Moreover, if $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}^n_{d_i}$ and $\boldsymbol{\alpha}+\boldsymbol{\beta} \in 2\mathbb{N}^n$, then there exists $\boldsymbol{\gamma} \in \{0,1\}^n$ such that $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \Lambda^{(i)}_{\boldsymbol{\gamma}}$. It implies that all blocks on the diagonal of $\bar{\mathbf{G}}^{(i)}$ must be

$$(\bar{G}^{(i)}_{\boldsymbol{\alpha},\boldsymbol{\beta}})_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\Lambda^{(i)}_{\boldsymbol{\gamma}}} \,,\, \boldsymbol{\gamma} \in \{0,1\}^n \,. \tag{9.1.14}$$

This yields the desired results. □

Remark 9.3. *The block-diagonal structure in Proposition 9.1 can be obtained by using TSSOS [212]. For general input polynomials f, g_i, we cannot ensure that the maximal block size in this form is upper bounded or possibly goes to infinity as each d_i increases. However, as shown in Remark 9.2, we cannot obtain blocks of size one for this form. In order to improve this, we provide another representation with diagonal Gram matrices in the next corollary.*

Extension of Handelman's Positivstellensatz

The following corollary is a consequence of Theorem 9.1.

Corollary 9.3. *(Dense representation without multiplier) Let f, g_i, S be as in Corollary 9.1 such that $g_1 := R - \|\mathbf{x}\|_2^2$ for some $R > 0$ and $g_m := 1$. Denote $d_{g_i} := \lceil \deg(g_i)/2 \rceil$. Then the following statements hold:*

1. *For all $\varepsilon > 0$, there exists $K_\varepsilon \in \mathbb{N}$ such that for all $k \geq K_\varepsilon$, there exist SOS of monomials $\sigma_{i,j}$ satisfying*

$$\deg(\sigma_{i,j} g_1^j g_i) \leq 2k \tag{9.1.15}$$

and

$$f + \varepsilon = \sum_{i=1}^{m} \sum_{j=0}^{k-d_{g_i}} \sigma_{i,j} g_1^j g_i \,. \tag{9.1.16}$$

2. *If S has nonempty interior, then there exist positive constants $\bar{\mathfrak{c}}$ and $\mathfrak{c}$ depending on f, g_i such that for all $\varepsilon > 0$, one can take $K_\varepsilon = \bar{\mathfrak{c}} \varepsilon^{-\mathfrak{c}}$.*

Proof. Denote $d_f := \lfloor \deg(f)/2 \rfloor + 1$. With an additional variable x_{n+1}, we first define the following homogeneous polynomials:

$$\bar{f} := \|(\mathbf{x}, x_{n+1})\|_2^{2d_f} f(\tfrac{\mathbf{x}\sqrt{R}}{\|(\mathbf{x}, x_{n+1})\|_2}) \quad \text{and} \quad \bar{g}_i := \|(\mathbf{x}, x_{n+1})\|_2^{2d_{g_i}} g_i(\tfrac{\mathbf{x}\sqrt{R}}{\|(\mathbf{x}, x_{n+1})\|_2}) \,. \tag{9.1.17}$$

It is not hard to prove that $\bar{f}$ is nonnegative on the semialgebraic set $\{(\mathbf{x}, x_{n+1}) \in \mathbb{R}^{n+1} : \bar{g}_i(\mathbf{x}, x_{n+1}) \geq 0 \,,\, i \in [m]\}$, so that Theorem 9.1 yields the representation

$$\|(\mathbf{x}, x_{n+1})\|_2^{2k} (\bar{f} + \varepsilon \|(\mathbf{x}, x_{n+1})\|_2^{2d_f}) = \sigma_1 \bar{g}_1 + \cdots + \sigma_m \bar{g}_m \,, \tag{9.1.18}$$

for some SOS of monomials σ_i. By replacing x_{n+1} by $\sqrt{R - \|\mathbf{x}\|_2^2}$, we obtain the results. □

Remark 9.4. *The number of SOS of monomials in the representation* (9.1.16) *is* $\sum_{i=1}^m (k - d_{g_i} + 1)$ *which becomes larger when* k *increases, while the number of SOS of monomials in the representation* (9.1.6) *is* $m + 1$*, thus does not depend on* k*. However, a large number of Gram matrices is not a computational issue, since the complexity of interior-point methods mainly depend on the maximal block sizes of the Gram matrices and are still efficient when their number is large.*

Remark 9.5. *With* f *being defined as in Remark 9.2, the following decomposition is an instance of the extended Handelman's Positivstellensatz:*

$$f = \eta_0 + \eta_1(1 - x^2) + \eta_2(1 - x^2)^2 \,, \tag{9.1.19}$$

where $\eta_0 = \frac{1}{4}$*,* $\eta_1 = \eta_2 = 1$ *are SOS of monomials. Note that the degrees of these SOS of monomials are zero while the degrees of the ones from* (9.1.8) *for the extension of Pólya's Positivstellensatz are* 8 *and* 4*.*

Remark 9.6. *In Section 9.4.3, we provide some variations of Pólya's and Handelman's Positivstellensatz where the input polynomials are not required to be even in each variable. Moreover, the weighted SOS polynomials of these representations are still associated with Gram matrices of factor width one thanks to a change of monomial basis.*

9.2 Polynomial optimization on the nonnegative orthant: Compact case

This section is concerned with some applications of the extensions of Pólya's and Handelman's Positivstellensatz for polynomial optimization on compact semialgebraic subsets of the nonnegative orthant. The noncompact case is postponed to Section 9.4.4. Moreover, Section 9.4.6 is devoted to some applications of the sparse representation provided in Section 9.4.5 for polynomial optimization with correlative sparsity.

Consider the following POP:

$$f^\star := \inf_{\mathbf{x} \in S} f(\mathbf{x}) \,, \tag{9.2.1}$$

where $f \in \mathbb{R}[\mathbf{x}]$ and

$$S = \{\mathbf{x} \in \mathbb{R}^n : x_j \geq 0 \,,\, j \in [n] \,,\, g_i(\mathbf{x}) \geq 0 \,,\, i \in [m]\} \,, \tag{9.2.2}$$

for some $g_i \in \mathbb{R}[\mathbf{x}]$, $i \in [m]$, with $g_m = 1$.

Throughout this section, we assume that $f^\star > -\infty$ and problem (9.2.1) has an optimal solution $x^\star$.

Remark 9.7. *Every general POP in variable* $\mathbf{x} = (x_1, \ldots, x_n)$ *can be converted to the form* (9.2.1) *by replacing each variable* x_j *by the difference of two new nonnegative variables* $x_j^+ - x_j^-$*. In particular, if there are several constraints* $x_j \geq a_j$*, we can obtain an equivalent POP on the nonnegative orthant by defining new nonnegative variables* $y_j := x_j - a_j$*. However, we restrict ourselves to POPs on the nonnegative orthant in this chapter.*

Recall that $\check{q}(\mathbf{x}) := q(\mathbf{x}^2)$, for a given polynomial q. In this case, $\check{q}$ is even in each variable. Then POP (9.2.1) is equivalent to

$$f^\star := \inf_{\mathbf{x}\in\check{S}} \check{f}\,, \tag{9.2.3}$$

where

$$\check{S} = \{\mathbf{x}\in\mathbb{R}^n \,:\, \check{g}_i(\mathbf{x}) \geq 0\,,\, i\in[m]\}\,, \tag{9.2.4}$$

with $\mathbf{x}^{\star 2}$ being an optimal solution.

Let $\theta := 1 + \|\mathbf{x}\|_2^2$. Denote $d_f := \deg(f) + 1$, $d_{g_i} := \deg(g_i)$, $i\in[m]$ and let $\mathrm{diag}(\cdot)$ stand for the vector of diagonal entries of a square matrix.

9.2.1 Linear relaxations

Based on the extension of Pólya's Positivstellensatz

Consider the hierarchy of linear programs indexed by $k\in\mathbb{N}$:

$$\begin{aligned} \tau_k^{\mathrm{Pól}} := \inf_{\mathbf{y}} \quad & L_{\mathbf{y}}(\theta^k\check{f}) \\ \text{s.t.} \quad & \mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{2(d_f+k)}} \subset \mathbb{R}\,,\, L_{\mathbf{y}}(\theta^k) = 1\,, \\ & \mathrm{diag}(\mathbf{M}_{k_i}(\check{g}_i\mathbf{y})) \in \mathbb{R}_+^{b(n,k_i)}\,,\, i\in[m]\,, \end{aligned} \tag{9.2.5}$$

where $k_i := k + d_f - d_{g_i}$, $i\in[m]$. Note that $\check{g}_m = 1$.

Remark 9.8. *The optimal value $\tau_k^{\mathrm{Pól}}$ only depends on the subset of variables $\{y_{2\boldsymbol{\alpha}} : \boldsymbol{\alpha}\in\mathbb{N}^n_{d_f+k}\}$, i.e., the optimal value of LP* (9.2.5) *does not change when we assign each of the other variables with any real number. It is due to the fact that θ, $\check{f}$, and $\check{g}_i$ only have nonzero coefficients associated to the monomials $\mathbf{x}^{2\boldsymbol{\alpha}}$ for some $\boldsymbol{\alpha}\in\mathbb{N}^n$.*

Theorem 9.2. *Let $f, g_i\in\mathbb{R}[\mathbf{x}]$, $i\in[m]$, with $g_m = 1$ and $g_1 := R - \sum_{j\in[n]} x_j$ for some $R > 0$. Consider POP* (9.2.1) *with S being defined as in* (9.2.2)*. For every $k\in\mathbb{N}$, the dual of* (9.2.5) *reads as:*

$$\begin{aligned} \rho_k^{\mathrm{Pól}} := \sup_{\lambda,\mathbf{u}_i} \quad & \lambda \\ \text{s.t.} \quad & \lambda\in\mathbb{R}\,,\, \mathbf{u}_i\in\mathbb{R}_+^{b(n,k_i)}\,,\, i\in[m]\,, \\ & \theta^k(\check{f} - \lambda) = \textstyle\sum_{i\in[m]} \check{g}_i\mathbf{v}_{k_i}^\top \mathrm{diag}(\mathbf{u}_i)\mathbf{v}_{k_i}\,. \end{aligned} \tag{9.2.6}$$

The following statements hold:

1. *For all $k\in\mathbb{N}$,*
$$\rho_k^{\mathrm{Pól}} \leq \rho_{k+1}^{\mathrm{Pól}} \leq f^\star\,. \tag{9.2.7}$$
2. *The sequence $(\rho_k^{\mathrm{Pól}})_{k\in\mathbb{N}}$ converges to $f^\star$.*
3. *If S has nonempty interior, there exist positive constants $\bar{\mathfrak{c}}$ and $\mathfrak{c}$ depending on f, g_i such that $0 \leq f^\star - \rho_k^{\mathrm{Pól}} \leq \left(\frac{k}{\bar{\mathfrak{c}}}\right)^{-\frac{1}{\mathfrak{c}}}$.*

The proof of Theorem 9.2 relies on Corollary 9.2 and can be proved in almost the same way as the proof of [132, Theorem 4].

Based on the extension of Handelman's Positivstellensatz

Consider the hierarchy of linear programs indexed by $k\in\mathbb{N}$:

$$\begin{aligned} \tau_k^{\mathrm{Han}} := \inf_{\mathbf{y}} \quad & L_{\mathbf{y}}(\check{f}) \\ \text{s.t.} \quad & \mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{2k}} \subset \mathbb{R}\,,\, y_{\mathbf{0}} = 1\,, \\ & \mathrm{diag}(\mathbf{M}_{k_{ij}}((\check{g}_i\check{g}_1^j)\mathbf{y})) \in \mathbb{R}_+^{b(n,k_{ij})}\,,\, i\in[m]\,,\, j\in\{0\}\cup[k-d_{g_i}]\,, \end{aligned} \tag{9.2.8}$$

where $k_{ij} := k - d_{g_i} - j$, for $i\in[m]$, for $j\in\{0\}\cup[k-d_{g_i}]$. Note that $\check{g}_m = 1$.

Theorem 9.3. *Let $f, g_i \in \mathbb{R}[\mathbf{x}]$, $i \in [m]$, with $g_m = 1$ and $g_1 := R - \sum_{t\in[n]} x_t$ for some $R > 0$. Consider POP* (9.2.1) *with S being defined as in* (9.2.2). *For every $k \in \mathbb{N}$, the dual of* (9.2.8) *reads as:*

$$\begin{aligned} \rho_k^{\mathrm{Han}} := \sup_{\lambda, \mathbf{u}_{ij}} \quad & \lambda \\ \text{s.t.} \quad & \lambda \in \mathbb{R}\,,\, \mathbf{u}_{ij} \in \mathbb{R}_+^{b(n,k_{ij})}\,,\, i \in [m]\,,\, j \in \{0\} \cup [k - d_{g_i}]\,, \\ & \check{f} - \lambda = \sum_{i\in[m]} \sum_{j=0}^{k-d_{g_i}} \check{g}_i \check{g}_1^j \mathbf{v}_{k_{ij}}^\top \operatorname{diag}(\mathbf{u}_{ij}) \mathbf{v}_{k_{ij}}\,. \end{aligned} \tag{9.2.9}$$

The following statements hold:

1. *For all $k \in \mathbb{N}$, $\rho_k^{\mathrm{Han}} \le \rho_{k+1}^{\mathrm{Han}} \le f^\star$.*
2. *The sequence $(\rho_k^{\mathrm{Han}})_{k\in\mathbb{N}}$ converges to $f^\star$.*
3. *If S has nonempty interior, there exist positive constants $\bar{\mathfrak{c}}$ and $\mathfrak{c}$ depending on f, g_i such that $0 \le f^\star - \rho_k^{\mathrm{Han}} \le \left(\frac{k}{\bar{\mathfrak{c}}}\right)^{-\frac{1}{\mathfrak{c}}}$.*

The proof of Theorem 9.3 relies on Corollary 9.3 and can be proved in almost the same way as the proof of Theorem 9.2.

9.2.2 Semidefinite relaxations

In this subsection, we construct the sparsity pattern $\mathcal{A}_j^{(s,d)} \subset \mathbb{N}_d^n$ inspired by even symmetry reduction in Proposition 9.1.

We write $\mathbb{N}^n = \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_r, \boldsymbol{\alpha}_{r+1}, \dots\}$ such that

$$\boldsymbol{\alpha}_1 < \boldsymbol{\alpha}_2 < \cdots < \boldsymbol{\alpha}_r < \boldsymbol{\alpha}_{r+1} < \dots\,. \tag{9.2.10}$$

Let

$$W_j := \{i \in \mathbb{N} : i \ge j\,,\, \boldsymbol{\alpha}_i + \boldsymbol{\alpha}_j \in 2\mathbb{N}^n\}\,, \quad j \in \mathbb{N}_{>0}\,. \tag{9.2.11}$$

Then for all $j \in \mathbb{N}_{>0}$, $W_j \neq \emptyset$ since $j \in W_j$. For every $j \in \mathbb{N}$, we write $W_j := \{i_1^{(j)}, i_2^{(j)}, \dots\}$ such that $j = i_1^{(j)} < i_2^{(j)} < \dots$. Let

$$\mathcal{T}_j^{(s,d)} = \{\boldsymbol{\alpha}_{i_1^{(j)}}, \dots, \boldsymbol{\alpha}_{i_s^{(j)}}\} \cap \mathbb{N}_d^n\,, \quad j, s \in \mathbb{N}_{>0}\,,\, d \in \mathbb{N}\,. \tag{9.2.12}$$

For every $s \in \mathbb{N}_{>0}$ and $d \in \mathbb{N}$, define $\mathcal{A}_1^{(s,d)} := \mathcal{T}_1^{(s,d)}$ and for $j = 2, \dots, b(n,d)$, define

$$\mathcal{A}_j^{(s,d)} := \begin{cases} \mathcal{T}_j^{(s,d)} & \text{if } \mathcal{T}_j^{(s,d)} \backslash \mathcal{A}_l^{(s,d)} \neq \emptyset\,,\, \forall l \in [j-1]\,, \\ \emptyset & \text{otherwise}\,. \end{cases} \tag{9.2.13}$$

Note that $\cup_{j=1}^{b(n,d)} \mathcal{A}_j^{(s,d)} = \mathbb{N}_d^n$ and $|\mathcal{A}_j^{(s,d)}| \le s$. Here $|\cdot|$ stands for the cardinality of a set. Then the sequence

$$(\boldsymbol{\alpha} + \boldsymbol{\beta})_{(\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathcal{A}_j^{(s,d)})}\,,\, j \in [b(n,d)] \tag{9.2.14}$$

are overlapping blocks of size at most s in $(\boldsymbol{\alpha} + \boldsymbol{\beta})_{(\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}_d^n)}$. Note that $\boldsymbol{\alpha} + \boldsymbol{\beta} \in 2\mathbb{N}^n$ for all $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathcal{A}_j^{(s,d)}$.

Example 9.1. *Consider the case of $n = d = s = 2$. The matrix $(\boldsymbol{\alpha} + \boldsymbol{\beta})_{(\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}_2^2)}$ can be written explicitly as*

$$\begin{bmatrix} (\mathbf{0},\mathbf{0}) & (1,0) & (0,1) & (\mathbf{2},\mathbf{0}) & (1,1) & (\mathbf{0},\mathbf{2}) \\ (1,0) & (\mathbf{2},\mathbf{0}) & (1,1) & (3,0) & (2,1) & (1,2) \\ (0,1) & (1,1) & (\mathbf{0},\mathbf{2}) & (2,1) & (1,2) & (0,3) \\ (\mathbf{2},\mathbf{0}) & (3,0) & (2,1) & (\mathbf{4},\mathbf{0}) & (3,1) & (\mathbf{2},\mathbf{2}) \\ (1,1) & (2,1) & (1,2) & (3,1) & (\mathbf{2},\mathbf{2}) & (1,3) \\ (\mathbf{0},\mathbf{2}) & (1,2) & (0,3) & (\mathbf{2},\mathbf{2}) & (1,3) & (\mathbf{0},\mathbf{4}) \end{bmatrix}. \tag{9.2.15}$$

In this matrix, the entries in bold belong to $2\mathbb{N}^2$. *Then* $\mathcal{A}_1^{(2,2)} = \{(0,0),(2,0)\}$, $\mathcal{A}_2^{(2,2)} = \{(1,0)\}$, $\mathcal{A}_3^{(2,2)} = \{(0,1)\}$, $\mathcal{A}_4^{(2,2)} = \{(2,0),(0,2)\}$, $\mathcal{A}_5^{(2,2)} = \{(1,1)\}$ *and* $\mathcal{A}_6^{(2,2)} = \emptyset$. *The blocks* $(\boldsymbol{\alpha}+\boldsymbol{\beta})_{(\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathcal{A}_j^{(2,2)})}$, $j \in [5]$, *are as follows:*

$$\begin{bmatrix} (\mathbf{0},\mathbf{0}) & (\mathbf{2},\mathbf{0}) \\ (\mathbf{2},\mathbf{0}) & (\mathbf{4},\mathbf{0}) \end{bmatrix}, \big[(\mathbf{2},\mathbf{0})\big], \big[(\mathbf{0},\mathbf{2})\big], \begin{bmatrix} (\mathbf{4},\mathbf{0}) & (\mathbf{2},\mathbf{2}) \\ (\mathbf{2},\mathbf{2}) & (\mathbf{0},\mathbf{4}) \end{bmatrix}, \big[(\mathbf{2},\mathbf{2})\big]. \tag{9.2.16}$$

For all $\mathcal{B} = \{\boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_r\} \subset \mathbb{N}^n$ such that $\boldsymbol{\beta}_1 < \cdots < \boldsymbol{\beta}_r$, for every $h = \sum_{\boldsymbol{\gamma}} h_{\boldsymbol{\gamma}}\mathbf{x}^{\boldsymbol{\gamma}} \in \mathbb{R}[\mathbf{x}]$ and for every $\mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n} \subset \mathbb{R}$, let us define

$$\mathbf{v}_{\mathcal{B}} := \begin{bmatrix} \mathbf{x}^{\boldsymbol{\beta}_1} \\ \dots \\ \mathbf{x}^{\boldsymbol{\beta}_r} \end{bmatrix} \quad \text{and} \quad \mathbf{M}_{\mathcal{B}}(h\mathbf{y}) := \big(\textstyle\sum_{\boldsymbol{\gamma}} h_{\boldsymbol{\gamma}} y_{\boldsymbol{\gamma}+\boldsymbol{\beta}_i+\boldsymbol{\beta}_j}\big)_{i,j\in[r]}. \tag{9.2.17}$$

Based on the extension of Pólya's Positivstellensatz

Consider the hierarchy of semidefinite programs indexed by $s \in \mathbb{N}_{>0}$ and $k \in \mathbb{N}$:

$$\begin{aligned} \tau_{k,s}^{\mathrm{Pól}} := \inf_{\mathbf{y}} \quad & L_{\mathbf{y}}(\theta^k \check{f}) \\ \text{s.t.} \quad & \mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{2(d_f+k)}} \subset \mathbb{R},\, L_{\mathbf{y}}(\theta^k) = 1, \\ & \mathbf{M}_{\mathcal{A}_j^{(s,k_i)}}(\check{g}_i\mathbf{y}) \succeq 0,\, j \in [b(n,k_i)],\, i \in [m], \end{aligned} \tag{9.2.18}$$

where $k_i := k + d_f - d_{g_i}$, $i \in [m]$. Here $\check{g}_m = 1$.

Remark 9.9. *If we assume that* $\theta = 1$ *then* (9.2.18) *becomes a moment relaxation based on Putinar's Positivstellensatz for POP* (9.0.3). *Here each constraint* $\mathbf{M}_{k_i}(\check{g}_i\mathbf{y}) \succeq 0$ *is replaced by the constraint* $\mathbf{M}_{\mathcal{A}_j^{(s,k_i)}}(\check{g}_i\mathbf{y}) \succeq 0$. *If* s *is large enough,* (9.2.18) *corresponds to an SDP relaxation obtained after exploiting term sparsity (see [212]).*

Theorem 9.4. *Let* $f, g_i \in \mathbb{R}[\mathbf{x}]$, $i \in [m]$, *with* $g_m = 1$ *and* $g_1 := R - \sum_{j\in[n]} x_j$ *for some* $R > 0$. *Consider POP* (9.2.1) *with* S *being defined as in* (9.2.2). *For every* $s \in \mathbb{N}_{>0}$ *and for every* $k \in \mathbb{N}$, *the dual of* (9.2.18) *reads as:*

$$\begin{aligned} \rho_{k,s}^{\mathrm{Pól}} := \sup_{\lambda,\mathbf{G}_{ij}} \quad & \lambda \\ \text{s.t.} \quad & \lambda \in \mathbb{R},\, \mathbf{G}_{ij} \succeq 0,\, j \in [b(n,k_i)],\, i \in [m], \\ & \theta^k(\check{f} - \lambda) = \textstyle\sum_{i\in[m]} \check{g}_i\big(\sum_{j\in[b(n,k_i)]} \mathbf{v}_{\mathcal{A}_j^{(s,k_i)}}^{\top} \mathbf{G}_{ij} \mathbf{v}_{\mathcal{A}_j^{(s,k_i)}}\big). \end{aligned} \tag{9.2.19}$$

The following statements hold:

1. *For all* $k \in \mathbb{N}$ *and for every* $s \in \mathbb{N}_{>0}$, $\rho_k^{\mathrm{Pól}} = \rho_{k,1}^{\mathrm{Pól}} \le \rho_{k,s}^{\mathrm{Pól}} \le f^\star$.
2. *For every* $s \in \mathbb{N}_{>0}$, *the sequence* $(\rho_{k,s}^{\mathrm{Pól}})_{k\in\mathbb{N}}$ *converges to* $f^\star$.
3. *If* S *has nonempty interior, there exist positive constants* $\bar{\mathfrak{c}}$ *and* $\mathfrak{c}$ *depending on* f, g_i *such that for every* $s \in \mathbb{N}_{>0}$ *and for every* $k \in \mathbb{N}$, $0 \le f^\star - \rho_{k,s}^{\mathrm{Pól}} \le \big(\frac{k}{\bar{\mathfrak{c}}}\big)^{-\frac{1}{\mathfrak{c}}}$.
4. *If* S *has nonempty interior, for every* $k \in \mathbb{N}$ *and for every* $s \in \mathbb{N}_{>0}$, *strong duality holds for the primal-dual problems* (9.2.18)-(9.2.19).

