On Quadratically Constrained Quadratic Programs and their Semidefinite Program Relaxations

Alex L. Wang

CMU-CS-22-116
June 15, 2022

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee
Fatma Kılınç-Karzan (Chair)
Samuel Burer (University of Iowa)
Pravesh Kothari
Ryan O'Donnell
Levent Tunçel (University of Waterloo)

Submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

This research was sponsored in part by the National Science Foundation (CMMI 1454548) and the Office of Naval Research (N00014-19-1-2321). The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

Keywords: quadratically constrained quadratic programs · semidefinite programs · quadratic matrix programs · generalized trust-region subproblem · rank-one generated · simultaneous diagonalizability · first-order methods

For Mom, Dad, and Ben

Abstract

Quadratically constrained quadratic programs (QCQPs) are a fundamental class of optimization problems. In a QCQP, we are asked to minimize a (possibly nonconvex) quadratic function subject to a number of (possibly nonconvex) quadratic constraints. Quadratic matrix programs (QMPs) are a related class of optimization problems where the quadratic objective and constraints in the class of QCQPs are replaced by quadratic *matrix* functions. Both QCQPs and QMPs are frequently encountered in practice and arise naturally in diverse areas of operations research, computer science, and engineering. One may regard QMPs as QCQPs with additional structure. Although QCQPs are NP-hard to solve in general, they admit a natural convex relaxation via the standard (Shor) semidefinite program (SDP) relaxation.

The research in this thesis is guided by two fundamental questions related to the SDP relaxation of a general QCQP: (1) What structures within a QCQP ensure that its SDP relaxation is accurate? And, (2) What structures within a QCQP allow its SDP relaxation to be solved efficiently? These two questions comprise the two parts of this thesis.

In contrast to prior work on SDP relaxations of QCQPs (which has focused largely on *approximation* guarantees), Part 1 of this thesis asks when *exactness* occurs in the SDP relaxation of a QCQP. In this direction, we develop a framework for understanding various forms of exactness: (i) *objective value exactness*—the condition that the optimal value of the QCQP and the optimal value of its SDP relaxation coincide, (ii) *convex hull exactness*—the condition that the convex hull of the QCQP epigraph coincides with the (projected) SDP epigraph, and (iii) the *rank-one generated* (ROG) property—the condition that a particular conic subset of the positive semidefinite cone related to a given QCQP is generated by its rank-one matrices. Our analysis for objective value exactness and convex hull exactness stems from a geometric treatment of the projected SDP relaxation and crucially considers how the objective function interacts with the constraints. The ROG property complements these results by offering a sufficient condition for both objective value exactness and convex hull exactness which is oblivious to the objective function.

Part 2 of this thesis seeks to develop new methods for solving large-scale QCQPs and their SDP relaxations efficiently. In this direction, we develop new first-order methods (FOMs) for solving the generalized trust-region subproblem (GTRS) and a broader class of SDPs with exactness properties. Specifically, while the GTRS (the class of QCQPs with a single constraint) is known to have an exact SDP relaxation, the large computational complexity of SDP-based algorithms prevent them from being applied directly to the GTRS. We overcome this barrier by designing FOMs for the GTRS that operate in the original space and possess accelerated linear convergence rates. Perhaps surprisingly, we then show that similar algorithms can be extended to a wider class of SDPs with structured low-rank solutions (e.g., the SDP relaxation of a QCQP or QMP with exactness properties). These FOMs work in the space of the low-rank factorization of the matrix variable and completely avoid storing full matrix variables. In this way, this thesis provides new efficient FOMs for solving QCQPs and QMPs that can be applied whenever the SDP relaxation is known to be exact. Additional work in Part 2 of this thesis studies various notions of simultaneous diagonalizability of sets of quadratic forms. These new notions, specifically the almost SDC and d-restricted SDC properties, seek to understand when a QCQP can be diagonalized after arbitrarily small perturbations of the QCQP data or the introduction of additional variables. We give complete characterizations of these properties in a few settings.

Acknowledgments

First and foremost, thank you Fatma for your energy and effort into making me a well-rounded academic. You went above and beyond to show me the ropes of grad school and academia beyond just the research. Thank you for introducing me to this fascinating area of research, for collaborating with me on exciting projects, and for continuously challenging me to think of the impact of my research. I could always count on you for sage advice on presentations, slides, talks, interviews, and teaching and I know that I am better prepared for the next stage of my career because of the time we spent together.

I have had the pleasure to work with and learn from a number of exceptional colleagues, professors, and students. To my collaborators—Fatma, Rujun, Yunlei, C.J., Gary, Noel, Aravindan, and Abhratanu: Thank you for being so eager to work and for cultivating a friendly collaborative environment. Anupam: Thank you for being willing to chat about grad school and life whenever I knocked on your door. I am grateful for your consistently level-headed advice and hope that I can pay your kindness forward. To Deb: Thank you for all of the work that you put in behind the scenes and your extreme professionalism. To my thesis committee—Fatma, Levent, Pravesh, Ryan, and Sam: Thank you for your insightful comments and the wonderful conversations we have had. I hope to continue to learn from you all.

To my roommates and office mates—Anson, Ellis, Jalani, Kevin, Mark—and all of my friends at CMU and beyond: I cannot imagine surviving grad school without your company. Thank you for all the coffee breaks, phone calls, hikes, trips, and dinners. I cherish these memories fondly.

Last, but most importantly, to my family: Thank you Mom, Dad, and Ben for all of your warmth and love and for always being my biggest believers. Just as important as any of the lessons I learned at CMU were the life lessons I learned from you. Without you, this thesis would never have been possible. To Lucy: Thank you for your constant support, for your positivity and for always believing in me. My days are invariably brighter for having you in them.

Contents

List of Figures

List of Tables

Introduction

Convex optimization has been influential in shaping data science and modern computing. This subfield of optimization has found numerous applications in a variety of domains (e.g., machine learning, statistics, signal processing, and engineering). Unfortunately, a growing number of interesting problems encountered by data scientists, engineers, and the scientific community at large are by nature *highly nonconvex*. Simultaneously, the convex optimization community has begun to investigate more "high-powered" machinery (e.g., semidefinite programs or the sum-of-squares hierarchy), much of which is at present considered impractical in large-scale applications.

This thesis attempts to address this divide by answering theoretical questions underpinning the practical application of tools from convex optimization (specifically, semidefinite programs) to interesting structured nonconvex problems (specifically, structured quadratically constrained quadratic programs and quadratic matrix programs). The goal of this thesis is to understand when certain nonconvex problems may be solved both *accurately* and *efficiently*, with a particular view towards large-scale applications.

Quadratically constrained quadratic programs (QCQPs) are a fundamental class of nonconvex optimization problems that naturally arise in operations research, engineering, and computer science; see [181] for additional applications of QCQPs. The ubiquity of this class of optimization problems stems from its expressiveness: any $\{0, 1\}$ integer program or polynomial optimization problem may be recast as a QCQP (see [11, 25, 91] and references therein).

Quadratic matrix programs (QMPs) are a related class of optimization problems where the quadratic objective function and constraints are replaced by quadratic *matrix* functions. This class of problems finds additional applications in robust optimization, sphere-packing problems, and statistics [17, 20]. The class of QCQPs and QMPs are in fact equivalent, i.e., we can write any QCQP as a QMP and vice versa, however QCQPs derived from QMPs often possess additional structure so that it will be useful to treat them as their own class.

It is well known that QCQPs (and, by extension, QMPs) are NP-hard to solve in general—indeed, the NP-hard combinatorial problem MAX-CUT can be readily recast as a QCQP. On the other hand, the standard (Shor) semidefinite program (SDP) relaxation offers a natural tractable convex relaxation for a general QCQP [161]. This convex relaxation is obtained by first reformulating the QCQP in a lifted space with an additional rank constraint and then dropping the rank constraint.

In passing from the nonconvex QCQP to its convex SDP relaxation, there are two important questions that must be addressed if SDPs are to be of practical importance in this setting:

Question 1. *What structures within a QCQP ensure that its SDP relaxation is* accurate*?*

Question 2. *What structures within a QCQP allow its SDP relaxation to be solved* efficiently*?*

These questions constitute the two parts of this thesis.

A preview of what is to come

We now give an overview of the results and outline of this thesis. We will highlight only a small subset of the background literature and discuss related work in more detail within the individual chapters.

This thesis is predominantly interested in QCQPs and their SDP relaxations:

$$\begin{aligned}&\inf_{x\in\mathbb{R}^n}\left\{q_{\mathrm{obj}}(x):\ q_i(x)\leq 0,\ \forall i\in[m]\right\}\\ &\qquad\geq \inf_{Y\in\mathbb{S}^{n+1}}\left\{\left\langle M_{\mathrm{obj}},Y\right\rangle:\ \begin{array}{l}\langle M_i,Y\rangle\leq 0,\ \forall i\in[m]\\ Y=\begin{pmatrix}* & *\\ * & 1\end{pmatrix}\succeq 0\end{array}\right\}.\end{aligned}$$

Here, for each $i\in\{\mathrm{obj}\}\cup[m]$, we will write $q_i(x)=x^\intercal A_i x+2b_i^\intercal x+c_i$ for some $A_i\in\mathbb{S}^n$, $b_i\in\mathbb{R}^n$, and $c_i\in\mathbb{R}$. Then, defining $M_i:=\left(\begin{smallmatrix}A_i & b_i\\ b_i^\intercal & c_i\end{smallmatrix}\right)$, we have that $q_i(x)=\langle M_i,\left(\begin{smallmatrix}xx^\intercal & x\\ x^\intercal & 1\end{smallmatrix}\right)\rangle$ so that the SDP relaxation is indeed a relaxation of the QCQP. We emphasize that any or all of $q_{\mathrm{obj}},\ q_1,\ \dots,\ q_m$ may be nonconvex.

In a sense, all of the work in this thesis begins with the generalized trust-region subproblem (GTRS). The GTRS is the special class of QCQPs with exactly one constraint:

$$\inf_{x\in\mathbb{R}^n}\left\{q_{\mathrm{obj}}(x):\ q_1(x)\leq 0\right\}.$$

Furthermore, it is standard to assume[1] the existence of some $\hat\gamma\geq 0$ such that $A_{\mathrm{obj}}+\hat\gamma A_1\succ 0$. This unassuming problem is in fact quite intriguing from a theoretical perspective—the GTRS is one of few problems that, despite its nonconvex presentation, can be solved both *exactly* and *efficiently* even in large-scale settings via convex optimization tools.

We note some interesting properties related to *exactness* in the GTRS and how we extend them to the broader class of QCQPs:

- (Chapters 1 and 2). The optimum value of the GTRS coincides with the optimal value of its SDP relaxation. That is,

$$\min_{x\in\mathbb{R}^n}\left\{q_{\mathrm{obj}}(x):\ q_1(x)\leq 0\right\}=\min_{Y\in\mathbb{S}^{n+1}}\left\{\left\langle M_{\mathrm{obj}},Y\right\rangle:\ \begin{array}{l}\langle M_1,Y\rangle\leq 0\\ Y=\begin{pmatrix}* & *\\ * & 1\end{pmatrix}\succeq 0\end{array}\right\},$$

[1] This assumption is standard [2, 95, 180] and is equivalent to requiring that the dual of the SDP relaxation is strictly feasible.

where $M_i = \begin{pmatrix} A_i & b_i \\ b_i^\intercal & c_i \end{pmatrix}$ for each $i \in \{\text{obj}, 1\}$. Furthermore, the epigraph of the GTRS coincides with the epigraph of the (projected) SDP relaxation. That is

$$\operatorname{conv}\left(\left\{(x,t) :\ \begin{array}{l} q_{\text{obj}}(x) \leq t \\ q_1(x) \leq 0 \end{array}\right\}\right) = \left\{(x,t) :\ \begin{array}{l} \exists Y \in \mathbb{S}^{n+1} : \\ \left\langle M_{\text{obj}}, Y\right\rangle \leq t \\ \langle M_1, Y\rangle \leq 0 \\ Y = \begin{pmatrix} * & x \\ x^\intercal & 1 \end{pmatrix} \succeq 0 \end{array}\right\}.$$

In words, we say that *objective value exactness* and *convex hull exactness* hold for the GTRS. Objective value exactness is interesting for straightforward reasons—it allows us to reduce computing the optimum value of the nonconvex GTRS to solving an SDP. Convex hull exactness is a natural sufficient condition for objective value exactness but is also interesting in its own right. Specifically, such results are routinely used within state-of-the-art computational approaches for mixed integer linear and nonlinear programs [49, 171] to produce good convex relaxations of more complicated problems.

These equivalences can (and perhaps should) be quite surprising at first. Indeed, both forms of exactness break even in the presence of just two constraints.

A recent line of work investigates when these forms of exactness hold in the presence of more constraints [37, 38, 87, 111, 112, 167]. We continue this line of work in Chapters 1 and 2 where we develop a framework for analyzing these forms of exactness, unifying a number of previous results. As examples, we show that convex hull exactness holds for vectorized reformulations of certain QMPs or other highly symmetric QCQPs, and that objective value exactness holds for sign-definite QCQPs or random under-constrained QCQPs.

- (Chapter 3). One method of proving both objective value exactness and convex hull exactness for the GTRS relies on the celebrated S-lemma [67]. The S-lemma can be interpreted as the statement that the cone

$$\{Y \in \mathbb{S}^n_+ : \langle M_1, Y\rangle \leq 0\}$$

is *rank-one generated* for any $M_1 \in \mathbb{S}^n$. That is, all extreme rays of the above cone are generated by rank-one matrices. As before, the analogous statement does not necessarily hold when we move from a single linear matrix inequality (LMI) to even just two. Related work in this direction has proved bounds on the rank of extreme rays of cones defined by LMIs [16, 69, 141], proved ROG results for variants of the trust-region subproblem [33, 37], and given a complete characterization of ROG cones defined by linear matrix *equality* (LME) constraints [29, 83].

In Chapter 3, we continue this line of work by investigating the ROG property of cones defined by LMIs. We remark that the ROG property of cones defined by LMIs does not follow from the analogous theory for LMEs. As one of our main results, we give a complete characterization of the ROG property for cones defined by two LMIs. We additionally present some applications of the ROG property to QCQPs and optimization problems involving ratios of quadratic functions.

The GTRS also possesses a number of useful properties that allow it to be solved efficiently (without explicitly solving its SDP relaxation). We note some of these properties and how we extend to them to the broader class of QCQPs:

- (Chapters 4 and 5). In Chapters 4 and 5, we develop two new algorithms for solving the GTRS via FOMs. These methods are particularly suited to solving large-scale GTRS that are sparse or well-structured instances (so that matrix-vector multiplications are cheap).

  The first algorithm (Chapter 4) observes that both the optimizer and the optimum value of the GTRS (where both the objective function and constraint are nonconvex) can be recovered from the optimal solution to a convex quadratic-linear minimax problem:

  $$\min_{x\in\mathbb{R}^n}\max_{\gamma\in\Gamma}\left(q_{\text{obj}}(x)+\gamma q_1(x)\right).$$

  Here, $\Gamma := \left\{\gamma\in\mathbb{R}_+ :\ A_{\text{obj}}+\gamma A_1\succeq 0\right\} = [\gamma_-,\ \gamma_+]$ is a bounded interval. This reformulation was previously noted in [94]. We then show how to apply Nesterov's accelerated gradient descent method for minimax functions [132] to the resulting minimax problem. The FOM developed and analyzed in Chapter 4 is iterative and requires only a constant number of matrix-vector multiplications per iteration. As such, its running time scales linearly with the sparsity of the GTRS instance (i.e., the sparsity of the Hessians in q_{obj} and q_1). This FOM has a sublinear convergence rate of $\tilde{O}\left(\epsilon^{-1/2}\right)$ and improves upon previous state-of-the-art [94, 95]. Our convergence rate also matches the convergence rate of the Lanczos method for computing a minimum eigenvalue [103] or the accelerated gradient descent method for smooth convex functions [132] (up to log-factors) .

  Perhaps surprisingly, in Chapter 5 we show that it is almost always possible to improve these algorithms to achieve *linear* convergence rates, i.e., convergence rates of the form $O(\log(\epsilon^{-1}))$. The key observation here is that for almost all GTRS instances [5], the optimizer γ^* of the dual problem:

  $$\max_{\gamma\in\Gamma}\inf_{x\in\mathbb{R}^n}\left(q_{\text{obj}}(x)+\gamma q_1(x)\right)$$

  satisfies $A_{\text{obj}}+\gamma^* A_1\succ 0$. In particular, by picking a smaller interval $\gamma^*\in[\tilde{\gamma}_-,\tilde{\gamma}_+]\subseteq\left\{\gamma\in\mathbb{R}_+ :\ A_{\text{obj}}+\gamma A_1\succ 0\right\}$, we may introduce *strong convexity* into the minimax problem. These ideas lead to a new FOM that requires only $O\left((\mu^*)^{-1/2}\log(\epsilon^{-1})\right)$ iterations to converge to an ϵ-optimal solution. Here $\mu^*\approx\lambda_{\min}\left(A_{\text{obj}}+\gamma^* A_1\right)$. These rates match those of the Krylov subspace method for the trust-region subproblem [41] (up to log-factors).

- (Chapter 6). In Chapter 6, we extend the algorithm in Chapter 5 to a more general class of SDPs. This class of SDPs, which we refer to as *rank-k exact QMP-like SDPs*, is characterized by solutions with rank k, *a priori* knowledge of the restriction of the SDP solution to a k-dimensional subspace, and standard regularity assumptions such as strict complementarity. We show that similar to the GTRS setting, in the rank-k exact QMP-like SDP setting, it is possible to construct a strongly convex minimax problem whose optimizer coincides with a factorization of the SDP optimizer. We then develop FOMs for constructing the strongly convex minimax problem and subsequently solving it. The overall FOM requires roughly $O(\log(\epsilon^{-1}))$ calls to a prox-map oracle, or, roughly $O(\epsilon^{-1})$ matrix-vector multiplications. Furthermore, in contrast to standard methods for solving SDPs, which require $O(n^2)$ storage to keep track of matrix iterates, our algorithm requires only $O(nk)$ storage where k is the rank of the true SDP solution. This builds upon a recent line of work on storage-optimal FOMs for SDPs [60, 198] but significantly improves the convergence rate.

- (Chapter 7). Under a standard assumption, it is well-known that the GTRS is separable [21, 88]. That is, there exists an invertible $P \in \mathbb{R}^{n\times n}$ such that $P^\intercal A_{\text{obj}} P$ and $P^\intercal A_1 P$ are both diagonal. This property, the *simultaneously diagonalizable via congruence* (SDC) property, is useful from a computational perspective as the SDP relaxation of a diagonal QCQP (one where all A_i matrices are diagonal) can be rewritten as a second-order cone program (SOCP).

  In Chapter 7, we investigate variants of simultaneous diagonalizability. These variants allow us to extend the reach of the SDC property to QCQPs that are *a priori* not diagonalizable. Specifically, the almost SDC property seeks to understand QCQPs that may be diagonalized after arbitrarily small perturbations and the restricted SDC property seeks to understand QCQPs that that admit diagonalizable lifted formulations obtained by introducing a small number of additional variables. In this direction, we give complete characterization of these properties in a few settings. Of particular interest, we show that any pair of symmetric matrices may be diagonalized after arbitrarily small perturbations or with the introduction of a single additional variable.

What's new, what's old?

The work in this thesis has appeared in various forms and is almost entirely verbatim from the following articles.

Chapter 1:

A. L. Wang and F. Kılınç-Karzan. On convex hulls of epigraphs of QCQPs. In *Integer Programming and Combinatorial Optimization (IPCO 2020)*, pages 419–432, 2020

A. L. Wang and F. Kılınç-Karzan. On the tightness of SDP relaxations of QCQPs. *Math. Program.*, 193:33–73, 2022

Chapter 2:

> A. L. Wang and F. Kılınç-Karzan. A geometric view of SDP exactness in QCQPs and its applications. *arXiv preprint*, 2011.07155, 2020

Chapter 3:

> C.J. Argue, F. Kılınç-Karzan, and A. L. Wang. Necessary and sufficient conditions for rank-one generated cones. *Math. Oper. Res.*, 2022. Forthcoming, *arXiv preprint*, 2007.07433
>
> F. Kılınç-Karzan and A. L. Wang. Exactness in SDP relaxations of QCQPs: Theory and applications. Tut. in Oper. Res. 2021
>
> J. Wang, W. Huang, R. Jiang, X. Li, and A. L. Wang. Solving stackelberg prediction game with least squares loss via spherically constrained least squares reformulation. In *International Conference on Machine Learning*, 2022. Forthcoming

Chapter 4:

> A. L. Wang and F. Kılınç-Karzan. The generalized trust region subproblem: solution complexity and convex hull results. *Math. Program.*, 191:445–486, 2022

Chapter 5:

> A. L. Wang, Y. Lu, and F. Kılınç-Karzan. Implicit regularity and linear convergence rates for the generalized trust-region subproblem. *arXiv preprint*, 2112.13821, 2021

Chapter 6:

> A. L. Wang and F. Kılınç-Karzan. Accelerated first-order methods for a class of semidefinite programs. *arXiv preprint*, 2206.00224, 2022

Chapter 7:

> A. L. Wang and R. Jiang. New notions of simultaneous diagonalizability of quadratic forms with applications to QCQPs. *arXiv preprint*, 2101.12141, 2021

As much of the material has appeared previously, we explicitly indicate any new material in framed boxes …

…like so.

Notation

For nonnegative integers $m \leq n$, define $[n] \coloneqq \{1, \dots, n\}$ and $[m, n] \coloneqq \{m, m+1, \dots, n\}$. Let $\mathbb{R}_+$ denote the nonnegative reals and $\mathbb{R}_{++}$ the positive reals. For $i \in [n]$, let $e_i \in \mathbb{R}^n$ denote the ith standard basis vector. Let $\mathbf{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}$ denote the $n-1$ sphere. Let $\mathbb{S}^n$ denote the set of real symmetric $n \times n$ matrices and $\mathbb{S}^n_+$ (resp. $\mathbb{S}^n_{++}$) the cone of positive semidefinite (resp. positive definite) matrices. We write $A \succeq 0$ (resp. $A \succ 0$) if A is positive semidefinite (resp. positive definite). Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues of A. Given $M \in \mathbb{R}^{m\times n}$, let $\operatorname{range}(M)$ and $\ker(M)$ denote the range and kernel of M respectively. When $m = n$, let $\operatorname{tr}(M)$ denote the trace of M. Let $0_n,\ I_n \in \mathbb{S}^n$ denote the $n \times n$ zero matrix and identity matrix respectively; we will simply write 0 or I when the dimension is clear. We will also let $0_n \in \mathbb{R}^n$ denote the zero vector; whether 0 or 0_n is a scalar, vector, or matrix will be clear from context. For $a \in \mathbb{R}^n$, let $\operatorname{Diag}(a)$ denote the diagonal matrix $A \in \mathbb{S}^n$ with diagonal entries $A_{i,i} = a_i$. We endow $\mathbb{S}^n$ with the inner product $\langle A, B\rangle := \operatorname{tr}(A^\intercal B)$. Given W a subspace of $\mathbb{R}^n$ with dimension k, a surjective map $U : \mathbb{R}^k \to W$ and $A \in \mathbb{S}^n$, let A_W denote the restriction of A to W, i.e., $A_W = U^\intercal A U$. When U is inconsequential, we will omit specifying it. We will use $\tilde{O}$-notation to hide log log-factors. Given $u \in W$ and $v \in W^\perp$, let $u \oplus v$ denote their direct sum. For $A \in \mathbb{S}^n$ and $B \in \mathbb{S}^m$, let $A \oplus B \in \mathbb{S}^{n+m}$ and $A \otimes B \in \mathbb{S}^{nm}$ denote the direct sum and Kronecker product of A and B respectively. For a subset $\mathcal{D}$ of some Euclidean space (e.g., $\mathbb{R}^n$ or $\mathbb{S}^n$) let $\mathcal{D}^\circ$, $\operatorname{int}(\mathcal{D})$, $\operatorname{rint}(\mathcal{D})$, $\operatorname{extr}(\mathcal{D})$, $\operatorname{cl}(\mathcal{D})$, $\operatorname{conv}(\mathcal{D})$, $\operatorname{clconv}(\mathcal{D})$, $\operatorname{cone}(\mathcal{D})$, $\operatorname{clcone}(\mathcal{D})$, $\operatorname{span}(\mathcal{D})$, $\operatorname{aff}(\mathcal{D})$, $\dim(\mathcal{D})$, $\operatorname{aff}\dim(\mathcal{D})$ and $\mathcal{D}^\perp$ denote the polar, interior, relative interior, extreme points, closure, convex hull, closed convex hull, conic hull, closed conic hull, linear hull, affine hull, dimension, affine dimension, and orthogonal complement of $\mathcal{D}$, respectively.

Part I

What structures within a QCQP ensure that its SDP relaxation is accurate?

1 Objective value and convex hull exactness

This chapter is based on joint work [178, 181] with Fatma Kılınç-Karzan.

This chapter studies conditions under which the standard semidefinite program (SDP) relaxation of a quadratically constrained quadratic program (QCQP) possesses exactness properties. We begin by outlining a general framework for proving such sufficient conditions. Then, using this framework, we show that the SDP relaxation possesses *objective value exactness* whenever the *quadratic eigenvalue multiplicity*, a parameter capturing the amount of symmetry present in a given problem, is large enough. We present similar sufficient conditions for *convex hull exactness*, i.e., the condition that the projected epigraph of the SDP gives the convex hull of the epigraph in the original QCQP. Our results also imply new sufficient conditions for the tightness (as well as convex hull exactness) of a second order cone program (SOCP) relaxation of simultaneously diagonalizable QCQPs.

1.1 Introduction

This chapter will write a general QCQP in the following form

$$\text{Opt} := \inf_{x\in\mathbb{R}^N}\left\{ q_0(x) : \begin{array}{l} q_i(x) \leq 0,\ \forall i \in [m_I] \\ q_i(x) = 0,\ \forall i \in [m_I+1, m_I+m_E] \end{array} \right\}, \tag{1.1}$$

where for every $i \in [0, m_I + m_E]$, the function $q_i : \mathbb{R}^N \to \mathbb{R}$ is a (possibly nonconvex) quadratic function. We will write $q_i(x) = x^\intercal A_i x + 2b_i^\intercal x + c_i$ where $A_i \in \mathbb{S}^N$, $b_i \in \mathbb{R}^N$, and $c_i \in \mathbb{R}$. Here m_I and m_E are the number of inequality constraints and equality constraints respectively. We will assume that $m := m_I + m_E \geq 1$.

There is a vast literature on approximation guarantees for the standard (Shor) SDP relaxation [22, 121, 129, 161, 193], however, less is known about its exactness. Recently, a number of exciting results in phase retrieval [40] and clustering [1, 122, 153] have shown that under various assumptions on the data (or on the parameters in a random data model), the QCQP formulation of the corresponding problem has a tight SDP relaxation. See also [117] and references therein for more examples of exactness results regarding SDP relaxations. In contrast to these results, which address QCQPs arising from particular problems, Burer and Ye [38] very recently gave some appealing deterministic sufficient conditions under which the standard SDP relaxation of *general* QCQPs is tight. In this chapter, we continue this vein of research for general QCQPs initiated by Burer and Ye [38]. More precisely, we will provide sufficient conditions under which the following two types of results hold: 1) The convex hull of the epigraph of the QCQP is given by the projection of the epigraph of its SDP relaxation, 2) the optimal objective value of the QCQP is equal to the optimal objective value of its SDP relaxation. We will refer to these two types of results as "convex hull results" and "SDP tightness results."

The convex hull results will necessarily require stronger assumptions than the SDP tightness results, however they are also more broadly applicable because such convex hull results are typically used as building blocks to derive strong convex relaxations for complex problems. In fact, the convexification of commonly occurring substructures has been critical in advancing the state-of-the-art computational approaches and software packages for mixed integer linear programs and general nonlinear nonconvex programs [49, 171]. For computational purposes, conditions guaranteeing simple convex hull descriptions are particularly favorable. As we will discuss later, a number of our sufficient conditions will guarantee not only the desired convex hull results but also that these convex hulls are given by a finite number of easily computable convex quadratic constraints in the original space of variables.

1.1.1 Related work

Convex hull results

Convex hull results are well-known for simple QCQPs such as the Trust Region Subproblem (TRS) and the Generalized Trust Region Subproblem (GTRS). Recall that the TRS is a QCQP with a single strictly convex inequality constraint and that the GTRS is a QCQP with a single (possibly nonconvex) inequality constraint. A celebrated result due to Fradkov and Yakubovich [67] implies that the SDP relaxation of the GTRS is tight. More recently, Ho-Nguyen and Kılınç-Karzan [87] showed that the convex hull of the TRS epigraph is given exactly by the projection of the SDP epigraph. Follow-up work by Wang and Kılınç-Karzan [180] showed that the (closed) convex hull of the GTRS epigraph is also given exactly by the projection of the SDP epigraph. In both cases, the projections of the SDP epigraphs can be described in the original space of variables with at most two convex quadratic inequalities. As a result, the TRS and the GTRS can be solved without explicitly running costly SDP-based algorithms; see [2, 94, 95] for other algorithmic ideas to solve the TRS and GTRS.

A different line of research has focused on providing explicit descriptions for the convex hull of the intersection of a single nonconvex quadratic region with convex sets (such as convex quadratic regions, second-order cones (SOCs), or polytopes) or with another single nonconvex quadratic region. For example, the convex hull of the intersection of a two-term disjunction, which is a nonconvex quadratic constraint under mild assumptions, and the second-order cone (SOC) or its cross sections has received much attention in mixed integer programming (see [35, 101, 196] and references therein). Burer and Kılınç-Karzan [35] also studied the convex hull of the intersection of a general nonconvex quadratic region with the SOC or its cross sections. Yıldıran [195] gave an explicit description of the convex hull of the intersection of two *strict* quadratic inequalities (note that the resulting set is open) under the mild regularity condition that there exists $\mu \in [0, 1]$ such that $(1-\mu)A_0 + \mu A_1 \succeq 0$. Follow-up work by Modaresi and Vielma [123] gave sufficient conditions guaranteeing a closed version of the same result. More recently, Santana and Dey [156] gave an explicit description of the convex hull of the intersection of a nonconvex quadratic region with a polytope; this convex hull was further shown to be second-order cone representable. In contrast to these results, we will not limit the number of nonconvex quadratic constraints in our

QCQPs. Additionally, the nonconvex sets that we study in this chapter will arise as epigraphs of QCQPs. In particular, the epigraph variable will play a special role in our analysis. Therefore, we view our developments as complementary to these results.

The convex hull question has also received attention for certain strengthened relaxations of simple QCQPs [33, 34, 37, 167]. In this line of work, the standard SDP relaxation is strengthened by additional inequalities derived using the Reformulation-Linearization Technique (RLT). Sturm and Zhang [167] showed that the standard SDP relaxation strengthened with an additional SOC constraint derived from RLT gives the convex hull of the epigraph of the TRS with one additional linear inequality. Burer and Yang [37] extended this result to the case of an arbitrary number of additional linear inequalities as long as the linear constraints do not intersect inside the trust region domain. See [33] for a survey of some results in this area. Note that in this chapter, we restrict our attention to the standard SDP relaxation of QCQPs. Nevertheless, establishing exactness conditions for strengthened SDP relaxations of QCQPs is clearly of great interest and is a direction for future research.

SDP tightness results

A number of SDP tightness results are known for variants of the TRS.

Jeyakumar and Li [92] showed that the standard SDP relaxation of the TRS with additional linear inequalities is tight under a condition regarding the dimension of the minimum eigenvalue[1] of A_0. These results were extended in the same paper to handle multiple convex quadratic inequality constraints with the same sufficiently rank-deficient quadratic form (see [92, Section 6]). Ho-Nguyen and Kılınç-Karzan [87] presented a sufficient condition for tightness of the SDP relaxation that is slightly more general than [92, Section 6] (see Ho-Nguyen and Kılınç-Karzan [87, Section 2.2] for a comparison of these conditions). A related line of work by Ye and Zhang [194] and Beck and Eldar [18] gives sufficient conditions under which the TRS with one additional quadratic inequality constraint admits a tight SDP relaxation. In contrast to this line of work, our results will address the SDP tightness question in the context of more general QCQPs.

In terms of SDP tightness results, simultaneously diagonalizable QCQPs (SD-QCQPs) have received separate attention [21, 93, 111, 112]. It is shown in [112, Theorem 2.1] that for SD-QCQPs, the SDP relaxation is equivalent to a SOC program (SOCP) relaxation (see also Proposition 1). In particular, the KKT-based sufficient conditions that have been presented for SOCP tightness in [21, 111] also guarantee SDP tightness. We will present SDP tightness results (Theorems 3 and 4) that generalize some of the conditions presented in this line of work. More specifically, our results will not make use of simultaneous diagonalizability assumptions.

A series of articles beginning with Beck [17] and Beck et al. [20] has derived SDP tightness results for quadratic matrix programs (QMPs). A QMP is an optimization problem of the form

$$\inf_{X\in\mathbb{R}^{n\times k}} \left\{ \operatorname{tr}(X^\intercal \mathcal{A}_0 X) + 2\operatorname{tr}(B_0^\intercal X) + c_0 : \begin{array}{r} \operatorname{tr}(X^\intercal \mathcal{A}_i X) + 2\operatorname{tr}(B_i^\intercal X) + c_i \leq 0,\\ \forall i\in[m_I]\\ \operatorname{tr}(X^\intercal \mathcal{A}_i X) + 2\operatorname{tr}(B_i^\intercal X) + c_i = 0,\\ \forall i \in [m_I+1, m]\end{array}\right\},$$

[1] More precisely, this is the minimum generalized eigenvalue of A_0 with respect to the positive definite quadratic form in the constraint.

where $\mathcal{A}_i \in \mathbb{S}^n$, $B_i \in \mathbb{R}^{n\times k}$, and $c_i \in \mathbb{R}$, and arises often in robust least squares or as a result of Burer-Monteiro reformulations for rank-constrained semidefinite programming [17, 36]. In this research vein, Beck [17] showed that a carefully constructed SDP relaxation of QMP is tight whenever $m \leq k$. Note that by replacing the matrix variable $X \in \mathbb{R}^{n\times k}$ by the vector variable $x \in \mathbb{R}^{nk}$, we may reformulate any QMP as a QCQP of a very particular form. Working backwards, if a QCQP can be reformulated as a QMP with $m \leq k$, then we may apply the SDP relaxation proposed in [17] to solve it exactly. We will discuss how such a condition compares with our assumptions in Section 1.3.

In a recent intriguing paper, Burer and Ye [38] gave a sufficient condition guaranteeing that the standard SDP relaxation of general QCQPs is tight. We emphasize that in contrast to prior work, the condition proposed in [38] can be applied to *general* QCQPs. Then, motivated by recent results on exactness guarantees for specific recovery problems with random data and sampling, Burer and Ye [38] also examined a class of random QCQPs and established that if the number of constraints m grows no faster than any fixed polynomial in the number of variables N, then their sufficient condition holds with probability approaching one. In particular, the SDP relaxation is tight with probability approaching one. The SDP tightness results that we present (Theorems 3 and 4) will generalize their deterministic sufficient condition [38, Theorem 1]. As such, their proofs directly imply that our sufficient conditions also hold with probability approaching one in their random data model.

Remark 1. In Chapter 2, we will see new exactness results for both random and semi-random QCQPs (see Propositions 9 and 10). □

1.1.2 Overview and outline of chapter

In contrast to the literature, which has mainly focused on simple QCQPs or QCQPs under certain structural assumptions, in this chapter, we will consider general QCQPs and develop sufficient conditions for both the convex hull result and the SDP tightness result.

We first introduce the epigraph of the QCQP by writing

$$\mathrm{Opt} = \inf_{(x,t)\in\mathbb{R}^{N+1}} \{2t :\ (x,t) \in \mathcal{D}\},$$

where $\mathcal{D}$ is the epigraph of the QCQP in (1.1), i.e.,

$$\mathcal{D} := \left\{ (x,t) \in \mathbb{R}^N \times \mathbb{R} :\ \begin{array}{l} q_0(x) \leq 2t \\ q_i(x) \leq 0,\ \forall i \in [m_I] \\ q_i(x) = 0,\ \forall i \in [m_I + 1, m] \end{array} \right\}. \tag{1.2}$$

As $(x,t) \mapsto 2t$ is linear, we may replace the (potentially nonconvex) epigraph $\mathcal{D}$ with its convex hull $\mathrm{conv}(\mathcal{D})$. Then,

$$\mathrm{Opt} = \inf_{(x,t)\in\mathbb{R}^{N+1}} \{2t :\ (x,t) \in \mathrm{conv}(\mathcal{D})\}.$$

A summary of our contributions, along with an outline of the chapter, is as follows:

i In Section 1.2, we introduce and study the standard SDP relaxation of QCQPs [161] along with its optimal value $\mathrm{Opt}_{\mathrm{SDP}}$ and projected epigraph $\mathcal{D}_{\mathrm{SDP}}$. We set up a framework for deriving sufficient conditions for the "convex hull result," $\mathrm{conv}(\mathcal{D}) = \mathcal{D}_{\mathrm{SDP}}$, and the "SDP tightness result," $\mathrm{Opt} = \mathrm{Opt}_{\mathrm{SDP}}$. This framework is based on the Lagrangian function $(\gamma, x) \mapsto q_0(x) + \sum_{i=1}^m \gamma_i q_i(x)$ and the eigenvalue structure of a dual object $\Gamma \subseteq \mathbb{R}^m$. This object Γ, which consists of the convex Lagrange multipliers, has been extensively studied in the literature (see [188, Chapter 13.4] and more recently [159]).

ii In Section 1.3, we define an integer parameter k, the quadratic eigenvalue multiplicity, that captures the amount of symmetry in a given QCQP. We then give examples where the quadratic eigenvalue multiplicity is large. Specifically, vectorized reformulations of quadratic matrix programs [17] are such an example.

iii In Section 1.4, we use our framework to derive sufficient conditions for the convex hull result: $\mathrm{conv}(\mathcal{D}) = \mathcal{D}_{\mathrm{SDP}}$. Theorem 2 states that if Γ is polyhedral and k is sufficiently large, then $\mathrm{conv}(\mathcal{D}) = \mathcal{D}_{\mathrm{SDP}}$. This theorem actually follows as a consequence of Theorem 1, which replaces the assumption on the quadratic eigenvalue multiplicity with a weaker assumption regarding the dimension of zero eigenspaces related to the matrices A_i. Furthermore, our results in this section establish that if Γ is polyhedral, then $\mathcal{D}_{\mathrm{SDP}}$ is SOC representable; see Remark 8. In particular, when the assumptions of Theorems 1 or 2 hold, we have that $\mathrm{conv}(\mathcal{D}) = \mathcal{D}_{\mathrm{SDP}}$ is SOC representable. In Section 1.4.1, we provide several classes of problems that satisfy the assumptions of these theorems. In particular, we recover a number of results regarding the TRS [87], the GTRS [180], and the solvability of systems of quadratic equations [15]. In Section 1.4.2, we compare our assumption that Γ is polyhedral with the assumption that the QCQP is an SD-QCQP and show that our assumption is strictly more general. In Section 1.4.2, we prove that the SOCP relaxation of SD-QCQPs considered by [21] and Locatelli [111, 112] is indeed equivalent to the SDP relaxation. Consequently, this allows us to recover some of the results from [21, 111, 112] as a consequence of our sufficient conditions (see Section 1.5.1). In Section 1.4.3, we conclude by showing that the dependence we prove on the quadratic eigenvalue multiplicity k is optimal (Propositions 2 and 3).

iv In Section 1.5, we use our framework to derive sufficient conditions for the SDP tightness result: $\mathrm{Opt} = \mathrm{Opt}_{\mathrm{SDP}}$. Specifically, Theorems 3 and 4 give generalizations of the conditions introduced by Locatelli [112] for SDP tightness in a variant of the TRS and Burer and Ye [38] for SDP tightness in diagonal QCQPs.

v In Section 1.6, we discuss the assumption that the dual object Γ is polyhedral. In particular, we show that it is possible to recover both a convex hull result (Theorem 7) and an SDP tightness result (Theorem 8) when this assumption is dropped as long as the quadratic eigenvalue multiplicity k is sufficiently large.

To the best of our knowledge, our results are the first to provide a unified explanation of many of the exactness guarantees present in the literature. Moreover, our results also provide significant generalizations in a number of settings. We discuss the relevant comparisons in detail in the corresponding sections as outlined above. Finally, our results present the first sufficient conditions under which the convex hull of the epigraph of a general QCQP is SOC representable.

1.1.3 Additional notation

Let $\mathbf{B}(\bar{x}, r) = \{x \in \mathbb{R}^n : \|x - \bar{x}\| \leq r\}$ denote the n-ball with radius r and center $\bar{x}$. For a subspace of V of $\mathbb{R}^n$ and $x \in \mathbb{R}^n$, let $\Pi_V x$ denote the projection of x onto V.

1.2 A general framework

In this section, we introduce a general framework for analyzing the standard Shor SDP relaxation of QCQPs. We will examine how both the objective value and the feasible domain change when moving from a QCQP to its SDP relaxation.

We make an assumption that can be thought of as a primal feasibility and dual strict feasibility assumption. This assumption (or a slightly stronger version of it) is standard and is routinely made in the literature on QCQPs (see for example [17, 24, 194]).

Assumption 1. Assume the feasible region of (1.1) is nonempty and there exists $\gamma^* \in \mathbb{R}^m$ such that $\gamma_i^* \geq 0$ for all $i \in [m_I]$ and $A_0 + \sum_{i=1}^m \gamma_i^* A_i \succ 0$. □

Remark 2. By the continuity of $\gamma \mapsto \lambda_{\min}(A_0 + \sum_{i=1}^m \gamma_i A_i)$, we may assume without loss of generality that $\gamma_i^* > 0$ for all $i \in [m_I]$. □

The standard SDP relaxation of (1.1) takes the following form

$$\mathrm{Opt}_{\mathrm{SDP}} := \inf_{x \in \mathbb{R}^N, X \in \mathbb{S}^N} \left\{ \langle Q_0, Y \rangle : \begin{array}{l} Y := \begin{pmatrix} 1 & x^\intercal \\ x & X \end{pmatrix} \\ \langle Q_i, Y \rangle \leq 0,\ \forall i \in [m_I] \\ \langle Q_i, Y \rangle = 0,\ \forall i \in [m_I + 1, m] \\ Y \succeq 0 \end{array} \right\}. \tag{1.3}$$

Here, $Q_i \in \mathbb{S}^{N+1}$ is the matrix $Q_i := \begin{pmatrix} c_i & b_i^\intercal \\ b_i & A_i \end{pmatrix}$. Let $\mathcal{D}_{\mathrm{SDP}}$ denote the epigraph of (1.3) projected onto the (x, t) variables, i.e., define

$$\mathcal{D}_{\mathrm{SDP}} := \left\{ (x, t) \in \mathbb{R}^{N+1} : \begin{array}{l} \exists X \in \mathbb{S}^N : \\ Y := \begin{pmatrix} 1 & x^\intercal \\ x & X \end{pmatrix} \\ \langle Q_0, Y \rangle \leq 2t \\ \langle Q_i, Y \rangle \leq 0,\ \forall i \in [m_I] \\ \langle Q_i, Y \rangle = 0,\ \forall i \in [m_I + 1, m] \\ Y \succeq 0 \end{array} \right\}. \tag{1.4}$$

By taking $X = xx^\intercal$ in both (1.3) and (1.4), we see that $\mathcal{D} \subseteq \mathcal{D}_{\mathrm{SDP}}$ and $\mathrm{Opt} \geq \mathrm{Opt}_{\mathrm{SDP}}$. Noting that $\mathcal{D}_{\mathrm{SDP}}$ is convex (it is the projection of a convex set), we further have that $\mathrm{conv}(\mathcal{D}) \subseteq \mathcal{D}_{\mathrm{SDP}}$. The framework that we set up in the remainder of this section allows us to reason about when equality occurs in both relations, i.e., when $\mathrm{conv}(\mathcal{D}) = \mathcal{D}_{\mathrm{SDP}}$ and/or $\mathrm{Opt} = \mathrm{Opt}_{\mathrm{SDP}}$. We will refer to these two types of result as "convex hull results" and "SDP tightness results."

1.2.1 Rewriting the SDP in terms of a dual object

For $\gamma \in \mathbb{R}^m$, define

$$A(\gamma) := A_0 + \sum_{i=1}^m \gamma_i A_i, \quad b(\gamma) := b_0 + \sum_{i=1}^m \gamma_i b_i, \quad c(\gamma) := c_0 + \sum_{i=1}^m \gamma_i c_i,$$
$$q(\gamma, x) := q_0(x) + \sum_{i=1}^m \gamma_i q_i(x).$$

It is easy to verify that $q(\gamma, x) = x^\intercal A(\gamma)x + 2b(\gamma)^\intercal x + c(\gamma)$. Our framework for analyzing (1.3) is based on the dual object

$$\Gamma := \left\{ \gamma \in \mathbb{R}^m : \begin{array}{l} A(\gamma) \succeq 0 \\ \gamma_i \geq 0,\ \forall i \in [m_I] \end{array} \right\}.$$

We begin by rewriting both $\mathcal{D}_{\mathrm{SDP}}$ and $\mathrm{Opt}_{\mathrm{SDP}}$ to highlight the role played by Γ.

Lemma 1. *Suppose Assumption 1 holds. Then*

$$\mathcal{D}_{\mathrm{SDP}} = \left\{ (x,t) : \sup_{\gamma\in\Gamma} q(\gamma, x) \leq 2t \right\} \quad \textit{and} \quad \mathrm{Opt}_{\mathrm{SDP}} = \min_{x\in\mathbb{R}^N} \sup_{\gamma\in\Gamma} q(\gamma, x).$$

We note that the second identity in Lemma 1 is well-known and was first recorded by Fujie and Kojima [71].

Proof. The second identity follows immediately from the first identity, thus it suffices to prove only the former.

Fix $\hat{x}$ and consider the SDP

$$\inf_{X\in\mathbb{S}^N} \left\{ \langle Q_0, Y\rangle : \begin{array}{l} Y := \begin{pmatrix} 1 & \hat{x}^\intercal \\ \hat{x} & X \end{pmatrix} \\ \langle Q_i, Y\rangle \leq 0,\ \forall i \in [m_I] \\ \langle Q_i, Y\rangle = 0,\ \forall i \in [m_I + 1, m] \\ Y \succeq 0 \end{array} \right\}. \tag{1.5}$$

Comparing programs (1.4) and (1.5), we see that $(\hat{x}, \hat{t}) \in \mathcal{D}_{\mathrm{SDP}}$ if and only if the value $2\hat{t}$ is achieved in (1.5). The dual SDP to (1.5) is given by

$$\sup_{\gamma\in\mathbb{R}^m, t\in\mathbb{R}, y\in\mathbb{R}^N} \left\{ 2t + 2\langle y, \hat{x}\rangle : \begin{array}{l} \begin{pmatrix} c(\gamma) - 2t & b(\gamma)^\intercal - y^\intercal \\ b(\gamma) - y & A(\gamma) \end{pmatrix} \succeq 0 \\ \gamma_i \geq 0,\ \forall i \in [m_I] \end{array} \right\}. \tag{1.6}$$

Note that the first constraint in the dual SDP can only be satisfied if $A(\gamma) \succeq 0$. We may thus rewrite

$$\begin{aligned}
(1.6) &= \sup_{\gamma\in\mathbb{R}^m, t\in\mathbb{R}, y\in\mathbb{R}^N} \left\{ 2t + 2\langle y, \hat{x}\rangle : \begin{array}{l} \begin{pmatrix} 1 \\ x \end{pmatrix}^\intercal \begin{pmatrix} c(\gamma) - 2t & b(\gamma)^\intercal - y^\intercal \\ b(\gamma) - y & A(\gamma) \end{pmatrix} \begin{pmatrix} 1 \\ x \end{pmatrix} \geq 0,\ \forall x \in \mathbb{R}^N \\ \gamma \in \Gamma \end{array} \right\} \\
&= \sup_{\gamma\in\mathbb{R}^m, t\in\mathbb{R}, y\in\mathbb{R}^N} \left\{ 2t + 2\langle y, \hat{x}\rangle : \begin{array}{l} q(\gamma, x) - 2\langle y, x\rangle \geq 2t,\ \forall x \in \mathbb{R}^N \\ \gamma \in \Gamma \end{array} \right\} \\
&= \sup_{\gamma\in\Gamma, y\in\mathbb{R}^N} \inf_{x\in\mathbb{R}^N} q(\gamma, x) + 2\langle y, \hat{x} - x\rangle.
\end{aligned}$$

We first consider the case that the value of the dual SDP (1.6) is bounded. Assumption 1 and Remark 2 imply that (1.6) is strictly feasible. Then by strong conic duality, the primal SDP (1.5) achieves its optimal value and in particular must be feasible. Let γ^* be such that $A(\gamma^*) \succ 0$ (this exists by Assumption 1) and let $y^* = 0$. Then,

$$\lim_{\|x\|\to\infty} q(\gamma^*, x) + 2\langle y^*, \hat{x} - x\rangle = \lim_{\|x\|\to\infty} q(\gamma^*, x) = \infty.$$

In other words, $x \mapsto q(\gamma^*, x) + 2\langle y^*, \hat{x} - x\rangle$ is coercive and we may apply the Minimax Theorem [62, Chapter VI, Proposition 2.3] to get

$$(1.5) = (1.6) = \min_{x\in\mathbb{R}^N} \sup_{\gamma\in\Gamma, y\in\mathbb{R}^N} q(\gamma, x) + 2\langle y, \hat{x} - x\rangle = \sup_{\gamma\in\Gamma} q(\gamma, \hat{x}).$$

The last equation follows as for any $x \neq \hat{x}$, the supremum may take y arbitrarily large in the direction of $\hat{x} - x$. We conclude that if the value of the dual SDP (1.6) is bounded, then

$$(\hat{x}, \hat{t}) \in \mathcal{D}_{\text{SDP}} \quad \iff \quad \sup_{\gamma\in\Gamma} q(\gamma, \hat{x}) \leq 2\hat{t}.$$

Now suppose the value of the dual SDP (1.6) is unbounded. In this case $(\hat{x}, \hat{t}) \notin \mathcal{D}_{\text{SDP}}$ for any value of $\hat{t}$. It remains to observe that

$$\sup_{\gamma\in\Gamma} q(\gamma, \hat{x}) \geq \sup_{\gamma\in\Gamma, y\in\mathbb{R}^N} \inf_{x\in\mathbb{R}^N} q(\gamma, x) + 2\langle y, \hat{x} - x\rangle = \infty.$$

In particular, $(\hat{x}, \hat{t})$ does not satisfy $\sup_{\gamma\in\Gamma} q(\gamma, \hat{x}) \leq 2\hat{t}$ for any value of $\hat{t}$. We conclude that if the value of the dual SDP (1.6) is unbounded, then for all $\hat{t}$,

$$(\hat{x}, \hat{t}) \notin \mathcal{D}_{\text{SDP}} \quad \text{and} \quad \sup_{\gamma\in\Gamma} q(\gamma, \hat{x}) \not\leq 2\hat{t}. \qquad \blacksquare$$

Remark 3. It is not hard to show[2] that the questions $\text{conv}(\mathcal{D}) \overset{?}{=} \mathcal{D}_{\text{SDP}}$ and $\text{Opt} \overset{?}{=} \text{Opt}_{\text{SDP}}$ are invariant under invertible affine transformations of the x-space. In particular, the sufficient

[2] A short proof follows from Lemma 1.

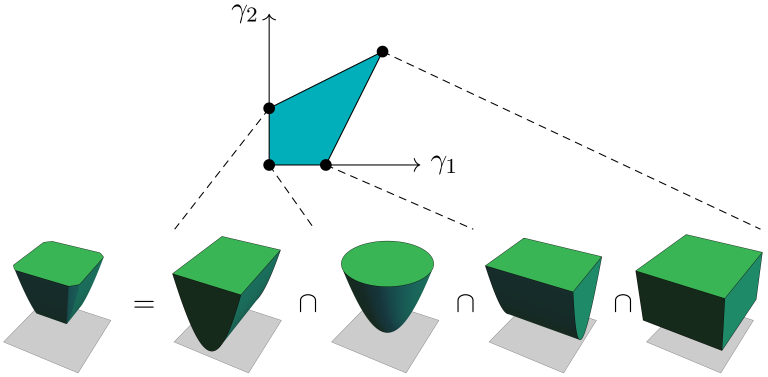

Figure 1.1: The blue region (first row) is an example of the set Γ for some QCQP with two constraints. Lemma 1 then states that $\mathcal{D}_{\mathrm{SDP}}$ (the leftmost set on the second row) is equal to the intersection of the sets $\{(x,t) \in \mathbb{R}^n \times \mathbb{R} : q(\gamma, x) \leq 2t\}$ (the remaining sets on the bottom row) over the extreme points γ of this blue region.

conditions that we will present in this chapter only need to hold after some invertible affine transformation. In this sense, the SDP relaxation will "find" structure in a given QCQP even if it is "hidden" by an affine transformation. □

1.2.2 The eigenvalue structure of Γ

We will make a technical assumption on Γ and $q(\gamma, x)$ in the remainder of our framework.

Assumption 2. Assume that for all $\hat{x} \in \mathbb{R}^n$, if $\sup_{\gamma\in\Gamma} q(\gamma, \hat{x})$ is finite then its maximum value is achieved in Γ. □

Remark 4. As $\gamma \mapsto q(\gamma, \hat{x})$ is linear in γ and Γ is closed, Assumption 2 holds for example whenever Γ is polyhedral or bounded. □

Under Assumption 2, the following definition is well-defined.

Definition 1. Suppose Assumption 2 holds. For any $\hat{x} \in \mathbb{R}^N$ such that $\sup_{\gamma\in\Gamma} q(\gamma, \hat{x})$ is finite, define $\mathcal{F}(\hat{x})$ to be the face of Γ achieving $\sup_{\gamma\in\Gamma} q(\gamma, \hat{x})$, i.e.,

$$\mathcal{F}(\hat{x}) \coloneqq \arg\max_{\gamma\in\Gamma} q(\gamma, x). \qquad \square$$

Definition 2. Let $\mathcal{F}$ be a face of Γ. We say that $\mathcal{F}$ is a *definite face* if there exists $\gamma \in \mathcal{F}$ such that $A(\gamma) \succ 0$. Otherwise, we say that $\mathcal{F}$ is a *semidefinite face* and let $\mathcal{V}(\mathcal{F})$ denote the shared zero eigenspace of $\mathcal{F}$, i.e.,

$$\mathcal{V}(\mathcal{F}) \coloneqq \left\{ v \in \mathbb{R}^N : A(\gamma)v = 0,\ \forall \gamma \in \mathcal{F} \right\}. \qquad \square$$

Note that under Definition 2, each face of Γ is *either* a definite face or a semidefinite face. Specifically, a definite face is not also a semidefinite face.

The following lemma shows that $\mathcal{V}(\mathcal{F})$, which *a priori* may be the trivial subspace $\{0\}$, in fact contains nonzero elements when $\mathcal{F}$ is a semidefinite face.

Lemma 2. *Let $\mathcal{F}$ be a semidefinite face of Γ. Then $\mathcal{V}(\mathcal{F}) \cap \mathbf{S}^{N-1}$ is nonempty.*

Proof. Let $\hat\gamma$ denote a vector in the relative interior of $\mathcal{F}$. By the assumption that $\mathcal{F}$ is a semidefinite face, there exists $v \in \mathbf{S}^{N-1}$ such that $v^\intercal A(\hat\gamma)v = 0$. We claim that $v \in \mathcal{V}(\mathcal{F})$. As $A(\gamma) \succeq 0$ for all $\gamma \in \mathcal{F}$, it suffices to show that $v^\intercal A(\gamma)v \leq 0$ for all $\gamma \in \mathcal{F}$. Suppose $v^\intercal A(\gamma')v > 0$ for some $\gamma' \in \mathcal{F}$. Then, as $\hat\gamma$ is in the relative interior of $\mathcal{F}$, there exists $\epsilon > 0$ small enough such that $\gamma_\epsilon := \hat\gamma + \epsilon(\hat\gamma - \gamma') \in \mathcal{F}$. Finally, by the linearity of $\gamma \mapsto v^\intercal A(\gamma)v$ in γ, we conclude that $v^\intercal A(\gamma_\epsilon)v < 0$, a contradiction. ∎

1.2.3 The framework

Our framework for analyzing the SDP relaxation consists of an "easy part" and a "hard part." The former only requires Assumptions 1 and 2 to hold while the latter may require additional assumptions. We detail the "easy part" in the remainder of this section.

Begin by making the following observations.

Lemma 3. *Suppose Assumptions 1 and 2 hold and let $(\hat x, \hat t) \in \mathcal{D}_{\text{SDP}}$. If $\mathcal{F}(\hat x)$ is a definite face of Γ, then $(\hat x, \hat t) \in \mathcal{D}$.*

Proof. Let $\mathcal{F} := \mathcal{F}(\hat x)$. Because $\mathcal{F}$ is a definite face, there exists $\gamma^* \in \mathcal{F}$ such that $A(\gamma^*) \succ 0$. We verify that $(\hat x, \hat t)$ satisfies each of the constraints in (1.2).

1. By continuity, there exists $\epsilon > 0$ such that $A((1+\epsilon)\gamma^*) \succ 0$. We claim that $(1+\epsilon)\gamma^* \in \mathcal{F}$. Indeed, $A(\gamma^*)$ and $A((1+\epsilon)\gamma^*)$ are both positive definite, thus the constraint $A(\gamma) \succeq 0$ is inactive at both γ^* and $(1+\epsilon)\gamma^*$. Furthermore, for all $i \in [m_I]$, the constraint $\gamma_i \geq 0$ is active at γ^* if and only if it is active at $(1+\epsilon)\gamma^*$. We conclude that $(1+\epsilon)\gamma^* \in \mathcal{F}$ and in particular $0 \in \operatorname{aff}(\mathcal{F})$. This implies
$$q_0(\hat x) = q(0, \hat x) = q(\gamma^*, \hat x) \leq 2\hat t.$$
2. Let $i \in [m_I]$. By continuity there exists $\epsilon > 0$ such that $A(\gamma^* + \epsilon e_i) \succ 0$. Thus, $\gamma^* + \epsilon e_i \in \Gamma$. In particular, since $q(\gamma, \hat x)$ is maximized on $\mathcal{F}$ in Γ, we have that
$$q_i(\hat x) = \frac{q(\gamma^* + \epsilon e_i, \hat x) - q(\gamma^*, \hat x)}{\epsilon} \leq 0.$$
3. Let $i \in [m_I + 1, m]$. By continuity, there exists $\epsilon > 0$ such that $A(\gamma^* \pm \epsilon e_i) \succ 0$. Thus, $\gamma^* \pm \epsilon e_i \in \Gamma$. In particular, since $q(\gamma, \hat x)$ is maximized on $\mathcal{F}$ in Γ, we have that
$$q_i(\hat x) = \frac{q(\gamma^* + \epsilon e_i, \hat x) - q(\gamma^*, \hat x)}{\epsilon} \leq 0.$$
Repeating this calculation with $-\epsilon$ gives $q_i(\hat x) \geq 0$. We deduce that $q_i(\hat x) = 0$. ∎

Observation 1. *Suppose Assumption 1 holds, and let $\mathcal{F}$ be a face of Γ. If $\operatorname{aff}\dim(\mathcal{F}) = m$, then $\mathcal{F}$ is definite.*

Together, Lemma 3 and Observation 1 give a sufficient condition for a point $(\hat{x}, \hat{t}) \in \mathcal{D}_{\text{SDP}}$ to belong to $\mathcal{D}$, namely when $\operatorname{aff}\dim(\mathcal{F}(\hat{x})) = m$. Concretely, we can use the quantity $\operatorname{aff}\dim(\mathcal{F}(\hat{x}))$ to measure the progress of a convex decomposition algorithm.

Lemma 4. *Suppose Assumptions 1 and 2 hold. Suppose furthermore that:*

$$\text{For every } (\hat{x}, \hat{t}) \in \mathcal{D}_{\text{SDP}} \text{ with } \mathcal{F}(\hat{x}) \text{ semidefinite, we can write } (\hat{x}, \hat{t}) \text{ as a convex combination of points } (x_\alpha, t_\alpha) \in \mathcal{D}_{\text{SDP}} \text{ such that } \operatorname{aff}\dim(\mathcal{F}(x_\alpha)) > \operatorname{aff}\dim(\mathcal{F}(\hat{x})). \tag{1.7}$$

Then $\operatorname{conv}(\mathcal{D}) = \mathcal{D}_{\text{SDP}}$ *and* $\text{Opt} = \text{Opt}_{\text{SDP}}$.

Proof. Suppose for the sake of contradiction that $\operatorname{conv}(\mathcal{D}) \neq \mathcal{D}_{\text{SDP}}$. Let

$$(\hat{x}, \hat{t}) \in \operatorname*{arg\,max}_{\mathcal{D}_{\text{SDP}} \setminus \operatorname{conv}(\mathcal{D})} (\operatorname{aff}\dim(\mathcal{F}(\hat{x}))).$$

The point $(\hat{x}, \hat{t})$ is well-defined as $\operatorname{aff}\dim(\mathcal{F}(\hat{x}))$ is a nonnegative integer bounded above by $m-1$ (this follows from Lemma 3 and Observation 1). By Lemma 3, we must have that $\mathcal{F}(\hat{x})$ is semidefinite. By (1.7), there exist points $(x_\alpha, t_\alpha) \in \mathcal{D}_{\text{SDP}}$ such that $\operatorname{aff}\dim(\mathcal{F}(x_\alpha)) > \operatorname{aff}\dim(\mathcal{F}(\hat{x}))$. Then by construction of $(\hat{x}, \hat{t})$ and the fact that $\operatorname{aff}\dim(\mathcal{F}(x_\alpha)) > \operatorname{aff}\dim(\mathcal{F}(\hat{x}))$, we have that $(x_\alpha, t_\alpha) \in \operatorname{conv}(\mathcal{D})$. We conclude that $(\hat{x}, \hat{t}) \in \operatorname{conv}(\{(x_\alpha, t_\alpha)\}_\alpha) \subseteq \operatorname{conv}(\mathcal{D})$, a contradiction. ■

Equivalently, when the assumptions of Lemma 4 hold, the following convex decomposition procedure is guaranteed to terminate and succeed: Given $(\hat{x}, \hat{t}) \in \mathcal{D}_{\text{SDP}}$, if $(\hat{x}, \hat{t}) \in \mathcal{D}$ return $(\hat{x}, \hat{t})$, else decompose $(\hat{x}, \hat{t})$ as a finite convex combination of points $(x_\alpha, t_\alpha) \in \mathcal{D}_{\text{SDP}}$ with $\operatorname{aff}\dim(\mathcal{F}(x_\alpha)) > \operatorname{aff}\dim(\mathcal{F}(\hat{x}))$ and recursively compute convex decompositions of (x_α, t_α).

A similar proof gives the following sufficient condition in the context of the SDP tightness result.

Lemma 5. *Suppose Assumptions 1 and 2 hold. Suppose furthermore that:*

$$\text{For every optimal } (\hat{x}, \hat{t}) \in \mathcal{D}_{\text{SDP}} \text{ with } \mathcal{F}(\hat{x}) \text{ semidefinite, there exists a point } (x', t') \in \mathcal{D}_{\text{SDP}} \text{ such that } t' \leq \hat{t} \text{ and } \operatorname{aff}\dim(\mathcal{F}(x')) > \operatorname{aff}\dim(\mathcal{F}(\hat{x})). \tag{1.8}$$

Then $\text{Opt} = \text{Opt}_{\text{SDP}}$.

The proof of this statement follows the proof of Lemma 4 almost exactly and is omitted.

The "hard part" of our framework for the convex hull result is to give sufficient conditions for (1.7). We give examples of such conditions in Section 1.4. Similarly, the "hard part" of our framework for the SDP tightness result is to give sufficient conditions for (1.8). We give examples of such conditions in Section 1.5.

1.3 Symmetries in QCQPs

In this section, we examine a parameter k that captures the amount of symmetry present in a QCQP of the form (1.1).

Definition 3. The *quadratic eigenvalue multiplicity* of a QCQP of the form (1.1) is the largest integer k such that for every $i \in [0, m]$ there exists $\mathcal{A}_i \in \mathbb{S}^n$ for which $A_i = I_k \otimes \mathcal{A}_i$. □

The quadratic eigenvalue multiplicity k is always at least 1 as we can write each A_i as $A_i = I_1 \otimes \mathcal{A}_i$. On the other hand, it is clear that k must be a divisor of N. In particular, k is always well defined.

For $\gamma \in \mathbb{R}^m$, we also define $\mathcal{A}(\gamma) \coloneqq \mathcal{A}_0 + \sum_{i=1}^m \gamma_i \mathcal{A}_i$.

Example 1. Consider the following optimization problem

$$\inf_{x\in\mathbb{R}^4} \left\{ -\|x\|_2^2 : \begin{array}{l} x_1^2 - x_2^2 + x_3^2 - x_4^2 - 1 \leq 0 \\ 2x_1^2 + x_2^2 + 2x_3^2 + x_4^2 - 1 \leq 0 \end{array} \right\}.$$

The quadratic forms in this problem are

$$A_0 = I_2 \otimes \begin{pmatrix} -1 & \\ & -1 \end{pmatrix}, \qquad A_1 = I_2 \otimes \begin{pmatrix} 1 & \\ & -1 \end{pmatrix}, \quad \text{and} \quad A_2 = I_2 \otimes \begin{pmatrix} 2 & \\ & 1 \end{pmatrix}.$$

Thus, this QCQP has quadratic eigenvalue multiplicity $k \geq 2$. Recalling that k must be a divisor of N and noting that A_1 cannot be written as $A_1 = I_4 \otimes \mathcal{A}_1$ for any $\mathcal{A}_1 \in \mathbb{S}^1$, we conclude that $k = 2$. □

Remark 5. Suppose we have access to some $\mu \in \mathbb{R}^m$ such that $\mathcal{A}(\mu)$ has distinct eigenvalues. Then, by simply performing a spectral decomposition of $A(\mu)$ and counting the multiplicities of the eigenvalues, we can correctly output the value k. □

Remark 6. The quadratic eigenvalue multiplicity can be viewed as a particular *group symmetry* in $\{A_0, A_1, \dots, A_m\}$. Group symmetric SDPs have been studied in more generality with the goal of reducing the size of large SDPs (and in turn their solve-times) [54, 72]. See also [53] for an application of such ideas to solving large real-world instances of the quadratic assignment problem.

Specifically, the *Wedderburn decomposition* of the matrix $\mathbb{C}^*$-algebra generated by $\{A_0, A_1, \dots, A_m\}$ plays a prominent role in the analysis of such symmetries. In this setting, our parameter k can be compared to the "block multiplicity" of a basic algebra in the Wedderburn decomposition. This decomposition can be computed efficiently given access to a generic element from the center of the algebra (see [55, 73] and references therein). □

Recall that in Lemma 2, we showed that $\dim(\mathcal{V}(\mathcal{F})) \geq 1$ whenever $\mathcal{F}$ is a semidefinite face of Γ. The following lemma will show that when the quadratic eigenvalue multiplicity is large, we can in fact lower bound $\dim(\mathcal{V}(\mathcal{F})) \geq k$. This is the main property of the quadratic eigenvalue multiplicity that we will use in Sections 1.4 and 1.5.

Lemma 6. *If $\mathcal{F}$ is a semidefinite face of Γ, then $\dim(\mathcal{V}(\mathcal{F})) \geq k$.*

Proof. By Lemma 2, there exists $\hat{v} \in \mathcal{V}(\mathcal{F}) \cap \mathbf{S}^{N-1}$. We can write $\hat{v}$ as the concatenation of k-many n-dimensional vectors $v_1, \dots, v_k \in \mathbb{R}^n$. Then for $\gamma \in \mathcal{F}$,

$$0 = A(\gamma)\hat{v} = \begin{pmatrix} \mathcal{A}(\gamma) & & & \\ & \mathcal{A}(\gamma) & & \\ & & \ddots & \\ & & & \mathcal{A}(\gamma) \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{pmatrix} = \begin{pmatrix} \mathcal{A}(\gamma)v_1 \\ \mathcal{A}(\gamma)v_2 \\ \vdots \\ \mathcal{A}(\gamma)v_k \end{pmatrix}.$$

Hence, $\mathcal{A}(\gamma)v_i = 0$ for all $i \in [k]$. As $\hat{v} \neq 0$, there exists some $i \in [k]$ such that $v_i \neq 0$. Finally, note that for all $y \in \mathbb{R}^k$,

$$A(\gamma)(y \otimes v_i) = (I_k \otimes \mathcal{A}(\gamma))(y \otimes v_i) = y \otimes (\mathcal{A}(\gamma)v_i) = 0.$$

In other words, $\mathbb{R}^k \otimes v_i \subseteq \mathcal{V}(\mathcal{F})$ and thus $\dim(\mathcal{V}(\mathcal{F})) \geq k$. ■

Remark 7. In quadratic matrix programming [17, 20], we are asked to optimize

$$\inf_{X \in \mathbb{R}^{n\times k}} \left\{ \operatorname{tr}(X^\intercal \mathcal{A}_0 X) + 2\operatorname{tr}(B_0^\intercal X) + c_0 : \begin{array}{r} \operatorname{tr}(X^\intercal \mathcal{A}_i X) + 2\operatorname{tr}(B_i^\intercal X) + c_i \leq 0, \\ \forall i \in [m_I] \\ \operatorname{tr}(X^\intercal \mathcal{A}_i X) + 2\operatorname{tr}(B_i^\intercal X) + c_i = 0, \\ \forall i \in [m_I + 1, m] \end{array} \right\}, \tag{1.9}$$

where $\mathcal{A}_i \in \mathbb{S}^n$, $B_i \in \mathbb{R}^{n\times k}$ and $c_i \in \mathbb{R}$ for all $i \in [0, m]$. We can transform this program to an equivalent QCQP in the vector variable $x \in \mathbb{R}^{nk}$ by identifying

$$X = \begin{pmatrix} x_1 & \dots & x_{(k-1)n+1} \\ \vdots & \ddots & \vdots \\ x_n & \dots & x_{kn} \end{pmatrix}.$$

Then

$$\operatorname{tr}(X^\intercal \mathcal{A}_i X) + 2\operatorname{tr}(B_i^\intercal X) + c_i = x^\intercal (I_k \otimes \mathcal{A}_i)x + 2b_i^\intercal x + c_i,$$

where, $b_i \in \mathbb{R}^{nk}$ has entries $(b_i)_{(t-1)n+s} = (B_i)_{s,t}$. In particular, the vectorized reformulation of (1.9) has quadratic eigenvalue value multiplicity k. □

1.4 Convex hull results

In this section, we present new sufficient conditions for the convex hull result $\mathcal{D}_{\text{SDP}} = \operatorname{conv}(\mathcal{D})$. We will first analyze the case where the geometry of Γ is particularly nice.

Assumption 3. Assume that Γ is polyhedral. □

We remark that although Assumption 3 is rather restrictive, it is general enough to cover the case where the set of quadratic forms $\{A_i\}_{i\in[0,m]}$ is diagonal or simultaneously diagonalizable—a class of QCQPs which has been studied extensively in the literature (see Section 1.4.2 for references). We will present examples and non-examples of Assumption 3 in Sections 1.4.1 and 1.4.2 and discuss the difficulties in removing this assumption in Section 1.4.3. Finally, we will recover weaker results without this assumption in Section 1.6.

Note that Assumption 3 immediately implies Assumption 2 so that we may apply the framework from Section 1.2.

Our main result in this section is the following theorem.

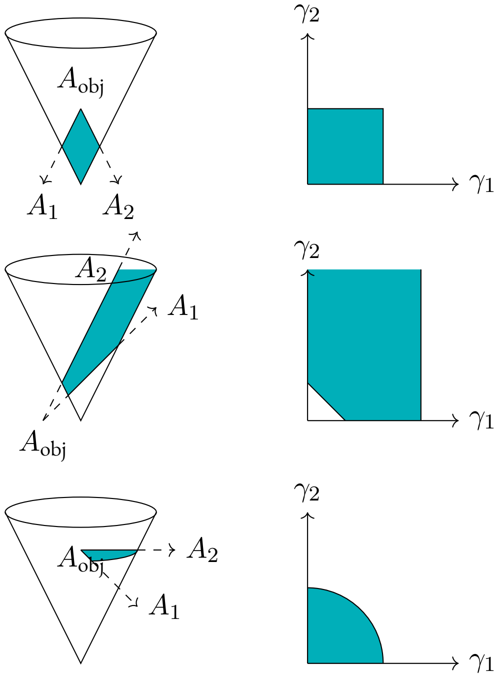

Figure 1.2: In each row above, we illustrate first the set $\left\{A(\gamma) \in \mathbb{S}^2 : \gamma \in \mathbb{R}^2_+\right\}$ on the left and the set Γ on the right.

Theorem 1. *Suppose Assumptions 1 and 3 hold. Furthermore, suppose that for every semidefinite face $\mathcal{F}$ of Γ we have*

$$\dim(\mathcal{V}(\mathcal{F})) \geq \operatorname{aff}\dim(\{b(\gamma) : \gamma \in \mathcal{F}\}) + 1.$$

Then,

$$\operatorname{conv}(\mathcal{D}) = \mathcal{D}_{\text{SDP}} \quad \textit{and} \quad \operatorname{Opt} = \operatorname{Opt}_{\text{SDP}}.$$

As before, the second identity follows immediately from the first identity, thus it suffices to prove only the former. The main effort in this section will be the proof of the following lemma.

Lemma 7. *Suppose Assumptions 1 and 3 hold. Furthermore, suppose that for every semidefinite face $\mathcal{F}$ of Γ we have*

$$\dim(\mathcal{V}(\mathcal{F})) \geq \operatorname{aff}\dim(\{b(\gamma) : \gamma \in \mathcal{F}\}) + 1.$$

Let $(\hat{x}, \hat{t}) \in \mathcal{D}_{\text{SDP}}$ and let $\mathcal{F} = \mathcal{F}(\hat{x})$. If $\mathcal{F}$ is a semidefinite face of Γ, then $(\hat{x}, \hat{t})$ can be written as a convex combination of points $(x_\alpha, t_\alpha) \in \mathcal{D}_{\text{SDP}}$ such that $\operatorname{aff}\dim(\mathcal{F}(x_\alpha)) > \operatorname{aff}\dim(\mathcal{F}(\hat{x}))$.

The proof of Theorem 1 follows at once from Lemma 7 and Lemma 4.

Before proving Lemma 7, we introduce some new notation for handling the recessive directions of Γ and prove a straightforward lemma about decomposing Γ. Let

$$\breve{A}(\gamma) \coloneqq \sum_{i=1}^{m} \gamma_i A_i, \quad \breve{b}(\gamma) \coloneqq \sum_{i=1}^{m} \gamma_i b_i, \quad \breve{c}(\gamma) \coloneqq \sum_{i=1}^{m} \gamma_i c_i, \quad \breve{q}(\gamma, x) \coloneqq \sum_{i=1}^{m} \gamma_i q_i(x).$$

Lemma 8. *Suppose Assumption 3 holds. Then Γ can be written as*

$$\Gamma = \Gamma_e + \operatorname{cone}(\Gamma_r)$$

where both Γ_e and Γ_r are polytopes. Here, Γ_r may be the trivial set $\{0\}$. Furthermore, for $\hat{x} \in \mathbb{R}^N$ such that $\sup_{\gamma\in\Gamma} q(\gamma, \hat{x})$ is finite, we have

$$\mathcal{F}(\hat{x}) = \mathcal{F}_e(\hat{x}) + \operatorname{cone}(\mathcal{F}_r(\hat{x}))$$

where $\mathcal{F}_e(\hat{x})$ is the face of Γ_e maximizing $q(\gamma, \hat{x})$ and $\mathcal{F}_r(\hat{x})$ is the face of Γ_e satisfying $\breve{q}(\gamma, \hat{x}) = 0$.

Proof. This follows immediately from the Minkowski-Weyl Theorem and noting that $\breve{q}(\gamma_r, \hat{x}) \leq 0$ for all $\gamma_r \in \Gamma_r$ when $\sup_{\gamma\in\Gamma} q(\gamma, \hat{x})$ is finite. ■

Proof of Lemma 7. Without loss of generality, we may assume that $\sup_{\gamma\in\Gamma} q(\gamma, \hat{x}) = 2\hat{t}$. Otherwise, we can decrease $\hat{t}$ and note that $\mathcal{D}$ is closed upwards in the t-direction. In particular, we have that $q(\gamma, \hat{x})$ achieves the value $2\hat{t}$ on $\mathcal{F}$.

We claim that the following system in variables v and s

$$\begin{cases} \langle b(\gamma), v\rangle = s, \ \forall \gamma \in \mathcal{F} \\ v \in \mathcal{V}(\mathcal{F}), \ s \in \mathbb{R} \end{cases}$$

has a nonzero solution. Indeed, we may replace the constraint $\langle b(\gamma), v\rangle = s, \ \forall\gamma \in \mathcal{F}$ with at most

$$\operatorname{aff\,dim}(\{b(\gamma) : \gamma \in \mathcal{F}\}) + 1 \leq \dim(\mathcal{V}(\mathcal{F}))$$

homogeneous linear equalities in the variables v and s. The claim then follows by noting that the equivalent system is an under-constrained homogeneous system of linear equalities and thus has a nonzero solution (v, s). It is easy to verify that $v \neq 0$, hence by scaling we may take $v \in \mathbf{S}^{N-1}$. In the remainder of the proof, let $v \in \mathcal{V}(\mathcal{F}) \cap \mathbf{S}^{N-1}$ and $s \in \mathbb{R}$ denote a solution pair to the above system.

Apply Lemma 8 to decompose $\Gamma = \Gamma_e + \operatorname{cone}(\Gamma_r)$ and $\mathcal{F} = \mathcal{F}_e + \operatorname{cone}(\mathcal{F}_r)$.

We will modify $(\hat{x}, \hat{t})$ in the (v, s) direction. For $\alpha \in \mathbb{R}$, we define

$$(x_\alpha, t_\alpha) \coloneqq \left(\hat{x} + \alpha v, \ \hat{t} + \alpha s\right).$$

First, for any fixed $\gamma_f \in \mathcal{F}$, we consider how $q(\gamma_f, x_\alpha) - 2t_\alpha$ changes with α. We can expand

$$\begin{aligned} q(\gamma_f, x_\alpha) - 2t_\alpha &= \left(q(\gamma_f, \hat{x}) - 2\hat{t}\right) + 2\alpha(\hat{x}^\intercal A(\gamma_f)v + b(\gamma_f)^\intercal v - s) + \alpha^2 v^\intercal A(\gamma_f)v \\ &= q(\gamma_f, \hat{x}) - 2\hat{t} \\ &= 0, \end{aligned}$$

where the second line follows as $A(\gamma_f)v = 0$ (recall $v \in \mathcal{V}(\mathcal{F})$) and $b(\gamma_f)^\intercal v = s$ for all $\gamma_f \in \mathcal{F}$, and the third line follows as $q(\gamma_f, \hat{x}) = 2\hat{t}$ for all $\gamma_f \in \mathcal{F}$. Now consider any $\gamma_e \in \mathcal{F}_e$ and $\gamma_r \in \mathcal{F}_r$. Note that γ_e and $\gamma_e + \gamma_r$ both lie in $\mathcal{F}$. Then by the above calculation, both $\alpha \mapsto q(\gamma_e, x_\alpha) - 2t_\alpha$ and $\alpha \mapsto q(\gamma_e + \gamma_r, x_\alpha) - 2t_\alpha$ are identically zero. In particular, we also have that $\alpha \mapsto \breve{q}(\gamma_r, x_\alpha) = q(\gamma_e + \gamma_r, x_\alpha) - q(\gamma_e, x_\alpha) = 0$ is identically zero.

On the other hand, for $\gamma_e \in \Gamma_e \setminus \mathcal{F}_e$, we can expand

$$q(\gamma_e, x_\alpha) - 2t_\alpha = \left(q(\gamma_e, \hat{x}) - 2\hat{t}\right) + 2\alpha(\hat{x}^\intercal A(\gamma_e)v + b(\gamma_e)^\intercal v - s) + \alpha^2 v^\intercal A(\gamma_e)v,$$

and note that $v^\intercal A(\gamma_e)v \geq 0$ holds because $A(\gamma_e)$ is positive semidefinite. Hence, for $\gamma_e \in \Gamma_e \setminus \mathcal{F}_e$, we have that $\alpha \mapsto q(\gamma_e, x_\alpha) - 2t_\alpha$ is a (possibly non-strictly) convex quadratic function taking the value $q(\gamma_e, \hat{x}) - 2\hat{t} < 0$ at $\alpha = 0$ (the strict inequality here follows from the fact that $\gamma_e \in \Gamma_e \setminus \mathcal{F}_e$).

Similarly, for $\gamma_r \in \Gamma_r \setminus \mathcal{F}_r$, we can expand

$$\breve{q}(\gamma_r, x_\alpha) = \breve{q}(\gamma_r, \hat{x}) + 2\alpha\left(\hat{x}^\intercal \breve{A}(\gamma_r)v + \breve{b}(\gamma_r)^\intercal v\right) + \alpha^2 v^\intercal \breve{A}(\gamma_r)v.$$

Note that $\breve{A}(\gamma) \succeq 0$ for all $\gamma \in \Gamma_r$. Hence, for $\gamma_r \in \Gamma_r \setminus \mathcal{F}_r$, we have that $\alpha \mapsto \breve{q}(\gamma_r, x_\alpha)$ is a (possibly non-strictly) convex quadratic function taking the value $\breve{q}(\gamma_r, \hat{x}) < 0$ at $\alpha = 0$ (the strict inequality here follows from the fact that $\gamma_r \in \Gamma_r \setminus \mathcal{F}_r$).

We have shown that the following finite set of univariate quadratic functions in α,

$$\mathcal{Q} := \left(\{q(\gamma_e, x_\alpha) - 2t_\alpha :\ \gamma_e \in \operatorname{extr}(\Gamma_e)\} \cup \{\breve{q}(\gamma_r, x_\alpha) :\ \gamma_r \in \operatorname{extr}(\Gamma_r)\}\right) \setminus \{0\},$$

consists of (possibly non-strictly) convex quadratic functions which are negative at $\alpha = 0$. The finiteness of this set follows from the assumption that Γ is polyhedral.

We claim that there exists a quadratic function in $\mathcal{Q}$ which is strictly convex: Note γ^* from Assumption 1 satisfies $\gamma^* \in \Gamma$. Thus, we can decompose $\gamma^* = \gamma_e + \alpha\gamma_r$ for $\gamma_e \in \Gamma_e$, $\gamma_r \in \Gamma_r$, and $\alpha \geq 0$. Then,

$$0 < v^\intercal A(\gamma^*)v = [v^\intercal A(\gamma_e)v] + \alpha\left[v^\intercal \breve{A}(\gamma_r)v\right].$$

Hence, one of the square-bracketed terms must be positive. The claim then follows by linearity in γ of the functions $\gamma \mapsto v^\intercal A(\gamma)v$ and $\gamma \mapsto v^\intercal \breve{A}(\gamma)v$.

As $\mathcal{Q}$ is a finite set by Assumption 3, there exists an $\alpha_+ > 0$ such that $q(\alpha_+) \leq 0$ for all $q \in \mathcal{Q}$ with at least one equality. Then because $\Gamma_e = \operatorname{conv}(\operatorname{extr}(\Gamma_e))$ and $\Gamma_r = \operatorname{conv}(\operatorname{extr}(\Gamma_r))$, we have $q(\gamma_e, x_{\alpha_+}) \leq 2t_{\alpha_+}$ for all $\gamma_e \in \Gamma_e$ and $\breve{q}(\gamma_r, x_{\alpha_+}) \leq 0$ for all $\gamma_r \in \Gamma_r$. Thus, $(x_{\alpha_+}, t_{\alpha_+}) \in \mathcal{D}_{\text{SDP}}$.

It remains to show that $\operatorname{aff\,dim}(\mathcal{F}(x_{\alpha_+})) > \operatorname{aff\,dim}(\mathcal{F}(\hat{x}))$. The discussion in the previous paragraph implies that $\sup_{\gamma\in\Gamma} q(\gamma, x_{\alpha_+}) \leq 2t_{\alpha_+}$. This value is achieved by $\gamma_f \in \mathcal{F}(\hat{x})$: Note $q(\gamma_f, x_{\alpha_+}) - 2t_{\alpha_+} = q(\gamma_f, \hat{x}) - 2\hat{t} = 0$. In particular, $\mathcal{F}(\hat{x}) \subseteq \mathcal{F}(x_{\alpha_+})$. Thus, it suffices to show that there exists $\gamma_+ \in \mathcal{F}(x_{\alpha_+}) \setminus \mathcal{F}(\hat{x})$.

Suppose the quadratic function in $\mathcal{Q}$ with α_+ as a root is of the form $q(\gamma_+, x_\alpha) - 2t_\alpha$. Then $\gamma_+ \in \mathcal{F}(x_{\alpha_+})$ as $q(\gamma_+, x_{\alpha_+}) - 2t_{\alpha_+} = 0$. On the other hand, $\gamma_+ \notin \mathcal{F}(\hat{x})$ by the construction of $\mathcal{Q}$.

Suppose the quadratic function in $\mathcal{Q}$ with α_+ as a root is of the form $\breve{q}(\gamma_r, x_\alpha)$. Select any $\gamma_f \in \mathcal{F}(\hat{x})$ and recall that $q(\gamma_f, x_\alpha) - 2t_\alpha$ is identically zero as an expression in α. Define $\gamma_+ = \gamma_f + \gamma_r$. Then,

$$q(\gamma_+, x_{\alpha_+}) - 2t_{\alpha_+} = \left(q(\gamma_f, x_{\alpha_+}) - 2t_{\alpha_+}\right) + \breve{q}(\gamma_r, x_{\alpha_+}) = 0$$

and hence $\gamma_+ \in \mathcal{F}(x_{\alpha_+})$. On the other hand, $\breve{q}(\gamma_r, \hat{x}) < 0$ by the construction of $\mathcal{Q}$. In particular,

$$q(\gamma_+, \hat{x}) - 2\hat{t} = \left(q(\gamma_f, \hat{x}) - 2\hat{t}\right) + \breve{q}(\gamma_r, \hat{x}) < 0$$

and thus $\gamma_+ \notin \mathcal{F}(\hat{x})$.

The existence of an $\alpha_- < 0$ satisfying the same properties is proved analogously. Then we may write $(\hat{x}, \hat{t})$ as a convex combination of $(x_{\alpha_+}, t_{\alpha_+})$ and $(x_{\alpha_-}, t_{\alpha_-})$. ■

The next theorem follows as a corollary to Theorem 1.

Theorem 2. *Suppose Assumptions 1 and 3 hold. Furthermore, suppose that for every semidefinite face $\mathcal{F}$ of Γ we have*

$$k \geq \operatorname{aff\,dim}(\{b(\gamma) : \gamma \in \mathcal{F}\}) + 1.$$

Then,

$$\operatorname{conv}(\mathcal{D}) = \mathcal{D}_{\text{SDP}} \quad \textit{and} \quad \text{Opt} = \text{Opt}_{\text{SDP}}.$$

Proof. This theorem follows from Lemma 6 and Theorem 1. ■

Remark 8. We remark that when Γ is polyhedral (Assumption 3), the set $\mathcal{D}_{\text{SDP}}$ is actually SOC representable: By Lemmas 1 and 8 we can write

$$\begin{aligned}\mathcal{D}_{\text{SDP}} &= \left\{(x,t) : \sup_{\gamma\in\Gamma} q(\gamma, x) \leq 2t\right\}\\ &= \left\{(x,t) : \begin{array}{l} q(\gamma_e, x) \leq 2t, \ \forall \gamma_e \in \operatorname{extr}(\Gamma_e)\\ \breve{q}(\gamma_f, x) \leq 0, \ \forall \gamma_f \in \operatorname{extr}(\Gamma_r)\end{array}\right\}.\end{aligned}$$

In other words, $\mathcal{D}_{\mathrm{SDP}}$ is defined by finitely many convex quadratic inequalities. In particular, the assumptions of Theorem 1 and 2 imply that $\operatorname{conv}(\mathcal{D})$ is SOC representable. □

1.4.1 Applications of Theorems 1 and 2

We now state some classes of problems where the assumptions of Theorems 1 and 2 hold.

The most basic setup we can cover via these theorems is the case of convex quadratic programs.

Corollary 1. *Suppose Assumption 1 holds. If $A_0 \succ 0$, $m_E = 0$ and $A_i \succeq 0$ for all $i \in [m_I]$, then*

$$\operatorname{conv}(\mathcal{D}) = \mathcal{D}_{\mathrm{SDP}} \quad \textit{and} \quad \mathrm{Opt} = \mathrm{Opt}_{\mathrm{SDP}}\,.$$

Proof. Assumption 3 holds in this case as

$$\Gamma = \left\{ \gamma \in \mathbb{R}^m : \begin{array}{c} A(\gamma) \succeq 0 \\ \gamma \geq 0 \end{array} \right\} = \{\gamma \in \mathbb{R}^m : \gamma \geq 0\}.$$

Furthermore, each face of Γ contains the origin. Thus noting that $A(0) = A_0 \succ 0$ is positive definite, we conclude that Γ does not have any semidefinite face. This allows us to apply Theorem 2. ■

Remark 9. It is possible to apply a standard limit argument (see for example [38]) to handle additionally the case where A_0 is only positive semidefinite. □

Next, we discuss a number of results on TRS and GTRS.

Corollary 2. *Suppose $m = 1$ and Assumption 1 holds. Then,*

$$\operatorname{conv}(\mathcal{D}) = \mathcal{D}_{\mathrm{SDP}} \quad \textit{and} \quad \mathrm{Opt} = \mathrm{Opt}_{\mathrm{SDP}}\,.$$

Proof. The set Γ will either be a bounded interval $[\gamma_1, \gamma_2]$, a semi-infinite interval $[\gamma_1, \infty)$, or the entire line $(-\infty, \infty)$. In all three cases, Γ is polyhedral and Assumption 3 holds.

By Observation 1, any semidefinite face of Γ must have affine dimension at most $m - 1 = 0$. In particular $\operatorname{aff}\dim(\{b(\gamma) : \gamma \in \mathcal{F}\}) = 0$ and the assumption on the quadratic eigenvalue multiplicity in Theorem 2 holds as k is always at least 1. ■

Corollary 2 in particular recovers the well-known results associated with the epigraph set of the TRS[3] and the GTRS (see [87, Theorem 13] and [180, Theorems 1 and 2]).

Corollary 3. *Suppose Assumptions 1 and 3 hold. If $b_i = 0$ for all $i \in [m]$, then*

$$\operatorname{conv}(\mathcal{D}) = \mathcal{D}_{\mathrm{SDP}} \quad \textit{and} \quad \mathrm{Opt} = \mathrm{Opt}_{\mathrm{SDP}}\,.$$

[3] Corollary 2 fails to fully recover [87, Theorem 13]. Indeed, [87, Theorem 13] also gives a description of the convex hull of the epigraph of the TRS with an additional conic constraint under some assumptions. We do not consider these additional conic constraints in our setup.

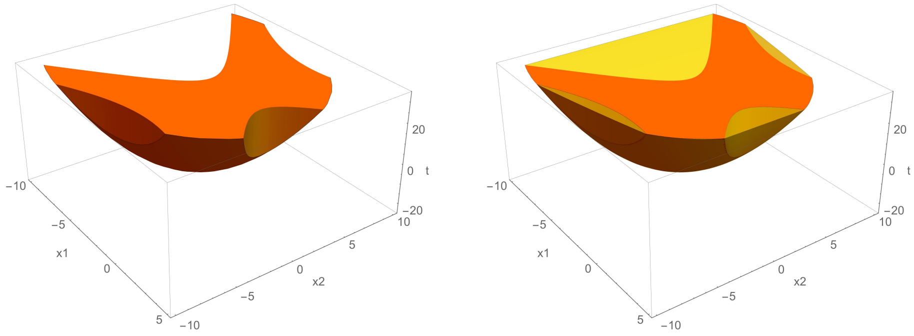

Figure 1.3: The sets $\mathcal{D}$ (in orange) and $\operatorname{conv}(\mathcal{D})$ (in yellow) from Example 2

Proof. Note that $b(\gamma) = b_0 + \sum_{i=1}^m \gamma_i b_i = b_0$ for any $\gamma \in \mathbb{R}^m$. Thus, for any face $\mathcal{F}$ of Γ, we have

$$\operatorname{aff\,dim}(\{b(\gamma) : \gamma \in \mathcal{F}\}) + 1 = \operatorname{aff\,dim}(\{b_0\}) + 1 = 1.$$

In particular, the assumptions on the quadratic eigenvalue multiplicity in Theorem 2 holds as k is always at least 1. ■

Example 2. Consider the following optimization problem

$$\inf_{x\in\mathbb{R}^2}\left\{x_1^2 + x_2^2 + 10x_1 : \begin{array}{l} x_1^2 - x_2^2 - 5 \leq 0 \\ -x_1^2 + x_2^2 - 50 \leq 0 \end{array}\right\}.$$

We check that the conditions of Corollary 3 hold. Assumption 1 holds as $A(0) = A_0 = I \succ 0$ and $x = 0$ is feasible. Next, Assumption 3 holds as

$$\Gamma = \left\{\gamma \in \mathbb{R}^2 : \begin{array}{l} 1 + \gamma_1 - \gamma_2 \geq 0 \\ 1 - \gamma_1 + \gamma_2 \geq 0 \\ \gamma \geq 0 \end{array}\right\}.$$

One can verify that

$$\Gamma = \operatorname{conv}(\{(0,0),(1,0),(0,1)\}) + \operatorname{cone}(\{1,1\}).$$

Finally, we note that $b_1 = b_2 = 0$. Hence, Corollary 3 and Remark 8 imply that

$$\operatorname{conv}(\mathcal{D}) = \mathcal{D}_{\text{SDP}} = \left\{(x,t) : \begin{array}{l} x_1^2 + x_2^2 + 10x_1 \leq 2t \\ 2x_1^2 + 10x_1 - 5 \leq 2t \\ 2x_2^2 + 10x_1 - 50 \leq 2t \end{array}\right\}.$$

We plot $\mathcal{D}$ and $\operatorname{conv}(\mathcal{D}) = \mathcal{D}_{\text{SDP}}$ in Figure 1.3. □

Remark 10 (Joint zero of a finite set of quadratic forms)**.** Barvinok [15] shows that one can decide in polynomial time (in N) whether a constant number, m_E, of quadratic forms $\{A_i\}_{i\in[m_E]}$ has a joint nontrivial zero. That is, whether the system $x^\intercal A_i x = 0$ for $i \in [m_E]$ and $x^\intercal x = 1$ is feasible. We can recast this as asking whether the following optimization problem

$$\min_x \left\{ -x^\intercal x : \begin{array}{l} x^\intercal x \leq 1 \\ x^\intercal A_i x = 0, \forall i \in [m_E] \end{array} \right\}$$

has objective value -1 or 0.

Thus, the feasibility problem studied in [15] reduces to a QCQP of the form we study in this chapter. Note that Assumption 1 for a QCQP of this form holds, for example, by taking $\gamma^* = 2e_1$ so that $A(\gamma^*) = -I + 2I \succ 0$ and noting that $x = 0$ is a feasible solution to this QCQP. Then when Γ is polyhedral (Assumption 3), Corollary 3 implies that the feasibility problem (in even a variable number of quadratic forms) can be decided using a semidefinite programming approach. Nevertheless, Assumption 3 may not necessarily hold, and thus Corollary 3 does not recover the full result of [15]. □

Corollary 4. *Suppose Assumption 1 holds and for every $i \in [0, m]$, there exists α_i such that $A_i = \alpha_i I_N$. If $m \leq N$, then*

$$\operatorname{conv}(\mathcal{D}) = \mathcal{D}_{\text{SDP}} \quad \textit{and} \quad \text{Opt} = \text{Opt}_{\text{SDP}} .$$

Proof. Assumption 3 holds in this case as

$$\Gamma := \left\{ \gamma \in \mathbb{R}^m : \begin{array}{l} A(\gamma) \succeq 0 \\ \gamma_i \geq 0,\ \forall i \in [m_I] \end{array} \right\} = \left\{ \gamma \in \mathbb{R}^m : \begin{array}{l} \alpha_0 + \sum_{i=1}^m \gamma_i \alpha_i \geq 0 \\ \gamma_i \geq 0,\ \forall i \in [m_I] \end{array} \right\}$$

is defined by $m_I + 1$ linear inequalities.

As each $A_i = \alpha_i I_N$, we have that the quadratic eigenvalue multiplicity satisfies $k = N$. By Observation 1, any semidefinite face of Γ must have affine dimension at most $m - 1$. In particular $\operatorname{aff}\dim(\{b(\gamma) : \gamma \in \mathcal{F}\}) + 1 \leq m$ and the assumption on the quadratic eigenvalue multiplicity in Theorem 2 holds as $k = N \geq m$. The final inequality $N \geq m$ holds by the assumptions of the corollary. ■

Remark 11. Consider the problem of finding the distance between the origin $0 \in \mathbb{R}^N$ and a piece of "Swiss cheese" $C \subseteq \mathbb{R}^N$. We will assume that C is nonempty and defined as

$$C = \left\{ x \in \mathbb{R}^N : \begin{array}{l} \|x - y_i\| \leq s_i,\ \forall i \in [m_1] \\ \|x - z_i\| \geq t_i,\ \forall i \in [m_2] \\ \langle x, b_i \rangle \geq c_i,\ \forall i \in [m_3] \end{array} \right\},$$

where $y_i, z_i, b_i \in \mathbb{R}^N$ and $s_i, t_i, c_i \in \mathbb{R}$ are arbitrary. In words, C is defined by m_1-many "inside-ball" constraints, m_2-many "outside-ball" constraints, and m_3-many linear inequalities. Note that each of these constraints may be written as a quadratic inequality with a quadratic form I, $-I$, or 0. In particular, Corollary 4 implies that if $m_1 + m_2 + m_3 \leq N$, then the value

$$\inf_{x \in \mathbb{R}^N} \left\{ \|x\|^2 : x \in C \right\}$$

may be computed using the standard SDP relaxation of the problem.

Bienstock and Michalka [26] give sufficient conditions under which a related problem

$$\inf_{x\in\mathbb{R}^N} \{q_0(x) : x \in C\},$$

is polynomial-time solvable. Here, $q_0 : \mathbb{R}^N \to \mathbb{R}$ is an arbitrary quadratic function but m_1 and m_2 are constant. Specifically, they devise an enumerative algorithm for problems of this form and prove its correctness under different assumptions. In contrast, our work deals only with the standard SDP relaxation and does not assume that the number of quadratic forms is constant.

Yang et al. [191] consider QCQPs with additional "hollow" type constraints. Formally, they consider a QCQP with domain $\mathcal{G} \coloneqq F \setminus \bigcup_\alpha \operatorname{int}(E_\alpha)$ where F is a quadratically constrained domain and $\{E_\alpha\}$ is a finite collection of *non-intersecting* ellipsoids completely contained within F. They show that if the SDP relaxation for a QCQP over the domain F is exact, then the SDP relaxation strengthened by additional linear constraints is exact for the same QCQP over the domain $\mathcal{G}$. In contrast, Corollary 4 makes no assumption on how the constraints defining $\mathcal{G}$ intersect but deals only with linearly many (in the dimension) spherical constraints. □

1.4.2 SD-QCQPs and the polyhedrality assumption

A natural class of QCQPs where Assumption 3 is immediately satisfied is the class of simultaneously diagonalizable QCQPs (SD-QCQPs) (see Definition 5 below). In this section, we first discuss how the simultaneously diagonalizable (SD) assumption relates to the polyhedrality assumption. Then, in Section 1.4.2, we show that under the SD assumption, the standard SDP relaxation is in fact equivalent to the lifted SOCP relaxation (both in terms of optimal value and projected epigraph). Consequently, our framework automatically generates sufficient conditions for SOCP-based tightness and convex hull results. Such sufficient conditions have been studied in the literature and we will compare our conditions with sufficient conditions proposed by Ben-Tal and den Hertog [21] and Locatelli [112] in Section 1.5.1.

Recall the following definition.

Definition 4. A set of matrices $\{A_i\}_{i\in[0,m]} \subseteq \mathbb{S}^N$ is said to be *simultaneously diagonalizable* (SD) if there exists an invertible matrix $U \in \mathbb{R}^{N\times N}$ such that the set $\{U^\intercal A_i U\}_{i\in[0,m]}$ consists of diagonal matrices. □

We note that this condition, sometimes referred to as simultaneously diagonalizable *by congruence*, is weaker than the notion of being simultaneously diagonalizable *by similarity* which further requires that U be an orthonormal matrix.

Definition 5. A *simultaneously diagonalizable QCQP* (SD-QCQP) is a QCQP of the form (1.1) where $\{A_i\}_{i\in[0,m]}$ is SD. □

Lemma 9. *For any SD-QCQP, we have that Γ is polyhedral.*

Proof. Let $U \in \mathbb{R}^{N\times N}$ be an invertible matrix such that $U^\intercal A_i U = \Lambda_i$ is diagonal for each $i \in [0, m]$. Note that $A(\gamma) \succeq 0$ if and only if $U^\intercal A(\gamma) U \succeq 0$ if and only if $\Lambda_0 + \sum_{i=1}^m \gamma_i \Lambda_i \succeq 0$. It is clear that

$$\Gamma = \left\{ \gamma \in \mathbb{R}^m : \begin{array}{l} \Lambda_0 + \sum_{i=1}^m \gamma_i \Lambda_i \succeq 0 \\ \gamma_i \geq 0, \ \forall i \in [m_I] \end{array} \right\}$$

is polyhedral. ■

The following example shows that changing a given constraint in a QCQP from an inequality into an equality constraint can alter whether Γ is polyhedral or not. As a consequence, we will deduce by Lemma 9 that Assumption 3 is strictly weaker than the simultaneous diagonalizability assumption.

Example 3. Consider the matrices

$$A_0 = \begin{pmatrix} 1 & & \\ & \sqrt{2} & 0 \\ & 0 & \sqrt{2} \end{pmatrix}, \quad A_1 = \begin{pmatrix} -1 & & \\ & 1 & 1 \\ & 1 & -1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} -1 & & \\ & 1 & -1 \\ & -1 & -1 \end{pmatrix}.$$

Note that $A(\gamma) \succeq 0$ if and only if each of its two blocks are positive semidefinite. Recall that a 2×2 matrix is positive semidefinite if and only if both its trace and determinant are nonnegative.

Suppose first that A_1 and A_2 correspond to equality constraints. Then

$$\begin{aligned} \Gamma &= \left\{ \gamma \in \mathbb{R}^2 : \begin{array}{l} 1 - \gamma_1 - \gamma_2 \geq 0 \\ (\sqrt{2} + (\gamma_1 + \gamma_2))(\sqrt{2} - (\gamma_1 + \gamma_2)) - (\gamma_1 - \gamma_2)^2 \geq 0 \end{array} \right\} \\ &= \left\{ \gamma \in \mathbb{R}^2 : \begin{array}{l} \gamma_1 + \gamma_2 \leq 1 \\ 2 - (\gamma_1 + \gamma_2)^2 - (\gamma_1 - \gamma_2)^2 \geq 0 \end{array} \right\} \\ &= \left\{ \gamma \in \mathbb{R}^2 : \begin{array}{l} \gamma_1 + \gamma_2 \leq 1 \\ \gamma_1^2 + \gamma_2^2 \leq 1 \end{array} \right\}. \end{aligned}$$

is not polyhedral (see Figure 1.4 left). In particular by Lemma 9, we deduce that the set $\{A_0, A_1, A_2\}$ is not simultaneously diagonalizable.

Now suppose that A_1 and A_2 correspond to inequality constraints. Then

$$\Gamma = \left\{ \gamma \in \mathbb{R}^2 : \begin{array}{l} \gamma_1 + \gamma_2 \leq 1 \\ \gamma_1^2 + \gamma_2^2 \leq 1 \\ \gamma \geq 0 \end{array} \right\} = \left\{ \gamma \in \mathbb{R}^2 : \begin{array}{l} \gamma_1 + \gamma_2 \leq 1 \\ \gamma \geq 0 \end{array} \right\}$$

is polyhedral (see Figure 1.4 right). Thus, we have constructed an example where the set $\{A_0, A_1, A_2\}$ is not simultaneously diagonalizable but Γ is polyhedral. We deduce that Assumption 3 is strictly weaker than the simultaneous diagonalizability assumption. □

Remark 12. Ramana [150] showed that deciding whether a given spectrahedron is polyhedral is coNP-hard. In particular, it is coNP-hard to decide whether Assumption 3 holds in general.

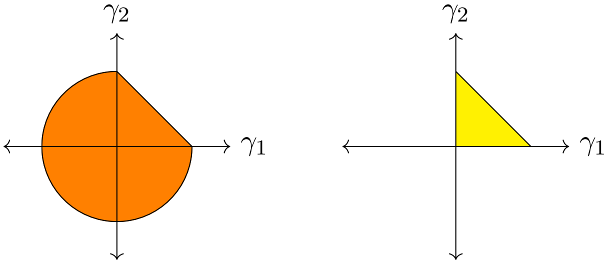

Figure 1.4: The set Γ with equality (orange) and inequality (yellow) constraints from Example 3

Nevertheless, it is possible to prove that this assumption holds for specific classes of interesting QCQPs (for example see Corollaries 2 and 4). □

The equivalence of SDP and SOCP relaxations of SD-QCQPs

Given an SD-QCQP and the invertible matrix U, we may perform a change of variables to arrive at a diagonal QCQP, i.e., a QCQP of the form (1.1) where each A_i is diagonal. In the remainder of this section, we assume that we have already made this change of variables and are left with

$$\inf_{x\in\mathbb{R}^N}\left\{q_0(x):\ \begin{array}{l} q_i(x)\le 0,\ \forall i\in[m_I]\\ q_i(x)=0,\ \forall i\in[m_I+1,m]\end{array}\right\}, \tag{1.10}$$

where $q_i(x)=\langle a_i,x^2\rangle+2\langle b_i,x\rangle+c_i$, $a_i\in\mathbb{R}^N$, $b_i\in\mathbb{R}^N$, and $c_i\in\mathbb{R}$ for each $i\in[0,m]$. Here, $x^2\in\mathbb{R}^N$ denotes the vector with $(x^2)_j=(x_j)^2$ for all $j\in[N]$.

Ben-Tal and den Hertog [21] and Locatelli [112] study the following SOCP relaxation

$$\mathrm{Opt}_{\mathrm{SOCP}}:=\inf_{x\in\mathbb{R}^N,\,y\in\mathbb{R}^N}\left\{\langle a_0,y\rangle+2\langle b_0,x\rangle+c_0:\ \begin{array}{r} \langle a_i,y\rangle+2\langle b_i,x\rangle+c_i\le 0,\\ \forall i\in[m_I]\\ \langle a_i,y\rangle+2\langle b_i,x\rangle+c_i=0,\\ \forall i\in[m_I+1,m]\\ y_j\ge x_j^2,\ \forall j\in[N]\end{array}\right\}. \tag{1.11}$$

Let $\mathcal{D}_{\mathrm{SOCP}}$ denote the epigraph of (1.11) projected onto the (x,t) variables, i.e., define

$$\mathcal{D}_{\mathrm{SOCP}}:=\left\{(x,t)\in\mathbb{R}^{N+1}:\ \begin{array}{l} \exists y\in\mathbb{R}^N:\\ \langle a_0,y\rangle+2\langle b_0,x\rangle+c_0\le 2t\\ \langle a_i,y\rangle+2\langle b_i,x\rangle+c_i\le 0,\ \forall i\in[m_I]\\ \langle a_i,y\rangle+2\langle b_i,x\rangle+c_i=0,\ \forall i\in[m_I+1,m]\\ y_j\ge x_j^2,\ \forall j\in[N]\end{array}\right\}. \tag{1.12}$$

The following proposition states that the SDP and the SOCP relaxations are equivalent for both the convex hull question and the tightness question. In particular, we may apply the sufficient conditions of this chapter for either result directly to the SOCP relaxation as well.

Proposition 1. *For any SD-QCQP, we have*

$$\mathcal{D}_{\text{SOCP}} = \mathcal{D}_{\text{SDP}} \quad \textit{and} \quad \text{Opt}_{\text{SOCP}} = \text{Opt}_{\text{SDP}}\,.$$

The second identity in Proposition 1 was first recorded by Locatelli [112]. The first identity, while straightforward, is to the best of our knowledge not present in the literature prior to our work. The proof of this result is deferred to Appendix A.1.

Remark 13. Remark 8 implies that for any SD-QCQP satisfying Assumption 1, the set $\mathcal{D}_{\text{SDP}}$ is SOC representable in the original space. However, this representation may potentially involve exponentially many quadratics—this follows as Γ may have exponentially many extreme points and rays. Moreover, identifying these extreme points and rays may require non-trivial computational effort. In contrast, Proposition 1 implies that $\mathcal{D}_{\text{SDP}} = \mathcal{D}_{\text{SOCP}}$ is SOCP representable in a lifted space (with only N new variables) using only linearly many convex quadratic constraints. Consequently, the $\mathcal{D}_{\text{SOCP}}$ representation is perhaps more interesting from a computational view. □

1.4.3 On the sharpness of Theorems 1 and 2

In this section we construct QCQPs that show that the assumptions made in Theorem 2 (and hence in Theorem 1) cannot be weakened individually.

We first examine the quadratic eigenvalue multiplicity assumption in Theorems 1 and 2, and show that both of these theorems break when the assumption on the lower bound on the value of the quadratic eigenvalue multiplicity k,

$$k \geq \operatorname{aff\,dim}(\{b(\gamma) : \gamma \in \mathcal{F}\}) + 1$$

is relaxed to $k \geq \operatorname{aff\,dim}(\{b(\gamma) : \gamma \in \mathcal{F}\})$.

Proposition 2. *For any positive integers n and k, there exists a QCQP in $N := nk$ variables with $m := k+1$ constraints such that*

- *Assumptions 1 and 3 are satisfied,*
- *the quadratic eigenvalue multiplicity of the QCQP is k, and*
- *k satisfies*

  $$k \geq \operatorname{aff\,dim}(\{b(\gamma) : \gamma \in \mathcal{F}\})$$

  for all semidefinite faces $\mathcal{F}$ of Γ, but
- $\text{Opt} \neq \text{Opt}_{\text{SDP}}$ *(and hence* $\operatorname{conv}(\mathcal{D}) \neq \mathcal{D}_{\text{SDP}}$*).*

Proof. Consider the following QCQP

$$\min_{x\in\mathbb{R}^N} \left\{ -x_1^2 - x_{n+1}^2 - \cdots - x_{(k-1)n+1}^2 : \begin{array}{l} \|x\|^2 - 1 \leq 0 \\ x_{(j-1)n+1} = 0,\ \forall j \in [k] \end{array} \right\}. \tag{1.13}$$

Here, $A_0 = I_k \otimes (-e_1 e_1^\intercal)$, $A_1 = I$, and $A_i = 0$ for all $i \in [m]$.

Assumption 1 holds because $A_1 = I \succ 0$ and $x = 0$ is feasible in (1.13). Moreover, Assumption 3 holds because

$$\Gamma \coloneqq \{\gamma \in \mathbb{R}^m : \gamma_1 \geq 0,\ A(\gamma) \succeq 0\} = \{\gamma \in \mathbb{R}^m : \gamma_1 \geq 1\}.$$

We compute: $\operatorname{aff\,dim}(\{b(\gamma) : \gamma_1 = 1\}) = k$.

By Lemma 1,

$$\mathrm{Opt}_{\mathrm{SDP}} = \min_{x \in \mathbb{R}^N} \sup_{\gamma \in \Gamma} q(\gamma, x) \leq \sup_{\gamma \in \Gamma} q(\gamma, 0) = -1.$$

On the other hand, it is clear from (1.13) that $\mathrm{Opt} = 0$. ■

We next provide a construction that illustrates that Theorems 1 and 2 both break when Assumption 3 is dropped.

Proposition 3. *There exists a QCQP in $n = 2$ variables with $m = 2$ constraints such that*

- *Assumptions 1 and 2 are satisfied,*
- *the quadratic eigenvalue multiplicity of the QCQP is $k = 1$, and*
- *k satisfies*

  $$k \geq \operatorname{aff\,dim}(\{b(\gamma) : \gamma \in \mathcal{F}\}) + 1$$

  for all semidefinite faces $\mathcal{F}$ of Γ, but
- $\mathrm{Opt} \neq \mathrm{Opt}_{\mathrm{SDP}}$ *(and hence* $\mathrm{conv}(\mathcal{D}) \neq \mathcal{D}_{\mathrm{SDP}}$*).*

Proof. Consider the following QCQP

$$\min_{x \in \mathbb{R}^2} \left\{ \|x - e_1\|^2 : \begin{array}{l} x_1^2 - x_2^2 + 2x_1x_2 = 0 \\ x_1^2 - x_2^2 - 2x_1x_2 = 0 \end{array} \right\}. \tag{1.14}$$

Here

$$A_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad A_1 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \qquad A_2 = \begin{pmatrix} 1 & -1 \\ -1 & -1 \end{pmatrix}.$$

Assumption 1 holds since $A(0) = I \succ 0$ and $x = 0$ is feasible in (1.14).

We will describe Γ explicitly. For a 2×2 matrix $A(\gamma)$, we have that $A(\gamma) \succeq 0$ if and only if $\mathrm{tr}(A(\gamma)) \geq 0$ and $\det(A(\gamma)) \geq 0$. Note that $\mathrm{tr}(A(\gamma)) = \mathrm{tr}(A_0) \geq 0$ for all γ, thus

$$\begin{aligned} \Gamma &= \left\{ \gamma \in \mathbb{R}^2 : (1 + \gamma_1 + \gamma_2)(1 - \gamma_1 - \gamma_2) - (\gamma_1 - \gamma_2)^2 \geq 0 \right\} \\ &= \left\{ \gamma \in \mathbb{R}^2 : 1 - 2\|\gamma\|^2 \geq 0 \right\} \\ &= \mathbf{B}(0, 2^{-1/2}). \end{aligned}$$

Then Assumption 2 holds as Γ is bounded.

It is clear that $k \geq 1$. To see that $k = 1$, note that A_1 has eigenvalues 1 and -1. Furthermore, as $b_1 = b_2 = 0$, we have that aff $\dim(\{b(\gamma) : \gamma \in \mathbb{R}^2\}) + 1 = 1$. In particular, the same is true for any semidefinite face $\mathcal{F}$ of Γ.

Next we compute $\mathrm{Opt}_{\mathrm{SDP}}$. By our explicit description of Γ, for any fixed $\hat{x}$ we have

$$\begin{aligned}\sup_{\gamma\in\Gamma} q(\gamma,\hat{x}) &= q_0(x) + \max_{\gamma\in\mathbf{B}(0,1/\sqrt{2})} \left\langle \gamma, \begin{pmatrix} q_1(x) \\ q_2(x)\end{pmatrix}\right\rangle \\ &= q_0(x) + \sqrt{(q_1(x)^2 + q_2(x)^2)/2} \\ &= q_0(x) + \|x\|^2.\end{aligned}$$

Then, by Lemma 1

$$\begin{aligned}\mathrm{Opt}_{\mathrm{SDP}} &= \min_x \sup_{\gamma\in\Gamma} q(\gamma, x) \\ &= \min_x \left(\|x - e_1\|^2 + \|x\|^2\right) \\ &= 1/2.\end{aligned}$$

On the other hand, it is clear from (1.14) that Opt $= 1$. ■

1.5 Exactness of the SDP relaxation

In this section, we use our framework to give new conditions under which $\mathrm{Opt}_{\mathrm{SDP}} = \mathrm{Opt}$.

Theorem 3. *Suppose Assumptions 1 and 3 hold. If for every semidefinite face $\mathcal{F}$ of Γ we have*

$$0 \notin \Pi_{\mathcal{V}(\mathcal{F})}\{b(\gamma) : \gamma \in \mathcal{F}\},$$

then any optimizer (x^*, t^*) in $\arg\min_{(x,t)\in\mathcal{D}_{\mathrm{SDP}}} 2t$ satisfies $(x^*, t^*) \in \mathcal{D}$. In particular, $\mathrm{Opt} = \mathrm{Opt}_{\mathrm{SDP}}$.

In other words, under the assumptions of Theorem 3, given any optimizer

$$\begin{pmatrix} 1 & x^\intercal \\ x & X\end{pmatrix}$$

of (1.3), we can simply return x as an optimizer for (1.1).

Proof. Let

$$(x^*, t^*) \in \operatorname*{arg\,min}_{(x,t)\in\mathcal{D}_{\mathrm{SDP}}} 2t.$$

Let $\mathcal{F} = \mathcal{F}(x^*)$. We claim that $\mathcal{F}$ will always be definite under the assumptions of this theorem. In particular, we will be able to apply Lemma 3 to conclude that $(x^*, t^*) \in \mathcal{D}$. To this end, we will show that $\mathcal{F}$ is definite by first assuming that $\mathcal{F}$ is semidefinite and then deriving a contradiction to the assumption that $(x^*, t^*) \in \arg\min_{(x,t)\in\mathcal{D}_{\text{SDP}}} 2t$.

Assume for contradiction that $\mathcal{F}$ is a semidefinite face of Γ. By Lemma 2, $\mathcal{V}(\mathcal{F})$ has a nonzero element. For the sake of convenience, let $\mathcal{P} := \Pi_{\mathcal{V}(\mathcal{F})}\{b(\gamma) : \gamma \in \mathcal{F}\}$. Assumption 3 implies that $\mathcal{P}$ is a nonempty closed convex set. Indeed, $\mathcal{P}$ is an affine transformation of $\mathcal{F}$, which is a face of the polyhedral set Γ, and is thus itself polyhedral.

Under our assumption, the compact set $\{0\}$ and the nonempty closed convex set $\mathcal{P}$ are disjoint. Thus, by the hyperplane separation theorem, there exists a nonzero vector $v \in \mathcal{V}(\mathcal{F})$ and $\epsilon > 0$ such that $v^\intercal b(\gamma) \leq -\epsilon$ for all $\gamma \in \mathcal{F}$.

Apply Lemma 8 to decompose $\Gamma = \Gamma_e + \text{cone}(\Gamma_r)$ and $\mathcal{F} = \mathcal{F}_e + \mathcal{F}_r$.

We will modify (x^*, t^*) in the $(v, -\epsilon)$ direction. Define

$$(x_\alpha, t_\alpha) := (x^* + \alpha v, t^* - \alpha\epsilon),$$

where $\alpha > 0$ will be chosen later.

First, consider how $q(\gamma, x_\alpha) - 2t_\alpha$ changes with α for fixed $\gamma_f \in \mathcal{F}$. We can expand

$$\begin{aligned} q(\gamma_f, x_\alpha) - 2t_\alpha &= (q(\gamma_f, x^*) - 2t^*) + 2\alpha(x^{*\intercal}A(\gamma_f)v + b(\gamma_f)^\intercal v + \epsilon) + \alpha^2 v^\intercal A(\gamma_f)v \\ &\leq (q(\gamma_f, x^*) - 2t^*) \\ &= 0. \end{aligned}$$

The second line follows as $A(\gamma_f)v = 0$ and $b(\gamma_f)^\intercal v \leq -\epsilon$ for all $\gamma_f \in \mathcal{F}$. The third line follows as $q(\gamma_f, x^*) = 2t^*$ for all $\gamma_f \in \mathcal{F}$.

On the other hand, for $\gamma_e \in \Gamma_e \setminus \mathcal{F}_e$, the function $\alpha \mapsto q(\gamma_e, x_\alpha) - 2t_\alpha$ is a continuous function taking the value $q(\gamma_e, x^*) - 2t^* < 0$ at $\alpha = 0$ (the strict inequality follows from the fact that $\gamma_e \in \Gamma_e \setminus \mathcal{F}_e$).

Similarly, for $\gamma_r \in \Gamma_r \setminus \mathcal{F}_r$, the function $\alpha \mapsto \breve{q}(\gamma_r, x_\alpha)$ is a continuous function taking the value $\breve{q}(\gamma_r, x^*) < 0$ at $\alpha = 0$ (the strict inequality follows from the fact that $\gamma_r \in \Gamma_r \setminus \mathcal{F}_r$).

We have shown that the following finite set of continuous functions in α,

$$\mathcal{Q} := \{q(\gamma_e, x_\alpha) - 2t_\alpha : \gamma_e \in \text{extr}(\Gamma_e) \setminus \mathcal{F}_e\} \cup \{\breve{q}(\gamma_r, x_\alpha) : \gamma_r \in \text{extr}(\Gamma_r) \setminus \mathcal{F}_r\},$$

consists of continuous functions which are negative at $\alpha = 0$. The finiteness of this set follows from the assumption that Γ is polyhedral.

Fix an $\alpha > 0$ such that $q(\alpha) \leq 0$ for every $q \in \mathcal{Q}$ — this is possible by the finiteness of $\mathcal{Q}$ and the continuity of each $q \in \mathcal{Q}$. Then because $\Gamma_e = \text{conv}(\text{extr}(\Gamma_e))$ and $\Gamma_r = \text{conv}(\text{extr}(\Gamma_r))$, we have $q(\gamma_e, x_\alpha) \leq 2t_\alpha$ for all $\gamma_e \in \Gamma_e$ and $\breve{q}(\gamma_r, x_\alpha) \leq 0$ for all $\gamma_r \in \Gamma_r$. Thus, $(x_\alpha, t_\alpha) \in \mathcal{D}_{\text{SDP}}$. In particular, $\min_{(x,t)\in\mathcal{D}_{\text{SDP}}} 2t \leq 2t_\alpha < 2t^*$, a contradiction. ■

The following theorem will follow from Theorem 3 by a perturbation argument.

Theorem 4. *Suppose Assumptions 1 and 3 hold. If there exists a sequence $(h_j)_{j\in\mathbb{N}}$ in $\mathbb{R}^N$ such that $\lim_{j\to\infty} h_j = 0$ and for every semidefinite face $\mathcal{F}$ of Γ and $j \in \mathbb{N}$ we have*

$$0 \notin \Pi_{\mathcal{V}(\mathcal{F})}\{b(\gamma) + h_j : \gamma \in \mathcal{F}\},$$

then $\mathrm{Opt} = \mathrm{Opt}_{\mathrm{SDP}}$.

Proof. Consider the following sequence of QCQPs indexed by $j \in \mathbb{N}$:

$$\mathrm{Opt}_j := \min_{x\in\mathbb{R}^N}\left\{ q_0(x) + 2h_j^\intercal x : \begin{array}{l} q_i(x) \le 0,\ \forall i \in [m_I] \\ q_i(x) = 0,\ \forall i \in [m_I+1, m] \end{array} \right\}.$$

We will use the subscript j to denote all quantities corresponding to the perturbed QCQP. By construction, each of the QCQPs in this sequence satisfies the assumptions of Theorem 3 and thus $\mathrm{Opt}_{\mathrm{SDP},j} = \mathrm{Opt}_j$. For $j \in \mathbb{N}$, let

$$(x_j, t_j) \in \operatorname*{arg\,min}_{(x,t)\in\mathcal{D}_j} 2t.$$

Let x^* be a subsequential limit of $\{x_j\}_{j\in\mathbb{N}}$ (this exists as we can bound the sequence $\{x_j\}_{j\in\mathbb{N}}$ using Assumption 1). Noting that the feasible domain of the original QCQP is closed, we have that x^*, a subsequential limit of feasible points, is also feasible. Finally, by continuity of q_0 and the optimality of $(x_j, t_j) \in \mathcal{D}_j$, we have that

$$q_0(x^*) = \lim_{j\to\infty} q_0(x_j) = \lim_{j\to\infty} \mathrm{Opt}_j = \lim_{j\to\infty} \mathrm{Opt}_{\mathrm{SDP},j} = \mathrm{Opt}_{\mathrm{SDP}}\,.$$

Here, the final equality holds by a simple boundedness argument and Assumption 1. ■

The following example shows that SDP tightness (for example via Theorem 4) may hold even when the convex hull result does not.

Example 4. Consider the following QCQP

$$\inf_{x\in\mathbb{R}^2}\left\{ x_1^2 + x_2^2 : \begin{array}{l} x_1^2 - x_2^2 \le 0 \\ 2x_2 \le 0 \end{array} \right\}.$$

We verify that the conditions of Theorem 4 hold. It is clear that Assumption 1 holds: $A(0) = I \succ 0$ and $x = 0$ is feasible. It is easy to verify that $\Gamma = [0,1] \times \mathbb{R}_+$, thus Assumption 3 also holds. Finally, pick $h_j = e_2/j$ for $j \in \mathbb{N}$. Note that the only semidefinite face of Γ is $\mathcal{F} = \{1\} \times \mathbb{R}_+$ and that $\mathcal{V}(\mathcal{F}) = \operatorname{span}\{e_2\}$. In particular,

$$\Pi_{\mathcal{V}(\mathcal{F})}\{b(\gamma) + h_j : \gamma \in \mathcal{F}\} = \{0\} \times [1/j, \infty),$$

which does not contain 0. We deduce that $\mathrm{Opt} = \mathrm{Opt}_{\mathrm{SDP}}$.

Next, we claim that $\operatorname{conv}(\mathcal{D}) \neq \mathcal{D}_{\mathrm{SDP}}$. First note that $\mathcal{D}$ is actually convex in this example.

$$\mathcal{D} = \left\{ (x,t) : \begin{array}{l} x_1^2 + x_2^2 \le 2t \\ x_1^2 - x_2^2 \le 0 \\ 2x_2 \le 0 \end{array} \right\} = \left\{ (x,t) : \begin{array}{l} x_1^2 + x_2^2 \le 2t \\ |x_1| \le -x_2 \\ 2x_2 \le 0 \end{array} \right\}$$

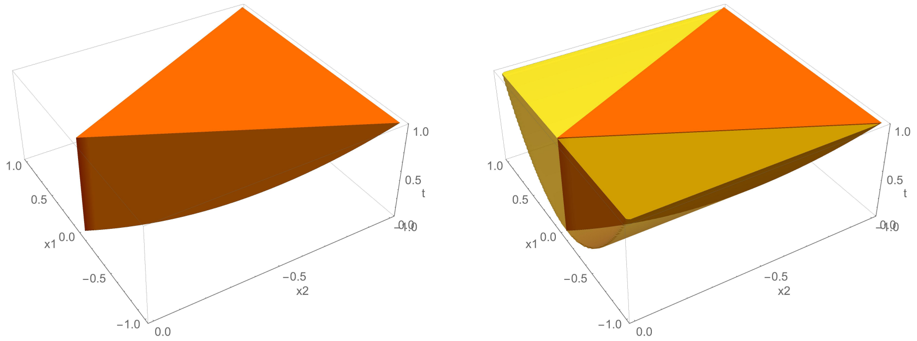

Figure 1.5: The sets $\text{conv}(\mathcal{D})$ (in orange) and $\mathcal{D}_{\text{SDP}}$ (in yellow) from Example 4

Next by Lemma 1 and the description of Γ above, we have that

$$\mathcal{D}_{\text{SDP}} = \left\{ (x,t) : \begin{array}{l} x_1^2 + x_2^2 \leq 2t \\ 2x_1^2 \leq 2t \\ 2x_2 \leq 0 \end{array} \right\}.$$

Then we may check, for example, that

$$((1,0),1) \in \mathcal{D}_{\text{SDP}} \qquad \text{but} \qquad ((1,0),1) \notin \mathcal{D} = \text{conv}(\mathcal{D}).$$

We conclude that $\text{Opt} = \text{Opt}_{\text{SDP}}$ but $\text{conv}(\mathcal{D}) \neq \mathcal{D}_{\text{SDP}}$. We plot $\mathcal{D}$ and $\mathcal{D}_{\text{SDP}}$ in Figure 1.5. □

1.5.1 Comparison with related conditions in the literature

Several sufficient conditions for SDP tightness results have been examined in the literature. In this section, we compare these conditions with our Theorems 3 and 4.

Locatelli [112] considers the SDP relaxation of a variant of the TRS,

$$\inf_{x\in\mathbb{R}^N} \left\{ q_0(x) : \begin{array}{l} b_i^\intercal x + c_i \leq 0,\ \forall i \in [m-1] \\ x^\intercal x - 1 \leq 0 \end{array} \right\}. \tag{1.15}$$

We assume that $A_0 = \text{Diag}(a_0)$ without loss of generality. Indeed, if A_0 is not diagonal, we can reformulate the problem in the eigenbasis of A_0. Furthermore, we will assume that A_0 has at least one negative eigenvalue as otherwise (1.15) is already convex.

Let $J \subseteq [N]$ be the set of coordinates corresponding to $\lambda_{\min}(A_0)$, i.e., define

$$J \coloneqq \left\{ j \in [N] : (a_0)_j = \min_{i\in[N]} (a_0)_i \right\},$$

and let $\mathcal{V}_J \coloneqq \text{span}(\{e_j : j \in J\})$.

Locatelli [112] derives a sufficient condition for SDP tightness by reasoning about the nonexistence of certain KKT multipliers in the SOCP relaxation of (1.15). For the sake of completeness, we restate this result in our language.

Theorem 5 ([112, Theorem 3.1])**.** *Consider the problem* (1.15) *and assume that A_0 has at least one negative eigenvalue. Suppose the feasible region of* (1.15) *is strictly feasible. If there exists a sequence $(h_j)_{j\in\mathbb{N}}$ in $\mathbb{R}^N$ such that $\lim_{j\to\infty} h_j = 0$ and for every $j \in \mathbb{N}$ we have*

$$0 \notin \Pi_{\mathcal{V}_J}\{b(\gamma) + h_j : \gamma \in \mathbb{R}^m_+\},$$

then $\mathrm{Opt} = \mathrm{Opt}_{\mathrm{SDP}}$.

Proposition 4. *Suppose the assumptions of Theorem 5 hold, then the assumptions of Theorem 4 also hold.*

Proof. Consider a QCQP of the form (1.15) satisfying the assumptions of Theorem 5. We will verify that the assumptions of Theorem 4 are also satisfied. Note the feasible region of (1.1) is nonempty. Furthermore, by taking $\eta \in \mathbb{R}$ large enough, we can ensure that $A(\eta e_m) = A_0 + \eta I \succ 0$. Thus, Assumption 1 is satisfied. Assumption 3 is satisfied as well because

$$\Gamma = \left\{\gamma \in \mathbb{R}^m : \begin{array}{l} A(\gamma) \succeq 0 \\ \gamma \geq 0 \end{array}\right\} = \left\{\gamma \in \mathbb{R}^m : \begin{array}{l} \gamma_m \geq -\lambda_{\min}(A_0) \\ \gamma \geq 0 \end{array}\right\} \tag{1.16}$$

is polyhedral.

Let $\mathcal{F}$ be a semidefinite face of Γ. By Lemma 2, $A(\gamma)$ must have a zero eigenvalue for every $\gamma \in \mathcal{F}$. In particular, we can deduce from the description of Γ in (1.16) that

$$\mathcal{F} = \left\{\gamma \in \mathbb{R}^m : \begin{array}{l} \gamma_m = -\lambda_{\min}(A_0) \\ \gamma \geq 0 \end{array}\right\}.$$

Therefore, $\mathcal{V}(\mathcal{F}) = \mathcal{V}_J$. Then the assumption $0 \notin \Pi_{\mathcal{V}_J}\{b(\gamma) + h_j : \gamma \in \mathbb{R}^m_+\}$ for every $j \in \mathbb{N}$ immediately implies that

$$0 \notin \Pi_{\mathcal{V}(\mathcal{F})}\{b(\gamma) + h_j : \gamma \in \mathcal{F}\}$$

for every $j \in \mathbb{N}$ as $\mathbb{R}^m_+ \supseteq \mathcal{F}$. Hence, we conclude that the third condition in Theorem 4 also holds. ■

Remark 14. Ho-Nguyen and Kılınç-Karzan [87] study a particular convex relaxation of the TRS with additional conic constraints. For such problems, they suggest a particular assumption under which their relaxation is tight; see [87, Theorem 2.4]. It was also shown in [87, Lemma 2.10] that when the conic constraints are in a particular linear form, then their assumption is indeed an equivalent form of Locatelli [112]'s assumption from Theorem 5. It is of interest to compare our assumptions with the one from [87]. We note however that our Theorem 4 and the result due to [87, Theorem 2.4] are incomparable. To see this, note that the former covers some optimization problems with nonconvex quadratic constraints while the latter covers some optimization problems with non-quadratic conic constraints. In addition, we note that the relaxation studied in Ho-Nguyen and Kılınç-Karzan [87] is weaker than the SDP relaxation that we study here. □

Burer and Ye [38] consider the standard SDP relaxation of diagonal QCQPs[4] and show that under an assumption on the input data $\{A_i\}_{i\in[0,m]}$ and $\{b_i\}_{i\in[0,m]}$ that the SDP relaxation is tight. For the sake of completeness, we first restate[5] [38, Theorem 1] as it relates to SDP tightness in our language.

Theorem 6 ([38, Theorem 1])**.** *Consider a diagonal QCQP with no equality constraints. Suppose the feasible region of* (1.1) *is nonempty and there exists $\gamma^* \geq 0$ such that $\breve{A}(\gamma^*) \succ 0$. Suppose the SDP relaxation* (1.3) *is strictly feasible. If for every $j \in [N]$ the set*

$$\left\{ \gamma \in \mathbb{R}^m : \begin{array}{l} \gamma \geq 0 \\ A(\gamma) \succeq 0 \\ A(\gamma)_{j,j} = 0 \\ b(\gamma)_j = 0 \end{array} \right\}$$

is empty, then any optimizer (x^*, t^*) in $\arg\min_{(x,t)\in\mathcal{D}_{\text{SDP}}} 2t$ satisfies $(x^*, t^*) \in \mathcal{D}$.

Proposition 5. *Suppose the assumptions of Theorem 6 hold, then the assumptions of Theorem 3 also hold.*

Proof. Consider a QCQP satisfying the assumptions of Theorem 6. We will verify that the assumptions of Theorem 3 are also satisfied. Note the feasible region of (1.1) is nonempty. Furthermore, by taking $\eta \in \mathbb{R}$ large enough, we can ensure $A(\eta\gamma^*) = A_0 + \eta\breve{A}(\gamma^*) \succ 0$. Thus, Assumption 1 is satisfied. Assumption 3 holds as all of the quadratic forms $A_0, \ldots, A_m$ are diagonal. The condition on the input data in Theorem 6 is equivalent to requiring that

$$A(\gamma)_{j,j} = 0 \implies b(\gamma)_j \neq 0$$

for all $\gamma \in \Gamma$ and $j \in [N]$. Consider a semidefinite face $\mathcal{F}$ of Γ, and any $\gamma \in \mathcal{F}$. As $A(\gamma)$ is diagonal, we deduce that

$$\mathcal{V}(\mathcal{F}) = \operatorname{span}(\{e_j : A(\gamma)_{j,j} = 0\}).$$

Then, the final assumption in Theorem 3, namely

$$0 \notin \Pi_{\mathcal{V}(\mathcal{F})}\{b(\gamma) : \gamma \in \mathcal{F}\},$$

holds immediately. ■

The following example shows that Theorem 3 is strictly more general than Theorem 6 even in the case of diagonal QCQPs with strictly convex constraints.

[4] Burer and Ye [38] address general QCQPs in their paper by first transforming them into diagonal QCQPs and then applying the standard SDP relaxation. In particular, the standard Shor SDP relaxation is only analyzed in the context of diagonal QCQPs.

[5] The original statement of this theorem gives additional guarantees, which are weaker than SDP tightness, when the conditions of Theorem 6 fail.

Example 5. Consider the following QCQP

$$\min_{x\in\mathbb{R}^2}\left\{-\|x\|^2 : \begin{array}{l}\|x-e_1\|^2 \leq 1\\ \|x-e_2\|^2 \leq 1\end{array}\right\}.$$

We first verify that the assumptions of Theorem 3 hold. It is clear that this problem satisfies Assumption 1: the origin is feasible and $A(e_1+e_2) = I \succ 0$. Next, we compute Γ.

$$\Gamma = \left\{\gamma \in \mathbb{R}^2_+ : A(\gamma) \succeq 0\right\} = \left\{\gamma \in \mathbb{R}^2_+ : \gamma_1+\gamma_2 \geq 1\right\}.$$

We conclude that Assumption 3 also holds. Furthermore, the only semidefinite face of Γ is $\mathcal{F} = \left\{\gamma \in \mathbb{R}^2_+ : \gamma_1+\gamma_2 = 1\right\}$. For this semidefinite face, we have that $\mathcal{V}(\mathcal{F})$ is the entire space $\mathbb{R}^2$. Consequently,

$$\Pi_{\mathcal{V}(\mathcal{F})}\{b(\gamma) : \gamma \in \mathcal{F}\} = \left\{\gamma_1 e_1 + \gamma_2 e_2 : \gamma \in \mathbb{R}^2_+,\, \gamma_1+\gamma_2 = 1\right\}$$

is the set of all convex combinations of e_1 and e_2. This set does not contain the origin and thus the assumptions of Theorem 3 are satisfied.

On the other hand, by picking $j = 1$ in Theorem 6 and $\gamma = e_2$, we have that $\gamma \geq 0$, $A(\gamma) \succeq 0$, and $A(\gamma)_{j,j} = 0$ but $b(\gamma)_j = (e_2)_1 = 0$. We see that the assumptions of Theorem 6 are not satisfied. □

1.6 Removing the polyhedrality assumption

One of the main assumptions we use in our proof of the convex hull results (Theorems 1 and 2) and the SDP tightness results (Theorems 3 and 4) is that the set Γ is polyhedral (Assumption 3). In this section we show that one can remove Assumption 3 in Theorem 2 when k is sufficiently large[6]. The results in this section do not use the framework described in Section 1.2 and in particular do not require the technical assumption (Assumption 2).

Theorem 7. *Suppose Assumption 1 holds. If the quadratic eigenvalue multiplicity k satisfies $k \geq m+2$, then* $\operatorname{conv}(\mathcal{D}) = \mathcal{D}_{\text{SDP}}$.

Proof. Suppose $(\hat{x},\hat{t}) \in \mathcal{D}_{\text{SDP}}$. Therefore,

$$\begin{aligned} 2\hat{t} &\geq \sup_{\gamma\in\mathbb{R}^m}\left\{q(\gamma,\hat{x}) : \begin{array}{l} A(\gamma)\succeq 0\\ \gamma_i \geq 0,\ \forall i \in [m_I]\end{array}\right\}\\ &= \sup_{\gamma\in\mathbb{R}^m}\left\{q(\gamma,\hat{x}) : \begin{array}{l} \mathcal{A}(\gamma)\succeq 0\\ \gamma_i \geq 0,\ \forall i \in [m_I]\end{array}\right\}. \end{aligned}$$

[6] Recall the example constructed in Proposition 3. This example shows that both the convex hull result and SDP tightness result fail when Assumption 3 is dropped from Theorem 2. In particular, the SDP tightness and convex hull results we recover in this section will require assumptions on k that are strictly stronger than in the polyhedral case.

The second line follows as $A(\gamma) \succeq 0$ if and only if $\mathcal{A}(\gamma) \succeq 0$. Note that Assumption 1 allows us to apply strong conic duality to the program on the second line. Furthermore, this dual SDP achieves its optimal value, i.e., there exists $Z \in \mathbb{S}^n$ such that $(\hat{x}, \hat{t}, Z)$ satisfies

$$\begin{cases} q_0(\hat{x}) + \langle \mathcal{A}_0, Z\rangle \leq 2\hat{t} \\ q_i(\hat{x}) + \langle \mathcal{A}_i, Z\rangle \leq 0,\ \forall i \in [m_I] \\ q_i(\hat{x}) + \langle \mathcal{A}_i, Z\rangle = 0,\ \forall i \in [m_I + 1, m] \\ Z \succeq 0. \end{cases} \tag{1.17}$$

We will show by induction on $\operatorname{rank}(Z)$ that for any $(\hat{x}, \hat{t}, Z)$ satisfying (1.17), we have $(\hat{x}, \hat{t}) \in \operatorname{conv}(\mathcal{D})$. The claim clearly holds when $\operatorname{rank}(Z) = 0$.

Now suppose $r := \operatorname{rank}(Z) \geq 1$. Let $(\hat{x}, \hat{t}, Z)$ satisfy (1.17). Write $Z = \sum_{i=1}^r z_i z_i^\intercal$ where each z_i is nonzero. Fix $z := z_1$.

We claim that the following system in y is feasible:

$$\begin{cases} \langle A_0\hat{x} + b_0, y \otimes z\rangle = 0 \\ \langle A_i\hat{x} + b_i, y \otimes z\rangle = 0,\ \forall i \in [m] \\ y \in \mathbf{S}^{k-1}. \end{cases} \tag{1.18}$$

Indeed, the first two constraints impose at most $m+1$ homogeneous linear equalities in $k \geq m+2$ variables. In particular, there exists a nonzero solution y to the first two constraints. This y may then be scaled to satisfy $y \in \mathbf{S}^{k-1}$.

Note then that for all $i \in [0, m]$,

$$\begin{aligned} q_i(\hat{x} \pm y \otimes z) &= (\hat{x} \pm y \otimes z)^\intercal A_i (\hat{x} \pm y \otimes z) + 2b_i^\intercal(\hat{x} \pm y \otimes z) + c_i \\ &= q_i(\hat{x}) \pm 2\langle A_i\hat{x} + b_i, y \otimes z\rangle + \langle \mathcal{A}_i, zz^\intercal\rangle \\ &= q_i(\hat{x}) + \langle \mathcal{A}_i, zz^\intercal\rangle. \end{aligned}$$

Consequently, $(\hat{x} \pm y \otimes z, \hat{t}, Z - zz^\intercal)$ satisfies (1.17). Furthermore, we have $\operatorname{rank}(Z - zz^\intercal) = r - 1$. By induction, $(\hat{x} \pm y \otimes z, \hat{t}) \in \operatorname{conv}(\mathcal{D})$. We conclude that $(\hat{x}, \hat{t}) \in \operatorname{conv}(\mathcal{D})$. ■

A similar proof leads to an SDP tightness result without Assumption 3.

Theorem 8. *Suppose Assumption 1 holds. Define the hyperplane* $H = \left\{(x, t) \in \mathbb{R}^{N+1} :\ 2t = \mathrm{Opt}_{\mathrm{SDP}}\right\}$. *If the quadratic eigenvalue multiplicity k satisfies* $k \geq m + 1$, *then* $\operatorname{conv}(\mathcal{D} \cap H) = \mathcal{D}_{\mathrm{SDP}} \cap H$. *In particular,* $\mathrm{Opt} = \mathrm{Opt}_{\mathrm{SDP}}$.

The proof of this statement follows the proof of Theorem 7 almost exactly and is deferred to Appendix A.2. For now, we will simply sketch how to modify the proof of Theorem 7 to get a proof for Theorem 8: We will only consider points $(\hat{x}, \hat{t}) \in \mathcal{D}_{\mathrm{SDP}} \cap H$. In this situation, it is easy to show that the first two constraints in (1.18) are dependent and impose at most m homogeneous linear equalities. Thus we may carry out the procedure in the proof of Theorem 7 as long as $k \geq m + 1$. At the end of the procedure, we will have decomposed $(\hat{x}, \hat{t})$ as a convex combination of points $(x_\alpha, \hat{t}) \in \mathcal{D}$.

Remark 15. Beck [17, Corollary 4.4] shows that under Assumption 1, the conclusion $\mathrm{Opt} = \mathrm{Opt}_{\mathrm{SDP}}$ holds even when $k = m$. Thus, recalling the definition of H from Theorem 8, we can summarize Theorems 7 and 8 and [17, Corollary 4.4] as follows. Under Assumption 1, we have:

Assumption	Result	Reference
$k \geq m + 2$	$\mathrm{conv}(\mathcal{D}) = \mathcal{D}_{\mathrm{SDP}}$	Theorem 7
$k \geq m + 1$	$\mathrm{conv}(\mathcal{D} \cap H) = \mathcal{D}_{\mathrm{SDP}} \cap H$	Theorem 8
$k \geq m$	$\mathcal{D} \cap H \neq \varnothing$	[17, Corollary 4.4]

We conjecture, but are unable to prove at the moment, that the values required of k for these three results are sharp. □

Remark 16. We vastly generalize the framework of this chapter in Chapter 2 where we replace the polyhedrality assumption with a *facially exposed* condition. We furthermore improve Theorem 7 from $k \geq m + 2$ to $k \geq m$.
In retrospect, the fact that $k \geq m$ should suffice for convex hull exactness seems obvious: Intuitively, convex hull exactness asks whether objective value exactness holds for any choice of the linear term in the objective b_0. Thus, as objective value exactness holds [17, Corollary 4.4] for $k \geq m$ conditioned only on Assumption 1 (and irregardless of b_0), it should follow that convex hull exactness holds for $k \geq m$. □

2 A geometric view of SDP exactness in QCQPs and its applications

This chapter is based on joint work [179] with Fatma Kılınç-Karzan.

This chapter extends the work of Chapter 1 towards understanding objective value and convex hull exactness and completely removes the polyhedrality assumption made in Chapter 1. In this chapter, we view the *cone* of convex Lagrange multipliers

$$\Gamma \coloneqq \left\{ (\gamma_{\text{obj}}, \gamma) \in \mathbb{R} \times \mathbb{R}^m : \begin{array}{l} \gamma_{\text{obj}} A_{\text{obj}} + \sum_{i=1}^m \gamma_i A_i \\ \gamma_i \geq 0,\ \forall i \in [m_I] \end{array} \right\}$$

as being the natural dual object to study. We then define the slice $\Gamma_1 \coloneqq \{\gamma \in \mathbb{R}^m : (1, \gamma) \in \Gamma\}$. Note that the definition of Γ in Chapter 1 coincides with the definition of Γ_1 in the present chapter. Our conditions for exactness are again based on geometric properties of Γ_1 and its relatives Γ and Γ°. These tools form the basis of our main message: questions of exactness can be treated systematically whenever Γ_1, Γ, or Γ° is well-understood. As further evidence of this message, we apply our tools to address questions of exactness for a prototypical QCQP involving a binary on-off constraint, quadratic matrix programs, the QCQP formulation of the partition problem, and random and semi-random QCQPs.

2.1 Introduction

Quadratically constrained quadratic programs (QCQPs) are a fundamental class of *nonconvex* optimization problems of the form

$$\text{Opt} \coloneqq \inf_{x \in \mathbb{R}^n} \left\{ q_{\text{obj}}(x) : \begin{array}{l} q_i(x) \leq 0,\ \forall i \in [m_I] \\ q_i(x) = 0,\ \forall i \in [m_I + 1, m] \end{array} \right\},$$

where $q_{\text{obj}}, q_1, \dots, q_m : \mathbb{R}^n \to \mathbb{R}$ are each (possibly nonconvex) quadratic functions. For each $i \in [m]$, we will write $q_i(x) = x^\top A_i x + 2b_i^\top x + c_i$ for $A_i \in \mathbb{S}^n$, $b_i \in \mathbb{R}^n$, and $c_i \in \mathbb{R}$. Similarly, write $q_{\text{obj}}(x) = x^\top A_{\text{obj}} x + 2b_{\text{obj}}^\top x + c_{\text{obj}}$.