Proof. It is not hard to prove the first statement. The second and third one are due to the first statement of Theorem 9.2. The final statement is proved similarly to the third statement of [132, Theorem 4]. □

Remark 9.10. *In order to construct the semidefinite relaxation* (9.2.19), *the SOS of monomials in the linear relaxation* (9.2.6) *are replaced by a sum of several SOS polynomials associated to Gram matrices of small sizes. This idea is inspired by [215], where the authors replace the first nonnegative scalar by an SOS polynomial in the linear relaxation based on Krivine–Stengle's Positivstellensatz.*

Remark 9.11. *At fixed $s \in \mathbb{N}_{>0}$, the sequence $(\rho_{k,s}^{\text{Pól}})_{k\in\mathbb{N}}$ may not be monotonic w.r.t. k, and similarly at fixed $k \in \mathbb{N}$.*

Example 9.2. *(AM-GM inequality) Consider the case where $n = 3$, $f = x_1 + x_2 + x_3$ and $S = \{\mathbf{x} \in \mathbb{R}^3 : x_j \geq 0\,,\, j \in [3]\,,\, x_1x_2x_3 - 1 \geq 0\,,\, 3 - x_1 - x_2 - x_3 \geq 0\}$. Using AM-GM inequality, we have*

$$f(\mathbf{x}) \geq 3(x_1x_2x_3)^{1/3} \geq 3\,, \quad \forall x \in S\,, \tag{9.2.20}$$

yielding $f^\star = 3$. We solve SDP (9.2.18) with `Mosek` *and report the corresponding numerical results in Table 9.1. The table displays $\tau_{2,4}^{\text{Pól}} = 2.9999$ which is very close to $f^\star$. However, $\tau_{17}^{\text{Pól}} = \tau_{17,1}^{\text{Pól}} = 1.5030$ is smaller than $\tau_{16}^{\text{Pól}} = \tau_{16,1}^{\text{Pól}} = 2.4000$, which violates the theoretical inequality (9.2.7). The underlying reason is that the matrix A used to define the convex polytope $P = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq 0\,,\, \mathbf{Ax} \leq \mathbf{b}\}$ in the equivalent form $\min_{\mathbf{x}\in P} \mathbf{c}^\top\mathbf{x}$ of LP (9.2.5) is ill-conditioned, and the solver is not able to solve accurately the LP corresponding to $\tau_{17}^{\text{Pól}}$.*

Based on the extension of Handelman's Positivstellensatz

Consider the hierarchy of semidefinite programs indexed by $s \in \mathbb{N}_{>0}$ and $k \in \mathbb{N}$:

$$\begin{aligned} \tau_{k,s}^{\text{Han}} := \inf_{\mathbf{y}} \quad & L_{\mathbf{y}}(\check{f}) \\ \text{s.t.} \quad & \mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{2k}} \subset \mathbb{R}\,,\, y_{\mathbf{0}} = 1\,, \\ & \mathbf{M}_{\mathcal{A}_r^{(s,k_{ij})}}((\check{g}_i\check{g}_1^j)\mathbf{y}) \succeq 0\,,\, r \in [b(n,k_{ij})]\,,\, i \in [m]\,,\, j \in \{0\} \cup [k - d_{g_i}]\,, \end{aligned} \tag{9.2.21}$$

where $k_{ij} := k - d_{g_i} - j$, for $i \in [m]$, for $j \in \{0\} \cup [k - d_{g_i}]$. Note that $\check{g}_m = 1$.

Theorem 9.5. *Let $f, g_i \in \mathbb{R}[\mathbf{x}]$, $i \in [m]$, with $g_m = 1$ and $g_1 := R - \sum_{j\in[n]} x_j$ for some $R > 0$. Consider POP (9.2.1) with S being defined as in (9.2.2). For every $s \in \mathbb{N}_{>0}$ and for every $k \in \mathbb{N}$, the dual of (9.2.21) reads as:*

$$\begin{aligned} \rho_{k,s}^{\text{Han}} := \sup_{\lambda,\mathbf{G}_{ijr}} \quad & \lambda \\ \text{s.t.} \quad & \lambda \in \mathbb{R}\,,\, \mathbf{G}_{ijr} \succeq 0\,,\, j \in [b(n,k_{ij})]\,,\, i \in [m]\,,\, j \in \{0\} \cup [k - d_{g_i}]\,, \\ & \check{f} - \lambda = \textstyle\sum_{i\in[m]} \sum_{j=0}^{k-d_{g_i}} \check{g}_i\check{g}_1^j \big(\sum_{r\in[b(n,k_{ij})]} \mathbf{v}_{\mathcal{A}_r^{(s,k_{ij})}}^\top \mathbf{G}_{ijr} \mathbf{v}_{\mathcal{A}_r^{(s,k_{ij})}} \big)\,. \end{aligned} \tag{9.2.22}$$

The following statements hold:

1. *For all $k \in \mathbb{N}$ and for every $s \in \mathbb{N}_{>0}$, $\rho_k^{\text{Han}} = \rho_{k,1}^{\text{Han}} \leq \rho_{k,s}^{\text{Han}} \leq f^\star$.*
2. *For every $s \in \mathbb{N}_{>0}$, the sequence $(\rho_{k,s}^{\text{Han}})_{k\in\mathbb{N}}$ converges to $f^\star$.*
3. *If S has nonempty interior, there exist positive constants $\bar{\mathfrak{c}}$ and $\mathfrak{c}$ depending on f, g_i such that for every $s \in \mathbb{N}_{>0}$ and for every $k \in \mathbb{N}$, $0 \leq f^\star - \rho_{k,s}^{\text{Han}} \leq \big(\frac{k}{\bar{\mathfrak{c}}}\big)^{-\frac{1}{\mathfrak{c}}}$.*
4. *If S has nonempty interior, for every $k \in \mathbb{N}$ and for every $s \in \mathbb{N}_{>0}$, strong duality holds for the primal-dual problems (9.2.21)-(9.2.22).*

The proof of Theorem 9.5 is based on Theorem 9.3 and similar to the proof of Theorem 9.3.

Remark 9.12. *To make the use of the extended Handelman's Positivstellensatz, we need at least one ball constraint $\check{g}_1 := R - \|\mathbf{x}\|_2^2$ for some $R > 0$. Thus, Theorem 9.5 is applicable only when the domain S of POP (9.2.1) is compact. To deal with the noncompact case, we might combine it with the so-called "big ball trick".*

9.2.3 Obtaining an optimal solution

A real sequence $(y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_t}$ has a representing measure if there exists a finite Borel measure μ such that $y_{\boldsymbol{\alpha}} = \int_{\mathbb{R}^n} \mathbf{x}^{\boldsymbol{\alpha}} d\mu(\mathbf{x})$ is satisfied for every $\boldsymbol{\alpha} \in \mathbb{N}^n_t$.

Next, we discuss about the extraction of an optimal solution $\mathbf{x}^\star$ of POP (9.2.1) from the optimal solution $\mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{2(d_f+k)}}$ of the semidefinite relaxations (9.2.18).

Table 9.1: Numerical values (in the first subtable) and computing time (in the second subtable) for $\tau_{k,s}^{\text{Pól}}$ in Example 9.2

k \ s	1	2	3	4	5	6	7	8
0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.0000	0.0000	0.4999	**2.9999**	**2.9999**	**2.9999**	**2.9999**	**2.9999**
3	1.0000	0.9999	0.9999	2.7454	2.8368	2.8383	**2.9999**	**2.9999**
4	1.4399	1.4999	1.4999	1.4999	1.4999	1.4999	**2.9999**	**2.9999**
5	1.8615	1.9961	1.9999	1.9999	1.9999	1.9999	**2.9999**	**2.9999**
6	2.1999	2.4526	2.4998	2.4999	2.4999	2.4999	2.4999	2.4999
7	2.3971	2.8090	2.9633	2.9950	2.9996	**2.9999**	**2.9999**	**2.9999**
8	2.4109	2.9022	2.9989	**2.9999**	**2.9999**	**2.9999**	**2.9999**	**2.9999**
9	2.5161	2.9137	2.9997	**2.9999**	**2.9999**	**2.9999**	**2.9999**	**2.9999**
10	2.5896	2.9520	2.9993	**2.9999**	**2.9999**	**2.9999**	**2.9999**	**2.9999**
11	2.6210	2.9607	2.9983	**2.9999**	**2.9999**	**2.9999**	**2.9999**	**2.9999**
12	2.6937	2.9615	2.9973	2.9998	**2.9999**	**2.9999**	**2.9999**	**2.9999**
13	2.7330	2.9662	2.9977	**2.9999**	**2.9999**	**2.9999**	**2.9999**	**2.9999**
14	2.7390	2.9687	2.9974	**2.9999**	**2.9999**	**2.9999**	**2.9999**	**2.9999**
15	2.3704	2.9697	2.9972	2.9998	**2.9999**	**2.9999**	**2.9999**	**2.9999**
16	2.4000	2.9710	2.9971	2.9997	**2.9999**	**2.9999**	**2.9999**	**2.9999**
17	1.5030	2.9723	2.9968	**2.9999**	**2.9999**	**2.9999**	**2.9999**	**2.9999**
18	0.5833	2.9732	2.9966	2.9996	**2.9999**	**2.9999**	**2.9999**	**2.9999**
19	0.8121	0.0000	0.0000	2.9995	**2.9999**	**2.9999**	**2.9999**	**2.9999**
20	0.7457	0.0000	0.0000	2.9994	**2.9999**	**2.9999**	**2.9999**	**2.9999**

k \ s	1	2	3	4	5	6	7	8
0	1.1	1.3	1.0	1.0	1.0	1.1	1.0	1.0
1	1.1	1.1	1.1	1.2	1.1	1.1	1.1	1.1
2	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1
3	1.1	1.1	1.1	1.1	1.1	1.1	1.5	1.1
4	1.1	1.2	1.1	1.1	1.1	1.1	1.2	1.2
5	1.1	1.1	1.1	1.2	1.2	1.3	1.2	1.3
6	1.2	1.2	1.2	1.3	1.3	1.5	1.3	1.2
7	1.4	1.2	1.2	1.4	1.4	1.6	1.6	1.4
8	1.2	1.2	1.3	1.3	1.5	1.4	1.8	1.9
9	1.3	1.2	1.3	1.4	1.5	1.7	1.7	1.6
10	1.3	1.3	1.5	1.8	1.7	1.9	2.2	1.9
11	1.3	1.5	1.4	1.9	1.9	2.0	2.2	2.3
12	1.3	1.7	1.8	2.1	2.2	2.3	2.7	2.6
13	1.4	1.6	1.9	2.2	2.3	2.4	2.9	3.4
14	1.2	1.5	2.0	2.5	2.6	2.9	3.5	3.8
15	1.2	1.6	2.3	2.8	3.1	3.5	4.2	5.0
16	1.3	2.5	2.8	3.5	3.9	4.4	5.9	7.1
17	1.4	2.3	3.8	5.3	6.2	7.2	7.9	9.7
18	1.6	2.9	5.2	7.2	7.9	9.7	10.6	12.3
19	1.5	2.7	4.1	9.8	13.3	14.0	14.1	16.6
20	1.4	3.4	4.8	12.6	16.5	20.8	24.5	27.2

Remark 9.13. *A naive idea is to define the new sequence of moments* $\mathbf{u} = (u_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{2(d_f+k)}}$ *given by* $u_{\boldsymbol{\alpha}} := y^2_{\boldsymbol{\alpha}}$*, for* $\boldsymbol{\alpha} \in \mathbb{N}^n_{2(d_f+k)}$*. Obviously, if* $\mathbf{y}$ *has a representing Dirac measure* $\delta_{\mathbf{z}^\star}$*, then* u *has a representing Dirac measure* $\delta_{\mathbf{z}^{\star 2}}$*. In this case, we take* $\mathbf{x}^\star := \mathbf{z}^{\star 2}$*. However, there is no guarantee that* $\mathbf{u}$ *has a representing measure in general even if* $\mathbf{y}$ *has one.*

Based on Remark 4.8, we use the following heuristic extraction algorithm:

Algorithm 14 Extraction algorithm for POPs on the nonnegative orthant

Input: precision parameter $\varepsilon > 0$ and an optimal solution $(\lambda, \mathbf{G}_{ij})$ of SDP (9.2.19).
Output: an optimal solution $\mathbf{x}^\star$ of POP (9.2.1).

1: For $j \in [b(n,k_m)]$, let $\bar{\mathbf{G}}_j = (w^{(j)}_{\mathbf{pq}})_{\mathbf{p},\mathbf{q}\in\mathbb{N}^n_{k_m}}$ such that $(w^{(j)}_{\mathbf{pq}})_{\mathbf{p},\mathbf{q}\in\mathcal{A}_j^{(s,k_m)}} = \mathbf{G}_j$ and $w^{(j)}_{\mathbf{pq}} = 0$ if $(\mathbf{p},\mathbf{q}) \notin (\mathcal{A}_j^{(s,k_m)})^2$. Then $\bar{\mathbf{G}}_j \succeq 0$ and

$$\mathbf{v}^\top_{\mathbb{N}^n_{k_m}} \bar{\mathbf{G}}_j \mathbf{v}_{\mathbb{N}^n_{k_m}} = \mathbf{v}^\top_{\mathcal{A}_j^{(s,k_m)}} \mathbf{G}_j \mathbf{v}_{\mathcal{A}_j^{(s,k_m)}}\,; \tag{9.2.23}$$

2: Let $\mathbf{G} := \sum_{j\in[b(n,k_m)]} \bar{\mathbf{G}}_j$. Then $\mathbf{G}$ is the Gram matrix corresponding to σ_m in the SOS decomposition

$$\theta^k(\check{f} - \lambda) = \textstyle\sum_{i\in[m]} \check{g}_i\sigma_i\,, \tag{9.2.24}$$

where σ_i are SOS polynomials and $\check{g}_m = 1$;
3: Obtain an atom $\mathbf{z}^\star \in \mathbb{R}^n$ by using the extraction algorithm of Henrion and Lasserre in [77], where the matrix $\mathbf{V}$ in [77, (6)] is taken such that the columns of $\mathbf{V}$ form a basis of the null space $\{\mathbf{u} \in \mathbb{R}^{\omega_k} : \mathbf{G}\mathbf{u} = 0\}$;
4: Verify that $\mathbf{z}^\star$ is an approximate optimal solution of POP (9.2.3) by checking the following inequalities:

$$|\check{f}(\mathbf{z}^\star) - \lambda| \le \varepsilon\|\check{f}\|_{\max} \text{ and } \check{g}_i(\mathbf{z}^\star) \ge -\varepsilon\|\check{g}_i\|_{\max}\,,\ i \in [m]\,, \tag{9.2.25}$$

where $\|q\|_{\max} := \max_{\boldsymbol{\alpha}} |q_{\boldsymbol{\alpha}}|$ for any $q \in \mathbb{R}[\mathbf{x}]$.
5: If the inequalities (9.2.25) hold, set $\mathbf{x}^\star := \mathbf{z}^{\star 2}$.

9.3 Numerical experiments

In this section we report results of numerical experiments obtained by solving the Moment-SOS relaxations of some random and nonrandom instances of POP (9.0.1). Other results for some random instances of POP (9.0.1) can be found in Section 9.4.7. Notice that our relaxations from Section 9.2 are to deal with dense POPs while the ones from Section 9.4.6 are for POPs with correlative sparsity.

For numerical comparison purposes, recall the semidefinte relaxation based on Putinar's Positivstellensatz for solving POP (9.0.1) indexed by $k \in \mathbb{N}$:

$$\begin{aligned} \tau_k^{\mathrm{Put}} := \inf_{\mathbf{y}} \quad & L_{\mathbf{y}}(f) \\ \text{s.t.} \quad & \mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{2k}} \subset \mathbb{R}\,,\ y_{\mathbf{0}} = 1\,, \\ & \mathbf{M}_{k-\lceil g_i\rceil}(g_i\mathbf{y}) \succeq 0\,,\ i \in [\bar{m}]\,, \end{aligned} \tag{9.3.1}$$

and its sparse version:

$$\begin{aligned} \tau_k^{\mathrm{spPut}} := \inf_{\mathbf{y}} \quad & L_{\mathbf{y}}(f) \\ \text{s.t.} \quad & \mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{2k}} \subset \mathbb{R}\,,\ y_{\mathbf{0}} = 1\,, \\ & \mathbf{M}_{\mathbb{N}^{I_c}_{k-\lceil g_i\rceil}}(g_i\mathbf{y}) \succeq 0\,,\ i \in J_c\,,\ c \in [p]\,. \end{aligned} \tag{9.3.2}$$

Here $\bar{m} := m + n$ and $g_{m+j} := x_j$, $j \in [n]$. Here $g_m := 1$ and $m \in J_c \subset [\bar{m}]$, for $c \in [p]$. As shown by Baldi and Mourrain [13], the sequence $(\tau_k^{\mathrm{Put}})_{k\in\mathbb{N}}$ converges to $f^\star$ with the rate of at

Table 9.2: The notation

Pb	the ordinal number of a POP instance
Id	the ordinal number of an SDP instance
n	the number of nonnegative variables in POP (9.0.1)
m_{ineq}	the number of inequality constraints of the form $g_i \geq 0$ in POP (9.0.1)
m_{eq}	the number of equality constraints of the form $g_i = 0$ in POP (9.0.1)
Put	the SDP relaxation based on Putinar's Positivstellensatz (9.3.1) modeled by `TSSOS` and solved by `Mosek` 9.1
Pól	the SDP relaxation based on the extension of Pólya's Positivstellensatz (9.2.19) modeled by our software `InterRelax` and solved by `Mosek` 9.1
Han	the SDP relaxation based on the extension of Handelman's Positivstellensatz (9.2.22) modeled by our software `InterRelax` and solved by `Mosek` 9.1
spPut	the SDP relaxation for a sparse POP based on Putinar's Positivstellensatz (9.3.2) modeled by `TSSOS` and solved by `Mosek` 9.1
spPól	the SDP relaxation for a sparse POP based on the extension of Pólya's Positivstellensatz (9.4.81) modeled by our software `InterRelax` and solved by `Mosek` 9.1
spHan	the SDP relaxation for a sparse POP based on the extension of Handelman's Positivstellensatz (9.4.83) modeled by our software `InterRelax` and solved by `Mosek` 9.1
k	the relaxation order
s	the factor width upper bound used in SDP (9.0.7) and SDP (9.4.81)
d	the sparsity order of the SDP relaxation (9.4.81)
nmat	the number of matrix variables of an SDP
msize	the largest size of matrix variables of an SDP
nscal	the number of scalar variables of an SDP
naff	the number of affine constraints of an SDP
val	the value returned by the SDP relaxation
*	there exists at least one optimal solution of the POP, which can be extracted by Algorithm 14 or 9.4.6
time	the running time in seconds (including modeling and solving time)
∞	the SDP relaxation is unbounded or infeasible
$-$	the calculation runs out of space

least $\mathcal{O}(\varepsilon^{-\mathfrak{c}})$ when POP (9.0.1) has a ball constraint, e.g., $g_1 := R - \|\mathbf{x}\|_2^2$ for some $R > 0$. If $g_1 = R - \sum_{j\in[n]} x_j$ for some $R > 0$, then $(\tau_k^{\text{Put}})_{k\in\mathbb{N}}$ still converges to $f^\star$ due to Jacobi-Prestel [86, Theorem 4.2] (see also [8, Theorem 1 (JP)]).

Remark 9.14. *If we assume that $g_1 := R - \sum_{j\in[n]} x_j$ for some $R > 0$, SDP* (9.3.1) *may be unbounded when k is too small since its variables $\mathbf{y}$ are possibly unbounded. This issue occurs later on, see, e.g., Section 9.3.1. However, if we assume that $g_1 := R - \|\mathbf{x}\|_2^2$ for some $R > 0$, then SDP* (9.3.1) *is feasible for any order $k \geq 1$ (see Section 9.4.7).*

The experiments are performed in Julia 1.3.1. We rely on `TSSOS` [212] to solve the Moment-SOS relaxations of sparse POPs.

The implementation of our method is available online via the link:

https://github.com/maihoanganh/InterRelax.

We use a desktop computer with an Intel(R) Core(TM) i7-8665U CPU @ 1.9GHz × 8 and 31.2 GB of RAM. The notation for the numerical results is given in Table 9.2.

9.3.1 Dense QCQPs

Test problems: We construct randomly generated dense quadratically constrained quadratic programs (QCQPs) in the form (9.0.1)-(9.0.2) as follows:

Table 9.3: Numerical results for randomly generated dense QCQPs.

Id	Pb	POP size			Put			Pól				Han			
		n	m_{ineq}	m_{eq}	k	val	time	k	s	val	time	k	s	val	time
1 2	1	20	2	0	1 2	∞ **-1.99792**	0.0 92	0	17	**-1.99792***	1	2	5	**-1.99792**	1
3 4	2	20	5	0	1 2	∞ -0.350601	0.03 342	1	20	**-0.265883***	9	3	1	**-0.265883**	1
5 6	3	20	5	4	1 2	∞ -0.431543	0.02 356	1	7	-0.429442	5	3	7	**-0.429430**	9
7 8	4	30	2	0	1 2	∞ **-2.31695**	0.0 3545	0	20	**-2.31695***	2	2	10	**-2.31695**	1
9 10	5	30	7	0	1 2	∞ -2.13423	0.2 15135	0	31	**-1.79295**	45	3	20	**-1.79295**	238
11 12	6	30	7	6	1 2	∞ **-1.56374**	0.1 12480	1	31	**-1.56374**	54	3	15	**-1.56374**	236

Id	Put				Pól				Han			
	nmat	msize	nscal	naff	nmat	msize	nscal	naff	nmat	msize	nscal	naff
1 2	1 22	21 231	22 1	231 10626	5	17	232	231	17	5	255	231
3 4	1 25	21 231	25 1	231 10626	44	20	1604	1771	0	1	2344	1771
5 6	1 25	21 231	29 925	231 10626	330	7	1688	1771	345	7	1945	1771
7 8	1 32	31 496	32 1	496 46376	11	21	497	496	22	10	530	496
9 10	1 37	31 496	37 1	496 46376	32	31	5116	5456	396	20	5650	5456
11 12	1 37	31 496	43 2977	496 46376	32	31	5302	5456	561	15	5836	5456

1. Take a in the simplex

$$\Delta_n := \{\mathbf{x} \in \mathbb{R}^n \,:\, x_j \geq 0\,,\, j \in [n]\,,\, \textstyle\sum_{j\in[n]} x_j \leq 1\} \tag{9.3.3}$$

   w.r.t. the uniform distribution.

2. Let $g_1 := 1 - \sum_{j\in[n]} x_j$ and $g_2 := 1$.

3. Take every coefficient of f and g_i, $i = 2, \ldots, m$, in $(-1, 1)$ w.r.t. the uniform distribution.

4. Update $g_i(\mathbf{x}) := g_i(\mathbf{x}) - g_i(\mathbf{a}) + 0.125$, for $i = 2, \ldots, m_{\text{ineq}}$.

5. Update $g_{i+m_{\text{ineq}}}(\mathbf{x}) := g_{i+m_{\text{ineq}}}(\mathbf{x}) - g_{i+m_{\text{ineq}}}(\mathbf{a})$ and set $g_{i+m_{\text{eq}}+m_{\text{ineq}}} = -g_{i+m_{\text{ineq}}}$, for $i \in [m_{\text{eq}}]$.

Here $m = m_{\text{ineq}} + 2m_{\text{eq}}$ with m_{ineq} (resp. m_{eq}) being the number of inequality (resp. equality) constraints except the orthogonal constraints $x_j \geq 0$. If $m_{\text{ineq}} = 2$ and $m_{\text{eq}} = 0$, we obtain the case of the minimization of a polynomial on the simplex Δ_n. The point a is a feasible solution of POP (9.0.1).

The numerical results are displayed in Table 9.3.

Discussion: Table 9.3 shows that Pól and Han are typically faster and more accurate than Put. For instance, when $n = 20$, $m_{\text{ineq}} = 5$ and $m_{\text{eq}} = 0$, Put takes 342 seconds to return the lower bound -0.350601 for $f^\star$, while Pól only takes 9 seconds to return the better bound -0.265883 and

an approximate optimal solution. It is due to the fact that Pól has 44 matrix variables with the maximal matrix size 20, while Put has 25 matrix variables with the maximal matrix size 231 in this case. In addition, Han provides slightly better bounds than Pól in Pb 3 and the same bounds with Pól in the others. Moreover, Pól runs about five times faster than Han in Pb 5 and 6.

9.3.2 Sparse QCQPs

Test problems: We construct randomly generated QCQPs in the form (9.0.1)-(9.0.2) with correlative sparsity as follows:

1. Take a positive integer u, $p := \lfloor n/u \rfloor + 1$ and let

$$I_c = \begin{cases} [u], & \text{if } c = 1\,, \\ \{u(c-1), \ldots, uc\}, & \text{if } c \in \{2, \ldots, p-1\}\,, \\ \{u(p-1), \ldots, n\}, & \text{if } c = p\,; \end{cases} \tag{9.3.4}$$

2. Generate a quadratic polynomial objective function $f = \sum_{c\in[p]} f_c$ such that for each $c \in [p]$, $f_c \in \mathbb{R}[x(I_c)]_2$, and the coefficient $f_{c,\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \mathbb{N}_2^{I_c}$ of f_c is randomly generated in $(-1, 1)$ w.r.t. the uniform distribution;

3. Take a random point a such that for every $c \in [p]$, $a(I_c)$ belongs to the simplex

$$\Delta^{(c)} := \{\mathbf{x}(I_c) \in \mathbb{R}^{n_c} \,:\, x_j \geq 0\,,\, j \in I_c\,,\, \textstyle\sum_{j\in I_c} x_j \leq 1\} \tag{9.3.5}$$

4. Let $q := \lfloor m_{\text{ineq}}/p \rfloor$ and

$$J_c := \begin{cases} \{(c-1)q+1, \ldots, cq\}, & \text{if } c \in [p-1]\,, \\ \{(p-1)q+1, \ldots, l\}, & \text{if } c = p\,. \end{cases} \tag{9.3.6}$$

   For every $c \in [p]$ and every $i \in J_c$, generate a quadratic polynomial $g_i \in \mathbb{R}[\mathbf{x}(I_c)]_2$ by

   (a) for each $\boldsymbol{\alpha} \in \mathbb{N}_2^{I_c}\backslash\{\mathbf{0}\}$, taking a random coefficient $\mathbf{G}_{i,\boldsymbol{\alpha}}$ of h_i in $(-1, 1)$ w.r.t. the uniform distribution;

   (b) setting $g_{i,0} := 0.125 - \sum_{\boldsymbol{\alpha}\in\mathbb{N}_2^{I_c}\backslash\{\mathbf{0}\}} g_{j,\boldsymbol{\alpha}}\mathbf{a}^{\boldsymbol{\alpha}}$.