These optimization problems arise naturally in a variety of application areas (see [11, 22, 181]). Indeed, one fundamental reason for the ubiquity of QCQPs is their expressiveness—any polynomial optimization problem or $\{0, 1\}$-integer program may be reformulated as a QCQP.

Although QCQPs are NP-hard in general, they admit a natural tractable convex relaxation known as the standard semidefinite program (SDP) relaxation [161],

$$\mathrm{Opt}_{\mathrm{SDP}} := \inf_{x\in\mathbb{R}^n}\left\{\left\langle A_{\mathrm{obj}}, X\right\rangle + 2b_{\mathrm{obj}}^\top x + c_{\mathrm{obj}} :\ \begin{array}{l} \exists X \succeq xx^\top : \\ \langle A_i, X\rangle + 2b_i^\top x + c_i \leq 0,\ \forall i\in[m_I] \\ \langle A_i, X\rangle + 2b_i^\top x + c_i = 0,\ \forall i\in[m_I+1, m] \end{array}\right\}.$$

This relaxation is also referred to as the Shor SDP relaxation. In contrast to the vast literature on the *approximation* quality of this relaxation [22, 121, 129, 193], the question of when *exactness* occurs in this relaxation is much more limited and recent.

One interesting line of work has offered deterministic conditions under which the SDP relaxation of a general QCQP is exact for various definitions of exactness. In their celebrated paper, Fradkov and Yakubovich [67] prove the S-lemma, which implies that the problem of minimizing an arbitrary quadratic objective function over the unit ball (or any single quadratic constraint) can be solved via SDP techniques. Specifically, the S-lemma implies that *objective value exactness*—the condition that the optimal value of the QCQP and the optimal value of its SDP relaxation coincide—holds for QCQPs with a single constraint; see also [180]. In contrast, Burer and Ye [38] study diagonal QCQPs—those QCQPs for which $A_{\mathrm{obj}}, A_1, \ldots, A_m$ are diagonal matrices—with a *general* number of constraints and give sufficient conditions for objective value exactness. Wang and Kılınç-Karzan [178, 181] continue this line of work by developing a general framework for deriving sufficient conditions for both objective value exactness and *convex hull exactness*—the condition that the convex hull of the QCQP epigraph coincides with the (projected) SDP epigraph—for QCQPs where a specific dual set, Γ_1, is polyhedral (see Section 2.2). Beyond being a natural sufficient condition for objective value exactness, convex hull exactness has its own far-reaching applications and motivation. Such results find use for example in deriving strong relaxations of certain critical substructures in nonconvex problems. Specifically, the convexification of commonly occurring substructures in complex nonconvex problems has been critical in advancing the state-of-the-art computational approaches for general nonlinear nonconvex programs and mixed integer linear programs [49, 171]. (See [7, 163, 181] and references therein for additional work in this direction.)

While the framework presented by Wang and Kılınç-Karzan [181] can at once cover and extend many existing results on objective value and convex hull exactness [35, 38, 67, 87, 111, 123, 180, 195], it is still quite limited. In particular, the assumption that Γ_1 is polyhedral is rarely satisfied outside of simultaneously diagonalizable QCQPs and precludes the results in [181] from being applicable to a wider range of interesting QCQPs.

Additional work in this direction [48] studies objective value exactness from an algebraic point of view. Specifically, Cifuentes et al. [48] consider QCQPs with fixed equality constraints and study the semialgebraic region of objective functions for which objective value exactness holds. As an example of their results, they give a formula for the degree of the algebraic boundary of this region in the setting of Euclidean distance minimization problems.

A related line of work has explored sufficient conditions for the *rank-one-generated* (ROG) property [7, 29, 33, 34, 83]. A conic subset of the positive semidefinite cone is said to be ROG if it is the convex hull of its rank-one elements. This property can be thought of as the SDP–QCQP analogue to the integrality property in the context of linear program relaxations of integer linear programs [7] and can be shown to imply both convex hull exactness and objective value exactness. Research in this direction has established explicit descriptions of the ROG cones related to quadratic programs over low-dimensional polytopes [34] and ellipsoids with missing caps [37]. Other work in this direction [29, 83] explores the ROG property from an algebro-geometric perspective and establishes results related to the degree and representation of such sets. Importantly, Blekherman et al. [29] completely characterized the ROG cones defined by linear matrix *equalities*. More recently, Argue et al. [7] gave general sufficient conditions for this property and completely characterized the ROG cones defined by at most two linear matrix *inequalities*.

See [100] for an overview and comparison of objective value exactness, convex hull exactness, and the rank-one generated property.

SDP exactness has been studied in the context of quadratic matrix programs (QMPs) as well. A QMP is an optimization problem over a matrix variable $X \in \mathbb{R}^{r\times k}$, where the objective function and constraints are each of the form

$$\operatorname{tr}(X^\top AX) + 2\operatorname{tr}(B^\top X) + c$$

for $A \in \mathbb{S}^r$, $B \in \mathbb{R}^{r\times k}$ and $c \in \mathbb{R}$, and can be thought of as a natural generalization to QCQPs.[1] This class of problems has been used to model robust least squares problems, the orthogonal Procrustes problem [17], and sphere packing [20]. QMPs and their SDP relaxations were first studied by Beck [17], Beck et al. [20] who showed that objective value exactness holds as long as the number of constraints is small compared to k. Similarly, Wang and Kılınç-Karzan [181] show that both objective value exactness and convex hull exactness hold for (vectorized reformulations of) QMPs whenever the number of constraints is small enough and Γ_1 is polyhedral.

Finally, a number of results have shown that various *random* QCQPs have exact SDP relaxations with high probability. For example, such results have been proved for signal-recovery tasks such as phase retrieval [40], sensor-network localization [157], max-likelihood angular synchronization [10], and clustering [1, 122, 153]. In these settings, the goal is to recover some ground-truth solution (the solution to some QCQP) via observations (constraints in a QCQP). These results then show that once an application-specific signal-to-noise ratio is large enough (for example, given enough observations/constraints), that the SDP relaxation is exact. In contrast, a second line of work [38, 113] addresses random QCQPs which do not assume the existence of a ground-truth solution. In this direction, it is shown that when the number of constraints is *small enough* that the SDP relaxation has a rank-one optimal solution.

2.1.1 Overview and outline of the chapter

In this chapter, we generalize the framework first introduced in [178, 181] by eliminating its reliance on the polyhedrality assumption. Specifically, we give a broad set of sufficient conditions for both convex hull exactness and objective value exactness that are phrased in terms of the *cone of convex Lagrange multipliers* Γ (or the closely related sets Γ_1 and Γ°; see Section 2.2). In particular, these sufficient conditions can be checked in a systematic manner whenever Γ, Γ_1, or Γ° is sufficiently simple. Furthermore, we show that our sufficient conditions for convex hull exactness

[1] In fact, these two problem classes are equivalent. Any QCQP is a QMP with $k = 1$. In the reverse direction, any QMP in a variable $X \in \mathbb{R}^{n\times k}$ can be written as a QCQP in the variable obtained by stacking the columns of X on top of each other. Note that in this second direction, the value k induces additional structure on the QCQP.

are additionally *necessary* under a technical assumption (see Assumption 5). We complement our theory with a number of explicit examples illustrating our tools on QCQPs from various settings, including a basic QCQP originating from modeling big-M constraints, quadratic matrix programs, the partition QCQP, and two random QCQP models.

Collectively, these results and examples offer evidence for the main message of this chapter that *questions of exactness can be treated systematically whenever the convex Lagrange multipliers are well-understood.*

A summary of our contributions, along with an outline of the remainder of the chapter, is as follows:

1. In Section 2.2, we formally define our setup and assumptions and recall basics regarding Lagrangian aggregation and the SDP relaxation of a QCQP. We then define and examine a number of faces of the cone of convex Lagrange multipliers Γ and its polar cone Γ° that play key roles in our analysis.

2. In Section 2.3, we present a sufficient condition for convex hull exactness that generalizes [181, Theorem 1]. This sufficient condition (Theorem 9) is based on an analysis of the "rounding directions" inside $\mathcal{S}_{\text{SDP}}$ and is performed in the original space. Specifically, we show that convex hull exactness holds as long as certain systems of equations (that depend on Γ, Γ_1, or Γ°) contain nontrivial solutions. In contrast to [181, Theorem 1], our sufficient condition does not make any assumptions on the geometry of Γ or Γ_1 and can be used to cover additional interesting QCQPs (see Section 2.4). One of our main technical contributions (Theorem 10) shows that our sufficient condition for convex hull exactness is in fact also *necessary* under the assumption that Γ° is facially exposed (see Assumption 5 and its surrounding discussion). We end Section 2.3 by revisiting the polyhedral setting. We derive necessary and sufficient conditions for convex hull exactness (Theorem 11) and compare it to the sufficient condition presented in [181, Theorem 1]. To the best of our knowledge, this is the first necessary and sufficient condition for convex hull exactness even in the context of diagonal QCQPs (where Γ, Γ_1 and Γ° are automatically polyhedral).

3. In Section 2.4, we present example applications of our general results from Section 2.3 to a prototypical set containing big-M constraints, quadratic matrix programs, and the partition problem. In all of these applications, the resulting Γ sets are non-polyhedral, and thus the sufficient conditions from [181] that work under the polyhedrality assumption of Γ fail to cover these applications.

   In Section 2.4.1, we apply our framework to show that convex hull exactness holds for a well-studied set involving convex quadratics, binary variables and big-M relations. This set occurs as a substructure commonly studied in sparse regression applications. The convex hull characterization of this set is well-known in the literature and is often shown as a consequence of the perspective formulation trick due to Ceria and Soares [42] (see also [61, 68, 79]).

   In Section 2.4.2, we show that the SDP relaxation of a quadratic matrix program satisfies convex hull exactness whenever the number of constraints is small (when compared to the rank of the matrix variable). This strengthens separate results first presented in [181] and

[17]; see Remark 24. In contrast to the *ad hoc* proof given in [181], the proof we present in Section 2.4.2 follows the outline of our general framework.

In Section 2.4.3, we consider the QCQP formulation of the NP-hard partition problem and its SDP relaxation. Using our framework, we give an explicit description of the optimal value and epigraph of the SDP relaxation. Consequently, we recover a result due to Laurent and Poljak [106] stating that *deciding* whether objective value exactness holds for the partition QCQP is NP-hard. In contrast, we show that convex hull exactness never holds for the partition QCQP (as long as there are at least two nonzero weights). This then implies that deciding whether convex hull exactness holds for the partition QCQP is trivial.

4. In Section 2.5, we present a number of sufficient conditions for objective value exactness. In fact, our sufficient conditions further imply *optimizer exactness*, i.e., that the optimizers of the QCQP and its (projected) SDP relaxation coincide. Section 2.5.1 presents a general sufficient condition (Theorem 12) for objective value exactness based on a primal analysis. Similarly, Section 2.5.2 presents a general sufficient condition (Theorem 13) for objective value exactness based on a dual analysis. These results recover known sufficient conditions [38, 181] for objective value exactness and explain the roles played by polyhedrality in prior settings. We additionally specialize these abstract conditions to derive more concrete conditions (see Corollaries 9 to 12) for objective value exactness.

5. In Section 2.6, we present example applications of our general results from Section 2.5 to two models of random QCQPs. The results in this section show that ideas from Section 2.5 can be applied even when Γ, Γ_1, or Γ° is only known approximately. The models in this section are inspired by recent work on objective value exactness [38, 113] where random QCQPs have been used as a testing ground for understanding the strength or explanatory power of various sufficient conditions. In Section 2.6.2, we consider a fully random model of QCQPs and show that objective value exactness (in fact optimizer exactness) holds with probability $1 - o(1)$ in the regime where m (the number of constraints) is fixed and n (the number of variables) diverges to $+\infty$. In Section 2.6.3, we consider a semi-random model of QCQPs where, for each quadratic function, the quadratic terms are randomly generated and the linear and constant terms can be chosen adversarially. In this setting, we show that a perturbed notion of exactness holds again with probability $1 - o(1)$ as $n \to +\infty$.

2.1.2 Additional notation

For $x, y \in \mathbb{R}$, let $[x \pm y] \coloneqq [x - y, x + y]$, $x_+ \coloneqq \max(0, x)$ and $x_+^2 \coloneqq (x_+)^2$. For $\delta \geq 0$ and $x \in \mathbb{R}^n$, let $B_n(x, \delta) \coloneqq \{y \in \mathbb{R}^n : \|x - y\| \leq \delta\}$. When n is clear from context, we will simply write 0 and $B(x, \delta)$. For $M \in \mathbb{S}^n$, let $\lambda_{\min}(M) = \lambda_1(M) \leq \cdots \leq \lambda_n(M) = \lambda_{\max}(M)$ denote the spectrum of M. Let $K \subseteq \mathbb{E}$ be a cone. Let K° denote the polar cone of K. The notation $F \trianglelefteq K$ denotes that F is a face of K. By convention, faces of cones are always nonempty. $C_c^\infty(\mathbb{R}^n)$ denotes the smooth functions with compact support on $\mathbb{R}^n$. Let ∇ denote the gradient operator. Let $N(\mu, \Sigma)$ denote the multivariate normal distribution with mean μ and covariance Σ.

2.2 Preliminaries

2.2.1 Setup

We will consider quadratically constrained quadratic programs (QCQPs) in $\mathbb{R}^n$ defined by m-many quadratic constraints

$$\mathrm{Opt} := \inf_{x\in\mathbb{R}^n}\left\{q_{\mathsf{obj}}(x) : \begin{array}{l} q_i(x) \leq 0,\ \forall i\in[m_I] \\ q_i(x) = 0,\ \forall i\in[m_I+1,m] \end{array}\right\}. \tag{2.1}$$

Here, m_I is the number of inequality constraints and $m_E := m - m_I$ is the number of equality constraints. For each $i \in [m]$, we will write $q_i(x) = x^\top A_i x + 2b_i^\top x + c_i$ for some $A_i \in \mathbb{S}^n$, $b_i \in \mathbb{R}^n$, and $c_i \in \mathbb{R}$. Similarly, we will write $q_{\mathsf{obj}}(x) = x^\top A_{\mathsf{obj}} x + 2b_{\mathsf{obj}}^\top x + c_{\mathsf{obj}}$.

We will also consider the epigraph, $\mathcal{S}$, of this QCQP, i.e.,

$$\mathcal{S} := \left\{(x,t)\in\mathbb{R}^n\times\mathbb{R} : \begin{array}{l} q_{\mathsf{obj}}(x) \leq 2t \\ q_i(x) \leq 0,\ \forall i\in[m_I] \\ q_i(x) = 0,\ \forall i\in[m_I+1,m] \end{array}\right\}.$$

2.2.2 Aggregation and the (projected) SDP relaxation

It is well known in the QCQP literature [22, 71, 181] that the SDP relaxation of a QCQP is equivalent (under a minor assumption) to the double-Lagrangian-dual. We will state this formally in Lemma 10 but will first need to introduce notation related to Lagrangian aggregation.

Let $q : \mathbb{R}^n \to \mathbb{R}^{1+m}$ be indexed by $\{\mathsf{obj}\}\cup[m]$ where $q(x)_{\mathsf{obj}} = q_{\mathsf{obj}}(x)$ and $q(x)_i = q_i(x)$ for $i\in[m]$. Let $e_{\mathsf{obj}}, e_1, \dots, e_m$ denote the corresponding unit vectors in $\mathbb{R}^{1+m}$. We will work extensively with the aggregated quadratic functions $\left\langle(\gamma_{\mathsf{obj}},\gamma), q(x)\right\rangle$ for $(\gamma_{\mathsf{obj}},\gamma)\in\mathbb{R}^{1+m}$. For notational convenience, define $A(\gamma_{\mathsf{obj}},\gamma) := \gamma_{\mathsf{obj}}A_{\mathsf{obj}} + \sum_{i\in[m]}\gamma_i A_i$. Similarly define $b(\gamma_{\mathsf{obj}},\gamma)$, and $c(\gamma_{\mathsf{obj}},\gamma)$. We will at times work on the slice of $\mathbb{R}^{1+m}$ where the variable γ_{obj} is taken to be one. Let $A[\gamma] := A(1,\gamma)$ and similarly define $b[\gamma]$ and $c[\gamma]$. Set $[\gamma, q(x)] := \langle(1,\gamma), q(x)\rangle$. Note that

$$\begin{aligned}\left\langle(\gamma_{\mathsf{obj}},\gamma), q(x)\right\rangle &= \gamma_{\mathsf{obj}}q_{\mathsf{obj}}(x) + \sum_{i=1}^m \gamma_i q_i(x)\\ &= x^\top A(\gamma_{\mathsf{obj}},\gamma)x + 2b(\gamma_{\mathsf{obj}},\gamma)^\top x + c(\gamma_{\mathsf{obj}},\gamma), \quad \text{and}\\ [\gamma, q(x)] &= q_{\mathsf{obj}}(x) + \sum_{i=1}^m \gamma_i q_i(x)\\ &= x^\top A[\gamma]x + 2b[\gamma]^\top x + c[\gamma].\end{aligned}$$

We recall and extend the following definition from [181].

Definition 6. The *cone of convex Lagrange multipliers* for (2.1) is

$$\Gamma \coloneqq \left\{ (\gamma_{\text{obj}}, \gamma) \in \mathbb{R} \times \mathbb{R}^m : \begin{array}{l} A(\gamma_{\text{obj}}, \gamma) \succeq 0 \\ \gamma_{\text{obj}} \geq 0 \\ \gamma_i \geq 0,\ \forall i \in [m_I] \end{array} \right\}.$$

The *set of convex Lagrange multipliers* for (2.1) is

$$\Gamma_1 \coloneqq \{\gamma \in \mathbb{R}^m : (1, \gamma) \in \Gamma\} = \left\{ \gamma \in \mathbb{R}^m : \begin{array}{l} A[\gamma] \succeq 0 \\ \gamma_i \geq 0,\ \forall i \in [m_I] \end{array} \right\}. \qquad \square$$

We will see soon (see Remark 19) that Γ_1 can be thought of as the feasible domain of a partial dual of the SDP relaxation of a QCQP (see (2.2)).

Note that given $(\gamma_{\text{obj}}, \gamma) \in \Gamma$, the quadratic function $x \mapsto \left\langle (\gamma_{\text{obj}}, \gamma), q(x) \right\rangle$ is convex. Similarly, given $\gamma \in \Gamma_1$, the quadratic function $x \mapsto [\gamma, q(x)]$ is convex.

We will make the following *blanket assumption* for the remainder of the chapter. This assumption can be interpreted as a dual strict feasibility condition and is standard in the literature [17, 24, 38, 181, 194].

Assumption 4. There exists $(\gamma^*_{\text{obj}}, \gamma^*) \in \Gamma$ such that $A(\gamma^*_{\text{obj}}, \gamma^*) \succ 0$. Equivalently, there exists $\gamma^* \in \Gamma_1$ such that $A[\gamma^*] \succ 0$. □

Remark 17. Note that under Assumption 4, we have that Γ is the closed cone generated by its slice at $\gamma_{\text{obj}} = 1$, i.e., $\Gamma = \operatorname{clcone}(\{(1, \gamma) : \gamma \in \Gamma_1\})$. (See discussion following [181, Assumption 2]) □

Recall that the (projected) *SDP relaxation* of $\mathcal{S}$ is given by

$$\mathcal{S}_{\text{SDP}} \coloneqq \left\{ (x, t) \in \mathbb{R}^n \times \mathbb{R} : \begin{array}{l} \exists X \succeq xx^\top : \\ \left\langle A_{\text{obj}}, X \right\rangle + 2b_{\text{obj}}^\top x + c_{\text{obj}} \leq 2t \\ \langle A_i, X\rangle + 2b_i^\top x + c_i \leq 0,\ \forall i \in [m_I] \\ \langle A_i, X\rangle + 2b_i^\top x + c_i = 0,\ \forall i \in [m_I + 1, m] \end{array} \right\}, \qquad (2.2)$$

and $\text{Opt}_{\text{SDP}} \coloneqq \inf_{(x,t)\in\mathcal{S}_{\text{SDP}}} 2t$. By taking $X = xx^\top$ in (2.2), we see that $\text{Opt} \geq \text{Opt}_{\text{SDP}}$ and $\operatorname{conv}(\mathcal{S}) \subseteq \mathcal{S}_{\text{SDP}}$.

The following lemma states that under Assumption 4, we can rewrite $\mathcal{S}_{\text{SDP}}$ in terms of Γ. This lemma follows from a straightforward duality argument.

Lemma 10. *Suppose Assumption 4 holds. Then*

$$\begin{aligned} \mathcal{S}_{\text{SDP}} &= \left\{ (x,t) \in \mathbb{R}^{n+1} : [\gamma, q(x)] \leq 2t,\ \forall \gamma \in \Gamma_1 \right\} \\ &= \left\{ (x,t) \in \mathbb{R}^{n+1} : \left\langle (\gamma_{\text{obj}}, \gamma), q(x) \right\rangle \leq 2\gamma_{\text{obj}} t,\ \forall (\gamma_{\text{obj}}, \gamma) \in \Gamma \right\} \\ &= \left\{ (x,t) \in \mathbb{R}^{n} : q(x) - 2te_{\text{obj}} \in \Gamma^\circ \right\}. \end{aligned}$$

Here, Γ° *denotes the polar cone of* Γ.

Proof. Fix $(x,t) \in \mathbb{R}^{n+1}$. Note that

$$\begin{aligned}\sup_{\gamma\in\Gamma_1}[\gamma, q(x)] &= \sup_{\gamma\in\mathbb{R}^m}\left\{[\gamma, q(x)] : \begin{array}{l} A[\gamma] \succeq 0 \\ \gamma_i \geq 0,\ \forall i \in [m_I] \end{array}\right\} \\ &= \inf_{\xi\in\mathbb{S}^n}\left\{q_{\mathrm{obj}}(x) + \left\langle A_{\mathrm{obj}}, \xi\right\rangle : \begin{array}{l} q_i(x) + \langle A_i, \xi\rangle \leq 0,\ \forall i \in [m_I] \\ q_i(x) + \langle A_i, \xi\rangle = 0,\ \forall i \in [m_I+1, m] \\ \xi \succeq 0 \end{array}\right\},\end{aligned}$$

where the second equation follows from the strong conic duality theorem and Assumption 4. Taking $X := xx^\top + \xi$, we deduce that the first equality in Lemma 10 holds.

Note that by Assumption 4, $\Gamma = \operatorname{clcone}\{(1,\gamma) : \gamma \in \Gamma_1)\}$ so that $[\gamma, q(x)] \leq 2t$ for all $\gamma \in \Gamma_1$ if and only if $\left\langle (\gamma_{\mathrm{obj}}, \gamma), q(x)\right\rangle \leq 2\gamma_{\mathrm{obj}} t$ for all $(\gamma_{\mathrm{obj}}, \gamma) \in \Gamma$; this gives the second equality. The third equality holds by definition of the polar cone. ■

Corollary 5. *Suppose Assumption 4 holds. Then*

$$\mathrm{Opt}_{\mathrm{SDP}} = \inf_{x\in\mathbb{R}^n}\sup_{\gamma\in\Gamma_1}[\gamma, q(x)]. \tag{2.3}$$

Corollary 6. *Suppose Assumption 4 holds. Then, $\mathcal{S}_{\mathrm{SDP}}$ is closed.*

Remark 18. In comparison with (2.2), the expressions for $\mathcal{S}_{\mathrm{SDP}}$ given in Lemma 10 make the roles played by Γ, Γ_1, and Γ° explicit. In particular, these expressions for $\mathcal{S}_{\mathrm{SDP}}$ lend themselves to a clean analysis whenever the corresponding dual set Γ, Γ_1, or Γ° is sufficiently simple. □

Remark 19. Phrased differently, one may minimize $\mathrm{Opt}_{\mathrm{SDP}}$ in the form (2.2) by minimizing over $x \in \mathbb{R}^n$ the value of an inner minimization problem over the matrix variables $X \succeq xx^\top \in \mathbb{S}^n$. Writing $X = xx^\top + \xi$ and taking the SDP dual in the ξ variable then results in the same saddle-point structure $\mathrm{Opt}_{\mathrm{SDP}} = \inf_{x\in\mathbb{R}^n}\sup_{\gamma\in\Gamma_1}[\gamma, q(x)]$ observed in Corollary 5. In other words, Γ_1 *is the feasible domain to this partial dual of* (2.2). □

Let us consider a concrete example to help materialize these definitions.

Example 6. Consider the following QCQP epigraph,

$$\mathcal{S} := \left\{(x,t) \in \mathbb{R}^2 \times \mathbb{R} : \begin{array}{l} q_{\mathrm{obj}}(x) \leq 2t \\ q_1(x) \leq 0 \\ q_2(x) \leq 0 \end{array}\right\},$$

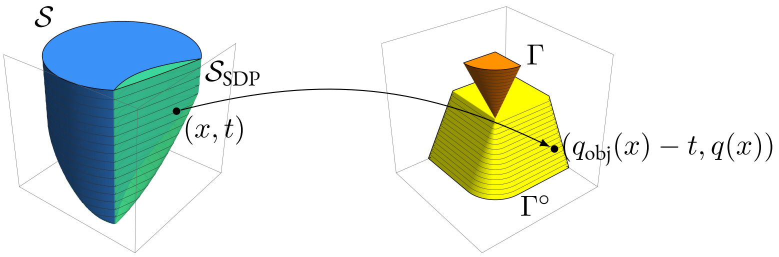
$\mathcal{S}$
$\mathcal{S}_{\mathrm{SDP}}$
(x,t)
Γ
$(q_{\mathrm{obj}}(x) - t, q(x))$
Γ°

Figure 2.1: The sets $\mathcal{S}$, $\mathcal{S}_{\mathrm{SDP}}$, Γ, and Γ° from Example 6 are shown in blue, green, orange, and yellow respectively. By Lemma 10, $(x,t) \in \mathcal{S}_{\mathrm{SDP}}$ if and only if $q(x) - 2te_{\mathrm{obj}} \in \Gamma^\circ$.

where $q_{\mathrm{obj}}(x) \coloneqq 2x_1x_2 - x_2 - 1/4$, $q_1(x) \coloneqq x_1^2 - x_2^2 - x_1 + x_2 - 1$, and $q_2(x) \coloneqq x_1^2 + x_2^2 - 1$. Through a straightforward calculation, we obtain

$$\Gamma = \left\{ (\gamma_{\mathrm{obj}}, \gamma) \in \mathbb{R}^3 : \begin{array}{l} \gamma_2 \geq \sqrt{\gamma_{\mathrm{obj}}^2 + \gamma_1^2} \\ \gamma_{\mathrm{obj}}, \gamma_1, \gamma_2 \geq 0 \end{array} \right\},$$

$$\Gamma^\circ = \left\{ (\gamma_{\mathrm{obj}}, \gamma) \in \mathbb{R}^3 : -\gamma_2 \geq \sqrt{(\gamma_{\mathrm{obj}})_+^2 + (\gamma_1)_+^2} \right\}, \text{ and}$$

$$\mathcal{S}_{\mathrm{SDP}} = \left\{ (x,t) \in \mathbb{R}^2 : -q_2(x) \geq \sqrt{(q_{\mathrm{obj}}(x) - 2t)_+^2 + q_1(x)_+^2} \right\}.$$

See Figure 2.1 for the plots of the sets corresponding to $\mathcal{S}$, Γ, Γ°, and $\mathcal{S}_{\mathrm{SDP}}$. □

2.2.3 Faces of Γ and Γ°

In this section we define key faces of Γ and Γ° that will play important roles in our analysis. We will additionally recall a number of elementary properties of convex cones and their faces specialized to our setting. See [12, 13, 142] for a more in-depth treatment of general convex cones and their faces.

Recall the following definitions.

Definition 7. Given a face $\mathcal{G} \trianglelefteq \Gamma^\circ$ and $(g_{\mathrm{obj}}, g) \in \operatorname{rint}(\mathcal{G})$, the *conjugate face of* $\mathcal{G}$ is

$$\mathcal{G}^\triangle \coloneqq \Gamma \cap \mathcal{G}^\perp = \Gamma \cap (g_{\mathrm{obj}}, g)^\perp.$$

Similarly, define the *conjugate face of* $\mathcal{F}$ for a face $\mathcal{F} \trianglelefteq \Gamma$. □

Definition 8. For a face $\mathcal{G} \trianglelefteq \Gamma^\circ$, we say that $\mathcal{G}$ is *exposed* if there exists $(\gamma_{\mathrm{obj}}, \gamma) \in \Gamma$ such that $\mathcal{G} = \Gamma^\circ \cap (\gamma_{\mathrm{obj}}, \gamma)^\perp$. □

We will additionally associate faces of Γ and Γ° to points $(x,t) \in \mathcal{S}_{\mathrm{SDP}}$ as follows.

Definition 9. Given $(x,t) \in \mathcal{S}_{\mathrm{SDP}}$, let $\mathcal{G}(x,t) \trianglelefteq \Gamma^\circ$ denote the minimal face of Γ° containing $q(x) - 2te_{\mathrm{obj}}$ and define $\mathcal{F}(x,t) \coloneqq \mathcal{G}(x,t)^\triangle$. □

The next fact follows from Definition 9.

Fact 1. *Given* $(x,t) \in \mathcal{S}_{\mathrm{SDP}}$*, we have that* $q(x) - 2te_{\mathrm{obj}} \in \operatorname{rint}(\mathcal{G}(x,t))$ *and* $\mathcal{F}(x,t) = \Gamma \cap (q(x) - 2te_{\mathrm{obj}})^\perp$.

2.3 Convex hull exactness

In this section, we present necessary and sufficient conditions for convex hull exactness, i.e., the property that $\operatorname{conv}(\mathcal{S}) = \mathcal{S}_{\text{SDP}}$. These results form the basis of our assertion that *exactness can be treated systematically whenever* Γ, Γ_1, *or* Γ° *is well-understood.*

We begin by rephrasing convex hull exactness as a question regarding the existence of certain "rounding directions." The following result follows from basic convex analysis.

Lemma 11. *Suppose Assumption 4 holds. Then,* $\operatorname{conv}(\mathcal{S}) = \mathcal{S}_{\text{SDP}}$ *if and only if for every* $(x, t) \in \mathcal{S}_{\text{SDP}} \setminus \mathcal{S}$*, there exists a nonzero* $(x', t') \in \mathbb{R}^{n+1}$ *and* $\alpha > 0$ *such that*

$$[(x, t) \pm \alpha(x', t')] \subseteq \mathcal{S}_{\text{SDP}}.$$

Proof. Note that $\mathcal{S}_{\text{SDP}}$ is a closed convex set containing no lines. Also, one can easily check that $(0_n, 1)$ is indeed a recessive direction of $\mathcal{S}_{\text{SDP}}$. Furthermore, $(0_n, 1)$ is the only recessive direction of $\mathcal{S}_{\text{SDP}}$. To see this, let γ^* be such that $A[\gamma^*] \succ 0$ (which exists by Assumption 4) and consider any (x', t') where x' is nonzero. Then, for any $(\tilde{x}, \tilde{t}) \in \mathcal{S}_{\text{SDP}}$ and all $\alpha > 0$ large enough, $2(\tilde{t}+\alpha t') < [\gamma^*, q(\tilde{x}+\alpha x')]$. Therefore, we deduce by [152, Theorem 18.5], that $\mathcal{S}_{\text{SDP}}$ is the sum of the convex hull of its extreme points and the direction $(0_n, 1)$. In particular $\operatorname{conv}(\mathcal{S}) = \mathcal{S}_{\text{SDP}}$ if and only if (x, t) is not extreme for every $(x, t) \in \mathcal{S}_{\text{SDP}} \setminus \mathcal{S}$. By definition, (x, t) is not extreme if and only if there exists (x', t') and $\alpha > 0$ such that $[(x, t) \pm \alpha(x', t')] \subseteq \mathcal{S}_{\text{SDP}}$. ■

We capture the relevant set in Lemma 11 in the following definition.

Definition 10. The subspace of *rounding directions* at $(x, t) \in \mathcal{S}_{\text{SDP}}$ is

$$\mathcal{R}(x, t) := \left\{(x', t') \in \mathbb{R}^{n+1} : \exists \alpha > 0 \text{ s.t. } [(x, t) \pm \alpha(x', t')] \subseteq \mathcal{S}_{\text{SDP}}\right\}.$$

This set is *nontrivial* if it contains a nonzero element. □

Note that $\mathcal{R}(x, t)$ is in fact a subspace so that its name is justified. Indeed, $\mathcal{R}(x, t)$ is a convex cone as $\mathcal{S}_{\text{SDP}}$ is convex. Furthermore, it holds that $-\mathcal{R}(x, t) = \mathcal{R}(x, t)$.

Remark 20. One may compare our *rounding directions* to other similar definitions from elementary convex analysis [85, Section 5.1]. Fix a point $(x, t) \in \mathcal{S}_{\text{SDP}}$ and $(x', t') \in \mathbb{R}^{n+1}$. Recall that (x', t') is a *feasible direction* if there exists $\alpha > 0$ such that $[(x, t), (x, t) + \alpha(x', t')] \subseteq \mathcal{S}_{\text{SDP}}$. In particular, feasible directions are a *unidirectional* notion, whereas rounding directions are *bidirectional*. Next, recall that (x', t') is a *tangent direction* if it is a limit of feasible directions. Again, tangent directions are unidirectional. □

Remark 21. Suppose $(x, t) \in \mathcal{S}_{\text{SDP}}$ and $2t > \sup_{\gamma \in \Gamma_1} [\gamma, q(x)]$. Then, by Lemma 10 there exists $\alpha > 0$ such that $[(x, t) \pm \alpha(0_n, 1)] \subseteq \mathcal{S}_{\text{SDP}}$. In particular, it suffices to verify the condition of Lemma 11 for points $(x, t) \in \mathcal{S}_{\text{SDP}} \setminus \mathcal{S}$ for which $2t = \sup_{\gamma \in \Gamma_1} [\gamma, q(x)]$. □

2.3.1 Sufficient conditions for convex hull exactness

In this section we identify a particular subset of the rounding directions at $(x, t) \in \mathcal{S}_{\text{SDP}}$. This then leads to a sufficient condition for convex hull exactness, i.e., the condition that $\operatorname{conv}(\mathcal{S}) = \mathcal{S}_{\text{SDP}}$.

Definition 11. Given $(x,t) \in \mathcal{S}_{\mathrm{SDP}}$, define

$$\mathcal{R}'(x,t) := \left\{(x',t') \in \mathbb{R}^{n+1} :\ q(x+\alpha x') - 2(t+\alpha t')e_{\mathrm{obj}} \in \operatorname{span}(\mathcal{G}(x,t)),\ \forall \alpha \in \mathbb{R}\right\}. \square$$

Lemma 12. *Suppose Assumption 4 holds and $(x,t) \in \mathcal{S}_{\mathrm{SDP}}$. Then, $\mathcal{R}'(x,t) \subseteq \mathcal{R}(x,t)$.*

Proof. Let $(x,t) \in \mathcal{S}_{\mathrm{SDP}}$ and $(x',t') \in \mathcal{R}'(x,t)$. Then, by continuity and the fact that $q(x) - 2te_{\mathrm{obj}} \in \operatorname{rint}(\mathcal{G}(x,t))$, there exists $\alpha > 0$ such that

$$q(x+\epsilon x') - 2(t+\epsilon t')e_{\mathrm{obj}} \in \mathcal{G}(x,t) \subseteq \Gamma^\circ$$

for all $\epsilon \in [\pm\alpha]$. By the third characterization of $\mathcal{S}_{\mathrm{SDP}}$ in Lemma 10, we have that $[(x,t) \pm \alpha(x',t')] \subseteq \mathcal{S}_{\mathrm{SDP}}$. ■

Lemmas 11 and 12 immediately imply the following sufficient condition for convex hull exactness.

Theorem 9. *Suppose Assumption 4 holds and that for all $(x,t) \in \mathcal{S}_{\mathrm{SDP}} \setminus \mathcal{S}$, the set $\mathcal{R}'(x,t)$ is nontrivial. Then, $\operatorname{conv}(\mathcal{S}) = \mathcal{S}_{\mathrm{SDP}}$.*

We will see applications of Theorem 9 in Section 2.4.

In Lemma 13 below, we will record an alternate description of $\mathcal{R}'(x,t)$. We will require the following observation.

Observation 2. *Suppose Assumption 4 holds. Let $(x,t) \in \mathcal{S}_{\mathrm{SDP}}$ where $2t = \sup_{\gamma\in\Gamma_1}[\gamma, q(x)]$. Then, $\operatorname{span}(\mathcal{G}(x,t)) \not\supseteq \mathbb{R} \times 0_m$. In particular, $\mathcal{G}(x,t)^\perp = \operatorname{span}\Big(\mathcal{G}(x,t)^\perp \cap \big\{\gamma_{\mathrm{obj}} = 1\big\}\Big)$.*

Proof. Suppose $\operatorname{span}(\mathcal{G}(x,t)) \supseteq \mathbb{R} \times 0_m$ so that $(0_n, 1) \in \mathcal{R}'(x,t)$. By Lemma 12, there exists $\alpha > 0$ such that $(x, t-\alpha) \in \mathcal{S}_{\mathrm{SDP}}$. This contradicts $2t = \sup_{\gamma\in\Gamma_1}[\gamma, q(x)]$.

We deduce that $\operatorname{span}(\mathcal{G}(x,t)) \not\supseteq \mathbb{R} \times 0_m$. Equivalently, $\mathcal{G}(x,t)^\perp \not\subseteq 0 \times \mathbb{R}^m$ and there exists $(1,\bar\gamma) \in \mathcal{G}(x,t)^\perp$. Then, for any $(\gamma_{\mathrm{obj}}, \gamma) \in \mathcal{G}(x,t)^\perp$, we can write $(\gamma_{\mathrm{obj}}, \gamma)$ as a linear combination of

$$(\gamma_{\mathrm{obj}}, \gamma) + (1-\gamma_{\mathrm{obj}})(1,\gamma) \qquad \text{and} \qquad (1,\gamma). \qquad ■$$

Lemma 13. *Suppose Assumption 4 holds and let $(x,t) \in \mathcal{S}_{\mathrm{SDP}}$. Then,*

$$\mathcal{R}'(x,t) = \left\{(x',t') \in \mathbb{R}^{n+1} :\ \begin{array}{l} (x')^\top A(\gamma_{\mathrm{obj}},\gamma)x' = 0,\ \forall (\gamma_{\mathrm{obj}},\gamma) \in \mathcal{G}(x,t)^\perp \\ \Big\langle A(\gamma_{\mathrm{obj}},\gamma)x + b(\gamma_{\mathrm{obj}},\gamma), x'\Big\rangle - \gamma_{\mathrm{obj}}t' = 0,\ \forall (\gamma_{\mathrm{obj}},\gamma) \in \mathcal{G}(x,t)^\perp \end{array}\right\}$$

If furthermore $2t = \sup_{\gamma\in\Gamma_1}[\gamma, q(x)]$, then

$$\mathcal{R}'(x,t) = \left\{(x',t') \in \mathbb{R}^{n+1} :\ \begin{array}{l} (x')^\top A[\gamma]x' = 0,\ \forall (1,\gamma) \in \mathcal{G}(x,t)^\perp \\ \langle A[\gamma]x + b[\gamma], x'\rangle - t' = 0,\ \forall (1,\gamma) \in \mathcal{G}(x,t)^\perp \end{array}\right\}.$$

Proof. Note that $(x', t') \in \mathcal{R}'(x,t)$ if and only if for all $(\gamma_{\text{obj}}, \gamma) \in \mathcal{G}(x,t)^\perp$, we have that

$$\begin{aligned}&\left\langle (\gamma_{\text{obj}}, \gamma), q(x+\alpha x') - 2(t+\alpha t')e_{\text{obj}}\right\rangle\\ &\quad = \alpha^2 (x')^\top A(\gamma_{\text{obj}}, \gamma)(x') + 2\alpha\left(\left\langle A(\gamma_{\text{obj}}, \gamma)x + b(\gamma_{\text{obj}}, \gamma), x'\right\rangle - t'\right)\\ &\qquad + \left\langle (\gamma_{\text{obj}}, \gamma), q(x) - 2te_{\text{obj}}\right\rangle\end{aligned}$$

is identically zero in α. This occurs if and only if for all $(\gamma_{\text{obj}}, \gamma) \in \mathcal{G}(x,t)^\perp$, we have

$$\begin{gathered}(x')^\top A(\gamma_{\text{obj}}, \gamma)x' = 0, \quad \text{and}\\ \left\langle A(\gamma_{\text{obj}}, \gamma)x + b(\gamma_{\text{obj}}, \gamma), x'\right\rangle - t' = 0.\end{gathered}$$

This proves the first assertion. The second assertion follows from the first and Observation 2. ■

2.3.2 Necessary conditions for convex hull exactness

In Section 2.3.1, we gave a sufficient condition for convex hull exactness by identifying a subset of directions $\mathcal{R}'(x,t) \subseteq \mathcal{R}(x,t)$ and invoking Lemma 11. In this section, we show that under a technical assumption (Assumption 5), we have $\mathcal{R}'(x,t) = \mathcal{R}(x,t)$. This then leads to a necessary and sufficient condition for convex hull exactness under the technical assumption.

Assumption 5. Suppose Γ° is facially exposed, i.e., every face of Γ° is exposed. □

This assumption holds for any cone isomorphic to a slice of the nonnegative orthant, the second-order cone, or the positive semidefinite cone. See [143] for a longer discussion of this assumption and its connections to the *nice cones*. In general, all *nice cones* are facially exposed. Our analysis will be based on the following property of exposed faces $\mathcal{G} \trianglelefteq \Gamma^\circ$ (see [13, Definition 2.A.9] and its surrounding discussion):

Fact 2. *A face $\mathcal{G} \trianglelefteq \Gamma^\circ$ is exposed if and only if $\mathcal{G} = (\mathcal{G}^\triangle)^\triangle$.*

We are now ready to prove a partial converse to Lemma 12.

Lemma 14. *Suppose Assumptions 4 and 5 hold and let $(x,t) \in \mathcal{S}_{\text{SDP}}$. Then $\mathcal{R}'(x,t) = \mathcal{R}(x,t)$.*

Proof. Fix $(x', t') \in \mathcal{R}(x,t)$. As $\mathcal{R}(x,t)$ is a convex cone, we may without loss of generality assume that $[(x,t) \pm (x',t')] \subseteq \mathcal{S}_{\text{SDP}}$. Our goal is to show that $(x',t') \in \mathcal{R}'(x,t)$, i.e., that

$$q(x+\alpha x') - 2(t+\alpha t')e_{\text{obj}} \in \operatorname{span}(\mathcal{G}(x,t)),\ \forall \alpha \in \mathbb{R}.$$

As each coordinate of this vector is quadratic in α, it suffices to show instead that

$$q(x+\alpha x') - 2(t+\alpha t')e_{\text{obj}} \in \mathcal{G}(x,t),\ \forall \alpha \in [-1,1].$$

Let $(f_{\text{obj}}, f) \in \operatorname{rint}(\mathcal{F}(x,t))$ so that by Assumption 5 and Fact 2, we may write $\mathcal{G}(x,t) = \Gamma^\circ \cap (f_{\text{obj}}, f)^\perp$. As $[(x,t) \pm (x',t')] \subseteq \mathcal{S}_{\text{SDP}}$, we immediately have that $q(x+\alpha x') - 2(t+\alpha t')e_{\text{obj}} \in \Gamma^\circ$ for all $\alpha \in [-1,1]$. It remains to verify that the map

$$\alpha \mapsto \left\langle (f_{\text{obj}}, f), q(x+\alpha x') - 2(t+\alpha t')e_{\text{obj}} \right\rangle$$

evaluates to zero on $\alpha \in [-1,1]$. Again, as $[(x,t) \pm (x',t')] \subseteq \mathcal{S}_{\text{SDP}}$, this map is nonpositive for all $\alpha \in [-1,1]$. Next, note that $(f_{\text{obj}}, f) \in \mathcal{F}(x,t) = \Gamma \cap (q(x) - 2te_{\text{obj}})^\perp$ so that this map evaluates to zero at $\alpha = 0$. Finally, $(f_{\text{obj}}, f) \in \Gamma$ implies that this map is also convex. We conclude that this map is identically zero. ■

The following necessary and sufficient condition for convex hull exactness then follows from Lemma 14.

Theorem 10. *Suppose Assumptions 4 and 5 hold. Then,* $\operatorname{conv}(\mathcal{S}) = \mathcal{S}_{\text{SDP}}$ *if and only if for all* $(x,t) \in \mathcal{S}_{\text{SDP}} \setminus \mathcal{S}$*, the set* $\mathcal{R}'(x,t)$ *is nontrivial.*

To close this subsection, we give a compact description of $\mathcal{R}'(x,t)$ under Assumption 5.

Proposition 6. *Suppose Assumptions 4 and 5 hold. Let* $(x,t) \in \mathcal{S}_{\text{SDP}}$ *where* $2t = \sup_{\gamma\in\Gamma_1}[\gamma, q(x)]$ *and let* $(1,f) \in \operatorname{rint}(\mathcal{F}(x,t))$*. Then,*

$$\mathcal{R}'(x,t) = \left\{ (x',t') \in \mathbb{R}^{n+1} : \begin{array}{l} x' \in \ker(A[f]) \\ \langle A[\eta]x + b[\eta], x'\rangle - t' = 0,\ \forall (1,\eta) \in \mathcal{G}(x,t)^\perp \end{array} \right\}.$$

Proof. Let $(1,f) \in \operatorname{rint}(\mathcal{F}(x,t))$. By Lemma 13, it suffices to show that $x' \in \ker(A[f])$ if and only if

$$(x')^\top A[\gamma]x' = 0,\ \forall (1,\gamma) \in \mathcal{G}(x,t)^\perp.$$

The reverse direction holds immediately as $(1,f) \in \mathcal{F}(x,t) \subseteq \mathcal{G}(x,t)^\perp$ and $A[f] \succeq 0$.

To see the forward direction: Let $x' \in \ker(A[f])$ and set $v_{\text{obj}} = (x')^\top A_{\text{obj}} x'$. Similarly, set $v_i = (x')^\top A_i x'$. Then,

$$\left\langle (v_{\text{obj}}, v), (\gamma_{\text{obj}}, \gamma) \right\rangle = (x')^\top A(\gamma_{\text{obj}}, \gamma)x' \geq 0,\ \forall (\gamma_{\text{obj}}, \gamma) \in \Gamma.$$

Thus $(-v_{\text{obj}}, -v) \in \Gamma^\circ$. On the other hand, $\left\langle (v_{\text{obj}}, v), (1, f) \right\rangle = (x')^\top A[f]x' = 0$. We deduce that $(-v_{\text{obj}}, -v) \in \Gamma^\circ \cap (1,f)^\perp = \mathcal{F}(x,t)^\triangle = \mathcal{G}(x,t)$. In particular, $(x')^\top A(\gamma_{\text{obj}}, \gamma)x' = \left\langle (v_{\text{obj}}, v), (\gamma_{\text{obj}}, \gamma) \right\rangle = 0$ for all $(\gamma_{\text{obj}}, \gamma) \in \mathcal{G}(x,t)^\perp$. ■

2.3.3 Revisiting the setting of polyhedral Γ

Wang and Kılınç-Karzan [181] give sufficient conditions for convex hull exactness under the assumption that Γ_1 is polyhedral.[2] This assumption holds, for example, when the set of quadratic

[2] Equivalently (see Remark 17), under the assumption that Γ is polyhedral.

forms $\left\{A_{\text{obj}}, A_1, \dots, A_m\right\}$ is simultaneously diagonalizable. Specializing Theorem 10 to this setting, we prove the following *necessary and sufficient* counterpart to [181, Theorem 1].

Theorem 11. *Suppose Assumption 4 holds and that Γ is polyhedral. Then,* $\operatorname{conv}(\mathcal{S}) = \mathcal{S}_{\text{SDP}}$ *if and only if*

$$\left\{(x', t') \in \mathbb{R}^{n+1} : \begin{array}{l} x' \in \ker(A[f]) \\ \langle b[\gamma], x'\rangle - t' = 0,\ \forall (1,\gamma) \in \mathcal{F} \end{array}\right\}$$

is nontrivial for every $\mathcal{F} \trianglelefteq \Gamma$ which is exposed by some vector $q(x) - 2te_{\text{obj}}$ for $(x,t) \in \mathcal{S}_{\text{SDP}} \setminus \mathcal{S}$. Here, f is any vector such that $(1, f) \in \operatorname{rint}(\mathcal{F})$.

Proof. We begin by noting that when Γ is polyhedral, so too is Γ° so that Assumption 5 holds. Next, we claim that for every face $\mathcal{G} \trianglelefteq \Gamma^\circ$ we have $\mathcal{G}^\perp = \operatorname{span}(\mathcal{G}^\triangle)$. By definition, $\operatorname{span}(\mathcal{G}^\triangle) = \operatorname{span}(\Gamma \cap \mathcal{G}^\perp) \subseteq \mathcal{G}^\perp$. On the other hand, as Γ and Γ° are polyhedral, we have that [169, Theorem 3]

$$\dim(\mathcal{G}) + \dim(\mathcal{G}^\triangle) = m.$$

Rearranging this equation, we have $\dim(\mathcal{G}^\triangle) = m - \dim(\mathcal{G}) = \dim(\mathcal{G}^\perp)$. We conclude that $\mathcal{G}^\perp = \operatorname{span}(\mathcal{G}^\triangle)$.

Let $(x,t) \in \mathcal{S}_{\text{SDP}}$ such that $2t = \sup_{\gamma\in\Gamma_1}[\gamma, q(x)]$ and let $(1, f) \in \operatorname{rint}(\mathcal{F}(x,t))$. Then, Observation 2 and Proposition 6 imply that

$$\begin{aligned}
\mathcal{R}'(x,t) &= \left\{(x', t') \in \mathbb{R}^{n+1} : \begin{array}{l} x' \in \ker(A[f]) \\ \langle A[\gamma]x + b[\gamma], x'\rangle - t' = 0,\ \forall (1,\gamma) \in \mathcal{G}(x,t)^\perp \end{array}\right\} \\
&= \left\{(x', t') \in \mathbb{R}^{n+1} : \begin{array}{l} x' \in \ker(A[f]) \\ \langle A[\gamma]x + b[\gamma], x'\rangle - t' = 0,\ \forall (1,\gamma) \in \mathcal{F}(x,t) \end{array}\right\} \\
&= \left\{(x', t') \in \mathbb{R}^{n+1} : \begin{array}{l} x' \in \ker(A[f]) \\ \langle b[\gamma], x'\rangle - t' = 0,\ \forall (1,\gamma) \in \mathcal{F}(x,t) \end{array}\right\}.
\end{aligned}$$

Here, the second line follows because we have shown $\mathcal{G}^\perp = \operatorname{span}(\mathcal{G}^\triangle)$ holds for every face $\mathcal{G} \trianglelefteq \Gamma^\circ$ and by definition $\mathcal{F}(x,t) = \mathcal{G}(x,t)^\triangle$. The third line follows from the fact that $(1, f) \in \operatorname{rint}(\mathcal{F}(x,t))$ implies $\ker(A[f]) \subseteq \ker(A[\gamma])$ for every $(1,\gamma) \in \mathcal{F}(x,t)$. The result then follows from Theorem 10. ■

Remark 22. The main difference between Theorem 11 and [181, Theorem 1] is that Theorem 11 only considers certain (*a fortiori* semidefinite) faces of Γ whereas [181, Theorem 1] imposes a constraint on *every* semidefinite face of Γ. This idea of restricting the analysis to certain faces of Γ was previously investigated by [38, 113] who used it to provide sufficient conditions for objective value exactness. □

2.4 Applications: Convex hull exactness

In this section, we apply the results of Section 2.3 to a number of problems. These examples provide further evidence towards the message that *exactness can be treated systematically whenever* Γ, Γ_1, *or* Γ° *is well-understood.*

2.4.1 Mixed binary programming

To begin, we apply our results to a well-studied prototypical set involving a convex quadratic function, a binary variable and a big-M relation. The example in this subsection highlights the systematic nature of our approach.

Consider the epigraph set

$$\mathcal{S} = \left\{ (x,t)\in\mathbb{R}^2\times\mathbb{R} :\ \begin{array}{l} q_{\text{obj}}(x) := x_2^2 \leq 2t \\ q_1(x) := x_1(x_1-1) = 0 \\ q_2(x) := \sqrt{2}x_2(x_1-1) = 0 \end{array} \right\}.$$

In words, x_1 is a binary on-off variable, x_2 is a continuous variable which is constrained to be off whenever x_1 is off, and t is the epigraph variable corresponding to x_2^2. The normalization of $q_2(x)$ is not important here and is made only for notational convenience in the calculations.

It is well-known that $\operatorname{conv}(\mathcal{S})$ is given by the perspective reformulation of $\mathcal{S}$ (see e.g., [68, 79]), i.e.,

$$\operatorname{conv}(\mathcal{S}) = \left\{ (x,t)\in\mathbb{R}^2\times\mathbb{R} :\ x_2^2 - 2tx_1 \leq 0,\ 0\leq x_1\leq 1\right\}. \tag{2.4}$$

We give an alternative proof of (2.4). We will show that $\operatorname{conv}(\mathcal{S}) = \mathcal{S}_{\text{SDP}}$, the projected SDP relaxation, using Theorem 10. Then, using an explicit description of Γ°, we will give a description of $\operatorname{conv}(\mathcal{S}) = \mathcal{S}_{\text{SDP}}$ in the original space.

A simple computation shows that in this setting, we have

$$\Gamma = \left\{ (\gamma_{\text{obj}},\gamma)\in\mathbb{R}^3 :\ \gamma_{\text{obj}} + \gamma_1 \geq \sqrt{(\gamma_{\text{obj}}-\gamma_1)^2 + \left(\sqrt{2}\gamma_2\right)^2}\right\} \text{ and}$$
$$\Gamma^\circ = \left\{ (\ell_{\text{obj}},\ell)\in\mathbb{R}^3 :\ -\ell_{\text{obj}} - \ell_1 \geq \sqrt{(\ell_{\text{obj}}-\ell_1)^2 + \left(\sqrt{2}\ell_2\right)^2}\right\}.$$

In words, Γ and Γ° are both (rotated) second-order cones and Assumptions 4 and 5 hold.

It remains to show that for all $(x,t)\in\mathcal{S}_{\text{SDP}}\setminus\mathcal{S}$, the set $\mathcal{R}'(x,t)$ is nontrivial. To this end, let $(x,t)\in\mathcal{S}_{\text{SDP}}\setminus\mathcal{S}$. Recall that Γ° has three types of faces: the two trivial faces (the apex and the cone itself) and the one-dimensional proper faces. Thus, there are three cases to consider: (i) $\mathcal{G}(x,t) = \{0\}$, (ii) $\mathcal{G}(x,t) = \Gamma^\circ$, and (iii) $\mathcal{G}(x,t)$ is a one-dimensional face of Γ°.

In case (i), $q(x)-2te_{\text{obj}} = 0$ implying that $(x,t) \in \mathcal{S}$, a contradiction. In case (ii), $\operatorname{span}(\mathcal{G}(x,t)) = \mathbb{R}^3$ so that $\mathcal{R}'(x,t) = \mathbb{R}^3$ and is nontrivial. In the final case, a mechanical but slightly tedious application of Proposition 6 (see Section B.1) gives

$$\mathcal{R}'(x,t) = \left\{ \begin{pmatrix} 2t \\ -x_2 \\ 0 \end{pmatrix}, \begin{pmatrix} -x_2 \\ x_1 \\ 0 \end{pmatrix}, \begin{pmatrix} t \\ -x_2 \\ x_1 \end{pmatrix}, \begin{pmatrix} x_2(x_1 - 1 + 2t) \\ -x_1^2 + x_1 - 2tx_1 - 2t \\ 2x_2 \end{pmatrix} \right\}^{\perp}. \tag{2.5}$$

Finally, one may verify that $(x,t) \in \mathcal{R}'(x,t)$ is nonzero.

Remark 23. Here, the motivation for the final step of checking that $(x,t) \in \mathcal{R}'(x,t)$ is as follows: One can show that in case (iii), the first three vectors in (2.5) span the 2-dimensional subspace orthogonal to (x,t). In particular, $\mathcal{R}'(x,t)$ is nontrivial if and only if $(x,t) \in \mathcal{R}'(x,t)$.

□

We conclude that

$$\begin{aligned} &\operatorname{conv}(\mathcal{S}) = \mathcal{S}_{\text{SDP}} = \\ &\quad \left\{ (x,t) \in \mathbb{R}^3 : \, -(q_{\text{obj}}(x) - 2t) - q_1(x) \geq \sqrt{(q_{\text{obj}}(x) - 2t - q_1(x))^2 + 2q_2(x)^2} \right\}. \end{aligned}$$

This example highlights the systematic nature of the approach outlined in Theorem 10 for proving convex hull exactness. In contrast to *ad hoc* proofs of convex hull exactness which may rely on *guessing and verifying* a nonzero rounding direction, the system of equations defining $\mathcal{R}'(x,t)$ gives a principled way of *deducing* a direction. While guessing such a rounding direction may be possible in low-dimensional settings (for example, the setting of the current subsection), this becomes more difficult in higher-dimensional settings where $\mathcal{S}$ and $\mathcal{S}_{\text{SDP}}$ are difficult to visualize. We illustrate this in the following subsection.

2.4.2 Quadratic matrix programs

Quadratic matrix programs (QMPs) [17, 181] are a generalization of QCQPs where the decision variable $x \in \mathbb{R}^n$ is replaced by a decision matrix $X \in \mathbb{R}^{r\times k}$. These problems find a variety of application and have been used to model robust least squares problems, the orthogonal Procrustes problem [17], and certain sphere packing problems [20]. Formally, a QMP is an optimization problem in the variable $X \in \mathbb{R}^{r\times k}$, where the constraints and objective function are each of the form

$$\operatorname{tr}(X^\top \mathbb{A} X) + 2\operatorname{tr}(B^\top X) + c$$

for some $\mathbb{A} \in \mathbb{S}^r$, $B \in \mathbb{R}^{r\times k}$, and $c \in \mathbb{R}$.

Alternatively, letting $x \in \mathbb{R}^n$ (resp. $b \in \mathbb{R}^n$) denote the vector formed by stacking the columns of X (resp. B) on top of each other, we can rewrite the above expression as

$$x^\top (I_k \otimes \mathbb{A})x + 2\langle b, x\rangle + c.$$

We will choose to view QMPs as the special class of QCQPs where the quadratic forms $A_{\text{obj}}, A_1, \dots, A_m$ are each of the form $I_k \otimes \mathbb{A}$ for some $\mathbb{A} \in \mathbb{S}^r$.

The following lemma establishes that if the number of constraints is small compared to k (originally the width of the matrix variable), then convex hull exactness holds.

Proposition 7. *Suppose Assumption 4 holds and that $A_{\text{obj}} = I_k \otimes \mathbb{A}_{\text{obj}}$, $A_1 = I_k \otimes \mathbb{A}_1$, …, $A_m = I_k \otimes \mathbb{A}_m$ for some $\mathbb{A}_{\text{obj}}, \mathbb{A}_1, \dots, \mathbb{A}_m \in \mathbb{S}^r$. Furthermore, suppose $k \geq m$. Then, $\mathcal{R}(x,t)$ is nontrivial for every $(x,t) \in \mathcal{S}_{\text{SDP}} \setminus \mathcal{S}$. In particular, convex hull exactness holds, i.e.,* $\operatorname{conv}(\mathcal{S}) = \mathcal{S}_{\text{SDP}}$.

Proof. Fix $(x,t) \in \mathcal{S}_{\text{SDP}} \setminus \mathcal{S}$. Based on Theorem 9 and Lemma 13, our goal is to prove that

$$\mathcal{R}'(x,t) = \left\{ (x',t') \in \mathbb{R}^{n+1} : \begin{array}{l} x'^\top A(\gamma_{\text{obj}},\gamma)x' = 0,\ \forall (\gamma_{\text{obj}},\gamma) \in \mathcal{G}(x,t)^\perp \\ \left\langle A(\gamma_{\text{obj}},\gamma)x + b(\gamma_{\text{obj}},\gamma), x' \right\rangle - \gamma_{\text{obj}} t' = 0,\ \forall (\gamma_{\text{obj}},\gamma) \in \mathcal{G}(x,t)^\perp \end{array} \right\} \tag{2.6}$$

is nontrivial. We claim that it suffices to show how to construct a nonzero $y \in \mathbb{R}^r$ such that

$$y^\top \mathbb{A}(\gamma_{\text{obj}},\gamma) y = 0,\ \forall (\gamma_{\text{obj}},\gamma) \in \mathcal{G}(x,t)^\perp. \tag{2.7}$$

To see that this suffices, note that for any $w \in \mathbb{R}^k$, the vector $x' := w \otimes y$ satisfies the first constraint in (2.6) since for $(\gamma_{\text{obj}},\gamma) \in \mathcal{G}(x,t)^\perp$, we have

$$(w \otimes y)^\top A(\gamma_{\text{obj}},\gamma)(w \otimes y) = (w^\top w)(y^\top \mathbb{A}(\gamma_{\text{obj}},\gamma)y) = 0.$$

Then, $(w \otimes y, t') \in \mathcal{R}'(x,t)$ if and only if

$$\left\langle A(\gamma_{\text{obj}},\gamma)x + b(\gamma_{\text{obj}},\gamma), w \otimes y \right\rangle - \gamma_{\text{obj}} t' = 0,\ \forall (\gamma_{\text{obj}},\gamma) \in \mathcal{G}(x,t)^\perp.$$

This is a system of $\dim(\mathcal{G}(x,t)^\perp)$-many homogeneous linear equations in the variables $(w,t) \in \mathbb{R}^{k+1}$. Note that as $\mathcal{G}(x,t) \ni q(x) - 2te_{\text{obj}}$, which is nonzero by assumption, we have that $\dim(\mathcal{G}(x,t)^\perp) \leq m$. As $k+1 > m$ by assumption, we deduce that this system has a nontrivial solution. Thus, we conclude that (2.6) is nontrivial if there exists a nonzero $y \in \mathbb{R}^r$ satisfying (2.7).

It remains to construct y. By definition of $\mathcal{S}_{\text{SDP}}$, there exists $Y \succeq 0$ such that

$$\begin{cases} q_{\text{obj}}(x) + \left\langle A_{\text{obj}}, Y \right\rangle \leq 2t, \\ q_i(x) + \langle A_i, Y \rangle \leq 0,\ \forall i \in [m_I], \text{ and} \\ q_i(x) + \langle A_i, Y \rangle = 0,\ \forall i \in [m_I+1, m]. \end{cases} \tag{2.8}$$

Without loss of generality, $Y = (\frac{1}{k} I_k) \otimes \mathbb{Y}$. As $(x,t) \notin \mathcal{S}$, we have that $\mathbb{Y} \in \mathbb{S}^r_+ \setminus \{0\}$ and we may pick a nonzero $y \in \mathbb{R}^r$ such that $yy^\top \preceq \mathbb{Y}$. For notational convenience, let $\ell_{\text{obj}} := y^\top \mathbb{A}_{\text{obj}} y$ and $\ell_i := y^\top \mathbb{A}_i y$ for $i \in [m]$. Note that for any $(\gamma_{\text{obj}},\gamma) \in \Gamma$, we have $A(\gamma_{\text{obj}},\gamma) \succeq 0$, or equivalently $\mathbb{A}(\gamma_{\text{obj}},\gamma) \succeq 0$. Thus, $yy^\top \preceq \mathbb{Y}$ implies that $\left\langle (\gamma_{\text{obj}},\gamma), (\ell_{\text{obj}},\ell) \right\rangle = y^\top \mathbb{A}(\gamma_{\text{obj}},\gamma) y \leq$

$\left\langle \mathbb{A}(\gamma_{\text{obj}}, \gamma), \mathbb{Y} \right\rangle$. Also, from $Y = (\frac{1}{k} I_k) \otimes \mathbb{Y}$ and the relation between the matrices A_{obj}, A_i and $\mathbb{A}_{\text{obj}}, \mathbb{A}_i$, we have $\left\langle \mathbb{A}(\gamma_{\text{obj}}, \gamma), \mathbb{Y} \right\rangle = \left\langle A(\gamma_{\text{obj}}, \gamma), Y \right\rangle$. We deduce that for $(\gamma_{\text{obj}}, \gamma) \in \Gamma$,

$$\left\langle \begin{pmatrix} \gamma_{\text{obj}} \\ \gamma \end{pmatrix}, q(x) - 2te_{\text{obj}} + \begin{pmatrix} \ell_{\text{obj}} \\ \ell \end{pmatrix} \right\rangle \leq \left\langle \begin{pmatrix} \gamma_{\text{obj}} \\ \gamma \end{pmatrix}, q(x) - 2te_{\text{obj}} + \begin{pmatrix} \left\langle A_{\text{obj}}, Y \right\rangle \\ (\langle A_i, Y \rangle)_i \end{pmatrix} \right\rangle \leq 0,$$

where the last inequality follows from (2.8) and $(\gamma_{\text{obj}}, \gamma) \in \Gamma$. This then shows that $q_{\text{obj}}(x) - 2te_{\text{obj}} + (\ell_{\text{obj}}, \ell) \in \Gamma^\circ$. Moreover, because $\mathbb{A}(\gamma_{\text{obj}}, \gamma) \succeq 0$, we have

$$0 \geq -y^\top \mathbb{A}(\gamma_{\text{obj}}, \gamma) y = \left\langle \begin{pmatrix} \gamma_{\text{obj}} \\ \gamma \end{pmatrix}, \begin{pmatrix} -\ell_{\text{obj}} \\ -\ell \end{pmatrix} \right\rangle,$$

which implies $-(\ell_{\text{obj}}, \ell) \in \Gamma^\circ$. We have shown that $q_{\text{obj}}(x) - 2te_{\text{obj}} + (\ell_{\text{obj}}, \ell)$ and $-(\ell_{\text{obj}}, \ell)$ both lie in Γ°. Then, as $q_{\text{obj}}(x) - 2te_{\text{obj}} \in \operatorname{rint}(\mathcal{G}(x,t))$, we deduce that $(\ell_{\text{obj}}, \ell) \in \operatorname{span}(\mathcal{G}(x,t))$. In particular, $y^\top \mathbb{A}(\gamma_{\text{obj}}, \gamma) y = \left\langle (\gamma_{\text{obj}}, \gamma), (\ell_{\text{obj}}, \ell) \right\rangle = 0$ for all $(\gamma_{\text{obj}}, \gamma) \in \mathcal{G}(x,t)^\perp$. ■

Remark 24. SDP exactness in the context of QMPs was previously studied by Beck [17], Beck et al. [20], Wang and Kılınç-Karzan [181]. Specifically, Beck [17] shows that objective value exactness holds whenever $k \geq m$ and Wang and Kılınç-Karzan [181] show that convex hull exactness holds whenever $k \geq m + 2$. Proposition 7 strengthens both of these results by showing that convex hull exactness holds whenever $k \geq m$. □

2.4.3 The partition problem

We next consider the partition QCQP and its SDP relaxation. Recall the partition QCQP: Given $a \in \mathbb{R}^n$, we want to minimize

$$\text{Opt} := \min_{x \in \mathbb{R}^n} \left\{ (a^\top x)^2 : x_i^2 = 1, \, \forall i \in [n] \right\}.$$

Note that $\text{Opt} = 0$ if and only if the vector a can be *partitioned* into two sets of equal weight. Thus, deciding whether $\text{Opt} = 0$ is NP-hard [99]. In this section, we will first give an explicit description of $\mathcal{S}_{\text{SDP}}$ under a minor assumption. This explicit $\mathcal{S}_{\text{SDP}}$ description will then let us conclude that $\operatorname{conv}(\mathcal{S}) \neq \mathcal{S}_{\text{SDP}}$ under the same minor assumption.

Assumption 6. $a \in \mathbb{R}^n_{++}$ and $n \geq 2$. □

Remark 25. Assumption 6 is essentially without loss of generality: It is straightforward to derive a closed form description of $\mathcal{S}_{\text{SDP}}$ when $n = 1$. Similarly, one can relate $\mathcal{S}_{\text{SDP}}$ corresponding to an arbitrary $a \in \mathbb{R}^n$ with the set $\mathcal{S}_{\text{SDP}}$ corresponding to some $a' \in \mathbb{R}^{n'}_{++}$ for $n' := |\{i \in [n] : a_i \neq 0\}|$. □

Proposition 8. *Suppose Assumption 6 holds. Then,*

$$\Gamma = \left\{ (\gamma_{\text{obj}}, \gamma) \in \mathbb{R} \times \mathbb{R}^n : \begin{array}{l} \gamma_{\text{obj}} a a^\top + \operatorname{Diag}(\gamma) \succeq 0 \\ \gamma_{\text{obj}} \geq 0 \end{array} \right\}, \textit{ and}$$

$$\mathcal{S}_{\text{SDP}} = \left\{ (x,t) \in [-1,1]^n \times \mathbb{R} : \, (a^\top x)^2 + \max_{i \in [n]} \left(a_i \sqrt{1 - x_i^2} - \sum_{j \neq i} a_j \sqrt{1 - x_j^2} \right)_+^2 \leq 2t \right\}.$$

See Section B.1 for a proof of this statement.

Recall from [106] that a vector $a \in \mathbb{R}^n_{++}$ is said to be *balanced* if for all $i \in [n]$, $a_i \leq \sum_{j \neq i} a_j$. The following result then follows as a corollary to Proposition 8. (See Section B.1.)

Corollary 7. *Suppose Assumption 6 holds. Then,* $\operatorname{Opt}_{\text{SDP}} = 0$ *if and only if a is balanced.*

As a consequence of Corollary 7 (and the NP-hardness of deciding whether $\operatorname{Opt} = 0$ for the partition QCQP), we see that it is NP-hard to decide whether objective value exactness holds for the partition QCQP. This recovers a result due to Laurent and Poljak [106].

In contrast to the NP-hardness of checking *objective value exactness* for the partition QCQP, the following corollary states that checking *convex hull exactness* for the partition QCQP is a trivial task.

Corollary 8. *Suppose Assumption 6 holds. Then,* $\operatorname{conv}(\mathcal{S}) \neq \mathcal{S}_{\text{SDP}}$.

The proof of Corollary 8 follows from the observation that $\operatorname{conv}(\mathcal{S})$ is polyhedral and that $\mathcal{S}_{\text{SDP}}$ is not polyhedral. See Section B.1 for details.

2.5 Objective value exactness

In this section, we present sufficient conditions for objective value exactness, i.e., the property that $\operatorname{Opt} = \operatorname{Opt}_{\text{SDP}}$. In fact, all of our sufficient conditions imply the stronger condition, which we refer to as *optimizer exactness*, that the optimizers of the QCQP and its SDP relaxation coincide, i.e.,

$$\mathop{\arg\min}_{(x,t) \in \mathcal{S}_{\text{SDP}}} 2t = \mathop{\arg\min}_{(x,t) \in \mathcal{S}} 2t.$$

We begin by presenting sufficient conditions stemming from a primal analysis. These sufficient conditions generalize [181, Theorem 3]. Our second set of sufficient conditions are based on a dual analysis and require the additional assumption that the dual optimum is achieved. These conditions imply further that the optimizers are unique.

2.5.1 Sufficient conditions based on a primal analysis

We begin by presenting a very general sufficient condition for optimizer exactness.

Theorem 12. *Suppose Assumption 4 holds. Furthermore, suppose that for all $(x, t) \in \mathcal{S}_{\mathrm{SDP}} \setminus \mathcal{S}$, there exists closed cones $K_1, K_2 \subseteq \mathbb{R}^{1+m}$ and $(x', t') \in \mathbb{R}^{n+1}$ satisfying*

$$\begin{cases} K_1 \subseteq (q(x) - 2te_{\mathrm{obj}})^\perp \\ -K_2 \cap (q(x) - 2te_{\mathrm{obj}})^\circ = \{0\} \\ K_1 + K_2 \supseteq \Gamma \\ (x')^\top A(\gamma_{\mathrm{obj}}, \gamma)x' = 0,\ \forall(\gamma_{\mathrm{obj}}, \gamma) \in K_1 \\ \left\langle A(\gamma_{\mathrm{obj}}, \gamma)x + b(\gamma_{\mathrm{obj}}, \gamma), x' \right\rangle - \gamma_{\mathrm{obj}} t' \leq 0,\ \forall(\gamma_{\mathrm{obj}}, \gamma) \in K_1 \\ t' < 0 \end{cases}. \tag{2.9}$$

Then, optimizer exactness holds, i.e., $\arg\min_{(x,t)\in\mathcal{S}_{\mathrm{SDP}}} 2t = \arg\min_{(x,t)\in\mathcal{S}} 2t$.

Proof. Let $(x, t) \in \mathcal{S}_{\mathrm{SDP}} \setminus \mathcal{S}$. It suffices to show that $(x, t) \notin \arg\min_{(x,t)\in\mathcal{S}_{\mathrm{SDP}}} 2t$. Let K_1, K_2, x', t' denote the quantities furnished by the assumption.

We claim that for all $\alpha > 0$ small enough, $(x + \alpha x', t + \alpha t') \in \mathcal{S}_{\mathrm{SDP}}$. Indeed, for all $\alpha > 0$ small enough and $(\gamma_{\mathrm{obj}}, \gamma) \in K_1$,

$$\begin{aligned} &\left\langle (\gamma_{\mathrm{obj}}, \gamma), q(x + \alpha x') - 2(t + \alpha t')e_{\mathrm{obj}} \right\rangle \\ &= \alpha^2 \underbrace{(x')^\top A(\gamma_{\mathrm{obj}}, \gamma)x'}_{=0} + 2\alpha \underbrace{\left(\left\langle A(\gamma_{\mathrm{obj}}, \gamma)x + b(\gamma_{\mathrm{obj}}, \gamma), x' \right\rangle - \gamma_{\mathrm{obj}} t' \right)}_{\leq 0} \\ &\quad + \underbrace{\left\langle (\gamma_{\mathrm{obj}}, \gamma), q(x) - 2te_{\mathrm{obj}} \right\rangle}_{=0} \\ &\leq 0. \end{aligned}$$

Next, set $\mathcal{B} \coloneqq K_2 \cap \mathbf{S}^{(1+m)-1}$ so that $\mathrm{cone}(\mathcal{B}) = K_2$. By definition of K_2 and $\mathcal{B}$, we have $-\mathcal{B} \cap (q(x) - 2te_{\mathrm{obj}})^\circ = \varnothing$ so that the map

$$\alpha \mapsto \max_{(\gamma_{\mathrm{obj}}, \gamma)\in\mathcal{B}} \left\langle (\gamma_{\mathrm{obj}}, \gamma), q(x + \alpha x') - 2(t + \alpha t')e_{\mathrm{obj}} \right\rangle$$

is negative at $\alpha = 0$. Note also that this map is a continuous function of α. Then, by continuity, this map is negative for some $\alpha > 0$.

Finally, by linearity and the fact that $K_1 + K_2 \supseteq \Gamma$, we deduce that $(x + \alpha x', t + \alpha t') \in \mathcal{S}_{\mathrm{SDP}}$ for some $\alpha > 0$. This shows $(x, t) \notin \arg\min_{(x,t)\in\mathcal{S}_{\mathrm{SDP}}} 2t$. ∎

We next recover more concrete sufficient conditions by picking K_1 and K_2 appropriately. The following corollary recovers the sufficient condition for objective value exactness (in the setting of polyhedral Γ) presented in [181, Theorem 3].

Corollary 9. *Suppose Assumption 4 holds and that* Γ *is polyhedral. Furthermore, suppose that for all* $(x,t) \in \mathcal{S}_{\text{SDP}} \setminus \mathcal{S}$*, there exists* $(x',t') \in \mathbb{R}^{n+1}$ *satisfying*

$$\begin{cases} (x')^\top A(\gamma_{\text{obj}},\gamma)x' = 0, \ \forall(\gamma_{\text{obj}},\gamma) \in \mathcal{F}(x,t) \\ \left\langle b(\gamma_{\text{obj}},\gamma), x' \right\rangle - \gamma_{\text{obj}} t' \leq 0, \ \forall(\gamma_{\text{obj}},\gamma) \in \mathcal{F}(x,t) \\ t' < 0 \end{cases}. \tag{2.10}$$

Then, $\arg\min_{(x,t)\in\mathcal{S}} 2t = \arg\min_{(x,t)\in\mathcal{S}_{\text{SDP}}} 2t$*.*

Proof. Let $(x,t) \in \mathcal{S}_{\text{SDP}} \setminus \mathcal{S}$. Since Γ is polyhedral, we can write $\Gamma := \operatorname{cone}\left\{(\gamma_{\text{obj}}^{(i)}, \gamma^{(i)})\right\}_{i\in[T]}$ for a finite set of generators. Take,

$$K_1 = \operatorname{cone}\left\{(\gamma_{\text{obj}}^{(i)}, \gamma^{(i)}) : \left\langle(\gamma_{\text{obj}}^{(i)}, \gamma^{(i)}), q(x) - 2te_{\text{obj}}\right\rangle = 0\right\} = \mathcal{F}(x,t)$$

and

$$K_2 = \operatorname{cone}\left\{(\gamma_{\text{obj}}^{(i)}, \gamma^{(i)}) : \left\langle(\gamma_{\text{obj}}^{(i)}, \gamma^{(i)}), q(x) - 2te_{\text{obj}}\right\rangle < 0\right\}.$$

Note that K_1 and K_2 are polyhedral and thus closed. Moreover, the first three requirements of (2.9) are satisfied for this choice of K_1 and K_2. Moreover, note that for every $(\gamma_{\text{obj}},\gamma) \in \mathcal{F}(x,t) \subseteq \Gamma$ we have $A(\gamma_{\text{obj}},\gamma) \succeq 0$ and for any $A \succeq 0$, $x^\top A x = 0$ implies $Ax = 0$. Thus, from $K_1 = \mathcal{F}(x,t)$, we deduce $\left\langle A(\gamma_{\text{obj}},\gamma)x, x'\right\rangle = 0$ for every $(\gamma_{\text{obj}},\gamma) \in K_1$ so that the last three requirements of (2.9) coincide with (2.10). ■

The following corollary derives a sufficient condition for objective value exactness without the assumption that Γ is polyhedral. In words, this assumption supposes that for any $(x,t) \in \mathcal{S}_{\text{SDP}} \setminus \mathcal{S}$, there exists a direction $(x',t') \in \mathbb{R}^{n+1}$ such that $q(x+\alpha x') - 2(t+\alpha t')e_{\text{obj}}$ varies only along the line containing $q(x) - 2te_{\text{obj}}$. In particular, by picking α appropriately, we can achieve $q(x+\alpha x') - 2(t+\alpha t')e_{\text{obj}} = 0$.

Corollary 10. *Suppose Assumption 4 holds. Furthermore, suppose that for all* $(x,t) \in \mathcal{S}_{\text{SDP}} \setminus \mathcal{S}$*, there exists* $(x',t') \in \mathbb{R}^{n+1}$ *satisfying*

$$\begin{cases} (x')^\top A(\gamma_{\text{obj}},\gamma)x' = 0, \ \forall(\gamma_{\text{obj}},\gamma) \in (q(x) - 2te_{\text{obj}})^\perp \\ \left\langle A(\gamma_{\text{obj}},\gamma)x + b(\gamma_{\text{obj}},\gamma), x' \right\rangle - \gamma_{\text{obj}} t' = 0, \ \forall(\gamma_{\text{obj}},\gamma) \in (q(x) - 2te_{\text{obj}})^\perp \\ t' < 0 \end{cases}. \tag{2.11}$$

Then, $\arg\min_{(x,t)\in\mathcal{S}_{\text{SDP}}} 2t = \arg\min_{(x,t)\in\mathcal{S}} 2t$*.*

Proof. Take $K_1 = (q(x) - 2te_{\text{obj}}))^\perp$ and $K_2 = -\operatorname{cone}(q(x) - 2te_{\text{obj}})$. Then, K_1 and K_2 are both closed convex cones and we can easily observe that the first three requirements in (2.9) are automatically satisfied for this choice of K_1 and K_2. The last three requirements in (2.9) coincide with (2.11). ■

2.5.2 Sufficient conditions based on a dual analysis

Next, we give a sufficient condition for objective value exactness depending on a dual analysis. To this end, we define the concave extended-real valued function $\mathbf{d} : \mathbb{R}^m \to \mathbb{R} \cup \{-\infty\}$ by

$$\mathbf{d}(\gamma) \coloneqq \inf_{x\in\mathbb{R}^n} [\gamma, q(x)].$$

Remark 26. Recall here that by Corollary 5, we can write $\mathrm{Opt}_{\mathrm{SDP}}$ in the saddle-point form $\mathrm{Opt}_{\mathrm{SDP}} = \inf_{x\in\mathbb{R}^n} \sup_{\gamma\in\Gamma_1} [\gamma, q(x)]$ given in (2.3). Whence, by coercivity [62, Proposition VI.2.3] we can also write $\mathrm{Opt}_{\mathrm{SDP}} = \sup_{\gamma\in\Gamma_1} \mathbf{d}(\gamma)$. □

The following theorem states that if $\mathbf{d}(\gamma)$ is maximized at a point γ where $A[\gamma] \succ 0$ (e.g., on the interior of Γ_1), then optimizer exactness holds. This theorem can be interpreted as the observation that if the dual to (2.2) in $\mathbb{S}^{n+1}$ has a rank-n optimizer, then (2.2) has a unique rank-1 solution. *This is well-known and has been vastly explored in the literature.* See [38] for one recent example. We state it as a theorem not because it is new or difficult to prove but because of its importance in deriving additional sufficient conditions (see Corollaries 11 and 12).

Theorem 13. *Suppose Assumption 4 holds and that* $\sup_{\gamma\in\Gamma_1} \mathbf{d}(\gamma)$ *is achieved at some* γ^* *for which* $A[\gamma^*] \succ 0$ *(e.g.,* $\gamma^* \in \mathrm{int}(\Gamma)$*). Then,* $\arg\min_{(x,t)\in\mathcal{S}} 2t = \arg\min_{(x,t)\in\mathcal{S}_{\mathrm{SDP}}} 2t$*. Furthermore, the minimizers of these two optimization problems are unique.*

Proof. It suffices to show that $\min_{(x,t)\in\mathcal{S}_{\mathrm{SDP}}} 2t$ has a unique solution (x^*, t^*) and that $(x^*, t^*) \in \mathcal{S}$. Let $(x^*, t^*) \in \arg\min_{(x,t)\in\mathcal{S}_{\mathrm{SDP}}} 2t$ so that $x^* \in \arg\min_x \sup_{\gamma\in\Gamma_1} [\gamma, q(x)]$ and $2t^* = \sup_{\gamma\in\Gamma_1} [\gamma, q(x^*)]$. By the Saddle Point Theorem applied to (2.3), we have

$$0_n = \nabla_x [\gamma^*, q(x^*)] = 2(A[\gamma^*]x^* + b[\gamma^*]).$$

As $A[\gamma^*] \succ 0$, we have $x^* = -A[\gamma^*]^{-1} b[\gamma^*]$. This proves uniqueness of (x^*, t^*).

Note that there exists $\alpha > 0$ such that $[\gamma^*, \gamma^* + \alpha e_i] \subseteq \Gamma_1$ for all $i \in [m_I]$ and $[\gamma^* \pm \alpha e_i] \subseteq \Gamma_1$ for all $i \in [m_I + 1, m]$. Then, by the Saddle Point Theorem we have

$$\begin{aligned} q_i(x^*) &= \nabla_{\gamma_i} [\gamma^*, q(x^*)] \leq 0, \ \forall i \in [m_I], \text{ and} \\ q_i(x^*) &= \nabla_{\gamma_i} [\gamma^*, q(x^*)] = 0, \ \forall i \in [m_I + 1, m]. \end{aligned}$$

We deduce that $q_{\mathrm{obj}}(x^*) \leq \sup_{\gamma\in\Gamma_1} [\gamma, q(x^*)] = 2t^*$. Hence, we conclude that $(x^*, t^*) \in \mathcal{S}$. ■

Remark 27. Note that for any γ for which $A[\gamma] \succ 0$, the dual function $\mathbf{d}(\gamma)$ is the sum of a linear function $c[\gamma]$ and a concave function $-b[\gamma]^\top A[\gamma]^{-1} b[\gamma]$, i.e.,

$$\mathbf{d}(\gamma) = -b[\gamma]^\top A[\gamma]^{-1} b[\gamma] + c[\gamma].$$

We will use this structure to derive more concrete sufficient conditions ensuring that $\mathbf{d}(\gamma)$ is maximized at some point γ^* for which $A[\gamma^*] \succ 0$. □

The following sufficient condition can be interpreted as requiring $-b[\gamma]^\top A[\gamma]^{-1}b[\gamma]$ (and hence also $\mathbf{d}(\gamma)$) to diverge to $-\infty$ as γ approaches a point $\hat{\gamma} \in \Gamma_1$ for which $A[\hat{\gamma}] \not\succ 0$.

Corollary 11. *Suppose Assumption 4 holds and that* $\sup_{\gamma\in\Gamma_1} \mathbf{d}(\gamma)$ *is achieved. Furthermore, suppose that for all* $\gamma \in \Gamma_1$*, we have*

$$A[\gamma] \not\succ 0 \qquad \implies \qquad \exists v \in \ker(A[\gamma]) \textit{ s.t. } \langle v, b[\gamma] \neq 0\rangle.$$

Then, $\arg\min_{(x,t)\in\mathcal{S}} 2t = \arg\min_{(x,t)\in\mathcal{S}_{\text{SDP}}} 2t$*. Furthermore, the minimizers of these two optimization problems are unique.*

Proof. Let $\gamma^* \in \arg\max_{\gamma\in\Gamma_1} \mathbf{d}(\gamma)$. By Theorem 13, it suffices to show that $A[\gamma^*] \succ 0$. Suppose otherwise, so that $A[\gamma^*] \not\succ 0$. Then, the assumptions of the corollary furnish a $v \in \ker(A[\gamma^*])$ such that $\langle v, b[\gamma^*]\rangle \neq 0$. Let $(x^*, t^*) \in \arg\min_{(x,t)\in\mathbb{S}_{\text{SDP}}} 2t$. By the Saddle Point Theorem applied to (2.3), we deduce

$$0 = \langle v, 0_n\rangle = \langle v, \nabla_x[\gamma^*, q(x^*)]\rangle = 2\langle v, A[\gamma^*]x^* + b[\gamma^*]\rangle \neq 0,$$

a contradiction. ■

Remark 28. Burer and Ye [38] study diagonal QCQPs and show [38, Theorem 1] that objective value exactness holds whenever certain systems of equations are infeasible. Specifically, their sufficient condition for diagonal QCQPs can be rewritten as the condition that for any $i \in [n]$, the system $\{\gamma \in \Gamma_1,\ e_i \in \ker(A[\gamma]),\ b[\gamma]_i = 0\}$ is infeasible. Corollary 11 generalizes [38, Theorem 1] by considering general matrices A_i as opposed to diagonal matrices considered in [38].

□

Alternatively, one may impose the slightly weaker condition that $-b[\gamma]^\top A[\gamma]^{-1}b[\gamma]$ gets "sufficiently steep near points at which $A[\gamma] \not\succ 0$" compared to $\|(c_1, \ldots, c_m)\|_2$.

Corollary 12. *Suppose Assumption 4 holds and that* $\sup_{\gamma\in\Gamma_1} \mathbf{d}(\gamma)$ *is achieved. Furthermore, suppose that for all* $\gamma \in \Gamma_1$ *such that* $A[\gamma] \not\succ 0$*, there exists* $\delta \in \mathbb{R}^m$ *such that* $\gamma_\delta := \gamma + \delta \in \operatorname{int}(\Gamma_1)$ *and* $\gamma_{2\delta} := \gamma + 2\delta \in \operatorname{int}(\Gamma_1)$ *and*

$$\left(-b[\gamma_\delta]^\top A[\gamma_\delta]^{-1} b[\gamma_\delta]\right) - \left(-b[\gamma_{2\delta}]^\top A[\gamma_{2\delta}]^{-1} b[\gamma_{2\delta}]\right) \leq -\|\delta\|_2 \sqrt{\sum_{i=1}^m c_i^2}.$$

Then, $\arg\min_{(x,t)\in\mathcal{S}} 2t = \arg\min_{(x,t)\in\mathcal{S}_{\text{SDP}}} 2t$*. Furthermore, the minimizers of these two optimization problems are unique.*

Proof. Let $\gamma^* \in \arg\max_{\gamma\in\Gamma_1} \mathbf{d}(\gamma)$. We will construct an optimizer $\tilde{\gamma} \in \arg\max_{\gamma\in\Gamma_1} \mathbf{d}(\gamma)$ for which $A[\tilde{\gamma}] \succ 0$. The result will then follow from Theorem 13.

If $A[\gamma^*] \succ 0$ then we may take $\tilde{\gamma} = \gamma^*$. Else, let δ be furnished by the assumption of the corollary and note that $\delta \neq 0$. We will set $\tilde{\gamma} = \gamma^*_\delta$. Then, $\tilde{\gamma} \in \operatorname{int}(\Gamma)$ and thus $A[\tilde{\gamma}] \succ 0$. By optimality of γ^*, it suffices to show that $\mathbf{d}(\gamma^*) \leq \mathbf{d}(\gamma^*_\delta)$. As $\mathbf{d}(\gamma)$ is concave and $\gamma^*, \gamma^*_\delta, \gamma^*_{2\delta}$

lie on a line, it suffices in turn to show that $\mathbf{d}(\gamma^*_\delta) \leq \mathbf{d}(\gamma^*_{2\delta})$. Finally, as γ^*_δ and $\gamma^*_{2\delta}$ both lie in $\operatorname{int}(\Gamma_1)$, we may expand

$$\begin{aligned}\mathbf{d}(\gamma^*_\delta) - \mathbf{d}(\gamma^*_{2\delta}) &= \left(-b[\gamma^*_\delta]A[\gamma^*_\delta]^{-1}b[\gamma^*_\delta] + c[\gamma^*_\delta]\right) - \left(-b[\gamma^*_{2\delta}]A[\gamma^*_{2\delta}]^{-1}b[\gamma^*_{2\delta}] + c[\gamma^*_{2\delta}]\right)\\ &\leq -\|\delta\|_2\sqrt{\sum_{i=1}^m c_i^2} - \sum_{i=1}^m \delta_i c_i \leq 0.\end{aligned}$$

Applying Theorem 13 concludes the proof. ■

2.6 Applications: Objective value exactness

In this section, we apply the results of Section 2.5 to random and semi-random QCQPs. Again, these examples offer further evidence that *questions of exactness can be treated systematically whenever* Γ, Γ_1, *or* Γ° *is well-understood*. In fact, the results in this section show that the ideas of Section 2.5 can be applied (at times with additive errors) even when the dual set Γ, Γ_1, or Γ° is not known exactly, but only approximately. The random and semi-random QCQPs considered in this section are motivated by recent work [38, 113], which has treated random QCQPs as a testing ground for understanding the strength or explanatory power of various sufficient conditions for objective value exactness.

We will fix m, the number of quadratic constraints, and take n, the number of variables, to $+\infty$ independently. We will abbreviate "with probability $1 - o(1)$ as $n \to +\infty$" as "asymptotically almost surely" (*a.a.s.*).

The random and semi-random QCQPs we will consider in this section will involve data generated according to the normalized Gaussian Orthogonal Ensemble (NGOE). We collect some basic facts on the NGOE in the following section.

2.6.1 Preliminaries on the (normalized) Gaussian Orthogonal Ensemble

Here, we recall the normalized GOE and a few of its basic properties.

Definition 12. Let $A \in \mathbb{S}^n$ be a random matrix where: each diagonal entry $A_{i,i}$ is i.i.d. $N(0, 1/2n)$; each superdiagonal entry $A_{i,j}$ is i.i.d. $N(0, 1/4n)$; and each subdiagonal entry $A_{i,j}$ is defined by symmetry. We will refer to this distribution as the *normalized Gaussian Orthogonal Ensemble* (NGOE). We will write

$$A \sim \mathrm{NGOE}(n)$$

to denote the fact that A is drawn according to this distribution. □

Remark 29. A different procedure for generating the same distribution is: sample $M \in \mathbb{R}^{n\times n}$ with every entry i.i.d. $N(0, 1/2n)$ and return $A = (M + M^\top)/2$. □

The NGOE is a very well-understood distribution [170]. We will only need a few basic facts. The first two facts state that the NGOE is invariant under various notions of rotation.

Fact 3. *Fix* $U \in \mathbb{R}^{n\times n}$ *orthogonal and let* $A \sim \mathrm{NGOE}(n)$*. Then,* $U^\top AU \sim \mathrm{NGOE}(n)$*.*

Fact 4. *Fix $U \in \mathbb{R}^{k\times k}$ orthogonal and let $A_1, \dots, A_k \overset{i.i.d.}{\sim} \mathrm{NGOE}(n)$. Define $\tilde{A}_i \coloneqq \sum_{j=1}^k U_{i,j} A_j$. Then, $\tilde{A}_1, \dots, \tilde{A}_k \overset{i.i.d.}{\sim} \mathrm{NGOE}(n)$.*

Define also the *normalized semicircular measure*

$$\mu_{\mathrm{nsc}} \coloneqq \frac{2}{\pi}\sqrt{(1-x^2)_+}.$$

The next fact states that the NGOE obeys the semicircle law.

Fact 5. *For any $\psi \in C_c^\infty(\mathbb{R})$ and $\epsilon > 0$,*

$$\lim_{n\to\infty} \Pr\left[\left|\int \psi d\mu_n - \int \psi d\mu_{\mathrm{nsc}}\right| > \epsilon\right] = 0.$$

Here, μ_n is the random *measure constructed by sampling $A \sim \mathrm{NGOE}(n)$ and setting $\mu_n \coloneqq \frac{1}{n}\sum_{j=1}^n \delta_{\lambda_j(A)}$, where $\delta_{\lambda_j(A)}$ is the Dirac measure at $\lambda_j(A)$.*

Finally, we recall that the operator norm of $A \sim \mathrm{NGOE}(n)$ is ≈ 1 asymptotically almost surely.

Fact 6. *Fix $\epsilon > 0$ and let $A \sim \mathrm{NGOE}(n)$. Then, $-\lambda_{\min}(A), \lambda_{\max}(A) \in [1 \pm \epsilon]$ a.a.s..*

2.6.2 Exactness in the fully Gaussian setting

This subsection considers random Euclidean distance minimization problems of the form

$$\inf_{x\in\mathbb{R}^n}\left\{\|x\|_2^2 : \, q_i(x) = 0, \, \forall i \in [m]\right\}. \tag{2.12}$$

In words, we are looking for minimum norm solutions to random quadratic systems.

We will sample each quadratic constraint $q_i(x) = x^\top A_i x + 2b_i^\top x + c_i$ independently where $A_i \sim \mathrm{NGOE}(n)$, $b_i \sim N(0, I_n/n)$, and $c_i \sim N(0,1)$. Here, the normalization on the A_is and b_is are chosen so that $\|A_i\|_2 \approx 1$ and $\|b_i\|_2 \approx 1$.

Below, we will show that for any fixed m and $n \to \infty$, (2.12) has an exact SDP relaxation *a.a.s.*. Specifically, we will apply ideas from Corollary 11 to prove:

Proposition 9. *Let $A_1 \dots, A_m \overset{i.i.d.}{\sim} \mathrm{NGOE}(n)$, $b_1, \dots, b_m \overset{i.i.d.}{\sim} N(0, I_n/n)$ and $c_1, \dots, c_m \overset{i.i.d.}{\sim} N(0,1)$ be independent. Then, a.a.s., optimizer exactness holds in* (2.12)*, i.e.,* $\arg\min_{(x,t)\in\mathcal{S}} 2t = \arg\min_{(x,t)\in\mathcal{S}_{\mathrm{SDP}}} 2t$.

We will highlight the very simple geometric ideas underlying the proof of this result and defer proofs of the more technical lemmas to Section B.2.

We will Proposition 9 using Theorem 13; specifically, we will show that $\mathbf{d}(\gamma)$ is maximized on the interior of Γ_1. As a first step, we observe that Γ_1 contains the unit ball (shrunk by ϵ) *a.a.s.*. The following lemma follows from an ϵ-net argument, concavity of $\lambda_{\min}(A[\gamma])$ as a function of γ, and Facts 5 and 6.

Lemma 15. *Fix* $r \geq 0$ *and* $\epsilon > 0$*. Let* $A_1 \ldots, A_m \overset{\text{i.i.d.}}{\sim} \mathrm{NGOE}(n)$*. Then, a.a.s.,*

$$\lambda_{\min}(A[\gamma]) \in [1 - r \pm \epsilon],\ \forall \gamma \in r\mathbf{S}^{m-1}.$$

In particular, $\mathrm{int}(\Gamma_1) \supseteq B(0, 1-\epsilon)$ *a.a.s..*

Recall Remark 27 that for $\gamma \in \mathrm{int}(\Gamma_1)$, we can write

$$\mathbf{d}(\gamma) = -b[\gamma]^\top A[\gamma]^{-1} b[\gamma] + c[\gamma].$$

The next lemma notes that the first term in $\mathbf{d}(\gamma)$, i.e., $-b[\gamma]^\top A[\gamma]^{-1} b[\gamma]$, concentrates to a *sphere cap* and follows from Fact 5.

Lemma 16. *Fix* $r \in (0, 1)$ *and* $\epsilon > 0$*. Let* $A_1 \ldots, A_m \overset{\text{i.i.d.}}{\sim} \mathrm{NGOE}(n)$*. Then, a.a.s.,*

$$-b[\gamma]^\top A[\gamma]^{-1} b[\gamma] \in [\phi(r) \pm \epsilon],\ \forall \gamma \in r\mathbf{S}^{m-1},$$

where $\phi(r) \coloneqq 2(\sqrt{1 - r^2} - 1)$.

We are now ready to prove Proposition 9. The proof will observe that the gradient of $-b[\gamma]^\top A[\gamma]^{-1} b[\gamma]$ gets "arbitrarily steep at the boundary of Γ_1" so that any maximizer of $\mathbf{d}(\gamma)$ must lie in $\mathrm{int}(\Gamma_1)$. One may compare the proof of Proposition 9 to Corollary 12.

Proof of Proposition 9. For convenience, let $c \in \mathbb{R}^m$ denote the vector with ith coordinate c_i.

Fix $\delta > 0$ and let $M > 0$ such that $\Pr_c[\|c\|_2 \leq M] \geq 1 - \delta/2$. Let $0 < r_1 < r_2 < 1$ and $\epsilon \in (0, 1 - r_2)$ such that

$$\frac{\phi(r_1) - \phi(r_2) - 2\epsilon}{r_2 - r_1} \geq M.$$

In the remainder of the proof, we will condition on the events that $\|c\|_2 \leq M$,

$$\begin{gathered}\lambda_{\min}(A[\gamma]) \geq 1 - r_2 - \epsilon,\ \forall \gamma \in r_2\mathbf{S}^{m-1},\\ -b[\gamma]^\top A[\gamma]^{-1} b[\gamma] \geq \phi(r_1) - \epsilon,\ \forall \gamma \in r_1\mathbf{S}^{m-1},\ \text{and}\\ -b[\gamma]^\top A[\gamma]^{-1} b[\gamma] \leq \phi(r_2) + \epsilon,\ \forall \gamma \in r_2\mathbf{S}^{m-1}.\end{gathered}$$

By Lemmas 15 and 16, this holds with probability $1 - \delta$ for all n large enough.

Let $\gamma \in \Gamma_1 \setminus B(0, r_2)$ and let $\gamma^{(1)}, \gamma^{(2)}$ denote the projections of γ onto $B(0, r_1)$ and $B(0, r_2)$ respectively. We claim that $\mathbf{d}(\gamma^{(2)}) \geq \mathbf{d}(\gamma)$. By concavity of $\mathbf{d}(\gamma)$, it suffices to show that $\mathbf{d}(\gamma^{(1)}) \geq \mathbf{d}(\gamma^{(2)})$. We compute,

$$\begin{aligned}\mathbf{d}(\gamma^{(1)}) &= -b\left[\gamma^{(1)}\right] A\left[\gamma^{(1)}\right]^{-1} b\left[\gamma^{(1)}\right] + \left\langle c, \gamma^{(1)} \right\rangle\\ &\geq \mathbf{d}(\gamma^{(2)}) + (\phi(r_1) - \epsilon) - (\phi(r_2) + \epsilon) + \left\langle c, \gamma^{(1)} - \gamma^{(2)} \right\rangle\\ &\geq \mathbf{d}(\gamma^{(2)}) + (\phi(r_1) - \phi(r_2) - 2\epsilon) - M(r^{(2)} - r^{(1)})\\ &\geq \mathbf{d}(\gamma^{(2)}).\end{aligned}$$

We conclude that $\mathbf{d}(\gamma)$ is maximized on the interior of Γ_1. ■

2.6.3 Almost exactness in a semi-random setting

This section considers semi-random QCQPs of the form

$$\inf_{x\in\mathbb{R}^n}\left\{q_{\text{obj}}(x):\ \begin{array}{l} q_i(x)=0,\ \forall i\in[m] \\ \|x\|_2^2\leq 1\end{array}\right\}. \tag{2.13}$$

For notational convenience, define $q_{m+1}(x) \coloneqq \|x\|_2^2 - 1$.

We will consider the following semi-random model: First, $A_{\text{obj}}, A_1, \dots, A_m$ are independently sampled from $\text{NGOE}(n)$. Then, $b_{\text{obj}}, b_1, \dots, b_m$ and $c_{\text{obj}}, c_1, \dots, c_m$ are chosen arbitrarily (possibly adversarially depending on the A_is).

Below, we will show that for any fixed m, (2.13) has an "almost" exact SDP relaxation *a.a.s.*. Specifically, we will apply ideas from Corollary 10 to prove:

Proposition 10. *Fix $\epsilon > 0$ and let Let $A_{\text{obj}}, \dots, A_m \overset{i.i.d.}{\sim} \text{NGOE}(n)$. Then, a.a.s., for all $b_{\text{obj}}, \dots, b_m \in \mathbb{R}^n$ and $c_{\text{obj}}, \dots, c_m$, we have*

$$\text{Opt} \geq \text{Opt}_{\text{SDP}} \geq \inf_{x\in\mathbb{R}^n}\left\{q_{\text{obj}}(x)-\epsilon:\ \begin{array}{l} q_i(x)\in[\pm\epsilon],\ \forall i\in[m] \\ \|x\|_2^2\leq 1\end{array}\right\}.$$

In a slight departure from previous notation, we will write our dual vector as $(\gamma_{\text{obj}}, \gamma, \gamma_{m+1}) \in \mathbb{R}^{1+m+1}$ where $\gamma_{m+1} \in \mathbb{R}$ corresponds to the constraint $\|x\|_2^2 \leq 1$. As in Section 2.6.2, we will emphasize the main ideas in the proof of Proposition 10 and leave the proofs of more technical lemmas to Section B.2.

The following lemma says that in this random model, Γ will again converge to the second-order cone. This lemma follows from Lemma 15.

Lemma 17. *Fix $r \geq 0$ and $\epsilon > 0$. Let $A_{\text{obj}}, A_1 \dots, A_m \overset{i.i.d.}{\sim} \text{NGOE}(n)$. Then, a.a.s.,*

$$\lambda_{\min}(A(\gamma_{\text{obj}}, \gamma, 1)) \in [1 - r \pm \epsilon],\ \forall (\gamma_{\text{obj}}, \gamma) \in r\mathbf{S}^m.$$

In particular, a.a.s.,

$$\begin{aligned}\left\{(\gamma_{\text{obj}}, \gamma, \gamma_{m+1}):\ \left\|(\gamma_{\text{obj}}, \gamma)\right\|_2 \leq (1-\epsilon)\gamma_{m+1}\right\} &\subseteq \Gamma \\ \subseteq \left\{(\gamma_{\text{obj}}, \gamma, \gamma_{m+1}):\ \left\|(\gamma_{\text{obj}}, \gamma)\right\|_2 \leq (1+\epsilon)\gamma_{m+1}\right\}&.\end{aligned}$$

The following lemma says that a version of Corollary 10 with errors holds in this setting. This lemma follows from an ϵ-net argument along with Fact 5.

Lemma 18. *Fix $\epsilon > 0$ and $N \in \mathbb{N}$. Then, a.a.s., for every $(\gamma_{\text{obj}}, \gamma) \in \mathbf{S}^m$, there exists an N-dimensional vector space $W \subseteq \mathbb{R}^n$ such that*

$$w^\top A(\gamma_{\text{obj}}, \gamma, 1)w \in [\pm\epsilon]\|w\|_2^2,\ \forall w \in W.$$

With Lemmas 17 and 18, we may now prove Proposition 10.

Proof of Proposition 10. Without loss of generality, we assume $\epsilon \in (0, 1/2)$ and $b_{\text{obj}}, b_1, \dots, b_m$, $c_{\text{obj}}, c_1, \dots, c_m$ are picked so that the SDP relaxation is feasible, i.e.,

$$\infty > \inf_{x \in \mathbb{R}^n} \sup_{(\gamma, \gamma_{m+1}) \in \Gamma_1} [(\gamma, \gamma_{m+1}), q(x)]. \tag{2.14}$$

Let x^* denote an optimizer of (2.14) with value $2t^*$. Consider the vector $q(x^*) - 2t^* e_{\text{obj}} \in \mathbb{R}^{1+m+1}$. Without loss of generality, we may assume that $q(x^*) - 2t^* e_{\text{obj}}$ is both nonzero and on the boundary of Γ°. By Lemma 17 and the assumption that $q(x^*) - 2t^* e_{\text{obj}} \in \text{bd}(\Gamma^\circ)$, we have

$$\tau := \sqrt{(q_{\text{obj}}(x^*) - 2t^*)^2 + \sum_{i=1}^m q_i(x^*)^2} \in [1 \pm \epsilon] q_{m+1}(x^*).$$

Next, as $q(x^*) - 2t^* e_{\text{obj}}$ is nonzero, we have that $0 < q_{m+1}(x^*) = 1 - \|x^*\|^2$, i.e., $\|x^*\|^2 < 1$. Hence, by definition of τ, we have $|\tau| \leq 1 + \epsilon$.

Set $(f_{\text{obj}}, f, f_{m+1}) := \left(\frac{q_{\text{obj}}(x^*) - 2t^*}{\tau}, \frac{q_1(x^*)}{\tau}, \dots, \frac{q_m(x^*)}{\tau}, 1\right)$ so that $\left\|(f_{\text{obj}}, f)\right\|_2 = 1$.

Note that by Lemma 18, there exists a subspace W of dimension $m + 3$ such that

$$w^\top A(f_{\text{obj}}, f, f_{m+1}) w \in [\pm\epsilon] \|w\|_2^2, \ \forall w \in W.$$

By a dimension counting argument, there exists a unit $w \in W$ satisfying

$$\left\langle A(\gamma_{\text{obj}}, \gamma, \gamma_{m+1}) x^* + b(\gamma_{\text{obj}}, \gamma, \gamma_{m+1}), w \right\rangle = 0, \ \forall (\gamma_{\text{obj}}, \gamma, \gamma_{m+1}) \in \mathbb{R}^{1+m+1}. \tag{2.15}$$

Then, for this vector w we have

$$\begin{cases} w^\top A(f_{\text{obj}}, f, 1) w \in [\pm\epsilon], \\ w^\top A(0, 0_m, 1) w = 1, \text{ and} \\ w^\top A(\gamma_{\text{obj}}, \gamma, 1) w \geq 0, \ \forall (\gamma_{\text{obj}}, \gamma) \in (1 - \epsilon) \mathbf{S}^m. \end{cases} \tag{2.16}$$

Here, the first two relations follow from $\|w\|_2^2 = 1$. The third relation follows from Lemma 17, which implies that $A(\gamma_{\text{obj}}, \gamma, 1) \succeq 0$ for all $(\gamma_{\text{obj}}, \gamma) \in (1 - \epsilon) \mathbf{S}^m$.

Set $v_{\text{obj}} := w^\top A_{\text{obj}} w$ and $v \in \mathbb{R}^m$ where $v_i := w^\top A_i w$ for $i \in [m]$. Note that by (2.15), we have

$$q(x^* + \alpha w) - 2t^* e_{\text{obj}} = \left(q(x^*) - 2t^* e_{\text{obj}}\right) + \alpha^2 (v_{\text{obj}}, v, 1).$$

Then, by the first two lines of (2.16),

$$\begin{aligned} \left\langle (v_{\text{obj}}, v), (f_{\text{obj}}, f) \right\rangle &= f_{\text{obj}} w^\top A_{\text{obj}} w + \sum_{i=1}^m f_i w^\top A_i w \\ &= w^\top A(f_{\text{obj}}, f, 1) w - w^\top w \in [-1 \pm \epsilon]. \end{aligned}$$

Next, by the third line of (2.16), we have $\left\|(v_{\mathrm{obj}}, v)\right\|_2 \leq 1/(1-\epsilon)$. Set $(\delta_{\mathrm{obj}}, \delta) \coloneqq (v_{\mathrm{obj}}, v) + (f_{\mathrm{obj}}, f)$. We will argue that $(\delta_{\mathrm{obj}}, \delta)$ is small by bounding its components along (f_{obj}, f) and orthogonal to (f_{obj}, f),

$$\left\|(\delta_{\mathrm{obj}}, \delta)\right\|^2 \leq \epsilon^2 + \left(\frac{1}{(1-\epsilon)^2} - (1-\epsilon)^2\right) = O(\epsilon).$$

Finally, set $\tilde{x} = x^* + \alpha w$ where $\alpha = \sqrt{1 - \|x^*\|^2}$ and note that

$$\begin{aligned} q(\tilde{x}) - 2t^* e_{\mathrm{obj}} &= q(x^*) - 2t^* e_{\mathrm{obj}} + (1 - \|x^*\|_2^2)\Big(v_{\mathrm{obj}}, v, 1\Big) \\ &= q(x^*) - 2t^* e_{\mathrm{obj}} + (1 - \|x^*\|_2^2) e_{m+1} + \tau(v_{\mathrm{obj}}, v, 0) \\ &\qquad + (1 - \|x^*\|_2^2 - \tau)(v_{\mathrm{obj}}, v, 0) \\ &= q(x^*) - 2t^* e_{\mathrm{obj}} - (\tau f_{\mathrm{obj}}, \tau f, \|x^*\|_2^2 - 1) \\ &\qquad + \tau(\delta_{\mathrm{obj}}, \delta, 0) + (1 - \|x^*\|_2^2 - \tau)(v_{\mathrm{obj}}, v, 0) \\ &= \tau(\delta_{\mathrm{obj}}, \delta, 0) + (1 - \|x^*\|_2^2 - \tau)(v_{\mathrm{obj}}, v, 0). \end{aligned}$$

The conclusion then follows from the bounds $|\tau| \leq (1+\epsilon)$, $\left\|(\delta_{\mathrm{obj}}, \delta)\right\|_2 = O(\sqrt{\epsilon})$, $\left|1 - \|x^*\|_2^2 - \tau\right| \leq \epsilon$ and $\left\|(v_{\mathrm{obj}}, v)\right\|_2 \leq 1/(1-\epsilon)$. ■

3 Rank-one-generated cones

This chapter is based on joint work [7] with C.J. Argue and Fatma Kılınç-Karzan, [100] with Fatma Kılınç-Karzan.

A closed convex conic subset $\mathcal{S}$ of the positive semidefinite (PSD) cone is rank-one generated (ROG) if all of its extreme rays are generated by rank-one matrices. The ROG property of $\mathcal{S}$ is closely related to the exactness of SDP relaxations of nonconvex quadratically constrained quadratic programs (QCQPs) related to $\mathcal{S}$. In this chapter, we consider the case where $\mathcal{S}$ is obtained as the intersection of the PSD cone with finitely many homogeneous linear matrix inequalities and conic constraints and identify sufficient conditions that guarantee that $\mathcal{S}$ is ROG. In the case of two linear matrix inequalities, we also establish the necessity of our sufficient conditions. This extends one of the few settings from the literature—the case of one linear matrix inequality and the S-lemma—where an explicit characterization for the ROG property exists. We additionally show how to apply ROG results to derive exactness properties of QCQPs as well as optimization problems involving ratios of quadratic functions.

3.1 Introduction

Let $\mathbb{S}^n$ denote the real vector space of $n \times n$ real symmetric matrices and $\mathbb{S}^n_+$ the cone of positive semidefinite matrices. We will say that a closed convex cone $\mathcal{S} \subseteq \mathbb{S}^n_+$ is *rank-one generated* (ROG)[1] if

$$\mathcal{S} = \operatorname{conv}(\mathcal{S} \cap \{xx^\intercal : x \in \mathbb{R}^n\}),$$

where $\operatorname{conv}(\cdot)$ is the convex hull operation. In words, a closed convex cone $\mathcal{S}$ is ROG if and only if it is equal to the convex hull of its rank-one matrices.

In most applications, the cone $\mathcal{S} \subseteq \mathbb{S}^n_+$ will be represented as the intersection of $\mathbb{S}^n_+$ with a (possibly infinite) system of linear matrix inequalities (LMIs). Specifically, we will consider cones of the form

$$\mathcal{S}(\mathcal{M}) := \{X \in \mathbb{S}^n_+ : \langle M, X\rangle \geq 0,\, \forall M \in \mathcal{M}\},$$

where $\mathcal{M} \subseteq \mathbb{S}^n$. Note also that any closed convex cone $\mathcal{S} \subseteq \mathbb{S}^n_+$ can be expressed in this form. An obvious question then is: What does the ROG property of $\mathcal{S}(\mathcal{M})$ correspond to in terms of $\mathcal{M}$, its defining LMIs?

[1] We will see in Lemma 19 that the definitions of ROG cones given in the first sentence of the abstract and the second sentence of the main body are equivalent. For the purposes of our developments, we will begin with the definition given in the main body.

While our main focus will be on closed convex cones, our results also have implications in the more general setting of arbitrary closed convex sets $\mathcal{S} \subseteq \mathbb{S}^n_+$ and their defining LMIs.

3.1.1 Motivation

The ROG property is important in studying semidefinite program (SDP) relaxations of quadratically constrained quadratic programs (QCQPs).

QCQPs are a fundamental class of optimization problems that arise naturally in many areas. Indeed, many problems including binary integer linear programs, max-cut, max-clique, certain robust optimization problems and polynomial optimization problems can be readily recast as QCQPs (see [11, 25, 91] and references therein).

It is well known that any QCQP can be reformulated as an SDP in a lifted space with an additional nonconvex rank constraint. Dropping this rank constraint leads to the standard SDP relaxation [161]. A general QCQP and its SDP relaxation are given by

$$\begin{aligned}\inf_{y\in\mathbb{R}^{n-1}}\{q_0(y) :\ q_i(y) \geq 0,\ \forall i \in [m]\} &= \inf_{x\in\mathbb{R}^n}\left\{x^\intercal M_0 x :\ \begin{array}{l} x^\intercal M_i x \geq 0,\ \forall i \in [m] \\ x_1^2 = 1 \end{array}\right\}\\ &\geq \inf_{X\in\mathbb{S}^n_+}\left\{\langle M_0, X\rangle :\ \begin{array}{l} \langle M_i, X\rangle \geq 0,\ \forall i \in [m] \\ X_{1,1} = 1 \end{array}\right\}.\end{aligned} \tag{3.1}$$

Here, $[m] := \{1, \dots, m\}$, the functions q_i are quadratic functions of the form $q_i(y) = y^\intercal A_i y + 2b_i^\intercal y + c_i$, the vector x should be thought of as $\left(\begin{smallmatrix}1\\ y\end{smallmatrix}\right)$, and the matrices M_i are defined as $M_i := \left(\begin{smallmatrix}c_i & b_i^\intercal\\ b_i & A_i\end{smallmatrix}\right)$.

In general, it is NP-hard to determine whether the SDP relaxation of a given QCQP is *exact*, i.e., when equality holds in (3.1) (see [106]). Nevertheless, sufficient conditions that ensure equality in (3.1) are of great interest, and thus establishing such conditions has attracted a lot of attention in the literature.

Geometrically, SDP exactness occurs if and only if there exist rank-one matrices in the feasible domain of the SDP approaching its optimum value. The ROG property is a similar but stronger notion of exactness. Specifically, if the cone

$$\mathcal{S}(\{M_1, \dots, M_m\}) = \left\{X \in \mathbb{S}^n_+ :\ \ \langle M_i, X\rangle \geq 0,\ \forall i \in [m]\ \right\} \tag{3.2}$$

is ROG, then there exist rank-one matrices in the right hand side of (3.1) approaching its optimum value *for every* choice of M_0 such that the right hand side of (3.1) is finite. In other words, if the cone in (3.2) is ROG, then equality holds in (3.1) *for every* choice of objective function such that the SDP value is finite. In the case of homogeneous QCQPs, i.e., where all $b_i = 0$ and $c_i = 0$ for $i = 0, 1, \dots, m$, then (3.2) is ROG if and only if the underlying SDP relaxation is exact for every choice of objective function. See Section 3.5.1 for a more detailed discussion of how equality holding in (3.1) relates to the ROG property of $\mathcal{S}(\{M_1, \dots, M_m\})$.

The ROG property is a natural strengthening of SDP exactness. Consider, for example, the problem of minimizing an arbitrary quadratic function over an ellipsoid. The celebrated S-lemma

[190] guarantees that the SDP relaxation of this problem is exact regardless of the choice of objective function. One way of reinterpreting this statement is as the fact that

$$\mathcal{S}(\{M_1\}) = \left\{ X \in \mathbb{S}^n_+ : \quad \langle M_1, X \rangle \geq 0 \ \right\}$$

is ROG when M_1 corresponds to an ellipsoid constraint.[2] From a different perspective, the ROG property of spectrahedra can be thought of as an analogue of the *integrality* property of polyhedra for linear programming relaxations of integer programs. While there are well-known sufficient conditions such as total unimodularity or total dual integrality for the integrality property of polyhedra (see [52] for recent developments and earlier references), the research on sufficient conditions for the ROG property of spectrahedra is much more recent and limited.

The ROG property is also relevant in the context of sum-of-squares (SOS) programming. Consider a real homogeneous quadratic variety $V := \{x \in \mathbb{R}^n : x^\intercal M_i x = 0, \ \forall i \in [m]\}$. Let $\mathcal{P}_V$ denote the set of nonnegative quadratic forms on V, i.e., $\mathcal{P}_V := \{M \in \mathbb{S}^n : x^\intercal M x \geq 0, \ \forall x \in V\}$. Let Σ_V denote the set of quadratic forms that are "immediately nonnegative" on V, i.e., $\Sigma_V := \mathbb{S}^n_+ + \operatorname{span}\{M_i : i \in [m]\}$, where $\operatorname{span}(\cdot)$ is the span (linear hull) of the given elements.

It is clear that $\Sigma_V \subseteq \mathcal{P}_V$. A direct calculation shows that the dual cones of $\mathcal{P}_V$ and Σ_V are given by

$$\begin{aligned} \mathcal{P}_V^* &= \operatorname{conv}\{xx^\intercal : \langle M_i, xx^\intercal \rangle = 0, \ \forall i \in [m]\} \quad \text{and} \\ \Sigma_V^* &= \{X \in \mathbb{S}^n_+ : \langle M_i, X \rangle = 0, \ \forall i \in [m]\}, \end{aligned}$$

respectively. Therefore, $\Sigma_V = \mathcal{P}_V$ if and only if $\Sigma_V^* = \mathcal{P}_V^*$, which holds if and only if Σ_V^* is rank-one generated. In other words, every nonnegative quadratic form on V is "immediately nonnegative" if and only if Σ_V^* is ROG. See [29, Section 6] for further connections and applications of the ROG property in the context of real algebraic geometry and statistics.

3.1.2 Related literature

Bounds on the rank of extreme points of general spectrahedra. A rich line of research has proved optimal worst-case bounds on the rank of extreme points of a spectrahedron (an affine slice of the PSD cone) in terms of the number of its defining linear matrix equalities (LMEs) [14, 69, 141]; see also [16, Chapter II.13]. It is known that given m LMEs, if there exists a positive semidefinite (PSD) solution to the LMEs, then there also exists a PSD solution with rank at most r for any integral r such that

$$m < \binom{r+2}{2}.$$

From this, we may deduce[3] that any spectrahedron defined by m LMEs has only extreme points of rank at most r for any integral r satisfying $m + 1 < \binom{r+2}{2}$. In particular, taking $r = 1$, this bound

[2] Along with the observation that the SDP relaxation of this problem is always bounded.

[3] After taking into account an additional LME due to the objective function and applying Strasziewicz Theorem (see [152, Theorem 18.6]).

implies that any spectrahedron defined by a single LME is ROG. Unfortunately, this bound does not shed much light onto (even the existence of) ROG spectrahedra in the case where $m > 1$. Although this bound is tight in general, it does not exploit potential structure in the defining LMEs. In other words, it is possible to achieve stronger bounds on the rank of extreme points of spectrahedra with additional structure. Our work complements this line of research by examining properties of systems of LMEs and LMIs that guarantee the ROG property beyond the case of $m = 1$.

SDP EXACTNESS. The question of when equality holds in (3.1) has attracted significant interest. Within this line of research, a number of papers study the classical trust region subproblem (TRS)—the problem of minimizing a nonconvex quadratic function over an ellipsoid—and its variants, and identify cases under which an exact SDP reformulation is possible. This line of work can be traced back to Yakubovich's S-procedure [67, 190] (also known as the S-lemma) and the work of Sturm and Zhang [167]. We refer the interested readers to the excellent survey by Burer [33] and references therein.

It is worth noting that although the results in [33] are stated in terms of the exactness of (strengthened) SDP relaxations, the underlying arguments in fact establish the ROG property for the corresponding SDP feasible domains. For example, the domain of the SDP relaxation associated with the classical TRS is the intersection of $\mathbb{S}^n_+$ with a single LMI, which is well known to be ROG via S-lemma. In the other variants of TRS examined in [33], the domain of the associated exact SDP reformulation involves at least one problem specific conic constraint (in fact a second-order cone constraint), and consequently is described by an infinite family of well-structured LMIs.

These lines of work can be thought of as addressing the special case where there are only a few (usually one or two) nonconvex quadratic functions in the QCQP on the left of (3.1). In contrast, Burer and Ye [38] and Wang and Kılınç-Karzan [181] recently introduced more general sufficient conditions for SDP exactness which do not make explicit assumptions on the number of nonconvex quadratic functions. As an example, it can be shown that SDP exactness holds whenever a natural symmetry parameter of the QCQP is large enough and the set of convex Lagrange (dual) multipliers is polyhedral [181]. See also [179] for sufficient conditions that make weaker assumptions on the geometry of the set of convex Lagrange multipliers. Some of these sufficient conditions for SDP exactness [179, 181] have also been shown to guarantee that the (projection of the) epigraph of the SDP relaxation coincides exactly with the convex hull of the epigraph of the QCQP. In particular, the convex hulls of epigraphs of "highly-symmetric" QCQPs with favorable geometry are semidefinite-representable. Results in this line of work generally depend heavily on how the objective function interacts with the constraints. Our work complements this line of research by establishing conditions for SDP exactness which are oblivious to the objective function.

ALGEBRO-GEOMETRIC PROPERTIES OF ROG SPECTRAHEDRA. The ROG property has also been studied from a more algebro-geometric perspective [29, 83].

Hildebrand [83] studies algebraic properties of ROG cones obtained by adding homogeneous LMEs to $\mathbb{S}^n_+$, and proves important facts about their representations. The study begins by exploring the minimal defining polynomials and facial structure of ROG cones. These properties are then used to build the main contribution of [83]: The geometry of an ROG cone determines its representation as a linear section of a PSD cone (of any dimension) uniquely up to an isomorphism on the underlying vector space. Additional results in this paper include a complete classification of ROG cones of degree[4] at most four as well as a number of operations on ROG cones (the direct product, full extension, and intertwining operations) that preserve the ROG property.

Blekherman et al. [29] study the ROG property of the cones Σ_V^* (see Section 3.1.1) using techniques from real algebraic geometry and establish a connection between the geometry of Σ_V^* and the property $N_{2,p}$ of the defining ideal of V.[5] Specifically, one of the main results in [29] is that, for general real projective varieties V, if Σ_V^* has an extreme ray of rank $p > 1$ then V does not satisfy the property $N_{2,p}$. This result is then strengthened in [29, Theorem 20] to show that a spectrahedral cone $\mathcal{S}$ defined by LMEs is ROG if and only if $\mathcal{S} = \Sigma_V^*$ for a non-degenerate, reduced, 2-regular, totally real scheme V. Finally, [29] also examines consequences of this connection to problems from real algebraic geometry, convex geometry, statistics, and real analysis, such as the positive semidefinite matrix completion problem.

In contrast to [29, 83], our results deal with possibly infinitely many LMIs. The ROG property of such sets is not obvious and does not follow immediately from the ROG property of spectrahedral cones defined by LMEs. Indeed, we will see that both replacing equalities with inequalities (Remark 39) and lifting inequalities to equalities (Example 9) can destroy the ROG property of a spectrahedral cone. In addition, our more general setup allows us to handle additional interesting spectrahedral cones that have conic constraints, for example those arising from variants of the TRS. We also discuss implications of the ROG property in terms of the exactness of SDP relaxations of QCQPs and explicit convex hull characterizations of sets defined by quadratic inequality constraints. Finally, all of the proofs in this chapter follow from elementary linear algebra and convex analysis. In particular, we hope that our results and their proofs shed light on the ROG property for readers less familiar with algebraic geometry.

ROG spectrahedra arising from PSD matrix completion. The ROG property has also been studied for spectrahedra arising in the matrix completion literature. PSD matrix completion arises in a number of areas—for example in statistics, this problem is related to maximum likelihood estimation in Gaussian graphical models [56]. Let E denote the edge set of an undirected graph on n vertices that contains all self-loops. Let $K \subseteq \mathbb{S}^n$ denote the projection of $\mathbb{S}^n_+$ onto the indices in E. Then, a matrix Y that is specified only on E has a PSD completion if and only if it lies in the cone K. A short calculation shows that

$$K = \left\{ Y \in \mathbb{S}^n : \begin{array}{l} Y_{i,j} = 0,\ \forall (i,j) \notin E \\ \langle X, Y \rangle \geq 0,\ \forall X \in \mathcal{S} \end{array} \right\}, \qquad \text{where}$$
$$\mathcal{S} = \left\{ X \in \mathbb{S}^n_+ : X_{i,j} = 0,\ \forall (i,j) \notin E \right\}.$$

Consequently, the condition that every fully specified submatrix of Y is positive semidefinite is necessary and sufficient for Y to have a PSD completion if and only $\mathcal{S}$ is ROG. It is well-known that $\mathcal{S}$ is ROG if and only if E is the edge set of a chordal graph[6] on n vertices [3, 78, 145].

[4] This is the degree of the minimal defining polynomial. This quantity is shown to be equivalent to the maximum rank over matrices in the ROG cone.

[5] A real projective variety V satisfies property $N_{2,p}$ for an integer $p \geq 1$ if the jth syzygy module of the homogeneous ideal of V is generated in degree at most $j + 2$ for all $j < p$.

[6] A graph is chordal if every minimal cycle in the graph has at most 3 edges.

3.1.3 Overview and outline of the chapter

In this chapter, we study necessary and/or sufficient conditions under which the intersection of the positive semidefinite cone with a set of homogeneous LMIs is an ROG cone. A summary of our contributions, along with an outline of the chapter, is as follows:

i In Section 3.2, we introduce our main terminology and basic tools. Specifically, we show how the ROG property behaves when we switch from linear matrix inequalities (LMIs) to linear matrix equalities (LMEs) and how the ROG property for LMEs is characterized by the existence of solutions of quadratic systems. In Section 3.2.5, using our basic tools, we recover the well-known fact that a set defined by a single LMI/LME is ROG, i.e., the S-lemma, and discuss a few implications for a simple sufficient condition in the case of two LMIs/LMEs.

ii In Section 3.3, we establish a number of new sufficient conditions for the ROG property. As an example, we show that $\mathcal{S}$ is ROG when $\mathcal{S} = \{X \in \mathbb{S}^n_+ : Xc \in K\}$ for a fixed vector c and an arbitrary closed convex cone K. We also provide a number of examples to demonstrate that even simple extensions of our sufficient conditions are not possible. We conclude this section by recovering the well-known result that the SDP relaxation strengthened with a second-order cone reformulation-linearization technique (SOC-RLT) inequality is exact for the variant of the TRS with a single linear inequality constraint.

iii A well-known consequence of the S-lemma is that the set $\mathcal{S}(\mathcal{M})$ is ROG whenever $\mathcal{M} = \{M\}$ is a single LMI; see e.g., Ye and Zhang [194, Lemma 2.2]. In Section 3.4, we give a complete characterization of ROG cones defined by two LMIs. One of our main results states a necessary and sufficient condition on the matrices M_1 and M_2 which ensures that the set $\mathcal{S}$ is ROG. In particular, we establish in Theorem 16 that such a set is ROG if and only if the LMIs defined by M_1 and M_2 either "only interact" on a single face of $\mathbb{S}^n_+$ where they induce the same inequality constraint or both M_1 and M_2 have a specific indefinite rank-two structure. We conclude that in the case of $m = 2$, there exist simple certificates of the ROG property.

iv In Section 3.5, we give a few applications of ROG cones. In particular, we show how results on the ROG property of convex cones can be translated into inhomogeneous SDP exactness results and SDP-based convex hull descriptions of quadratically constrained sets. We then apply our ROG-based sufficient condition for exactness of the SDP relaxation to a simple set involving binary and continuous variables linked through a complementarity constraint. This gives a new method for deriving the well-known perspective reformulation for the convex hull of this set. We additionally present a number of examples that highlight how our ROG-based sufficient conditions for the SDP exactness and convex hull descriptions differ from other SDP exactness conditions in the literature. We close this chapter by showing how to combine our ROG results with a "re-homogenization" trick to minimize ratios of quadratic functions over ROG domains. We give applications to the regularized total least squares problem and a Stackelberg prediction game with a least squares loss function. The results in this section are self-contained and serve as additional motivation for the main study.

We will compare our results with the literature in further detail in the sections as outlined above.

3.1.4 Additional notation

For $M \in \mathbb{R}^{n\times n}$, let $\mathrm{Sym}(M) := (M + M^\intercal)/2 \in \mathbb{S}^n$. For a cone K in a Euclidean space $\mathcal{E}$, let $\mathrm{extr}(K)$ denote its extreme rays and define $K^* := \{y \in \mathbb{E} : \langle x, y\rangle \geq 0,\ \forall x \in K\}$ to be the dual cone of K. Given a subspace $W \subseteq \mathbb{R}^n$ and $x \in \mathbb{R}^n$, let $x_W \in W$ denote the projection of x onto W. For $x \in W$ and $y \in W^\perp$, let $x \oplus y$ denote their direct sum. For $X \in \mathbb{S}^W$ and $Y \in \mathbb{S}^{W^\perp}$, let $X \oplus Y$ denote their direct sum, i.e., the unique matrix in $\mathbb{S}^n$ such that $(x\oplus y)^\intercal (X \oplus Y)(x \oplus y) = x^\intercal X x + y^\intercal Y y$ for all $x \in W$ and $y \in W^\perp$.

3.2 Properties of ROG cones

3.2.1 Definitions

Given $\mathcal{M} \subseteq \mathbb{S}^n$, define

$$\mathcal{S}(\mathcal{M}) := \{X \in \mathbb{S}^n_+ : \langle M, X\rangle \geq 0,\ \forall M \in \mathcal{M}\}.$$

Note that $\mathcal{S}(\mathcal{M})$ is a closed convex cone. We are interested in the following property of such sets.

Definition 13. A closed convex cone $\mathcal{S} \subseteq \mathbb{S}^n_+$ is *rank-one generated* (ROG) if

$$\mathcal{S} = \mathrm{conv}(\mathcal{S} \cap \{xx^\intercal : x \in \mathbb{R}^n\}). \qquad \square$$

Remark 30. Note that when $\mathcal{S} \subseteq \mathbb{S}^n_+$ is a closed convex cone, we have $\mathrm{conv}(\mathcal{S}\cap\{xx^\intercal : x \in \mathbb{R}^n\}) = \mathrm{clconv}(\mathcal{S} \cap \{xx^\intercal : x \in \mathbb{R}^n\})$. □

We will make extensive use of the following definitions and basic facts.

Definition 14. For $X \in \mathbb{S}^n$ nonzero, the ray spanned by X is

$$\mathbb{R}_+ X := \{\alpha X : \alpha \geq 0\}.$$

Let $\mathcal{S} \subseteq \mathbb{S}^n_+$ be a closed convex cone and suppose $X \in \mathcal{S}$ is nonzero. We say that $\mathbb{R}_+X$ is an extreme ray of $\mathcal{S}$ if for any $Y, Z \in \mathcal{S}$ such that $X = (Y + Z)/2$, we must have $Y, Z \in \mathbb{R}_+X$. □

Fact 7. *Let $X \in \mathbb{S}^n_+$. Then, $x \in \mathrm{range}(X)$ if and only if there exists $\epsilon > 0$ such that $X - \epsilon xx^\intercal \in \mathbb{S}^n_+$.*

Fact 8. *Let $\mathcal{S} \subseteq \mathbb{S}^n_+$ be a closed convex cone. Then, for $X \neq 0$, $\mathbb{R}_+X$ is an extreme ray of $\mathcal{S}$ if and only if for every Y,*

$$[X - Y, X + Y] \subseteq \mathcal{S} \implies \exists\, \alpha \in \mathbb{R} \text{ such that } Y = \alpha X.$$

The following fact follows immediately from Facts 7 and 8.

Fact 9. *Let $\mathcal{S} \subseteq \mathbb{S}^n_+$ be a closed convex cone. If $X \in \mathcal{S}$ has $\mathrm{rank}(X) = 1$, then $\mathbb{R}_+X$ is an extreme ray of $\mathcal{S}$.*

Lemma 19. *Let $\mathcal{S} \subseteq \mathbb{S}^n_+$ be a closed convex cone. Then, $\mathcal{S}$ is ROG if and only if for each extreme ray $\mathbb{R}_+ X$ of $\mathcal{S}$ we have* $\operatorname{rank}(X) = 1$.

Proof. ($\Leftarrow$) Note that as $\mathcal{S}$ is a subset of $\mathbb{S}^n_+$, it must be pointed. Then, as a closed convex pointed cone is the convex hull of its extreme rays, we have that $\mathcal{S} = \operatorname{conv}(\mathcal{S} \cap \{xx^\intercal : x \in \mathbb{R}^n\})$.

($\Rightarrow$) Let $\mathbb{R}_+ X$ denote an extreme ray of $\mathcal{S}$. As $\mathcal{S}$ is ROG, we may by assumption write $X = \sum_{i=1}^k x_i x_i^\intercal$ where $x_i x_i^\intercal \in \mathcal{S}$ for every $i \in [k]$. Then, as $\mathbb{R}_+ X$ is an extreme ray of $\mathcal{S}$, we must have $x_i x_i^\intercal \in \mathbb{R}_+ X$ for every $i \in [k]$. Thus, we deduce that X is rank-one. ■

The following fact allows us to decompose positive semidefinite matrices which are identically zero on a given subspace.

Lemma 20. *Let $X \in \mathbb{S}^n_+$. Suppose $W \subseteq \mathbb{R}^n$ is a subspace on which $X_W = 0$. Then, we can write $X = 0_W \oplus X_{W^\perp}$.*

Proof. By performing an orthonormal change of variables, we may assume without loss of generality that W corresponds to the first k coordinates of $\mathbb{R}^n$ and $W^\perp$ corresponds to the last $n-k$ coordinates of $\mathbb{R}^n$. We can then write X as a block matrix

$$X = \begin{pmatrix} X_W & Y \\ Y^\intercal & X_{W^\perp} \end{pmatrix}.$$

Then, as $X \in \mathbb{S}^n_+$ and $X_W = 0$, we deduce that $Y = 0$. In particular, $X = 0_W \oplus X_{W^\perp}$. ■

3.2.2 Relating LMIs to LMEs

Given a set $\mathcal{M} \subseteq \mathbb{S}^n$, we will quickly switch from studying $\mathcal{S}(\mathcal{M})$ to sets defined by LMEs, i.e., sets of the form

$$\mathcal{T}(\mathcal{M}) \coloneqq \{X \in \mathbb{S}^n_+ : \langle M, X \rangle = 0,\ \forall M \in \mathcal{M}\}.$$

Sets of the form $\mathcal{T}(\mathcal{M})$ are simpler to analyze than sets of the form $\mathcal{S}(\mathcal{M})$.

Remark 31. It is clear that given any $\mathcal{M} \subseteq \mathbb{S}^n$, we have $\mathcal{S}(\mathcal{M}) = \mathcal{S}(\operatorname{clcone}(\mathcal{M}))$ and $\mathcal{T}(\mathcal{M}) = \mathcal{T}(\operatorname{span}(\mathcal{M}))$. In particular, we may without loss of generality assume that $\mathcal{M}$ is finite when analyzing sets of the form $\mathcal{T}(\mathcal{M})$—simply replace $\mathcal{M}$ with a finite basis of $\operatorname{span}(\mathcal{M})$. On the other hand, $\operatorname{clcone}(\mathcal{M})$ is not necessarily finitely generated. □

We now present a series of lemmas relating $\mathcal{S}(\mathcal{M})$ and $\mathcal{T}(\mathcal{M})$ and their facial structures in terms of the ROG property. These results are particularly instrumental when we analyze the spectrahedral sets defined by finitely many LMIs/LMEs.

Lemma 21. *For any set $\mathcal{M} \subseteq \mathbb{S}^n$, the following are equivalent:*

1. *$\mathcal{S}(\mathcal{M})$ is ROG.*
2. *Every face of $\mathcal{S}(\mathcal{M})$ is ROG.*
3. *$\mathcal{S}(\mathcal{M}) \cap \mathcal{T}(\mathcal{M}')$ is ROG for every $\mathcal{M}' \subseteq \mathcal{M}$.*

Proof. (1. ⇒ 2.) Note that every extreme ray of a face of $\mathcal{S}(\mathcal{M})$ is also an extreme ray of $\mathcal{S}(\mathcal{M})$.

(2. ⇒ 3.) First, suppose $\mathcal{M}' = \varnothing$. Then, $\mathcal{T}(\mathcal{M}') = \mathbb{S}^n_+$ and thus $\mathcal{S}(\mathcal{M}) \cap \mathcal{T}(\mathcal{M}') = \mathcal{S}(\mathcal{M})$. Since $\mathcal{S}(\mathcal{M})$ is a face of itself, by part 2. we deduce it is ROG. Now consider any $\varnothing \neq \mathcal{M}' \subseteq \mathcal{M}$. Note that $\mathcal{T}(\mathcal{M}')$ only depends on the linear span of $\mathcal{M}'$, thus without loss of generality we may assume that $\mathcal{M}'$ is a basis of $\operatorname{span}(\mathcal{M}')$. Take Y to be the average of $\mathcal{M}'$, i.e., $Y = \frac{1}{|\mathcal{M}'|}\sum_{M\in\mathcal{M}'} M$. Note that $Y \in \operatorname{cone}(\mathcal{M}')$ so that $Y \in \mathcal{S}(\mathcal{M})^*$. We claim that $\mathcal{S}(\mathcal{M}) \cap \mathcal{T}(\mathcal{M}') = \mathcal{S}(\mathcal{M}) \cap Y^\perp$. Indeed, for all $X \in \mathcal{S}(\mathcal{M})$, we have that $\langle Y, X\rangle = 0$ if and only if $\langle Y, M\rangle = 0$ for all $M \in \mathcal{M}'$ if and only if $X \in \mathcal{T}(\mathcal{M}')$. We deduce that $\mathcal{S}(\mathcal{M}) \cap \mathcal{T}(\mathcal{M}')$ is a face of $\mathcal{S}(\mathcal{M})$, and thus it is ROG.

(3. ⇒ 1.) Take $\mathcal{M}' = \varnothing$. ■

We have the following immediate corollary of Lemma 21.

Corollary 13. *For any set $\mathcal{M} \subseteq \mathbb{S}^n$, if $\mathcal{S}(\mathcal{M})$ is ROG then $\mathcal{T}(\mathcal{M})$ is ROG.*

Proof. Take $\mathcal{M}' = \mathcal{M}$ in Lemma 21. ■

Informally, an extreme ray of $\mathcal{S}(\mathcal{M})$ should also be an extreme ray of $\mathcal{S}(\mathcal{M}')$ for $\mathcal{M}' \subseteq \mathcal{M}$ as long as $\mathcal{M}'$ contains the "relevant" inequalities in $\mathcal{M}$. The following technical lemma makes this notion precise.

Lemma 22. *Let $\mathcal{M} \subseteq \mathbb{S}^n$ and let $\mathbb{R}_+X$ be an extreme ray of $\mathcal{S}(\mathcal{M})$. Let $\mathcal{M}' \subseteq \mathcal{M}$ contain all of the constraints that are tight at X, i.e., $\{M \in \mathcal{M} : \langle M, X\rangle = 0\} \subseteq \mathcal{M}'$. If $\mathcal{M} \setminus \mathcal{M}'$ is compact, then $\mathbb{R}_+X$ is an extreme ray of $\mathcal{S}(\mathcal{M}')$. If additionally $\mathcal{M}' = \{M \in \mathcal{M} : \langle M, X\rangle = 0\}$, then $\mathbb{R}_+X$ is an extreme ray of $\mathcal{T}(\mathcal{M}')$.*

Proof. Suppose $Y \in \mathbb{S}^n$ is such that $[X - Y, X + Y] \subseteq \mathcal{S}(\mathcal{M}')$. By compactness of $\mathcal{M} \setminus \mathcal{M}'$, we have that $\langle M, X\rangle$ achieves a positive minimum value on $\mathcal{M} \setminus \mathcal{M}'$. Furthermore, by compactness, $\langle M, Y\rangle$ is bounded on $\mathcal{M} \setminus \mathcal{M}'$. In particular, there exists $\epsilon > 0$ small enough guaranteeing that $\langle M, X \pm \epsilon Y\rangle > 0$ for all $M \in \mathcal{M} \setminus \mathcal{M}'$. This together with $[X - Y, X + Y] \subseteq \mathcal{S}(\mathcal{M}')$ implies that $[X - \epsilon Y, X + \epsilon Y] \subseteq \mathcal{S}(\mathcal{M})$. Thus, as $\mathbb{R}_+X$ is an extreme ray of $\mathcal{S}(\mathcal{M})$ we conclude that $Y = \alpha X$ for some $\alpha \in \mathbb{R}$. This then implies that $\mathbb{R}_+X$ is extreme in $\mathcal{S}(\mathcal{M}')$.

The second statement follows by replacing $\mathcal{S}(\mathcal{M}')$ with $\mathcal{T}(\mathcal{M}')$ in the argument above. ■

Lemma 22 allows us to strengthen Lemma 21 in a few ways.

Lemma 23. *Let $\mathcal{M} \subseteq \mathbb{S}^n$ be compact. Then, $\mathcal{S}(\mathcal{M})$ is ROG if and only if $\mathcal{S}(\mathcal{M}) \cap \mathcal{T}(\mathcal{M}')$ is ROG for every $\varnothing \neq \mathcal{M}' \subseteq \mathcal{M}$.*

Proof. (⇒) This direction follows Lemma 21.

(⇐) Let $\mathbb{R}_+X$ be an extreme ray of $\mathcal{S}(\mathcal{M})$ and define $\mathcal{M}' := \{M \in \mathcal{M} : \langle M, X\rangle = 0\}$. First suppose $\mathcal{M}' \neq \varnothing$. As $\mathbb{R}_+X$ is also an extreme ray of $\mathcal{S}(\mathcal{M}) \cap \mathcal{T}(\mathcal{M}')$, which by assumption is ROG, we have that $\operatorname{rank}(X) = 1$. Now suppose $\mathcal{M}' = \varnothing$. By Lemma 22 and the assumption that $\mathcal{M}$ is compact, we deduce that $\mathbb{R}_+X$ is an extreme ray of $\mathcal{T}(\varnothing) = \mathbb{S}^n_+$. We conclude that $\operatorname{rank}(X) = 1$. ■

We note that given Lemma 23, it may be tempting to try to strengthen the third condition in Lemma 21 to the condition that $\mathcal{S}(\mathcal{M}) \cap \mathcal{T}(\mathcal{M}')$ is ROG for every $\varnothing \neq \mathcal{M}' \subseteq \mathcal{M}$. The following example shows that this is not possible without making the compactness assumption of Lemma 23.

Example 7. Suppose $n = 2$ and $\mathcal{M} = \bigcup_{i\in[4]} \mathcal{M}_i$, where

$$\mathcal{M}_1 = \left\{ \begin{pmatrix} 1 & \\ & -1+\epsilon \end{pmatrix} : \epsilon > 0 \right\}, \quad \mathcal{M}_2 = \left\{ \begin{pmatrix} -1 & \\ & 1+\epsilon \end{pmatrix} : \epsilon > 0 \right\},$$
$$\mathcal{M}_3 = \left\{ \begin{pmatrix} 0 & 1 \\ 1 & \epsilon \end{pmatrix} : \epsilon > 0 \right\}, \qquad \mathcal{M}_4 = \left\{ \begin{pmatrix} 0 & -1 \\ -1 & \epsilon \end{pmatrix} : \epsilon > 0 \right\}.$$

Noting that $\mathcal{S}(\mathcal{M})$ is unchanged upon taking the closure of $\mathcal{M}$ and that for all $i \in [4]$ and the constraints $\langle M_\epsilon, X\rangle \geq 0$ for $M_\epsilon \in \mathcal{M}_i$ get only more restrictive as $\epsilon \to 0$, we deduce

$$\mathcal{S}(\mathcal{M}) = \mathcal{S}\left(\left\{ \begin{pmatrix} 1 & \\ & -1 \end{pmatrix}, \begin{pmatrix} -1 & \\ & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} \right\} \right) = \mathbb{R}_+ I.$$

We conclude $\mathcal{S}(\mathcal{M}) = \mathbb{R}_+ I$ is not ROG. On the other hand, for any $\varnothing \neq \mathcal{M}' \subseteq \mathcal{M}$, we have $\mathcal{S}(\mathcal{M}) \cap \mathcal{T}(\mathcal{M}') = \{0\}$ (because $\langle M, I\rangle \neq 0$ for any $M \in \mathcal{M}$) and is ROG. □

Lemma 24. *Let $\mathcal{M} \subseteq \mathbb{S}^n$ be finite. If $\mathcal{T}(\mathcal{M}')$ is ROG for every $\mathcal{M}' \subseteq \mathcal{M}$, then $\mathcal{S}(\mathcal{M})$ is ROG.*

Proof. Let $\mathbb{R}_+ X$ be an extreme ray of $\mathcal{S}(\mathcal{M})$. Define $\mathcal{M}' := \{M \in \mathcal{M} : \langle M, X\rangle = 0\}$. By Lemma 22 and the fact that any finite set is compact, we deduce that $\mathbb{R}_+ X$ is an extreme ray of $\mathcal{T}(\mathcal{M}')$. We conclude that $\operatorname{rank}(X) = 1$. ■

The following lemma shows that the ROG property of $\mathcal{T}(\mathcal{M})$ is equivalent to the ROG property of $\mathcal{T}(\overline{\mathcal{M}})$ where $\overline{\mathcal{M}}$ is the restriction of $\mathcal{M}$ onto the joint range of the matrices $M \in \mathcal{M}$.