5. Take $g_{i_c} := 1 - \sum_{i\in I_c} x_i$, for some $i_c \in J_t$, for $c \in [p]$;

6. Let $r := \lfloor m_{\text{eq}}/p \rfloor$ and

$$W_c := \begin{cases} \{(c-1)r+1, \ldots, cr\}, & \text{if } c \in [p-1]\,, \\ \{(p-1)r+1, \ldots, l\}, & \text{if } c = p\,. \end{cases} \tag{9.3.7}$$

   For every $c \in [p]$ and every $i \in W_c$, generate a quadratic polynomial $h_i \in \mathbb{R}[\mathbf{x}(I_c)]_2$ by

   (a) for each $\boldsymbol{\alpha} \in \mathbb{N}_2^{I_c}\backslash\{\mathbf{0}\}$, taking a random coefficient $h_{i,\boldsymbol{\alpha}}$ of h_i in $(-1, 1)$ w.r.t. the uniform distribution;

   (b) setting $h_{i,\mathbf{0}} := -\sum_{\boldsymbol{\alpha}\in\mathbb{N}_2^{I_c}\backslash\{\mathbf{0}\}} h_{i,\boldsymbol{\alpha}}\mathbf{a}^{\boldsymbol{\alpha}}$.

7. Take $g_{i+m_{\text{ineq}}}(\mathbf{x}) := h_i$ and set $g_{i+m_{\text{eq}}+m_{\text{ineq}}} = -h_i$, for $i \in [m_{\text{eq}}]$.

Here $m = m_{\text{ineq}} + 2m_{\text{eq}}$ with m_{ineq} (resp. m_{eq}) being the number of inequality (resp. equality) constraints except the orthogonal constraints $x_j \geq 0$. The point $\mathbf{a}$ is a feasible solution of POP (9.0.1).

The numerical results are displayed in Table 9.4.

Table 9.4: Numerical results for randomly generated QCQPs with correlative sparsity of $n = 1000$ and $d = \deg(f) = 2$.

Id	Pb	POP size			spPut			spPól				spHan			
		u	m_{ineq}	m_{eq}	k	val	time	k	s	val	time	k	s	val	time
1 2	1	10	201	0	1 2	∞ -129.061	1.5 385	0	10	-128.906	15	2	7	**-128.660**	20
3 4	2	10	201	200	1 2	∞ -66.0696	2.0 475	1	12	-65.3195	51	3	10	**-65.3050**	283
5 6	3	20	201	0	1 2	∞ -66.1306	3.6 56360	0	15	-65.9794	19	2	15	**-65.8646**	24
7 8	4	20	201	200	1 2	∞ –	9 –	1	22	-38.2061	319	3	20	**-38.2035**	2146

Id	spPut				spPól				spHan			
	nmat	msize	nscal	naff	nmat	msize	nscal	naff	nmat	msize	nscal	naff
1 2	100 1300	12 78	1201 1	7491 135641	299	10	7889	7491	599	7	9288	7491
3 4	100 1300	12 78	1401 15577	7491 135641	1299	12	39920	43813	4184	10	41419	35926
5 6	50 1250	22 253	1201 1	12481 630231	399	15	25407	25109	399	15	13978	12481
7 8	50 1250	22 253	1401 50513	12481 630231	1149	22	108641	113428	3574	20	109990	100751

Discussion: Similarly to the previous discussion, spPól and spHan in Table 9.4 are also much faster and more accurate than spPut. For instance, when $u = 20$, $m_{\text{ineq}} = 201$ and $m_{\text{eq}} = 0$, spPól takes 20 seconds to return the lower bound -65.9794 of $f^\star$, while spPut takes 56360 seconds to return a worse bound of -66.1306. In this case, spPól has 399 matrix variables with maximal matrix size 15, while spPut has 1250 matrix variables with maximal matrix size 253. In particular, spHan provides slightly better bounds than spPól for Pb 1, 2, 4 while it is seven (resp. five) times slower than spPól in Pb 4 (resp. Pb 2).

9.3.3 Stability number of a graph

In order to compute the stability number $\alpha(G)$ of a given graph G, we solve the following POP on the unit simplex:

$$\tfrac{1}{\alpha(G)} = \min_{\mathbf{x}\in\mathbb{R}^n_+} \{\mathbf{x}^\top(\mathbf{A}+\mathbf{I})\mathbf{x} : \textstyle\sum_{j\in[n]} x_j = 1\}, \tag{9.3.8}$$

where $\mathbf{A}$ is the adjacency matrix of G and $\mathbf{I}$ is the identity matrix.

Test problems: We take some adjacency matrices of known graphs from [180]. The numerical results are displayed in Tables 9.5 and 9.6. Note that in Table 9.6 we solve POP (9.3.8) with an additional unit ball constraint $1 - \|\mathbf{x}\|_2^2 \geq 0$. The columns under "val" show the approximations of $\alpha(G)$.

Discussion: The graphs from Table 9.5 are relatively dense so that we cannot exploit term sparsity or correlative sparsity for POP (9.3.8) in these cases. For the graph GD02_a in Table 9.5, Pól and Han provide better bounds for $\alpha(G)$ compared to the ones returned by the second order relaxations of Put. In Table 9.6, Put provides negative values for the first order relaxations. The additional unit ball constraint does not help to improve the bound for the second order relaxation for Id 2. Besides, Table 9.5 shows that Han provides slightly better bounds than Pól for johnson16-2-4, but its value is less accurate than the corresponding one from Table 9.6.

Table 9.5: Numerical results for stability number of some known graphs in [180].

Id	Pb	POP size	Put			Pól				Han			
		n	k	val	time	k	s	val	time	k	s	val	time
1 2	GD02_a	23	1 2	∞ 13.0110	0.02 394	0	25	**13.0000**	1	2	25	**13.0000**	1
3 4	johnson8-2-4	28	1 2	∞ **7.00000**	0.03 2098	0	30	**7.00000**	1	2	30	**6.99999**	1
5 6	johnson8-4-4	70	1 2	∞ –	1 –	0	72	**5.00000**	5	2	72	**5.00001**	8
7 8	hamming6-2	64	1 2	∞ –	0.5 –	0	66	**1.99999**	3	2	66	**1.99999**	6
9 10	hamming6-4	64	1 2	∞ –	0.6 –	0	66	**12.0000**	3	2	66	**12.0000**	5
11 12	johnson16-2-4	120	1 2	∞ –	0.6 –	0	122	**15.0001**	54	2	122	**15.0000**	78

Id	Put				Pól				Han			
	nmat	msize	nscal	naff	nmat	msize	nscal	naff	nmat	msize	nscal	naff
1 2	1 24	24 300	25 301	300 17550	1	24	301	300	1	24	326	300
3 4	1 29	29 435	30 436	435 35960	1	29	436	435	1	29	466	435
5 6	1 71	71 2556	72 2557	2556 1150626	1	71	2557	2556	1	71	2629	2556
7 8	1 65	65 2145	66 2146	2145 814385	1	65	2146	2145	1	65	2212	2145
9 10	1 65	65 2145	66 2146	2145 814385	1	65	2146	2145	1	65	2212	2145
11 12	1 121	121 7381	122 7380	7381 9381251	1	121	7382	7381	1	121	7504	7381

Table 9.6: Numerical results for stability number of some known graphs in [180] with an additional unit ball constraint.

Id	Pb	POP size	Put			Pól				Han			
		n	k	val	time	k	s	val	time	k	s	val	time
1 2	GD02_a	23	1 2	-0.62896 13.0170	0.02 442	0	13	**13.0000**	1	2	13	**13.0000**	1
3 4	johnson8-2-4	28	1 2	-0.30434 **7.00000**	0.03 3010	0	23	**7.00000**	1	2	23	**7.00000**	1
5 6	johnson8-4-4	70	1 2	-0.14056 –	1 –	0	70	**5.00000**	10	2	70	**5.00000**	8
7 8	hamming6-2	64	1 2	-0.32989 –	1 –	0	64	**2.00000**	7	2	64	**2.00000**	7
9 10	hamming6-4	64	1 2	-0.11764 –	0.6 –	0	64	**12.0000**	6	2	64	**12.0000**	7
11 12	johnson16-2-4	120	1 2	-0.08982 –	26 –	0	121	**15.0000**	75	2	121	15.0026	74

Id	Put				Pól				Han			
	nmat	msize	nscal	naff	nmat	msize	nscal	naff	nmat	msize	nscal	naff
1 2	1 25	24 300	26 301	300 17550	12	13	302	300	12	13	327	300
3 4	1 30	29 435	31 436	435 35960	7	23	437	435	7	23	467	435
5 6	1 72	71 2556	73 2557	2556 1150626	2	70	2558	2556	2	70	2630	2556
7 8	1 66	65 2145	67 2146	2145 814385	2	64	2146	2145	2	64	2213	2145
9 10	1 66	65 2145	67 2146	2145 814385	2	64	2146	2145	2	64	2213	2145
11 12	1 122	121 7381	123 7380	7381 9381251	1	121	7383	7381	1	121	7505	7381

Table 9.7: Numerical results for some instances of MAXCUT problems.

Id	Pb	POP size	Put			Pól				Han			
		n	k	val	time	k	s	val	time	k	s	val	time
1 2	burma14	14	1 2	30310.915 **30301.999**	0.2 4	1	16	**30302.000**	1	3	16	**30301.999**	1
3 4	gr17	17	1 2	25089.044 **24985.999**	0.2 24	1	19	**24986.000**	1	3	19	**24985.999**	2
5 6	fri26	26	1 2	22220.657 **22217.999**	0.4 1970	1	28	**22218.000**	12	3	28	**22217.999**	28
7 8	att48	48	1 2	799281.420 –	1 –	1	50	**798857.049**	1129	3	50	798890.722	3600

Id	Put				Pól				Han			
	nmat	msize	nscal	naff	nmat	msize	nscal	naff	nmat	msize	nscal	naff
1 2	1 15	15 120	29 1681	130 3060	15	15	666	680	16	15	787	680
3 4	1 18	18 171	35 2908	171 5985	18	18	1123	1140	19	18	1295	1140
5 6	1 27	27 378	53 9829	378 27405	27	27	3628	3654	28	27	4007	3654
7 8	1 30	49 465	97 13486	1225 40920	49	49	20777	20825	50	49	22003	20825

9.3.4 The MAXCUT problems

The MAXCUT problem is given by:

$$\max_{\mathbf{x}\in\{0,1\}^n} \mathbf{x}^\top \mathbf{W}(\mathbf{e}-\mathbf{x})\,, \tag{9.3.9}$$

where $\mathbf{e} = (1,\ldots,1)$ and W is the matrix of edge weights associated with a graph (see [38, Theorem 1]).

Test problems: The data of graphs is taken from TSPLIB [173].

The numerical results are displayed in Table 9.7. Note that all instance of matrix W are dense.

Discussion: The behavior of our method is similar to that in Section 9.3.1.

9.3.5 Robustness certification of deep neural networks

In [172], the robustness certification problem of a multi-layer neural network with ReLU activation function is formulated as the following QCQP:

$$\begin{aligned} l_y^\star(\bar{\mathbf{x}},\bar{y}) := \max_{\mathbf{x}^0,\ldots,\mathbf{x}^L} \quad & (\mathbf{c}_y - \mathbf{c}_{\bar{y}})^\top \mathbf{x}^L \\ \text{s.t.} \quad & x_t^i(x_t^i - \mathbf{W}_t^{i-1}\mathbf{x}^{i-1}) = 0\,,\ x_t^i \geq 0\,,\ x_t^i \geq \mathbf{W}_t^{i-1}\mathbf{x}^{i-1}\,, \\ & \qquad\qquad t\in[m_i]\,,\ i\in[L] \\ & -\varepsilon \leq x_t^0 - \bar{x}_t \leq \varepsilon\,,\ t\in[m_0]\,, \end{aligned} \tag{9.3.10}$$

where we use the same notation as in [172, Section 2] and write $\mathbf{W}^{i-1} = \begin{bmatrix} \mathbf{W}_1^{i-1} \\ \ldots \\ \mathbf{W}_{m_i}^{i-1} \end{bmatrix}$.

We say that the network is certifiably ε-robust on $(\bar{\mathbf{x}},\bar{y})$ if $l_y^\star(\bar{\mathbf{x}},\bar{y}) < 0$ for all $y \neq \bar{y}$.

Table 9.8: Information for the training model (9.3.11).

Dataset	BHPD
Number of hidden layers	$L = 2$
Length of an input	13
Number of inputs	506
Test size	20%
Number of classes	$k = 3$
Numbers of units in layers	$m = (13, 20, 10)$
Number of weights	490
Opimization method	Adadelta algorithm[2]
Accuracy	70%
Batch size	128
Epochs	200

Table 9.9: Numerical results for robustness certification on BHPD, $n = 43$, $m_{\text{ineq}} = 43$, $m_{\text{eq}} = 30$ and $d = \deg(f) = 2$.

Id	Pb	spPut			spPól				spHan			
		k	val	time	k	s	val	time	k	s	val	time
1 2	$y = 1$	1 2	88.1571 –	0.4 –	1	35	**-11.8706**	625	3	35	–	1364
3 4	$y = 2$	1 2	208.934 –	0.4 –	1	35	**-13.3240**	518	3	35	–	1270

Id	spPut				spPól				spHan			
	nmat	msize	nscal	naff	nmat	msize	nscal	naff	nmat	msize	nscal	naff
1,3 2,4	23 97	22 595	117 14431	737 86285	297	35	46233	28195	308	35	47629	9670

Test problems: To obtain an instance of weights $\mathbf{W}^i$, we train a classification model by using Keras[1]. Explicitly, we minimize a loss function as follows:

$$\min_{\mathbf{W}^0,\ldots,\mathbf{W}^{L-1}} \quad \tfrac{1}{2}\sum_{(\mathbf{x}^0,y^0)\in\mathcal{D}} \|f(\mathbf{x}^0) - \mathbf{e}_{y^0}\|_2^2\,, \tag{9.3.11}$$

where the network f is defined as in [172, Section 2] and $\mathbf{e}_{y^0}$ has 1 at the y^0-th element and zeros at the others. Here the input set $\mathcal{D}$ is a part of Boston House Price Dataset (BHPD). The class label y^0 is assigned to the input $\mathbf{x}^0$. We classify the inputs from BHPD into 3 classes according to the MEDian Value of owner-occupied homes (MEDV) in \$1000 as follows:

$$y^0 = \begin{cases} 1 & \text{if MEDV} < 10\,, \\ 2 & \text{if } 10 \le \text{MEDV} < 20\,, \\ 3 & \text{otherwise}\,. \end{cases} \tag{9.3.12}$$

We also take a clean input label pair $(\bar{\mathbf{x}}, \bar{y}) \notin \mathcal{D}$ with $\bar{y} = 3$ from BHPD.

As shown in [35, Section 4.2], POP (9.3.10) has correlative sparsity. To use our method, we convert (9.3.10) to a POP on the nonnegative orthant by defining new nonnegative variables $\bar{z}_t := x_t^0 - \bar{x}_t + \varepsilon$. Here we choose $\varepsilon = 0.1$. More detailed information for our training model are available in Table 9.8.

The numerical results are displayed in Table 9.9.

Discussion: Compared to spPut, spPól and spHan provide better upper bounds in less total time. Moreover, in Table 9.9, the values returned by spPut with $k = 1$ are positive and are much

[1] `https://keras.io/api/models/model_training_apis/`

larger than the negative ones returned by spPut with $k = 2$. Since in Table 9.9, the upper bounds on $l_y^\star(\bar{\mathbf{x}}, \bar{y})$ are negative, for all $y \neq \bar{y}$, $l_y^\star(\bar{\mathbf{x}}, \bar{y})$ must be negative. Thus, we conclude that this network is certifiably ε-robust on $(\bar{\mathbf{x}}, \bar{y})$.

9.4 Appendix

9.4.1 Preliminary material

For each $q = \sum_{\boldsymbol{\alpha}} q_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}} \in \mathbb{R}[\mathbf{x}]$, we note $\|q\| := \max_{\boldsymbol{\alpha}} \frac{|q_{\boldsymbol{\alpha}}|}{c_{\boldsymbol{\alpha}}}$ with $c_{\boldsymbol{\alpha}} := \frac{|\boldsymbol{\alpha}|!}{\alpha_1! \ldots \alpha_n!}$ for each $\alpha \in \mathbb{N}^n$. This defines a norm on the real vector space $\mathbb{R}[\mathbf{x}]$. Moreover, for $p_1, q_2 \in \mathbb{R}[\mathbf{x}]$, we have

$$\|q_1 q_2\| \leq \|q_1\| \|q_2\|\,, \tag{9.4.1}$$

according to [188, Lemma 8].

We recall the following bound for central binomial coefficient stated in [94, page 590]:

Lemma 9.1. *For all $t \in \mathbb{N}_{>0}$, it holds that $\binom{2t}{t} \frac{1}{2^{2t}} \leq \frac{1}{\sqrt{\pi t}}$.*

Define the simplex

$$\Delta_n := \{\mathbf{x} \in \mathbb{R}^n \,:\, x_j \geq 0\,,\, j \in [n]\,,\, \textstyle\sum_{j\in[n]} x_j = 1\}\,. \tag{9.4.2}$$

We recall the degree bound for Pólya's Positivstellensatz [165]:

Lemma 9.2. *(Powers and Reznick [167]) If q is a homogeneous polynomial of degree d positive on Δ_n, then for all $k \in \mathbb{N}$ satisfying*

$$k \geq \frac{d(d-1)\|q\|}{2\min_{\mathbf{x}\in\Delta_n} q(\mathbf{x})} - d\,, \tag{9.4.3}$$

$(\sum_{j\in[n]} x_j)^k q$ has positive coefficients.

Let us recall the concept and the properties of polynomials even in each variable in [182, Definition 3.3]. A polynomial q is even in each variable if for every $j \in [n]$,

$$q(x_1, \ldots, x_{j-1}, -x_j, x_{j+1}, \ldots, x_n) = q(x_1, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_n)\,. \tag{9.4.4}$$

If q is even in each variable, then there exists a polynomial $\tilde{q}$ such that $q = \tilde{q}(x_1^2, \ldots, x_n^2)$. Indeed, let $q = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^n} q_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}}$ be a polynomial even in each variable. Let $j \in [n]$ be fixed. Then $q(\mathbf{x}) = \frac{1}{2}(q(\mathbf{x}) + q(x_1, \ldots, x_{j-1}, -x_j, x_{j+1}, \ldots, x_n))$. It implies that $q_{\boldsymbol{\alpha}} = 0$ if α_j is odd. Thus, $q = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^n} q_{2\boldsymbol{\alpha}} \mathbf{x}^{2\boldsymbol{\alpha}}$ since j is arbitrary in $[n]$. This yields $\tilde{q} = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^n} q_{2\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}}$.

For convenience, we denote $\mathbf{x}^2 := (x_1^2, \ldots, x_n^2)$. Moreover, if q is even in each variable and homogeneous of degree $2d_q$, then $\tilde{q}$ is homogeneous of degree d_q. Conversely, if q is a polynomial of degree at most $2d$ such that q is even in each variable, the degree-$2d$ homogenization of q is even in each variable.

9.4.2 The proof of Theorem 9.1

Proof. Let $\varepsilon > 0$. By assumption, $\deg(f) = 2d_f$, $\deg(g_i) = 2d_{g_i}$ for some $d_f, d_{g_i} \in \mathbb{N}$, for $j \in [m]$.

Step 1: Converting to polynomials on the nonnegative orthant. We claim that $\tilde{f}$ is nonnegative on the semialgebraic set

$$\tilde{S} := \{\mathbf{x} \in \mathbb{R}^n \,:\, x_j \geq 0\,,\, j \in [n]\,,\, \tilde{g}_i(\mathbf{x}) \geq 0\,,\, i \in [m]\}\,. \tag{9.4.5}$$

Let $y \in \tilde{S}$. Set $\mathbf{z} = (\sqrt{y_1}, \ldots, \sqrt{y_n})$. Then $g_i(\mathbf{z}) = \tilde{g}_i(\mathbf{z}^2) = \tilde{g}_i(\mathbf{y}) \geq 0$, for $i \in [m]$. By assumption, $\tilde{f}(\mathbf{y}) = \tilde{f}(\mathbf{z}^2) = f(\mathbf{z}) \geq 0$. It implies that $\tilde{f} + \varepsilon(\sum_{j=1}^n x_j)^{d_f}$ is homogeneous and positive on $\tilde{S}\backslash\{\mathbf{0}\}$.

To prove the first statement, we proceed exactly as in the proof of [45, Theorem 2.4] for $\tilde{f} + \varepsilon(\sum_{j=1}^n x_j)^{d_f}$ and derive the bound on the degree of polynomials having positive coefficients

when apply Pólya's Positivstellensatz. To obtain (9.1.3), we replace $\mathbf{x}$ by $\mathbf{x}^2$ in the representation of $\tilde{f} + \varepsilon(\sum_{j=1}^n x_j)^{d_f}$.

We shall prove the second statement. Assume that S has nonempty interior. Set $\bar{m} := m + n$ and $g_{m+j} := x_j^2$ with $d_{g_{m+j}} := 1$, $j \in [n]$. Then $\tilde{g}_{m+j} := x_j$, $j \in [n]$, and

$$\tilde{S} := \{\mathbf{x} \in \mathbb{R}^n \,:\, \tilde{g}_i(\mathbf{x}) \geq 0\,,\, i \in [\bar{m}]\}\,. \tag{9.4.6}$$

Note that $\deg(\tilde{g}_i) = d_{g_i}$, $i \in [\bar{m}]$. Since S has nonempty interior and $\cup_{j=1}^n \{\mathbf{x} \in \mathbb{R}^n \,:\, x_j = 0\}$ has zero Lebesgue measure in $\mathbb{R}^n$, $S \backslash (\cup_{j=1}^n \{\mathbf{x} \in \mathbb{R}^n \,:\, x_j = 0\})$ also has nonempty interior. Then there exists $\mathbf{a} \in S \backslash (\cup_{j=1}^n \{\mathbf{x} \in \mathbb{R}^n \,:\, x_j = 0\})$ such that $g_i(\mathbf{a}) > 0$, $i \in [m]$. Let $\mathbf{b} = (\sqrt{|a_1|}, \ldots, \sqrt{|a_n|})$. Then $\mathbf{b} \in (0, \infty)^n$ and $\mathbf{b}^2 = (|a_1|, \ldots, |a_n|)$. Since each g_i is even in each variable, $\tilde{g}_i(\mathbf{b}) = g_i(\mathbf{b}^2) = g_i(\mathbf{a}) > 0$, $i \in [m]$, yielding $\tilde{S}$ has nonempty interior.

Step 2: Construction of the positive weight functions. We process similarly to the proof of [132, Theorem 1] (see [132, Appendix A.2.1]) to obtain functions $\bar{\varphi}_j : \mathbb{R}^n \to \mathbb{R}$, $j \in [\bar{m}]$, such that,

1. $\bar{\varphi}_j$ is positive and bounded from above by $C_{\bar{\varphi}_j} = \bar{r}_j \varepsilon^{-r_j}$ on $B(\mathbf{0}, \sqrt{n} + j)$ for some positive constants $\bar{r}_j$ and r_j independent of ε.
2. $\bar{\varphi}_j$ is Lipschitz with Lipschitz constant $L_{\bar{\varphi}_j} = \bar{t}_j \varepsilon^{-t_j}$ for some positive constants $\bar{t}_j$ and t_j independent of ε.
3. The inequality
$$\tilde{f} + \varepsilon - \sum_{i=1}^{\bar{m}} \bar{\varphi}_i^2 \tilde{g}_i \geq \frac{\varepsilon}{2^{\bar{m}}} \text{ on } [-1,1]^n\,, \tag{9.4.7}$$
holds.

Note that we do not need to prove the even property for each weight $\bar{\varphi}_i$ above.