Lemma 25. *Let $W := \operatorname{span}(\bigcup_{M\in\mathcal{M}} \operatorname{range}(M))$. For $M \in \mathcal{M}$, let $\overline{M} = M_W$ denote the restriction of M to W. Let $\overline{\mathcal{M}} = \left\{ \overline{M} : M \in \mathcal{M} \right\}$. Then, $\mathcal{T}(\mathcal{M})$ is ROG if and only if $\mathcal{T}(\overline{\mathcal{M}})$ is ROG.*

Proof. ($\Rightarrow$) Note that $\mathcal{T}(\overline{\mathcal{M}})$ is isomorphic to $\mathcal{T}(\overline{\mathcal{M}}) \oplus 0_{W^\perp}$ via the rank-preserving map $X_W \mapsto X_W \oplus 0_{W^\perp}$. We claim that $\mathcal{T}(\overline{\mathcal{M}}) \oplus 0_{W^\perp}$ is a face of $\mathcal{T}(\mathcal{M})$. Indeed, we can write

$$\mathcal{T}(\overline{\mathcal{M}}) \oplus 0_{W^\perp} = \mathcal{T}(\mathcal{M}) \cap \{X \in \mathbb{S}^n_+ : \langle 0_W \oplus I_{W^\perp}, X\rangle = 0\}$$

and note that $0_W \oplus I_{W^\perp} \in \mathbb{S}^n_+$. Then, $\mathcal{T}(\overline{\mathcal{M}}) \oplus 0_{W^\perp}$ is ROG by Lemma 21. We conclude that $\mathcal{T}(\overline{\mathcal{M}})$ is ROG.

($\Leftarrow$) Let $\mathbb{R}_+(X)$ be an extreme ray of $\mathcal{T}(\mathcal{M})$ and set $\overline{X} := X_W$. We will show that $\operatorname{rank}(X) = 1$ by considering two cases. First, suppose $\overline{X} = 0$, then $\operatorname{range}(X) \subseteq W^\perp$. We deduce that as $X \neq 0$, there exists a nonzero vector $y \in \operatorname{range}(X) \subseteq W^\perp$. Note that $\langle M, yy^\intercal\rangle = \langle M_W, (yy^\intercal)_W\rangle = 0$. Furthermore, $X \pm \epsilon yy^\intercal \in \mathbb{S}^n_+$ for all small enough $\epsilon > 0$. By the

$$\boxed{\mathcal{M} \text{ is finite and } \forall \mathcal{M}' \subseteq \mathcal{M},\ \mathcal{T}(\mathcal{M}') \text{ ROG}} \Longrightarrow \boxed{\mathcal{S}(\mathcal{M}) \text{ ROG}} \Longrightarrow \boxed{\mathcal{T}(\mathcal{M}) \text{ ROG}}$$

Figure 3.1: A summary of Lemma 24 and Corollary 13

assumption that $\mathbb{R}_+(X)$ is an extreme ray, we then conclude that X is a scalar multiple of $yy^\intercal$ and is rank-one.

Next, suppose $\overline{X} \neq 0$. As $\langle M, X\rangle = \left\langle \overline{M}, \overline{X}\right\rangle$ for every $M \in \mathcal{M}$, we have that $\overline{X} \in \mathcal{T}(\overline{\mathcal{M}})$. By the assumption that $\mathcal{T}(\overline{\mathcal{M}})$ is ROG, we may write $\overline{X} = \sum_{i=1}^k \overline{y}_i\overline{y}_i^\intercal$ where $\overline{y}_i\overline{y}_i^\intercal \in \mathcal{T}(\overline{\mathcal{M}})$ are each nonzero. Fix $\overline{y} := \overline{y}_1$ and define $\overline{z}$ such that $\overline{y} = \overline{X}\overline{z}$. This is possible as $\overline{y} \in \operatorname{range}(\overline{X})$. Finally, define

$$y := X(\overline{z} \oplus 0_{W^\perp}).$$

We claim that $X \pm \epsilon yy^\intercal \in \mathcal{T}(\mathcal{M})$ for all $\epsilon > 0$ small enough. Indeed, as $y \in \operatorname{range}(X)$ we have that $X \pm \epsilon yy^\intercal \in \mathbb{S}^n_+$ for all $\epsilon > 0$ small enough. Furthermore, for all $M \in \mathcal{M}$ we have

$$\langle M, yy^\intercal\rangle = \left\langle \overline{M}, \overline{y}\overline{y}^\intercal \right\rangle = 0,$$

where the second equality follows from the fact that $\overline{y} \in \mathcal{T}(\overline{\mathcal{M}})$. Additionally note that $\overline{y}$ is nonzero and $y_W = \overline{y}$ so that y is nonzero. We deduce that $X \pm \epsilon yy^\intercal \in \mathcal{T}(\mathcal{M})$ for all $\epsilon > 0$ small enough. By the assumption that $\mathbb{R}_+(X)$ is an extreme ray, we then conclude that X is a scalar multiple of $yy^\intercal$ and is rank-one. ■

Remark 32. The characterizations given in Lemmas 21 to 25 and Corollary 13 are based on the facial structure of the sets $\mathcal{S}(\mathcal{M})$ and $\mathcal{T}(\mathcal{M})$ and in a sense are analogous to characterizations of integral polyhedra. □

Remark 33. The ROG property is not preserved under trivial liftings. When $\mathcal{M} = \{M_1, \dots, M_k\}$ is finite, one may attempt to replace all of the inequalities defining $\mathcal{S}(\mathcal{M})$ with equalities by adding new slack variables. Specifically, for $i \in [k]$, let $\overline{M}_i \in \mathbb{S}^{n+k}$ be the following block matrix

$$\overline{M}_i := \begin{pmatrix} M_i & \\ & e_ie_i^\intercal \end{pmatrix}$$

and let $\overline{\mathcal{M}} := \left\{\overline{M}_1, \dots, \overline{M}_k\right\}$. It is straightforward to show that the ROG property is preserved under the projection of $\mathbb{S}^{n+k}$ onto $\mathbb{S}^n$. Thus, if $\mathcal{T}\left(\overline{\mathcal{M}}\right)$ is ROG, then $\mathcal{S}(\mathcal{M})$ is also ROG. Unfortunately the reverse implication is not true in general. We will give a counterexample in Section 3.4.4 (see Example 9). □

3.2.3 Simple operations preserving ROG property

We now present a few lemmas that are useful in reasoning about extreme rays of $\mathcal{S}(\mathcal{M})$. The following lemma states that an extreme ray $\mathbb{R}_+X$ "only cares about" constraints "in the range of X."

Lemma 26. *Let $\mathcal{M} \subseteq \mathbb{S}^n$ and let $\mathbb{R}_+X$ be an extreme ray of $\mathcal{S}(\mathcal{M})$. Let $W := \operatorname{range}(X)$ and let $\mathcal{M}_W := \{M_W : M \in \mathcal{M}\}$. Then $\mathbb{R}_+(X_W)$ is an extreme ray of $\mathcal{S}(\mathcal{M}_W)$. In particular, if $\mathcal{S}(\mathcal{M}_W)$ is ROG, then $\operatorname{rank}(X) = \operatorname{rank}(X_W) = 1$.*

Proof. Suppose $Y_W \in \mathbb{S}^W$ is such that $[X_W - Y_W, X_W + Y_W] \subseteq \mathcal{S}(\mathcal{M}_W)$. Let $Y = 0_{W^\perp} \oplus Y_W$. Then, $X + Y = 0_{W^\perp} \oplus (X_W + Y_W)$, and for any $M \in \mathcal{M}$ we have $\langle M, X+Y\rangle = \langle M_W, X_W + Y_W\rangle \geq 0$. We deduce that $X + Y \in \mathcal{S}(\mathcal{M})$. Similarly $X - Y \in \mathcal{S}(\mathcal{M})$ whence $[X - Y, X + Y] \subseteq \mathcal{S}(\mathcal{M})$. As $\mathbb{R}_+X$ is extreme in $\mathcal{S}(\mathcal{M})$, we deduce that $Y = \alpha X$ for some $\alpha \in \mathbb{R}$. Consequently, $Y_W = \alpha X_W$ for some $\alpha \in \mathbb{R}$ and $\mathbb{R}_+(X_W)$ is extreme in $\mathcal{S}(\mathcal{M}_W)$. ■

The following lemma addresses the case when $\mathcal{M}$ can be partitioned into "non-interacting" sets of constraints.

Lemma 27. *Let $\mathcal{M} \subseteq \mathbb{S}^n$ be a finite union of compact sets $\mathcal{M} = \bigcup_{i=1}^k \mathcal{M}_i$. Further, suppose that for all nonzero $X \in \mathbb{S}^n_+$ and $i \in [k]$, if $\langle M_i, X\rangle = 0$ for some $M_i \in \mathcal{M}_i$, then $\langle M, X\rangle > 0$ for all $M \in \mathcal{M} \setminus \mathcal{M}_i$. Then, $\mathcal{S}(\mathcal{M})$ is ROG if and only if $\mathcal{S}(\mathcal{M}_i)$ is ROG for all $i \in [k]$.*

Proof. ($\Rightarrow$) Fix $i \in [k]$ and let $\mathbb{R}_+X$ be an extreme ray of $\mathcal{S}(\mathcal{M}_i)$. If $\langle M_i, X\rangle > 0$ for all $M_i \in \mathcal{M}_i$, then Lemma 22 implies that $\mathbb{R}_+X$ is an extreme ray of $\mathbb{S}^n_+$ and so $\operatorname{rank}(X) = 1$. Now suppose $\langle M_i, X\rangle = 0$ for some $M_i \in \mathcal{M}_i$. By assumption, $\langle M, X\rangle > 0$ for all $M \in \mathcal{M} \setminus \mathcal{M}_i$ so that $X \in \mathcal{S}(\mathcal{M})$. As $\mathcal{S}(\mathcal{M}) \subseteq \mathcal{S}(\mathcal{M}_i)$, we have that $\mathbb{R}_+X$ must also be an extreme ray of $\mathcal{S}(\mathcal{M})$. We deduce that $\operatorname{rank}(X) = 1$.

($\Leftarrow$) Let $\mathbb{R}_+X$ be an extreme ray of $\mathcal{S}(\mathcal{M})$. Define $\mathcal{M}' := \{M \in \mathcal{M} : \langle M, X\rangle = 0\}$. If $\mathcal{M}' = \varnothing$ then Lemma 22 implies that $\mathbb{R}_+X$ is an extreme ray of $\mathcal{T}(\varnothing) = \mathbb{S}^n_+$ and so $\operatorname{rank}(X) = 1$.

Now suppose $\mathcal{M}'$ is nonempty. Then, by assumption, $\mathcal{M}' \subseteq \mathcal{M}_i$ for some i. By Lemma 22 and the assumption that $\mathcal{M} \setminus \mathcal{M}_i$ is compact, we deduce that $\mathbb{R}_+X$ is an extreme ray of $\mathcal{S}(\mathcal{M}_i)$. We conclude that $\operatorname{rank}(X) = 1$. ■

Finally, the following lemma states that an arbitrary intersection of ROG cones is ROG if and only if no new extreme rays are introduced.

Lemma 28. *Let $\mathcal{M} \subseteq \mathbb{S}^n$ be a union $\mathcal{M} = \bigcup_{\alpha \in A} \mathcal{M}_\alpha$. Suppose that $\mathcal{S}(\mathcal{M}_\alpha)$ is ROG for every $\alpha \in A$. Then, $\mathcal{S}(\mathcal{M})$ is ROG if and only if*

$$\operatorname{extr}(\mathcal{S}(\mathcal{M})) \subseteq \bigcap_{\alpha \in A} \operatorname{extr}(\mathcal{S}(\mathcal{M}_\alpha)).$$

Proof. ($\Leftarrow$) Let $\mathbb{R}_+X$ be an extreme ray of $\mathcal{S}(\mathcal{M})$. Then, by assumption, $\mathbb{R}_+X$ is an extreme ray of $\mathcal{S}(\mathcal{M}_\alpha)$ for each $\alpha \in A$. By recalling that each $\mathcal{S}(\mathcal{M}_\alpha)$ is ROG, we deduce $\operatorname{rank}(X) = 1$.

($\Rightarrow$) Let $\mathbb{R}_+X$ be an extreme ray of $\mathcal{S}(\mathcal{M})$. Then, by the assumption that $\mathcal{S}(\mathcal{M})$ is ROG, we have $\operatorname{rank}(X) = 1$. Next, note that $X \in \mathcal{S}(\mathcal{M}) = \bigcap_{\alpha \in A} \mathcal{S}(\mathcal{M}_\alpha)$, whence $X \in \mathcal{S}(\mathcal{M}_\alpha)$ for all $\alpha \in A$. Then as $\operatorname{rank}(X) = 1$, we deduce that $\mathbb{R}_+ X$ is extreme in $\mathcal{S}(\mathcal{M}_\alpha)$ for all $\alpha \in A$ by Fact 9. ■

3.2.4 The ROG property and solutions of quadratic systems

We next examine the ROG property of a set and its connection to the existence of nonzero solutions of underlying quadratic systems of inequalities and/or equations.

Definition 15. Given $\mathcal{M} \subseteq \mathbb{S}^n$ and $X \in \mathcal{S}(\mathcal{M})$, we define

$$\mathcal{E}(X, \mathcal{M}) := \{x \in \mathbb{R}^n : |x^\intercal M x| \leq \langle M, X\rangle, \forall M \in \mathcal{M}\}. \qquad \square$$

Lemma 29. *$\mathcal{S}(\mathcal{M})$ is ROG if and only if for every nonzero $X \in \mathcal{S}(\mathcal{M})$ we have* $\operatorname{range}(X) \cap \mathcal{E}(X, \mathcal{M}) \neq \{0\}$.

Proof. ($\Rightarrow$) Suppose $X \in \mathcal{S}(\mathcal{M})$ is nonzero. Because $\mathcal{S}(\mathcal{M})$ is ROG, we can write $X = \sum_{i=1}^k x_i x_i^\intercal$ using nonzero matrices $x_i x_i^\intercal \in \mathcal{S}(\mathcal{M})$. As X is a nonzero matrix, we have $k \geq 1$ and thus $\bar{x} := x_1$ exists. Then, for every $M \in \mathcal{M}$ and $i \in [k]$, we have $x_i^\intercal M x_i \geq 0$. In particular, $0 \leq \bar{x}^\intercal M \bar{x} \leq \sum_{i=1}^k x_i^\intercal M x_i = \langle M, X\rangle$. Furthermore, $\bar{x} \in \operatorname{range}(X)$. We conclude that $\operatorname{range}(X) \cap \mathcal{E}(X, \mathcal{M})$ contains the nonzero element $\bar{x}$.

($\Leftarrow$) Let $\mathbb{R}_+ X$ be an extreme ray of $\mathcal{S}(\mathcal{M})$. By assumption, there exists a nonzero $x \in \operatorname{range}(X)$ such that

$$|x^\intercal M x| \leq \langle M, X\rangle, \forall M \in \mathcal{M}.$$

By picking $\epsilon > 0$ small enough, we can simultaneously ensure that $X \pm \epsilon x x^\intercal \in \mathbb{S}^n_+$ and that

$$\langle M, X \pm \epsilon x x^\intercal\rangle \geq (1 - \epsilon)\langle M, X\rangle \geq 0, \forall M \in \mathcal{M}.$$

Hence, we conclude that the interval $[X - \epsilon x x^\intercal, X + \epsilon x x^\intercal]$ is contained in $\mathcal{S}(\mathcal{M})$. In particular, because $\mathbb{R}_+ X$ is an extreme ray of $\mathcal{S}(\mathcal{M})$, we deduce that $\epsilon x x^\intercal$ is a scalar multiple of X and hence $\operatorname{rank}(X) = 1$. ■

When studying $\mathcal{T}(\mathcal{M})$, we can replace the set $\mathcal{E}(X, \mathcal{M})$ in Lemma 29 with a simpler set corresponding to solutions to a homogeneous system of quadratic equations.[7]

Definition 16. Given $\mathcal{M} \subseteq \mathbb{S}^n$, we define

$$\mathcal{N}(\mathcal{M}) := \{x \in \mathbb{R}^n : x^\intercal M x = 0, \forall M \in \mathcal{M}\}. \qquad \square$$

Remark 34. Note that for every $\mathcal{M} \subseteq \mathbb{S}^n$ and every $X \in \mathcal{S}(\mathcal{M})$, we have $\mathcal{N}(\mathcal{M}) \subseteq \mathcal{E}(X, \mathcal{M})$. □

Corollary 14. *$\mathcal{T}(\mathcal{M})$ is ROG if and only if for every nonzero $X \in \mathcal{T}(\mathcal{M})$ we have* $\operatorname{range}(X) \cap \mathcal{N}(\mathcal{M}) \neq \{0\}$.

Proof. Note that $\mathcal{S}(-\mathcal{M} \cup \mathcal{M}) = \mathcal{T}(\mathcal{M})$ and apply Lemma 29. ■

[7] Readers familiar with algebraic geometry will recognize this as the variety defined by $\mathcal{M}$.

$$\begin{tikzcd}
\boxed{\mathcal{S}(\mathcal{M})\text{ ROG}} \arrow[r, Leftrightarrow] \arrow[d, Rightarrow] & \boxed{\forall X \in \mathcal{S}(\mathcal{M}) \setminus \{0\},\ \operatorname{range}(X) \cap \mathcal{E}(X, \mathcal{M}) \neq \{0\}} \arrow[d, Rightarrow] \\
\boxed{\mathcal{T}(\mathcal{M})\text{ ROG}} \arrow[r, Leftrightarrow] & \boxed{\forall X \in \mathcal{T}(\mathcal{M}) \setminus \{0\},\ \operatorname{range}(X) \cap \mathcal{N}(\mathcal{M}) \neq \{0\}}
\end{tikzcd}$$

Figure 3.2: A summary of Lemma 29 and Corollary 14.

Remark 35. When applying Lemma 29, it suffices to check the right hand side only for matrices X with rank at least two. Indeed if $X = xx^\intercal$, then $x \in \operatorname{range}(X) \cap \mathcal{E}(X, \mathcal{M})$. The same is true for Corollary 14. □

3.2.5 Known ROG sets

In order to familiarize the reader with our notation and setup, we now recover three known results in our language. We begin with a result due to Sturm and Zhang [167] regarding spectrahedral cones defined by a single LMI.

Lemma 30. *Consider any $M \in \mathbb{S}^n$, and let $\mathcal{M} = \{M\}$. Then $\mathcal{S}(\mathcal{M})$ is ROG.*

Proof. By Lemma 23, $\mathcal{S}(\mathcal{M})$ is ROG if and only if $\mathcal{T}(\mathcal{M})$ is ROG. We will show that $\mathcal{T}(\mathcal{M})$ is ROG by appealing to Corollary 14.

Let $X \in \mathcal{T}(\mathcal{M})$ have rank at least two. Begin by performing a spectral decomposition $X = \sum_{i=1}^r \lambda_i x_i x_i^\intercal$, where $r = \operatorname{rank}(X) \geq 2$, the x_i are orthonormal eigenvectors of X, and $\lambda_i > 0$ for all $i \in [r]$.

If one of the eigenvectors x_i is in $\mathcal{N}(\mathcal{M})$, then $\operatorname{range}(X) \cap \mathcal{N}(\mathcal{M})$ contains x_i and is clearly nontrivial.

Else, there exist distinct eigenvectors, without loss of generality x_1 and x_2, such that $\langle M, x_1 x_1^\intercal\rangle > 0 > \langle M, x_2 x_2^\intercal\rangle$. By continuity, there exists $x \in [x_1, x_2]$ such that $\langle M, xx^\intercal\rangle = 0$. Note that x is nonzero as $0 \notin [x_1, x_2]$ (this follows as x_1 and x_2 are orthonormal). Furthermore, $x \in \operatorname{range}(X)$. This concludes the proof as we have constructed a nonzero $x \in \operatorname{range}(X) \cap \mathcal{N}(\mathcal{M})$. ■

Based on Lemmas 24 and 30 and Corollary 13, we have the following characterization of ROG sets defined by two inequalities.

Corollary 15. *Suppose $|\mathcal{M}| = 2$, then $\mathcal{S}(\mathcal{M})$ is ROG if and only if $\mathcal{T}(\mathcal{M})$ is ROG.*

The characterization given in Corollary 15 for the case of $|\mathcal{M}| = 2$ is, at the moment, unsatisfactory as we have yet to analyze when $\mathcal{T}(\mathcal{M})$ is itself ROG. Our developments in the remainder of this chapter will make this implicit characterization much more explicit (see Section 3.4).

Next, we recover a result related to the S-lemma [67] and a convexity theorem due to Dines [59].

Lemma 31. *Let $\mathcal{M} = \{M_1, M_2\}$ and suppose there exists $(\alpha_1, \alpha_2) \neq (0, 0)$ such that $\alpha_1 M_1 + \alpha_2 M_2 \in \mathbb{S}^n_+$. Then, $\mathcal{S}(\mathcal{M})$ is ROG.*

Proof. By Corollary 15, it suffices to show that $\mathcal{T}(\mathcal{M})$ is ROG. Recall also that $\mathcal{T}(\mathcal{M})$ depends only on $\text{span}(\mathcal{M})$ (see Remark 31), thus we may without loss of generality suppose $M_1 \in \mathbb{S}^n_+$.

Let $W := \text{range}(M_1)$. We claim that $X_W = 0$ for all $X \in \mathcal{T}(\mathcal{M})$. Indeed, suppose $X \in \mathcal{T}(\mathcal{M})$ so that $\langle M_1, X\rangle = 0$. Noting that both $M_1, X \in \mathbb{S}^n_+$, we deduce that $M_1X = 0$ so that $X_W = 0$. Then, applying Lemma 20 allows us to write $X = 0_W \oplus X_{W^\perp}$.

Let $\overline{M}_2 := (M_2)_{W^\perp}$. Then,

$$\mathcal{T}(\mathcal{M}) = \left\{ 0_W \oplus X_{W^\perp} : \begin{array}{l} \left\langle \overline{M}_2, X_{W^\perp} \right\rangle = 0 \\ X_{W^\perp} \in \mathbb{S}^{W^\perp}_+ \end{array} \right\} = 0_W \oplus \mathcal{T}(\overline{M}_2). \tag{3.3}$$

By Lemma 30 and Corollary 13, $\mathcal{T}(\overline{M}_2)$ is ROG. Then as $\mathcal{T}(\mathcal{M})$ is isomorphic to $\mathcal{T}(\overline{M}_2)$ via the rank-preserving map $0_W \oplus X_{W^\perp} \mapsto X_{W^\perp}$, we conclude that $\mathcal{T}(\mathcal{M})$ is ROG. ■

Remark 36. The condition that there exists $(\alpha_1, \alpha_2) \neq (0,0)$ such that $\alpha_1 M_1 + \alpha_2 M_2 \in \mathbb{S}^n_+$ has a simple geometric interpretation. Specifically, this condition guarantees that the two LMEs defining $\mathcal{T}(\{M_1, M_2\})$ only interact with each other on a single (possibly trivial) face of the positive semidefinite cone. Furthermore, on this face, the two LMEs impose the same (possibly trivial) constraint. □

3.3 Sufficient conditions

The following observation generalizes the key step in Lemma 31.

Observation 3. *Let $\mathcal{M} \subseteq \mathbb{S}^n$. Suppose there exists a nonzero $M \in \text{span}(\mathcal{M}) \cap \mathbb{S}^n_+$. Let $W := \text{range}(M)$ and define $\mathcal{M}_{W^\perp} := \{M_{W^\perp} : M \in \mathcal{M}\}$. Then,*

$$\mathcal{T}(\mathcal{M}) = 0_W \oplus \mathcal{T}(\mathcal{M}_{W^\perp}).$$

In particular, $\mathcal{T}(\mathcal{M})$ is isomorphic to $\mathcal{T}(\mathcal{M}_{W^\perp})$ via the rank-preserving map $0_W \oplus Y \mapsto Y$ and $\mathcal{T}(\mathcal{M})$ is ROG if and only if $\mathcal{T}(\mathcal{M}_{W^\perp})$ is ROG.

Remark 37. Observation 3 simply notes that $\mathcal{T}(\mathcal{M})$ is a subset of the face $0_W \oplus \mathbb{S}^{W^\perp}_+$ of the positive semidefinite cone and then applies Lemma 20. This idea is linked to facial reduction [30, 110, 144], a technique which has been used previously in the literature to simplify semidefinite programs and more general conic programs. □

Applying Observation 3 repeatedly gives the following generalization of Lemma 31 as a sufficient condition for the ROG property.

Proposition 11. *Let $\mathcal{M} = \{M_1, \dots, M_k\}$ for some $k \geq 2$. Suppose for all distinct indices $i, j \in [k]$, there exists $(\alpha, \beta) \neq (0,0)$ such that $\alpha M_i + \beta M_j$ is positive semidefinite. Then, $\mathcal{S}(\mathcal{M})$ is ROG.*

Proof. By Lemmas 24 and 30, it suffices to show that $\mathcal{T}(\mathcal{M}')$ is ROG for every $\mathcal{M}' \subseteq \mathcal{M}$ with size at least two.

Let $\mathcal{M}' \subseteq \mathcal{M}$. Consider repeatedly applying Observation 3 to get a chain of subspaces $W_1 \subset W_2 \subset \cdots \subset W$ such that

$$\mathcal{T}(\mathcal{M}') = 0_{W_1} \oplus \mathcal{T}(\mathcal{M}'_{W_1^\perp}) = 0_{W_2} \oplus \mathcal{T}(\mathcal{M}'_{W_2^\perp}) = \cdots = 0_W \oplus \mathcal{T}(\mathcal{M}'_{W^\perp}).$$

We will repeat this process until $\operatorname{span}(\mathcal{M}'_{W^\perp}) \cap \mathbb{S}_+^{W^\perp} = \{0\}$. This process necessarily terminates as the subspaces W_i strictly increase in dimension. Let $\overline{M}_i \coloneqq (M_i)_{W^\perp}$ and $\overline{\mathcal{M}'} \coloneqq \left\{\overline{M}_i :\ M_i \in \mathcal{M}'\right\}$.

We claim that $\dim(\operatorname{span}(\overline{\mathcal{M}'})) \leq 1$. Suppose otherwise and let $M_i, M_j \in \mathcal{M}'$ such that $\overline{M}_i$ and $\overline{M}_j$ are independent. By assumption, there exists $(\alpha, \beta) \neq (0,0)$ such that $\alpha M_i + \beta M_j$ is positive semidefinite. Then,

$$\alpha \overline{M}_i + \beta \overline{M}_j = (\alpha M_i + \beta M_j)_{W^\perp}$$

is positive semidefinite. Furthermore, this linear combination is nonzero by independence of $\overline{M}_i$ and $\overline{M}_j$. This contradicts the assumption that $\operatorname{span}(\overline{\mathcal{M}'}) \cap \mathbb{S}_+^{W^\perp} = \{0\}$.

Note that $\mathcal{T}(\mathcal{M}')$ is isomorphic to $\mathcal{T}(\overline{\mathcal{M}'})$ via the rank-preserving map $0_W \oplus X_{W^\perp} \mapsto X_{W^\perp}$. Furthermore, by Remark 31 and Lemma 30, we have that $\mathcal{T}(\overline{\mathcal{M}'})$ is ROG. We conclude that $\mathcal{T}(\mathcal{M})$ is ROG. ■

Intuitively, the conditions in this proposition have a similar geometric interpretation to the conditions in Lemma 31 (see Remark 36). Specifically, the proof shows that for any $\mathcal{M}' \subseteq \mathcal{M}$ of size at least two, there exists a subspace $W \subseteq \mathbb{R}^n$ such that $\mathcal{T}(\mathcal{M}')$ is contained in the face $0_W \oplus \mathbb{S}_+^{W^\perp}$ of the positive semidefinite cone. Furthermore, on this face, the LMEs in $\mathcal{M}'$ all impose the same constraint.

Next, we present a new sufficient condition for the ROG property suggested by Lemma 29 and Remark 34.

Theorem 14. *Suppose $\mathcal{M} = \{\operatorname{Sym}(ab^\intercal) : b \in \mathcal{B}\}$ for some $a \in \mathbb{R}^n$ and $\mathcal{B} \subseteq \mathbb{R}^n$. Then, for every positive semidefinite X of rank at least two, we have $\operatorname{range}(X) \cap \mathcal{N}(\mathcal{M}) \neq \{0\}$. In particular, $\mathcal{S}(\mathcal{M})$ is ROG.*

Proof. For any $v \in a^\perp$, we have $v^\intercal \operatorname{Sym}(ab^\intercal) v = v^\intercal ab^\intercal v = 0$. We deduce that $a^\perp \subseteq \mathcal{N}(\mathcal{M})$, i.e., $\mathcal{N}(\mathcal{M})$ contains a vector space of codimension one.

Let X be a positive semidefinite matrix with rank at least two. As $\dim(\operatorname{range}(X)) = \operatorname{rank}(X)$, we see that $\operatorname{range}(X) \cap \mathcal{N}(\mathcal{M})$ must contain a vector space of dimension at least one. In particular, $\operatorname{range}(X) \cap \mathcal{E}(X, \mathcal{M}) \supseteq \operatorname{range}(X) \cap \mathcal{N}(\mathcal{M})$ and is nonempty. Lemma 29 then implies that $\mathcal{S}(\mathcal{M})$ is ROG. ■

We list two immediate corollaries of Theorem 14.

Corollary 16. *Let $K \subseteq \mathbb{R}^n$ be any closed convex cone and consider an arbitrary vector $c \in \mathbb{R}^n$. Then, the set $\left\{X \in \mathbb{S}_+^n :\ Xc \in K\right\}$ is ROG.*

Proof. Define $\mathcal{M} \coloneqq \{\operatorname{Sym}(cb^\intercal) :\ b \in K^*\}$ where K^* is the dual cone of K. Then $\left\{X \in \mathbb{S}_+^n :\ Xc \in K\right\} = \mathcal{S}(\mathcal{M})$, whence Theorem 14 implies the result. ■

Corollary 17. *Let $a, b, c \in \mathbb{R}^n$. Then the set $\{X \in \mathbb{S}^n_+ : a^\intercal Xc \geq 0,\ b^\intercal Xc \geq 0\}$ is ROG.*

By applying Lemma 24 once more, we next give a sufficient condition which is not covered by Theorem 14.

Theorem 15. *Let $a, b, c \in \mathbb{R}^n$. Then the set $\{X \in \mathbb{S}^n_+ : a^\intercal Xb \geq 0,\ b^\intercal Xc \geq 0,\ a^\intercal Xc \geq 0\}$ is ROG.*

Proof. Let $\mathcal{M} = \{\operatorname{Sym}(ab^\intercal), \operatorname{Sym}(ac^\intercal), \operatorname{Sym}(bc^\intercal)\}$. By Lemma 24 and Corollary 17, it suffices to show that $\mathcal{T}(\mathcal{M})$ is ROG.

We will show that $\mathcal{T}(\mathcal{M})$ is ROG by appealing to Corollary 14. Let $X \in \mathcal{T}(\mathcal{M})$ have rank at least two.

Note that $\mathcal{N}(\operatorname{Sym}(ab^\intercal)) = a^\perp \cup b^\perp$. Hence,

$$\mathcal{N}(\mathcal{M}) = \left(a^\perp \cup b^\perp\right) \cap \left(a^\perp \cup c^\perp\right) \cap \left(b^\perp \cup c^\perp\right) = \{a,b\}^\perp \cup \{a,c\}^\perp \cup \{b,c\}^\perp.$$

If $Xa = Xb = Xc = 0$, then $\operatorname{range}(X) \subseteq \{a,b,c\}^\perp$ and thus $\operatorname{range}(X) \cap \mathcal{N}(\mathcal{M}) = \operatorname{range}(X)$ is clearly nontrivial. Else, without loss of generality suppose $y = Xa \neq 0$. Because $X \in \mathcal{T}(\mathcal{M})$, we have $b^\intercal y = c^\intercal y = 0$, and thus $y \in \mathcal{N}(\mathcal{M})$. Noting that $y \neq 0$ and $y \in \operatorname{range}(X)$, we have concluded $0 \neq y \in \operatorname{range}(X) \cap \mathcal{N}(\mathcal{M})$ as desired. ■

Remark 38. By picking $n = 3$ and $\{a, b, c\} = \{e_1, e_2, e_3\}$ in Theorem 15, we recover the well-known fact that the set of doubly nonnegative matrices (i.e., the set of matrices which are both entry-wise nonnegative and positive semidefinite) in $\mathbb{S}^3$ is ROG. In particular, this states that $X \in \mathbb{S}^3$ is doubly nonnegative if and only if it can be written as $X = \sum_i x_i x_i^\intercal$ where $x_i \in \mathbb{R}^3$ are each entry-wise nonnegative. In other words, the set of doubly nonnegative matrices and the set of completely positive matrices in $\mathbb{S}^3$ coincide. □

Remark 39. A graph $G = (V, E)$ is *chordal* if every minimal cycle has at most 3 edges. It is well-known that the set of positive semidefinite matrices with a fixed chordal support is ROG [3, 78, 145]. Specifically, if $G = ([n], E)$ is a chordal graph containing all self-loops, then

$$\{X \in \mathbb{S}^n_+ : X_{i,j} = 0,\ \forall (i,j) \notin E\} \tag{3.4}$$

is ROG.

Unfortunately, the set in (3.4) does not necessarily remain ROG when the equality constraints are replaced with inequality constraints. Using our toolset, we illustrate this point below with an example. From this point of view, Theorem 15 and Remark 38 highlight a special chordal graph for which the inequality version of the set is also ROG.

Consider the path graph on four vertices with all self-loops. We will show that the following set is not ROG:

$$\mathcal{S} = \left\{ X \in \mathbb{S}^4_+ : \begin{array}{c} X_{1,2} \geq 0 \\ X_{2,3} \geq 0 \\ X_{3,4} \geq 0 \end{array} \right\}.$$

We will apply Lemma 29 to show that $\mathcal{S}$ is not ROG. Let $\mathcal{M} = \{\operatorname{Sym}(e_1e_2^\intercal), \operatorname{Sym}(e_2e_3^\intercal), \operatorname{Sym}(e_3e_4^\intercal)\}$ so that $\mathcal{S} = \mathcal{S}(\mathcal{M})$. Let $x = (1,\ 0,\ 1,\ 1)^\intercal$ and $y = (0,\ 1,\ 1,\ -1)^\intercal$. Note that the following rank-two matrix

$$X \coloneqq xx^\intercal + yy^\intercal = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & -1 \\ 1 & 1 & 2 & 0 \\ 1 & -1 & 0 & 2 \end{pmatrix}$$

satisfies $X \in \mathcal{S}$. We compute

$$\operatorname{range}(X) \cap \mathcal{E}(X, \mathcal{M}) = \operatorname{span}\{x, y\} \cap \left\{ z \in \mathbb{R}^4 : \begin{array}{c} z_1z_2 = 0 \\ |z_2z_3| \leq 1 \\ z_3z_4 = 0 \end{array} \right\}.$$

Let $z \in \operatorname{range}(X) \cap \mathcal{E}(X, \mathcal{M})$. Then, writing $z = \alpha x + \beta y = (\alpha,\ \beta,\ \alpha + \beta,\ \alpha - \beta)^\intercal$, we deduce that $0 = z_1z_2 = \alpha\beta$ and $0 = z_3z_4 = \alpha^2 - \beta^2$ so that $\alpha = \beta = 0$. Thus, $\operatorname{range}(X) \cap \mathcal{E}(X, \mathcal{M}) = \{0\}$. □

Finally, we show how our results can be used to recover a result due to Sturm and Zhang [167]; see also [33, Section 6.1]. Let $\text{Ł}^n \subseteq \mathbb{R}^n$ denote the second order cone (SOC)

$$\text{Ł}^n \coloneqq \left\{ x = (y, t) \in \mathbb{R}^{n-1} \times \mathbb{R} :\ \|y\|_2 \leq t \right\}.$$

Defining $L \coloneqq \operatorname{Diag}(-1, \dots, -1, 1) \in \mathbb{S}^n$, we can write $\text{Ł}^n = \{x \in \mathbb{R}^n :\ x^\intercal Lx \geq 0,\ x_n \geq 0\}$.

Lemma 32. *Let $c \in \mathbb{R}^n$ and define*

$$\mathcal{S} \coloneqq \left\{ X \in \mathbb{S}^n_+ : \begin{array}{l} Xc \in \text{Ł}^n \\ \langle L, X \rangle \geq 0 \end{array} \right\}.$$

Then, $\mathcal{S}$ is ROG.

Proof. We begin by rewriting $\mathcal{S}$ so that we may apply Lemma 22. Let $\mathcal{B}$ denote a compact base of $\text{Ł}^n = (\text{Ł}^n)^*$. Then,

$$\mathcal{S} = \mathcal{S}(\{L\} \cup \{\operatorname{Sym}(cb^\intercal) :\ b \in \mathcal{B}\}).$$

For the sake of contradiction suppose there exists an extreme ray $\mathbb{R}_+X$ of $\mathcal{S}$ with $\operatorname{rank}(X) \geq 2$.

If $\langle L, X \rangle > 0$ then $\mathbb{R}_+X$ is an extreme ray of $\mathcal{S}(\{\operatorname{Sym}(cb^\intercal) :\ b \in \mathcal{B}\}) = \{X \in \mathbb{S}^n_+ :\ Xc \in \text{Ł}^n\}$, contradicting Corollary 16. If $Xc \in \operatorname{int}(\text{Ł}^n)$ then $\mathbb{R}_+X$ is an extreme ray of $\mathcal{S}(\{L\}) = \{X \in \mathbb{S}^n_+ : \langle X, L \rangle \geq 0\}$, contradicting Lemma 30. Finally, suppose $Xc = 0$ and let $W = \operatorname{range}(X) \subseteq c^\perp$. Note that X_W and X have the same rank and $\operatorname{Sym}(cb^\intercal)_W = 0$ for all $b \in \mathcal{B}$. Then, by Lemma 26, we have that $\mathbb{R}_+(X_W)$ is an extreme ray of $\mathcal{S}(\{L_W\})$, contradicting Lemma 30.

In the remainder of the proof, we will assume that $\langle L, X \rangle = 0$ and $y \coloneqq Xc$ is a nonzero element in $\operatorname{bd}(\text{Ł}^n)$, i.e., $y^\intercal Ly = 0$.

Then, for all $\epsilon > 0$ small enough, we have $X \pm \epsilon yy^\intercal \succeq 0$, $\langle L, X \pm \epsilon yy^\intercal \rangle = \langle L, X \rangle = 0$, and $(X \pm \epsilon yy^\intercal)c = (1 \pm \epsilon y^\intercal c)y \in \L^n$. This contradicts the assumption that $\mathbb{R}_+ X$ is extreme. Thus, all extreme rays $\mathbb{R}_+ X$ of $\mathcal{S}$ have $\operatorname{rank}(X) \leq 1$. ■

3.4 Necessary conditions

In this section, we give a complete characterization of ROG cones defined by two LMIs.

Theorem 16. *Let $\mathcal{M} = \{M_1, M_2\}$. Then, $\mathcal{S}(\mathcal{M})$ is ROG if and only if one of the following holds:*

i there exists $(\alpha_1, \alpha_2) \neq (0, 0)$ such that $\alpha_1 M_1 + \alpha_2 M_2 \in \mathbb{S}^n_+$, or

ii there exists $a, b, c \in \mathbb{R}^n$ such that $M_1 = \operatorname{Sym}(ac^\intercal)$ and $M_2 = \operatorname{Sym}(bc^\intercal)$.

Note that the *if* direction of Theorem 16 is a direct consequence of the sufficient conditions identified in Proposition 11 and Corollary 17. Furthermore, recall from Corollary 15 that when $|\mathcal{M}| = 2$, the set $\mathcal{S}(\mathcal{M})$ is ROG if and only if $\mathcal{T}(\mathcal{M})$ is ROG. Thus, Theorem 16 follows as a corollary to the following necessary condition.

Theorem 17. *Let $\mathcal{M} = \{M_1, M_2\}$. If $\mathcal{T}(\mathcal{M})$ is ROG, then one of the following holds:*

i there exists $(\alpha_1, \alpha_2) \neq (0, 0)$ such that $\alpha_1 M_1 + \alpha_2 M_2 \in \mathbb{S}^n_+$, or

ii there exists $a, b, c \in \mathbb{R}^n$ such that $M_1 = \operatorname{Sym}(ac^\intercal)$ and $M_2 = \operatorname{Sym}(bc^\intercal)$.

Remark 40. The conic Gordan–Stiemke Theorem (see Equation 2.3 in [166] and its surrounding comments) implies that for any subspace $W \subseteq \mathbb{S}^n$,

$$W \cap \mathbb{S}^n_+ = \{0\} \iff W^\perp \cap \mathbb{S}^n_{++} \neq \varnothing.$$

In particular, applying the conic Gordan–Stiemke Theorem in the context of Theorem 17 we deduce that if M_1, M_2 are linearly independent, then condition (i) in Theorem 17 fails if and only if $\mathcal{T}(\{M_1, M_2\})$ contains a positive definite matrix. □

Conditions (i) and (ii) in Theorems 16 and 17 have simple geometric interpretations. See Remark 36 for a geometric interpretation of (i). We describe an interpretation of condition (ii) in Theorem 17, i.e., in the case of two LMEs. Condition (ii) covers the important case when the two LMEs interact in a nontrivial manner inside $\mathbb{S}^n_+$. Suppose for the sake of presentation that $a = e_1$, $b = e_2$, $c = e_n$. Then, Corollary 17 implies that

$$\begin{aligned}
\mathcal{T}(\mathcal{M}) &= \operatorname{conv}(\{xx^\intercal : x_1 x_n = 0,\ x_2 x_n = 0\}) \\
&= \operatorname{conv}(\operatorname{conv}\{xx^\intercal : x_1 = x_2 = 0\} \cup \operatorname{conv}\{xx^\intercal : x_n = 0\}) \\
&= \operatorname{conv}\Big((0_2 \oplus \mathbb{S}^{n-2}_+) \cup (\mathbb{S}^{n-1}_+ \oplus 0_1)\Big).
\end{aligned}$$

In other words, condition (ii) covers the case where $\mathcal{T}(\mathcal{M})$ is the convex hull of the union of two faces of the positive semidefinite cone with a particular intersection structure. Theorem 17 states that these are the only ways for $\mathcal{T}(\mathcal{M})$ to be ROG when $|\mathcal{M}| = 2$.

The proof of Theorem 17 is nontrivial and will be the focus of the remainder of the section. Before completing this proof, let us first work out in detail a prototypical example. This example will highlight a number of the steps of our proof.

Example 8. Suppose $\mathcal{M} = \{M_1, M_2\}$ where $M_1 = \operatorname{Diag}(1, -1, 0)$ and $M_2 = \operatorname{Diag}(0, 1, -1)$ so that

$$\mathcal{T}(\mathcal{M}) = \left\{X \in \mathbb{S}^3_+ : X_{1,1} = X_{2,2} = X_{3,3}\right\}.$$

We first verify that neither condition (i) nor (ii) from Theorem 17 hold. Indeed, $\alpha_1 M_1 + \alpha_2 M_2 = \operatorname{Diag}(\alpha_1, \alpha_2 - \alpha_1, -\alpha_2)$ is positive semidefinite if and only if $(\alpha_1, \alpha_2) = (0, 0)$ so that condition (i) is violated. Next, note that $2M_1 + M_2 = \operatorname{Diag}(2, -1, 1)$ has rank three so that condition (ii) is also violated. We next demonstrate that $\mathcal{T}(\mathcal{M})$ is not ROG.

Let $w := \left(1, 1, \sqrt{2}\right)^\intercal$. We claim there exists a vector z such that

$$\begin{pmatrix} z^\intercal M_1 z \\ z^\intercal M_2 z \end{pmatrix} = -\begin{pmatrix} w^\intercal M_1 w \\ w^\intercal M_2 w \end{pmatrix}.$$

Indeed for this example, $z = (-1, 1, 0)^\intercal$ is such a vector. It is clear that w and z are linearly independent so that $X := ww^\intercal + zz^\intercal$ is a rank-two matrix contained in $\mathcal{T}(\mathcal{M})$. By Corollary 14, it suffices to show that $\operatorname{range}(X) \cap \mathcal{N}(\mathcal{M}) = \{0\}$. We will write a generic element from $\operatorname{range}(X)$ as $\left(\alpha - \beta, \alpha + \beta, \sqrt{2}\alpha\right)^\intercal$. Then

$$\operatorname{range}(X) \cap \mathcal{N}(\mathcal{M}) = \left\{ \begin{pmatrix} \alpha - \beta \\ \alpha + \beta \\ \sqrt{2}\alpha \end{pmatrix} : \quad (\alpha - \beta)^2 = (\alpha + \beta)^2 = 2\alpha^2 \right\}.$$

The first equality implies $\alpha\beta = 0$. The second equality then implies that $\alpha = \beta = 0$. We conclude $\operatorname{range}(X) \cap \mathcal{N}(\mathcal{M}) = \{0\}$ and that $\mathcal{T}(\mathcal{M})$ is not ROG. □

We now begin on the proof of Theorem 17. We first make a simplifying assumption that holds without loss of generality.

Lemma 33. *Let $W := \operatorname{span}(\bigcup_{M \in \mathcal{M}} \operatorname{range}(M))$. For $M \in \mathcal{M}$, let $\overline{M} = M_W$ denote the restriction of M to W. Let $\overline{\mathcal{M}} = \left\{\overline{M} : M \in \mathcal{M}\right\}$. Then, $\mathcal{T}(\mathcal{M})$ is ROG if and only $\mathcal{T}(\overline{\mathcal{M}})$ is ROG. Furthermore, if $\mathcal{M} = \{M_1, M_2\}$ and $\overline{\mathcal{M}} = \left\{\overline{M}_1, \overline{M}_2\right\}$, then each of conditions (i) and (ii) in Theorem 17 hold for $\mathcal{M}$ if and only if they hold for $\overline{\mathcal{M}}$.*

Proof. The first part of this statement follows immediately from Lemma 25. The last statement of the lemma follows from definition of W. ■

We will henceforth assume that $\mathcal{M}$ spans $\mathbb{R}^n$ in the following sense.

Assumption 7. Assume that $\operatorname{span}(\bigcup_{M \in \mathcal{M}} \operatorname{range}(M)) = \mathbb{R}^n$. □

Proof of Theorem 17. By Lemma 33, we may without loss of generality assume that Assumption 7 holds. We will split the proof of Theorem 17 into a number of cases depending on the dimension n.

- The case $n = 1$ holds vacuously as we can set (α_1, α_2) to either $(1, 0)$ or $(-1, 0)$ to satisfy (i).

- For $n = 2$, we will suppose condition (i) is not satisfied and explicitly construct an extreme ray of $\mathcal{T}(\mathcal{M})$ with rank two. The construction crucially uses the geometry of $\mathbb{R}^2$ (and $\mathbb{S}^2$). See Proposition 12.

- For $n = 3$, we will suppose that neither conditions (i) nor (ii) are satisfied and explicitly construct extreme rays of $\mathcal{T}(\mathcal{M})$ with rank two. The construction is based on understanding what the corresponding $\mathcal{N}(\mathcal{M})$ set looks like. This construction crucially use the geometry of $\mathbb{R}^3$. See Proposition 13.

- Finally, we will show how to reduce the case of $n \geq 4$ to the case of $n = 3$. Specifically, supposing that $\mathcal{T}(\mathcal{M})$ is a ROG cone, with $n \geq 4$, violating (i), we will construct $\overline{\mathcal{M}}$ such that $\mathcal{T}(\overline{\mathcal{M}})$ is a ROG cone, with $n = 3$, violating both (i) and (ii). See Proposition 14. ■

Remark 41. Suppose Assumption 7 holds. In this case, condition (ii) necessarily fails if $n \geq 4$. On the other hand if $n \leq 2$ and condition (ii) holds, then in fact condition (i) also holds. In particular, condition (i) itself completely characterizes the ROG property of a cone defined by two LMIs whenever $n \neq 3$.

Expanding Assumption 7, we have that condition (i) completely characterizes the ROG property of a cone defined by two LMIs whenever $\dim(\operatorname{span}(\operatorname{range}(M_1) \cup \operatorname{range}(M_2))) \neq 3$. □

Remark 42. Both directions of Theorems 16 and 17 admit small certificates.

- Suppose $\mathcal{S}(\mathcal{M})$ is ROG. Then Theorem 16 implies that there exists either aggregation weights $(\alpha_1, \alpha_2) \neq (0, 0)$ for which $\alpha_1 M_1 + \alpha_2 M_2 \in \mathbb{S}^n_+$ or vectors $a, b, c \in \mathbb{R}^n$ for which $M_1 = \operatorname{Sym}(ac^\intercal)$ and $M_2 = \operatorname{Sym}(bc^\intercal)$.

- Suppose $\mathcal{S}(\mathcal{M})$ is not ROG. Then by Theorem 16, it suffices to certify that neither conditions (i) nor (ii) hold. As $\mathcal{S}(\mathcal{M})$ is not ROG, we may assume that M_1 and M_2 are linearly independent. Then, the Gordan–Stiemke Theorem (see Remark 40) implies that condition (i) fails if and only if there exists a positive definite matrix X in $\mathcal{T}(\mathcal{M})$. In other words, we can certify that condition (i) fails by presenting a positive definite matrix in $\mathcal{T}(\mathcal{M})$. If either $\operatorname{rank}(M_1) \geq 3$ or $\operatorname{rank}(M_2) \geq 3$, then the spectral decomposition of the corresponding M_i certifies that condition (ii) does not hold. Else, M_1 and M_2 are both indefinite rank-two matrices and we can write $M_1 = \eta_1 \operatorname{Sym}(ab^\intercal)$ and $M_2 = \eta_2 \operatorname{Sym}(cd^\intercal)$ where $\eta_i \in \mathbb{R}$, $a, b, c, d \in \mathbf{S}^{n-1}$. This decomposition is unique up to renaming a and b or c and d. Then condition (ii) does not hold if and only if a, b, c, d are distinct. In particular, this decomposition certifies that condition (ii) does not hold. □

In the proof of Theorem 17, we will make use of the following theorem related to the convexity of the joint image of two quadratic maps.

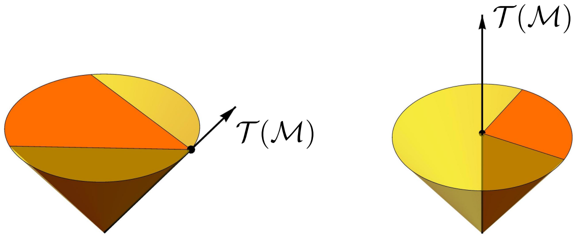

Figure 3.3: For $n = 2$, every point on the interior of $\mathbb{S}^n_+$ has rank two and every point on the boundary of $\mathbb{S}^n_+$ has rank at most one. Condition (i) implies that $\mathcal{T}(\mathcal{M})$, is either trivial or a ray in the boundary of $\mathbb{S}^n_+$—this corresponds to the picture on the left. Proposition 12 shows that when condition (i) is violated, $\mathcal{T}(\mathcal{M})$ is a ray on the interior of $\mathbb{S}^n_+$—this corresponds to the picture on the right.

Theorem 18 (Dines [59])**.** *Let $M_1, M_2 \in \mathbb{S}^n$ and suppose that for all $(\alpha_1, \alpha_2) \neq (0, 0)$, we have $\alpha_1 M_1 + \alpha_2 M_2 \notin \mathbb{S}^n_+$. Then,*

$$\left\{ \begin{pmatrix} x^\intercal M_1 x \\ x^\intercal M_2 x \end{pmatrix} \in \mathbb{R}^2 : x \in \mathbb{R}^n \right\} = \mathbb{R}^2,$$

i.e., for every $y \in \mathbb{R}^2$, there exists an $x \in \mathbb{R}^n$ such that $x^\intercal M_1 x = y_1$ and $x^\intercal M_2 x = y_2$.

3.4.1 Dimension $n = 2$

We now prove Theorem 17 for the case $n = 2$.

Proposition 12. *Let $\mathcal{M} = \{M_1, M_2\}$. Suppose Assumption 7 holds and $n = 2$. If $\mathcal{T}(\mathcal{M})$ is ROG then there exists $(\alpha_1, \alpha_2) \neq (0, 0)$ such that $\alpha_1 M_1 + \alpha_2 M_2 \in \mathbb{S}^n_+$.*

Proof. Suppose for all $(\alpha_1, \alpha_2) \neq (0, 0)$, the linear combination $\alpha_1 M_1 + \alpha_2 M_2$ is not positive semidefinite. In particular, M_1 and M_2 are linearly independent in $\mathbb{S}^2$. Then, by Gordan–Stiemke Theorem (see Remark 40), we deduce the existence of a positive definite matrix $X \in \mathcal{T}(\mathcal{M})$.

Finally, as $\mathbb{S}^2$ has dimension three, the space orthogonal to both M_1 and M_2 has dimension one, so that in fact $\mathcal{T}(\mathcal{M}) = \mathbb{R}_+(X)$. We conclude that $\mathbb{R}_+(X)$ is an extreme ray with $\operatorname{rank}(X) = 2$. ■

3.4.2 Dimension $n = 3$

We will make use of the following lemma from Hildebrand [83, Lemma 3.13]. The lemma states that the Carathéodory number of an element X of $\mathcal{T}(\mathcal{M})$ is equal to $\operatorname{rank}(X)$ when $\mathcal{T}(\mathcal{M})$ is ROG.

Lemma 34 ([83, Lemma 3.13])**.** *Suppose $\mathcal{T}(\mathcal{M})$ is ROG. For every $X \in \mathcal{T}(\mathcal{M})$, we can write $X = \sum_{i=1}^r x_i x_i^\intercal$ where $x_i \in \mathcal{N}(\mathcal{M})$ for all $i \in [r]$ and $r = \operatorname{rank}(X)$.*

The next lemma states that when neither conditions (i) nor (ii) hold, the set $\mathcal{N}(\mathcal{M})$ is extremely sparse in $\mathbb{R}^3$.

Lemma 35. *Let $\mathcal{M} = \{M_1, M_2\}$. Suppose Assumption 7 holds and $n = 3$. If neither conditions (i) nor (ii) of Theorem 17 hold, then $\mathcal{N}(\mathcal{M})$ is the union of at most four one-dimensional subspaces of $\mathbb{R}^3$.*

Readers familiar with algebraic geometry will recognize this as a consequence of Bézout's theorem.[8] For completeness, we provide an elementary proof of this lemma using only linear algebraic tools in Section C.1.

We are now ready to prove Theorem 17 for the case of $n = 3$. We will assume that neither conditions (i) nor (ii) hold and use Lemma 35 and Theorem 18 to construct a rank-two matrix contained in $\mathcal{T}(\mathcal{M})$. We will then apply Lemma 34 to derive a contradiction.

Proposition 13. *Let $\mathcal{M} = \{M_1, M_2\}$. Suppose Assumption 7 holds and $n = 3$. If $\mathcal{T}(\mathcal{M})$ is ROG, then one of conditions (i) or (ii) of Theorem 17 must hold.*

Proof. Suppose $\mathcal{T}(\mathcal{M})$ is ROG but neither conditions (i) nor (ii) hold. Consider the subset of $\mathbb{R}^3$ given by

$$\mathcal{R} := \bigcup_{x,y\in\mathcal{N}(\mathcal{M})} \operatorname{span}(\{x, y\}).$$

By Lemma 35, we have that $\mathcal{R}$ is the union of a finite number of planes and lines in $\mathbb{R}^3$, and thus there exists $w \notin \mathcal{R}$. By Theorem 18, we can pick z such that

$$\begin{pmatrix} z^\intercal M_1 z \\ z^\intercal M_2 z \end{pmatrix} = -\begin{pmatrix} w^\intercal M_1 w \\ w^\intercal M_2 w \end{pmatrix}.$$

As $w \notin \mathcal{R}$, we deduce at least one of $w^\intercal M_1 w$ and $w^\intercal M_2 w$ is nonzero. Then, it is clear that w and z are linearly independent, and thus $X := ww^\intercal + zz^\intercal$ is a rank-two matrix contained in $\mathcal{T}(\mathcal{M})$.

As $\mathcal{T}(\mathcal{M})$ is ROG, we can apply Lemma 34. In particular, we can write $X = xx^\intercal + yy^\intercal$ for some $x, y \in \mathcal{N}(\mathcal{M})$. Then, $w \in \operatorname{range}(X) = \operatorname{span}(x, y) \subseteq \mathcal{R}$. This contradicts our choice of $w \notin \mathcal{R}$. ∎

3.4.3 Dimensions $n \geq 4$

We will now reduce the case of $n \geq 4$ to $n = 3$. The proof will show that if $\mathcal{M}$ violates condition (i) then there exists a three-dimensional subspace W for which the restriction of $\mathcal{M}$ to W fails both conditions (i) and (ii).

We begin by showing that there exists a linear combination of M_1 and M_2 with rank at least three.

[8] Assuming that neither conditions (i) nor (ii) hold, the plane curves defined by M_1 and M_2 cannot share a common component. Then Bézout's theorem implies that $\mathcal{N}(\mathcal{M})$ consists of at most four lines (or equivalently, four points in projective space).

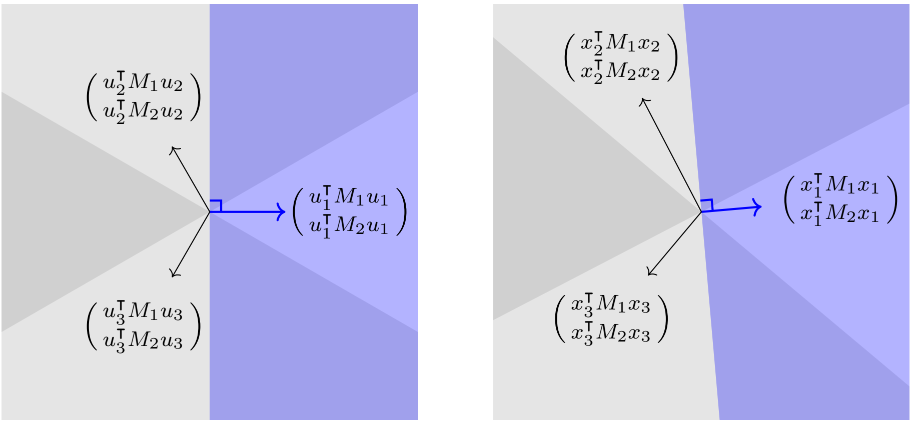
$\begin{pmatrix} u_2^\intercal M_1 u_2 \\ u_2^\intercal M_2 u_2 \end{pmatrix}$
$\begin{pmatrix} u_1^\intercal M_1 u_1 \\ u_1^\intercal M_2 u_1 \end{pmatrix}$
$\begin{pmatrix} u_3^\intercal M_1 u_3 \\ u_3^\intercal M_2 u_3 \end{pmatrix}$
$\begin{pmatrix} x_2^\intercal M_1 x_2 \\ x_2^\intercal M_2 x_2 \end{pmatrix}$
$\begin{pmatrix} x_1^\intercal M_1 x_1 \\ x_1^\intercal M_2 x_1 \end{pmatrix}$
$\begin{pmatrix} x_3^\intercal M_1 x_3 \\ x_3^\intercal M_2 x_3 \end{pmatrix}$

Figure 3.4: The proof of Proposition 14 assumes that condition (i) in Theorem 17 does not hold for $\{M_1, M_2\}$ and constructs $u_1, u_2, u_3 \in \mathbb{R}^n$ such that the vectors $\{(u_i^\intercal M_1 u_i, u_i^\intercal M_2 u_i)\} \subseteq \mathbb{R}^2$ are located as shown in the left figure. These vectors certify that condition (i) in Theorem 17 does not hold for $\{M_1, M_2\}$. Indeed, if $\alpha_1 M_1 + \alpha_2 M_2 \in \mathbb{S}^n_+$, then (α_1, α_2) must lie in the intersection of the three halfspaces defined by the u_i vectors (one such halfspace is shaded), whence $(\alpha_1, \alpha_2) = (0, 0)$. A key observation in the proof of Proposition 14 is that for all $x_1, x_2, x_3 \in \mathbb{R}^n$ close enough to u_1, u_2, u_3, the vectors $\{(x_i^\intercal M_1 x_i, x_i^\intercal M_2 x_i)\} \subseteq \mathbb{R}^2$ certify that condition (i) in Theorem 17 also does not hold for $\{(M_1)_W, (M_2)_W\}$ where $W = \operatorname{span}(\{x_i\})$. Again, the intersection of the corresponding halfspaces is trivial.

Lemma 36. *Let $\mathcal{M} = \{M_1, M_2\}$. Suppose Assumption 7 holds and $n \geq 4$. If condition (i) in Theorem 17 does not hold, then there exists (α_1, α_2) such that $\operatorname{rank}(\alpha_1 M_1 + \alpha_2 M_2) \geq 3$.*

Proof. Suppose $\operatorname{rank}(\alpha_1 M_1 + \alpha_2 M_2) \leq 2$ for all (α_1, α_2). Because condition (i) does not hold, we conclude that for all $(\alpha_1, \alpha_2) \neq (0, 0)$, the linear combination $\alpha_1 M_1 + \alpha_2 M_2$ has exactly one positive and one negative eigenvalue. Then, we can write $M_1 = \operatorname{Sym}(ab^\intercal)$ and $M_2 = \operatorname{Sym}(cd^\intercal)$. By Assumption 7, we have that a, b, c, d are linearly independent. By independence, there exists an x such that $x^\intercal b = 1$ and $x^\intercal a = x^\intercal c = x^\intercal d = 0$; we deduce that $(M_1 + M_2)x = a \in \operatorname{range}(M_1 + M_2)$. Similarly, $b, c, d \in \operatorname{range}(M_1 + M_2)$. Then $\operatorname{rank}(M_1 + M_2) = 4$, a contradiction. ■

We are now ready to prove Theorem 17 for the case of $n \geq 4$.

Proposition 14. *Let $\mathcal{M} = \{M_1, M_2\}$. Suppose Assumption 7 holds and $n \geq 4$. If $\mathcal{T}(\mathcal{M})$ is ROG, then there exists $(\alpha_1, \alpha_2) \neq (0, 0)$ such that $\alpha_1 M_1 + \alpha_2 M_2 \in \mathbb{S}^n_+$.*

Proof. Suppose for the sake of contradiction that $\mathcal{T}(\mathcal{M})$ is ROG but condition (i) in Theorem 17 does not hold.

Let $\theta_1 := 0$, $\theta_2 := 2\pi/3$ and $\theta_3 := 4\pi/3$. Then, using Theorem 18 we can find three vectors $u_1, u_2, u_3 \in \mathbb{R}^n$ satisfying

$$\begin{pmatrix} u_i^\intercal M_1 u_i \\ u_i^\intercal M_2 u_i \end{pmatrix} = \begin{pmatrix} \cos(\theta_i) \\ \sin(\theta_i) \end{pmatrix} \qquad \forall i \in [3]. \tag{3.5}$$

Note that u_1, u_2, u_3 certify that condition (i) does not hold for $\mathcal{M}$ (see also Figure 3.4):

$$\begin{aligned}\{(\alpha_1,\alpha_2) :\ \alpha_1 M_1 + \alpha_2 M_2 \succeq 0\} &\subseteq \{(\alpha_1,\alpha_2) :\ u_i^\intercal(\alpha_1 M_1 + \alpha_2 M_2)u_i \geq 0,\ \forall i \in [3]\}\\ &= \left\{(\alpha_1,\alpha_2) :\ \left\langle \begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}, \begin{pmatrix}u_i^\intercal M_1 u_i\\ u_i^\intercal M_2 u_i\end{pmatrix}\right\rangle \geq 0,\ \forall i \in [3]\right\} = \{(0,0)\}.\end{aligned}$$

Next, by Lemma 36, there exists $M_\beta := \beta_1 M_1 + \beta_2 M_2$ with rank at least three. Let $v_1, v_2, v_3 \in \mathbb{R}^n$ be orthonormal eigenvectors of M_β corresponding to nonzero eigenvalues. Note that v_1, v_2, v_3 certify that condition (ii) does not hold for $\mathcal{M}$:

$$\det\left(\begin{pmatrix}v_1^\intercal\\ v_2^\intercal\\ v_3^\intercal\end{pmatrix} M_\beta \begin{pmatrix}v_1 & v_2 & v_3\end{pmatrix}\right) \neq 0 \implies \operatorname{rank}(M_\beta) \geq 3.$$

We will use the vectors $\{u_i\}$ and $\{v_i\}$ to construct a three-dimensional subspace $W \subseteq \mathbb{R}^n$ and show that the certificates of neither conditions (i) nor (ii) holding in $\mathcal{M}$ can be used to find certificates of neither conditions (i) nor (ii) holding in $\{(M_1)_W, (M_2)_W\}$.

Let $\mu \in (0,1]$ to be fixed later. Define $x_i := (1-\mu)u_i + \mu v_i$ and set $W := \operatorname{span}\{x_1, x_2, x_3\}$. Let $\overline{M}_i := (M_i)_W$ and set $\overline{\mathcal{M}} := \left\{\overline{M}_1, \overline{M}_2\right\}$. Similarly define $\overline{M}_\beta$.

We first show that W is a three-dimensional subspace for all $\mu > 0$ small enough. It is clear that $\dim(W) \leq 3$. To see that $\dim(W) \geq 3$ for all $\mu > 0$ small enough, consider the determinant of the orthogonal projections of the x_i vectors onto $\operatorname{span}\{v_1, v_2, v_3\}$,

$$\det\left(\begin{pmatrix}v_1^\intercal\\ v_2^\intercal\\ v_3^\intercal\end{pmatrix}\begin{pmatrix}x_1 & x_2 & x_3\end{pmatrix}\right) = \det\begin{pmatrix}v_1^\intercal x_1 & v_1^\intercal x_2 & v_1^\intercal x_3\\ v_2^\intercal x_1 & v_2^\intercal x_2 & v_2^\intercal x_3\\ v_3^\intercal x_1 & v_3^\intercal x_2 & v_3^\intercal x_3\end{pmatrix}.$$

Recalling that the x_is are each linear in μ, we deduce that this determinant is a degree-3 polynomial in μ which is not identically zero (taking $\mu = 1$ gives the determinant of the identity matrix), and thus $\{x_i\}$ are linearly independent for all $\mu > 0$ small enough.

Next, we show that condition (i) does not hold for $\overline{\mathcal{M}}$ for all $\mu > 0$ small enough. Note that

$$\begin{aligned}\left\{(\alpha_1,\alpha_2) :\ \alpha_1 \overline{M}_1 + \alpha_2 \overline{M}_2 \succeq 0\right\} &\subseteq \{(\alpha_1,\alpha_2) :\ x_i^\intercal(\alpha_1 M_1 + \alpha_2 M_2)x_i \geq 0,\ \forall i \in [3]\}\\ &= \left\{(\alpha_1,\alpha_2) :\ \left\langle \begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}, \begin{pmatrix}x_i^\intercal M_1 x_i\\ x_i^\intercal M_2 x_i\end{pmatrix}\right\rangle \geq 0,\ \forall i \in [3]\right\},\end{aligned}$$

where the first relation follows from the definition of $\overline{M}_i$ and noting that $x_i \in W$. By continuity of the quadratic forms $x_i^\intercal M_1 x_i$ and $x_i^\intercal M_2 x_i$ in the variable μ, and the choice of the u_i in Equation (3.5), the set on the second line above is the trivial set $\{0\}$ for all $\mu > 0$ small enough. Thus, $\overline{\mathcal{M}}$ does not satisfy condition (i) for all $\mu > 0$ small enough.

Next, we will show that $\overline{M}_\beta$ has rank three for all $\mu > 0$ small enough. Note that $\overline{M}_\beta$ is singular if and only if $\det(\overline{M}_\beta) = 0$. Picking the basis $\{x_1, x_2, x_3\}$ of W, we have that $\det(\overline{M}_\beta) = 0$ if and only if

$$\det\left(\begin{pmatrix} x_1^\intercal \\ x_2^\intercal \\ x_3^\intercal \end{pmatrix} M_\beta \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix}\right) = \det\begin{pmatrix} x_1^\intercal M_\beta x_1 & x_1^\intercal M_\beta x_2 & x_1^\intercal M_\beta x_3 \\ x_2^\intercal M_\beta x_1 & x_2^\intercal M_\beta x_2 & x_2^\intercal M_\beta x_3 \\ x_3^\intercal M_\beta x_1 & x_3^\intercal M_\beta x_2 & x_3^\intercal M_\beta x_2 \end{pmatrix} = 0.$$

This is a degree-6 polynomial in μ (recall that x_is are linear in μ) that is not identically zero: for $\mu = 1$, this determinant evaluates to the product of three nonzero eigenvalues of M_β. Then, for all $\mu > 0$ small enough, this polynomial is nonzero and hence $\operatorname{rank}(\overline{M}_\beta) = 3$. Thus, we deduce that $\overline{\mathcal{M}}$ does not satisfy condition (ii) for all $\mu > 0$ small enough.

We now fix μ such that $\overline{\mathcal{M}}$ does not satisfy either condition (i) or (ii). Note that this also fixes W.

To complete the proof we will show that $\mathcal{T}(\overline{\mathcal{M}})$ is ROG. This will contradict Proposition 13. Note that

$$\mathcal{T}(\overline{\mathcal{M}}) \oplus 0_{W^\perp} = \mathcal{T}(\mathcal{M}) \cap \{X \in \mathbb{S}^n_+ : \langle 0_W \oplus I_{W^\perp}, X\rangle = 0\},$$

which is a face of $\mathcal{T}(\mathcal{M})$. Then, as $\mathcal{T}(\mathcal{M})$ is ROG, Lemma 21 implies that $\mathcal{T}(\overline{\mathcal{M}}) \oplus 0_{W^\perp}$ is ROG. Next, note that $\mathcal{T}(\overline{\mathcal{M}}) \oplus 0_{W^\perp}$ is isomorphic to $\mathcal{T}(\overline{\mathcal{M}})$ via the rank-preserving map $X_W \oplus 0_{W^\perp} \mapsto X_W$. We conclude that $\mathcal{T}(\mathcal{M})$ is ROG. ■

Proposition 14, together with Propositions 12 and 13, concludes the proof of Theorem 17.

3.4.4 Lifting LMIs into LMEs

In this section, we will show that a simple lifting of an LMI set $\mathcal{S}$ into an LME set $\mathcal{T}$ in a larger dimension may not preserve the ROG property.

Example 9. Consider the set

$$\mathcal{S} := \left\{X \in \mathbb{S}^3_+ : \begin{array}{c} X_{1,2} = 0 \\ X_{1,3} \geq 0 \end{array}\right\}.$$

This set is ROG by Theorem 16 and Lemma 21. We can replace the LMIs defining $\mathcal{S}$ with LMEs in a lifted space as follows: Let $\Pi : \mathbb{S}^4 \to \mathbb{S}^3$ denote the projection of a 4×4 matrix onto its top-left 3×3 principal submatrix. Then,

$$\mathcal{S} = \Pi\left(\left\{X \in \mathbb{S}^4 : \begin{array}{c} X_{1,2} = 0 \\ X_{1,3} - X_{4,4} = 0 \end{array}\right\}\right) = \Pi(\mathcal{T}(\{M_1', M_2'\})),$$

where

$$M_1' := \begin{pmatrix} 0 & 1/2 & 0 & 0 \\ 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \qquad \text{and} \qquad M_2' := \begin{pmatrix} 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

Define $\mathcal{M}' := \{M_1', M_2'\}$. By Theorem 16, we see that $\mathcal{T}(\mathcal{M}')$ is not ROG. We conclude that the obvious lifting of LMIs into LMEs can take ROG sets $\mathcal{S}(\mathcal{M})$ to non-ROG sets $\mathcal{T}(\mathcal{M}')$ (even when there is only a single inequality to lift). □

3.5 Applications of ROG cones

3.5.1 Exactness of SDP relaxations of QCQPs

In this subsection, we relate the ROG property of a cone $\mathcal{S}$ to exactness results for both homogeneous and inhomogeneous QCQPs and their relaxations.

The following lemma states that a cone $\mathcal{S} \subseteq \mathbb{S}^n_+$ is ROG if and only if the SDP relaxation of the corresponding homogeneous QCQP is exact for all choices of objective function.

Lemma 37. *Let $\mathcal{M} \subseteq \mathbb{S}^n$. Then $\mathcal{S}(\mathcal{M})$ is ROG if and only if for every $M_0 \in \mathbb{S}^n$,*

$$\inf_{X \in \mathcal{S}(\mathcal{M})} \langle M_0, X \rangle = \inf_{x \in \mathbb{R}^n} \{\langle M_0, xx^\intercal \rangle : xx^\intercal \in \mathcal{S}(\mathcal{M})\}. \tag{3.6}$$

Proof. By Definition 13, $\mathcal{S}(\mathcal{M})$ is ROG if and only if $\mathcal{S}(\mathcal{M}) = \operatorname{conv}(\mathcal{S}(\mathcal{M}) \cap \{xx^\intercal : x \in \mathbb{R}^n\})$. Moreover, both $\mathcal{S}(\mathcal{M})$ and $\operatorname{conv}(\mathcal{S}(\mathcal{M}) \cap \{xx^\intercal : x \in \mathbb{R}^n\})$ are closed convex cones so that they are equal if and only if their dual cones are equal. Note that

$$M_0 \in \mathcal{S}(\mathcal{M})^* \iff \inf_{X \in \mathcal{S}(\mathcal{M})} \langle M_0, X \rangle = 0.$$

Similarly,

$$M_0 \in (\operatorname{conv}(\mathcal{S}(\mathcal{M}) \cap \{xx^\intercal : x \in \mathbb{R}^n\}))^* \iff \inf_{x \in \mathbb{R}^n} \{\langle M_0, xx^\intercal \rangle : xx^\intercal \in \mathcal{S}(\mathcal{M})\} = 0.$$

Noting that both sides of (3.6) can only take the values 0 or $-\infty$ completes the proof. ■

Next, we consider a general QCQP and its SDP relaxation. Recall that in the general form given in (3.1), a QCQP and its SDP relaxation both contain exactly one inhomogeneous equality constraint. The following lemma relates the ROG property of a cone to SDP exactness results for its affine slices. This will allow us to apply our main results on spectrahedral cones to spectrahedra arising as the feasible domain of the SDP relaxations in (3.1).

Lemma 38. *Let $\mathcal{M} \subseteq \mathbb{S}^n$ and $B \in \mathbb{S}^n$. If $\mathcal{S}(\mathcal{M})$ is ROG, then*

$$\inf_{x\in\mathbb{R}^n}\left\{x^\intercal M_0 x : \begin{array}{l} x^\intercal M x \geq 0,\ \forall M \in \mathcal{M} \\ x^\intercal B x = 1 \end{array}\right\} = \inf_{X\in\mathbb{S}^n}\left\{\langle M_0, X\rangle : \begin{array}{l} \langle M, X\rangle \geq 0,\ \forall M \in \mathcal{M} \\ \langle B, X\rangle = 1 \\ X \succeq 0 \end{array}\right\}$$

for all $M_0 \in \mathbb{S}^n$ for which the optimum SDP objective value is bounded from below. In particular, this equality holds whenever the SDP feasible domain is bounded.

Proof. Let $\mathcal{S} := \mathcal{S}(\mathcal{M})$.

($\geq$) This direction is immediate as the SDP gives a relaxation of the QCQP.

($\leq$) We may assume without loss of generality that the SDP is feasible. Let X be a feasible SDP solution. As $X \in \mathcal{S}$ and $\mathcal{S}$ is an ROG cone, there exist $x_1, \ldots, x_r \in \mathbb{R}^n$ such that $x_i x_i^\intercal \in \mathcal{S}$ for all $i \in [r]$ and $X = \sum_{i=1}^r x_i x_i^\intercal$. That is, we have $x_i^\intercal M x_i \geq 0$ for all $M \in \mathcal{M}$ and $i \in [r]$. Without loss of generality, $x_i^\intercal B x_i$ is non-increasing in i and there exists some $k \in [r]$ such that $x_1^\intercal B x_1, \ldots, x_k^\intercal B x_k$ are positive scalars summing to one. Indeed, if this were to fail, we could first rearrange the indices in $[r]$ to get $x_i^\intercal B x_i$ in non-increasing order and then subdivide the first term $x_k x_k^\intercal$ for which $\sum_{i=1}^k x_i^\intercal B x_i \geq 1$ into two terms $(\sqrt{\alpha}x_k)(\sqrt{\alpha}x_k)^\intercal + (\sqrt{1-\alpha}x_k)(\sqrt{1-\alpha}x_k)^\intercal$ (naturally, also increasing r to $r+1$) so that the first k-many values of $x_i^\intercal B x_i$ are positive and sum to one. From here on we assume that such a transformation has been done (if needed), and r reflects the final number of summands in this decomposition of X.

We may then write

$$X = \hat{X} + \tilde{X} := \left(\sum_{i=1}^k x_i x_i^\intercal\right) + \left(\sum_{i=k+1}^r x_i x_i^\intercal\right).$$

Note that $\left\langle B, \tilde{X}\right\rangle = \langle B, X\rangle - \left\langle B, \hat{X}\right\rangle = 1 - 1 = 0$. Moreover, because the optimum SDP objective value is bounded from below, we must have $\left\langle M_0, \tilde{X}\right\rangle \geq 0$.

For $i \in [k]$, define $\mu_i := x_i^\intercal B x_i > 0$ and $\hat{x}_i := x_i/\sqrt{\mu_i}$. Then, $\hat{x}_i^\intercal B \hat{x}_i = 1$ and $\hat{x}_i^\intercal M \hat{x}_i \geq 0$ for all $M \in \mathcal{M}$ and $i \in [k]$. Finally, note that $1 = \sum_{i=1}^k x_i^\intercal B x_i = \sum_{i=1}^k \mu_i$. Using these facts, we deduce

$$\begin{aligned} \langle M_0, X\rangle \geq \left\langle M_0, \hat{X}\right\rangle &= \sum_{i=1}^k x_i^\intercal M_0 x_i = \sum_{i=1}^k \mu_i \hat{x}_i^\intercal M_0 \hat{x}_i \\ &\geq \min_{i\in[k]} \hat{x}_i^\intercal M_0 \hat{x}_i \geq \inf_{x\in\mathbb{R}^n}\left\{x^\intercal M_0 x : \begin{array}{l} x^\intercal M x \geq 0,\ \forall M \in \mathcal{M} \\ x^\intercal B x = 1 \end{array}\right\}. \end{aligned}$$

The desired result follows by taking the infimum of this inequality over feasible solutions X to the SDP. ■

Remark 43. Lemma 38 extends [83, Lemma 1.2], which shows that the same statement holds in the case of finitely many LMEs. The proof we present is new and immediately shows how to construct a QCQP feasible solution achieving the SDP value (or a sequence approaching the SDP value). □

Example 10. The reverse implication in Lemma 38 is not true in general. In particular, consider the following example. Let

$$\mathcal{S} = \left\{ \begin{pmatrix} \alpha & & \\ & \beta & \\ & & \beta \end{pmatrix} : \alpha, \beta \geq 0 \right\} \subseteq \mathbb{S}^3_+,$$

and set $B = e_1 e_1^\intercal$. Note that $\mathcal{S}$ has a rank-two extreme ray and thus is not ROG. Let $M_0 \in \mathbb{S}^3$. A short calculation shows that the SDP relaxation of the QCQP defined by M_0 and $\mathcal{M}$ associated with $\mathcal{S}$ satisfies

$$\inf_{X \in \mathbb{S}^3} \left\{ \langle M_0, X \rangle : \begin{array}{c} X \in \mathcal{S} \\ X_{1,1} = 1 \end{array} \right\} = \begin{cases} (M_0)_{1,1} & \text{if } (M_0)_{2,2} + (M_0)_{3,3} \geq 0, \\ -\infty & \text{else.} \end{cases}$$

In particular, if $M_0 \in \mathbb{S}^3$ is such that the optimum value of the SDP relaxation is bounded below, then the SDP relaxation takes the value $(M_0)_{1,1}$. On the other hand, $e_1 e_1^\intercal \in \mathcal{S}$ is a rank-one matrix achieving the same objective value. We deduce that

$$\inf_{x \in \mathbb{R}^3} \left\{ x^\intercal M_0 x : \begin{array}{c} xx^\intercal \in \mathcal{S} \\ (xx^\intercal)_{1,1} = 1 \end{array} \right\} = \inf_{X \in \mathbb{S}^3} \left\{ \langle M_0, X \rangle : \begin{array}{c} X \in \mathcal{S} \\ X_{1,1} = 1 \end{array} \right\}$$

for all $M_0 \in \mathbb{S}^3$ for which the right hand side is bounded below. □

Lemma 38 implies that equality holds in (3.1) whenever $\mathcal{S}(\{M_1, \dots, M_m\})$ is ROG and the SDP optimum value is bounded from below. It may be natural to ask whether the boundedness assumption can be dropped in the case where B is specialized to $B = e_1 e_1^\intercal$. Indeed, this is the only case we need when analyzing (3.1). The following example shows that this is not possible.

Example 11. Let $n = 2$ and $\mathcal{M} = \{\mathrm{Sym}(e_1 e_2^\intercal), -\mathrm{Sym}(e_1 e_2^\intercal)\}$ so that

$$\mathcal{S}(\mathcal{M}) = \left\{ \begin{pmatrix} x_1^2 & 0 \\ 0 & x_2^2 \end{pmatrix} : x \in \mathbb{R}^2 \right\} = \mathrm{conv}\left(\left\{ \begin{pmatrix} x_1 \\ 0 \end{pmatrix} \begin{pmatrix} x_1 \\ 0 \end{pmatrix}^\intercal : x_1 \in \mathbb{R} \right\} \cup \left\{ \begin{pmatrix} 0 \\ x_2 \end{pmatrix} \begin{pmatrix} 0 \\ x_2 \end{pmatrix}^\intercal : x_2 \in \mathbb{R} \right\} \right).$$

The representation on the right shows that $\mathcal{S}(\mathcal{M})$ is ROG. On the other hand, taking $B = e_1 e_1^\intercal$ and $M_0 = -e_2 e_2^\intercal$, we have

$$\inf_{x \in \mathbb{R}^2} \left\{ x^\intercal M_0 x : \begin{array}{c} xx^\intercal \in \mathcal{S}(\mathcal{M}) \\ x^\intercal B x = 1 \end{array} \right\} = \inf_{x \in \mathbb{R}^2} \left\{ -x_2^2 : \begin{array}{c} x_1 x_2 = 0 \\ x_1^2 = 1 \end{array} \right\} = 0,$$

which is not equal to

$$\inf_{X \in \mathbb{S}^2} \left\{ \langle M_0, X \rangle : \begin{array}{c} X \in \mathcal{S}(\mathcal{M}) \\ \langle B, X \rangle = 1 \end{array} \right\} = \inf_{x \in \mathbb{R}^2} \left\{ -x_2^2 : \ x_1^2 = 1 \right\} = -\infty. \qquad \square$$

In a sense, Example 11 exhibits a particular worst-case behavior. Specifically, adding an arbitrary inhomogeneous constraint to a ROG cone produces a set that is rank-two generated.

Lemma 39. *Let $\mathcal{M} \subseteq \mathbb{S}^n$. If $\mathcal{S}(\mathcal{M})$ is ROG, then for all $B \in \mathbb{S}^n$,*

$$\operatorname{conv}\left(\left\{X \in \mathbb{S}^n : \begin{array}{l} \langle M, X\rangle \geq 0,\ \forall M \in \mathcal{M} \\ \langle B, X\rangle = 1 \\ X \succeq 0 \\ \operatorname{rank}(X) \leq 2 \end{array}\right\}\right) = \left\{X \in \mathbb{S}^n : \begin{array}{l} \langle M, X\rangle \geq 0,\ \forall M \in \mathcal{M} \\ \langle B, X\rangle = 1 \\ X \succeq 0 \end{array}\right\}.$$

In particular, when $\mathcal{S}(\mathcal{M})$ is ROG, for any $M_0 \in \mathbb{S}^n$, there exists a sequence of rank-two solutions approaching the SDP optimum value in (3.1)*.*

Proof. Let $\mathcal{L}$ denote the inner set on the left hand side so that the left hand side is $\operatorname{conv}(\mathcal{L})$ and let $\mathcal{R}$ denote the right hand set.

($\subseteq$) This follows upon noting that $\mathcal{L} \subseteq \mathcal{R}$ and $\mathcal{R}$ is convex.

($\supseteq$) Let $X \in \mathcal{R}$. As $\mathcal{R} \subseteq \mathcal{S}(\mathcal{M})$, we may decompose $X = \sum_{i=1}^r x_i x_i^\intercal$ where $x_i x_i^\intercal \in \mathcal{S}(\mathcal{M})$ for all $i \in [r]$. We may assume that $r = \operatorname{rank}(X)$ by Lemma 34. Let $\beta_i := \langle B, x_i x_i^\intercal\rangle$.

If $\beta_i > 0$ for all $i \in [r]$, then we are done. Else, without loss of generality $\beta_1 > 0 \geq \beta_2$. Consider the value of $\mu := \alpha_1\beta_1 + \alpha_2\beta_2$ as (α_1, α_2) moves continuously on the line segments $(1, 0) \to (1, 1) \to (0, 1)$. Noting that $\beta_1 > 0$ and $\beta_2 \leq 0$, we may fix (α_1, α_2) on this path such that $\mu \in (0, 1)$. Then, we can decompose

$$X = \mu\left(\frac{\alpha_1 x_1 x_1^\intercal + \alpha_2 x_2 x_2^\intercal}{\mu}\right) + (1-\mu)\left(\frac{X - \alpha_1 x_1 x_1^\intercal - \alpha_2 x_2 x_2^\intercal}{1-\mu}\right) =: \mu X_\ell + (1-\mu) X_r.$$

We have written X as a convex combination of two matrices X_ℓ and X_r. It can be verified easily that $X_\ell \in \mathcal{L}$ and $X_r \in \mathcal{R}$. As at least one of α_1 or α_2 takes the value 1, the element X_r has rank strictly less than r. Iterating this procedure completes the proof. ■

Remark 44. A result similar to Lemma 39 in the case of a single *homogeneous* constraint is presented in [33, Lemma 5]. Specifically, it is shown that for an arbitrary closed convex cone $\mathcal{S}$, the extreme rays of the set obtained by intersecting $\mathcal{S}$ with a hyperplane through the origin can be written as convex combinations of at most two extreme rays of $\mathcal{S}$. □

3.5.2 Convex hulls of bounded quadratically constrained sets

Consider a set

$$\mathcal{Y} := \left\{y \in \mathbb{R}^{n-1} : q_i(y) \geq 0,\ \forall i \in [m]\right\},$$

where q_is are quadratic functions of the form $q_i(y) = y^\intercal A_i y + 2b_i^\intercal y + c_i$. Let $M_i := \begin{pmatrix} c_i & b_i^\intercal \\ b_i & A_i \end{pmatrix}$ and $\mathcal{M} := \{M_1, \dots, M_m\}$.

We begin by proving a technical lemma that will be useful in the remainder of this section. This lemma states that under a definiteness assumption, the set $\mathcal{Y}$, its projected SDP relaxation, and its SDP relaxation are each compact.

Lemma 40. *Suppose there exists $\lambda^* \in \mathbb{R}^m_+$ such that $\sum_{i=1}^m \lambda^* A_i$ is negative definite. Then, the following three sets are each compact:*

$$\left\{y \in \mathbb{R}^{n-1} :\quad y^\intercal A_i y + 2\langle b_i, y\rangle + c_i \geq 0,\ \forall i \in [m] \right\},$$
$$\left\{y \in \mathbb{R}^{n-1} :\ \begin{array}{l} \exists Y \succeq yy^\intercal : \\ \langle A_i, Y\rangle + 2\langle b_i, y\rangle + c_i \geq 0,\ \forall i \in [m] \end{array}\right\},\ \textit{and}$$
$$\left\{X \in \mathbb{S}^n_+ :\ \begin{array}{l} \langle M_i, X\rangle \geq 0,\ \forall i \in [m] \\ \langle e_1 e_1^\intercal, X\rangle = 1 \end{array}\right\}.$$

Proof. For convenience, let $\mathcal{Y}, \mathcal{Y}', \mathcal{Y}''$ denote the three sets in the lemma statement. Let $A^* := \sum_{i=1}^m \lambda_i^* A_i$. Similarly define b^* and c^*. Note in particular that A^* is negative definite.

To see that $\mathcal{Y}''$ is compact, note that if $X \in \mathcal{Y}''$, then for all $\mu \in \mathbb{R}$ we have

$$\left\langle \begin{pmatrix} c^* - \mu & (b^*)^\intercal \\ b^* & A^* \end{pmatrix}, X \right\rangle \geq -\mu.$$

By picking μ large enough, we can ensure that the matrix on the left hand side of this inequality is negative definite. We conclude that $\mathcal{Y}''$ is bounded, whence compact.

Note that the $\mathcal{Y}'$ is the image of the compact set $\mathcal{Y}''$ under the continuous map $\begin{pmatrix} 1 & y^\intercal \\ y & Y \end{pmatrix} \mapsto y$ so that $\mathcal{Y}'$ is compact.

Finally, note that $\mathcal{Y} \subseteq \mathcal{Y}'$ so that $\mathcal{Y}$ is bounded. As $\mathcal{Y}$ is closed, it is also compact. ■

The following lemma gives an explicit description of $\operatorname{conv}(\mathcal{Y})$ under the assumption that $\mathcal{S}(\mathcal{M})$ is ROG and $\mathcal{Y}$ satisfies the above definiteness assumption.

Proposition 15. *Suppose there exists $\lambda^* \in \mathbb{R}^m_+$ such that $\sum_{i=1}^m \lambda_i^* A_i$ is negative definite. If $\mathcal{S}(\mathcal{M})$ is ROG, then $\operatorname{conv}(\mathcal{Y})$ is a semidefinite-representable set given by*

$$\operatorname{conv}(\mathcal{Y}) = \left\{y \in \mathbb{R}^{n-1} :\ \begin{array}{l} \exists Y \succeq yy^\intercal : \\ \langle A_i, Y\rangle + 2\langle b_i, y\rangle + c_i \geq 0,\ \forall i \in [m] \end{array}\right\}. \tag{3.7}$$

Proof. As the assumptions of Lemma 40 hold, we have that both sides of (3.7) are compact. Therefore, it suffices to verify that the support function of $\mathcal{Y}$ and the support function of the set on the right hand side of (3.7) are equal.

Let $b_0 \in \mathbb{R}^{n-1}$. Then,

$$\begin{aligned}
\inf_{y \in \mathcal{Y}} \langle b_0, y\rangle &= \frac{1}{2} \inf_{x \in \mathbb{R}^n} \left\{ x^\intercal \begin{pmatrix} 0 & b_0^\intercal \\ b_0 & 0_{n-1} \end{pmatrix} x :\ \begin{array}{l} x^\intercal M_i x \geq 0,\ \forall i \in [m] \\ x^\intercal (e_1 e_1^\intercal) x = 1 \end{array} \right\} \\
&= \frac{1}{2} \inf_{X \in \mathbb{S}^n} \left\{ \left\langle \begin{pmatrix} 0 & b_0^\intercal \\ b_0 & 0_{n-1} \end{pmatrix}, X \right\rangle :\ \begin{array}{l} \langle M_i, X\rangle \geq 0,\ \forall i \in [m] \\ \langle e_1 e_1^\intercal, X\rangle = 1 \\ X \succeq 0 \end{array} \right\} \\
&= \inf_{y \in \mathbb{R}^{n-1}} \left\{ \langle b_0, y\rangle :\ \begin{array}{l} \exists Y \succeq yy^\intercal : \\ \langle A_i, Y\rangle + 2\langle b_i, y\rangle + c_i \geq 0,\ \forall i \in [m] \end{array} \right\}.
\end{aligned}$$

Here, the first equality follows by writing $x = (1,\, y)$, the second equality follows by Lemma 38, and the third equality follows by writing $X = \left(\begin{smallmatrix} 1 & y^\intercal \\ y & Y \end{smallmatrix}\right)$. ■

We next turn our attention to the closed convex hull of epigraph sets. Let q_0 be a quadratic function of the form $q_0(y) = y^\intercal A_0 y + 2b_0^\intercal y + c_0$ and define $M_0 := \left(\begin{smallmatrix} c_0 & b_0^\intercal \\ b_0 & A_0 \end{smallmatrix}\right)$.

Proposition 16. *Suppose there exists $\lambda^* \in \mathbb{R}^m_+$ such that $A_0 - \sum_{i=1}^m \lambda_i^* A_i$ is positive definite. If $\mathcal{S}(\mathcal{M})$ is ROG, then the closed convex hull of*

$$\mathrm{epi} := \left\{ (y,t) \in \mathbb{R}^{n-1} \times \mathbb{R} : \begin{array}{l} q_0(y) \le t \\ y \in \mathcal{Y} \end{array} \right\}$$

is a semidefinite-representable set given by

$$\mathrm{clconv}(\mathrm{epi}) = \left\{ (y,t) \in \mathbb{R}^{n-1} \times \mathbb{R} : \begin{array}{l} \exists Y \succeq yy^\intercal : \\ \langle A_0, Y\rangle + 2\langle b_0, y\rangle + c_0 \le t \\ \langle A_i, Y\rangle + 2\langle b_i, y\rangle + c_i \ge 0,\ \forall i \in [m] \end{array} \right\}.$$

Proof. Let $\mathcal{R}$ denote the set on the right.

($\subseteq$) By taking $Y = yy^\intercal$, we have that epi $\subseteq \mathcal{R}$. It suffices to show that $\mathcal{R}$ is both convex and closed. As $\mathcal{R}$ is the projection of the SDP relaxation (a convex set) of epi, it is itself convex. Next, consider a sequence $(y^{(i)}, t^{(i)}) \in \mathcal{R}$ converging to (y,t). Let $Y^{(i)}$ denote a sequence of matrices certifying $(y^{(i)}, t^{(i)}) \in \mathcal{R}$. As there exists a $\lambda^* \in \mathbb{R}^m_+$ such that $A_0 - \sum_{i=1}^m \lambda_i^* A_i$ is positive definite, the sequence $Y^{(i)}$ is bounded and hence has a convergent subsequence with limit Y. By continuity, we deduce that $(y,t) \in \mathcal{R}$ and hence $\mathcal{R}$ is closed.

($\supseteq$) Suppose $(y,t) \notin$ clconv(epi). We will show that $(y,t) \notin \mathcal{R}$.

First, we claim that $q_0(y)$ is bounded below on $\mathcal{Y}$. Let $A^* := A_0 - \sum_{i=1}^m \lambda_i^* A_i$ and similarly define b^* and c^*. Then, for all $y \in \mathcal{Y}$, we have

$$q_0(y) \ge q_0(y) - \sum_{i=1}^m \lambda_i^* q_i(y) = y^\intercal A^* y + 2\langle b^*, y\rangle + c^* \ge -(b^*)^\intercal (A^*)^{-1} b^* + c^*.$$

We deduce that $q_0(y)$ is bounded below on $\mathcal{Y}$.

By the strict hyperplane separation theorem, there exists $(\mu, \nu) \ne (0,0) \in \mathbb{R}^{n-1} \times \mathbb{R}$ such that

$$\langle \mu, y\rangle + \nu t < \inf_{(y',t') \in \mathrm{clconv}(\mathrm{epi})} \langle \mu, y'\rangle + \nu t' = \inf_{(y',t') \in \mathrm{epi}} \langle \mu, y'\rangle + \nu t'. \tag{3.8}$$

We claim that we may assume $\nu > 0$ without loss of generality. First, suppose $\mathcal{Y} = \varnothing$. In this case, epi $= \varnothing$ and any arbitrary $(\mu, \nu) \ne (0,0)$ satisfies (3.8). On the other hand, if $\mathcal{Y}$ is nonempty then e_n is a recessive direction for epi. In particular, as the objective value of the program on the right is finite (by the bound on the left), we deduce that $\nu \ge 0$. Finally, as $q_0(y)$ is bounded below on $\mathcal{Y}$, we may increase ν by some positive amount without affecting (3.8).

Then,

$$\begin{aligned}\langle \mu, y\rangle + \nu t &< \min_{y'}\{\langle \mu, y'\rangle + \nu q_0(y') :\, y' \in \mathcal{Y}\}\\ &= \min_{y',Y'}\left\{\langle \mu, y'\rangle + \nu(\langle A_0, Y'\rangle + 2\langle b_0, y'\rangle + c_0) :\, \begin{array}{l} Y' \succeq y'y'^\intercal \\ \langle A_i, Y'\rangle + 2\langle b_i, y'\rangle + c_i \geq 0,\, \forall i \in [m]\end{array}\right\}\\ &\leq \min_{Y}\left\{\langle \mu, y\rangle + \nu(\langle A_0, Y\rangle + 2\langle b_0, y\rangle + c_0) :\, \begin{array}{l} Y \succeq yy^\intercal \\ \langle A_i, Y\rangle + 2\langle b_i, y\rangle + c_i \geq 0,\, \forall i \in [m]\end{array}\right\}.\end{aligned}$$

Here, the first line follows by substituting the optimal value of t' in (3.8), the second line follows from Lemma 38 (which we can apply as $\mathcal{S}(\mathcal{M})$ is ROG and the SDP on the second line has finite objective value), and the third line follows by selecting $y' = y$.

Subtracting $\langle \mu, y\rangle$ from both sides and dividing by $\nu > 0$ leads to the desired conclusion that $(y, t) \notin \mathcal{R}$ and completes the proof. ■

Applying a perturbation argument to Proposition 16 allows us to additionally relax the assumption that $A_0 - \sum_{i=1}^m \lambda_i^* A_i$ is positive definite.

Corollary 18. *Suppose there exists $\lambda^* \in \mathbb{R}_+^m$ such that $A_0 - \sum_{i=1}^m \lambda_i^* A_i$ is positive semidefinite. If $\mathcal{S}(\mathcal{M})$ is ROG, then the closed convex hull of*

$$\text{epi} := \left\{(y,t) \in \mathbb{R}^{n-1} \times \mathbb{R} :\, \begin{array}{l} q_0(y) \leq t \\ y \in \mathcal{Y}\end{array}\right\}$$

is the closure of a semidefinite-representable set:

$$\text{clconv}(\text{epi}) = \text{cl}\left(\left\{(y,t) \in \mathbb{R}^{n-1} \times \mathbb{R} :\, \begin{array}{l} \exists Y \succeq yy^\intercal : \\ \langle A_0, Y\rangle + 2\langle b_0, y\rangle + c_0 \leq t \\ \langle A_i, Y\rangle + 2\langle b_i, y\rangle + c_i \geq 0,\, \forall i \in [m]\end{array}\right\}\right).$$

Proof. Let $\mathcal{R}$ denote the set inside the right hand side so that the desired conclusion is $\text{clconv}(\text{epi}) = \overline{\mathcal{R}}$.

($\subseteq$) This direction follows simply from observing that epi $\subseteq \mathcal{R}$ and that $\mathcal{R}$ is convex.

($\supseteq$) Let $(\hat{y}, \hat{t}) \in \mathcal{R}$ and let $\hat{Y}$ be a matrix certifying $(\hat{y}, \hat{t}) \in \mathcal{R}$. It suffices to show that $(\hat{y}, \hat{t} + \epsilon) \in \text{clconv}(\text{epi})$ for all $\epsilon > 0$. Let $A_0' := A_0 + \delta I$ where we have set $\delta := \epsilon / \text{tr}(\hat{Y})$. Define $q_0'(y) := q_0(y) + \delta\|y\|^2 = y^\intercal A_0' y + 2\langle b_0, y\rangle + c_0$. Note that by construction,

$$\left\langle A_0', \hat{Y}\right\rangle + 2\langle b_0, \hat{y}\rangle + c_0 = \left(\left\langle A_0, \hat{Y}\right\rangle + 2\langle b_0, \hat{y}\rangle + c_0\right) + \epsilon \leq \hat{t} + \epsilon$$

so that

$$(\hat{y}, \hat{t} + \epsilon) \in \left\{(y,t) \in \mathbb{R}^{n-1} \times \mathbb{R} :\, \begin{array}{l} \exists Y \succeq yy^\intercal : \\ \langle A_0', Y\rangle + 2\langle b_0, y\rangle + c_0 \leq t \\ \langle A_i, Y\rangle + 2\langle b_i, y\rangle + c_i \geq 0,\, \forall i \in [m]\end{array}\right\}.$$

Next, as $\mathcal{S}(\mathcal{M})$ is ROG and $A_0' - \sum_{i=1}^m \lambda_i^* A_i = (A_0 - \sum_{i=1}^m \lambda_i^* A_i) + \delta I$ is positive definite, we may apply Proposition 16 with $q_0'(y)$ to deduce that

$$\begin{aligned}(\hat{y}, \hat{t} + \epsilon) &\in \operatorname{clconv}\left(\left\{(y,t) : \begin{array}{l} q_0'(y) \leq t \\ y \in \mathcal{Y} \end{array}\right\}\right)\\ &\subseteq \operatorname{clconv}(\mathrm{epi}).\end{aligned}$$

Here, the second containment follows by noting that $q_0(y) \leq q_0'(y)$ for all y. ■

The following example shows how to recover [180, Theorem 4] as an immediate corollary of Lemma 30 and Corollary 18.

Example 12. Consider a set $\mathcal{Y}$ defined by a single quadratic inequality constraint

$$\mathcal{Y} = \left\{y \in \mathbb{R}^{n-1} : \, q_1(y) \geq 0\right\}.$$

The associated cone $\mathcal{S}(\{M_1\})$ is ROG by Lemma 30. Next, suppose $q_0(x)$ is a quadratic objective function for which there exists $\lambda \geq 0$ such that $A_0 - \lambda A_1 \succeq 0$. Then, Corollary 18 implies that

$$\begin{aligned}&\operatorname{clconv}\left(\left\{(y,t) \in \mathbb{R}^{n-1} \times \mathbb{R} : \begin{array}{l} q_0(y) \leq t \\ q_1(y) \geq 0 \end{array}\right\}\right)\\ &\quad = \operatorname{cl}\left(\left\{(y,t) \in \mathbb{R}^{n-1} \times \mathbb{R} : \begin{array}{l} \exists Y \succeq yy^\intercal \, : \\ \langle A_0, Y\rangle + 2\langle b_0, y\rangle + c_0 \leq t \\ \langle A_1, Y\rangle + 2\langle b_1, y\rangle + c_1 \geq 0 \end{array}\right\}\right). \qquad \square\end{aligned}$$

We next examine a classical example related to the "perspective reformulation/relaxation" trick [42, 68, 79] and demonstrate how this convex hull result can be recovered using our ROG toolsets. The nonconvex set in this example will involve both binary and continuous variables and complementarity constraints.

Example 13. Consider the quadratically constrained set

$$\mathcal{Y} = \left\{y \in \mathbb{R}^2 : \begin{array}{l} (1-y_1)y_1 = 0 \\ (1-y_1)y_2 = 0 \end{array}\right\}.$$

In words, y_1 is constrained to be a binary variable, y_2 is allowed to be arbitrary when $y_1 = 1$ is "on" and forced to be zero when $y_1 = 0$ is "off."

Letting $M_1 \coloneqq \operatorname{Sym}((e_3 - e_1)e_1^\intercal)$ and $M_2 \coloneqq \operatorname{Sym}((e_3 - e_1)e_2^\intercal)$, we have that

$$\mathcal{Y} = \left\{y \in \mathbb{R}^2 : \begin{array}{l} \begin{pmatrix} y \\ 1 \end{pmatrix}^\intercal M_1 \begin{pmatrix} y \\ 1 \end{pmatrix} = 0 \\ \begin{pmatrix} y \\ 1 \end{pmatrix}^\intercal M_2 \begin{pmatrix} y \\ 1 \end{pmatrix} = 0 \end{array}\right\}.$$

Let $\mathcal{M} = \{M_1, M_2\}$ and note that $\mathcal{T}(\mathcal{M})$ is ROG by Corollary 17.

Next, we rewrite $\mathcal{Y}$ using inequality constraints so that we may apply Proposition 16. Letting $q_1(y) = (1-y_1)y_1$, $q_2(y) = -(1-y_1)y_1$, $q_3(y) = (1-y_1)y_2$, and $q_4(y) = -(1-y_1)y_2$, we may write

$$\mathcal{Y} = \left\{ y \in \mathbb{R}^2 : \quad q_i(y) \leq 0,\ \forall i \in [4] \right\}.$$

Note that $A_1 = -e_1e_1^\intercal$, $A_2 = e_1e_1^\intercal$, $A_3 = -\operatorname{Sym}(e_1e_2^\intercal)$, and $A_4 = \operatorname{Sym}(e_1e_2^\intercal)$. Setting $q_0(y) = y_2^2$, we have that $A_0 = e_2e_2^\intercal$. Then, as $A_0 + A_2 \succ 0$, we deduce that the assumptions of Proposition 16 hold. Applying Proposition 16 then gives

$$\begin{aligned}
&\operatorname{clconv}\left\{ (y,t) \in \mathbb{R}^2 \times \mathbb{R} : \begin{array}{l} y_2^2 \leq t \\ (1-y_1)y_1 = 0 \\ (1-y_1)y_2 = 0 \end{array} \right\} \\
&= \left\{ (y,t) \in \mathbb{R}^2 \times \mathbb{R} : \begin{array}{l} \exists Y \succeq yy^\intercal \\ Y_{2,2} \leq t \\ y_1 - Y_{1,1} = 0 \\ y_2 - Y_{1,2} = 0 \end{array} \right\} \\
&= \left\{ (y,t) \in \mathbb{R}^2 \times \mathbb{R} : \begin{pmatrix} y_1 & y_2 \\ y_2 & t \end{pmatrix} \succeq yy^\intercal \right\} \\
&= \left\{ (y,t) \in \mathbb{R}^2 \times \mathbb{R} : \begin{array}{l} y_1 \geq y_1^2 \\ t \geq y_2^2 \\ (y_1 - y_1^2)(t - y_2^2) \geq (y_2 - y_1y_2)^2 \end{array} \right\}.
\end{aligned}$$

Note that the first constraint in the last formulation implies that $y_1 \in [0,1]$. By expanding and rearranging, we can write the last constraint as

$$0 \leq (y_1 - y_1^2)(t - y_2^2) - (y_2 - y_1y_2)^2 = y_1t + y_1y_2^2 - y_1^2t - y_2^2 = (y_1t - y_2^2)(1 - y_1).$$

When $y_1 \in [0,1)$, this constraint is equivalent to $y_1t - y_2^2 \geq 0$. On the other hand when $y_1 = 1$, the constraint $y_1t - y_2^2 \geq 0$ is redundant. Hence, we deduce that

$$\operatorname{clconv}\left\{ (y,t) \in \mathbb{R}^2 \times \mathbb{R} : \begin{array}{l} y_2^2 \leq t \\ (1-y_1)y_1 = 0 \\ (1-y_1)y_2 = 0 \end{array} \right\} = \left\{ (y,t) \in \mathbb{R}^2 \times \mathbb{R} : \begin{array}{l} y_1 \in [0,1] \\ y_1t \geq y_2^2 \end{array} \right\}.$$

This gives the well-known perspective formulation of $\operatorname{clconv}(\mathcal{Y})$. □

Remark 45. There are few known sufficient conditions guaranteeing that the convex hull of the epigraph of a QCQP is given by its SDP relaxation. The conditions presented by Wang and Kılınç-Karzan [181, Theorems 1 and 7] are among the most general in this direction. We claim that both [181, Theorems 1 and 7] are incomparable with Proposition 16. Note that [181, Theorem 1] cannot be applied directly to Example 13: the set of *convex Lagrange multipliers* (see [181, Section 2.1]) for this example is

$$\begin{aligned}\Gamma &:= \left\{\gamma \in \mathbb{R}^2 :\ \begin{pmatrix} 0 & \\ & 1 \end{pmatrix} + \gamma_1 \begin{pmatrix} -1 & \\ & 0 \end{pmatrix} + \gamma_2 \begin{pmatrix} 0 & -1/2 \\ -1/2 & 0 \end{pmatrix} \succeq 0 \right\}\\ &= \left\{\gamma \in \mathbb{R}^2 :\ \gamma_1 \leq 0,\ |\gamma_2| \leq \sqrt{-\gamma_1}\right\},\end{aligned}$$

which is not polyhedral. On the other hand, [181, Theorem 1] can be applied to QCQPs where the A_is satisfy a "symmetry" condition. The following QCQP is such an example. Consider

$$\inf_{y\in\mathbb{R}^4} \left\{ \|y\|^2 :\ \begin{array}{l} y^\intercal \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & -1 & \\ & & & -1 \end{pmatrix} y + 1 \geq 0 \\ y^\intercal \begin{pmatrix} -2 & & & \\ & -2 & & \\ & & 1 & \\ & & & 1 \end{pmatrix} y + 1 \geq 0 \end{array} \right\}.$$

The corresponding set $\mathcal{M}$ for this example is $\mathcal{M} = \{\mathrm{Diag}(1, 1, -1, -1, 1), \mathrm{Diag}(-2, -2, 1, 1, 1)\}$. Theorem 16 implies that $\mathcal{S}(\mathcal{M})$ is not ROG and thus Proposition 16 cannot be applied to this example. We conclude that [181, Theorem 1] and Proposition 16 are incomparable. Similar examples can be constructed to show that [181, Theorem 7] and Proposition 16 are incomparable. □

3.5.3 Minimizing ratios of quadratic functions over ROG domains

In this section, we show how a "re-homegenization" trick can be combined with our toolset (specifically Lemma 38) to minimize the ratio of two quadratic functions over a ROG domain. Let $M_{\mathrm{obj}}, B \in \mathbb{S}^n$ and let $\mathcal{M} \subseteq \mathbb{S}^n$. We will consider the following optimization problem:

$$\inf_{\tilde{z}\in\mathbb{R}^n} \left\{ \frac{\tilde{z}^\intercal M_{\mathrm{obj}} \tilde{z}}{\tilde{z}^\intercal B \tilde{z}} :\ \begin{array}{l} \tilde{z}\tilde{z}^\intercal \in \mathcal{S}(\mathcal{M}) \\ \tilde{z}^\intercal B \tilde{z} > 0 \\ \tilde{z}_1^2 = 1 \end{array} \right\}. \tag{3.9}$$

Remark 46. The variant of (3.9) where the constraint $\tilde{z}^\intercal B \tilde{z} > 0$ is replaced with $\tilde{z}^\intercal B \tilde{z} \neq 0$ can be decomposed as two instances of (3.9) based on the sign of $\tilde{z}^\intercal B \tilde{z}$ (and negating both M_{obj} and B on the portion of the domain where $\tilde{z}^\intercal B \tilde{z}$ is negative). □

We derive an SDP relaxation to (3.9) as follows:

$$\begin{aligned}\inf_{\tilde{z}\in\mathbb{R}^n} \left\{ \frac{\tilde{z}^\intercal M_{\mathrm{obj}} \tilde{z}}{\tilde{z}^\intercal B \tilde{z}} :\ \begin{array}{l} \tilde{z}\tilde{z}^\intercal \in \mathcal{S}(\mathcal{M}) \\ \tilde{z}^\intercal B \tilde{z} > 0 \\ \tilde{z}_1^2 = 1 \end{array} \right\} &= \inf_{z\in\mathbb{R}^n} \left\{ z^\intercal M_{\mathrm{obj}} z :\ \begin{array}{l} zz^\intercal \in \mathcal{S}(\mathcal{M}) \\ z^\intercal B z = 1 \\ z_1^2 > 0 \end{array} \right\} && (3.10)\\ &\geq \inf_{z\in\mathbb{R}^n} \left\{ z^\intercal M_{\mathrm{obj}} z :\ \begin{array}{l} zz^\intercal \in \mathcal{S}(\mathcal{M}) \\ z^\intercal B z = 1 \end{array} \right\} && (3.11)\\ &\geq \inf_{Z\in\mathbb{S}^n} \left\{ \left\langle M_{\mathrm{obj}}, Z \right\rangle :\ \begin{array}{l} Z \in \mathcal{S}(\mathcal{M}) \\ \langle B, Z\rangle = 1 \end{array} \right\}. && (3.12)\end{aligned}$$

Lemma 38 implies that the inequality between (3.11) and (3.12) holds with equality whenever $\mathcal{S}(\mathcal{M})$ is ROG and (3.12) is bounded below. This boundedness holds under relatively minor assumptions. Similarly, a variety of different assumptions may be used to guarantee that the inequality relation between (3.10) and (3.11) holds with equality. The following lemma demonstrates one such pair of sufficient conditions.

Lemma 41. *Let $M_{\text{obj}}, B \in \mathbb{S}^n$ and $\mathcal{M} \subseteq \mathbb{S}^n$. Suppose $\mathcal{S}(\mathcal{M})$ is ROG, there exists $M^* \in \operatorname{clcone}(\mathcal{M})$ and $\lambda \in \mathbb{R}$ such that $M_{\text{obj}} + M^* + \lambda B \succeq 0$, and*

$$\operatorname{cl}\left\{z \in \mathbb{R}^n : \begin{array}{c} zz^\intercal \in \mathcal{S}(\mathcal{M}) \\ z_1^2 > 0 \end{array}\right\} = \{z \in \mathbb{R}^n : zz^\intercal \in \mathcal{S}(\mathcal{M})\}. \tag{3.13}$$

Then, equality holds throughout (3.10) to (3.12).

Example 14 (Regularized total least squares)**.** The total least squares problem (TLS) adapts least squares regression to the setting where both the independent and dependent variables may be corrupted by noise [76]. A variant of the TLS, known as the regularized total least squares problem (RTLS), introduces an additional regularization constraint that protects against poorly behaved solutions which arise when the data matrix has small singular values. This regularization is well studied from both theoretical and practical points of view (see [19, 189] and references therein).

By eliminating variables, the RTLS can be rewritten as minimizing the ratio of a nonnegative quadratic function and a positive quadratic function over a nonempty ellipsoid (see for example [76]). In particular, the RTLS can be written in the form of (3.9) where $M_{\text{obj}}, B \in \mathbb{S}^n_+$ and $|\mathcal{M}| = 1$. It is then straightforward to verify that the assumptions of (41) are satisfied so that the RTLS admits an exact SDP relaxation in the sense of objective value exactness. □

Example 15 (Stackelberg prediction game with least-squares loss)**.** In [186], we show that a Stackelberg prediction game with a least-squares loss function (SPG-LS) can be written in the form of (3.9). This game is played between a learner (us) and a number of data providers, who each come from some fixed but unknown distribution. We assume we have access to m-many tuples $\{(x_i, y_i, z_i)\}_{i=1}^m$ from this distribution where $x_i \in \mathbb{R}^n$ is the feature vector of the ith data provider, $y_i \in \mathbb{R}$ is the label that we would like to assign to the ith data provider and z_i is the label that the ith data provider would like us to assign to it. Our goal is then to perform least-squares regression where the data provider has some penalty (controlled by $\gamma > 0$) for "lying" or "altering" their data:

$$\min_{w \in \mathbb{R}^n}\left\{\sum_{i=1}^m (w^\intercal x_i^*(w) - y_i)^2 : x_i^*(w) \in \arg\min_{\tilde{x} \in \mathbb{R}^n} (w^\intercal \tilde{x} - z_i)^2 + \gamma\|\tilde{x} - x_i\|^2\right\}.$$

Here, the inner minimization problem says that the ith data provider (with full knowledge of w) chooses $x_i^*(w) \in \mathbb{R}^n$ to minimize its own loss function. Letting $X \in \mathbb{R}^{m \times n}$ denote the matrix with ith row $x_i^\intercal$, and $y,\ z \in \mathbb{R}^m$ the vectors with ith entry y_i and z_i, we may rewrite this problem more compactly as

$$\min_{w\in\mathbb{R}^n}\left\{\|X^*(w)w - y\|^2 :\ X^*(w) \in \operatorname*{arg\,min}_{\tilde{X}\in\mathbb{R}^{m\times n}}\left(\left\|\tilde{X}w - z\right\|^2 + \gamma\left\|\tilde{X} - X\right\|_F^2\right)\right\}.$$

Then, applying the Sherman-Morrison formula to solve for $X^*(w)$ and introducing a new variable $\alpha = \frac{w^\intercal w}{\gamma}$, we get

$$\begin{aligned}
&\min_{w\in\mathbb{R}^n}\left\|\frac{Xw + \frac{1}{\gamma}zw^\intercal w}{1 + \frac{1}{\gamma}w^\intercal w} - y\right\|^2 \\
&\quad = \min_{w\in\mathbb{R}^n,\,\alpha\in\mathbb{R}}\left\{\left\|\frac{Xw + \alpha z}{1+\alpha} - y\right\|^2 :\ \alpha = \frac{w^\intercal w}{\gamma}\right\} \\
&\quad = \min_{w\in\mathbb{R}^n,\,\alpha\in\mathbb{R}}\left\{\frac{\|Xw + \alpha z - (1+\alpha)y\|^2}{(1+\alpha)^2} :\ \alpha = \frac{w^\intercal w}{\gamma}\right\}.
\end{aligned}$$

Thus, we may solve the SPG-LS by solving its SDP relaxation (see Lemma 41). In [186], we go a step further and note that we may apply algorithms for the GTRS to the SPG-LS. Numerical results show that this method for solving the SPG-LS is orders of magnitudes faster than previous state-of-the-art algorithms from [28, 185] for the SPG-LS. □

Part II

What structures within a QCQP allow its SDP relaxation to be solved efficiently?

4 The Generalized Trust Region Subproblem: solution complexity and convex hull results

This chapter is based on joint work [180] with Fatma Kılınç-Karzan.

We consider the Generalized Trust Region Subproblem (GTRS) of minimizing a nonconvex quadratic objective over a nonconvex quadratic constraint. The epigraph representation of this problem recasts the GTRS as minimizing a linear objective subject to two nonconvex quadratic constraints. Our first main contribution is structural: we give an explicit description of the convex hull of this nonconvex set in terms of the generalized eigenvalues of an associated matrix pencil. This result may be of interest in building relaxations for nonconvex quadratic programs. Moreover, this result allows us to reformulate the GTRS as the minimization of two convex quadratic functions in the original space. Our next set of contributions is algorithmic: we present a first-order method for solving the GTRS up to an ϵ additive error based on this reformulation in $\approx O\left(\epsilon^{-1/2}\right)$ iterations. We carefully handle numerical issues that arise from inexact generalized eigenvalue and eigenvector computations and establish explicit running time guarantees for these algorithms. Notably, our algorithms run in *linear (in the size of the input) time*. Furthermore, our algorithm for computing an ϵ-optimal solution has a slightly-improved running time dependence on ϵ over the state-of-the-art algorithm. Our analysis shows that the dominant cost in solving the GTRS lies in solving a generalized eigenvalue problem—establishing a natural connection between these problems. Finally, generalizations of our convex hull results allow us to apply our algorithms and their theoretical guarantees directly to equality-, interval-, and hollow-constrained variants of the GTRS. This gives the first linear-time algorithm in the literature for these variants of the GTRS.

4.1 Introduction

In this chapter, we study the *Generalized Trust-Region Subproblem* (GTRS), which is defined as

$$\mathrm{Opt} := \inf_{x\in\mathbb{R}^n} \{q_0(x) :\ q_1(x) \leq 0\}, \tag{4.1}$$

where $q_0 : \mathbb{R}^n \to \mathbb{R}$ and $q_1 : \mathbb{R}^n \to \mathbb{R}$ are general quadratic functions of the form $q_i(x) = x^\intercal A_i x + 2b_i^\intercal x + c_i$. Here, $A_i \in \mathbb{R}^{n\times n}$ are symmetric matrices, $b_i \in \mathbb{R}^n$ and $c_i \in \mathbb{R}$. We are interested, in particular, in the case where q_0 and q_1 are both nonconvex, i.e., A_i has at least one negative eigenvalue for both $i = 0, 1$.

Problem (4.1), introduced and studied by Moré [124], Stern and Wolkowicz [165], generalizes the classical *Trust-Region Subproblem* (TRS) [50] in which one is asked to optimize a nonconvex quadratic objective over a Euclidean ball. The TRS is an essential ingredient of trust-region methods that are commonly used to solve continuous nonconvex optimization problems [50, 135, 148] and also arises in applications such as robust optimization [21, 87]. On the other hand, the GTRS has applications in nonconvex quadratic integer programs, signal processing, and compressed sensing; see [2, 32, 95] and references therein for more applications.

Although the TRS, as stated, is nonlinear and nonconvex, it is well-known that its semidefinite programming (SDP) relaxation is exact. Consequently, the TRS and a number of its variants can be solved in polynomial time via SDP-based techniques [66, 151] or using specialized nonlinear algorithms [77, 125]. In fact, custom iterative methods with linear (in the size of the input) running times have been shown in a few works. Hazan and Koren [82] proposed an algorithm to solve the TRS (as well as the GTRS when A_1 is positive definite) based on repeated approximate eigenvector computations. This algorithm runs in time

$$\tilde{O}\left(\frac{N\sqrt{\kappa_{\text{HK}}}}{\sqrt{\epsilon}}\log\left(\frac{n}{p}\right)\log\left(\frac{\kappa_{\text{HK}}}{\epsilon}\right)\right), \tag{4.2}$$

where N is the number of nonzero entries in the matrices A_0 and A_1, ϵ is the additive error, n is the dimension of the problem, p is the failure probability, and κ_{HK} is a condition number. This was the first algorithm in the literature shown to achieve a linear time complexity. Here, and in the remainder of the chapter, the term "linear" is used to describe running times that scale at most linearly with N but may depend arbitrarily on its other parameters. Afterwards, Ho-Nguyen and Kılınç-Karzan [87] presented another linear-time algorithm for the TRS with a slightly better overall complexity, eliminating the $\log(\kappa_{\text{HK}}/\epsilon)$ term. Their approach reformulates the TRS as minimizing a *convex* quadratic objective over the Euclidean ball, and solving the resulting smooth convex optimization problem via Nesterov's accelerated gradient descent method. In contrast to [82], this convex reformulation approach requires only a single minimum eigenvalue computation. Wang and Xia [184] also suggested using Nesterov's algorithm in the case of the interval-constrained TRS.

The GTRS shares a number of nice properties of the TRS. For example, by the S-lemma, it is well-known that the GTRS also admits an exact SDP reformulation under the Slater condition [67, 146]. Thus, while quadratically-constrained quadratic programming is NP-hard in general, there are polynomial-time SDP-based algorithms for solving the GTRS. Nevertheless, the relatively large computational complexity of SDP-based algorithms prevents them from being applied as a black box to solve large-scale instances of the GTRS. A variety of custom approaches have been developed to solve the GTRS; for earlier work on this domain see [65, 124, 165] and references therein.

One line of work has developed algorithms for solving the GTRS when the matrices A_0 and A_1 are simultaneously diagonalizable (SD) (see Jiang and Li [93] and references therein for background on the SD condition). Under the SD condition, along with certain restrictions on the quadratics q_0 and q_1, Ben-Tal and Teboulle [24] provide a reformulation of the interval-constrained GTRS as a convex minimization problem with linear constraints. More recently, Ben-Tal and den Hertog [21] show that there is a second order cone programming (SOCP) reformulation of the GTRS in a lifted space under the SD condition. Subsequent work by Locatelli [111] extends Ben-Tal and den Hertog [21] by illustrating some additional settings in which the SOCP reformulation is tight. Under the SD condition, Fallahi et al. [64] exploit the separable structure of the problem and, using Lagrangian duality, they suggest a solution procedure based on solving a univariate convex minimization problem. Salahi and Taati [155] derive an algorithm for solving the interval-constrained GTRS by exploiting the structure of the dual problem under the SD condition. By applying a simultaneous block diagonalization approach, Jiang et al. [96] generalize Ben-Tal and den Hertog [21] and provide an SOCP reformulation for the GTRS in a lifted space when the problem has a finite optimal value. Their methods apply even when q_0 and q_1 do *not* satisfy the SD condition. They further derive a closed-form solution when the SD condition fails and examine the case of interval- or equality-constrained GTRS. In this line of work, it is often assumed implicitly that A_0 and A_1 are already diagonal or that a simultaneously-diagonalizing basis can be computed. The only method that we know of for computing such a basis relies on exact matrix eigen-decomposition. Thus, although experiments have been presented [96, 155] suggesting that such algorithms (where exact procedures are replaced by numerical ones) may perform well, theoretical guarantees have yet to be established. Furthermore, the large cost of matrix eigen-decomposition prevents the application of these algorithms to large-scale instances of the GTRS.

A second line of work has explored the connections between the GTRS and generalized eigenvalues of the matrix pencil $A_0 + \gamma A_1$. These works all assume a *regularity condition* about the matrix pencil: there exists a $\gamma \geq 0$ such that $A_0 + \gamma A_1$ is either positive definite or positive semidefinite.[1] Pong and Wolkowicz [148] study the optimality structure of the GTRS and propose a generalized-eigenvalue-based algorithm which exploits this structure. Unfortunately, an explicit running time is not presented in [148]. Adachi and Nakatsukasa [2] present another generalized-eigenvalue-based algorithm motivated by similar observations. The dominant costs present in this algorithm come from computing a pair of generalized eigenvalues and solving a linear system. Ignoring issues of exact computations, the runtime of this algorithm is $O(n^3)$. Jiang and Li [94] show how to reformulate the GTRS as a convex quadratic program in terms of generalized eigenvalues. They establish that a saddle-point-based first-order algorithm can be used to solve the reformulation within an ϵ additive error in $O(1/\epsilon)$ time. In this line of work, it is often assumed that the generalized eigenvalues are given or can be computed exactly. In particular, theoretical guarantees have not yet been given regarding how these algorithms perform when only approximate generalized eigenvalue computations are available. This is of interest as, in practice, we cannot hope to numerically compute generalized eigenvalues exactly; see also the discussion in Section 4.1 in [95]. We would like to remark that numerical experiments in these papers [2, 94, 148] have suggested that algorithms motivated by these ideas may perform well even using only approximate generalized eigenvalue computations.

The very recent work of Jiang and Li [95] presents an algorithm for solving the GTRS up to an ϵ additive error in the objective with high probability under the regularity condition. This algorithm relies on machinery developed by [82] for solving the TRS and differs from previous algorithms in

[1] In fact, this assumption can be made without loss of generality; see Remark 49.

that it does not assume the ability to compute a simultaneously-diagonalizing basis or generalized eigenvalues. The running time of this algorithm is

$$\tilde{O}\left(\frac{N\phi^3}{\sqrt{\epsilon\,\xi_{\mathrm{JL}}^5}}\log\left(\frac{n}{p}\right)\log\left(\frac{\phi}{\epsilon\,\xi_{\mathrm{JL}}}\right)^2\right), \tag{4.3}$$

where N is the number of nonzero entries in A_0 and A_1, ϵ is the additive error, n is the dimension, p is the failure probability, and $(\phi, \xi_{\mathrm{JL}})$ are a pair of parameters measuring the regularity of the GTRS. In particular, this algorithm is able to take advantage of sparsity in the description of the quadratic functions. To our knowledge, this is the first provably linear-time algorithm for the GTRS to be presented in the literature.

In this chapter, we derive a new algorithm for the GTRS based on a convex quadratic reformulation in the original space. This algorithm can also be applied to variants of the GTRS with interval, equality, or hollow constraints. The basic idea in our approach relies on the fact that we can provide exact (closed) convex hull characterizations of the epigraph of the GTRS. We summarize our results below and provide an outline of the chapter.

i) We rewrite the GTRS with a linear objective

$$\mathrm{Opt} = \inf_{(x,t)}\{t :\ (x,t) \in \mathcal{S}\}, \tag{4.4}$$

where the set $\mathcal{S}$ is defined as

$$\mathcal{S} := \left\{(x,t) \in \mathbb{R}^{n+1} :\ \begin{array}{l} q_0(x) \le t \\ q_1(x) \le 0 \end{array}\right\}. \tag{4.5}$$

As the objective in (4.4) is linear, we can take either the convex hull or closed convex hull of the feasible domain. Then,

$$\mathrm{Opt} = \inf_{x,t}\{t :\ (x,t) \in \mathrm{conv}(\mathcal{S})\} = \inf_{x,t}\{t :\ (x,t) \in \overline{\mathrm{conv}}(\mathcal{S})\}.$$

In Section 4.2, we give an explicit description of the set $\mathrm{conv}(\mathcal{S})$ (respectively, $\mathrm{clconv}(\mathcal{S})$). Specifically, we show that when the respective assumptions are satisfied, $\mathrm{conv}(\mathcal{S})$ and $\mathrm{clconv}(\mathcal{S})$ can both be described in terms of two convex quadratic functions determined by the generalized eigenvalue structure of the matrix pencil $A_0 + \gamma A_1$. We note that these convex hull results may be of independent interest in building relaxations and/or algorithms for nonconvex quadratic programs with or without integer variables.

Remark 47. These convex hull results are stated in strictly more general terms in Chapters 1 and 2. □

As an immediate consequence of these (closed) convex hull results, we can reformulate the GTRS as the minimization of the maximum of two convex quadratics. This convex reformulation was previously discovered by Jiang and Li [94] by considering the Lagrangian dual and proving a zero duality gap. Our approach shows that the reformulation is tight for a very intuitive reason — the convex hull of the epigraph is exactly characterized by the convex quadratics used in the reformulation.

ii) The proofs in Section 4.2 actually imply stronger convex hull results: under the same assumptions, the (closed) convex hull of $\mathcal{S}$ is generated by points in $\mathcal{S}$ where the constraint $q_1(x) \leq 0$ is tight. This observation immediately leads to interesting consequences, which we detail in Section 4.3. Specifically, we extend our (closed) convex hull results to handle epigraph sets that arise when additional nonintersecting constraints are imposed on the GTRS. This will allow us to extend our algorithms to variants of the GTRS present in the literature [21, 24, 87, 94, 96, 124, 148, 155, 165, 191]. Specifically, this generalization allows us to handle interval-, equality-, and hollow-constrained GTRS.

iii) In Section 4.4, we give a careful analysis of the numerical issues that come up for an algorithm based on the above ideas. At a high level, we show that by approximating the generalized eigenvalues sufficiently well, the perturbed convex reformulation is within a small additive error of the true convex reformulation. Then, by leveraging the concavity of the function $\lambda_{\min}(A_0+\gamma A_1)$, in the variable γ, we show how to approximate the necessary generalized eigenvalues efficiently. We believe this subroutine and the theoretical guarantees we present for it may also be of independent interest in other contexts. Next, we utilize an algorithm proposed by Nesterov [132, Section 2.3.3] for solving general minimax problems with smooth components to solve our convex reformulation with a convergence rate of $\tilde{O}(1/\sqrt{\epsilon})$. This contrasts the approach taken by Jiang and Li [94] that analyzes a saddle-point-based first-order algorithm and results in a convergence rate of $O(1/\epsilon)$. In order to apply the algorithm proposed by Nesterov, we establish that the gradient mapping step can be computed efficiently in our context. Finally, relying on our convex hull characterization, we show how to recover an approximate solution of the GTRS using only approximate eigenvectors.

We present two algorithms (Algorithms 1 and 4). The former finds an ϵ-optimal value and the latter finds an ϵ-optimal feasible solution. In other words, the former returns a scalar in $[\mathrm{Opt}, \mathrm{Opt}+\epsilon]$ and the latter returns a vector x in the feasible region with $q_0(x) \in [\mathrm{Opt}, \mathrm{Opt}+\epsilon]$. Their running times are

$$\tilde{O}\left(\frac{N\kappa^{3/2}\sqrt{\zeta}}{\sqrt{\epsilon}}\log\left(\frac{n}{p}\right)\log\left(\frac{\kappa}{\epsilon}\right)\right),\quad \tilde{O}\left(\frac{N\kappa^{2}\sqrt{\zeta}}{\sqrt{\epsilon}}\log\left(\frac{n}{p}\right)\log\left(\frac{\kappa}{\epsilon}\right)\right), \tag{4.6}$$

respectively. Here, ξ, ζ, and κ are regularity parameters of the matrix pencil $A_0 + \gamma A_1$ (see Definition 17). Comparing (4.6) and (4.2), we see that our running times match the dependences on N, n, ϵ, and p from the algorithm for the TRS presented by Hazan and Koren [82]. Comparing (4.6) (specifically the running time for finding an ϵ-optimal solution) and (4.3), we see that our running time matches the linear dependence on N and improves the dependence on ϵ by a logarithmic factor from the running time presented by Jiang and Li [95]. The dependences on the regularity parameters in the two running times are incomparable (see Remark 56) but there exist examples where our running time gives a polynomial-order improvement upon the running time presented by Jiang and Li [95] (see Remark 60).

In comparison to the approach taken by Jiang and Li [95], we believe our approach is conceptually simpler and more straightforward to implement. In particular our approach directly solves the GTRS in the primal space as opposed to solving a feasibility version of the dual problem. Moreover, our analysis highlights the connection between the GTRS and generalized eigenvalue problems, and in fact demonstrates that the dominant cost in solving the GTRS is the cost of solving a generalized eigenvalue problem.

In our running times (4.6), the large dependence on the regularity parameters arises from the error that is introduced as a result of inexact generalized eigenvalue and eigenvector computations. We illustrate that our algorithms can be substantially sped up if we have access to exact generalized eigenvalue and eigenvector methods. In particular, we show that when A_0 and A_1 are diagonal, we can compute an ϵ-optimal solution to the GTRS in time

$$O\left(\frac{N\kappa\sqrt{\zeta}}{\sqrt{\epsilon}}\right).$$

As mentioned previously, the generalizations of our convex hull results allow us to apply our algorithms to variants of the GTRS. In particular, our algorithms can be applied without change to interval-, equality-, or hollow-constrained GTRS.

Remark 48. In Chapter 5, we will see a second *faster* algorithm for the GTRS that works under a stronger assumption. □

Our study of the convex hull of the epigraph of GTRS is inspired by convex hull results in related contexts. The recent work of Ho-Nguyen and Kılınç-Karzan [87] gives a characterization on the convex hull of the epigraph of the TRS. In particular, under the assumption that A_1 is positive definite, Ho-Nguyen and Kılınç-Karzan [87, Theorem 3.5] give the explicit closed convex hull characterization of the set $\mathcal{S}$. In this respect, one can view our developments on the (closed) convex hull of $\mathcal{S}$ when neither A_0 nor A_1 is positive semidefinite as complementary to the results of Ho-Nguyen and Kılınç-Karzan [87, Section 3]. Notably, in contrast to [87, Section 3], we have to handle a number of issues that arise due to the recessive directions of the nonconvex domain. The papers by Modaresi and Vielma [123], Yıldıran [195] are also closely related to our convex hull results. Yıldıran [195] studies the convex hull of the intersection of two *strict* quadratic inequalities (note that the resulting set is open) under the milder regularity condition that there exists $\gamma \geq 0$ such that $A_0 + \gamma A_1$ is positive semidefinite, and Modaresi and Vielma [123] analyze conditions under which one can safely take the closure of the sets in Yıldıran [195] and still obtain the desired closed convex hull results. In contrast, our analysis leverages the additional structure present in an epigraph set to give a more direct proof of the convex hull result. Furthermore, as our analysis is constructive (given $x \in \operatorname{conv}(\mathcal{S})$, we show how to find two points $x_1, x_2 \in \mathcal{S}$ such that $x \in [x_1, x_2]$), it immediately suggests a rounding procedure (given a solution to the convex reformulation, we show how to find a solution to the original GTRS). This contrasts the analysis in Yıldıran [195], where such a rounding procedure is not obvious. Moreover, our analysis provides a more refined result that easily extends to variants of the GTRS with non-intersecting constraints. Finally, we would like to mention related work on convex hulls of sets defined by second-order cones (SOCs). Burer and Kılınç-Karzan [35] study the convex hull of the intersection of a convex and nonconvex quadratic or the intersection of an SOC with a nonconvex quadratic. Similarly, the convex hull of the a two-term disjunction applied to an SOC or its cross section has received much attention (see [35, 101] and references therein). As our focus has been on the case where neither A_0 nor A_1 is positive semidefinite, we view our developments as complementary to these results.

4.1.1 Additional notation

Given $A \in \mathbb{S}^n$, let $\lambda_{\max}(A)$ denote the maximum eigenvalue of A. Let $\|A\|$ denote the spectral norm of A. For $x \in \mathbb{R}^n$ and $r \geq 0$, let $B(x, r)$ be the closed ball of radius r centered at x, i.e., $B(x, r) = \{y \in \mathbb{R}^n : \|x - y\| \leq r\}$.

4.2 Convex hull characterization

In this section we discuss our (closed) convex hull results. We will aggregate the objective function q_0 with the constraint q_1 using a nonnegative aggregation weight to derive relaxations of the set $\mathcal{S}$. We then show that under a mild assumption the (closed) convex hull of $\mathcal{S}$ can be described by two convex quadratic functions obtained from this aggregation technique.

Let $q : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ be defined as

$$q(\gamma, x) \coloneqq q_0(x) + \gamma q_1(x).$$

Let $A : \mathbb{R} \to \mathbb{R}^{n \times n}$ be defined as $A(\gamma) \coloneqq A_0 + \gamma A_1$. Similarly define $b(\gamma)$ and $c(\gamma)$. In particular, $q(\gamma, x) = x^\intercal A(\gamma) x + 2b(\gamma)^\intercal x + c(\gamma)$. We stress that while $q(0, x) = q_0(x)$, we have $q(1, x) = q_0(x) + q_1(x)$ which is not equal to $q_1(x)$ in general.

Note that $q(\gamma, x)$ is linear in its first argument and quadratic in its second argument. This structure plays a large role in our analysis.

In order to derive valid relaxations to $\mathcal{S}$ based on aggregation, we will consider only nonnegative γ in the remainder of the chapter. For $\gamma \geq 0$, define

$$\mathcal{S}(\gamma) \coloneqq \left\{(x, t) \in \mathbb{R}^{n+1} : \; q(\gamma, x) \leq t\right\}.$$

Note that $\mathcal{S} \subseteq \mathcal{S}(\gamma)$ holds for all $\gamma \geq 0$. Furthermore, it is clear that $q_0(x) \leq t$ and $q_1(x) \leq 0$ if and only if $q(\gamma, x) \leq t$ for all $\gamma \geq 0$. Thus, we can rewrite $\mathcal{S}$ as

$$\mathcal{S} \coloneqq \left\{(x, t) \in \mathbb{R}^{n+1} : \begin{array}{c} q_0(x) \leq t \\ q_1(x) \leq 0 \end{array}\right\} = \bigcap_{\gamma \geq 0} \mathcal{S}(\gamma).$$

Note that the set $\mathcal{S}(\gamma)$ is convex if and only if $A(\gamma) \succeq 0$. We will define Γ to be these γ values, i.e.,

$$\Gamma \coloneqq \{\gamma \in \mathbb{R}_+ : \; A(\gamma) \succeq 0\}.$$

Note that Γ is a closed (possibly empty) interval. When this interval is nonempty, we will write it as $\Gamma = [\gamma_-, \gamma_+]$.

We use the following two assumptions in our convex hull characterizations:

Assumption 8. The matrices A_0 and A_1 both have negative eigenvalues and there exists a $\gamma^* \geq 0$ such that $A(\gamma^*) \succ 0$. □

Assumption 9. The matrices A_0 and A_1 both have negative eigenvalues and there exists a $\gamma^* \geq 0$ such that $A(\gamma^*) \succeq 0$. □

Remark 49. We claim that the case where A_0 and A_1 both have negative eigenvalues but do not satisfy either of the above assumptions is not interesting. In particular if A_0 and A_1 both have negative eigenvalues and $A(\gamma) \not\succeq 0$ for all $\gamma \geq 0$, then it is easy to show (apply the S-lemma then note that A_0 has a negative eigenvalue) that $\operatorname{conv}(\mathcal{S}) = \mathbb{R}^{n+1}$. Consequently, the optimal value of the GTRS is always $-\infty$ in this case.

The assumption that there exists a $\gamma^* \geq 0$ such that $A(\gamma^*) \succeq 0$ is made in most of the present literature on the GTRS [2, 21, 93–95, 148, 155] and convex hulls of the intersection of two quadratics [123, 195] either implicitly (for example, by assuming that an optimizer exists or that the optimal value is finite) or explicitly.

It is well-known that Assumption 8 implies that A_0 and A_1 are simultaneously diagonalizable. Even so, we will refrain from assuming that our matrices are diagonal and opt to work on a general basis. We choose to do this as the proofs of our convex hull results will serve as the basis for our algorithms, which do not have access to a simultaneously-diagonalizing basis. □

Remark 50. Assumptions 8 and 9 each imply that Γ is nonempty and, consequently, that γ_- and γ_+ exist. In addition, as $A(\gamma_-)$ and $A(\gamma_+)$ are both on the boundary of the positive semidefinite cone, they both have zero as an eigenvalue.

Under Assumption 8, the existence of some $\gamma^* \geq 0$ such that $A(\gamma^*) \succ 0$ implies that $\gamma_- < \gamma^* < \gamma_+$ and hence γ_- and γ_+ are distinct. Furthermore, as $\gamma^* \in (\gamma_-, \gamma_+)$, we have $d^\intercal A(\gamma_-)d = d^\intercal A(\gamma_+)d = 0$ if and only if $d = 0$.

In contrast, under Assumption 9, it is possible to have $\gamma_- = \gamma^* = \gamma_+$ and $\Gamma = \{\gamma^*\}$. □

Finally, define $\mathfrak{S}$ to be the subset of $\mathcal{S}$ where the constraint $q_1(x) \leq 0$ is tight.

$$\mathfrak{S} := \left\{ (x,t) \in \mathbb{R}^{n+1} : \begin{array}{c} q_0(x) \leq t \\ q_1(x) = 0 \end{array} \right\}.$$

When either Assumption 8 or 9 holds, A_1 has both positive and negative eigenvalues so that $\mathfrak{S}$ is nonempty.

We now state our (closed) convex hull results:

Theorem 19. *Under Assumption 8, we have*

$$\operatorname{conv}(\mathcal{S}) = \operatorname{conv}(\mathfrak{S}) = \mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+).$$

In particular,

$$\min_{x \in \mathbb{R}^n} \{q_0(x) : q_1(x) \leq 0\} = \min_{x \in \mathbb{R}^n} \max\{q(\gamma_-, x), q(\gamma_+, x)\}.$$

Theorem 20. *Under Assumption 9, we have*

$$\operatorname{clconv}(\mathcal{S}) = \operatorname{clconv}(\mathfrak{S}) = \mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+).$$

In particular,

$$\inf_{x \in \mathbb{R}^n} \{q_0(x) : q_1(x) \leq 0\} = \inf_{x \in \mathbb{R}^n} \max\{q(\gamma_-, x), q(\gamma_+, x)\}.$$

Remark 51. The convex reformulation given in the second part of Theorem 20 was first proved by Jiang and Li [94] using a different argument without relying on the convex hull structure of the underlying sets. In contrast, the first part of Theorem 20 establishes a fundamental convex hull result highlighting the crux of why such a convex reformulation is possible. □

We present the proof of Theorem 19 in Section 4.2.1. The proof of Theorem 20 is presented in Section 4.2.2 and relies on Theorem 19.

4.2.1 Proof of Theorem 19

Lemma 42. *The set $\mathcal{S}(\gamma)$ is convex and closed for all $\gamma \in \Gamma$.*

Proof. Let $\gamma \in \Gamma$ and recall the definition of $\mathcal{S}(\gamma)$.

$$\begin{aligned} \mathcal{S}(\gamma) &:= \left\{(x,t) \in \mathbb{R}^{n+1} : q(\gamma, x) \leq t\right\} \\ &= \left\{(x,t) \in \mathbb{R}^{n+1} : x^\intercal A(\gamma) x + 2b(\gamma)^\intercal x + c(\gamma) \leq t\right\} \end{aligned}$$

By the definition of Γ, we have $A(\gamma) \succeq 0$. Thus, the constraint defining $\mathcal{S}(\gamma)$ is convex in (x,t), and we conclude that $\mathcal{S}(\gamma)$ is convex. Closedness of $\mathcal{S}(\gamma)$ follows by noting that it is the preimage of $(-\infty, 0]$ under a continuous map. ■

Lemma 43. *Suppose Γ is nonempty and write $\Gamma = [\gamma_-, \gamma_+]$. Then,* $\operatorname{conv}(\mathcal{S}) \subseteq \mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+)$.

Proof. Note that $\mathcal{S} = \bigcap_{\gamma \geq 0} \mathcal{S}(\gamma) \subseteq \mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+)$. The result then follows by taking the convex hull of each side and noting that both $\mathcal{S}(\gamma_-)$ and $\mathcal{S}(\gamma_+)$ are convex by Lemma 42. ■

The bulk of the work in proving Theorem 19 lies in the following result.

Lemma 44. *Under Assumption 8, we have $\mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+) \subseteq$* $\operatorname{conv}(\mathfrak{S})$.

Proof. Let $(\hat{x}, \hat{t}) \in \mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+)$. We will show that $(\hat{x}, \hat{t}) \in \operatorname{conv}(\mathfrak{S})$. We split the analysis into three cases: (i) $q_1(\hat{x}) = 0$, (ii) $q_1(\hat{x}) > 0$, and (iii) $q_1(\hat{x}) < 0$.

i If $q_1(\hat{x}) = 0$, then $q_0(\hat{x}) = q_0(\hat{x}) + \gamma_- q_1(\hat{x}) = q(\gamma_-, \hat{x})$. As $(\hat{x}, \hat{t}) \in \mathcal{S}(\gamma_-)$ by assumption, we deduce that $q(\gamma_-, \hat{x}) \leq \hat{t}$. Combining these inequalities, we have that $q_0(\hat{x}) = q(\gamma_-, \hat{x}) \leq \hat{t}$ and that $(\hat{x}, \hat{t}) \in \mathfrak{S}$.

ii Now suppose $q_1(\hat{x}) > 0$. Let $d \neq 0$ such that $d^\intercal A(\gamma_+) d = 0$ (such a vector d exists as $A(\gamma_+)$ has zero as an eigenvalue; see Remark 50) and define $e := 2(\hat{x}^\intercal A(\gamma_+) d + b(\gamma_+)^\intercal d)$. We modify $(\hat{x}, \hat{t})$ along the direction (d, e): For $\alpha \in \mathbb{R}$, let $(\hat{x}_\alpha, \hat{t}_\alpha) := (\hat{x} + \alpha d, \hat{t} + \alpha e)$.

We will show that there exist $\alpha_1 < 0 < \alpha_2$ such that $(\hat{x}_{\alpha_i}, \hat{t}_{\alpha_i}) \in \mathfrak{S}$ for $i = 1, 2$, whence $(\hat{x}, \hat{t}) \in \operatorname{conv}(\mathfrak{S})$.