Step 3: Approximating with Bernstein polynomials. For each $i \in [\bar{m}]$, we now approximate $\bar{\varphi}_i$ on $[-1,1]^n$ with the following Bernstein polynomials defined as in [132, Definition 1]:

$$B_i^{(d)}(\mathbf{x}) = B_{\mathbf{y} \mapsto \bar{\varphi}_i(2\mathbf{y} - \mathbf{e}), d\mathbf{e}}\left(\frac{\mathbf{x} + \mathbf{e}}{2}\right)\,, \quad d \in \mathbb{N}\,, \tag{9.4.8}$$

with $\mathbf{e} = (1, \ldots, 1) \in \mathbb{R}^n$. By using [132, Lemma 6], for all $\mathbf{x} \in [-1,1]^n$, for $i \in [\bar{m}]$,

$$|B_i^{(d)}(\mathbf{x}) - \bar{\varphi}_i(\mathbf{x})| \leq L_{\bar{\varphi}_i}\left(\frac{n}{d}\right)^{\frac{1}{2}}\,, \quad d \in \mathbb{N}\,, \tag{9.4.9}$$

and for all $\mathbf{x} \in [-1,1]^n$, for $i \in [\bar{m}]$:

$$|B_i^{(d)}(\mathbf{x})| \leq \sup_{\mathbf{x} \in [-1,1]^n} |\bar{\varphi}_i(\mathbf{x})| \leq C_{\bar{\varphi}_i}\,. \tag{9.4.10}$$

For $i \in [\bar{m}]$, let

$$d_i := 2u_i \quad \text{with} \quad u_i = \left\lceil \frac{2C_{\tilde{g}_i}^2 C_{\bar{\varphi}_i}^2 n L_{\bar{\varphi}_i}^2 (\bar{m}+1)^2 2^{2\bar{m}}}{\varepsilon^2} \right\rceil, \tag{9.4.11}$$

where $C_{\tilde{g}_i}$ is an upper bound of $|\tilde{g}_i|$ on $B(\mathbf{0}, \sqrt{n} + i)$. Set $q_i := B_i^{(d_i)}$, $i \in [\bar{m}]$. Then for all $\mathbf{x} \in [-1,1]^n$,

$$\begin{aligned} |q_i(\mathbf{x}) - \bar{\varphi}_i(\mathbf{x})| &= |B_i^{(d_i)}(\mathbf{x}) - \bar{\varphi}_i(\mathbf{x})| \\ &\leq L_{\bar{\varphi}_i}\left(\tfrac{n}{d_i}\right)^{\frac{1}{2}} \\ &\leq L_{\bar{\varphi}_i}\left(\frac{n}{\frac{4C_{\tilde{g}_i}^2 C_{\bar{\varphi}_i}^2 n L_{\bar{\varphi}_i}^2 (\bar{m}+1)^2 2^{2\bar{m}}}{\varepsilon^2}}\right)^{\frac{1}{2}} \\ &= \frac{\varepsilon}{2C_{\tilde{g}_i} C_{\bar{\varphi}_i} (\bar{m}+1) 2^{\bar{m}}}\,. \end{aligned} \tag{9.4.12}$$

Step 4: Estimating the lower and upper bounds of $\tilde{f}(\mathbf{x}) + \varepsilon - \sum_{i=1}^{\bar{m}} q_i(\mathbf{x})^2 \tilde{g}_i(\mathbf{x})$ on Δ_n. From these and (9.4.7), for all $\mathbf{x} \in \Delta_n$,

$$\begin{array}{rl} & \tilde{f}(\mathbf{x}) + \varepsilon - \sum_{i=1}^{\bar{m}} q_i(\mathbf{x})^2 \tilde{g}_i(\mathbf{x}) \\ = & \tilde{f}(\mathbf{x}) + \varepsilon - \sum_{i=1}^{\bar{m}} \bar{\varphi}_i(\mathbf{x})^2 \tilde{g}_i(\mathbf{x}) + \sum_{i=1}^{\bar{m}} \tilde{g}_i(\mathbf{x})[\bar{\varphi}_i(\mathbf{x})^2 - q_i(\mathbf{x})^2] \\ \geq & \frac{\varepsilon}{2^{\bar{m}}} - \sum_{i=1}^{\bar{m}} |\tilde{g}_i(\mathbf{x})||\bar{\varphi}_i(\mathbf{x}) + q_i(\mathbf{x})||\bar{\varphi}_i(\mathbf{x}) - q_i(\mathbf{x})| \\ \geq & \frac{\varepsilon}{2^{\bar{m}}} - \sum_{i=1}^{\bar{m}} C_{\tilde{g}_i}(|\bar{\varphi}_i(\mathbf{x})| + |q_i(\mathbf{x})|) \frac{\varepsilon}{2C_{\tilde{g}_i} C_{\bar{\varphi}_i}(\bar{m}+1)2^{\bar{m}}} \\ \geq & \frac{\varepsilon}{2^{\bar{m}}} - \sum_{i=1}^{\bar{m}} 2C_{\tilde{g}_i} C_{\bar{\varphi}_i} \frac{\varepsilon}{2C_{\tilde{g}_i} C_{\bar{\varphi}_i}(\bar{m}+1)2^{\bar{m}}} \\ = & \frac{\varepsilon}{2^{\bar{m}}} - \frac{\bar{m}\varepsilon}{(\bar{m}+1)2^{\bar{m}}} = \frac{\varepsilon}{(\bar{m}+1)2^{\bar{m}}} . \end{array} \tag{9.4.13}$$

Thus,

$$\tilde{f} + \varepsilon - \textstyle\sum_{i=1}^{\bar{m}} q_i^2 \tilde{g}_i \geq \frac{\varepsilon}{(\bar{m}+1)2^{\bar{m}}} \text{ on } \Delta_n . \tag{9.4.14}$$

Step 5: Estimating the upper bound of $\|q_i\|$. For $i \in [\bar{m}]$, we write

$$\begin{array}{rl} q_i = B_i^{(2u_i)} = & \sum_{k_1=0}^{2u_i} \cdots \sum_{k_n=0}^{2u_i} \bar{\varphi}_i \left(\frac{k_1-u_i}{u_i}, \ldots, \frac{k_n-u_i}{u_i} \right) \\ & \times \prod_{j=1}^{n} \left[\binom{2u_i}{k_j} \left(\frac{x_j+1}{2} \right)^{k_j} \left(\frac{1-x_j}{2} \right)^{2u_i-k_j} \right] . \end{array} \tag{9.4.15}$$

Then

$$\deg(q_i) \leq 2nu_i , \tag{9.4.16}$$

for $i \in [\bar{m}]$. From (9.4.1), we have

$$\begin{array}{rl} \|q_i\| \leq & \sum_{k_1=0}^{2u_i} \cdots \sum_{k_n=0}^{2u_i} \left| \bar{\varphi}_i \left(\frac{k_1-u_i}{u_i}, \ldots, \frac{k_n-u_i}{u_i} \right) \right| \\ & \times \prod_{j=1}^{n} \left[\binom{2u_i}{k_j} \frac{1}{2^{2u_i}} \|x_j + 1\|^{k_j} \|1 - x_j\|^{2u_i-k_j} \right] \\ \leq & \sum_{k_1=0}^{2u_i} \cdots \sum_{k_n=0}^{2u_i} C_{\bar{\varphi}_i} \prod_{j=1}^{n} \left(\binom{2u_i}{u_i} \frac{1}{2^{2u_i}} \right) \\ = & C_{\bar{\varphi}_i} \left(\binom{2u_i}{u_i} \frac{2u_i+1}{2^{2u_i}} \right)^n \\ \leq & C_{\bar{\varphi}_i} \left(\frac{2u_i+1}{\sqrt{\pi u_i}} \right)^n =: T_{q_i} . \end{array} \tag{9.4.17}$$

The second inequality is due to $\|x_j + 1\| = \|1 - x_j\| = 1$ and $\binom{2u_i}{u_i} \geq \binom{2u_i}{k_j}$, for $k_j = 0, \ldots, 2u_i$. The third inequality is implied from Lemma 9.1.

Step 6: Converting to homogeneous polynomials. Thanks to (9.4.14), we get

$$\tilde{f} + 2\varepsilon - \textstyle\sum_{i \in [\bar{m}]} (q_i^2 + \frac{\varepsilon}{\bar{m} C_{\tilde{g}_i}}) \tilde{g}_i \geq \frac{\varepsilon}{(\bar{m}+1)2^{\bar{m}}} \quad \text{on } \Delta_n , \tag{9.4.18}$$

since $|\tilde{g}_i| \leq C_{\tilde{g}_i}$ on Δ_n. Note that $\tilde{f}, \tilde{g}_i$ are homogeneous polynomials of degree d_f, d_{g_i}, respectively.

For each $q \in \mathbb{R}[\mathbf{x}]_d$, $\hat{q}$ is a d-homogenization of q if

$$\hat{q} = \textstyle\sum_{t=0}^{d} h^{(t)} (\sum_{j \in [n]} x_j)^{d-t} , \tag{9.4.19}$$

for some $h^{(t)}$ is the homogeneous polynomial of degree t satisfying $q = \sum_{t=0}^{d} h^{(t)}$. In this case, $\hat{q} = q$ on Δ_n.

Let $p_i := \hat{q}_i^2 + \frac{\varepsilon}{\bar{m} C_{\tilde{g}_i}} (\sum_{j \in [n]} x_j)^{4nu_i}$ with $\hat{q}_i$ being a $2nu_i$-homogenization of q_i, for $i \in [\bar{m}]$. Then p_i is a homogeneous polynomial of degree $4nu_i$,

$$p_i = q_i^2 + \tfrac{\varepsilon}{\bar{m} C_{\tilde{g}_i}} \geq \tfrac{\varepsilon}{\bar{m} C_{\tilde{g}_i}} \text{ on } \Delta_n , \tag{9.4.20}$$

and

$$\|p_i\| \leq \|q_i\|^2 + \tfrac{\varepsilon}{\bar{m} C_{\tilde{g}_i}} \leq T_{q_i}^2 + \tfrac{\varepsilon}{\bar{m} C_{\tilde{g}_i}} =: T_{p_i} . \tag{9.4.21}$$

Set $D := \max\{d_f, 4nu_i + d_{g_i} \,:\, i \in [\bar{m}]\}$ and

$$\begin{array}{rl} F := & (\sum_{j \in [n]} x_j)^{D-d_f} (\tilde{f} + 2\varepsilon (\sum_{j \in [n]} x_j)^{d_f}) \\ & - \sum_{i \in [\bar{m}]} \tilde{g}_i p_i (\sum_{j \in [n]} x_j)^{D-4nu_i-d_{g_i}} . \end{array} \tag{9.4.22}$$

Then F is a homogeneous polynomial of degree D and

$$F = \tilde{f} + 2\varepsilon - \textstyle\sum_{i\in[\bar{m}]}(q_i^2 + \frac{\varepsilon}{\bar{m}C_{\tilde{g}_i}})\tilde{g}_i \geq \frac{\varepsilon}{(\bar{m}+1)2^{\bar{m}}} \quad \text{on } \Delta_n\,, \tag{9.4.23}$$

Moreover,

$$\begin{array}{rl} \|F\| \leq & \|\sum_{j\in[n]} x_j\|^{D-d_f}(\|\tilde{f}\| + 2\varepsilon\|\sum_{j\in[n]} x_j\|^{d_f}) \\ & + \sum_{i\in[\bar{m}]} \|\tilde{g}_i\|\|p_i\|\|\sum_{j\in[n]} x_j\|^{D-4nu_i-d_{g_i}} \\ \leq & \|\tilde{f}\| + 2\varepsilon + \sum_{i\in[\bar{m}]} T_{p_i}\|\tilde{g}_i\| =: T_F\,, \end{array} \tag{9.4.24}$$

since $\|\sum_{j\in[n]} x_j\| = 1$.

Step 7: Applying the degree bound of Pólya's Positivstellensatz. Using Lemma 9.2, we obtain:

- For all $k \in \mathbb{N}$ satisfying
$$k \geq \frac{D(D-1)T_F}{\frac{\varepsilon}{(\bar{m}+1)2^{\bar{m}}}} =: K_0\,, \tag{9.4.25}$$
$(\sum_{j\in[n]} x_j)^k F$ has positive coefficients.
- For each $i \in [\bar{m}]$ and for all $k \in \mathbb{N}$ satisfying
$$k \geq \frac{4nu_i(4nu_i-1)T_{p_i}}{\frac{\varepsilon}{\bar{m}C_{\tilde{g}_i}}} =: K_i\,, \tag{9.4.26}$$
$(\sum_{j\in[n]} x_j)^k p_i$ has positive coefficients.

Notice that K_i, $i = 0,\ldots,\bar{m}$, are obtained by composing finitely many times the following operators: "$+$", "$-$", "$\times$", "$\div$", "$|\cdot|$", "$\lceil\cdot\rceil$", "$(x_1,x_2)\mapsto \max\{x_1,x_2\}$", "$(x_1,x_2)\mapsto \min\{x_1,x_2\}$", "$(\cdot)^{\alpha_m}$" and "$\sqrt{\cdot}$", where all arguments possibly depend on ε. Without loss of generality, let $\bar{\mathfrak{c}}, \mathfrak{c}$ be positive constants independent of ε such that $\bar{\mathfrak{c}}\varepsilon^{-\mathfrak{c}} \geq \max\{K_0,\ldots,K_{\bar{m}}\}$.

Let $k \geq \bar{\mathfrak{c}}\varepsilon^{-\mathfrak{c}}$ be fixed. Multiplying two sides of (9.4.22) with $(\sum_{j\in[n]} x_j)^k$, we get

$$\begin{array}{rl} s_0 = & (\sum_{j\in[n]} x_j)^{D-d_f+k}(\tilde{f} + 2\varepsilon(\sum_{j\in[n]} x_j)^{d_f}) \\ & - \sum_{i\in[\bar{m}]} \tilde{g}_i s_i (\sum_{j\in[n]} x_j)^{D-4nu_i-d_{g_i}}\,, \end{array} \tag{9.4.27}$$

where $s_0 := (\sum_{j\in[n]} x_j)^k F$ and $s_i := (\sum_{j\in[n]} x_j)^k p_i$ are homogeneous polynomials having nonnegative coefficients. Replacing $\mathbf{x}$ by $\mathbf{x}^2$, we obtain:

$$\|\mathbf{x}\|_2^{2(D-d_f+k)}(f + 2\varepsilon\|\mathbf{x}\|_2^{2d_f}) = \sigma_0 + \textstyle\sum_{i\in[m]} g_i\sigma_i\,, \tag{9.4.28}$$

where

$$\begin{array}{rl} \sigma_0 & = s_0(\mathbf{x}^2) + \sum_{j\in[n]} \tilde{g}_{j+m}(\mathbf{x}^2)s_{j+m}(\mathbf{x}^2)\|\mathbf{x}\|_2^{2(D-4nu_{j+m}-d_{g_{j+m}})} \\ & = s_0(\mathbf{x}^2) + \sum_{j\in[n]} x_j^2 s_{j+m}(\mathbf{x}^2)\|\mathbf{x}\|_2^{2(D-4nu_{j+m}-d_{g_{j+m}})}\,, \end{array} \tag{9.4.29}$$

and

$$\sigma_i = s_i(\mathbf{x}^2)\|\mathbf{x}\|_2^{2(D-4nu_i-d_{g_i})}\,,\ i \in [m]\,, \tag{9.4.30}$$

are SOS of monomials. Set $K = D - d_f + K$. Then $\|\mathbf{x}\|_2^{2K}(f + 2\varepsilon\|\mathbf{x}\|_2^{2d_f}) = \sigma_0 + \sum_{i=1}^m g_i\sigma_i$ with $\deg(\sigma_0) = \deg(g_i\sigma_i) = 2(K + d_f)$, for $i \in [m]$. This completes the proof of Theorem 9.1. □

9.4.3 Variations of Pólya's and Handelman's Positivstellensatz

For every $t \in \mathbb{N}$, denote

$$\bar{\mathbf{v}}_t(\mathbf{x}) := \mathbf{v}_t(\tfrac{1}{2}(\mathbf{x}+\mathbf{e}), \tfrac{1}{2}(\mathbf{x}-\mathbf{e})) = (\tfrac{1}{2^{|\boldsymbol{\alpha}+\boldsymbol{\beta}|}}(\mathbf{x}+\mathbf{e})^{\boldsymbol{\alpha}}(\mathbf{x}-\mathbf{e})^{\boldsymbol{\beta}})_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathbb{N}^{2n}_t}\,, \tag{9.4.31}$$

where $\mathbf{e} := (1,\ldots,1) \in \mathbb{R}^n$.

As a consequence of Corollary 9.2, the next proposition shows that the weighted SOS polynomials in Putinar–Vasilescu's Positivstellensatz can be associated with diagonal Gram matrices via a change of monomial basis.

Proposition 9.2. *(Putinar–Vasilescu's Positivstellensatz with diagonal Gram matrices) Let $g_1, \dots, g_m$ be polynomials such that $g_1 := R - \|\mathbf{x}\|_2^2$ for some $R > 0$ and $g_m := 1$. Let S be the semialgebraic set defined by*

$$S := \{\mathbf{x} \in \mathbb{R}^n \,:\, g_1(\mathbf{x}) \geq 0 \dots, g_m(\mathbf{x}) \geq 0\}\,. \tag{9.4.32}$$

Let f be a polynomial of degree at most $2d_f$ such that f is nonnegative on S. Denote $d_{g_i} := \lceil \deg(g_i)/2 \rceil$. Then the following statements hold:

1. *For all $\varepsilon > 0$, there exists $K_\varepsilon \in \mathbb{N}$ such that for all $k \geq K_\varepsilon$, there exist vectors $\boldsymbol{\eta}^{(i)} \in \mathbb{R}_+^{b(2n,k+d_f-d_{g_i})}$ satisfying*

$$(\|\mathbf{x}\|_2^2 + n + 2)^k (f + \varepsilon) = \textstyle\sum_{i=1}^m g_i \bar{\mathbf{v}}_{k+d_f-d_{g_i}}^\top \operatorname{diag}(\boldsymbol{\eta}^{(i)}) \bar{\mathbf{v}}_{k+d_f-d_{g_i}}\,. \tag{9.4.33}$$

2. *If S has nonempty interior, then there exist positive constants $\bar{\mathfrak{c}}$ and $\mathfrak{c}$ depending on f, g_i such that for all $\varepsilon > 0$, one can take $K_\varepsilon = \bar{\mathfrak{c}}\varepsilon^{-\mathfrak{c}}$.*

Proof. Take two vectors of n variables $\mathbf{y} = (y_1, \dots, y_n)$ and $\mathbf{z} = (z_1, \dots, z_n)$. Given $q \in \mathbb{R}[\mathbf{x}]$, denote the polynomial $\hat{q}(\mathbf{y}, \mathbf{z}) = q(\mathbf{y}^2 - \mathbf{z}^2) \in \mathbb{R}[\mathbf{y}, \mathbf{z}]$. Let $\hat{g}_{m+1} := \frac{1}{2}(L + n) - \|(\mathbf{y}, \mathbf{z})\|_2^2$ and $d_{g_{m+1}} := 1$. Define

$$\hat{S} := \{(\mathbf{y}, \mathbf{z}) \in \mathbb{R}^{2n} \,:\, \hat{g}_i(\mathbf{y}, \mathbf{z}) \geq 0\,,\, i \in [m+1]\} \tag{9.4.34}$$

Note that $\hat{g}_1 := R - \|\mathbf{y}^2 - \mathbf{z}^2\|_2^2$ and $\hat{g}_m := 1$. Since $f \geq 0$ on S, replacing $\mathbf{x}$ by $\mathbf{y}^2 - \mathbf{z}^2$ gives $\hat{f} \geq 0$ on $\hat{S}$. From this and Corollary 9.2, there exist $\boldsymbol{\eta}^{(i)} \in \mathbb{R}_+^{b(2n,k+d_f-d_{g_i})}$ such that

$$(\|(\mathbf{y}, \mathbf{z})\|_2^2 + 1)^k (\hat{f} + \varepsilon) = \textstyle\sum_{i=1}^{m+1} \hat{g}_i \mathbf{v}_{k+d_f-d_{g_i}}(\mathbf{y}, \mathbf{z})^\top \operatorname{diag}(\boldsymbol{\eta}^{(i)}) \mathbf{v}_{k+d_f-d_{g_i}}(\mathbf{y}, \mathbf{z})\,. \tag{9.4.35}$$

With $\mathbf{y} = \frac{1}{2}(\mathbf{x} + \mathbf{e})$ and $\mathbf{z} = \frac{1}{2}(\mathbf{x} - \mathbf{e})$, it becomes

$$\tfrac{1}{2^k}(\|\mathbf{x}\|_2^2 + n + 2)^k (f + \varepsilon) = \textstyle\sum_{i=1}^{m+1} g_i \bar{\mathbf{v}}_{k+d_f-d_{g_i}}^\top \operatorname{diag}(\boldsymbol{\eta}^{(i)}) \bar{\mathbf{v}}_{k+d_f-d_{g_i}}\,. \tag{9.4.36}$$

Here $g_{m+1}(\cdot) := \hat{g}_{m+1}(\frac{1}{2}(\cdot + \mathbf{e}), \frac{1}{2}(\cdot - \mathbf{e})) = \frac{1}{2} g_1(\cdot)$. Indeed, since $\mathbf{y}^2 - \mathbf{z}^2 = \mathbf{x}$, $\hat{f}(\mathbf{y}, \mathbf{z}) = f(\mathbf{x})$ and $\hat{g}_i(\mathbf{y}, \mathbf{z}) = g_i(\mathbf{x})$, for $i \in [m]$. Since $\mathbf{y}^2 + \mathbf{z}^2 = \frac{1}{2}(\mathbf{x}^2 + \mathbf{e})$, $\|\mathbf{y}\|_2^2 + \|\mathbf{z}\|_2^2 = \frac{1}{2}(\|\mathbf{x}\|_2^2 + n)$. This implies that

$$\hat{g}_{m+1}(\mathbf{y}, \mathbf{z}) = \tfrac{1}{2}(L + n) - \|(\mathbf{y}, \mathbf{z})\|_2^2 = \tfrac{1}{2}(R - \|\mathbf{x}\|_2^2) = \tfrac{1}{2} g_1(\mathbf{x})\,. \tag{9.4.37}$$

Moreover, if $\mathbf{a}$ belongs to the interior of S, then $(\frac{1}{2}(\mathbf{a} + \mathbf{e}), \frac{1}{2}(\mathbf{a} - \mathbf{e}))$ belongs to the interior of $\hat{S}$. Thus, the desired result follows. □

As a consequence of Corollary 9.3, the next proposition states a new representation associated with diagonal Gram matrices for a polynomial positive on a compact semialgebraic set without assumption on even property.

Proposition 9.3. *(Representation without even symmetry) Let f, g_i, S, d_{g_i} be as in Proposition 9.2. Then the following statements hold:*

1. *For all $\varepsilon > 0$, there exists $K_\varepsilon \in \mathbb{N}$ such that for all $k \geq K_\varepsilon$, there exist vectors $\boldsymbol{\eta}^{(i,r)} \in \mathbb{R}_+^{b(2n,k-d_{g_i}-r)}$ satisfying*

$$f + \varepsilon = \textstyle\sum_{i=1}^m \sum_{r=0}^{k-d_{g_i}} g_i g_1^r \bar{\mathbf{v}}_{k-d_{g_i}-r}^\top \operatorname{diag}(\boldsymbol{\eta}^{(i,r)}) \bar{\mathbf{v}}_{k-d_{g_i}-r}\,. \tag{9.4.38}$$

2. *If S has nonempty interior, then there exist positive constants $\bar{\mathfrak{c}}$ and $\mathfrak{c}$ depending on f, g_i such that for all $\varepsilon > 0$, one can take $K_\varepsilon = \bar{\mathfrak{c}}\varepsilon^{-\mathfrak{c}}$.*

The proof of Proposition 9.3 relies on Corollary 9.3 and can be proved in almost the same way as the proof of Proposition 9.2.

Remark 9.15. *Representation* (9.4.38) *in Proposition 9.3 is similar in spirit to the one of Roebers et al. in [179, (3), page 4]. The difference here is that the weight of each constrained polynomial* g_i, $i \notin \{1, m\}$, *in* (9.4.38) *is the polynomial*

$$\textstyle\sum_{r=0}^{k-d_{g_i}} g_1^r \bar{\mathbf{v}}_{k-d_{g_i}-r}^\top \operatorname{diag}(\boldsymbol{\eta}^{(i,r)}) \bar{\mathbf{v}}_{k-d_{g_i}-r}\,, \tag{9.4.39}$$

which does not involve g_i. *This is in contrast with the weight associated to each* g_i *in [179, (3), page 4], which is of the form* $\sigma_i(U_i - g_i)$, *where* U_i *is an upper bound of* g_i *on the ball* $\{\mathbf{x} \in \mathbb{R}^n : g_1(\mathbf{x}) \geq 0\}$ *and* σ_i *is a univariate polynomial, e.g.,* $\sigma_i(t) = t^{2\xi}$ *for some* $\xi \in \mathbb{N}$.

Remark 9.16. *In view of Propositions 9.2 and 9.3, replacing the standard monomial basis* $\mathbf{v}_t$ *by the new basis* $\bar{\mathbf{v}}_t$ *can provide a Positivstellensatz involving Gram matrix of factor width* 1. *Thus, ones can build up hierarchies of semidefinite relaxations with any maximal matrix size, based on both representations* (9.4.38) *and* (9.4.33). *However, expressing the entries of the basis* $\bar{\mathbf{v}}_t$ *is a time-consuming task within the modeling process. A potential workaround is to impose* (9.4.38) *and* (9.4.33) *on a set of generic points similarly to [109, Section 2.3]. This needs further study.*

9.4.4 Polynomial optimization on the nonnegative orthant: Noncompact case

Linear relaxations

Given $\varepsilon > 0$, consider the hierarchy of linear programs indexed by $k \in \mathbb{N}$:

$$\begin{aligned} \tau_k^{\text{Pól}}(\varepsilon) := \inf_{\mathbf{y}} \quad & L_{\mathbf{y}}(\theta^k(\check{f} + \varepsilon\theta^{d_f})) \\ \text{s.t.} \quad & \mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in \mathbb{N}^n_{2(d_f+k)}} \subset \mathbb{R}\,,\, L_{\mathbf{y}}(\theta^k) = 1\,, \\ & \operatorname{diag}(\mathbf{M}_{k_i}(\check{g}_i \mathbf{y})) \in \mathbb{R}_+^{b(n,k_i)}\,,\, i \in [m]\,, \end{aligned} \tag{9.4.40}$$

where $k_i := k + d_f - d_{g_i}$, $i \in [m]$. Here $\check{g}_m = 1$. Note that

$$\operatorname{diag}(\mathbf{M}_{k_i}(\check{g}_i \mathbf{y})) = (\textstyle\sum_{\boldsymbol{\gamma} \in \mathbb{N}^n_{2d_{g_i}}} y_{2\boldsymbol{\alpha}+\boldsymbol{\gamma}} \check{g}_{i,\boldsymbol{\gamma}})_{\boldsymbol{\alpha} \in \mathbb{N}^n_{k_i}}\,. \tag{9.4.41}$$

Theorem 9.6. *Let* $f, g_i \in \mathbb{R}[\mathbf{x}]$, $i \in [m]$, *with* $g_m = 1$. *Consider POP* (9.2.1) *with* S *being defined as in* (9.2.2). *Let* $\varepsilon > 0$ *be fixed. For every* $k \in \mathbb{N}$, *the dual of* (9.4.40) *reads as:*

$$\begin{aligned} \rho_k^{\text{Pól}}(\varepsilon) := \sup_{\lambda, \mathbf{u}_i} \quad & \lambda \\ \text{s.t.} \quad & \lambda \in \mathbb{R}\,,\, \mathbf{u}_i \in \mathbb{R}_+^{b(n,k_i)}\,,\, i \in [m]\,, \\ & \theta^k(\check{f} - \lambda + \varepsilon\theta^{d_f}) = \textstyle\sum_{i=1}^m \check{g}_i \mathbf{v}_{k_i}^\top \operatorname{diag}(\mathbf{u}_i) \mathbf{v}_{k_i}\,. \end{aligned} \tag{9.4.42}$$

Here $\check{g}_m = 1$. *The following statements hold:*

1. *For all* $k \in \mathbb{N}$, $\rho_k^{\text{Pól}}(\varepsilon) \leq \rho_{k+1}^{\text{Pól}}(\varepsilon) \leq f^\star$.
2. *There exists* $K \in \mathbb{N}$ *such that for all* $k \geq K$, $0 \leq f^\star - \rho_k^{\text{Pól}}(\varepsilon) \leq \varepsilon\theta(\mathbf{x}^{\star 2})^{d_f}$.
3. *If* S *has nonempty interior, there exist positive constants* $\bar{\mathfrak{c}}$ *and* $\mathfrak{c}$ *depending on* f, g_i *such that for all* $k \geq \bar{\mathfrak{c}}\varepsilon^{-\mathfrak{c}}$, $0 \leq f^\star - \rho_k^{\text{Pól}}(\varepsilon) \leq \varepsilon\theta(\mathbf{x}^{\star 2})^{d_f}$.