We study the behavior of the expressions $q(\gamma_-, \hat{x}_\alpha) - \hat{t}_\alpha$ and $q(\gamma_+, \hat{x}_\alpha) - \hat{t}_\alpha$ as functions of α. A short calculation shows that for any $\alpha \in \mathbb{R}$, we have

$$\begin{aligned} & q(\gamma_+, \hat{x}_\alpha) - \hat{t}_\alpha \\ &\qquad = \left(q(\gamma_+, \hat{x}) - \hat{t}\right) + 2\alpha(\hat{x}^\intercal A(\gamma_+)d + b(\gamma_+)^\intercal d - e/2) + \alpha^2 d^\intercal A(\gamma_+)d \\ &\qquad = q(\gamma_+, \hat{x}) - \hat{t}, \end{aligned} \tag{4.7}$$

where the last equation follows from the definition of e. Thus, $q(\gamma_+, \hat{x}_\alpha) - \hat{t}_\alpha$ is constant in α. Next, we compute

$$\begin{aligned} & q(\gamma_-, \hat{x}_\alpha) - \hat{t}_\alpha \\ &\qquad = \left(q(\gamma_-, \hat{x}) - \hat{t}\right) + 2\alpha(\hat{x}^\intercal A(\gamma_-)d + b(\gamma_-)^\intercal d - e/2) + \alpha^2 d^\intercal A(\gamma_-)d. \end{aligned}$$

As $d \neq 0$ and $d^\intercal A(\gamma_+)d = 0$, we deduce that $d^\intercal A(\gamma_-)d \neq 0$ (see Remark 50). Then, as $A(\gamma_-) \succeq 0$, we have that $d^\intercal A(\gamma_-)d > 0$. Hence, $q(\gamma_-, \hat{x}_\alpha) - \hat{t}_\alpha$ is strongly convex in α.

Note that

$$q(\gamma_-, \hat{x}) = q_0(\hat{x}) + \gamma_- q_1(\hat{x}) < q_0(\hat{x}) + \gamma_+ q_1(\hat{x}) = q(\gamma_+, \hat{x}),$$

where the inequality follows from the fact that $\gamma_- < \gamma_+$ and $q_1(\hat{x}) > 0$. Therefore, $q(\gamma_-, \hat{x}) - \hat{t} < q(\gamma_+, \hat{x}) - \hat{t}$. Thus, there are values $\alpha_1 < 0 < \alpha_2$ such that $q(\gamma_-, \hat{x}_{\alpha_i}) - \hat{t}_{\alpha_i} = q(\gamma_+, \hat{x}_{\alpha_i}) - \hat{t}_{\alpha_i}$ for $i = 1, 2$.

It remains to show that $(\hat{x}_{\alpha_i}, \hat{t}_{\alpha_i}) \in \mathfrak{S}$ for $i = 1, 2$. This follows immediately because for $i = 1, 2$, we have

$$q_1(\hat{x}_{\alpha_i}) = \frac{1}{\gamma_+ - \gamma_-}(q(\gamma_+, \hat{x}_{\alpha_i}) - q(\gamma_-, \hat{x}_{\alpha_i})) = 0.$$

Then, applying (4.7) and recalling that $q(\gamma_+, \hat{x}) \leq \hat{t}$, we have

$$q_0(\hat{x}_{\alpha_i}) = q(\gamma_+, \hat{x}_{\alpha_i}) - \gamma_+ q_1(\hat{x}_{\alpha_i}) = q(\gamma_+, \hat{x}_{\alpha_i}) \leq \hat{t}_{\alpha_i}.$$

iii The final case is symmetric to case (ii), thus we will only sketch its proof.

Suppose $q_1(\hat{x}) < 0$. Let $d \neq 0$ such that $d^\intercal A(\gamma_-)d = 0$ and define $e := 2(\hat{x}^\intercal A(\gamma_-)d + b(\gamma_-)^\intercal d)$. For $\alpha \in \mathbb{R}$, let $(\hat{x}_\alpha, \hat{t}_\alpha) := (\hat{x} + \alpha d, \hat{t} + \alpha e)$.

A short calculation shows that for any $\alpha \in \mathbb{R}$, we have

$$\begin{aligned} &q(\gamma_-, \hat{x}_\alpha) - \hat{t}_\alpha \\ &\quad = \left(q(\gamma_-, \hat{x}) - \hat{t}\right) + 2\alpha(\hat{x}^\intercal A(\gamma_-)d + b(\gamma_-)^\intercal d - e/2) + \alpha^2 d^\intercal A(\gamma_-)d \\ &\quad = q(\gamma_-, \hat{x}) - \hat{t}. \end{aligned}$$

Similarly, for any $\alpha \in \mathbb{R}$,

$$\begin{aligned} &q(\gamma_+, \hat{x}_\alpha) - \hat{t}_\alpha \\ &\quad = \left(q(\gamma_+, \hat{x}) - \hat{t}\right) + 2\alpha(\hat{x}^\intercal A(\gamma_+)d + b(\gamma_+)^\intercal d - e/2) + \alpha^2 d^\intercal A(\gamma_+)d. \end{aligned}$$

As $d^\intercal A(\gamma_-)d = 0$ and $d \neq 0$, Assumption 8 implies that $d^\intercal A(\gamma_+)d > 0$. We see that $q(\gamma_+, \hat{x}_\alpha) - \hat{t}_\alpha$ is strongly convex in α. As $q_1(\hat{x}) < 0$, we have $q(\gamma_+, \hat{x}) - \hat{t} < q(\gamma_-, \hat{x}) - \hat{t}$. Thus, there are values $\alpha_1 < 0 < \alpha_2$ such that $q(\gamma_+, \hat{x}_{\alpha_i}) - \hat{t}_{\alpha_i} = q(\gamma_-, \hat{x}_{\alpha_i}) - \hat{t}_{\alpha_i}$ for $i = 1, 2$.

Noting that $\gamma_- \neq \gamma_+$ and $q(\gamma_-, \hat{x}_{\alpha_i}) = q(\gamma_+, \hat{x}_{\alpha_i})$, we conclude that $q_0(\hat{x}_{\alpha_i}) = q(\gamma_i, \hat{x}_{\alpha_i}) \leq \hat{t}_{\alpha_i}$ and $q_1(\hat{x}_{\alpha_i}) = 0$. Thus, $(\hat{x}_{\alpha_i}, \hat{t}_{\alpha_i}) \in \mathfrak{S}$ for $i = 1, 2$. We conclude $(\hat{x}, \hat{t}) \in \operatorname{conv}(\mathfrak{S})$. ■

Remark 52. The proof of Lemma 44 suggests a simple rounding scheme from the convex relaxation to the original nonconvex problem: given $\hat{x} \in \mathbb{R}^n$, let d be an eigenvector of eigenvalue zero for either $A(\gamma_\pm)$ (depending on the sign of $q_1(\hat{x})$) and move $\alpha \geq 0$ units in the direction of either $\pm d$ (depending on the sign of e defined in the proof) until $q_1(\hat{x} \pm \alpha d) = 0$. This rounding scheme guarantees that $q_0(\hat{x} \pm \alpha d) \leq \max\{q(\gamma_-, \hat{x}), q(\gamma_+, \hat{x})\}$. □

We are now ready to prove Theorem 19.

Proof of Theorem 19. Lemmas 43 and 44 together imply

$$\mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+) \subseteq \operatorname{conv}(\mathfrak{S}) \subseteq \operatorname{conv}(\mathcal{S}) \subseteq \mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+).$$

Hence, we deduce that equality holds throughout the chain of inclusions.

In particular, the GTRS (4.4) can be rewritten

$$\begin{aligned} \inf_{(x,t)\in\mathbb{R}^{n+1}} \{t : (x,t) \in \mathcal{S}\} &= \inf_{(x,t)\in\mathbb{R}^{n+1}} \{t : (x,t) \in \operatorname{conv}(\mathcal{S})\} \\ &= \inf_{(x,t)\in\mathbb{R}^{n+1}} \{t : (x,t) \in \mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+)\} \\ &= \inf_{(x,t)\in\mathbb{R}^{n+1}} \left\{ t : \begin{array}{l} q(\gamma_-, x) \leq t \\ q(\gamma_+, x) \leq t \end{array} \right\} \\ &= \inf_{x\in\mathbb{R}^n} \max\{q(\gamma_-, x), q(\gamma_+, x)\}. \end{aligned}$$

It remains to prove that the minimum is achieved in each of the formulations of the GTRS above. It suffices to show that the minimum is achieved in the last formulation. Note $q(\gamma_-, x)$ and

$q(\gamma_+, x)$ are both continuous functions of x, hence $\max\{q(\gamma_-, x), q(\gamma_+, x)\}$ is continuous. Next, taking $u := \max\{c(\gamma_-), c(\gamma_+)\}$ we have that u is an upper bound on the optimal value. Moreover, because $\gamma^* \in (\gamma_-, \gamma_+)$, we can lower bound $\max\{q(\gamma_-, x), q(\gamma_+, x)\}$, by $q(\gamma^*, x)$. Consequently, it suffices to replace the feasible domain $\mathbb{R}^n$ in the last formulation with the set

$$\{x \in \mathbb{R}^n : \, q(\gamma^*, x) \leq u\}.$$

This set is bounded as $A(\gamma^*) \succ 0$ and it is closed as it is the inverse image of $(-\infty, u]$ under a continuous map. Recalling that a continuous function on a compact set achieves its minimum concludes the proof. ■

We next provide a numerical example illustrating Theorem 19.

Example 16. Define the homogeneous quadratic functions $q_i(x) := x^\intercal A_i x$ for $i = 0, 1$, where

$$A_0 := \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, \qquad A_1 := \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}.$$

As $\det(A_0) = -3$ and $\det(A_1) = -1$, the matrices A_0 and A_1 must both have negative eigenvalues. Furthermore,

$$A(2) = A_0 + 2A_1 = I \succ 0.$$

Thus, Assumption 8 is satisfied.

We now compute γ_- and γ_+. Note that as $A(\gamma)$ is a 2×2 matrix, $A(\gamma) \succeq 0$ if and only if $\operatorname{tr}(A(\gamma)) \geq 0$ and $\det(A(\gamma)) \geq 0$. Note that $\operatorname{tr}(A(\gamma)) = 2 \geq 0$ is satisfied for all γ. We compute

$$\det(A(\gamma)) = 1 - (2 - \gamma)^2.$$

This quantity is nonnegative if and only if $|2 - \gamma| \leq 1$. Thus $\gamma_- = 1$ and $\gamma_+ = 3$. Theorem 19 then implies

$$\operatorname{conv}\left(\left\{(x,t) \in \mathbb{R}^3 : \begin{array}{l} x_1^2 + 4x_1x_2 + x_2^2 \leq t \\ -2x_1x_2 \leq 0 \end{array}\right\}\right) = \left\{(x,t) \in \mathbb{R}^3 : \begin{array}{l} (x_1 + x_2)^2 \leq t \\ (x_1 - x_2)^2 \leq t \end{array}\right\}.$$

We plot the corresponding sets $\mathcal{S}$ and $\mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+)$ in Figure 4.1. □

4.2.2 Proof of Theorem 20

Next, we prove Theorem 20 using a limiting argument and reducing it to Theorem 19.

Lemma 45. *Suppose* Γ *is nonempty and write* $\Gamma = [\gamma_-, \gamma_+]$. *Then,* $\operatorname{clconv}(\mathcal{S}) \subseteq \mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+)$.

Proof. Note that $\mathcal{S} = \bigcap_{\gamma \geq 0} \mathcal{S}(\gamma) \subseteq \mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+)$. Containment then follows by taking the closed convex hull of both sides and noting that both $\mathcal{S}(\gamma_-)$ and $\mathcal{S}(\gamma_+)$ are closed and convex by Lemma 42. ■

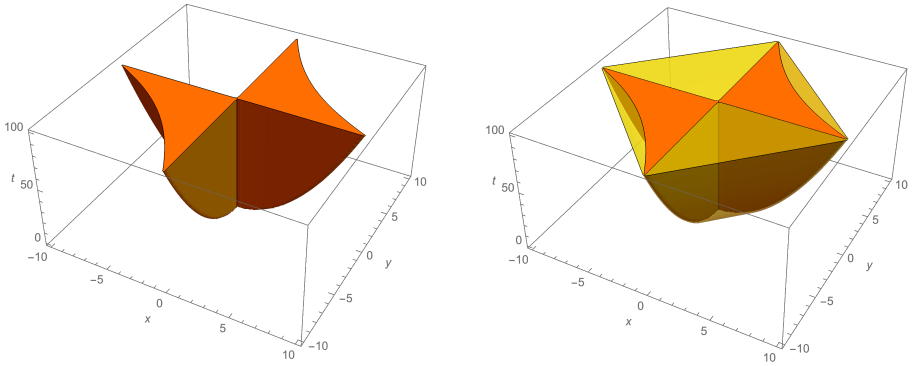

Figure 4.1: The sets $\mathcal{S}$ (in orange) and $\mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+)$ (in yellow) from Example 16

Lemma 46. *Under Assumption 9, we have that* $\mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+) \subseteq \operatorname{clconv}(\mathfrak{S})$.

Proof. Let $(\hat{x}, \hat{t}) \in \mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+)$. It suffices to show that $(\hat{x}, \hat{t} + \epsilon) \in \operatorname{conv}(\mathfrak{S})$ for all $\epsilon > 0$.

We will perturb A_0 slightly to create a new GTRS instance. Let $\delta > 0$ to be picked later. Define $A_0' = A_0 + \delta I_n$ and let all remaining data be unchanged, i.e.,

$$\begin{aligned} q_0'(x) &:= x^\intercal A_0' x + 2b_0'^\intercal x + c_0' := x^\intercal (A_0 + \delta I_n) x + 2b_0^\intercal x + c_0 \\ q_1'(x) &:= x^\intercal A_1' x + 2b_1'^\intercal x + c_1' := x^\intercal A_1 x + 2b_1^\intercal x + c_1. \end{aligned}$$

We will denote all quantities related to the perturbed system with an apostrophe.

We claim that it suffices to show that there exists a $\delta > 0$ small enough such that the GTRS defined by q_0' and q_1' satisfies Assumption 8 and $(\hat{x}, \hat{t} + \epsilon) \in \mathcal{S}'(\gamma_-') \cap \mathcal{S}'(\gamma_+')$. Indeed, suppose this is the case. Note that for any $x \in \mathbb{R}^n$, we have $q_1(x) = q_1'(x)$ and $q_0(x) \leq q_0'(x)$. Hence, $\mathfrak{S}' \subseteq \mathfrak{S}$ and $\operatorname{conv}(\mathfrak{S}') \subseteq \operatorname{conv}(\mathfrak{S})$. Then applying Theorem 19 gives $(\hat{x}, \hat{t} + \epsilon) \in \mathcal{S}'(\gamma_-') \cap \mathcal{S}'(\gamma_+') = \operatorname{conv}(\mathfrak{S}') \subseteq \operatorname{conv}(\mathfrak{S})$ as desired.

We pick $\delta > 0$ small enough such that

$$\lambda_{\min}(A_0') < 0, \quad \delta \|\hat{x}\|^2 \leq \frac{\epsilon}{2}, \quad |\gamma_+' - \gamma_+||q_1(\hat{x})| \leq \frac{\epsilon}{2}, \quad |\gamma_-' - \gamma_-||q_1(\hat{x})| \leq \frac{\epsilon}{2}.$$

This is possible as the expression on the left of each inequality is continuous in δ and is strictly satisfied if $\delta = 0$. Then, noting that $A'(\gamma^*) = A(\gamma^*) + \delta I_n \succ 0$, we have that the GTRS defined by q_0' and q_1' satisfies Assumption 8.

It remains to show that $q'(\gamma_+', \hat{x}) \leq (\hat{t} + \epsilon)$ and $q'(\gamma_-', \hat{x}) \leq (\hat{t} + \epsilon)$. We compute

$$\begin{aligned} q'(\gamma_+', \hat{x}) - (\hat{t} + \epsilon) &= q'(\gamma_+, \hat{x}) - (\hat{t} + \epsilon) + (\gamma_+' - \gamma_+) q_1(\hat{x}) \\ &\leq q(\gamma_+, \hat{x}) + \delta \|\hat{x}\|^2 - (\hat{t} + \epsilon) + |\gamma_+' - \gamma_+||q_1(\hat{x})| \\ &\leq q(\gamma_+, \hat{x}) - \hat{t} \\ &\leq 0. \end{aligned}$$

The first inequality follows by noting $q'(\gamma, x) = q(\gamma, x) + \delta\|x\|^2$, the second inequality follows from our assumptions on δ, and the third line follows from the assumption that $(\hat{x}, \hat{t}) \in \mathcal{S}(\gamma_+)$. A similar calculation shows $q'(\gamma'_-, \hat{x}) \leq (t + \epsilon)$. This concludes the proof. ■

We are now ready to prove Theorem 20.

Proof of Theorem 20. Lemmas 45 and 46 together imply

$$\mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+) \subseteq \operatorname{clconv}(\mathfrak{S}) \subseteq \operatorname{clconv}(\mathcal{S}) \subseteq \mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+).$$

Hence, we deduce that equality holds throughout the chain of inclusions.

In particular, the GTRS (4.4) can be rewritten

$$\begin{aligned}
\inf_{(x,t)\in\mathbb{R}^{n+1}} \{t : (x,t) \in \mathcal{S}\} &= \inf_{(x,t)\in\mathbb{R}^{n+1}} \{t : (x,t) \in \operatorname{clconv}(\mathcal{S})\} \\
&= \inf_{(x,t)\in\mathbb{R}^{n+1}} \{t : (x,t) \in \mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+)\} \\
&= \inf_{(x,t)\in\mathbb{R}^{n+1}} \left\{ t : \begin{array}{l} q(\gamma_-, x) \leq t \\ q(\gamma_+, x) \leq t \end{array} \right\} \\
&= \inf_{x\in\mathbb{R}^n} \max\{q(\gamma_-, x), q(\gamma_+, x)\}.
\end{aligned}$$ ■

4.2.3 Removing the nonconvex assumptions

As part of our Assumptions 8 and 9, we assume that A_0 and A_1 both have negative eigenvalues, i.e., that both q_0 and q_1 are nonconvex. These assumptions are made for ease of presentation and to highlight the novel contributions of this work. Indeed, the proofs of Theorems 19 and 20 can be modified to additionally cover all four cases of convex/nonconvex objective and constraint functions. We remark that the resulting theorem statement for the case of a nonconvex objective function and a strongly convex constraint function coincides with that of Ho-Nguyen and Kılınç-Karzan [87].

In this section we record more general versions Theorems 19 and 20. Their proofs are completely analogous to the original proofs and are deferred to Appendix D.1.

Theorem 21. *Suppose there exists $\gamma^* \geq 0$ such that $A(\gamma^*) \succ 0$. Consider the closed nonempty interval $\Gamma := \{\gamma \in \mathbb{R}_+ : A(\gamma) \succeq 0\}$. Let γ_- denote its leftmost endpoint.*

- *If Γ is bounded above, let γ_+ denote its rightmost endpoint. Then,*

  $$\operatorname{conv}(\mathcal{S}) = \mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+).$$

  In particular, we have $\min_{x\in\mathbb{R}^n}\{q_0(x) : q_1(x) \leq 0\} = \min_{x\in\mathbb{R}^n} \max\{q(\gamma_-, x),\ q(\gamma_+, x)\}$.

- *If Γ is not bounded above, then $q_1(x)$ is convex and*

  $$\operatorname{conv}(\mathcal{S}) = \mathcal{S}(\gamma_-) \cap \left\{(x,t) \in \mathbb{R}^{n+1} : q_1(x) \leq 0\right\}.$$

  In particular, we have $\min_{x\in\mathbb{R}^n}\{q_0(x) : q_1(x) \leq 0\} = \min_{x\in\mathbb{R}^n}\{q(\gamma_-, x) : q_1(x) \leq 0\}$.

Theorem 22. *Suppose there exists $\gamma^* \geq 0$ such that $A(\gamma^*) \succeq 0$. Consider the closed nonempty interval $\Gamma \coloneqq \{\gamma \in \mathbb{R}_+ : \ A(\gamma) \succeq 0\}$. Let γ_- denote its leftmost endpoint.*

- *If Γ is bounded above, let γ_+ denote its rightmost endpoint. Then,*
  $$\overline{\operatorname{conv}}(\mathcal{S}) = \mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+).$$
  In particular, $\inf_{x\in\mathbb{R}^n}\{q_0(x) : \ q_1(x) \leq 0\} = \inf_{x\in\mathbb{R}^n} \max\{q(\gamma_-, x), \ q(\gamma_+, x)\}$.

- *If Γ is not bounded above, then $q_1(x)$ is convex and*
  $$\overline{\operatorname{conv}}(\mathcal{S}) = \mathcal{S}(\gamma_-) \cap \left\{(x,t) \in \mathbb{R}^{n+1} : \ q_1(x) \leq 0\right\}.$$
  In particular, $\inf_{x\in\mathbb{R}^n}\{q_0(x) : \ q_1(x) \leq 0\} = \inf_{x\in\mathbb{R}^n}\{q(\gamma_-, x) : \ q_1(x) \leq 0\}$.

These results admit further nontrivial generalizations involving multiple quadratics; we refer the interested readers to our follow up work [181].

Remark 53. Yıldıran [195] proves a convex hull result for a set defined by two *strict* quadratic constraints. Modaresi and Vielma [123] then show that given a particular topological assumption, that the appropriate closed versions of Yıldıran [195]'s results also hold. We discuss these results in the context of the convex hull results we have presented thus far. Given q_0 and q_1 we will consider the quadratic functions $q_0(x) - t$ and $q_1(x)$ in the variables (x, t). As [195] works with homogeneous quadratics, we introduce an extra variable to get homogeneous quadratic forms. Define

$$Q_0 \coloneqq \begin{pmatrix} A_0 & 0 & b_0 \\ 0^\intercal & 0 & -1/2 \\ b_0^\intercal & -1/2 & c_0 \end{pmatrix}, \quad Q_1 \coloneqq \begin{pmatrix} A_1 & 0 & b_1 \\ 0^\intercal & 0 & 0 \\ b_1^\intercal & 0 & c_1 \end{pmatrix}, \quad Q(\gamma) \coloneqq \begin{pmatrix} A(\gamma) & 0 & b(\gamma) \\ 0^\intercal & 0 & -1/2 \\ b(\gamma)^\intercal & -1/2 & c(\gamma) \end{pmatrix}.$$

Yıldıran [195] uses the aggregation weights γ where $Q(\gamma)$ has exactly one negative eigenvalue. Note that for all $\gamma \geq 0$, the lower right 2×2 block of $Q(\gamma)$ is invertible. Thus, we may take the Schur complement of this block in $Q(\gamma)$:

$$Q(\gamma)/\begin{pmatrix} 0 & -1/2 \\ -1/2 & c(\gamma) \end{pmatrix} = A(\gamma) - \begin{pmatrix} 0 & b(\gamma) \end{pmatrix} \begin{pmatrix} 0 & -1/2 \\ -1/2 & c(\gamma) \end{pmatrix}^{-1} \begin{pmatrix} 0^\intercal \\ b(\gamma)^\intercal \end{pmatrix} = A(\gamma).$$

Recall that Schur complements preserve inertia. In other words, $Q(\gamma)$ and

$$\begin{pmatrix} A(\gamma) & & \\ & 0 & -1/2 \\ & -1/2 & c(\gamma). \end{pmatrix}$$

have the same number of negative eigenvalues. Noting that the lower right 2×2 block has exactly one negative eigenvalue, we conclude that $Q(\gamma)$ has exactly one negative eigenvalue if and only if $A(\gamma) \succeq 0$. The result presented by Yıldıran [195] then implies

$$\operatorname{conv}\left(\left\{(x,t) : \begin{array}{c} q_0(x) < t \\ q_1(x) < 0 \end{array}\right\}\right) = \{(x,t) : q(\gamma_-, x) < t\} \cap \{(x,t) : q(\gamma_+, x) < t\}$$

when γ_+ exists and

$$\operatorname{conv}\left(\left\{(x,t) : \begin{array}{c} q_0(x) < t \\ q_1(x) < 0 \end{array}\right\}\right) = \{(x,t) : q(\gamma_-, x) < t\} \cap \{(x,t) : q_1(x) < 0\}$$

otherwise.

One can then verify the topological assumption of Modaresi and Vielma [123], namely that $\mathcal{S} \subseteq \overline{\operatorname{int}(\mathcal{S})}$ the closure of the interior of $\mathcal{S}$. Thus, combining these two results gives an alternate proof of Theorems 21 and 22.

We believe our analysis is simpler and more direct. In particular, our analysis takes advantage of the epigraph structure present in our sets and immediately implies a rounding procedure via Lemma 44. In addition, our results are more refined when Assumption 8 or 9 hold as we can also characterize the (closed) convex hull of the set $\mathfrak{S}$ and show that it is equal to that of $\mathcal{S}$. This particular distinction between $\mathfrak{S}$ and $\mathcal{S}$ has a number of interesting implications in equality-, interval-, or hollow-constrained GTRS, and we discuss these results in the following section. □

4.3 Nonintersecting constraints

There have been a number of works considering interval-, equality-, or hollow-constrained variants of the GTRS [21, 24, 94, 96, 124, 148, 155, 165, 191] (see [87, Section 3.3] and references therein for extensions of the TRS and their applications). In this section, we extend our (closed) convex hull results in the presence of a general nonintersecting constraint. This allows us to handle multiple variants of the GTRS simultaneously.

Specifically, we will impose an additional requirement $x \in \Omega$. The new form of the GTRS will be

$$\inf_{x\in\mathbb{R}^n}\left\{q_0(x) : \begin{array}{c} q_1(x) \le 0 \\ x \in \Omega \end{array}\right\} = \inf_{(x,t)\in\mathbb{R}^{n+1}}\left\{t : \begin{array}{c} q_0(x) \le t \\ q_1(x) \le 0 \\ x \in \Omega \end{array}\right\}.$$

Let $\mathcal{S}_\Omega$ denote the set of feasible points (x,t), i.e.,

$$\mathcal{S}_\Omega := \left\{(x,t) \in \mathbb{R}^{n+1} : \begin{array}{c} q_0(x) \le t \\ q_1(x) \le 0 \\ x \in \Omega \end{array}\right\}.$$

We will assume that $\Omega \subseteq \mathbb{R}^n$ satisfies the following *nonintersecting* condition.

Assumption 10. The set $\Omega \subseteq \mathbb{R}^n$ satisfies $\{x \in \mathbb{R}^n : q_1(x) = 0\} \subseteq \Omega$. □

The following two corollaries to Theorems 19 and 20 follow immediately by noting that $\mathfrak{S} \subseteq \mathcal{S}_\Omega \subseteq \mathcal{S}$ holds under Assumption 10.

Corollary 19. *Suppose Assumptions 8 and 10 hold. Then,*

$$\operatorname{conv}(\mathcal{S}_\Omega) = \mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+).$$

Proof. Under Assumptions 8 and 10, we get the following chain of inclusions

$$\operatorname{conv}(\mathcal{S}_\Omega) \subseteq \operatorname{conv}(\mathcal{S}) = \operatorname{conv}(\mathfrak{S}) \subseteq \operatorname{conv}(\mathcal{S}_\Omega),$$

where the first subset relation follows $\mathcal{S}_\Omega \subseteq \mathcal{S}$ (by definition of the set $\mathcal{S}_\Omega$), the equality relation follows from Theorem 19, and the last subset relation follows from $\mathfrak{S} \subseteq \mathcal{S}_\Omega$ (by Assumption 10). We conclude that $\operatorname{conv}(\mathcal{S}_\Omega) = \operatorname{conv}(\mathcal{S})$. By Theorem 19, we know that $\operatorname{conv}(\mathcal{S}) = \mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+)$. ■

Corollary 20. *Suppose Assumptions 9 and 10 hold. Then,*

$$\overline{\operatorname{conv}}(\mathcal{S}_\Omega) = \mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+).$$

Proof. Applying Assumptions 9 and 10 and Theorem 20, we get the following chain of inclusions

$$\overline{\operatorname{conv}}(\mathcal{S}_\Omega) \subseteq \overline{\operatorname{conv}}(\mathcal{S}) = \overline{\operatorname{conv}}(\mathfrak{S}) \subseteq \overline{\operatorname{conv}}(\mathcal{S}_\Omega).$$

We conclude that $\overline{\operatorname{conv}}(\mathcal{S}_\Omega) = \overline{\operatorname{conv}}(\mathcal{S})$. By Theorem 20, we know that $\overline{\operatorname{conv}}(\mathcal{S}) = \mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+)$. ■

Remark 54. These two corollaries show that nonintersecting constraints in the GTRS may be ignored. Consider for example the interval-constrained GTRS. Define

$$\Omega := \{x \in \mathbb{R}^n : \, q_1(x) \geq -1\}.$$

Then, clearly Assumption 10 is satisfied. Under Assumption 9, we have

$$\begin{aligned}
\inf_{x \in \mathbb{R}^n} \{q_0(x) : \, -1 \leq q_1(x) \leq 0\} &= \inf_{(x,t) \in \mathbb{R}^{n+1}} \{t : \, (x,t) \in \mathcal{S}_\Omega\} \\
&= \inf_{(x,t) \in \mathbb{R}^{n+1}} \{t : \, (x,t) \in \overline{\operatorname{conv}}(\mathcal{S}_\Omega)\} \\
&= \inf_{(x,t) \in \mathbb{R}^{n+1}} \{t : \, (x,t) \in \mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+)\} \\
&= \inf_{x \in \mathbb{R}^n} \max\{q(\gamma_-, x), q(\gamma_+, x)\}.
\end{aligned}$$

Thus, the value of the interval-constrained GTRS is the same as the GTRS under Assumption 9. Similarly, the Ω sets arising from equality- or hollow-constrained GTRS also satisfy Assumption 10. Hence, under Assumption 9, the additional constraints in these variants of the GTRS can also be dropped. □

4.4 Solving the convex reformulation in linear time

In this section we present algorithms, inspired by Theorem 19, for approximately solving the GTRS. Note that Theorem 19 gives a tight convex reformulation of the GTRS: under Assumption 8,

$$\mathrm{Opt} := \min_{x\in\mathbb{R}^n}\{q_0(x) :\ q_1(x) \leq 0\} = \min_{x\in\mathbb{R}^n} \max\{q(\gamma_-, x), q(\gamma_+, x)\}.$$

Then given a solution to the convex reformulation on the right, Lemma 44 gives a rounding scheme to recover a solution to the original GTRS on the left.

In order to establish an explicit running time of an algorithm based on the above idea, we must carefully handle a number of numerical issues. In practice, we cannot expect to compute $\gamma_\pm$ exactly. Instead, we will show how to compute estimates $\tilde{\gamma}_\pm$ of $\gamma_\pm$ up to some accuracy δ. We will take care to pick $\tilde{\gamma}_\pm$ satisfying the relation $[\tilde{\gamma}_-, \tilde{\gamma}_+] \subseteq [\gamma_-, \gamma_+]$ so that the quadratic forms defined by $A(\tilde{\gamma}_-)$ and $A(\tilde{\gamma}_+)$ are convex. Based on the estimates $\tilde{\gamma}_\pm$, we will then formulate and solve the convex optimization problem

$$\widetilde{\mathrm{Opt}} := \min_{x\in\mathbb{R}^n} \max\{q(\tilde{\gamma}_-, x), q(\tilde{\gamma}_+, x)\}.$$

Finally, given an (approximate) solution to the convex problem $\widetilde{\mathrm{Opt}}$, Lemma 44 tells us how to construct a solution to the original nonconvex GTRS using specific eigenvectors. Again, we will need to handle numerical issues that arise from not being able to compute these eigenvectors exactly.

Throughout this section, we will work under the following assumption.

Assumption 11.

- There exists some $\gamma^* \geq 0$ such that $A(\gamma^*) \succ 0$,
- $\|A_0\|, \|A_1\|, \|b_0\|, \|b_1\|, |c_1| \leq 1$. □

Remark 55. Note that the first part of Assumption 11 is simply Assumption 8. We make this assumption so that we may use the convex reformulation guaranteed by Theorem 19. Assumption 8 is commonly used in GTRS algorithms; see e.g., Jiang and Li [95, Assumption 2.3] and the discussion following it. The second part of Assumption 11 can be achieved for an arbitrary pair q_0 and q_1 by simply scaling each quadratic by a positive scalar. Note that any optimal (respectively feasible) solution remains optimal (respectively feasible) when q_0 (respectively q_1) is scaled by a positive scalar. □

We will analyze the running time of our algorithm in terms of N, the number of nonzero entries in A_0 and A_1, ϵ, the additive error, p, the failure probability, and n, the dimension. In addition, the running time of our algorithm depends on certain regularity parameters of the pair q_0 and q_1 defined below.

Definition 17. Let q_0, q_1 satisfy Assumption 11. Define

$$\zeta^* := \max\{1, \gamma_+\}, \quad \text{and} \quad \xi^* := \min\left\{1, \max_{\gamma\geq 0} \lambda_{\min}(A(\gamma))\right\}.$$

We say that q_0 and q_1 are (ξ, ζ) regular if $0 < \xi \leq \xi^*$ and $\zeta \geq \zeta^*$. Define $\kappa^* = \zeta^*/\xi^*$. When (ξ, ζ) are clear from context we will write $\kappa := \zeta/\xi$. □

In our analysis, we will frequently use the inequalities $\kappa, \zeta, \xi^{-1} \geq 1$, which for example imply $\kappa^2 \geq \kappa$ and $1 + \kappa \leq 2\kappa$, and the inequalities $\gamma_- \leq \gamma_+ \leq \zeta$, which for example under Assumption 11 imply $\|A(\gamma_+)\| \leq 1 + \zeta \leq 2\zeta$.

Remark 56. Jiang and Li [95] present a different linear-time algorithm for solving the GTRS. In their paper, they assume they are given a regularity parameter ξ_{JL} as input. This parameter must satisfy $\xi_{\text{JL}} \leq \xi^*_{\text{JL}}$ where

$$\xi^*_{\text{JL}} := \min\left\{1, -\lambda_{\min}(A_1), \max_{\mu \in (0,1]} \lambda_{\min}(\mu A_0 + (1-\mu)A_1)\right\}.$$

We now discuss how our regularity parameters, ξ^*_{us}, ζ^*, and $\kappa^* := \frac{\zeta^*}{\xi^*_{\text{us}}}$ relate to ξ^*_{JL}. For simplicity, we will assume

$$\xi^*_{\text{us}} = \max_{\gamma \geq 0}\{\lambda_{\min}(A(\gamma))\}, \qquad \zeta^* = \gamma_+,$$
$$\xi^*_{\text{JL}} = \min\left\{-\lambda_{\min}(A_1), \max_{\mu \in (0,1]} \lambda_{\min}(\mu A_0 + (1-\mu)A_1)\right\}.$$

We claim $\zeta^* \leq (-\lambda_{\min}(A_1))^{-1}$. Indeed, let x be a unit eigenvector corresponding to $\lambda_{\min}(A_1)$. Then, for any $\gamma > (-\lambda_{\min}(A_1))^{-1}$, we have

$$x^\intercal A(\gamma)x = x^\intercal A_0 x + \gamma x^\intercal A_1 x \leq 1 + \gamma\lambda_{\min}(A_1) < 0.$$

The role played by the bound $\gamma_+ \leq \zeta$ in our analysis is similar to the role of the bound $\xi_{\text{JL}} \leq -\lambda_{\min}(A_1)$ in the analysis presented by Jiang and Li [95].

We claim that

$$\frac{1}{2\kappa^*} \leq \max_{\mu \in (0,1]} \lambda_{\min}(\mu A_0 + (1-\mu)A_1) \leq \xi^*_{\text{us}},$$

and that the lower bound is sharp. Indeed, by performing the transformation $\mu = \frac{1}{1+\gamma}$, we can rewrite

$$\max_{\mu \in (0,1]} \lambda_{\min}(\mu A_0 + (1-\mu)A_1) = \max_{\gamma \geq 0} \frac{1}{1+\gamma}\lambda_{\min}(A(\gamma)),$$

which we can clearly bound above by ξ^*_{us}. On the other hand, noting that any optimizer, γ, of the above problem must lie in $[0, \gamma_+] = [0, \zeta^*]$, we can lower bound

$$\max_{\gamma \geq 0} \frac{1}{1+\gamma}\lambda_{\min}(A(\gamma)) \geq \frac{1}{1+\zeta^*} \max_{\gamma \geq 0} \lambda_{\min}(A(\gamma)) = \frac{\xi^*_{\text{us}}}{1+\zeta^*} \geq \frac{1}{2\kappa^*}.$$

We now construct a simple example for which the lower bound, $\xi^*_{\mathrm{JL}} \geq \frac{1}{2\kappa^*}$, is sharp. Let $\alpha > 0$ and define

$$A_0 = \mathrm{Diag}(1, 1, -1), \qquad A_1 = \mathrm{Diag}\Big(1, -(1+\alpha)^{-1}, 1\Big).$$

It is simple to see that $\|A_0\| = \|A_1\| = 1$, $\xi^*_{\mathrm{us}} = \frac{\alpha}{2+\alpha}$ and $\zeta^* = 1 + \alpha$. In particular, $\kappa^* = \frac{2+3\alpha+\alpha^2}{\alpha}$. On the other hand, we can compute

$$\xi^*_{\mathrm{JL}} = \max_{\mu \in (0,1]} \min\{\mu - (1-\mu)\alpha, \mu(-1+2\alpha) + (1-\mu)\alpha\} = \frac{\alpha}{4+3\alpha}.$$

Then, letting $\alpha \to 0$, we have $\kappa^* = \frac{2+o(1)}{\alpha}$ and $\xi^*_{\mathrm{JL}} = \frac{\alpha}{4+o(1)}$.

In view of the (closed) convex hull results presented in Theorems 19 and 20, we believe that the right notion of regularity should depend on the parameterization $A_0 + \gamma A_1$ as opposed to $\mu A_0 + (1-\mu)A_1$. We compare the running time of the algorithm presented by Jiang and Li [95] and the running time of our algorithms in Remark 60. □

We will assume that we have access to these regularity parameters within our algorithms.

Assumption 12. Assume we have algorithmic access to a pair (ξ, ζ) such that q_0 and q_1 are (ξ, ζ)-regular and a $\hat{\gamma}$ satisfying $\lambda_{\min}(A(\hat{\gamma})) \geq \xi$. □

Remark 57. Assumption 12 is quite reasonable. Indeed, there are simple and efficient binary search schemes to find constant factor approximations of ξ^* and ζ^* and a corresponding $\hat{\gamma}$. We detail one such algorithm in Appendix D.2. We remark that a similar assumption is made by Jiang and Li [95]: they assume they are given access to ξ_{JL} and present an algorithm for computing a corresponding $\hat{\mu}$ (see Remark 56). Another algorithm for finding $\hat{\gamma}$ is presented by Guo et al. [80] in the language of matrix pencil definiteness. □

We now fix the accuracy[2] to which we will compute our estimates $\tilde{\gamma}_\pm$. Define

$$\delta := \frac{\epsilon}{72\kappa^2}. \tag{4.8}$$

The framework for our approach is shown in Algorithm 1.

```
Algorithm 1 ApproxConvex($q_0, q_1, \xi, \zeta, \hat{\gamma}, \epsilon, p$)
Given $q_0$ and $q_1$ satisfying Assumption 11, $(\xi, \zeta)$ and $\hat{\gamma}$ satisfying Assumption 12, error parameter
$0 < \epsilon \leq \kappa^2 \xi$, and failure probability $p > 0$
  1. Pick $\delta$ as in (4.8).
  2. Find $\tilde{\gamma}_-$ and $\tilde{\gamma}_+$ such that
       $\tilde{\gamma}_- \in [\gamma_-, \gamma_- + \delta], \quad \tilde{\gamma}_+ \in [\gamma_+ - \delta, \gamma_+], \quad \lambda_{\min}(A(\tilde{\gamma}_\pm)) \leq \delta/\kappa,$   (4.9)
     with failure probability of at most $p$.
  3. Define $\widetilde{\mathrm{Opt}} := \min_{x \in \mathbb{R}^n} \max\{q(\tilde{\gamma}_-, x), q(\tilde{\gamma}_+, x)\}$. Solve $\widetilde{\mathrm{Opt}}$ up to accuracy $\epsilon/2$.
  4. Output $\tilde{\gamma}_-, \tilde{\gamma}_+$, and the approximate optimizer $\tilde{x}$.
```

[2] Our definition of accuracy is presented in (4.9).

Note that by Definition 17, we have $\kappa^2\xi = \zeta^2/\xi \geq 1$. Thus the requirement $0 < \epsilon \leq \kappa^2\xi$ in Algorithm 1 is not a practical issue: given $\epsilon > \kappa^2\xi$, we can simply run our algorithm with $\epsilon' = 1$ and return a solution with a better error guarantee.

This section is structured as follows. In Section 4.4.1, we prove that when δ is picked according to (4.8), $\widetilde{\text{Opt}}$ is within $\epsilon/2$ of Opt. In Section 4.4.2 we show how to compute $\tilde{\gamma}_-$ and $\tilde{\gamma}_+$ to satisfy (4.9). Then in Section 4.4.3, we present an algorithm due to Nesterov [132] and show that it can be used to efficiently solve for $\widetilde{\text{Opt}}$ up to accuracy $\epsilon/2$. At the end of Section 4.4.3, we present Theorem 24, which collects the results of the previous subsections and formally analyzes the runtime of Algorithm 1. In Section 4.4.4, we give a rounding scheme for finding a solution to the original GTRS (4.1) given a solution to the convex reformulation. Finally, in Section 4.4.5, we show that the running times of our algorithms can be significantly improved in situations where it is easy to compute $\gamma_\pm$ and zero eigenvectors of $A(\gamma_\pm)$.

4.4.1 Perturbation analysis of the convex reformulation

In this subsection, we show that the perturbed convex reformulation, $\widetilde{\text{Opt}}$, approximates the true convex reformulation, Opt, up to an additive error of $\epsilon/2$ when δ is picked as in (4.8). We will assume that step 2 of Algorithm 1 is successful, i.e., we have $\tilde{\gamma}_\pm$ satisfying (4.9).

Recall the definition of δ in (4.8). As we require $\epsilon \leq \kappa^2\xi$, we will have

$$\delta := \frac{\epsilon}{72\kappa^2} \leq \frac{\xi}{72} < \xi.$$

It is easy to see that $\lambda_{\min}(A(\gamma))$ is a 1-Lipschitz function in γ. Then recalling that $\lambda_{\min}(A(\gamma_\pm)) = 0$ and $\lambda_{\min}(A(\hat{\gamma})) \geq \xi$, we deduce the containment $\hat{\gamma} \in (\gamma_- + \delta, \gamma_+ - \delta)$. This, along with (4.9), implies

$$\hat{\gamma} \in (\tilde{\gamma}_-, \tilde{\gamma}_+) \subseteq [\gamma_-, \gamma_+], \qquad \tilde{\gamma}_- \in [\gamma_-, \gamma_- + \delta], \qquad \tilde{\gamma}_+ \in [\gamma_+ - \delta, \gamma_+]. \tag{4.10}$$

Recall the perturbed reformulation

$$\widetilde{\text{Opt}} := \min_{x\in\mathbb{R}^n} \max\{q(\tilde{\gamma}_-, x), q(\tilde{\gamma}_+, x)\}.$$

For notational convenience, let $f(x) := \max\{q(\gamma_-, x), q(\gamma_+, x)\}$ and let $\tilde{f}(x) := \max\{q(\tilde{\gamma}_-, x), q(\tilde{\gamma}_+, x)\}$. Let x^* and $\tilde{x}^*$ denote optimizers of Opt and $\widetilde{\text{Opt}}$ respectively.

Lemma 47. *For any fixed $x \in \mathbb{R}^n$, we have $\tilde{f}(x) \leq f(x)$. In particular, $\widetilde{\text{Opt}} \leq \text{Opt}$.*

Proof. Note that $q(\gamma, x)$ is a linear function in γ. Hence, for any fixed $x \in \mathbb{R}^n$, the containment $[\tilde{\gamma}_-, \tilde{\gamma}_+] \subseteq [\gamma_-, \gamma_+]$ implies $\tilde{f}(x) \leq f(x)$. We deduce

$$\widetilde{\text{Opt}} \leq \tilde{f}(x^*) \leq f(x^*) = \text{Opt}\,. \qquad \blacksquare$$

To show $\widetilde{\text{Opt}} \geq \text{Opt} - \epsilon/2$, we will show that x^* and $\tilde{x}^*$ lie in a ball of bounded radius and that $\tilde{f}$ approximates f uniformly on this ball.

Lemma 48. *Let x^* and $\tilde{x}^*$ be optimizers of* Opt *and* $\widetilde{\mathrm{Opt}}$ *respectively. Then* $x^*, \tilde{x}^* \in B(0, 5\kappa)$*.*

Proof. By picking the feasible solution $0 \in \mathbb{R}^n$ and Lemma 47, we have a trivial upper bound on $\widetilde{\mathrm{Opt}}$ and Opt:

$$\widetilde{\mathrm{Opt}} \leq \mathrm{Opt} \leq \max\{q(\gamma_-, 0), q(\gamma_+, 0)\} = \max\{c(\gamma_-), c(\gamma_+)\}. \tag{4.11}$$

By the first part of (4.10), we have

$$f(x) \geq \tilde{f}(x) \geq q(\hat{\gamma}, x) \geq \xi\|x\|^2 + 2b(\hat{\gamma})^\intercal x + c(\hat{\gamma}),$$

where the last inequality follows from the assumption that $\lambda_{\min}(A(\hat{\gamma})) \geq \xi$. Then,

$$\begin{aligned} x^*, \tilde{x}^* &\in \left\{x \in \mathbb{R}^n : \xi\|x\|^2 + 2b(\hat{\gamma})^\intercal x + c(\hat{\gamma}) \leq \max\{c(\gamma_-), c(\gamma_+)\}\right\} \\ &\subseteq \left\{x \in \mathbb{R}^n : \xi\|x\|^2 + 2b(\hat{\gamma})^\intercal x \leq \zeta\right\}. \end{aligned}$$

The last relation holds since $\max\{c(\gamma_-) - c(\hat{\gamma}), c(\gamma_+) - c(\hat{\gamma})\} = \max\{(\gamma_- - \hat{\gamma})c_1, (\gamma_+ - \hat{\gamma})c_1\} \leq |c_1|\gamma_+ \leq \zeta$. Then, by completing the square

$$\begin{aligned} x^*, \tilde{x}^* &\in B\left(-b(\hat{\gamma})\xi^{-1}, \sqrt{\|b(\hat{\gamma})\|^2\xi^{-2} + \kappa}\right) \\ &\subseteq B\left(0, 2\|b(\hat{\gamma})\|\xi^{-1} + \sqrt{\kappa}\right) \\ &\subseteq B(0, 4\kappa + \sqrt{\kappa}) \\ &\subseteq B(0, 5\kappa), \end{aligned}$$

where in the third line, we used Assumption 11 and the bound $\|b(\hat{\gamma})\| \leq \|b_0\| + \gamma_+\|b_1\| \leq 2\zeta$. ■

Lemma 49. *If $\|\hat{x}\| \leq 5\kappa$, then $\tilde{f}(\hat{x}) \geq f(\hat{x}) - \epsilon/2$. In particular,* $\widetilde{\mathrm{Opt}} \geq \mathrm{Opt} - \epsilon/2$*,*

Proof. Recall that $\delta := \frac{\epsilon}{72\kappa^2}$. Let $\hat{x} \in \mathbb{R}^n$ such that $\|\hat{x}\| \leq 5\kappa$. We compute

$$\begin{aligned} \tilde{f}(\hat{x}) &= \max\{q(\tilde{\gamma}_-, \hat{x}), q(\tilde{\gamma}_+, \hat{x})\} \\ &\geq \max\{q(\gamma_-, \hat{x}), q(\gamma_+, \hat{x})\} - \delta|q_1(\hat{x})| \\ &\geq f(\hat{x}) - \delta\left(\|\hat{x}\|^2 + 2\|\hat{x}\| + 1\right) \\ &\geq f(\hat{x}) - \delta(6\kappa)^2 \\ &= f(\hat{x}) - \epsilon/2, \end{aligned}$$

where the first inequality follows from (4.10), the second inequality follows from Assumption 11, and the third inequality follows from the bound $\|\hat{x}\| \leq 5\kappa$. ■

4.4.2 Approximating γ_- and γ_+

In this subsection, we show how to approximate γ_- and γ_+ and provide an explicit running time analysis of this procedure. Our developments rely on the fact that $\lambda_{\min}(A(\gamma))$ is a concave function in γ and that γ_- and γ_+ are the unique zeros of this function.

Lemma 50. *$\lambda_{\min}(A(\gamma))$ is a concave function in γ.*

Proof. By Courant-Fischer Theorem, $\lambda_{\min}(A(\gamma)) = \min_{\|x\|=1} x^\intercal A(\gamma)x$. Note that for any fixed $x \in \mathbb{R}^n$, the expression $x^\intercal A(\gamma)x$ is linear in γ. Then, the result follows upon recalling that the minimum of concave (in our case linear) functions is concave. ■

Let us also state a simple property of the function $\lambda_{\min}(A(\gamma))$.

Lemma 51.

i *Suppose $\gamma \leq \hat{\gamma}$, then $|\gamma - \gamma_-| \leq \kappa|\lambda_{\min}(A(\gamma))|$.*

ii *Suppose $\gamma \geq \hat{\gamma}$, then $|\gamma - \gamma_+| \leq \kappa|\lambda_{\min}(A(\gamma))|$.*

Proof. We only prove the first statement as the second statement follows similarly. Let $\gamma \leq \hat{\gamma}$. From the concavity of $\lambda_{\min}(A(\gamma))$, we have

$$|\lambda_{\min}(A(\gamma))| \geq |\gamma - \gamma_-|\frac{\lambda_{\min}(A(\hat{\gamma}))}{\hat{\gamma} - \gamma_-} \geq |\gamma - \gamma_-|\frac{\xi}{\zeta},$$

where in the second inequality we used the definition of ξ in Definition 17 and the bound $\hat{\gamma} - \gamma_- \leq \gamma_+ \leq \zeta$. Noting $\zeta/\xi = \kappa$ and rearranging terms completes the proof. ■

We will use the Lanczos method for approximating the most negative eigenvalue (and a corresponding eigenvector) of a sparse matrix. This algorithm, along with Lemma 51, will allow us to binary search over the range $[0, \zeta]$ for the zeros of the function $\lambda_{\min}(A(\gamma))$.

Lemma 52 ([103])**.** *There exists an algorithm, ApproxEig(A, ρ, η, p_{eig}), which given a symmetric matrix $A \in \mathbb{R}^{n\times n}$, ρ such that $\|A\| \leq \rho$, and parameters $\eta, p_{eig} > 0$, will, with probability at least $1 - p_{eig}$, return a unit vector $x \in \mathbb{R}^n$ such that $x^\intercal Ax \leq \lambda_{\min}(A) + \eta$. This algorithm runs in time*

$$O\left(\frac{N\sqrt{\rho}}{\sqrt{\eta}}\log\left(\frac{n}{p_{eig}}\right)\right),$$

where N is the number of nonzero entries in A.

Consider ApproxGammaPlus (Algorithm 2) for computing $\tilde{\gamma}_+$ up to accuracy δ. A similar algorithm can be used to compute $\tilde{\gamma}_-$ up to accuracy δ and is omitted.

Lemma 53. *Given q_0, q_1 satisfying Assumption 11, (ξ, ζ) and $\hat{\gamma}$ satisfying Assumption 12, $\delta > 0$, and $p_{\tilde{\gamma}_+}$, ApproxGammaPlus (Algorithm 2) outputs $\tilde{\gamma}_+$ satisfying*

$$\tilde{\gamma}_+ \in [\gamma_+ - \delta,\ \gamma_+], \qquad \lambda_{\min}(A(\tilde{\gamma}_+)) \leq \delta/\kappa$$

Algorithm 2 ApproxGammaPlus($q_0, q_1, \xi, \zeta, \hat{\gamma}, \delta, p_{\tilde{\gamma}_+}$)

Given q_0 and q_1 satisfying Assumption 11, (ξ, ζ) and $\hat{\gamma}$ satisfying Assumption 12, error parameter $\delta > 0$, and failure probability $p_{\tilde{\gamma}_+}$

```
1. Let $s_0 = \hat{\gamma}, t_0 = \zeta$
2. Let $T = \left\lceil \log\left(\frac{\zeta\kappa}{\delta}\right) \right\rceil + 2$
3. For $k = 0, \dots, T-1$
    a) Let $\gamma = (s_k + t_k)/2$
    b) Let $x = \mathrm{ApproxEig}(A(\gamma), 2\zeta, \frac{\delta}{4\kappa}, \frac{p_{\tilde{\gamma}}}{T})$
    c) If $x^\intercal A(\gamma)x < \frac{\delta}{4\kappa}$, set $s_{k+1} = s_k$ and $t_{k+1} = \gamma$
    d) Else if $x^\intercal A(\gamma)x > \frac{\delta}{\kappa}$, set $s_{k+1} = \gamma$ and $t_{k+1} = t_k$
    e) Else, stop and output $\tilde{\gamma}$
```

with probability $1 - p_{\tilde{\gamma}_+}$. *This algorithm runs in time*

$$\tilde{O}\left(\frac{N\sqrt{\kappa\zeta}}{\sqrt{\delta}} \log\left(\frac{n}{p_{\tilde{\gamma}_+}}\right) \log\left(\frac{\kappa}{\delta}\right)\right).$$

Proof. We condition on the event that ApproxEig succeeds every time it is called. By the union bound, this happens with probability at least $1 - p_{\tilde{\gamma}_+}$.

Suppose the algorithm outputs at step 3.(e). Let γ be the value of γ on the round in which the algorithm stops, and x the vector returned by ApproxEig in the corresponding iteration. Then, the stopping rule guarantees $x^\intercal A(\gamma)x \in [\delta/4\kappa, \delta/\kappa]$. As we have conditioned on ApproxEig succeeding, we deduce

$$x^\intercal A(\gamma)x - \frac{\delta}{4\kappa} \leq \lambda_{\min}(A(\gamma)) \leq x^\intercal A(\gamma)x.$$

In particular, $\lambda_{\min}(A(\gamma)) \in [0, \delta/\kappa]$ and $\gamma \leq \gamma_+$. Applying Lemma 51 gives

$$|\gamma - \gamma_+| \leq \kappa|\lambda_{\min}(A(\gamma))| \leq \delta.$$

We conclude $\gamma_+ - \delta \leq \gamma \leq \gamma_+$.

We now show that this algorithm outputs within T rounds. Let

$$P \coloneqq \{\gamma : \gamma \geq \hat{\gamma},\ \lambda_{\min}(A(\gamma)) \in [\delta/4\kappa, 3\delta/4\kappa]\}.$$

Recalling that $\lambda_{\min}(A(\gamma))$ is 1-Lipschitz in γ, we deduce that $|P| \geq \delta/2\kappa$. Note also that $\lambda_{\min}(A(\hat{\gamma})) \geq \xi \geq \delta \geq 3\delta/4\kappa$ thus P is a connected interval.

Suppose for the sake of contradiction that the algorithm fails to output in each of the T rounds. Note that $P \subseteq [s_0, t_0]$. We will show by induction that $P \subseteq [s_k, t_k]$ for every $k \in \{1, \dots, T\}$. Let $k \in \{0, \dots, T-1\}$. By assumption, the algorithm fails to output in round k. This can happen in two ways: If $x^\intercal A(\gamma)x < \delta/4\kappa$, then x certifies that $\gamma \notin P$ and $P \subseteq [s_k, \gamma]$. If $x^\intercal A(\gamma)x > \delta/\kappa$, then as we have conditioned on ApproxEig succeeding, $\lambda_{\min}(A(\gamma)) \geq \delta/\kappa - \delta/4\kappa$ and $P \subseteq [\gamma, t_k]$. In either case, we have that $P \subseteq [s_{k+1}, t_{k+1}]$.

We conclude that P, an interval of length at least $\delta/2\kappa$, is contained in $[s_T, t_T]$, an interval of length

$$t_T - s_T \leq \frac{\zeta}{2^T} \leq \delta/4\kappa,$$

a contradiction. Thus, the algorithm outputs within T rounds.

The running time of this algorithm follows from Lemma 52. ■

Remark 58. Similar algorithms for approximating $\gamma_\pm$ given $\hat{\gamma}$ have been proposed in the literature [2, 94, 124, 148]. However to our knowledge, this is the first analysis to establish an explicit convergence rate; see the discussion after Remark 2.11 in [94] on this issue. □

4.4.3 Minimizing the maximum of two quadratic functions

In this subsection, we will assume that Algorithm 1 has successfully found $\tilde{\gamma}_\pm$ satisfying (4.9) and show how to approximately solve

$$\min_{x\in\mathbb{R}^n} \max\{q(\tilde{\gamma}_-, x), q(\tilde{\gamma}_+, x)\}.$$

For the sake of readability, we will use the following notation in this subsection.

$$\tilde{f}_0(x) := q(\tilde{\gamma}_-, x) \quad \text{and} \quad \tilde{f}_1(x) := q(\tilde{\gamma}_+, x) \tag{4.12}$$

In particular we have $\tilde{f}(x) = \max\left\{\tilde{f}_0(x), \tilde{f}_1(x)\right\}$.

Our analysis is based on Nesterov [132, Section 2.3.3], which proposes a high level algorithm for minimizing general minimax problems with smooth components. We state this algorithm (Algorithm 3) and its corresponding convergence rate in our context.

Algorithm 3 Constant Step Scheme II for Smooth Minimax Problems [132, Algorithm 2.3.12]

```
Given continuously differentiable convex, $2L$-smooth functions $\tilde{f}_0, \tilde{f}_1$
    1. Let $x_0 = y_0 = 0$ and $\alpha_0 = 1/2$
    2. For $k = 0, 1, \dots$
        a) Compute $\tilde{f}_i(y_k)$ and $\nabla\tilde{f}_i(y_k)$ for $i = 0, 1$
        b) Compute
            $x_{k+1} = \arg\min_x \max_{i=0,1}\left(\tilde{f}_i(y_k) + \left\langle\nabla\tilde{f}_i(y_k), x - y_k\right\rangle + L\|x - y_k\|^2\right)$
            $\alpha_{k+1} = \frac{\sqrt{\alpha_k^4 + 4\alpha_k^2} - \alpha_k^2}{2}$
            $\beta_k = \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2 + \alpha_{k+1}}$
            $y_{k+1} = x_{k+1} + \beta_k(x_{k+1} - x_k)$
```

Theorem 23 ([132, Theorem 2.3.5])**.** *Let $\tilde{f}_0, \tilde{f}_1$ be $2L$-smooth[3] differentiable convex functions such that $\tilde{f}$ is bounded below. Let $\tilde{x}^*$ be an optimizer of $\tilde{f}$. Then the iterates x_k produced by Algorithm 3 satisfy*

$$\tilde{f}(x_k) - \tilde{f}(\tilde{x}^*) \leq \frac{8}{(k+1)^2}\left(\tilde{f}(0) - \tilde{f}(\tilde{x}^*) + \frac{L}{2}\|\tilde{x}^*\|^2\right).$$

Lemma 54. *Let $x \in \mathbb{R}^n$. Then for $i = 0, 1$, we have*

$$|q_i(0) - q_i(x)| \leq \|x\|^2 + 2\|x\|.$$

Proof. For $i = 0, 1$, we have

$$|q_i(0) - q_i(x)| = |q_i(x) - c_i| \leq \|A_i\|\|x\|^2 + 2\|b_i\|\|x\| \leq \|x\|^2 + 2\|x\|.$$

where the second inequality follows from Assumption 11. ■

Corollary 21. *Let $\tilde{f}_0$ and $\tilde{f}_1$ be the functions defined in (4.12). Let $\tilde{x}^*$ be an optimizer of $\tilde{f}$. Then the iterates x_k produced by Algorithm 3 satisfy*

$$\tilde{f}(x_k) - \tilde{f}(\tilde{x}^*) \leq \frac{760}{(k+1)^2}\kappa^2\zeta.$$

In particular, after $k = O\left(\kappa\sqrt{\zeta/\epsilon}\right)$ iterations, the solution x_k satisfies $\tilde{f}(x_k) - \tilde{f}(\tilde{x}^*) \leq \epsilon/2$.

Proof. We have that $\tilde{f}_0$ and $\tilde{f}_1$ are both $2(2\zeta)$-smooth by Assumption 11 and Definition 17. Moreover, $\tilde{f}(x) \geq q(\hat{\gamma}, x)$ is bounded below. Thus, we may apply Theorem 23.

We bound the initial primal gap as follows:

$$\begin{aligned}
\tilde{f}(0) - \tilde{f}(\tilde{x}^*) &= \max\left\{\tilde{f}_0(0), \tilde{f}_1(0)\right\} - \max\left\{\tilde{f}_0(\tilde{x}^*), \tilde{f}_1(\tilde{x}^*)\right\} \\
&\leq \max\left\{\tilde{f}_0(0) - \tilde{f}_0(\tilde{x}^*), \tilde{f}_1(0) - \tilde{f}_1(\tilde{x}^*)\right\} \\
&= q_0(0) - q_0(\tilde{x}^*) + \max\{\tilde{\gamma}_-(q_1(0) - q_1(\tilde{x}^*)),\ \tilde{\gamma}_+(q_1(0) - q_1(\tilde{x}^*))\} \\
&\leq |q_0(0) - q_0(\tilde{x}^*)| + \zeta|q_1(0) - q_1(\tilde{x}^*)| \\
&\leq (1+\zeta)\left(25\kappa^2 + 10\kappa\right) \\
&\leq 70\kappa^2\zeta,
\end{aligned}$$

where the third line follows from definition (see (4.12)), the fourth line follows from the ordering $\tilde{\gamma}_- \leq \tilde{\gamma}_+ \leq \zeta$, the fifth line follows from Lemmas 48 and 54, and the last line follows from the trivial bounds $\kappa \geq 1$ and $\zeta \geq 1$.

Using Lemma 48 again, we also have $\frac{L}{2}\|\tilde{x}^*\|^2 = \frac{2\zeta}{2}\|\tilde{x}^*\|^2 \leq 25\kappa^2\zeta$. The result follows by combining these bounds. ■

[3] Recall that a convex quadratic function $x^\intercal Ax + 2b^\intercal x + c$ is $2L$-smooth if and only if $A \preceq LI$.

It remains to analyze the runtime of each iteration. Aside from computation of x_{k+1}, it is clear that the quantities in each iteration can be computed in $O(N)$ time. Below, we derive a closed form expression for x_{k+1} where each of the quantities can be computed in $O(N)$ time.

Lemma 55. *For any $y \in \mathbb{R}^n$, the quantity*

$$\operatorname*{arg\,min}_x \max_{i=0,1} \left(\tilde{f}_i(y) + \left\langle \nabla \tilde{f}_i(y), x - y \right\rangle + L\|x - y\|^2 \right)$$

can be computed in $O(N)$ time.

Proof. Fix $y \in \mathbb{R}^n$. We begin by recentering the quadratic functions in the objective.

$$\begin{aligned}
&\max_{i=0,1} \left(\tilde{f}_i(y) + \left\langle \nabla \tilde{f}_i(y), x - y \right\rangle + L\|x - y\|^2 \right) \\
&= \max_{i=0,1} \left(L \left\| x - \left[y - \frac{1}{L}\frac{\nabla \tilde{f}_i(y)}{2} \right] \right\|^2 + \left[\tilde{f}_i(y) - \frac{1}{L}\left\| \frac{\nabla \tilde{f}_i(y)}{2} \right\|^2 \right] \right) \\
&=: \max_{i=0,1} \left(L\|x - z_i\|^2 + h_i \right)
\end{aligned}$$

Here, z_i and h_i are defined to be the square-bracketed terms from the preceding line. It is clear that the minimizing x must belong to the line segment $[z_0, z_1]$. We will parameterize $x = z_0 + \alpha(z_1 - z_0)$ where $\alpha \in [0, 1]$.

$$\begin{aligned}
&\min_x \max_{i=0,1} \left(L\|x - z_i\|^2 + h_i \right) \\
&= \min_{\alpha \in [0,1]} \max \left\{ \alpha^2 L\|z_0 - z_1\|^2 + h_0,\ (1 - \alpha)^2 L\|z_0 - z_1\|^2 + h_1 \right\}.
\end{aligned}$$

We solve for α by setting the two terms inside the maximum equal. A simple calculation yields that the two quadratics are equal when

$$\bar{\alpha} := \frac{1}{2} - \frac{h_0 - h_1}{2L\|z_0 - z_1\|^2}.$$

If $\bar{\alpha}$ is between $[0, 1]$, let $\alpha^* = \bar{\alpha}$. Else let $\alpha^* = 0$ (respectively $\alpha^* = 1$) when $\bar{\alpha} < 0$ (respectively $\bar{\alpha} > 1$).

Then,

$$\operatorname*{arg\,min}_x \max_{i=0,1} \left(\tilde{f}_i(y) + \left\langle \nabla \tilde{f}_i(y), x - y \right\rangle + L\|x - y\|^2 \right) = z_0 + \alpha^*(z_1 - z_0).$$

Each of the quantities on the right hand side (namely α^*, z_i) can be computed in $O(N)$ time. ■

Combining Corollary 21 and Lemma 55 gives the following corollary.

Corollary 22. *Let $\tilde{f}_0, \tilde{f}_1$ be the functions defined in (4.12). There exists an algorithm which outputs $\tilde{x}$ satisfying $\tilde{f}(\tilde{x}) \leq \widetilde{\mathrm{Opt}} + \epsilon/2$ running in time*

$$O\left(\frac{N\,\kappa\,\sqrt{\zeta}}{\sqrt{\epsilon}}\right).$$

Remark 59. Jiang and Li [94] present a saddle-point-based first-oder algorithm for approximating $\widetilde{\mathrm{Opt}}$. By instantiating their algorithm with the initial iterate $x_0 = 0$ and applying our Lemma 48 to bound $\|x_0 - \tilde{x}^*\|^2$, we have that [94, Algorithm 1] produces an $\epsilon/2$-optimal solution to the convex reformulation in time

$$O\left(\frac{N\kappa^2\zeta}{\epsilon}\right).$$

Therefore, the dependences on ϵ, κ, and ζ of this algorithm are worse than that of the algorithm described in Corollary 22. Note that [94] does not present an analysis of the complexity of finding the approximate generalized eigenvalue $\tilde{\gamma}_\pm$ (needed to construct $\widetilde{\mathrm{Opt}}$) or how $\widetilde{\mathrm{Opt}}$ relates to Opt. □

By combining Lemmas 47, 49, and 53 and Corollary 22, we arrive at the following main theorem on the overall computational complexity of our approach.

Theorem 24. *Given q_0, q_1 satisfying Assumption 11, (ξ, ζ) and $\hat{\gamma}$ satisfying Assumption 12, error parameter $0 < \epsilon \leq \kappa^2\xi$, and failure probability $p > 0$, ApproxConvex (Algorithm 1) outputs $\tilde{\gamma}_-$, $\tilde{\gamma}_+$ and $\tilde{x} \in \mathbb{R}^n$ such that*

$$\mathrm{Opt} \leq \max\{q(\tilde{\gamma}_-, \tilde{x}), q(\tilde{\gamma}_+, \tilde{x})\} \leq \widetilde{\mathrm{Opt}} + \epsilon/2 \leq \mathrm{Opt} + \epsilon$$

with probability $1 - p$. This algorithm runs in time

$$\tilde{O}\left(\frac{N\kappa^{3/2}\sqrt{\zeta}}{\sqrt{\epsilon}} \log\left(\frac{n}{p}\right) \log\left(\frac{\kappa}{\epsilon}\right)\right).$$

4.4.4 Finding an approximate optimizer of the GTRS

Let $\tilde{x} \in \mathbb{R}^n$ be the approximate optimizer output by Algorithm 1. In this subsection, we show how to use $\tilde{x}$ to construct an $\bar{x}$ approximately minimizing the original GTRS (4.1). Our algorithm will follow the proof of Theorem 19 (in particular Lemma 44).

We present our algorithm, ApproxGTRS, as Algorithm 4. ApproxGTRS will use ApproxConvex as a subroutine. Given an additive error $\epsilon_{\mathrm{round}}$, ApproxGTRS will call ApproxConvex with additive error $\epsilon_{\mathrm{convex}}$. We will write these parameters as ϵ_r and ϵ_c for short.

Note that by Definition 17, we have $\kappa^3\xi \geq 1$. Thus, as before, the requirement $0 < \epsilon_r \leq \kappa^3\xi$ in Algorithm 4 is not a practical issue: given $\epsilon_r > \kappa^3\xi$, we can simply run our algorithm with $\epsilon_r' = \kappa^3/\xi$ and return a solution with a better error guarantee.

Algorithm 4 ApproxGTRS$(q_0, q_1, \xi, \zeta, \hat\gamma, \epsilon_r, p_r)$

```
Given $q_0$ and $q_1$ satisfying Assumption 11, $(\xi, \zeta)$ and $\hat\gamma$ satisfying Assumption 12, error parameter
$0 < \epsilon_r \le \kappa^3\xi$, and failure probability $p_r > 0$
  1. Define $\epsilon_c := \epsilon_r/(28\kappa)$
  2. Let $\tilde\gamma_-$, $\tilde\gamma_+$ and $\tilde x$ be the output of ApproxConvex$(q_0, q_1, \xi, \zeta, \hat\gamma, \epsilon_c, p_r/2)$
  3. If $q_1(\tilde x) = 0$ then return $\bar x = \tilde x$
  4. Else if $q_1(\tilde x) > 0$
     a) Let $d :=$ ApproxEig$(A(\tilde\gamma_+), 2\zeta, \delta/\kappa, p_r/2)$
     b) Let $e := 2(\tilde x^\intercal A(\tilde\gamma_+)d + b(\tilde\gamma_+)^\intercal d)$
     c) If necessary, take $d \leftarrow -d$ and $e \leftarrow -e$ to ensure that $e \le 0$
     d) Let $\alpha \ge 0$ be the nonnegative solution to
            $q(\tilde\gamma_-, \tilde x + \alpha d) = q(\tilde\gamma_+, \tilde x + \alpha d)$
     e) Return $\bar x = \tilde x + \alpha d$
  5. Else carry out the computation in step 4 where the roles of $\tilde\gamma_-$ and $\tilde\gamma_+$ are interchanged
```

The next lemma bounds $\|\tilde x\|$. Its proof follows the proof of Lemma 48 with minor adjustments (in particular, the upper bound of (4.11) is replaced with $\tilde f(\tilde x) \le \max\{c(\gamma_-), c(\gamma_+)\} + \epsilon/2$; see Corollary 22) and is omitted.

Lemma 56. *Let $\tilde x \in \mathbb{R}^n$ satisfy $\tilde f(\tilde x) \le \widetilde{\mathrm{Opt}} + \epsilon/2$. Then $\tilde x \in B(0, 6\kappa)$.*

We are now ready to prove a formal guarantee on Algorithm 4.

Theorem 25. *Given q_0, q_1 satisfying Assumption 11, (ξ, ζ) and $\hat\gamma$ satisfying Assumption 12, error parameter $0 < \epsilon_r \le \kappa^3\xi$, and failure probability p_r, ApproxGTRS (Algorithm 4) outputs $\bar x$ such that*

$$q_0(\bar x) \le \mathrm{Opt} + \epsilon_r$$
$$q_1(\bar x) = 0$$

with probability $1 - p_r$. This algorithm runs in time

$$\tilde O\left(\frac{N\kappa^2\sqrt{\zeta}}{\sqrt{\epsilon_r}} \log\left(\frac{n}{p_r}\right) \log\left(\frac{\kappa}{\epsilon_r}\right)\right).$$

Proof. We condition on the event that Algorithm 1 succeeds and the ApproxEig call in step 4.(a) or 5.(a) succeeds. By the union bound, this happens with probability at least $1 - p_r$. As in Lemma 44, we will split the analysis into three cases: (i) $q_1(\tilde x) = 0$, (ii) $q_1(\tilde x) > 0$, and (iii) $q_1(\tilde x) < 0$.

i If $q_1(\tilde x) = 0$ then $q_0(\tilde x) = \tilde f(\tilde x) \le \widetilde{\mathrm{Opt}} + \epsilon_c/2 \le \mathrm{Opt} + \epsilon_c \le \mathrm{Opt} + \epsilon_r$.

ii Now suppose $q_1(\tilde x) > 0$, i.e., we are in step 4 of Algorithm 4. We will need an upper bound on the value of α found in step 4.(d).

Let $t \coloneqq q(\tilde{\gamma}_+, \tilde{x})$. Recall that $\lambda_{\min}(A(\tilde{\gamma}_+)) \in [0, \delta/\kappa]$ (see Lemma 53). Then, as we have conditioned on the ApproxEig call in step 4.(a) succeeding, we have

$$\begin{aligned} q(\tilde{\gamma}_+, \tilde{x} + \alpha d) - (t + \alpha e) &= (q(\tilde{\gamma}_+, \tilde{x}) - t) + \alpha^2 d^\intercal A(\tilde{\gamma}_+) d \\ &\leq \alpha^2 (2\delta/\kappa). \end{aligned} \tag{4.13}$$

Next, we give a lower bound on $q(\tilde{\gamma}_-, \tilde{x}+\alpha d) - (t+\alpha e)$ using the estimate $d^\intercal A(\tilde{\gamma}_-) d \geq \xi$, and routine estimates on $\|A(\gamma)\|$ and $\|b(\gamma)\|$:

$$\begin{aligned} &q(\tilde{\gamma}_-, \tilde{x} + \alpha d) - (t + \alpha e) \\ &\qquad = (q(\tilde{\gamma}_-, \tilde{x}) - t) + 2\alpha(\tilde{x}^\intercal A(\tilde{\gamma}_-) d + b(\tilde{\gamma}_-)^\intercal d - e/2) + \alpha^2 d^\intercal A(\tilde{\gamma}_-) d \\ &\qquad \geq -|\tilde{\gamma}_+ - \tilde{\gamma}_-| |q_1(\tilde{x})| \\ &\qquad\quad - 2\alpha(\|\tilde{x}\| \|A(\tilde{\gamma}_-)\| + \|b(\tilde{\gamma}_-)\| + \|\tilde{x}\| \|A(\tilde{\gamma}_+)\| + \|b(\tilde{\gamma}_+)\|) \\ &\qquad\quad + \alpha^2 \xi \\ &\qquad \geq -49\kappa^2 \zeta - 2\alpha(14\kappa\zeta) + \alpha^2 \xi, \end{aligned}$$

where the last inequality follows from the bounds $|\tilde{\gamma}_+ - \tilde{\gamma}_-| \leq \gamma_+ \leq \zeta$ (Definition 17), $\|\tilde{x}\| \leq 6\kappa$ (Lemma 56), and Lemma 54.