The proof of Theorem 9.6 relies on Corollary 9.1 and is exactly the same as the proof of Theorem 6.5.

Semidefinite relaxations

Given $\varepsilon > 0$, consider the hierarchy of semidefinite programs indexed by $s \in \mathbb{N}_{>0}$ and $k \in \mathbb{N}$:

$$\begin{aligned} \tau_{k,s}^{\text{Pól}}(\varepsilon) := \inf_{\mathbf{y}} \quad & L_{\mathbf{y}}(\theta^k(\check{f} + \varepsilon\theta^{d_f})) \\ \text{s.t.} \quad & \mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in \mathbb{N}^n_{2(d_f+k)}} \subset \mathbb{R}\,,\, L_{\mathbf{y}}(\theta^k) = 1\,, \\ & \mathbf{M}_{\mathcal{A}_j^{(s,k_i)}}(\check{g}_i \mathbf{y}) \succeq 0\,,\, j \in [b(n,k_i)]\,,\, i \in [m]\,, \end{aligned} \tag{9.4.43}$$

where $k_i := k + d_f - d_{g_i}$, $i \in [m]$. Here $\check{g}_m = 1$.

Theorem 9.7. *Let $f, g_i \in \mathbb{R}[\mathbf{x}]$, $i \in [m]$, with $g_m = 1$. Consider POP* (9.2.1) *with S being defined as in* (9.2.2). *Let $\varepsilon > 0$ be fixed. For every $k \in \mathbb{N}$ and for every $s \in \mathbb{N}_{>0}$, the dual of* (9.4.43) *reads as:*

$$\begin{aligned} \rho_{k,s}^{\mathrm{Pól}}(\varepsilon) := \sup_{\lambda, \mathbf{G}_{ij}} \quad & \lambda \\ \text{s.t.} \quad & \lambda \in \mathbb{R}\,,\ \mathbf{G}_{ij} \succeq 0\,,\ j \in [b(n,k_i)]\,,\ i \in [m]\,, \\ & \theta^k(\check{f} - \lambda + \varepsilon\theta^{d_f}) = \textstyle\sum_{i\in[m]} \check{g}_i\big(\sum_{j\in[b(n,k_i)]} \mathbf{v}_{\mathcal{A}_j^{(s,k_i)}}^\top \mathbf{G}_{ij} \mathbf{v}_{\mathcal{A}_j^{(s,k_i)}}\big)\,. \end{aligned} \tag{9.4.44}$$

The following statements hold:

1. *For all $k \in \mathbb{N}_{>0}$ and for every $s \in \mathbb{N}_{>0}$, $\rho_k^{\mathrm{Pól}} = \rho_{k,1}^{\mathrm{Pól}}(\varepsilon) \le \rho_{k,s}^{\mathrm{Pól}}(\varepsilon)$.*
2. *For every $s \in \mathbb{N}_{>0}$, there exists $K \in \mathbb{N}$ such that for every $k \in \mathbb{N}$ satisfying $k \ge K$, $0 \le f^\star - \rho_{k,s}^{\mathrm{Pól}}(\varepsilon) \le \varepsilon\theta(\mathbf{x}^{\star 2})^{d_f}$.*
3. *If S has nonempty interior, there exist positive constants $\bar{\mathfrak{c}}$ and $\mathfrak{c}$ depending on f, g_i such that for every $s \in \mathbb{N}_{>0}$ and for every $k \in \mathbb{N}$ satisfying $k \ge \bar{\mathfrak{c}}\varepsilon^{-\mathfrak{c}}$, $0 \le f^\star - \rho_{k,s}^{\mathrm{Pól}}(\varepsilon) \le \varepsilon\theta(\mathbf{x}^{\star 2})^{d_f}$.*
4. *If S has nonempty interior, for every $s \in \mathbb{N}_{>0}$ and for every $k \in \mathbb{N}$ strong duality holds for the primal-dual problems* (9.4.43)-(9.4.44).

The proof of Theorem 9.7 is based on Theorem 9.6, Theorem 7.2 and the inequalities $\rho_k^{\mathrm{Pól}}(\varepsilon) \le \rho_{k,s}^{\mathrm{Pól}}(\varepsilon) \le \rho_k^{(\varepsilon)}$, where $\rho_k^{(\varepsilon)}$ is defined as in (7.2.4). For each $q \in \mathbb{R}[\mathbf{x}]_d$, denote the degree-$d$ homogenization of q by $x_{n+1}^d q(\frac{\mathbf{x}}{x_{n+1}}) \in \mathbb{R}[\mathbf{x}, x_{n+1}]$.

Remark 9.17. *Let $(\lambda, \mathbf{G}_{ij})$ be a feasible solution of* (9.4.44) *and consider the case of $m = 1$. Then the equality constraint of* (9.4.44) *becomes*

$$\theta^k(\check{f} - \lambda + \varepsilon\theta^{d_f}) = \textstyle\sum_{j\in[b(n,k_m)]} \mathbf{v}_{\mathcal{A}_j^{(s,k_m)}}^\top \mathbf{G}_{mj} \mathbf{v}_{\mathcal{A}_j^{(s,k_m)}}\,. \tag{9.4.45}$$

It implies that the degree-$2d_f$ homogenization of $\check{f} - \lambda + \varepsilon\theta^{d_f}$ belongs to the cone k-DSOS$_{n+1,2d_f}$ (resp. k-SDSOS$_{n+1,2d_f}$) when $s = 1$ (resp. $s = 2$) according to [3, Definition 3.10]. More generally, the polynomial $\theta^k(\check{f} - \lambda + \varepsilon\theta^{d_f})$ belongs to the cone of SOS polynomials whose Gram matrix has factor width at most s (see [3, Section 5.3]).

9.4.5 Sparse representation theorem

For every $I = \{i_1, \ldots, i_r\} \subset [n]$ with $i_1 < \cdots < i_r$, denote $\mathbf{x}(I) = (x_{i_1}, \ldots, x_{i_r})$.
We will make the following assumptions:

Assumption 9.1. *With $p \in \mathbb{N}_{>0}$, the following conditions hold:*

1. *There exists $(I_c)_{c\in[p]}$ being a sequence of subsets of $[n]$ such that $\cup_{c\in[p]} I_c = [n]$ and*

$$\forall c \in \{2, \ldots, p\}\,,\ \exists r_c \in [c-1]\ :\ I_c \cap (\cup_{t=1}^{c-1} I_t) \subset I_{r_c}\,. \tag{9.4.46}$$

   Denote $n_c := |I_c|$, for $c \in [p]$.

2. *With $m \in \mathbb{N}_{>0}$ and $(g_i)_{i\in[m]} \subset \mathbb{R}[\mathbf{x}]$, there exists $(J_c)_{c\in[p]}$ being a finite sequence of subsets of $[m]$ such that $\cup_{c\in[p]} J_c = [m]$ and*

$$\forall c \in [p]\,,\ (g_i)_{i\in J_c} \subset \mathbb{R}[\mathbf{x}(I_c)]\,. \tag{9.4.47}$$

3. *For every $c \in [p]$, there exists $i_c \in J_c$ and $R_c > 0$ such that*

$$g_{i_c} := R_c - \|\mathbf{x}(I_c)\|_2^2\,. \tag{9.4.48}$$

The condition (9.4.46) is called the running intersection property (RIP).
Let $\theta_c := 1 + \|\mathbf{x}(I_c)\|_2^2$, $c \in [p]$.

Extension of Pólya's Positivstellensatz

We state the sparse representation in the following theorem:

Theorem 9.8. *Let $g_1, \ldots, g_m$ be polynomials such that $g_1, \ldots, g_m$ are even in each variable and Assumption 9.1 holds. Let S be the semialgebraic set defined by*

$$S := \{\mathbf{x} \in \mathbb{R}^n \,:\, g_1(\mathbf{x}) \geq 0 \ldots, g_m(\mathbf{x}) \geq 0\}\,. \tag{9.4.49}$$

Let $f = f_1 + \cdots + f_p$ be a polynomial such that f is positive on S and for every $c \in [p]$, $f_c \in \mathbb{R}[\mathbf{x}(I_c)]$ is even in each variable. Then there exist $d, k \in \mathbb{N}$, $h_c \in \mathbb{R}[\mathbf{x}(I_c)]$, $\sigma_{0,c}, \sigma_{j,c} \in \mathbb{R}[\mathbf{x}(I_c)]$, for $j \in J_c$ and $c \in [p]$, such that the following conditions hold:

1. *The equality $f = h_1 + \cdots + h_p$ holds and h_c is a polynomial of degree at most $2d$ which is even in each variable.*

2. *For all $i \in J_c$ and $c \in [p]$, $\sigma_{0,c}, \sigma_{i,c}$ are SOS of monomials satisfying*

$$\deg(\sigma_{0,c}) \leq 2(k+d) \quad \textit{and} \quad \deg(\sigma_{i,c} g_i) \leq 2(k+d) \tag{9.4.50}$$

and

$$\theta_c^k h_c = \sigma_{0,c} + \textstyle\sum_{i \in J_c} \sigma_{i,c} g_i\,. \tag{9.4.51}$$

Proof. Let $\varepsilon > 0$. Similarly as in Step 1 of the proof of Theorem 9.1, $\tilde{f} = \tilde{f}_1 + \cdots + \tilde{f}_m$ is positive on the semialgebraic set $\tilde{S}$ defined as in (9.4.5). For every $c \in [p]$, let $\tilde{J}_c := J_c \cup (m + I_c)$. Recall that $\tilde{g}_{m+j} := x_j$, $j \in [n]$. By applying [68, Lemma 4], there exist polynomials $s_c, q_{i,c} \in \mathbb{R}[\mathbf{x}(I_c)]$, for $j \in \tilde{J}_c$ and $c \in [p]$, such that

$$\tilde{f} = \textstyle\sum_{c=1}^p (s_c + \sum_{i \in \tilde{J}_c} q_{i,c}^2 \tilde{g}_i)\,, \tag{9.4.52}$$

and for all $c \in [p]$, s_c is positive on the set

$$\{\mathbf{x}(I_c) \in \mathbb{R}^{n_c} \,:\, x_j \geq 0\,,\, j \in I_c\,,\, \tilde{g}_{i_c}(\mathbf{x}) = R_c - \textstyle\sum_{j \in I_c} x_j \geq 0\}\,. \tag{9.4.53}$$

Set $h_c := s_c(\mathbf{x}^2) + \sum_{i \in \tilde{J}_c} q_{i,c}(\mathbf{x}^2)^2 \tilde{g}_i(\mathbf{x}^2)$, $c \in [p]$. Let $d \in \mathbb{N}$ such that $2d - 1 \geq \max\{\deg(h_c) \,:\, c \in [p]\}$. Then $f = \sum_{c=1}^p h_c$ with $h_c \in \mathbb{R}[\mathbf{x}(I_c)]_{2d}$ being even in each variable and positive on the semialgebraic set

$$S_c := \{\mathbf{x}(I_c) \in \mathbb{R}^{n_c} \,:\, g_i(\mathbf{x}) \geq 0\,,\, i \in J_c\}\,. \tag{9.4.54}$$

Note that $g_{i_c} := R_c - \|\mathbf{x}(I_c)\|_2^2$ with $i_c \in J_c$. By applying Corollary 9.2, there exists $k_c \in \mathbb{N}$ such that for all $K \geq k_c$, there exist $\sigma_{0,c}, \sigma_{i,c} \in \mathbb{R}[\mathbf{x}(I_c)]$, $i \in J_c$, such that $\sigma_{0,c}, \sigma_{i,c}$ are SOS of monomials satisfying

$$\deg(\sigma_{0,c}) \leq 2(K+d) \quad \text{and} \quad \deg(\sigma_{i,c} g_i) \leq 2(K+d) \tag{9.4.55}$$

for all $i \in J_c$, and

$$\theta_c^K h_c = \sigma_{0,c} + \textstyle\sum_{i \in J_c} \sigma_{i,c} g_i\,. \tag{9.4.56}$$

Set $k = \max\{k^{(c)} \,:\, c \in [p]\}$. Finally, we obtain the desired results. □

Remark 9.18. *In Theorem 9.8, it is not hard to see that f has a rational SOS decomposition*

$$f = \textstyle\sum_{c \in [p]} \frac{\sigma_{0,c} + \sum_{i \in J_c} \sigma_{i,c} g_i}{\theta_c^k}\,. \tag{9.4.57}$$

This decomposition is simpler than the ones provided in [133] and thus is more applicable to polynomial optimization.

Another sparse representation without denominators can be found in the next theorem. However, the number of SOS of monomials is not fixed in this case.

Extension of Handelman's Positivstellensatz

We process similarly to the proof of Theorem 9.8 and apply Corollary 9.3 to obtain the following theorem:

Theorem 9.9. *(Sparse representation without multiplier) Let $f, g_1, \dots, g_m$ be as in Theorem 9.8. Assume that $g_m := 1$ and $m \in J_c$, for all $c \in [p]$. Denote $d_{g_i} := \lceil \deg(g_i)/2 \rceil$. Then there exist $k \in \mathbb{N}$, SOS of monomials $\sigma_{i,j,c} \in \mathbb{R}[\mathbf{x}(I_c)]$, for $j = 0, \dots, k - d_{g_i}$, $i \in J_c$ and $c \in [p]$, satisfying*

$$\deg(\sigma_{i,j,c} g_{i_c}^j g_i) \le 2k \tag{9.4.58}$$

and

$$f = \textstyle\sum_{c\in[p]} \sum_{i\in J_c} \sum_{j=0}^{k-d_{g_i}} \sigma_{i,j,c} g_{i_c}^j g_i \,. \tag{9.4.59}$$

9.4.6 Sparse polynomial optimization on the nonnegative orthant

Consider the following POP:

$$f^\star := \inf_{\mathbf{x}\in S} f(\mathbf{x}) \,, \tag{9.4.60}$$

where $f \in \mathbb{R}[\mathbf{x}]$ and

$$S = \{\mathbf{x} \in \mathbb{R}^n \,:\, x_j \ge 0 \,,\, j \in [n] \,,\, g_i(\mathbf{x}) \ge 0 \,,\, i \in [m]\} \,, \tag{9.4.61}$$

for some $g_i \in \mathbb{R}[\mathbf{x}]$, $i \in [m]$, with $g_m = 1$. Assume that $f^\star > -\infty$ and problem (9.4.60) has an optimal solution $\mathbf{x}^\star$.

Then POP (9.4.60) is equivalent to

$$f^\star := \inf_{\mathbf{x}\in \check{S}} \check{f} \,, \tag{9.4.62}$$

where

$$\check{S} = \{\mathbf{x} \in \mathbb{R}^n \,:\, \check{g}_i(\mathbf{x}) \ge 0 \,,\, i \in [m]\} \,, \tag{9.4.63}$$

with optimal solution $\mathbf{x}^{\star 2}$.

We will make the following assumptions:

Assumption 9.2. *With $p \in \mathbb{N}_{>0}$, the first two conditions of Assumption 9.1 and the following conditions hold:*

1. *For every $c \in [p]$, there exist $i_c \in J_c$ and $R_c > 0$ such that*

$$g_{i_c} = R_c - \textstyle\sum_{j\in I_c} x_j \,. \tag{9.4.64}$$

2. *There exist $f_c \in \mathbb{R}[\mathbf{x}(I_c)]$, for $c \in [p]$, such that $f = f_1 + \cdots + f_p$.*

Linear relaxations

Based on the extension of Pólya's Positivstellensatz: Consider the hierarchy of linear programs indexed by $k, d \in \mathbb{N}$:

$$\begin{aligned} \tau_{k,d}^{\mathrm{spP\acute{o}l}} := \inf_{\mathbf{y},\mathbf{y}^{(t)}} \quad & L_{\mathbf{y}}(\check{f}) \\ \text{s.t.} \quad & \mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{2d}} \subset \mathbb{R} \,,\, \mathbf{y}^{(c)} = (y^{(c)}_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{2(d+k)}} \subset \mathbb{R} \,,\, c \in [p] \,, \\ & \operatorname{diag}(\mathbf{M}_d(\mathbf{y}, I_c)) = \operatorname{diag}(\mathbf{M}_d(\theta_c^k \mathbf{y}^{(c)}, I_c)) \,,\, c \in [p] \,, \\ & \operatorname{diag}(\mathbf{M}_{k_i^{(d)}}(\check{g}_i \mathbf{y}^{(c)}, I_c)) \in \mathbb{R}_+^{b(n_c, k_i^{(d)})} \,,\, i \in [m] \,,\, c \in [p] \,,\, y_{\mathbf{0}} = 1 \,, \end{aligned} \tag{9.4.65}$$

where $k_i^{(d)} := k + d - d_{g_i}$.

Theorem 9.10. *Let $f, g_i \in \mathbb{R}[\mathbf{x}]$, $i \in [m]$, with $g_m = 1$. Consider POP* (9.4.60) *with S being defined as in* (9.4.61)*. Let Assumption 9.2 hold. The dual of SDP* (9.4.65) *reads as:*

$$\begin{aligned}\rho_{k,d}^{\mathrm{spP\acute{o}l}} := \sup_{\lambda,\mathbf{u}_c,\mathbf{w}_i^{(c)}} \quad & \lambda \\ \text{s.t.} \quad & \lambda \in \mathbb{R}\,,\ \mathbf{u}_c \in \mathbb{R}^{b(n_c,d)}\,,\ \mathbf{w}_i^{(c)} \in \mathbb{R}_+^{b(n_c,k_i^{(d)})}\,,\ i \in J_c\,,\ c \in [p]\,, \\ & \check{f} - \lambda = \textstyle\sum_{c\in[p]} h_c\,,\ h_c = \mathbf{v}_{\mathbb{N}_d^{I_c}}^\top \operatorname{diag}(\mathbf{u}_c)\mathbf{v}_{\mathbb{N}_d^{I_c}}\,,\ c \in [p]\,, \\ & \theta_c^k h_c = \textstyle\sum_{i\in J_c} \check{g}_i \mathbf{v}_{\mathbb{N}^{I_c}_{k_i^{(d)}}}^\top \operatorname{diag}(\mathbf{w}_i^{(c)})\mathbf{v}_{\mathbb{N}^{I_c}_{k_i^{(d)}}}\,,\ c \in [p]\,.\end{aligned} \tag{9.4.66}$$

The following statements hold:

1. *For all $k \in \mathbb{N}$ and for every $s \in \mathbb{N}_{>0}$, $\rho_{k-1,d}^{\mathrm{spP\acute{o}l}} \le \rho_{k,d}^{\mathrm{spP\acute{o}l}} \le \rho_{k,d+1}^{\mathrm{spP\acute{o}l}} \le f^\star$.*
2. *One has*
$$\sup\{\rho_{k,d}^{\mathrm{spP\acute{o}l}} \,:\, (k,d) \in \mathbb{N}^2\} = f^\star\,. \tag{9.4.67}$$

Proof. It is fairly easy to see that the first statement holds. Let us prove the second one. Let $\check{g}_{i_c} := R_c - \|\mathbf{x}(I_c)\|_2^2$ and $\varepsilon > 0$. Then $\check{f} - (f^\star - \varepsilon) > 0$ on S. By applying Theorem 9.8, there exist $d, k \in \mathbb{N}$, $h_c \in \mathbb{R}[\mathbf{x}(I_c)]$, $\sigma_{0,c}, \sigma_{j,c} \in \mathbb{R}[\mathbf{x}(I_c)]$, for $j \in J_c$ and $c \in [p]$, such that the following conditions hold:

1. The equality $\check{f} - (f^\star - \varepsilon) = h_1 + \cdots + h_p$ holds and h_c is a polynomial of degree at most $2d$ which is even in each variable.
2. For all $i \in J_c$ and $c \in [p]$, $\sigma_{0,c}, \sigma_{i,c}$ are SOS of monomials satisfying
$$\deg(\sigma_{0,c}) \le 2(k+d) \quad \text{and} \quad \deg(\sigma_{i,c}\check{g}_i) \le 2(k+d) \tag{9.4.68}$$
and
$$\theta_c^k h_c = \sigma_{0,c} + \textstyle\sum_{i\in J_c} \sigma_{i,c}\check{g}_i\,. \tag{9.4.69}$$

It implies that there exists $\mathbf{u}_c \in \mathbb{R}^{b(n_c,d)}$, $\mathbf{w}_i^{(c)} \in \mathbb{R}_+^{b(n_c,k_i^{(d)})}$ such that

$$h_c = \mathbf{v}_{\mathbb{N}_d^{I_c}}^\top \operatorname{diag}(\mathbf{u}_c)\mathbf{v}_{\mathbb{N}_d^{I_c}} \quad \text{and} \quad \sigma_{i,c} := \mathbf{v}_{\mathbb{N}^{I_c}_{k_i^{(d)}}}^\top \operatorname{diag}(\mathbf{w}_i^{(c)})\mathbf{v}_{\mathbb{N}^{I_c}_{k_i^{(d)}}}\,, \tag{9.4.70}$$

for $i \in J_c$ and $c \in [p]$. It implies that $(f^\star - \varepsilon, \mathbf{u}_c, \mathbf{w}_i^{(c)})$ is an optimal solution of LP (9.4.66). Thus $\rho_{k,d}^{\mathrm{P\acute{o}l}} \ge f^\star - \varepsilon$, yielding (9.4.67). □

Based on the extension of Handelman's Positivstellensatz: Consider the hierarchy of linear programs indexed by $k \in \mathbb{N}$:

$$\begin{aligned}\tau_k^{\mathrm{spHan}} := \inf_{\mathbf{y}} \quad & L_{\mathbf{y}}(\check{f}) \\ \text{s.t.} \quad & \mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{2k}} \subset \mathbb{R}\,,\ y_{\mathbf{0}} = 1\,, \\ & \operatorname{diag}(\mathbf{M}_{k_{ij}}((\check{g}_i\check{g}_{i_c}^j)\mathbf{y}, I_c)) \in \mathbb{R}_+^{b(n_c,k_{ij})}\,,\ c \in [p]\,,\ i \in [m]\,,\ j \in \{0\} \cup [k - d_{g_i}]\,,\end{aligned} \tag{9.4.71}$$

where $k_{ij} := k - d_{g_i} - j$, for $i \in [m]$, for $j \in \{0\} \cup [k - d_{g_i}]$.

Theorem 9.11. *Let $f, g_i \in \mathbb{R}[\mathbf{x}]$, $i \in [m]$, with $g_m = 1$. Consider POP* (9.4.60) *with S being defined as in* (9.4.61)*. Let Assumption 9.2 hold. The dual of SDP* (9.4.71) *reads as:*

$$\begin{aligned}\rho_k^{\mathrm{spHan}} := \sup_{\lambda,\mathbf{w}_{ij}^{(c)}} \quad & \lambda \\ \text{s.t.} \quad & \lambda \in \mathbb{R}\,,\ \mathbf{w}_{ij}^{(c)} \in \mathbb{R}_+^{b(n_c,k_{ij})}\,,\ c \in [p]\,,\ i \in J_c\,,\ j \in \{0\} \cup [k - d_{g_i}]\,, \\ & \check{f} - \lambda = \textstyle\sum_{c\in[p]} \sum_{i\in J_c} \sum_{j=0}^{k-d_{g_i}} \check{g}_i\check{g}_{i_c}^j \mathbf{v}_{\mathbb{N}^{I_c}_{k_{ij}}}^\top \operatorname{diag}(\mathbf{w}_{ij}^{(c)})\mathbf{v}_{\mathbb{N}^{I_c}_{k_{ij}}}\,.\end{aligned} \tag{9.4.72}$$

The following statements hold:

1. *For all* $k \in \mathbb{N}$, $\rho_k^{\mathrm{spHan}} \leq \rho_{k+1}^{\mathrm{spHan}} \leq f^\star$.

2. *The sequence* $(\rho_k^{\mathrm{spHan}})_{k\in\mathbb{N}}$ *converges to* $f^\star$.

The proof of Theorem 9.11 relies on Theorem 9.9 and is similar to Theorem 9.3.

Semidefinite relaxations

For every $I \subset [n]$, we write $\mathbb{N}^I = \{\boldsymbol{\alpha}_1^{(I)}, \boldsymbol{\alpha}_2^{(I)}, \ldots, \boldsymbol{\alpha}_r^{(I)}, \boldsymbol{\alpha}_{r+1}^{(I)}, \ldots\}$ such that

$$\boldsymbol{\alpha}_1^{(I)} < \boldsymbol{\alpha}_2^{(I)} < \cdots < \boldsymbol{\alpha}_r^{(I)} < \boldsymbol{\alpha}_{r+1}^{(I)} < \ldots . \tag{9.4.73}$$

Let

$$W_j^{(I)} := \{i \in \mathbb{N} : i \geq j\,,\, \boldsymbol{\alpha}_i^{(I)} + \boldsymbol{\alpha}_j^{(I)} \in 2\mathbb{N}^I\}\,, \quad j \in \mathbb{N}_{>0}\,,\, I \subset [n]\,. \tag{9.4.74}$$

Then for all $j \in \mathbb{N}_{>0}$ and for all $I \subset [n]$, $W_j^{(I)} \neq \emptyset$ since $j \in W_j^{(I)}$. For every $j \in \mathbb{N}$ and for every $I \subset [n]$, we write $W_j^{(I)} := \{i_{1,I}^{(j)}, i_{2,I}^{(j)}, \ldots\}$ such that $i_{1,I}^{(j)} < i_{2,I}^{(j)} < \ldots$. Let

$$\mathcal{T}_{j,I}^{(s,d)} = \{\boldsymbol{\alpha}_{i_{1,I}^{(j)}}^{(I)}, \ldots, \boldsymbol{\alpha}_{i_{s,I}^{(j)}}^{(I)}\} \cap \mathbb{N}_d^I\,, \quad I \subset [n]\,,\, j, s \in \mathbb{N}_{>0}\,,\, d \in \mathbb{N}\,. \tag{9.4.75}$$

For every $s \in \mathbb{N}_{>0}$, for every $d \in \mathbb{N}$ and for every $I \subset [n]$, define $\mathcal{A}_{1,I}^{(s,d)} := \mathcal{T}_{1,I}^{(s,d)}$ and for $j = 2, \ldots, b(|I|, d)$, define

$$\mathcal{A}_{j,I}^{(s,d)} := \begin{cases} \mathcal{T}_{j,I}^{(s,d)} & \text{if } T_{j,I}^{(s,d)} \backslash \mathcal{A}_{l,I}^{(s,d)} \neq \emptyset\,,\, \forall l \in [j-1]\,, \\ \emptyset & \text{otherwise}\,. \end{cases} \tag{9.4.76}$$

Note that $\cup_{j=1}^{b(|I|,d)} \mathcal{A}_{j,I}^{(s,d)} = \mathbb{N}_d^I$ and $|\mathcal{A}_{j,I}^{(s,d)}| \leq s$. Then the sequence

$$(\boldsymbol{\alpha} + \boldsymbol{\beta})_{\left(\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathcal{A}_{j,I}^{(s,d)}\right)}\,,\, j \in [b(|I|, d)] \tag{9.4.77}$$

are overlapping blocks of size at most s in $(\boldsymbol{\alpha} + \boldsymbol{\beta})_{(\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{N}_d^I)}$.

Example 9.3. *Consider the case of* $n = d = s = 2$, $I_1 = \{1\}$ *and* $I_2 = \{2\}$. *Matrix* $(\boldsymbol{\alpha}+\boldsymbol{\beta})_{(\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{N}^2_2)}$ *is written explicitly as in* (9.2.15). *We obtain two blocks:*

$$(\boldsymbol{\alpha} + \boldsymbol{\beta})_{(\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{N}_2^{I_1})} = \begin{bmatrix} (\mathbf{0},\mathbf{0}) & (1,0) & (\mathbf{2},\mathbf{0}) \\ (1,0) & (\mathbf{2},\mathbf{0}) & (3,0) \\ (\mathbf{2},\mathbf{0}) & (3,0) & (\mathbf{4},\mathbf{0}) \end{bmatrix} \tag{9.4.78}$$

and

$$(\boldsymbol{\alpha} + \boldsymbol{\beta})_{(\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{N}_2^{I_2})} = \begin{bmatrix} (\mathbf{0},\mathbf{0}) & (0,1) & (\mathbf{0},\mathbf{2}) \\ (0,1) & (\mathbf{0},\mathbf{2}) & (0,3) \\ (\mathbf{0},\mathbf{2}) & (0,3) & (\mathbf{0},\mathbf{4}) \end{bmatrix} \tag{9.4.79}$$

Then $\mathcal{A}_{1,I_1}^{(2,2)} = \{(0,0),(2,0)\}$, $\mathcal{A}_{2,I_1}^{(2,2)} = \{(1,0)\}$, $\mathcal{A}_{3,I_1}^{(2,2)} = \emptyset$ *and* $\mathcal{A}_{1,2}^{(2,I_2)} = \{(0,0),(0,2)\}$, $\mathcal{A}_{2,I_2}^{(2,2)} = \{(0,1)\}$, $\mathcal{A}_{2,I_2}^{(3,2)} = \emptyset$.

For every $I \subset [n]$, with $\mathcal{B} = \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_r\} \subset \mathbb{N}^I$ such that $\boldsymbol{\beta}_1 < \cdots < \boldsymbol{\beta}_r$, for every $h = \sum_{\boldsymbol{\gamma}\in\mathbb{N}^I} h_{\boldsymbol{\gamma}} \mathbf{x}^{\boldsymbol{\gamma}} \in \mathbb{R}[\mathbf{x}(I)]$ and $\mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n} \subset \mathbb{R}$, denote $\mathbf{M}_{\mathcal{B}}(h\mathbf{y}, I) := (\sum_{\boldsymbol{\gamma}\in\mathbb{N}^I} h_{\boldsymbol{\gamma}} y_{\boldsymbol{\gamma}+\boldsymbol{\beta}_i+\boldsymbol{\beta}_j})_{i,j\in[r]}$.

Based on the extension of Pólya's Positivstellensatz: Consider the hierarchy of linear programs indexed by $k, d \in \mathbb{N}$ and $s \in \mathbb{N}_{>0}$:

$$\begin{aligned} \tau_{k,d,s}^{\mathrm{spPól}} := \inf_{\mathbf{y},\mathbf{y}^{(c)}} \quad & L_{\mathbf{y}}(\check{f}) \\ \text{s.t.} \quad & \mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{2d}} \subset \mathbb{R}\,,\, \mathbf{y}^{(c)} = (y_{\boldsymbol{\alpha}}^{(c)})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{2(d+k)}} \subset \mathbb{R}\,,\, c \in [p]\,, \\ & \mathrm{diag}(\mathbf{M}_d(\mathbf{y}, I_c)) = \mathrm{diag}(\mathbf{M}_d(\theta_c^k \mathbf{y}^{(c)}, I_c))\,,\, c \in [p]\,,\, y_{\mathbf{0}} = 1\,, \\ & \mathbf{M}_{\mathcal{A}_{j,I_c}^{(s,k_i^{(d)})}}(\check{g}_i \mathbf{y}^{(c)}, I_c) \succeq 0\,,\, j \in [b(n_c, k_i^{(d)})]\,,\, i \in J_c\,,\, c \in [p]\,, \end{aligned} \tag{9.4.80}$$

where $k_i^{(d)} := k + d - d_{g_i}$.

Theorem 9.12. *Let $f, g_i \in \mathbb{R}[\mathbf{x}]$, $i \in [m]$, with $g_m = 1$. Consider POP* (9.4.60) *with S being defined as in* (9.4.61)*. Let Assumption 9.2 hold. The dual of SDP* (9.4.80) *reads as:*

$$\begin{aligned}\rho^{\mathrm{spP\acute{o}l}}_{k,d,s} := \sup_{\lambda,\mathbf{u}_c,\mathbf{G}^{(c)}_{i,j}} \quad & \lambda \\ \text{s.t.} \quad & \lambda \in \mathbb{R}\,,\ \mathbf{u}_c \in \mathbb{R}^{b(n_c,d)}\,,\ \mathbf{G}^{(c)}_{i,j} \succeq 0\,,\ j \in [b(n_c,k^{(d)}_i)]\,,\ i \in J_c\,,\ c \in [p]\,, \\ & \check{f} - \lambda = \textstyle\sum_{c\in[p]} h_c\,,\ h_c = \mathbf{v}^{\top}_{\mathbb{N}^{I_c}_d}\operatorname{diag}(\mathbf{u}_c)\mathbf{v}_{\mathbb{N}^{I_c}_d}\,,\ c \in [p]\,, \\ & \theta^k_c h_c = \textstyle\sum_{i\in J_c} \check{g}_i\Big(\sum_{j\in[b(n_c,k^{(d)}_i)]} \mathbf{v}^{\top}_{\mathcal{A}^{(s,k^{(d)}_i)}_{j,I_c}} \mathbf{G}^{(c)}_{i,j}\mathbf{v}_{\mathcal{A}^{(s,k^{(d)}_i)}_{j,I_c}}\Big)\,,\ c \in [p]\,.\end{aligned} \tag{9.4.81}$$

The following statements hold:

1. *For all $k, d \in \mathbb{N}$ and for every $s \in \mathbb{N}_{>0}$, $\rho^{\mathrm{spP\acute{o}l}}_{k,d} = \rho^{\mathrm{spP\acute{o}l}}_{k,d,1} \le \rho^{\mathrm{spP\acute{o}l}}_{k,d,s} \le f^\star$.*

2. *For every $s \in \mathbb{N}_{>0}$, $\sup\{\rho^{\mathrm{spP\acute{o}l}}_{k,d,s} : (k,d) \in \mathbb{N}^2\} = f^\star$.*

Proof. It is not hard to prove the first statement. The second one is due to the second statement of Theorem 9.10 and the inequalities $\rho^{\mathrm{spP\acute{o}l}}_{k,d} \le \rho^{\mathrm{spP\acute{o}l}}_{k,d,s} \le f^\star$. □

Based on the extension of Handelman's Positivstellensatz: Consider the hierarchy of linear programs indexed by $k \in \mathbb{N}$ and $s \in \mathbb{N}_{>0}$:

$$\begin{aligned}\tau^{\mathrm{spHan}}_{k,s} := \inf_{\mathbf{y}} \quad & L_{\mathbf{y}}(\check{f}) \\ \text{s.t.} \quad & \mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^n_{2k}} \subset \mathbb{R}\,,\ y_{\mathbf{0}} = 1\,, \\ & \mathbf{M}_{\mathcal{A}^{(s,k_{ij})}_{r,I_c}}((\check{g}_i\check{g}^j_{i_c})\mathbf{y}, I_c) \succeq 0\,, \\ & \qquad\qquad c \in [p]\,,\ i \in J_c\,,\ j \in \{0\}\cup[k-d_{g_i}]\,,\ r \in [b(n_c,k_{ij})]\,,\end{aligned} \tag{9.4.82}$$

where $k_{ij} := k - d_{g_i} - j$, for $i \in [m]$, for $j \in \{0\}\cup[k-d_{g_i}]$.

Theorem 9.13. *Let $f, g_i \in \mathbb{R}[\mathbf{x}]$, $i \in [m]$, with $g_m = 1$. Consider POP* (9.4.60) *with S being defined as in* (9.4.61)*. Let Assumption 9.2 hold. The dual of SDP* (9.4.82) *reads as:*

$$\begin{aligned}\rho^{\mathrm{spHan}}_{k,s} := \sup_{\lambda,\mathbf{G}^{(c)}_{ijr}} \quad & \lambda \\ \text{s.t.} \quad & \lambda \in \mathbb{R}\,,\ \mathbf{G}^{(c)}_{ijr} \succeq 0\,,\ c \in [p]\,,\ i \in J_c\,,\ j \in \{0\}\cup[k-d_{g_i}]\,,\ r \in [b(n_c,k_{ij})]\,, \\ & \check{f} - \lambda = \sum_{c\in[p]}\sum_{i\in J_c}\textstyle\sum_{j=0}^{k-d_{g_i}} \check{g}_i\check{g}^j_{i_c}\Big(\sum_{r\in[b(n_c,k_{ij})]} \mathbf{v}^{\top}_{\mathcal{A}^{(s,k_{ij})}_{r,I_c}}\mathbf{G}^{(c)}_{ijr}\mathbf{v}_{\mathcal{A}^{(s,k_{ij})}_{r,I_c}}\Big)\,.\end{aligned} \tag{9.4.83}$$

The following statements hold:

1. *For all $k \in \mathbb{N}$ and for every $s \in \mathbb{N}_{>0}$,*

$$\rho^{\mathrm{spHan}}_{k} = \rho^{\mathrm{spHan}}_{k,1} \le \rho^{\mathrm{spHan}}_{k,s} \le f^\star\,. \tag{9.4.84}$$

2. *For every $s \in \mathbb{N}_{>0}$, the sequence $(\rho^{\mathrm{spHan}}_{k,s})_{k\in\mathbb{N}}$ converges to $f^\star$.*

Proof. It is not hard to prove the first statement. The second one is due to the second statement of Theorem 9.11 and the inequalities $\rho^{\mathrm{spHan}}_{k} \le \rho^{\mathrm{spHan}}_{k,s} \le f^\star$. □

Obtaining an optimal solution

In other to extract an optimal solution of POP (9.4.60) with correlative sparsity, we first extract atoms on each clique similarly to Algorithm 14 and then connect them together to obtain atoms in $\mathbb{R}^n$. Explicitly, we use the following heuristic extraction algorithm:

Algorithm 15 Extraction algorithm for sparse POPs on the nonnegative orthant

Input: precision parameter $\varepsilon > 0$ and an optimal solution $(\lambda, \mathbf{u}_c, \mathbf{G}_{i,j}^{(c)})$ of SDP (9.4.81).
Output: an optimal solution $\mathbf{x}^\star$ of POP (9.4.60).

1: For $c \in [p]$, do:

a: For $j \in [b(n_c, k_m^{(d)})]$, let $\bar{\mathbf{G}}_j^{(c)} = (w_{\mathbf{pq}}^{(c,j)})_{\mathbf{p},\mathbf{q} \in \mathbb{N}^{I_c}_{k_m^{(d)}}}$ such that $(w_{\mathbf{pq}}^{(c,j)})_{\mathbf{p},\mathbf{q} \in \mathcal{A}_{j,I_c}^{(s,k_m^{(d)})}} = \mathbf{G}_{m,j}^{(c)}$ and $w_{\mathbf{pq}}^{(c,j)} = 0$ if $(\mathbf{p},\mathbf{q}) \notin (\mathcal{A}_{j,I_c}^{(s,k_m^{(d)})})^2$. Then $\bar{\mathbf{G}}_j^{(c)} \succeq 0$ and

$$\mathbf{v}^\top_{\mathbb{N}^{I_c}_{k_m^{(d)}}} \bar{\mathbf{G}}_j^{(c)} \mathbf{v}_{\mathbb{N}^{I_c}_{k_m^{(d)}}} = \mathbf{v}^\top_{\mathcal{A}_{j,I_c}^{(s,k_m^{(d)})}} \mathbf{G}_{m,j}^{(c)} \mathbf{v}_{\mathcal{A}_{j,I_c}^{(s,k_m^{(d)})}} ; \tag{9.4.85}$$

b: Let $\mathbf{G}^{(c)} := \sum_{j \in [b(n_c,k_m^{(d)})]} \bar{\mathbf{G}}_j^{(c)}$. Then $\mathbf{G}^{(c)}$ is the Gram matrix corresponding to $\sigma_{m,c}$ in the rational SOS decomposition

$$\check{f} - \lambda = \sum_{c \in [p]} \frac{\sum_{i \in J_c} \sigma_{i,c} \check{g}_i}{\theta_c^k} . \tag{9.4.86}$$

where each $\sigma_{i,c}$ is an SOS polynomial and $\check{g}_m = 1$;

c: Obtain an atom $\mathbf{z}^{\star(c)} \in \mathbb{R}^{n_c}$ by using the extraction algorithm of Henrion and Lasserre in [77], where the matrix $\mathbf{V}$ in [77, (6)] is taken such that the columns of $\mathbf{V}$ form a basis of the null space $\{\mathbf{u} \in \mathbb{R}^{\omega_k} : \mathbf{G}^{(c)}\mathbf{u} = 0\}$;

2: Let $\mathbf{z}^\star \in \mathbb{R}^n$ such that $\mathbf{z}^\star(I_c) = \mathbf{z}^{\star(c)}$, for $c \in [p]$.

3: If $\mathbf{z}^\star$ exists, verify that $\mathbf{z}^\star$ is an approximate optimal solution of POP (9.4.62) by checking the following inequalities:

$$|\check{f}(\mathbf{z}^\star) - \lambda| \le \varepsilon \|\check{f}\|_{\max} \text{ and } \check{g}_i(\mathbf{z}^\star) \ge -\varepsilon \|\check{g}_i\|_{\max} , \; i \in [m] , \tag{9.4.87}$$

where $\|q\|_{\max} := \max_{\boldsymbol{\alpha}} |q_{\boldsymbol{\alpha}}|$ for any $q \in \mathbb{R}[\mathbf{x}]$.

4: If the inequalities (9.4.87) hold, set $\mathbf{x}^\star := \mathbf{z}^{\star 2}$.

Table 9.10: Numerical results for positive maximal singular values.

Id	Pb	POP size		Put			Pól			
		m	n	k	val	time	k	s	val	time
1 2	1	4	16	1 2	47.48110 **30.18791**	0.02 16	0	17	**30.18791**	1
3 4	2	5	25	1 2	168.4450 **91.28158**	0.04 877	0	26	**91.28158**	0.7
5 6	3	6	36	1 2	4759.12 –	0.2 –	0	37	**2462.03**	0.9
7 8	4	7	49	1 2	1777.53 –	0.5 –	0	50	**970.202**	2

Id	Put				Pól			
	nmat	msize	nscal	naff	nmat	msize	nscal	naff
1 2	1 17	17 153	38 154	153 4845	1	17	138	153
3 4	1 26	26 351	27 352	351 23751	1	26	327	351
5 6	1 37	37 703	38 704	703 91390	1	37	668	703
7 8	1 50	50 1275	51 1276	1275 292825	1	50	1227	1275

9.4.7 Numerical experiments

In this section we report results of numerical experiments for random instances with the same settings and notation as in Section 9.3.

Positive maximal singular values

Test problems: We generate a matrix $\mathbf{M}$ as in [54, (12)]. Explicitly,

$$\mathbf{M} := \begin{bmatrix} \mathbf{D} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{CB} & \mathbf{D} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{CAB} & \mathbf{CB} & \mathbf{D} & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots & \dots \\ \mathbf{CA}^{m-2}\mathbf{B} & \mathbf{CA}^{m-3}\mathbf{B} & \mathbf{CA}^{m-4}\mathbf{B} & \dots & \mathbf{D} \end{bmatrix}, \tag{9.4.88}$$

where $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ are square matrices of size $r = m$. Every entry of $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ is taken uniformly in $(-1, 1)$. In order to compute the positive maximal singular value $\sigma_+(\mathbf{M})$ of $\mathbf{M}$, we solve the following POP on the nonnegative orthant:

$$\sigma_+(\mathbf{M})^2 = \max_{\mathbf{x} \in \mathbb{R}^n_+} \{\mathbf{x}^\top (\mathbf{M}^\top \mathbf{M}) \mathbf{x} \,:\, \|\mathbf{x}\|_2^2 = 1\}. \tag{9.4.89}$$

Note that $n = m \times r = m^2$.

The numerical results are displayed in Table 9.10. The columns of val show the approximations of $\sigma_+(\mathbf{M})^2$.

Discussion: The behavior of our method is similar to that in Section 9.3.1.

Stability number of a graph

Let us consider POP (9.3.8) which returns the stability number of a graph G.

Table 9.11: Numerical results for stability number of randomly generated graphs.

Id	Pb	POP size	Put			Pól			
		n	k	val	time	k	s	val	time
1	1	10	1	∞	0.01	0	11	**3.00000**	1
2			2	3.02305	0.6				
3	2	15	1	∞	0.01	0	16	**5.00000**	1
4			2	5.01898	10				
5	3	20	1	∞	0.02	1	21	**5.00001**	4
6			2	5.02951	119				
7	4	25	1	∞	0.04	1	26	**6.00000**	10
8			2	6.05801	1064				

Id	Put				Pól			
	nmat	msize	nscal	naff	nmat	msize	nscal	naff
1	1	11	12	66	1	11	67	66
2	11	66	67	1001				
3	1	16	17	136	1	16	137	136
4	16	136	137	3876				
5	1	21	22	231	21	21	1562	1771
6	21	231	232	10626				
7	1	26	27	351	26	26	2952	3276
8	26	351	352	23751				

Test problems: We generate the adjacency matrix $\mathbf{A} = (a_{ij})_{j,j\in[n]}$ of the graph G by the following steps:

1. Set $a_{ii} = 0$, for $i \in [n]$.
2. For $i \in [n]$, for $j \in \{1, \ldots, i-1\}$, let us select $a_{ij} = a_{ji}$ uniformly $\{0, 1\}$.

The numerical results are displayed in Table 9.11.

Note that the columns of val show the approximations of $\alpha(G)$.

Discussion: The behavior of our method is similar to that in Section 9.3.1. Note that the graphs from Tables 9.11 are dense so that we cannot exploit term sparsity or correlative sparsity for POP (9.3.8) in these cases. Moreover, for all graphs in Table 9.11, spPól provides the better bounds for $\alpha(G)$ compared to the ones returned by the second order relaxations of Put.

Remark 9.19. *In Pb 3, 4 of Table 9.11, Pól with $k = 1$ provides a better bound than Pól with $k = 0$. As shown in Remark 9.9, each SDP relaxation of Pól with $k = 0$ and sufficiently large s corresponds to an SDP relaxation obtained after exploiting term sparsity.*

Deciding the copositivity of a real symmetric matrix

Given a symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, we say that $\mathbf{A}$ is copositive if $\mathbf{u}^\top \mathbf{A}\mathbf{u} \geq 0$ for all $\mathbf{u} \in \mathbb{R}^n_+$. Consider the following POP:

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^n_+} \{\mathbf{x}^\top \mathbf{A}\mathbf{x} : \textstyle\sum_{j\in[n]} x_j = 1\}. \tag{9.4.90}$$

The matrix $\mathbf{A}$ is copositive iff $f^\star \geq 0$.

Test problems: We construct several instances of the matrix $\mathbf{A}$ as follows:

1. Take B_{ij} randomly in $(-1, 1)$ w.r.t. the uniform distribution, for all $i, j \in \{1, \ldots, n\}$.
2. Set $\mathbf{B} := (B_{ij})_{1\leq i,j\leq n}$ and $\mathbf{A} := \frac{1}{2}(\mathbf{B} + \mathbf{B}^\top)$.

The numerical results are displayed in Table 9.12.

Table 9.12: Numerical results for deciding the copositivity of a real symmetric matrix.

Id	Pb	POP size	Put			Pól			
		n	k	val	time	k	s	val	time
1	1	10	1	-1.45876	0.004	0	8	**-0.94862***	1
2			2	**-0.94862**	0.2				
3	2	15	1	-1.41319	0.007	0	13	**-0.65197***	1
4			2	**-0.65197**	10				
5	3	20	1	-1.40431	0.02	0	20	**-0.98026***	1
6			2	**-0.98026**	89				
7	4	25	1	-1.34450	0.03	0	19	**-0.97345***	2
8			2	**-0.97345**	519				

Id	Put				Pól			
	nmat	msize	nscal	naff	nmat	msize	nscal	naff
1	1	11	12	66	4	8	67	66
2	11	66	67	1001				
3	1	16	17	136	4	13	137	136
4	16	136	137	3876				
5	1	21	22	231	2	20	232	231
6	21	231	232	10626				
7	1	26	27	351	8	19	352	351
8	26	351	352	23751				

Discussion: The behavior of our method is similar to that in Section 9.3.1. In all cases, we can extract the solutions of the resulting POP and certify that $\mathbf{A}$ is not copositive since $f^\star$ is negative.

Deciding the nonnegativity of an even degree form on the nonegative orthant

Given a form $q \in \mathbb{R}[\mathbf{x}]$, q is nonnegative on $\mathbb{R}^n_+$ iff q is nonnegative on the unit simplex

$$\Delta := \{\mathbf{x} \in \mathbb{R}^n_+ \,:\, \textstyle\sum_{j\in[n]} x_j = 1\}\,. \tag{9.4.91}$$

Given a form $f \in \mathbb{R}[\mathbf{x}]$ of degree $2d$, we consider the following POP:

$$f^\star := \min_{\mathbf{x}\in\Delta} f(\mathbf{x})\,. \tag{9.4.92}$$

Note that if $d = 1$, problem (9.4.92) boils down to deciding the copositivity of the Gram matrix associated to f. Thus, we consider the case where $d \geq 2$.

Test problems: We construct several instances of the form f of degree $2d$ as follows:

1. Take $f_{\boldsymbol{\alpha}}$ randomly in $(-1, 1)$ w.r.t. the uniform distribution, for each $\boldsymbol{\alpha} \in \mathbb{N}^n$ with $|\boldsymbol{\alpha}| = 2d$.
2. Set $f := \sum_{|\boldsymbol{\alpha}|=2d} f_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}}$.

The numerical results are displayed in Table 9.13.

Discussion: The behavior of our method is similar to that in Section 9.3.1. In these cases, we were able to extract the solution of the resulting POPs. One can then conclude that f is not nonnegative on the nonnegative orthant since it has negative value at its atoms.

Minimizing a polynomial over the boolean hypercube

Consider the optimization problem:

$$\min_{\mathbf{x}\in\{0,1\}^n} f(\mathbf{x})\,, \tag{9.4.93}$$

Table 9.13: Numerical results for deciding the nonnegativity of an even degree form on the nonegative orthant, with $d = 2$.

Id	Pb	POP size	Put			Pól			
		n	k	val	time	k	s	val	time
1	1	5	2	-1.87958	0.001	0	8	**-0.68020***	1
2			3	**-0.68020**	0.06				
3	2	10	2	-1.87491	0.1	0	11	**-0.87524***	5
4			3	**-0.87524**	10				
5	3	15	2	-2.01566	6	0	44	**-0.86938***	79
6			3	**-0.86938**	7675				

Id	Put				Pól			
	nmat	msize	nscal	naff	nmat	msize	nscal	naff
1	6	21	22	126	31	6	72	126
2	6	56	232	462				
3	11	66	67	1001	111	11	617	1001
4	11	268	2212	8008				
5	16	136	137	3876	213	44	2637	3876
6	16	816	9317	54264				

where f is a polynomial of degree at most $2d$. It is equivalent to the following POP on the nonnegative orthant:

$$\min_{\mathbf{x}\in\mathbb{R}^n_+} \{f(\mathbf{x}) \,:\, x_j(1-x_j) = 0\,,\; j \in [n]\}\,, \tag{9.4.94}$$

Test problems: We construct several instances by taking the coefficients of f randomly in $(-1, 1)$ w.r.t. to the uniform distribution.

The numerical results are displayed in Table 9.14.

Discussion: The behavior of our method is similar to that in Section 9.3.1. Note that Pól with order $k = 0$ provides worse bounds than Put with order $k = 2$. However, as shown in Table 9.14, Pól with order $k = 1$ provides the same bounds as Put with order $k = 2$.

Table 9.14: Numerical results for minimizing polynomials over the boolean hypercube, with $d = 1$.

Id	Pb	POP size	Put			Pól			
		n	k	val	time	k	s	val	time
1	1	10	1	-4.61386	0.008	1	11	**-4.34345**	1
2			2	**-4.34345**	0.2				
3	1	20	1	-15.4584	0.02	1	21	**-14.9455**	4
4			2	**-14.9455**	108				
5	3	30	1	-29.3433	0.1	1	31	**-27.6311**	41
6			2	**-27.6311**	8068				

Id	Put				Pól			
	nmat	msize	nscal	naff	nmat	msize	nscal	naff
1	1	11	21	66	11	11	276	286
2	6	56	232	462				
3	1	21	41	231	21	21	1751	1771
4	21	231	4621	10626				
5	1	31	61	496	31	31	5426	5456
6	31	496	14881	46376				

Chapter 10

Conclusion and Perspectives

10.1 Achievements

10.1.1 General discussion

We have provided several methods that use positivity certificates with and without denominators, for solving POPs on compact and non-compact basic semi-algebraic sets.

On the one hand, we have exploited correlative-term sparsity and constant trace property for the Moment-SOS hierarchy based on Putinar's Positivstellensatz without denominator. On the other hand, we have relied on Putinar–Vasilescu's Positivstellensatz with uniform denominator to provide an appropriate Moment-SOS hierarchy for solving POPs on noncompact semialgebraic sets.

It is worth pointing out that correlative-term sparsity enables us to reduce the matrix sizes involved in the SDP relaxations, and hence to solve them efficiently with interior-point methods. Besides, one advantage of exploiting constant trace property is to benefit from a class of first-order methods to speed up the resolution of the SDP relaxations. The sparse setting for methods based on Reznick's Positivstellensatz with uniform denominators has been also considered. We emphasize that the same uniform denominator allows us to control the sizes of the SDP relaxations for POPs on the nonnegative orthant.

On the theoretical side, we have proved that methods with uniform denominators can have polynomial time complexity, similarly to methods without denominators. On a practical side, the efficiency of our methods has been illustrated on real-life problems arising from engineering, machine learning, and power networks.

10.1.2 Discussion and perspectives specific to each chapter

We list the detailed conclusions for each chapter:

- In Chapter 3 we have introduced the CS-TSSOS hierarchy, a sparse variant of the Moment-SOS hierarchy, to solve large-scale real-world nonlinear optimization problems whose input data are sparse polynomials. In addition to its theoretical convergence guarantees, CS-TSSOS allows one to make a trade-off between the quality of optimal values and the computational efficiency by controlling the types of chordal extensions and the sparse order k.

  By fully exploiting sparsity, CS-TSSOS allows one to go beyond Shor's relaxation and solve the second-order Moment-SOS relaxation associated with large-scale POPs to obtain more accurate bounds. Indeed CS-TSSOS can handle second-order relaxations of POP instances with thousands of variables and constraints on a standard laptop in few minutes. Such instances include the optimal power flow (OPF) problem, an important challenge in the management of electricity networks. In particular, our plan is to perform advanced numerical experiments on HPC cluster, for OPF instances with larger numbers of buses [56].

  Some additional further potential research directions are listed below:

1) The standard procedure of extracting optimal solutions for the dense Moment-SOS hierarchy does not apply to the CS-TSSOS hierarchy. It would be interesting to develop a procedure for extracting (approximate) solutions from partial information of moment matrices.

2) Recall that chordal extension plays an important role for both correlative and term sparsity patterns. It turns out that the size of the resulting maximal cliques is crucial for the overall computational efficiency of the CS-TSSOS hierarchy. So far, we have only considered *maximal* chordal extensions (for convergence guarantee) and approximately *smallest* chordal extensions. It would be worth investigating more general choices of chordal extensions.

3) The CS-TSSOS strategy could be adapted to other applications involving sparse polynomial problems, including deep learning [34].

4) At last but not least, a challenging research issue is to establish serious computationally cheaper alternatives to interior-point methods for solving SDP relaxations of POPs. The recent work [219] which reports spectacular results for standard SDPs (and Max-Cut problems in particular) is a positive sign in this direction. Such a computationally cheaper alternative is presented in Chapter 5, Appendix 5.3.

- In Chapter 4 we have provided a nonsmooth hierarchy of SDP relaxations for optimization of polynomials on varieties contained in a Euclidean sphere. The advantage of this hierarchy is to circumvent the hard constraints involved in the standard SDP hierarchy (4.1.6) by minimizing the maximal eigenvalue of a matrix pencil. This in turn boils down to solving an unconstrained convex nonsmooth optimization problem by LMBM and to computing largest eigenvalues by means of the modified Lanczos's algorithm. Our numerical experiments indicate that solving this nonsmooth hierarchy is more efficient and more robust than solving the classical semidefinite hierarchy by interior-point methods, at least for a class of interesting POPs, including equality constrained QCQPs on the sphere, QCQPs with a single inequality (ball) constraint, and minimization of quartics on the sphere. Our CTP framework can be further applied for an interesting class of noncommutative polynomial optimization problems [129], in particular for eigenvalue maximization problems arising from quantum information theory, where the variables are unitary operators [143].

  Eventually, we have tried to use spectral methods to solve SDP relaxations of QCQPs involving inequalities only, systems of polynomial equations, MAXCUT problems, 0/1 linear constrained quadratic problems and computation of stability numbers of graphs. However, our preliminary experiments for these problems have not been convincing in terms of efficiency and accuracy. In order to improve upon these results, one possible remedy would be to index the moment matrices by alternative Legendre/Chebychev bases, rather than with the standard monomial basis.

- In Chapter 5, we have proposed a general framework for exploiting the constant trace property in solving large-scale SDPs, typically SDP-relaxations arising from the Moment-SOS hierarchy for POPs with CTP. Extensive numerical experiments strongly suggest that with this CTP formulation, the CGAL solver based on first-order methods is more efficient and more scalable than `Mosek` without exploiting CTP, especially when the block size is large. In addition, the optimal value returned by CGAL is typically within 1% w.r.t. the one returned by `Mosek`.

  We have also integrated sparsity-exploiting techniques into the CTP-exploiting framework in order to handle large-scale POPs. For SDP-relaxations of large-scale POPs with a term and/or correlative sparsity pattern, and in applications for which only a medium accuracy of optimal solutions is enough, we believe that our framework should be very useful.

  As a topic of further investigation, we would like to improve the LP-based formulation for verifying CTP, for instance by relying on more general second-order cone programming. We also would like to generalize the CTP-exploiting framework to noncommutative POPs [29, 100, 210] which have attracted a lot of attention in the quantum information community. Another line of research would be to investigate whether CTP could be efficiently exploited by interior-point solvers.

- In Chapter 6, we have established two representations of (i) globally nonnegative polynomials and (ii) polynomials nonnegative on semialgebraic sets, based on the homogeneous representations of [175] and [170]. These representations have a distinguishing feature. They can be converted into a practical numerical scheme for approximating the global minimum, yielding converging appropriate hierarchies of semidefinite relaxations for unconstrained and constrained polynomial optimization problems.

  We have also introduced a new method (based on adding spherical constraints (ASC)) to solve systems of polynomial equalities and inequalities, and to obtain global solutions of polynomial optimization problems as well. In view of its practical efficiency, a topic of further investigation is to provide a more detailed comparison with other methods for solving polynomial systems.

- In Chapter 7 we have provided a new degree bound on the sum-of-squares (SOS) polynomials involved in Putinar–Vasilescu's Positivstellensatz. The resulting associated Moment-SOS hierarchy provides a sequence of lower bounds that converges to the global minimum (with prescribed accuracy $\varepsilon > 0$) at an $\mathcal{O}(\varepsilon^{-\mathfrak{c}})$ rate.

  A topic of further investigation is the analysis of the convergence rate of the Moment-SOS hierarchy for lower bounds in some special cases of basic (compact) semialgebraic sets. A first natural idea is to find the explicit constant α in the Łojasiewicz inequality stated in Lemma 7.2. We could then proceed analogously to the proof of the rate $\mathcal{O}(\varepsilon^{-65})$ for the minimization of a polynomial on the unit ball. It would be interesting to investigate whether our convergence analysis would be improved after replacing Bernstein approximations involved in our proof by Jackson kernels from [11, 145].

- In Chapter 8, we have provided:

  1) a sparse version for Reznick's Positivstellensatz (resp. Putinar–Vasilescu's Positivstellensatz) for positive definite forms (resp. nonnegative polynomials).

  2) a sparse version of Putinar–Vasilescu's Positivstellensatz for polynomials that are nonnegative on a possibly noncompact basic semialgebraic set.
  All these certificates involve sums of squares of rational functions with uniform denominators. For additional efficiency, our positivity certificates have been combined with appropriate sampling (evaluation) techniques (to impose that two polynomials are identical). The full computational benefit of such sampling techniques remains to be investigated.

- In Chapter 9 we have considered the case of dense POPs on the nonnegative orthant. By applying a positivity certificate involving SOS of monomials for a POP with input polynomials being even in each variable, one obtains a specific hierarchy of linear relaxations. Afterwards we replace each SOS of monomials by an SOS associated with a block-diagonal Gram matrix, where each block has a prescribed size. This ensures a practical efficiency of the corresponding hierarchy of SDP relaxations. Its convergence is still maintained with a $\mathcal{O}(\varepsilon^{-c})$ rate, similar to the one of Baldi and Mourrain [13].

  As a topic of further applications, we would like to use this methodology for solving large-scale POPs for phase retrieval and feedforward neural networks.

10.2 Additional future research directions

10.2.1 Deep Neural networks

Robustness certification. Evaluation and certification of robustness of Deep Neural Networks (DNNs) has become an important issue (DNNs), especially in view of certain of their applications. We refer the interested readers to [160, 31, 117, 15] for some recent research advances on DNNs. Figure 10.1 illustrates a neural network with two hidden layers.

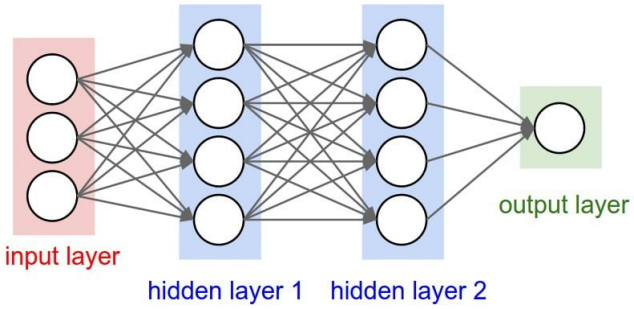


Figure 10.1: A two hidden layer neural network.

The widely used ReLU activation function ($x \mapsto \max(0, x)$) can be modeled by polynomial (in)equalities. As a result, robustness certification of such DNNs problem can be formulated as maximizing a linear function with linear and quadratic polynomial constraints, as described, e.g., in [172]. A promising research track is to take advantage of special structures of the objective and constraining polynomials to find a Positivstellensatz adapted to this situation. In Chapter 9, we have exploited nonnegativity of variables involved in the robustness certification problems. In the context of DNNs, the idea is to also exploit sparsity of the constraints which arise from their structure.

Matrix singular values and phase retrieval. We also intend to improve the approach of Chapter 9, to obtain tighter upper bounds for positive maximal singular values of a given matrix. Such bounds are very useful in the stability analysis of recurrent neural networks [55, 54]. Computing such singular values boils down to maximizing a quadratic polynomial over the intersection of the unit sphere and the nonnegative orthant. Another important application is phase retrieval, which can be similarly formulated as a (nonconvex) optimization problem involving quadratic polynomials, with several hidden structures. Another possibility is to develop a hierarchy of semidefinite relaxations similar to the one based on Pólya's Positivstellensatz in Chapter 9.

Finally, we also envision to adapt the method of chapters 4 and 5 that exploits the constant trace property to certify robustness of multi-layer neural networks, to compute positive maximal singular values, and to solve phase retrieval problems at global optimality.

10.2.2 Rates of convergence

An important challenge in polynomial optimization is to provide rates of convergence for various convex relaxations, and notably for the Moment-SOS hierarchy of semidefinite relaxations (and some variants described in this thesis). That is, the goal is to provide explicit bounds on the degrees of the polynomials involved in the various Positivstellensätze. We will particularly investigate (i) how different types of positive weights in the Positivstellensätze affect the convergence rate and (ii) if the corresponding hierarchies of semidefinite relaxations have finite convergence.

In addition to a theoretical interest in its own, analyzing the convergence rate also provides useful insights about the practical efficiency of the underlying semidefinite relaxations. A higher convergence rate is likely to yield a smaller computational effort to approximate closely the optimal value. Moreover, if the hierarchy has finite convergence, then the optimal value can be computed exactly.

One possibility is to follow the methodology that we have used to obtain the convergence rates of Putinar–Vasilescu's and Dickinson–Povh's Positivstellensatz in chapters 7 and 9, respectively, and which proceeds in two steps: (i) provide a constructive existence proof for the Positivstellensatz, and (ii) provide degree bounds of its explicit SOS weights.

Another possibility is to use Fritz–John optimality conditions as in our recent works [127, 128, 126]. Developed directly from the works of Nie–Demmel–Sturmfels [149] and Demmel–Nie–Powers [44], this method allows us to guarantee finite convergence and to compute global minimizers of POPs even when the Karush–Kuhn–Tucker conditions no longer hold.

Bibliography

[1] F. Acquistapace, C. Andradas, and F. Broglia. The positivstellensatz for definable functions on o-minimal structures. *Illinois Journal of Mathematics*, 46(3):685–693, 2002.

[2] J. Agler, W. Helton, S. McCullough, and L. Rodman. Positive semidefinite matrices with a given sparsity pattern. *Linear algebra and its applications*, 107:101–149, 1988.

[3] A. A. Ahmadi and A. Majumdar. DSOS and SDSOS optimization: more tractable alternatives to sum of squares and semidefinite optimization. *SIAM Journal on Applied Algebra and Geometry*, 3(2):193–230, 2019.

[4] E. D. Andersen and K. D. Andersen. The MOSEK interior point optimizer for linear programming: an implementation of the homogeneous algorithm. In *High performance optimization*, pages 197–232. Springer, 2000.

[5] M. ApS. Mosek optimization toolbox for matlab. *User's Guide and Reference Manual, Version*, 4, 2019.

[6] M. ApS. *The MOSEK optimization toolbox. Version 9.1.*, 2019.

[7] E. Artin. Über die zerlegung definiter funktionen in quadrate. *Abhandlungen aus dem mathematischen Seminar der Universität Hamburg*, 5(1):100–115, 1927.

[8] G. Averkov. Constructive proofs of some positivstellensätze for compact semialgebraic subsets of $\mathbb{R}^d$. *Journal of Optimization Theory and Applications*, 158(2):410–418, 2013.

[9] G. Averkov. Optimal size of linear matrix inequalities in semidefinite approaches to polynomial optimization. *SIAM Journal on Applied Algebra and Geometry*, 3(1):128–151, 2019.

[10] S. Babaeinejadsarookolaee, A. Birchfield, R. D. Christie, C. Coffrin, C. DeMarco, R. Diao, M. Ferris, S. Fliscounakis, S. Greene, R. Huang, et al. The power grid library for benchmarking ac optimal power flow algorithms. *arXiv preprint arXiv:1908.02788*, 2019.

[11] T. Bagby, L. Bos, and N. Levenberg. Multivariate simultaneous approximation. *Constructive approximation*, 18(4):569–577, 2002.

[12] T. Bajbar and O. Stein. Coercive polynomials and their Newton polytopes. *SIAM Journal on Optimization*, 25(3):1542–1570, 2015.

[13] L. Baldi and B. Mourrain. On moment approximation and the effective putinar's positivstellensatz. *arXiv preprint arXiv:2111.11258*, 2021.

[14] A. Barvinok. *A course in convexity*, volume 54. American Mathematical Soc., 2002.

[15] C. Baykal, L. Liebenwein, I. Gilitschenski, D. Feldman, and D. Rus. Sensitivity-informed provable pruning of neural networks. *SIAM Journal on Mathematics of Data Science*, 4(1):26–45, 2022.

[16] A. Ben-Tal and A. Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, volume 2. Siam, 2001.

[17] R. Berr and T. Wörmann. Positive polynomials on compact sets. *manuscripta mathematica*, 104(2):135–143, 2001.

[18] J. R. Blair and B. Peyton. An introduction to chordal graphs and clique trees. In *Graph theory and sparse matrix computation*, pages 1–29. Springer, 1993.

[19] G. Blekherman. There are significantly more nonegative polynomials than sums of squares. *Israel Journal of Mathematics*, 153(1):355–380, 2006.

[20] G. Blekherman, J. Gouveia, and J. Pfeiffer. Sums of squares on the hypercube. *Mathematische Zeitschrift*, 284(1-2):41–54, 2016.

[21] G. Blekherman, G. G. Smith, and M. Velasco. Sharp degree bounds for sum-of-squares certificates on projective curves. *Journal de Mathématiques Pures et Appliquées*, 129:61–86, 2019.

[22] J. Bochnak, M. Coste, and M.-F. Roy. *Real algebraic geometry*, volume 36. Springer Science & Business Media, 2013.

[23] E. G. Boman, D. Chen, O. Parekh, and S. Toledo. On factor width and symmetric H-matrices. *Linear algebra and its applications*, 405:239–248, 2005.

[24] S. Boyd, N. Parikh, and E. Chu. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

[25] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[26] D. Brosch and E. de Klerk. Jordan symmetry reduction for conic optimization over the doubly nonnegative cone: theory and software. *arXiv preprint arXiv:2001.11348*, 2020.

[27] S. Burer and R. D. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

[28] S. Burer and R. D. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005.

[29] S. Burgdorf, I. Klep, and J. Povh. *Optimization of polynomials in non-commuting variables*, volume 2. Springer, 2016.

[30] J. V. Burke, A. S. Lewis, and M. L. Overton. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM Journal on Optimization*, 15(3):751–779, 2005.

[31] A. Caragea, D. G. Lee, J. Maly, G. Pfander, and F. Voigtlaender. Quantitative approximation results for complex-valued neural networks. *SIAM Journal on Mathematics of Data Science*, 4(2):553–580, 2022.

[32] Y. Carmon and J. C. Duchi. First-order methods for nonconvex quadratic minimization. *SIAM Review*, 62(2):395–436, 2020.

[33] V. Chandrasekaran and P. Shah. Relative entropy relaxations for signomial optimization. *SIAM Journal on Optimization*, 26(2):1147–1173, 2016.

[34] T. Chen, J. B. Lasserre, V. Magron, and E. Pauwels. Semialgebraic optimization for lipschitz constants of relu networks. *Advances in Neural Information Processing Systems*, 33, 2020.

[35] T. Chen, J.-B. Lasserre, V. Magron, and E. Pauwels. A sublevel Moment-SOS hierarchy for polynomial optimization. *Computational Optimization and Applications*, 2021. Accepted for publication.

[36] T. Chen, J. B. Lasserre, V. Magron, and E. Pauwels. Semialgebraic representation of monotone deep equilibrium models and applications to certification. *Advances in Neural Information Processing Systems*, 34, 2021.

[37] M.-D. Choi, T.-Y. Lam, and B. Reznick. Even symmetric sextics. *Mathematische Zeitschrift*, 195(4):559–580, 1987.

[38] C. W. Commander. Maximum Cut Problem, MAX-CUT. *Encyclopedia of Optimization*, 2, 2009.

[39] F. E. Curtis and X. Que. A quasi-Newton algorithm for nonconvex, nonsmooth optimization with global convergence guarantees. *Mathematical Programming Computation*, 7(4):399–428, 2015.

[40] R. E. Curto and L. A. Fialkow. Truncated K-moment problems in several variables. *Journal of Operator Theory*, pages 189–226, 2005.

[41] J. Dahl. Semidefinite optimization using mosek. *ISMP, Berlin*, 2012.

[42] A. d'Aspremont and N. El Karoui. A stochastic smoothing algorithm for semidefinite programming. *SIAM Journal on Optimization*, 24(3):1138–1177, 2014.

[43] C. N. Delzell. Continous, piecewise-polynomial functions which solve Hilbert's 17th problem. *Journal für die Reine und Angewandte Mathematik*, 1993(440):157–174, 1993.

[44] J. Demmel, J. Nie, and V. Powers. Representations of positive polynomials on noncompact semialgebraic sets via KKT ideals. *Journal of pure and applied algebra*, 209(1):189–200, 2007.

[45] P. J. Dickinson and J. Povh. On an extension of Pólya's Positivstellensatz. *Journal of global optimization*, 61(4):615–625, 2015.

[46] P. J. Dickinson and J. Povh. A new approximation hierarchy for polynomial conic optimization. *Computational Optimization and Applications*, 73(1):37–67, 2019.

[47] L. Ding and B. Grimmer. Revisit of spectral bundle methods: Primal-dual (sub) linear convergence rates. *arXiv preprint arXiv:2008.07067*, 2020.

[48] L. Ding, A. Yurtsever, V. Cevher, J. A. Tropp, and M. Udell. An optimal-storage approach to semidefinite programming using approximate complementarity. *SIAM Journal on Optimization*, 31(4):2695–2725, 2021.

[49] T. L. Dinh and N. H. A. Mai. Comparing different subgradient methods for solving convex optimization problems with functional constraints. *arXiv preprint arXiv:2101.01045*, 2021.

[50] D. Djukić, V. Janković, I. Matić, and N. Petrović. *The IMO Compendium: A Collection of Problems Suggested for the International Mathematical Olympiads: 1959-2009 Second Edition*. Springer Science & Business Media, 2011.

[51] M. Dressler, S. Iliman, and T. De Wolff. A positivstellensatz for sums of nonnegative circuit polynomials. *SIAM Journal on Applied Algebra and Geometry*, 1(1):536–555, 2017.

[52] M. Dressler, S. Iliman, and T. De Wolff. An approach to constrained polynomial optimization via nonnegative circuit polynomials and geometric programming. *Journal of Symbolic Computation*, 91:149–172, 2019.

[53] I. Dunning, J. Huchette, and M. Lubin. Jump: A modeling language for mathematical optimization. *SIAM Review*, 59(2):295–320, 2017.

[54] Y. Ebihara, H. Waki, V. Magron, N. H. A. Mai, D. Peaucelle, and S. Tarbouriech. l2 induced norm analysis of discrete-time LTI systems for nonnegative input signals and its application to stability analysis of recurrent neural networks. *European Journal of Control*, 62:99–104, 2021.

[55] Y. Ebihara, H. Waki, V. Magron, N. H. A. Mai, D. Peaucelle, and S. Tarbouriech. Stability Analysis of Recurrent Neural Networks by IQC with Copositive Mutipliers. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 5098–5103. IEEE, 2021.

[56] A. Eltved, J. Dahl, and M. S. Andersen. On the robustness and scalability of semidefinite relaxation for optimal power flow problems. *Optimization and Engineering*, 21(2):375–392, 2020.

[57] K. Fang and H. Fawzi. The sum-of-squares hierarchy on the sphere and applications in quantum information theory. *Mathematical Programming*, pages 1–30, 2020.

[58] H. Fawzi, J. Saunderson, and P. A. Parrilo. Sparse sums of squares on finite abelian groups and improved semidefinite lifts. *Mathematical Programming*, 160(1):149–191, 2016.

[59] D. Fulkerson and O. Gross. Incidence matrices and interval graphs. *Pacific journal of mathematics*, 15(3):835–855, 1965.

[60] M. Garstka, M. Cannon, and P. Goulart. COSMO: A conic operator splitting method for convex conic problems. *Journal of Optimization Theory and Applications*, 190(3):779–810, 2021.

[61] K. Gatermann and P. A. Parrilo. Symmetry groups, semidefinite programs, and sums of squares. *Journal of Pure and Applied Algebra*, 192(1-3):95–128, 2004.

[62] S. B. Gershwin. KKT Examples. *min J*, 10:2, 2010.

[63] H. Godard, S. Elloumi, A. Lambert, J. Maeght, and M. Ruiz. Novel approach towards global optimality of optimal power flow using quadratic convex optimization. In *2019 6th International Conference on Control, Decision and Information Technologies (CoDIT)*, pages 1227–1232. IEEE, 2019.

[64] C. Goel, S. Kuhlmann, and B. Reznick. The analogue of Hilbert's 1888 theorem for even symmetric forms. *Journal of Pure and Applied Algebra*, 221(6):1438–1448, 2017.

[65] M. C. Golumbic. *Algorithmic graph theory and perfect graphs*. Elsevier, 2004.

[66] J. Gouveia, A. Kovacec, and M. Saee. On sums of squares of k-nomials. *Journal of Pure and Applied Algebra*, 226(1):106820, 2022.

[67] A. Greuet and M. Safey El Din. Probabilistic algorithm for polynomial optimization over a real algebraic set. *SIAM Journal on Optimization*, 24(3):1313–1343, 2014.

[68] D. Grimm, T. Netzer, and M. Schweighofer. A note on the representation of positive polynomials with structured sparsity. *Archiv der Mathematik*, 89(5):399–403, 2007.

[69] R. Grone, C. R. Johnson, E. M. Sá, and H. Wolkowicz. Positive definite completions of partial hermitian matrices. *Linear algebra and its applications*, 58:109–124, 1984.

[70] M. Haarala, K. Miettinen, and M. M. Mäkelä. New limited memory bundle method for large-scale nonsmooth optimization. *Optimization Methods and Software*, 19(6):673–692, 2004.

[71] N. Haarala, K. Miettinen, and M. M. Mäkelä. Globally convergent limited memory bundle method for large-scale nonsmooth optimization. *Mathematical Programming*, 109(1):181–205, 2007.

[72] W. R. Harris. Real even symmetric ternary forms. *Journal of Algebra*, 222(1):204–245, 1999.

[73] C. Heitzinger. *Simulation and inverse modeling of semiconductor manufacturing processes*. na, 2002.

[74] C. Helmberg, M. L. Overton, and F. Rendl. The spectral bundle method with second-order information. *Optimization Methods and Software*, 29(4):855–876, 2014.

[75] C. Helmberg and F. Rendl. A spectral bundle method for semidefinite programming. *SIAM Journal on Optimization*, 10(3):673–696, 2000.

[76] C. Helmberg, F. Rendl, R. J. Vanderbei, and H. Wolkowicz. An interior-point method for semidefinite programming. *SIAM Journal on Optimization*, 6(2):342–361, 1996.

[77] D. Henrion and J.-B. Lasserre. Detecting global optimality and extracting solutions in GloptiPoly. In *Positive polynomials in control*, pages 293–310. Springer, 2005.

[78] D. Henrion and J. Malick. Projection methods in conic optimization. In *Handbook on Semidefinite, Conic and Polynomial Optimization*, pages 565–600. Springer, 2012.

[79] D. Hilbert. Über die Darstellung Definiter Formen als Summe von Formenquadraten. *Mathematische Annalen*, 32(3):342–350, 1888.

[80] T. Hildebrandt and I. Schoenberg. On linear functional operations and the moment problem for a finite interval in one or several dimensions. *Annals of Mathematics*, pages 317–328, 1933.

[81] M. Hong and Z.-Q. Luo. On the linear convergence of the alternating direction method of multipliers. *Mathematical Programming*, 162(1-2):165–199, 2017.

[82] S. Hu, G. Li, and L. Qi. A tensor analogy of Yuan's theorem of the alternative and polynomial optimization with sign structure. *Journal of Optimization Theory and Applications*, 168(2):446–474, 2016.

[83] L. Huang, J. Nie, and Y.-X. Yuan. Homogenization for polynomial optimization with unbounded sets. *arXiv preprint arXiv:2112.15289*, 2021.

[84] S. Iliman and T. De Wolff. Amoebas, nonnegative polynomials and sums of squares supported on circuits. *Research in the Mathematical Sciences*, 3(1):1–35, 2016.

[85] T. Jacobi. A representation theorem for certain partially ordered commutative rings. *Mathematische Zeitschrift*, 237(2):259–273, 2001.

[86] T. Jacobi and A. Prestel. Distinguished representations of strictly positive polynomials. *Journal für die Reine und Angewandte Mathematik*, 2001.

[87] V. Jeyakumar, S. Kim, G. M. Lee, and G. Li. Solving global optimization problems with sparse polynomials and unbounded semialgebraic feasible sets. *J. Glob. Optim*, 65(2):175–190, 2016.

[88] V. Jeyakumar, J. B. Lasserre, and G. Li. On polynomial optimization over non-compact semi-algebraic sets. *Journal of Optimization Theory and Applications*, 163(3):707–718, 2014.

[89] D. Jibetean and M. Laurent. Semidefinite approximations for global unconstrained polynomial optimization. *SIAM Journal on Optimization*, 16(2):490–514, 2005.

[90] C. Josz, S. Fliscounakis, J. Maeght, and P. Panciatici. AC power flow data in MATPOWER and QCQP format: iTesla, RTE snapshots, and PEGASE. *arXiv preprint arXiv:1603.01533*, 2016.

[91] C. Josz and D. Henrion. Strong duality in Lasserre's hierarchy for polynomial optimization. *Optimization Letters*, 10(1):3–10, 2016.

[92] C. Josz and D. K. Molzahn. Lasserre hierarchy for large scale polynomial optimization in real and complex variables. *SIAM Journal on Optimization*, 28(2):1017–1048, 2018.

[93] M. Journée, F. Bach, P.-A. Absil, and R. Sepulchre. Low-rank optimization for semidefinite convex problems. *arXiv preprint arXiv:0807.4423*, 2008.

[94] S. Jukna. *Boolean function complexity: advances and frontiers*, volume 27. Springer Science & Business Media, 2012.

[95] N. Karmitsa. LMBM–FORTRAN subroutines for Large-Scale nonsmooth minimization: User's manual'. *TUCS Technical Report*, 77(856), 2007.

[96] M. Kirszbraun. Über die zusammenziehende und Lipschitzsche Transformationen. *Fundamenta Mathematicae*, 22(1):77–108, 1934.

[97] K. C. Kiwiel. Proximity control in bundle methods for convex nondifferentiable minimization. *Mathematical programming*, 46(1-3):105–122, 1990.

[98] K. C. Kiwiel. A tilted cutting plane proximal bundle method for convex nondifferentiable optimization. *Operations research letters*, 10(2):75–81, 1991.

[99] K. C. Kiwiel. Convergence of the gradient sampling algorithm for nonsmooth nonconvex optimization. *SIAM Journal on Optimization*, 18(2):379–388, 2007.

[100] I. Klep, V. Magron, and J. Povh. Sparse noncommutative polynomial optimization. *Mathematical Programming*, pages 1–41, 2021.

[101] J.-L. Krivine. Anneaux préordonnés. *Journal d'analyse mathématique*, 12(1):307–326, 1964.

[102] J. B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on optimization*, 11(3):796–817, 2001.

[103] J. B. Lasserre. Convergent SDP-relaxations in polynomial optimization with sparsity. *SIAM Journal on Optimization*, 17(3):822–843, 2006.

[104] J.-B. Lasserre. *Moments, positive polynomials and their applications*, volume 1. World Scientific, 2010.

[105] J. B. Lasserre. *An introduction to polynomial and semi-algebraic optimization*, volume 52. Cambridge University Press, 2015.

[106] J. B. Lasserre. A max-cut formulation of 0/1 programs. *Operations Research Letters*, 44(2):158–164, 2016.

[107] J. B. Lasserre. Homogeneous polynomials and spurious local minima on the unit sphere. *Optimization Letters*, pages 1–14, 2021.

[108] J. B. Lasserre, M. Laurent, and P. Rostalski. Semidefinite characterization and computation of zero-dimensional real radical ideals. *Foundations of Computational Mathematics*, 8(5):607–647, 2008.

[109] J. B. Lasserre, K.-C. Toh, and S. Yang. A bounded degree SOS hierarchy for polynomial optimization. *EURO Journal on Computational Optimization*, 5(1-2):87–117, 2017.

[110] M. Laurent. A comparison of the sherali-adams, lovász-schrijver, and lasserre relaxations for 0–1 programming. *Mathematics of Operations Research*, 28(3):470–496, 2003.

[111] M. Laurent. Revisiting two theorems of Curto and Fialkow on moment matrices. *Proceedings of the American Mathematical Society*, 133(10):2965–2976, 2005.

[112] M. Laurent and L. Slot. An effective version of Schmüdgen's Positivstellensatz for the hypercube. *arXiv preprint arXiv:2109.09528*, 2021.

[113] J. Lee, V. Balakrishnan, C.-K. Koh, and D. Jiao. From $O(k^2N)$ to $O(N)$: A fast complex-valued eigenvalue solver for large-scale on-chip interconnect analysis. In *2009 IEEE MTT-S International Microwave Symposium Digest*, pages 181–184. IEEE, 2009.

[114] D. Leep and C. Starr. Polynomials in R[X,Y] that are sums of squares in R(X,Y). *Proceedings of the American Mathematical Society*, 129(11):3133–3141, 2001.

[115] R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK users' guide: solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*. SIAM, 1998.

[116] A. S. Lewis and M. L. Overton. Nonsmooth optimization via quasi-Newton methods. *Mathematical Programming*, 141(1-2):135–163, 2013.

[117] Y. Li, T. Luo, and C. Ma. Nonlinear Weighted Directed Acyclic Graph and A Priori Estimates for Neural Networks. *SIAM Journal on Mathematics of Data Science*, 4(2):694–720, 2022.

[118] J. Lofberg and P. A. Parrilo. From coefficients to samples: a new approach to sos optimization. In *2004 43rd IEEE Conference on Decision and Control (CDC)(IEEE Cat. No. 04CH37601)*, volume 3, pages 3154–3159. IEEE, 2004.

[119] H. Lombardi, D. Perrucci, and M.-F. Roy. *An elementary recursive bound for effective Positivstellensatz and Hilbert's 17th problem*, volume 263. American mathematical society, 2020.

[120] V. Magron. Interval enclosures of upper bounds of roundoff errors using semidefinite programming. *ACM Transactions on Mathematical Software (TOMS)*, 44(4):1–18, 2018.

[121] V. Magron, G. Constantinides, and A. Donaldson. Certified roundoff error bounds using semidefinite programming. *ACM Transactions on Mathematical Software (TOMS)*, 43(4):1–31, 2017.

[122] V. Magron, N. H. A. Mai, Y. Ebihara, and H. Waki. Tractable semidefinite bounds of positive maximal singular values. *arXiv preprint arXiv:2202.08731*, 2022.

[123] V. Magron and J. Wang. SONC Optimization and Exact Nonnegativity Certificates via Second-Order Cone Programming. *Journal of Symbolic Computation*, 2021. Accepted for publication.

[124] V. Magron and J. Wang. TSSOS: a Julia library to exploit sparsity for large-scale polynomial optimization. *arXiv preprint arXiv:2103.00915*, 2021.

[125] N. H. A. Mai. A symbolic algorithm for exact polynomial optimization strengthened with Fritz John conditions. *arXiv preprint arXiv:2206.02643*, 2022.

[126] N. H. A. Mai. Complexity for exact polynomial optimization strengthened with Fritz John conditions. *arXiv preprint arXiv:2205.11797*, 2022.

[127] N. H. A. Mai. Exact polynomial optimization strengthened with Fritz John conditions. *arXiv preprint arXiv:2205.04254*, 2022.

[128] N. H. A. Mai. On the exactness for polynomial optimization strengthened with Fritz John conditions. *arXiv preprint arXiv:2205.08450*, 2022.

[129] N. H. A. Mai, A. Bhardwaj, and V. Magron. The constant trace property in noncommutative optimization. In *Proceedings of the 2021 on International Symposium on Symbolic and Algebraic Computation*, pages 297–304, 2021.

[130] N. H. A. Mai, J.-B. Lasserre, and V. Magron. Positivity certificates and polynomial optimization on non-compact semialgebraic sets. *Mathematical Programming*, pages 1–43, 2021.

[131] N. H. A. Mai, J.-B. Lasserre, V. Magron, and J. Wang. Exploiting constant trace property in large-scale polynomial optimization. *arXiv preprint arXiv:2012.08873*, 2020.

[132] N. H. A. Mai and V. Magron. On the complexity of Putinar–Vasilescu's Positivstellensatz. *Journal of Complexity*, page 101663, 2022.

[133] N. H. A. Mai, V. Magron, and J. Lasserre. A sparse version of Reznick's Positivstellensatz. *Mathematics of Operations Research*, 2022.

[134] N. H. A. Mai, V. Magron, and J.-B. Lasserre. A hierarchy of spectral relaxations for polynomial optimization. *arXiv preprint arXiv:2007.09027*, 2020.

[135] N. H. A. Mai, V. Magron, J.-B. Lasserre, and K.-C. Toh. Tractable hierarchies of convex relaxations for polynomial optimization on the nonnegative orthant. *Forthcoming*, 2021.

[136] M. Marshall. Extending the Archimedean Positivstellensatz to the non-compact case. *Canadian Mathematical Bulletin*, 44(2):223–230, 2001.

[137] M. Marshall. A general representation theorem for partially ordered commutative rings. *Mathematische Zeitschrift*, 242(2):217–225, 2002.

[138] M. Marshall. Approximating positive polynomials using sums of squares. *Canadian Mathematical Bulletin*, 46(3):400–418, 2003.

[139] M. Marshall. Representations of non-negative polynomials, degree bounds and applications to optimization. *Canadian Journal of Mathematics*, 61(1):205–221, 2009.

[140] A. Megretski. Systems polynomial optimization tools (SPOT). *Massachusetts Inst. Technol., Cambridge, MA, USA*, 2010.

[141] J. Miller, Y. Zheng, M. Sznaier, and A. Papachristodoulou. Decomposed structured subsets for semidefinite and sum-of-squares optimization. *Automatica*, 137:110125, 2022.

[142] R. Murray, V. Chandrasekaran, and A. Wierman. Signomial and polynomial optimization via relative entropy and partial dualization. *Mathematical Programming Computation*, 13(2):257–295, 2021.

[143] M. Navascués, S. Pironio, and A. Acín. A convergent hierarchy of semidefinite programs characterizing the set of quantum correlations. *New J. Phys.*, 10(7):073013, 2008.

[144] A. Nemirovsky and D. Yudin. Problem complexity and method efficiency in optimization. *Nauka (published in English by John Wiley, Chichester, 1983)*, 1983.

[145] D. Newman and H. Shapiro. Jackson's theorem in higher dimensions. In *On Approximation Theory/Über Approximationstheorie*, pages 208–219. Springer, 1964.

[146] J. Nie. Discriminants and nonnegative polynomials. *Journal of Symbolic Computation*, 47(2):167–191, 2012.

[147] J. Nie. Optimality conditions and finite convergence of Lasserre's hierarchy. *Mathematical programming*, 146(1-2):97–121, 2014.

[148] J. Nie. Tight relaxations for polynomial optimization and Lagrange multiplier expressions. *Mathematical Programming*, 178(1-2):1–37, 2019.

[149] J. Nie, J. Demmel, and B. Sturmfels. Minimizing polynomials via sum of squares over the gradient ideal. *Mathematical programming*, 106(3):587–606, 2006.

[150] J. Nie and M. Schweighofer. On the complexity of Putinar's Positivstellensatz. *Journal of Complexity*, 23(1):135–150, 2007.

[151] J. Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of computation*, 35(151):773–782, 1980.

[152] J. Nocedal and S. Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[153] P. J. Olver. On multivariate interpolation. *Studies in Applied Mathematics*, 116(2):201–240, 2006.

[154] A. Oustry, M. Tacchi, and D. Henrion. Inner approximations of the maximal positively invariant set for polynomial dynamical systems. *IEEE Control Systems Letters*, 3(3):733–738, 2019.

[155] M. L. Overton and R. S. Womersley. Second derivatives for optimizing eigenvalues of symmetric matrices. *SIAM Journal on Matrix Analysis and Applications*, 16(3):697–718, 1995.

[156] B. O'donoghue, E. Chu, N. Parikh, and S. Boyd. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068, 2016.

[157] P. A. Parrilo. *Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization*. PhD thesis, California Inst. Tech., 2000.

[158] D. Papp and S. Yildiz. Sum-of-squares optimization without semidefinite programming. *SIAM Journal on Optimization*, 29(1):822–851, 2019.

[159] D. Papp and S. Yıldız. alfonso: Matlab package for nonsymmetric conic optimization. *INFORMS Journal on Computing*, 2021.

[160] R. Parhi and R. D. Nowak. What kinds of functions do deep neural networks learn? Insights from variational spline theory. *SIAM Journal on Mathematics of Data Science*, 4(2):464–489, 2022.

[161] P. A. Parrilo. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. PhD thesis, California Institute of Technology, 2000.

[162] J. Pena, J. C. Vera, and L. F. Zuluaga. Completely positive reformulations for polynomial optimization. *Mathematical Programming*, 151(2):405–431, 2015.

[163] L. Perko. *Differential equations and dynamical systems*, volume 7. Springer Science & Business Media, 2013.

[164] T. S. Pham and H. H. Vui. *Genericity in polynomial optimization*, volume 3. World Scientific, 2016.

[165] G. Pólya. Über Positive Darstellung von Polynomen. *Vierteljschr. Naturforsch. Ges. Zürich*, 73:141–145, 1928.

[166] G. Pólya. Über positive darstellung von polynomen, vierteljahresschrift der naturforschenden gesellschaft in zürich 73 (1928), 141–145, reprinted in: Collected papers, volume 2, 309–313, 1974.

[167] V. Powers and B. Reznick. A new bound for Pólya's Theorem with applications to polynomials positive on polyhedra. *Journal of pure and applied algebra*, 164(1-2):221–229, 2001.

[168] M. Putinar. Positive polynomials on compact semi-algebraic sets. *Indiana University Mathematics Journal*, 42(3):969–984, 1993.

[169] M. Putinar and F.-H. Vasilescu. Positive polynomials on semi-algebraic sets. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 328(7):585–589, 1999.

[170] M. Putinar and F.-H. Vasilescu. Solving moment problems by dimensional extension. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 328(6):495–499, 1999.

[171] M. L. Quijorna. An experimental approach for global polynomial optimization based on moments and semidefinite programming. *arXiv preprint arXiv:1809.09043*, 2018.

[172] A. Raghunathan, J. Steinhardt, and P. S. Liang. Semidefinite relaxations for certifying robustness to adversarial examples. *Advances in Neural Information Processing Systems*, 31, 2018.

[173] G. Reinelt. TSPLIB—A traveling salesman problem library. *ORSA journal on computing*, 3(4):376–384, 1991.

[174] B. Reznick. Extremal psd forms with few terms. *Duke mathematical journal*, 45(2):363–374, 1978.

[175] B. Reznick. Uniform denominators in Hilbert's seventeenth problem. *Mathematische Zeitschrift*, 220(1):75–97, 1995.

[176] B. Reznick. Some concrete aspects of hilbert's 17th problem. *Contemporary mathematics*, 253:251–272, 2000.

[177] B. Reznick. On the absence of uniform denominators in Hilbert's 17th problem. *Proceedings of the American Mathematical Society*, 133(10):2829–2834, 2005.

[178] C. Riener, T. Theobald, L. J. Andrén, and J. B. Lasserre. Exploiting symmetries in SDP-relaxations for polynomial optimization. *Mathematics of Operations Research*, 38(1):122–141, 2013.

[179] L. M. Roebers, J. C. Vera, and L. F. Zuluaga. Sparse non-SOS Putinar-type Positivstellensätze. *arXiv preprint arXiv:2110.10079*, 2021.

[180] R. A. Rossi and N. K. Ahmed. The network data repository with interactive graph analytics and visualization. In *AAAI*, 2015.

[181] Y. Saad. *Numerical methods for large eigenvalue problems: revised edition.* SIAM, 2011.

[182] R. Schabert. Uniform denominators in Hilbert's 17th Problem Theorems by Pólya and Reznick. *Manuscript*, http://www.math.uni-konstanz.de/ infusino/FachseminarMA-SS19/Schabert-Handout.pdf, 2019.

[183] C. Schlosser and M. Korda. Sparse moment-sum-of-squares relaxations for nonlinear dynamical systems with guaranteed convergence. *arXiv preprint arXiv:2012.05572*, 2020.

[184] C. Schlosser and M. Korda. Converging outer approximations to global attractors using semidefinite programming. *Automatica*, 134:109900, 2021.

[185] K. Schmüdgen. The K-moment problem for compact semi-algebraic sets. *Mathematische Annalen*, 289(1):203–206, 1991.

[186] C. Schulze. Schmüdgen's theorem and results of positivity. *arXiv preprint arXiv:1411.4446*, 2014.

[187] M. Schweighofer. Iterated rings of bounded elements and generalizations of Schmüdgen's Positivstellensatz. *Journal für die Reine und Angewandte Mathematik*, 554:19–45, 2003.

[188] M. Schweighofer. On the complexity of Schmüdgen's Positivstellensatz. *Journal of Complexity*, 20(4):529–543, 2004.

[189] N. Z. Shor. Quadratic optimization problems. *Soviet Journal of Computer and Systems Sciences*, 25:1–11, 1987.

[190] A. M. Slinko. *USSR Mathematical Olympiads, 1989-1992*, volume 11. Australian Mathematics Trust, 1997.

[191] L. Slot. Sum-of-squares hierarchies for polynomial optimization and the Christoffel-Darboux kernel. *arXiv preprint arXiv:2111.04610*, 2021.

[192] G. Stengle. A Nullstellensatz and a Positivstellensatz in semialgebraic geometry. *Mathematische Annalen*, 207(2):87–97, 1974.

[193] M. Tacchi, C. Cardozo, D. Henrion, and J. B. Lasserre. Approximating regions of attraction of a sparse polynomial differential system. *IFAC-PapersOnLine*, 53(2):3266–3271, 2020.

[194] M. Tacchi, T. Weisser, J. B. Lasserre, and D. Henrion. Exploiting sparsity for semi-algebraic set volume computation. *Foundations of Computational Mathematics*, pages 1–49, 2021.

[195] K. C. Toh. Some numerical issues in the development of SDP algorithms. *Informs OS Today*, 8(2):7–20, 2018.

[196] K.-C. Toh, M. J. Todd, and R. H. Tütüncü. SDPT3—a MATLAB software package for semidefinite programming, version 1.3. *Optimization methods and software*, 11(1-4):545–581, 1999.

[197] M. Trnovska. Strong duality conditions in semidefinite programming. *Journal of Electrical Engineering*, 56(12):1–5, 2005.

[198] R. H. Tütüncü, K.-C. Toh, and M. J. Todd. Solving semidefinite-quadratic-linear programs using SDPT3. *Mathematical programming*, 95(2):189–217, 2003.

[199] L. Vandenberghe and M. S. Andersen. Chordal graphs and semidefinite optimization. *Foundations and Trends in Optimization*, 1(4):241–433, 2015.

[200] L. Vandenberghe, V. R. Balakrishnan, R. Wallin, A. Hansson, and T. Roh. Interior-point algorithms for semidefinite programming problems derived from the KYP lemma. In *Positive polynomials in control*, pages 195–238. Springer, 2005.

[201] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM review*, 38(1):49–95, 1996.

[202] N. Vreman, P. Pazzaglia, J. Wang, V. Magron, and M. Maggio. Stability of control systems under extended weakly-hard constraints. *Forthcoming*, 2020. Submitted.

[203] H. Waki, S. Kim, M. Kojima, and M. Muramatsu. Sums of squares and semidefinite program relaxations for polynomial optimization problems with structured sparsity. *SIAM Journal on Optimization*, 17(1):218–242, 2006.

[204] H. Waki, S. Kim, M. Kojima, M. Muramatsu, and H. Sugimoto. Algorithm 883: SparsePOP—A Sparse Semidefinite Programming Relaxation of Polynomial Optimization Problems. *ACM Transactions on Mathematical Software (TOMS)*, 35(2):15, 2008.

[205] I. Waldspurger and A. Waters. Rank optimality for the burer–monteiro factorization. *SIAM Journal on Optimization*, 30(3):2577–2602, 2020.

[206] J. Wang, H. Li, and B. Xia. A new sparse sos decomposition algorithm based on term sparsity. In *Proceedings of the 2019 on international symposium on symbolic and algebraic computation*, pages 347–354, 2019.

[207] J. Wang and V. Magron. A second order cone characterization for sums of nonnegative circuits. In *Proceedings of the 45th International Symposium on Symbolic and Algebraic Computation*, pages 450–457, 2020.

[208] J. Wang and V. Magron. Certifying global optimality of AC-OPF solutions via the CS-TSSOS hierarchy. *Forthcoming*, 2021. Submitted.

[209] J. Wang and V. Magron. Exploiting sparsity in complex polynomial optimization. *Journal of Optimization Theory and Applications*, pages 1–25, 2021.

[210] J. Wang and V. Magron. Exploiting term sparsity in noncommutative polynomial optimization. *Computational Optimization and Applications*, 80(2):483–521, 2021.

[211] J. Wang, V. Magron, and J.-B. Lasserre. Chordal-tssos: a moment-sos hierarchy that exploits term sparsity with chordal extension. *SIAM Journal on Optimization*, 31(1):114–141, 2021.

[212] J. Wang, V. Magron, and J.-B. Lasserre. TSSOS: A Moment-SOS hierarchy that exploits term sparsity. *SIAM Journal on Optimization*, 31(1):30–58, 2021.

[213] J. Wang, V. Magron, J. B. Lasserre, and N. H. A. Mai. CS-TSSOS: Correlative and term sparsity for large-scale polynomial optimization. *arXiv preprint arXiv:2005.02828*, 2020.

[214] J. Wang, C. Schlosser, M. Korda, and V. Magron. Exploiting term sparsity in Moment-SOS hierarchy for dynamical systems. *Forthcoming*, 2021. Submitted.

[215] T. Weisser, J. B. Lasserre, and K.-C. Toh. Sparse-BSOS: a bounded degree SOS hierarchy for large scale polynomial optimization with sparsity. *Mathematical Programming Computation*, 10(1):1–32, 2018.

[216] T. Weisser, B. Legat, C. Coey, L. Kapelevich, and J. P. Vielma. Polynomial and Moment Optimization in Julia and JuMP. In *JuliaCon*, 2019.

[217] H. Yang, L. Liang, L. Carlone, and K.-C. Toh. An inexact projected gradient method with rounding and lifting by nonlinear programming for solving rank-one semidefinite relaxation of polynomial optimization. *arXiv preprint arXiv:2105.14033*, 2021.

[218] A. Yurtsever, O. Fercoq, and V. Cevher. A conditional-gradient-based augmented lagrangian framework. In *International Conference on Machine Learning*, pages 7272–7281. PMLR, 2019.

[219] A. Yurtsever, J. A. Tropp, O. Fercoq, M. Udell, and V. Cevher. Scalable semidefinite programming. *SIAM Journal on Mathematics of Data Science*, 3(1):171–200, 2021.