We may combine our upper and lower bounds to deduce that for any $\alpha \in \mathbb{R}$,

$$\begin{aligned} q(\tilde{\gamma}_-, \tilde{x} + \alpha d) - q(\tilde{\gamma}_+, \tilde{x} + \alpha d) &\geq \alpha^2 (\xi - 2\delta/\kappa) - 2\alpha(14\kappa\zeta) - 49\kappa^2\zeta \\ &\geq \alpha^2 \left(\frac{35}{36}\xi\right) - 2\alpha(14\kappa\zeta) - 49\kappa^2\zeta, \end{aligned}$$

where the last relation follows from the definition of δ in (4.8), the definition of ϵ_c and the assumption on ϵ_r, we have $\epsilon_c \leq \kappa^2 \xi$, and the bound $\kappa \geq 1$. In particular, because the quadratic function on the left is negative at $\alpha = 0$ and is lower bounded by a strongly convex quadratic function, there must exist both positive and negative choices of α for which the left hand side takes the value zero. This justifies step 4.(d) of the algorithm.

We now fix α to be the positive solution to $q(\tilde{\gamma}_-, \tilde{x} + \alpha d) = q(\tilde{\gamma}_+, \tilde{x} + \alpha d)$ so that

$$0 \geq \alpha^2 \left(\frac{35}{36}\xi\right) - 2\alpha(14\kappa\zeta) - 49\kappa^2\zeta.$$

We get an upper bound on α by the quadratic formula

$$\alpha \leq \frac{14\kappa\zeta + \sqrt{(14\kappa\zeta)^2 + \frac{49\cdot 35}{36}\kappa^2\zeta\xi}}{\frac{35}{36}\xi} \leq 31\kappa^2.$$

Then, by defining $\bar{x} := \tilde{x} + \alpha d$, we have $q(\tilde{\gamma}_-, \bar{x}) = q(\tilde{\gamma}_+, \bar{x})$. Note that the containment $\hat{\gamma} \in (\tilde{\gamma}_-, \tilde{\gamma}_+)$ from (4.10) implies $\tilde{\gamma}_- \neq \tilde{\gamma}_+$. Then we deduce $q_1(\bar{x}) = 0$. Moreover, our upper bound (4.13) gives

$$\begin{aligned} q_0(\bar{x}) &= q(\tilde{\gamma}_+, \bar{x}) \\ &\leq t + \alpha e + \alpha^2(2\delta/\kappa). \end{aligned}$$

Then recalling that $t := q(\tilde{\gamma}_+, \tilde{x}) \leq \widetilde{\text{Opt}} + \epsilon_c/2 \leq \text{Opt} + \epsilon_c$ and that we picked $e \leq 0$, we bound

$$\begin{aligned} q_0(\bar{x}) &\leq \text{Opt} + \epsilon_c + (31\kappa^2)^2(2\delta/\kappa) \\ &\leq \text{Opt} + \epsilon_c + 27\kappa\epsilon_c \\ &\leq \text{Opt} + 28\kappa\epsilon_c \\ &= \text{Opt} + \epsilon_r. \end{aligned}$$

iii The final case is symmetric to case (ii) and is omitted.

The running time of this algorithm follows from Lemma 52 and Theorem 24. ■

Remark 60. Let us now compare the running time of our algorithms to the running time of the algorithm presented by Jiang and Li [95]. This algorithm takes as input a pair q_0, q_1 satisfying an assumption similar to our Assumption 11 and a regularity parameter ξ_{JL}. See Remark 56 for a discussion of how the parameter ξ_{JL} relates to our regularity parameters (ξ_{us}, ζ). Then given $\epsilon > 0$ and $p > 0$, this algorithm returns an ϵ-optimal feasible solution with probability at least $1 - p$. The running time of this algorithm is

$$\tilde{O}\left(\frac{N\phi^3}{\sqrt{\epsilon\, \xi_{\text{JL}}^5}} \log\left(\frac{n}{p}\right) \log\left(\frac{\phi}{\epsilon\, \xi_{\text{JL}}}\right)\right),$$

where ϕ is a computable regularity parameter.

Recall that in Remark 56, we constructed simple examples where $\xi_{\text{JL}} \approx 1/2\kappa$ and $\zeta \approx 1$. One can check that the regularity parameter ϕ is a constant on these examples. In particular, the analysis presented in Jiang and Li [95] implies a running time of

$$\tilde{O}\left(\frac{N\kappa^{5/2}}{\sqrt{\epsilon}} \log\left(\frac{n}{p}\right) \log\left(\frac{\kappa}{\epsilon}\right)\right)$$

on these instances. We contrast this with the running times

$$\tilde{O}\left(\frac{N\kappa^{3/2}}{\sqrt{\epsilon}} \log\left(\frac{n}{p}\right) \log\left(\frac{\kappa}{\epsilon}\right)\right), \qquad \tilde{O}\left(\frac{N\kappa^{2}}{\sqrt{\epsilon}} \log\left(\frac{n}{p}\right) \log\left(\frac{\kappa}{\epsilon}\right)\right)$$

of our Algorithms 1 and 4 for finding an ϵ-optimal value, and an ϵ-optimal feasible solution respectively on these instances. □

Remark 61. Algorithms 1 and 4 were designed and analyzed with worst-case guarantees in mind. Consequently, we have not been particularly careful about bounding the constants in our analysis (for example the bounds $\kappa, \zeta, \xi^{-1} \leq 1$ are routinely used). As such, there may be variants of our algorithms that achieve the *same* worst-case guarantees with significantly faster numerical performance. Similarly, the algorithm presented by Jiang and Li [95] is analyzed with worst-case guarantees in mind. They also remark that the the numerical performance of their algorithm may improve "with suitable modifications" (see Jiang and Li [95, Remark 4.2]).

We leave such implementation questions and a thorough comparison of the numerical performance of the algorithms present in the literature for future work. □

4.4.5 Further remarks

The algorithms given in the prior subsections can be sped up substantially if we know how to compute $\gamma_\pm$ and the corresponding zero eigenvectors exactly. As an example, we consider the special case where A_0 and A_1 are diagonal matrices.

Lemma 57. *There exists an algorithm which given q_0, q_1 satisfying Assumption 11 with A_0 and A_1 diagonal, returns $\gamma_\pm$, (ξ^*, ζ^*) and γ^* such that $\lambda_{\min}(A(\gamma^*)) = \xi^*$ in time $O(n)$.*

Proof. Let $a_0, a_1 \in \mathbb{R}^n$ be the diagonal entries of A_0 and A_1 respectively. Note that

$$\lambda_{\min}(A(\gamma)) = \min_{i\in[n]}\{a_{0,i} + \gamma a_{1,i}\}.$$

Thus, $\gamma_\pm$ and ζ^* can clearly be computed in $O(n)$ time. Note that

$$\xi^* = \max_{\gamma,\xi}\left\{\xi : \begin{array}{l} \forall i \in [n],\ a_{0,i} + \gamma a_{1,i} \geq \xi \\ \xi \geq 0 \end{array}\right\}.$$

Hence, ξ^* and γ^* are, respectively, the optimal value and solution to a two-variable linear program with n constraints. Applying the algorithm by Megiddo [120] for two-variable linear programming allows us to solve for ξ^* and γ^* in $O(n)$ time. ■

Corollary 23. *There exists an algorithm which given q_0, q_1 satisfying Assumption 11 with A_0 and A_1 diagonal and error parameter $\epsilon > 0$, outputs $\bar{x} \in \mathbb{R}^n$ such that*

$$\begin{aligned} q_0(\bar{x}) &\leq \mathrm{Opt} + \epsilon \\ q_1(\bar{x}) &= 0. \end{aligned}$$

This algorithm runs in time

$$O\left(\frac{n\kappa^*\sqrt{\zeta^*}}{\sqrt{\epsilon}}\right).$$

Proof. When A_0 and A_1 are diagonal we have $N \leq 2n$. By Lemma 57, we can compute all of the quantities needed for the exact convex reformulation in $O(n)$ time. Algorithm 3 can then be applied to the exact convex reformulation to find $\tilde{x} \in \mathbb{R}^n$ with

$$\max\{q(\gamma_-, \tilde{x}), q(\gamma_+, \tilde{x})\} \leq \mathrm{Opt} + \epsilon.$$

We can further carry out the modification procedure of Lemma 44 exactly in $O(n)$ time.

The running time of this algorithm follows from Corollary 22. ■

5 Implicit regularity in the generalized trust-region subproblem

This chapter is based on joint work [183] with Yunlei Lu and Fatma Kılınç-Karzan.

In this chapter we develop efficient first-order algorithms for the *generalized trust-region subproblem* (GTRS), which has applications in signal processing, compressed sensing, and engineering. Although the GTRS, as stated, is nonlinear and nonconvex, it is well-known that objective value exactness holds for its SDP relaxation under a Slater condition. While polynomial-time SDP-based algorithms exist for the GTRS, their relatively large computational complexity has motivated and spurred the development of custom approaches for solving the GTRS. In particular, recent work in this direction has developed first-order methods for the GTRS whose running times are linear in the sparsity (the number of nonzero entries) of the input data. In contrast to these algorithms, in this chapter we develop algorithms for computing ϵ-approximate solutions to the GTRS whose running times are linear in both the input sparsity *and* the precision $\log(1/\epsilon)$ whenever a regularity parameter is positive. We complement our theoretical guarantees with numerical experiments comparing our approach against algorithms from the literature. Our numerical experiments highlight that our new algorithms significantly outperform prior state-of-the-art algorithms on sparse large-scale instances.

5.1 Introduction

In this chapter we develop efficient first-order algorithms for the *generalized trust-region subproblem* (GTRS). Recall the GTRS,

$$\mathrm{Opt} := \inf_{x\in\mathbb{R}^n} \{q_0(x) : \, q_1(x) \leq 0\},$$

where $q_0(x)$ and $q_1(x)$ are quadratic functions in $x \in \mathbb{R}^n$. We will assume that for each $i \in \{0, 1\}$, the quadratic function $q_i(x)$ is given by $q_i(x) = x^\intercal A_i x + 2b_i^\intercal x + c_i$ for $A_i \in \mathbb{S}^n$, $b_i \in \mathbb{R}^n$ and $c_i \in \mathbb{R}$.

This problem generalizes the classical *trust-region subproblem* (TRS) where the general quadratic constraint $q_1(x) \leq 0$ is replaced with the unit ball constraint $\|x\|^2 \leq 1$. The TRS finds applications, for example, in robust optimization [21, 87] and combinatorial optimization [98, 140]. The TRS is additionally foundational in the area of nonlinear programming. Indeed, iterative algorithms based on the TRS (known sometimes as trust-region methods) [50] are among the most empirically successful techniques for general nonlinear programs.

Generalizing the TRS, the GTRS has applications in signal processing, compressed sensing, and engineering (see [180] and references therein). The problem of minimizing a quartic of the form $q(x, p(x))$, where $q : \mathbb{R}^{n+1} \to \mathbb{R}$ and $p : \mathbb{R}^n \to \mathbb{R}$ are both quadratic, can be cast in the equality-constrained variant of the GTRS. This approach has been used to address source localization [86] as well as the double-well potential functions [65]. More broadly, iterative ADMM-based algorithms for general QCQPs using the GTRS as a subprocedure have shown exceptional numerical performance [90] and outperform previous state-of-the-art approaches on a number of real world problems (e.g., multicast beamforming and phase retrieval). This application of the GTRS as a subprocedure within an iterative solver parallels the use of the TRS within trust-region methods.

Although the GTRS, as stated, is nonlinear and nonconvex, it is well-known that objective value exactness holds for its SDP relaxation under a Slater condition [67, 146]. Thus, unlike general QCQPs which are NP-hard, the GTRS can be solved in polynomial time via SDP-based algorithms. Nevertheless, the relatively large computational complexity of SDP-based approaches has motivated and spurred the development of alternative custom approaches for solving the GTRS. We restrict our discussion below to *recent* trends in GTRS algorithms and discuss *earlier* work [124, 125, 165] where appropriate in the main body.

One line of proposed algorithms for the GTRS assumes simultaneous diagonalizability (SD) of A_0 and A_1. It is well-known that SD holds under minor conditions—for example, if there exists a positive definite matrix in $\operatorname{span}\{A_0, A_1\}$ (see [177] for additional variants of this result). Ben-Tal and Teboulle [24] exploit the SD condition to provide a reformulation of the interval-constrained GTRS as a convex minimization problem with linear constraints. More recently, under the SD condition, Ben-Tal and den Hertog [21] provide a second-order cone program (SOCP) reformulation of the GTRS in a lifted space. This SOCP reformulation was generalized beyond the GTRS in [111]. Under the SD condition, a number of papers [64, 154] exploit the resulting problem structure of the primal or the dual formulation to derive solution procedures for the GTRS and interval-constrained GTRS. Generalizing [21], Jiang et al. [96] provide an SOCP reformulation for the GTRS in a lifted space whenever the problem has a finite optimal value even when the SD condition fails. Unfortunately, the algorithms in this line often assume implicitly that A_0 and A_1 are already diagonal or that a simultaneously-diagonalizing basis can be computed. In practice, however, computing such a basis requires a full eigen-decomposition and can be prohibitively expensive for large-scale instances.

A second line of research on the GTRS explores the connection between the GTRS and generalized eigenvalues of the matrix pencil $A_0 + \gamma A_1$. Pong and Wolkowicz [148] propose a generalized-eigenvalue-based algorithm which exploits the structure of optimal GTRS solutions, albeit without an explicit running time analysis. Adachi and Nakatsukasa [2] present another approach for solving the GTRS based on computing the minimum generalized eigenvalue (and corresponding eigenvector) of an associated *indefinite* $(2n + 1) \times (2n + 1)$ matrix pencil. Unfortunately, this approach suffers from the significant cost of computing a minimum generalized eigenvalue of an indefinite matrix pencil. Empirically, the complexity of this approach scales as $O(n^2)$ even for sparse instances of the GTRS with $O(n)$ nonzero entries in A_0 and A_1 (see [2, Section 4]). Jiang and Li [94] reformulate the GTRS as the problem of minimizing the maximum of two convex quadratic functions in the original space. This reformulation is constructed from a pair of generalized eigenvalues related to the matrix pencil $A_0 + \gamma A_1$. They then suggest a saddle-point-based first-order algorithm to solve this reformulation within an ϵ additive error in $O(1/\epsilon)$ time. These approaches are based on the assumption that the generalized eigenvalues are given or can be computed exactly, and offer no theoretical guarantees when only approximate generalized eigenvalue computations are available (as is the case in practice; see also the discussion in Section 4.1 in [95]). Despite this, the numerical experiments in [2, 94, 148] suggest that algorithms motivated by these ideas perform well even using only approximate generalized eigenvalue computations.

In contrast to these papers, recent work [95, 180] offers provably linear-time (in terms of the number of nonzero entries in the input data) algorithms for the GTRS using only approximate eigenvalue procedures. Jiang and Li [95] extend ideas developed in [82] for solving the TRS to derive an algorithm for solving the GTRS up to an ϵ additive error with high probability. This approach differs from the earlier literature in that it does not rely on the computation of a simultaneously-diagonalizing basis or exact generalized eigenvalues. The complexity of this approach is

$$\tilde{O}\left(\frac{N}{\sqrt{\epsilon}}\log\left(\frac{n}{p}\right)\log\left(\frac{1}{\epsilon}\right)^2\right),$$

where N is the number of nonzero entries in A_0 and A_1, ϵ is the additive error, and p is the failure probability. Here, we have elided quantities related to the condition number of the GTRS. Wang and Kılınç-Karzan [180] reexamine the convex quadratic reformulation idea of [94] and show formally that by approximating the generalized eigenvalues sufficiently well, the perturbed convex reformulation is within a small additive error of the true convex reformulation. Moreover, they establish that the resulting convex reformulation can be solved via Nesterov's accelerated gradient descent method [132, Section 2.3.3] for smooth minimax problems to achieve an overall run time guarantee of

$$\tilde{O}\left(\frac{N}{\sqrt{\epsilon}}\log\left(\frac{n}{p}\right)\log\left(\frac{1}{\epsilon}\right)\right).$$

A parallel line of work [41, 66, 77, 82, 87, 125] has developed custom first-order methods for the trust-region subproblem. Most relatedly, Carmon and Duchi [41] recently showed that a Krylov-based first-order method can achieve a convergence rate for the TRS that is linear in both N *and* the precision $\log(1/\epsilon)$ whenever a regularity parameter, μ^*, is positive. This contrasts with previous algorithms for the TRS whose guarantees scaled as $\approx 1/\sqrt{\epsilon}$.

In this chapter, we introduce and analyze a *new* algorithm for computing an ϵ-approximate solution to the GTRS whose running time is linear in both N and the precision $\log(1/\epsilon)$ whenever μ^* is positive. To be concrete, an ϵ-approximate solution is defined below.

Definition 18. We say $x \in \mathbb{R}^n$ is an *ϵ-approximate solution* to (5.1) if

$$q_0(x) \le \mathrm{Opt} + \epsilon \quad \text{and} \quad q_1(x) \le \epsilon. \qquad \square$$

Despite similar convergence guarantees, our approach for solving the GTRS does not share many algorithmic similarities with the approach of Carmon and Duchi [41] for the TRS.

5.1.1 Overview and outline of this chapter

A summary of our contributions, along with an outline of the remainder of the chapter, is as follows:

- In Section 5.2, we recall definitions and results related to the Lagrangian dual of the GTRS and define our notion of regularity. Specifically, we recall definitions and results in the literature [2, 65, 124, 125] regarding the dual function $\mathbf{d}(\gamma)$ and its derivative $\nu(\gamma)$. We then define a *regularity* parameter μ^*, which will play the role of strong convexity in our algorithms. We close with a key lemma (Lemma 60) that underpins the algorithms developed in this chapter. Intuitively, Lemma 60 says that when μ^* is positive, the unique optimizer of the GTRS is stable—an $\Omega(\mu^*)$-strongly convex reformulation of the GTRS, whose unique optimizer coincides with the GTRS optimizer, can be built using *inexact* estimates of the dual optimizer γ^*.

- In Section 5.3, we describe and analyze an approach for computing an ϵ-approximate optimizer of a nonconvex-nonconvex GTRS instance based on Lemma 60. Our approach consists of two algorithms, `ConstructReform` and `SolveRegular`. The first algorithm uses inexact estimates of $\nu(\gamma)$ to binary search for an inexact estimate of γ^*. `ConstructReform` will either return an *exact* $\Omega(\mu^*)$-strongly convex reformulation of the GTRS or an ϵ-approximate optimizer of the GTRS. In the former case, we may then apply `SolveRegular` to compute an ϵ-approximate optimizer. In the latter case, `ConstructReform` will additionally *attempt* to certify that $\mu^* = O(\epsilon)$ so that building an $\Omega(\mu^*)$-strongly convex reformulation may be undesirable. Together, these two algorithms achieve the following linear convergence rate (i.e., scaling as $\log(1/\epsilon)$) for the GTRS:

  $$\tilde{O}\left(\frac{N}{\sqrt{\phi}}\log\left(\frac{1}{\phi}\right)\log\left(\frac{n}{p}\right)\log\left(\frac{1}{\epsilon}\right)\right).$$

  Here, N is the number of nonzero entries in A_0 and A_1 combined, ϕ can be thought of as $\approx \max(\mu^*, \epsilon)$ (see Section 5.3 for a formal definition), p is the failure probability, and the $\tilde{O}$-notation hides log log-factors. This contrasts with previous algorithms [95, 180] that are described as "linear-time", referring to the fact that their running times scale linearly in only N. We close this section by examining in further detail the case where `ConstructReform` returns an ϵ-approximate optimizer but fails to certify that $\mu^* = O(\epsilon)$. Specifically, we show that this edge case can only happen if $\nu(\gamma)$ is "extremely flat," which in turn can only happen if a certain *coherence* parameter is small.

- In Section 5.4, we present numerical experiments comparing the algorithms of Section 5.3 to other algorithms proposed in the recent literature [2, 21, 94]. Our numerical experiments corroborate our theoretical understanding of the situation—the algorithms in this chapter significantly outperform prior state-of-the-art algorithms on sparse large-scale GTRS instances.

5.1.2 Additional notation

For $x \in \mathbb{R}$ and $y \geq 0$ let $[\pm y] \coloneqq [-y, +y]$ and $[x \pm y] \coloneqq [x - y, x + y]$. For $\gamma \in \mathbb{R}_+$, define $A(\gamma) \coloneqq A_0 + \gamma A_1$, $b(\gamma) \coloneqq b_0 + \gamma b_1$, and $c(\gamma) \coloneqq c_0 + \gamma c_1$. Let $q(\gamma, x) \coloneqq q_0(x) + \gamma q_1(x)$. For $A \in \mathbb{S}^n$, let $\|A\|$ be its spectral norm. For $b \in \mathbb{R}^n$, let $\|b\|$ be its Euclidean norm.

5.2 Implicit Regularity in the GTRS

Recall that the GTRS is the problem of minimizing a quadratic objective function subject to a single quadratic constraint, i.e.,

$$\mathrm{Opt} \coloneqq \inf_{x \in \mathbb{R}^n} \{q_0(x) : \, q_1(x) \leq 0\}, \tag{5.1}$$

where for each $i \in \{0, 1\}$, we have $q_i(x) = x^\intercal A_i x + 2b_i^\intercal x + c_i$ for some $A_i \in \mathbb{S}^n$, $b_i \in \mathbb{R}^n$, and $c_i \in \mathbb{R}$.

We will make the following *blanket* assumption, which is both natural and common in the literature on the GTRS [2, 93, 95, 180]. This assumption can be thought of as primal and dual strict feasibility assumptions or a Slater assumption.

Assumption 13. There exists $\bar{x} \in \mathbb{R}^n$ such that $q_1(\bar{x}) < 0$ and there exists $\bar{\gamma} \geq 0$ such that $A(\bar{\gamma}) \succ 0$. □

Remark 62. Note, for example, that Assumption 13 holds in the classical TRS setting where $q_1(x) = x^\intercal x - 1$. Indeed, $q_1(0) < 0$ and $A(\gamma) = A_0 + \gamma I \succ 0$ for all γ large enough. □

The results and definitions will assume only Assumption 13. In particular, they can be applied to both the classical TRS setting as well as the nonconvex-nonconvex GTRS setting of Section 5.3.

Let $\Gamma \coloneqq \{\gamma \in \mathbb{R}_+ : \, A(\gamma) \succeq 0\}$. This is a closed interval as the positive semidefinite cone is closed. If Γ is bounded, let $[\gamma_-, \gamma_+]$ denote its left and right endpoints. Else, let γ_- denote its left endpoint and define $\gamma_+ \coloneqq +\infty$. Note that for any $\gamma \in \Gamma$, $q(\gamma, x)$ is a convex function of x. Furthermore, by the existence of $\bar{\gamma} \geq 0$ such that $A(\bar{\gamma}) \succ 0$, we have that $0 \leq \gamma_- < \gamma_+$.

Definition 19. Let $\mathbf{d} : \mathbb{R}_+ \to \{-\infty\} \cup \mathbb{R}$ denote the extended-real-valued function defined by

$$\mathbf{d}(\gamma) \coloneqq \inf_{x \in \mathbb{R}^n} q(\gamma, x). \qquad \square$$

We make the following observations on $\mathbf{d}(\gamma)$.

Observation 4. *Suppose Assumption 13 holds. Then,*

- *The function $\mathbf{d}(\gamma)$ is concave as it is the infimum of affine functions of γ.*
- *The function $\mathbf{d}(\gamma)$ is continuous on $\operatorname{int}(\Gamma)$. Furthermore, $\lim_{\gamma \searrow \gamma_-} \mathbf{d}(\gamma) = \mathbf{d}(\gamma_-)$ and, if γ_+ is finite, then $\lim_{\gamma \nearrow \gamma_+} \mathbf{d}(\gamma) = \mathbf{d}(\gamma_+)$.*
- *For $\gamma \in \mathbb{R}_+ \setminus \Gamma$, the function $q(\gamma, x)$ is nonconvex in x so that $\mathbf{d}(\gamma) = -\infty$.*
- *As $q_1(\bar{x}) < 0$, we have $\mathbf{d}(\gamma) \leq q(\gamma, \bar{x}) \to -\infty$ as $\gamma \to \infty$.*

We comment on the connection between $\mathbf{d}(\gamma)$, the SDP relaxation of (5.1), and the Lagrangian dual of (5.1). One consequence of the S-lemma [67] is that the GTRS has an exact SDP relaxation. Furthermore, it is well-known that the SDP relaxation of a general quadratically constrained quadratic program is equivalent to its Lagrangian dual [22]. We will write this fact in our setting as the following identity (which holds under Assumption 13),

$$\mathrm{Opt} = \inf_{x\in\mathbb{R}^n} \sup_{\gamma\in\Gamma} q(\gamma, x). \tag{5.2}$$

We provide a short self-contained proof of this fact in Section E.3. Next, by coercivity [62, Proposition VI.2.3] we have that

$$\mathrm{Opt} = \sup_{\gamma\in\Gamma} \inf_{x\in\mathbb{R}^n} q(\gamma, x) = \sup_{\gamma\in\Gamma} \mathbf{d}(\gamma) = \sup_{\gamma\in\mathbb{R}_+} \mathbf{d}(\gamma). \tag{5.3}$$

In words, (5.2) shows that the GTRS can be written as a convex minimization problem. Specifically, we can write Opt in one of the two following ways, corresponding respectively to the cases $\gamma_+ < \infty$ and $\gamma_+ = \infty$:

$$\mathrm{Opt} = \inf_{x\in\mathbb{R}^n} \max(q(\gamma_-, x), q(\gamma_+, x)) \quad \text{or} \quad \mathrm{Opt} = \inf_{x\in\mathbb{R}^n} \{q(\gamma_-, x) : q_1(x) \leq 0\}. \tag{5.4}$$

Note in the latter case that $A_1 \succeq 0$ so that $q_1(x) \leq 0$ is a convex constraint. Similarly, (5.3) shows that the GTRS can be written as a concave maximization problem.

Remark 63. The reformulation of the GTRS given in (5.4) immediately suggests an algorithm for approximating Opt: Compute γ_- (and if necessary γ_+) up to some accuracy and solve the resulting convex reformulation. Convergence guarantees along with rigorous error analyses for such an algorithm were previously explored by Wang and Kılınç-Karzan [180]. One drawback to this approach is that the convex functions $q(\gamma_-, x)$ and $q(\gamma_+, x)$ are, by construction, *not* both strongly convex unless $A_0, A_1 \succ 0$. Thus, in view of oracle lower bounds for first-order-methods [132, Chapter 2.1.2], one should not expect to achieve linear convergence rates via this approach. Similarly, the reformulation of the GTRS given in (5.3) immediately suggests an algorithm for approximating Opt: apply a root-finding algorithm or binary search to find γ^*. This approach dates back to Moré and Sorenson [125] for the TRS and Moré [124] for the GTRS (see also [2, 65]). Unfortunately, theoretical convergence rates have not been established for algorithms of this form. □

We will combine both ideas above to construct strongly convex reformulations for instances of (5.1) possessing *regularity*. Our notion of regularity will correspond to properties of $\mathbf{d}(\gamma)$ and its optimizers. We will need the following notation.

Definition 20. For $\gamma \in \mathrm{int}(\Gamma)$, define

$$x(\gamma) := -A(\gamma)^{-1}b(\gamma), \quad \nu(\gamma) := q_1(x(\gamma)), \quad \text{and} \quad \mu(\gamma) := \lambda_{\min}(A(\gamma)). \qquad \square$$

The functions $\mathbf{d}(\gamma)$, $x(\gamma)$, and $\nu(\gamma)$ have been studied previously in the literature on algorithms for the TRS and the GTRS [2, 65, 124, 125]. In contrast to previous algorithms in this line of work, which propose methods for computing γ^* to high accuracy, the algorithms we present in this chapter will work with relatively inaccurate estimates of γ^*. Specifically, our algorithms are inspired by a key lemma, namely Lemma 60, which says that if (5.1) has positive regularity, then the optimal solution to (5.1) is stable to inaccurate estimates of γ^*. We begin by deriving some properties of $\mathbf{d}(\gamma)$ and its derivatives on $\operatorname{int}(\Gamma)$.

Lemma 58. *Suppose Assumption 13 holds. If $\gamma \in \operatorname{int}(\Gamma)$, then*

$$\mathbf{d}(\gamma) = q(\gamma, x(\gamma)) \quad \textit{and} \quad \tfrac{d}{d\gamma}\mathbf{d}(\gamma) = \nu(\gamma).$$

Proof. For $\gamma \in \operatorname{int}(\Gamma)$, we have $A(\gamma) \succ 0$ and thus $q(\gamma, x)$ is a strongly convex quadratic function in x. One may check that $\nabla_x q(\gamma, x) = 2(A(\gamma)x + b(\gamma))$, and thus $\mathbf{d}(\gamma) = q(\gamma, x(\gamma))$.

Next, from $\mathbf{d}(\gamma) = q(\gamma, x(\gamma))$ and $x(\gamma) = -A(\gamma)^{-1}b(\gamma)$, we deduce

$$\begin{aligned}
\tfrac{d}{d\gamma}\mathbf{d}(\gamma) &= \tfrac{d}{d\gamma}\left(-b(\gamma)^\intercal A(\gamma)^{-1}b(\gamma) + c(\gamma)\right)\\
&= b(\gamma)^\intercal A(\gamma)^{-1}A_1A(\gamma)^{-1}b(\gamma) - 2b_1^\intercal A(\gamma)^{-1}b(\gamma) + c_1\\
&= q_1(x(\gamma)). \qquad \blacksquare
\end{aligned}$$

Lemma 59. *Suppose Assumption 13 holds. Let $\hat\gamma \in \operatorname{int}(\Gamma)$, $P := A(\hat\gamma)^{-1/2}$, and $\Delta := (A_0P^2b_1 - A_1P^2b_0)$. Then, for $\gamma \in \operatorname{int}(\Gamma)$,*

$$\begin{aligned}
\tfrac{d}{d\gamma}\nu(\gamma) &= -2(A_1x(\gamma) + b_1)^\intercal A(\gamma)^{-1}(A_1x(\gamma) + b_1)\\
&= -2\Delta^\intercal\left(A(\gamma)P^2A(\gamma)P^2A(\gamma)\right)^{-1}\Delta.
\end{aligned}$$

Proof. Starting from $\nu(\gamma) = q_1(x(\gamma))$, we compute

$$\begin{aligned}
\tfrac{d}{d\gamma}\nu(\gamma) &= \left\langle \nabla_x q_1(x)\mid_{x=x(\gamma)}, \nabla_\gamma x(\gamma)\right\rangle\\
&= -2\left\langle A_1x(\gamma) + b_1, A(\gamma)^{-1}(A_1x(\gamma) + b_1)\right\rangle\\
&= -2(A_1x(\gamma) + b_1)^\intercal A(\gamma)^{-1}(A_1x(\gamma) + b_1).
\end{aligned}$$

Note also that

$$\begin{aligned}
A_1x(\gamma) + b_1 &= A(\gamma)A(\gamma)^{-1}b_1 - A_1A(\gamma)^{-1}b(\gamma)\\
&= \left(A_0A(\gamma)^{-1}b_1 + \gamma A_1A(\gamma)^{-1}b_1\right) - \left(A_1A(\gamma)^{-1}b_0 + \gamma A_1A(\gamma)^{-1}b_1\right)\\
&= A_0A(\gamma)^{-1}b_1 - A_1A(\gamma)^{-1}b_0.
\end{aligned}$$

Next, suppose $\hat\gamma \in \operatorname{int}(\Gamma)$ and let $P := A(\hat\gamma)^{-1/2}$. Then, PA_0P and PA_1P commute. Indeed, $PA_0P + \hat\gamma PA_1P = PA(\hat\gamma)P = I$. Then,

$$\begin{aligned}
A_0A(\gamma)^{-1}b_1 &= P^{-1}PA_0P(PA(\gamma)P)^{-1}Pb_1\\
&= P^{-1}(PA(\gamma)P)^{-1}PA_0P^2b_1\\
&= (A(\gamma)P^2)^{-1}A_0P^2b_1.
\end{aligned}$$

Similarly, $A_1 A(\gamma)^{-1} b_0 = (A(\gamma)P^2)^{-1} A_1 P^2 b_0$. We deduce

$$\tfrac{d}{d\gamma}\nu(\gamma) = -2\Big(A_0P^2b_1 - A_1P^2b_0\Big)^\intercal \Big(A(\gamma)P^2A(\gamma)P^2A(\gamma)\Big)^{-1}\Big(A_0P^2b_1 - A_1P^2b_0\Big). \blacksquare$$

Corollary 24. *Suppose Assumption 13 holds. Then, $\nu(\gamma)$ is either a strictly decreasing or constant function of γ.*

Proof. Fix $\hat\gamma \in \operatorname{int}(\Gamma)$. By Lemma 59, $\nu(\gamma)$ is strictly decreasing if $A_0A(\hat\gamma)^{-1}b_1 - A_1A(\hat\gamma)^{-1}b_0$ is nonzero. Else, $\nu(\gamma)$ is constant. ■

Corollary 25. *Suppose Assumption 13 holds. Then,* $\arg\max_{\gamma\in\mathbb{R}_+} \mathbf{d}(\gamma)$ *is either a unique point or is all of* Γ. *In the latter case, we furthermore have that* Γ *is compact.*

Proof. Note that by Assumption 13, $\sup_{\gamma\in\mathbb{R}_+} \mathbf{d}(\gamma)$ is achieved. Indeed, as noted in Observation 4, $\mathbf{d}(\gamma) \to -\infty$ as $\gamma \to \infty$. Thus, $\arg\max_{\gamma\in\mathbb{R}_+} \mathbf{d}(\gamma)$ is nonempty.

We will suppose that $\arg\max_{\gamma\in\mathbb{R}_+} \mathbf{d}(\gamma)$ contains at least two points, $\gamma^{(1)} < \gamma^{(2)}$, and show that $\mathbf{d}(\gamma)$ is constant on Γ. Note, by concavity of $\mathbf{d}(\gamma)$ and Lemma 58, we have that $\nu(\gamma) = 0$ for all $\gamma \in (\gamma^{(1)}, \gamma^{(2)})$. By Assumption 13 and Corollary 24, $\nu(\gamma) = 0$ on all of $\operatorname{int}(\Gamma)$ so that $\mathbf{d}(\gamma)$ is constant on $\operatorname{int}(\Gamma)$. By the limit behavior of $\mathbf{d}(\gamma)$ at γ_- and γ_+ (see Observation 4), $\mathbf{d}(\gamma)$ is then constant on all of Γ. This then implies that Γ is compact as again by Observation 4, we have $\mathbf{d}(\gamma) \to -\infty$ as $\gamma \to \infty$. ■

We now define our notion of regularity for the GTRS.

Definition 21. If $\sup_{\gamma\in\mathbb{R}_+} \mathbf{d}(\gamma)$ has a unique maximizer, then set γ^* to be the unique maximizer. Otherwise, $\arg\max_{\gamma\in\mathbb{R}_+} \mathbf{d}(\gamma) = \Gamma$ and let $\gamma^* \in \arg\max_{\gamma\in\Gamma} \mu(\gamma)$. Let $\mu^* := \mu(\gamma^*)$. We will say that the GTRS (5.1) has regularity μ^*. □

Remark 64. One may think of this notion of regularity as requiring *strict complementarity* between the desired rank-one solution of the SDP relaxation of the GTRS and its dual: Writing out the SDP relaxation of the GTRS in full gives

$$\begin{aligned}&\inf_{Y\in\mathbb{S}^{n+1}}\left\{\Big\langle M_{\text{obj}}, Y\Big\rangle : \begin{array}{l}\langle M_1, Y\rangle \le 0\\ Y = \begin{pmatrix} * & * \\ * & 1\end{pmatrix} \succeq 0\end{array}\right\}\\ &\quad\ge \sup_{\gamma\in\mathbb{R}^m,\, t\in\mathbb{R}}\left\{t : \begin{pmatrix} A(\gamma) & b(\gamma)\\ b(\gamma)^\intercal & c(\gamma)-t\end{pmatrix} \succeq 0\right\}.\end{aligned}$$

Strict complementarity asks that the primal SDP have a solution Y^* of rank k and the dual SDP have solution $\gamma^*,\ t^*$ satisfying $\operatorname{rank}\left(\begin{pmatrix} A(\gamma^*) & b(\gamma^*)\\ b(\gamma^*)^\intercal & c(\gamma^*)-t^*\end{pmatrix}\right) = n+1-k$. One may show that this holds with $k=1$ if and only if $A(\gamma^*) \succ 0$ for some maximizer of (γ) (see Lemma 73). □

Corollary 25 ensures that $\arg\max_{\gamma\in\mathbb{R}_+} \mathbf{d}(\gamma)$ and μ^* in Definition 21 are well-defined. Note that, technically, γ^* is *not* well-defined if $\arg\max_{\gamma\in\mathbb{R}_+} \mathbf{d}(\gamma) = \Gamma$ and $\mu(\gamma)$ has more than one maximizer. This is inconsequential and we may work with an arbitrary $\gamma \in \arg\max_{\gamma\in\Gamma} \mu(\gamma)$. For concreteness, one may take γ^* to be the minimum maximizer of $\mu(\gamma)$ in this case.

Remark 65. We make a few observations on our definition of regularity and compare it to the so-called "easy" and "hard" cases of the trust-region subproblem (TRS). Recall that the TRS is the special case of the GTRS (5.1) where $q_1(x) = x^\intercal x - 1$, i.e., the constraint $q_1(x) \leq 0$ corresponds to the unit ball constraint $\|x\|^2 \leq 1$. We will assume that $A_0 \not\succeq 0$. Let $V \subseteq \mathbb{R}^n$ denote the eigenspace corresponding to $\lambda_{\min}(A_0)$. The "easy" and "hard" cases of the TRS correspond to the cases $\Pi_V(b_0) \neq 0$ and $\Pi_V(b_0) = 0$ respectively. Here, Π_V is the projection onto V.

In the "easy" case, it is possible to show that $\lim_{\gamma\searrow-\lambda_{\min}(A_0)} \mathbf{d}(\gamma) = -\infty$ so that $\gamma^* > -\lambda_{\min}(A_0)$ and $\mu^* > 0$. On the other hand, it is possible for $\mu^* > 0$ even in the "hard" case. For example, taking $n = 2$ and

$$A_0 = \begin{pmatrix} 1 & \\ & -1 \end{pmatrix}, \qquad b_0 = \begin{pmatrix} 3 \\ 0 \end{pmatrix}, \qquad c_0 = 0,$$

we have $\Gamma = [1, +\infty)$ and $\mathbf{d}(\gamma) = -9(1+\gamma)^{-1} - \gamma$ on $\operatorname{int}(\Gamma)$. A simple computation then shows $\gamma^* = 2$ and $\mu^* = 1$. We conclude that $\mu^* = 0$ implies the "hard case" but not necessarily vice versa. □

We are now ready to present and prove our key lemma.

Lemma 60. *Suppose Assumption 13 holds, $\mu^* > 0$ and the interval $[\gamma^{(1)}, \gamma^{(2)}] \subseteq \mathbb{R}_+$ contains γ^*. Then, $\nu(\gamma^*) = 0$ and $x(\gamma^*)$ is the unique optimizer of both (5.1) and*

$$\inf_{x\in\mathbb{R}^n} \max\Big(q(\gamma^{(1)}, x), q(\gamma^{(2)}, x)\Big). \tag{5.5}$$

In particular, taking $[\gamma^{(1)}, \gamma^{(2)}] \subseteq \operatorname{int}(\Gamma)$, we have that $x(\gamma^*)$ is the unique optimizer to the strongly convex problem (5.5).

Proof. We show that $x(\gamma^*)$ is the unique minimizer of (5.5). Note that for all $x \in \mathbb{R}^n$, we have

$$\max\Big(q(\gamma^{(1)}, x), q(\gamma^{(2)}, x)\Big) \geq q(\gamma^*, x) \geq \inf_{x\in\mathbb{R}^n} q(\gamma^*, x) = \mathbf{d}(\gamma^*),$$

where the first inequality follows from the facts that $\gamma^* \in [\gamma^{(1)}, \gamma^{(2)}]$ and $q(\gamma, x)$ is an affine function of γ. On the other hand, as $\gamma^* \in \operatorname{int}(\Gamma)$ is a maximizer of the concave function $\mathbf{d}(\gamma)$, which is differentiable at γ^* (see Observation 4, Definition 17, and Lemma 58), we have that $0 = \frac{d}{d\gamma}\mathbf{d}(\gamma)|_{\gamma=\gamma^*} = \nu(\gamma^*) = q_1(x(\gamma^*))$ where the second equation follows from Lemma 58. Then, $q_1(x(\gamma^*)) = 0$ implies that $q(\gamma, x(\gamma^*)) = q_0(x(\gamma^*))$ for any γ. Hence, we deduce that

$$\max\Big(q(\gamma^{(1)}, x(\gamma^*)), q(\gamma^{(2)}, x(\gamma^*))\Big) = q(\gamma^*, x(\gamma^*)) = \mathbf{d}(\gamma^*)$$

so that $x(\gamma^*)$ is a minimizer of (5.5). Uniqueness of $x(\gamma^*)$ then follows from the fact that $q(\gamma^*, x)$ is a strongly convex function of x and it lower bounds the objective function $\max\Big(q(\gamma^{(1)}, x), q(\gamma^{(2)}, x)\Big)$ of (5.5).

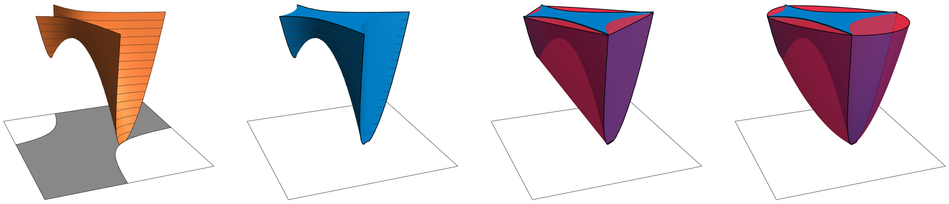

Figure 5.1: A comparison of the convex reformulations of the GTRS given in (5.4) and Lemma 60. The first two figures depict an instance of the GTRS and its epigraph (in blue). The third figure shows the epigraph of the convex reformulation of the GTRS given in (5.4) (in red). The fourth figure shows the epigraph of a strongly convex reformulation of the GTRS given by Lemma 60 (in red).

The proof that $x(\gamma^*)$ is the unique optimizer of (5.1) follows verbatim using the lower bound: $q_0(x) \geq q(\gamma^*, x)$ for all $x \in \mathbb{R}^n$ such that $q_1(x) \leq 0$. ■

5.3 Algorithms for the GTRS

We now turn to the GTRS and present an approach for computing Opt that exploits regularity in (5.1). Our approach will consist of two parts: constructing a convex reformulation of (5.1) and solving the convex reformulation. In conjunction, these two pieces will allow us to achieve *linear* convergence rates for (5.1) whenever $\mu^* > 0$.

Similar to other recent papers on the GTRS [95, 180], we will assume that we are given as input the problem data $(A_0, A_1, b_0, b_1, c_0, c_1)$, regularity parameters $(\xi, \zeta, \hat{\gamma})$, and error and failure parameters (ϵ, p). We will make the following assumption on our input data.

Assumption 14. Suppose that for both $i \in \{0, 1\}$, A_i has at least one negative eigenvalue, $\|A_i\|, \|b_i\|, |c_i| \leq 1$. Let N denote the number of nonzero entries in A_0 and A_1 combined and assume $N \geq n$. Furthermore, suppose $\gamma_+ \leq \zeta$, $A(\hat{\gamma}) \succeq \xi I$, $0 < \xi \leq 1 \leq \zeta$, and $\epsilon, p \in (0, 1)$. □

These assumptions are relatively minor. Indeed, $N \geq n$ without loss of generality. Furthermore, if any of the norms $\|A_i\|, \|b_i\|, |c_i|$ are larger than 1, we may scale the entire function $q_i(x)$ until Assumption 14 holds.

Remark 66. The regularity parameters ξ and ζ will appear in our error and running time bounds. We make no attempt to optimize constants in these bounds and will routinely apply the following bounds (following from Assumption 14) for $\gamma \in \Gamma$: $\|A(\gamma)\|, \|b(\gamma)\|, |c(\gamma)| \leq 1 + \zeta \leq 2\zeta$ □

Our first algorithm, `ConstructReform` (Algorithm 5), will attempt to construct a convex reformulation of (5.1) with strong convexity on the order of $\min(\mu^*, \xi)$. Note, however, that it may be undesirable to compute this reformulation if $\min(\mu^*, \xi) \lesssim \epsilon$. In view of this, we define

$$\phi := \max\Big(\min(\mu^*, \xi),\ \epsilon\xi^4/\zeta^4\Big).$$

To understand this quantity, note that $[\epsilon\xi^4/\zeta^4, \xi]$ is an interval and that ϕ is the closest point to μ^* in this interval. Then, `ConstructReform`, will either output an exact *strongly convex* reformulation of (5.1) with strong convexity on the order ϕ or an ϵ-approximate optimizer. In the former case, we will then apply our second algorithm, `SolveRegular` (Algorithm 8), to compute an ϵ-approximate optimizer.

Remark 67. `ConstructReform` only needs to successfully output an exact strongly convex reformulation *once* on any given instance. Specifically, imagine attempting to solve a single instance of the GTRS twice–once at low accuracy then a second time at a higher accuracy. If `ConstructReform` successfully outputs a strongly convex reformulation on the first run, then on the second run, we may skip `ConstructReform` and simply begin with `SolveRegular` with the new value of $\epsilon > 0$. □

Section E.2 contains useful algorithms and guarantees from the literature that we will use as building blocks in `ConstructReform` and `SolveRegular`. Specifically, Section E.2 recalls the running time of the conjugate gradient algorithm for minimizing a quadratic function (Lemma 97), the running time of the Lanczos method for finding a minimum eigenvalue (Lemma 96), and the running time of Nesterov's accelerated gradient descent method for minimax problems applied to the maximum of two quadratic functions (Lemma 98). We additionally present `ApproxGammaLeft`, a minor modification of [180, Algorithm 2] for finding an aggregation weight $\gamma \leq \hat{\gamma}$ such that $\mu(\gamma)$ falls in a specified range, and `ApproxNu`, a restatement of the conjugate gradient guarantee for the purpose of approximating $\nu(\gamma)$. We state the guarantees of `ApproxGammaLeft` and `ApproxNu` below and leave their proofs to Section E.2.

Lemma 61. *Suppose Assumption 14 holds, $\mu \in (0, \xi)$ and $p \in (0,1)$. Then, with probability at least $1-p$, `ApproxGammaLeft`(μ, p) (Algorithm 17) returns (γ, v) such that $\gamma \leq \hat{\gamma}$ and v is a unit vector satisfying $\mu/2 \leq \mu(\gamma) \leq v^\intercal A(\gamma) v \leq \mu$ in time*

$$\tilde{O}\left(\frac{N\sqrt{\zeta}}{\sqrt{\mu}} \log\left(\frac{n}{p}\right) \log\left(\frac{\zeta}{\mu}\right)\right).$$

Lemma 62. *Suppose Assumption 14 holds, $\mu \in (0, \xi]$, $\delta \in (0,1)$, and $A(\gamma) \succeq \mu I$. Then `ApproxNu`(μ, δ, γ) (Algorithm 18) returns $(\tilde{x}, \tilde{\nu})$ such that $\|\tilde{x} - x(\gamma)\| \leq \mu\delta/10\zeta$, and $\tilde{\nu} = q_1(\tilde{x}) \in [\nu(\gamma) \pm \delta]$ in time*

$$O\left(\frac{N\sqrt{\zeta}}{\sqrt{\mu}} \log\left(\frac{\zeta}{\mu\delta}\right)\right).$$

5.3.1 Constructing a strongly convex reformulation

We present and analyze `ConstructReform` (Algorithm 5). For the sake of presentation, we break `ConstructReform` into the following parts.

We will say that `ConstructReform` (similarly, `CRLeft`, `CRMid`, and `CRRight`) *succeeds* if it either outputs:

- `"regular"`, $\gamma^{(1)}, \gamma^{(2)}, \tilde{\mu}$ such that $\gamma^* \in [\gamma^{(1)}, \gamma^{(2)}]$ and $\mu(\gamma^{(i)}) \geq \tilde{\mu} \geq \min(\mu^*, \xi)/8$,
- `"maybe regular"`, x such that x is an ϵ-approximate optimizer, or
- `"not regular"`, x such that x is an ϵ-approximate optimizer.

The remainder of this subsection proves the following guarantee.

Algorithm 5 `ConstructReform`

```
Given $(A_0, A_1, b_0, b_1, c_0, c_1)$, $(\xi, \zeta, \hat{\gamma})$ and $\epsilon, p \in (0,1)$ satisfying Assumption 14
  1. Set $\gamma_0 = \hat{\gamma}$, $\mu_0 = \xi$
  2. Set $(x_0, \nu_0) =$ ApproxNu$(\mu_0, \epsilon/(4\zeta), \gamma_0)$
  3. If $\nu_0 + \epsilon/(4\zeta) < 0$, run CRLeft (Algorithm 6)
  4. Else if $\nu_0 - \epsilon/(4\zeta) > 0$, run CRRight
  5. Else, run CRMid (Algorithm 7)
```

Proposition 17. *Suppose Assumption 14 holds. With probability at least* $1 - p$, `ConstructReform` *(Algorithm 5) succeeds and runs in time*

$$\tilde{O}\left(\frac{N\sqrt{\zeta}}{\sqrt{\phi}}\log\left(\frac{1}{\phi}\right)\log\left(\frac{n}{p}\right)\log\left(\frac{\zeta}{\epsilon\xi}\right)\right).$$

Proposition 17 will follow as an immediate corollary to the corresponding guarantees for `CRLeft`, `CRRight`, and `CRMid`. The steps and analysis of `CRRight` are analogous to that of `CRLeft` and are omitted.

Our algorithms will attempt to binary search for γ^* using the sign of $\nu(\gamma)$. Unfortunately, as we can only approximate $\nu(\gamma)$ up to some accuracy, we will need to argue how to handle situations where our approximation of $\nu(\gamma)$ is close to zero.

Lemma 63. *Suppose Assumption 14 holds,* $\mu \in (0, \xi]$, $\epsilon \in (0,1)$, *and* $A(\gamma) \succeq \mu I$. *Let* $(\tilde{x}, \tilde{\nu}) =$ `ApproxNu`$(\mu, \epsilon/(4\zeta), \gamma)$. *If* $\tilde{\nu} \in [\pm\epsilon/(4\zeta)]$, *then* $\tilde{x}$ *is an* ϵ*-approximate optimizer of* (5.1).

Proof. By Lemma 62, we have that $q_1(x(\gamma)) = \nu(\gamma) \in [\tilde{\nu} \pm \epsilon/(4\zeta)] \subseteq [\pm\epsilon/(2\zeta)]$ where the last containment follows from $\tilde{\nu} \in [\pm\epsilon/(4\zeta)]$ in the premise of the lemma. Also, note that

$$q_0(x(\gamma)) = q(\gamma, x(\gamma)) - \gamma\nu(\gamma) \leq \mathrm{Opt} + \epsilon/2.$$

Here, the inequality follows from the bounds $\nu(\gamma) \in [\pm\epsilon/(2\zeta)]$, $\gamma \leq \gamma_+ \leq \zeta$ (as $A(\gamma) \succeq 0$ we have $\gamma \in \Gamma$ and Assumption 14 ensures $\gamma_+ \leq \zeta$), and $q(\gamma, x(\gamma)) = \mathbf{d}(\gamma) \leq \mathrm{Opt}$. Thus, we deduce that $x(\gamma)$ is an $\epsilon/2$-approximate optimizer.

Next, by Lemma 62, we have $\|x(\gamma) - \tilde{x}\| \leq \epsilon\mu/(40\zeta^2)$. Note that $\|x(\gamma)\| = \|-A(\gamma)^{-1}b(\gamma)\| \leq \|A(\gamma)^{-1}\|\|b(\gamma)\| \leq 2\zeta/\mu$ where the last inequality follows from $A(\gamma) \succeq \mu I$ and $\|b(\gamma)\| \leq 2\zeta$ (implied by Remark 66). Considering Assumption 14 and applying Lemma 94 with the bounds $\|x(\gamma)\| \leq 2\zeta/\mu$ and $\|x(\gamma) - \tilde{x}\| \leq \epsilon\mu/(40\zeta^2)$, we arrive at

$$\begin{aligned} q_0(\tilde{x}) &\leq q_0(x(\gamma)) + 5\epsilon\frac{\mu}{40\zeta^2}\frac{2\zeta}{\mu} \leq \mathrm{Opt} + \frac{\epsilon}{2} + \frac{\epsilon}{4\zeta} \leq \mathrm{Opt} + \epsilon \\ q_1(\tilde{x}) &\leq q_1(x(\gamma)) + 5\epsilon\frac{\mu}{40\zeta^2}\frac{2\zeta}{\mu} \leq \frac{\epsilon}{2} + \frac{\epsilon}{4\zeta} \leq \epsilon. \end{aligned}$$ ∎

Remark 68. In contrast to the TRS setting, where it is possible to show that $\nu(\gamma)$ "grows quickly" around γ^*, in the GTRS setting, $\nu(\gamma)$ may be "arbitrarily flat". In particular, it may not be possible to determine the sign of $\nu(\gamma)$ given only an inaccurate estimate. Correspondingly, `ConstructReform` may fail to differentiate between `"regular"` and `"not regular"` instances and return `"maybe` `regular"`. In view of Remark 67, we will think of `"maybe regular"` outputs as being less desirable than `"regular"` outputs. We will explore this issue in further detail in Section 5.3.4 and show that `ConstructReform` does not output `"maybe regular"` as long as the GTRS instance satisfies a *coherence* condition. □

Analysis of `CRLeft`

Algorithm 5 calls `CRLeft` if $\nu_0 + \epsilon/4\zeta < 0$. Note that in this case, from Lemma 62 we have $\nu(\hat{\gamma}) = \nu(\gamma_0) \in [\nu_0 \pm \epsilon/(4\zeta)]$ which implies $\nu(\hat{\gamma}) < 0$.

Algorithm 6 `CRLeft`

```
1. Let $T := \left\lceil \log\left(\frac{3200\zeta^4}{\epsilon\xi^3}\right) \right\rceil$. For $t = 1, \ldots, T$,
   a) Set $\mu_t = 2^{-t}\xi$
   b) Set $(\gamma_t, v_t) =$ ApproxGammaLeft$(\mu_t, p/T)$
   c) Set $(x_t, \nu_t) =$ ApproxNu$(\mu_t/2, \epsilon/(4\zeta), \gamma_t)$
   d) If $\nu_t - \epsilon/(4\zeta) > 0$, return "regular", $\gamma_t$, $\hat{\gamma}$, $\mu_t/4$
   e) Else if $\nu_t \in [-\epsilon/(4\zeta),\, \epsilon/(4\zeta)]$
      i. Set $\gamma' := \gamma_t - \mu_t/4$
      ii. Set $(x', \nu') =$ ApproxNu$(\mu_t/4, \epsilon/(4\zeta), \gamma')$
      iii. If $\nu' - \epsilon/(4\zeta) > 0$, return "regular", $\gamma'$, $\hat{\gamma}$, $\mu_t/4$
      iv. Else, return "maybe regular", $x_t$
2. If necessary, negate $v_T$ so that $\langle v_T, A(\gamma_T)x_T + b(\gamma_T)\rangle \leq 0$. Let $\alpha > 0$ such that $q_1(x_T + \alpha v_T) =$
   $0$, return "not regular", $x_T + \alpha v_T$.
```

Proposition 18. *Suppose Assumption 14 holds. With probability at least* $1 - p$, `CRLeft` *(Algorithm 6) succeeds and runs in time*

$$\tilde{O}\left(\frac{N\sqrt{\zeta}}{\sqrt{\phi}} \log\left(\frac{1}{\phi}\right) \log\left(\frac{n}{p}\right) \log\left(\frac{\zeta}{\epsilon\xi}\right)\right).$$

Proof. We condition on step 1.(b) of `CRLeft` succeeding in every iteration. This happens with probability at least $1 - p$.

We begin with the running time. Note that by Lemmas 61 and 62 and $\mu_t = 2^{-t}\xi$ (from step 1.(a)), iteration t of line 1 runs in time

$$\tilde{O}\left(\frac{N\sqrt{\zeta}}{\sqrt{\mu_t}} \log\left(\frac{n}{p}\right) \log\left(\frac{\zeta}{\epsilon\xi}\right)\right).$$

It suffices then to show that $\mu_t = \Omega(\phi)$ in every iteration before `CRLeft` outputs. Noting that $\mu_t \geq \mu_T = \Omega(\epsilon\xi^4/\zeta^4)$, we may instead show that $\mu_t = \Omega(\min(\mu^*, \xi))$ in the iteration at which `CRLeft` outputs.

It remains to show that the output of `CRLeft` satisfies the success criteria and that $\mu_t = \Omega(\min(\mu^*, \xi))$ for the iteration t at which `CRLeft` outputs. We split the remainder of the proof into three parts depending on which line `CRLeft` returns on.

Case 1: `CRLeft` terminates on either line 1.(d) or 1.(e).iii in iteration t Let $\tilde{\gamma} := \gamma_t$ in the first case and $\tilde{\gamma} := \gamma'$ in the second. As `CRLeft` did not terminate at time $t-1$, we have that $\nu(\gamma_{t-1}) < 0$. Indeed, if $\nu(\gamma_{t-1}) \geq 0$, then $\nu_{t-1} \geq -\epsilon/4\zeta$ by Lemma 62. Then, $\nu(\tilde{\gamma}) > 0 > \nu(\gamma_{t-1})$. We deduce by the fact that $\mathbf{d}(\gamma)$ is concave and Lemma 58 that $\gamma^* \in [\tilde{\gamma}, \gamma_{t-1}] \subseteq [\tilde{\gamma}, \hat{\gamma}]$. By construction in line 1.(b), we have that $\mu(\tilde{\gamma}) \geq \mu_t/4$.

It remains to show that $\mu_t \geq \min(\mu^*, \xi)/2$. This holds if $t = 1$, as then $\mu_1 = \xi/2$ by line 1.(a). On the other hand, if $t > 1$, then $\mu(\gamma)$ is an increasing function on the interval $(-\infty, \gamma_{t-1}]$. Indeed, this follows as $\gamma_{t-1} \leq \hat{\gamma}$, $\mu(\gamma_{t-1}) \leq \xi/2 < \mu(\hat{\gamma})$, and $\mu(\gamma) = \lambda_{\min}(A_0 + \gamma A_1)$ is a concave function of γ. Then, from $\gamma^* \in [\tilde{\gamma}, \gamma_{t-1}]$, we deduce that

$$\mu^* = \mu(\gamma^*) \leq \mu(\gamma_{t-1}) \leq \mu_{t-1} = 2\mu_t,$$

where the last inequality follows from line 1.(b).

Case 2: `CRLeft` terminates on line 1.(e).iv in iteration t In this case, we have that $(x_t, \nu_t) = \texttt{ApproxNu}(\mu_t/2, \epsilon/(4\zeta), \gamma_t)$ satisfies $\nu_t \in [\pm\epsilon/(4\zeta)]$. By Lemma 63, we have that x_t is an ϵ-approximate optimizer. It remains to note that the second paragraph of Case 1 holds in this case verbatim so that $\mu_t \geq \min(\mu^*, \xi)/2$.

Case 3: `CRLeft` terminates on line 2 Note that $q_1(x_T) = \nu_T < 0$ holds by line 1.(c), Lemma 62, and the fact that `CRLeft` did not terminate in a prior line. Furthermore,

$$v_T^\intercal A_1 v_T = v_T^\intercal \left(\frac{A(\hat{\gamma}) - A(\gamma_T)}{\hat{\gamma} - \gamma_T} \right) v_T \geq \frac{\xi - \mu_T}{\zeta} \geq \frac{\xi}{2\zeta} > 0,$$

where the first inequality follows from $\zeta \geq \hat{\gamma}$ (by Assumption 14), $v_T^\intercal A(\gamma_T) v_T \leq \mu_T$ (by line 1.(b) and Lemma 61) and $v_T^\intercal A(\hat{\gamma}) v_T = v_T^\intercal A_0 v_T + \hat{\gamma} \geq \xi$ (by Assumption 14 and $\hat{\gamma} \geq 0$), and the second inequality follows from $\mu_T = 2^{-T}\xi$ by line 1(a). This then implies that α in line 2 is well-defined. Thus, by construction in line 2, $q_1(x_T + \alpha v_T) = 0$. Our goal is to show that

$$q_0(x_T + \alpha v_T) = q(\gamma_T, x_T + \alpha v_T) \leq q(\gamma_T, x_T) + \alpha^2 \mu_T \leq \mathrm{Opt} + \epsilon.$$

The following sequence of inequalities allows us to bound $\|x(\gamma_T)\|$:

$$\xi \|x(\gamma_T)\|^2 - 4\zeta \|x(\gamma_T)\| - 2\zeta \leq q(\hat{\gamma}, x(\gamma_T)) \leq q(\gamma_T, x(\gamma_T)) \leq \mathrm{Opt}\,.$$

Here, the first inequality follows from $A(\hat{\gamma}) \succeq \xi I$, $\|b(\hat{\gamma})\| \leq 2\zeta$ and $|c(\hat{\gamma})| \leq 2\zeta$, the second inequality follows as $0 \geq \nu(\gamma_T) = q_1(x(\gamma_T))$ (by line 1.(c), Lemma 62 and the fact that `CRLeft` terminates on line 2) and $\hat{\gamma} \geq \gamma_T$, the third inequality follows as $q(\gamma_T, x(\gamma_T)) = \mathbf{d}(\gamma_T) \leq \mathrm{Opt}$ (by Lemma 58). Then, taking $x = 0$ in the expression $\mathrm{Opt} = \inf_x \sup_{\gamma \in \Gamma} q(\gamma, x)$ gives $\mathrm{Opt} \leq 2\zeta$. Applying Lemma 95 to $\xi\|x(\gamma_T)\|^2 - 4\zeta\|x(\gamma_T)\| - 4\zeta \leq 0$ gives $\|x(\gamma_T)\| \leq (2\sqrt{2} + 2)\zeta/\xi \leq 5\zeta/\xi$, and by Assumption 14 and line 1.(c) we have $\|A_1 x_T + b_1\| \leq \|A_1\|(\|x(\gamma_T)\| + \|x_T - x(\gamma_T)\|) + \|b_1\| \leq (5\zeta/\xi + 1) + 1 \leq 7\zeta/\xi$.

Next, we may bound

$$\begin{aligned} q(\gamma_T, x_T) &\leq q(\gamma_T, x(\gamma_T)) + \|A(\gamma_T)\| \|x(\gamma_T) - x_T\|^2 \\ &\leq \text{Opt} + (2\zeta)\left(\frac{\mu_T \epsilon}{80\zeta^2}\right)^2 \leq \text{Opt} + \epsilon/2. \end{aligned}$$

Similarly, $\nu(\gamma_T) \geq \nu(\hat{\gamma}) = q_1(x(\hat{\gamma})) \geq -\|x(\hat{\gamma})\|^2 - 2\|x(\hat{\gamma})\| - 1 \geq -(3\zeta/\xi)^2$, where the first inequality follows from Corollary 24 and the last from the bound $\|x(\hat{\gamma})\| \leq 2\zeta/\xi$. We deduce that $0 \geq q_1(x_T) \geq \nu(\gamma_T) - \epsilon/(4\zeta) \geq -10\zeta^2/\xi^2$. By line 2 and applying Lemma 95, we have that $\alpha \leq 40\zeta^2/\xi^2$.

We conclude that $\alpha^2 \mu_T \leq \alpha^2 \frac{\epsilon \xi^4}{3200\zeta^4} \leq \frac{\epsilon}{2}$ so that $q_0(x_T + \alpha v_T) = q(\gamma_T, x_T + \alpha v_T) \leq q(\gamma_T, x_T) + \alpha^2 \mu_T \leq \text{Opt} + \epsilon$, where the equation follows from the definition of α in line 2.

It remains to note that as $\nu(\gamma_T) < 0$, Corollary 24 implies $\gamma^* \leq \gamma_T$ and $\mu^* = \mu(\gamma^*) \leq \mu(\gamma_T) \leq \mu_T$. ■

Analysis of `CRMid`

Algorithm 5 calls `CRMid` if $\nu_0 \in [-\epsilon/(4\zeta),\ \epsilon/(4\zeta)]$. Note that in this case, we may deduce $|\nu(\hat{\gamma})| = |\nu(\gamma_0)| \leq \epsilon/(2\zeta)$.

Algorithm 7 `CRMid`

```
1. Let $\gamma' := \gamma_0 - \xi/2$ and $\gamma'' := \gamma_0 + \xi/2$
2. Set $(x', \nu') =$ ApproxNu$(\gamma', \epsilon/(4\zeta))$
3. Set $(x'', \nu'') =$ ApproxNu$(\gamma'', \epsilon/(4\zeta))$
4. If $\nu' - \epsilon/(4\zeta) > 0 > \nu'' + \epsilon/(4\zeta)$, return "regular", $\gamma', \gamma'', \xi/2$
5. Else if $\nu' - \epsilon/(4\zeta) \leq 0$, return "maybe regular", $x_0$
6. Else, return "maybe regular", $x_0$
```

Proposition 19. *Suppose Assumption 14 holds. Then,* `CRMid` *(Algorithm 7) succeeds and runs in time*

$$O\left(\frac{N\sqrt{\zeta}}{\sqrt{\xi}} \log\left(\frac{\zeta}{\epsilon\xi}\right)\right).$$

Proof. Suppose `CRMid` returns on line 4. Then, by Lemma 62 and lines 2 and 3 we have $\nu(\gamma') > 0 > \nu(\gamma'')$. We deduce by the fact that $\mathbf{d}(\gamma)$ is concave and Lemma 58, that $\gamma^* \in [\gamma', \gamma'']$. Furthermore, $\mu(\hat{\gamma} \pm \xi/2) \geq \mu(\hat{\gamma}) - \xi/2 \geq \xi/2$ as μ is 1-Lipschitz and $\mu(\hat{\gamma}) \geq \xi$.

If, `CRMid` returns on lines 5 or 6, then $(x_0, \nu_0) =$ `ApproxNu`$(\mu_0, \epsilon/(4\zeta), \gamma_0)$ satisfies $\nu_0 \in [\pm\epsilon/(4\zeta)]$. By Lemma 63, we have that x_0 is an ϵ-approximate optimizer.

The running time of `CRMid` follows from Lemma 97. ■

Algorithm 8 `SolveRegular`

```
Given $\gamma^{(1)}, \gamma^{(2)}, \tilde{\mu}$ such that $\gamma^* \in [\gamma^{(1)}, \gamma^{(2)}]$ and $\min_{i\in[2]}\left\{\mu(\gamma^{(i)})\right\} \geq \tilde{\mu} > 0$
  1. Apply Nesterov's accelerated minimax scheme for strongly convex smooth quadratic functions to
     compute a $\tilde{\mu}(\epsilon\tilde{\mu}/10\zeta)^2$-optimal solution $\bar{x}$ to
         $\min_{x\in\mathbb{R}^n} \max\left(q(\gamma^{(1)}, x), q(\gamma^{(2)}, x)\right)$
  2. Return $\bar{x}$
```

5.3.2 Solving the convex reformulation

Proposition 20. *Suppose Assumption 14 holds and $\tilde{\mu} \in (0, \xi]$. Then,* `SolveRegular` *(Algorithm 8) computes an ϵ-approximate solution to* (5.1) *in time*

$$O\left(\frac{N\sqrt{\zeta}}{\sqrt{\tilde{\mu}}}\log\left(\frac{\zeta}{\epsilon\tilde{\mu}}\right)\right).$$

Proof. For notational simplicity, let $q_{\max}(x) := \max\left(q(\gamma^{(1)}, x),\ q(\gamma^{(2)}, x)\right)$. Let $x^* := x(\gamma^*)$. Recall that $q_0(x^*) = \mathrm{Opt}$, $q_1(x^*) = 0$, and $q_{\max}(x^*) = \mathrm{Opt}$. Then, by definition of μ^* in Definition 21 and strong convexity of $q(\gamma^*, x)$, we have

$$\begin{aligned}\tilde{\mu}\left\|x^* - \bar{x}\right\|^2 \leq \mu^*\left\|x^* - \bar{x}\right\|^2 &\leq q(\gamma^*, \bar{x}) - q(\gamma^*, x^*) = q(\gamma^*, \bar{x}) - \mathrm{Opt}\\ &\leq q_{\max}(\bar{x}) - \mathrm{Opt} \leq \tilde{\mu}\left(\frac{\epsilon\tilde{\mu}}{10\zeta}\right)^2.\end{aligned}$$

Rearranging, we may bound $\left\|x^* - \bar{x}\right\| \leq \frac{\epsilon\tilde{\mu}}{10\zeta}$. Furthermore, $\left\|x^*\right\| = \left\|x(\gamma^*)\right\| = \left\|-A(\gamma^*)^{-1}b(\gamma^*)\right\|$ so that $\left\|x^*\right\| \leq 2\zeta/\tilde{\mu}$ holds by Assumption 14.

Then, as $\epsilon\tilde{\mu}/(10\zeta) \leq 1$ and $2\zeta/\tilde{\mu} \geq 1$ (by definition of $\tilde{\mu}$ and Assumption 14), we can apply Lemma 94 to get

$$\begin{aligned}q_0(\bar{x}) &\leq q_0(x^*) + 5\epsilon\frac{\tilde{\mu}}{10\zeta}\frac{2\zeta}{\tilde{\mu}} = \mathrm{Opt} + \epsilon\\ q_1(\bar{x}) &\leq q_1(x^*) + 5\epsilon\frac{\tilde{\mu}}{10\zeta}\frac{2\zeta}{\tilde{\mu}} = \epsilon.\end{aligned}$$

The running time follows from Lemma 98. ■

5.3.3 Putting the pieces together

The following theorem states the guarantee for applying `ConstructReform` (Algorithm 5) and `SolveRegular` (Algorithm 8). This guarantee follows as a corollary to Propositions 18 to 20

Theorem 26. *Suppose Assumption 14 holds. Then with probability* $1-p$, *the procedure outlined above returns an* ϵ*-approximate solution to* (5.1) *in time*

$$\tilde{O}\left(\frac{N}{\sqrt{\phi}}\log\left(\frac{1}{\phi}\right)\log\left(\frac{n}{p}\right)\log\left(\frac{\zeta}{\epsilon\xi}\right)\right).$$

5.3.4 Revisiting `"maybe regular"` outputs

We revisit `ConstructReform` (Algorithm 5) and show that `ConstructReform` does not output `"maybe regular"` on a successful run as long as a coherence condition is satisfied.

The following example shows that in the GTRS setting, $\nu(\gamma)$ may grow arbitrarily slowly near γ^*.

Example 17. Let $n=2$ and $\epsilon\in(0,1/4)$ and set

$$A_0=\begin{pmatrix}1 & \\ & -1/2\end{pmatrix},\quad A_1=\begin{pmatrix}-1 & \\ & 1\end{pmatrix},\quad b_0=\epsilon\cdot e_1,\quad b_1=0,\quad c_0=0,\quad c_1=16\epsilon^2.$$

Note that $\Gamma=[1/2,1]$ and $A(3/4)=I/4$ so that Assumption 14 holds with $\xi=1/4$ and $\zeta=1$. Then, we have

$$x(\gamma)=-\frac{\epsilon}{1-\gamma}e_1,\quad \nu(\gamma)=\epsilon^2\left(16-\frac{1}{(1-\gamma)^2}\right),\ \forall\gamma\in(1/2,1).$$

Taking $\epsilon\to 0$, we have that $\frac{d}{d\gamma}\nu(\gamma)$ may be arbitrarily close to zero around $\gamma^*=3/4$. We deduce that Assumption 14 alone is not enough to upper bound $\frac{d}{d\gamma}\nu(\gamma)$ over $\operatorname{int}(\Gamma)$. □

Lemma 64. *Suppose Assumption 14 holds and that*

$$\delta:=\left\|A_0A(\hat{\gamma})^{-1}b_1-A_1A(\hat{\gamma})^{-1}b_0\right\|>0.$$

Then, $\frac{d}{d\gamma}\nu(\gamma)\leq-\delta^2\xi^2/(4\zeta^3)$ *for any* $\gamma\in\operatorname{int}(\Gamma)$. *In particular,* $|\nu(\gamma)|\leq\epsilon/(2\zeta)$ *for an interval of length at most* $4\epsilon\zeta^2/(\delta^2\xi^2)$.

Proof. For convenience, let $P:=A(\hat{\gamma})^{-1/2}$ and $\Delta:=A_0P^2b_1-A_1P^2b_0$ so that $\delta=\|\Delta\|$. By Lemma 59,

$$\frac{d}{d\gamma}\nu(\gamma)=-2\Delta^\intercal(A(\gamma)P^2A(\gamma)P^2A(\gamma))^{-1}\Delta.$$

Assumption 14 implies $A(\hat{\gamma})\succeq\xi I$, and so $P^2\preceq(1/\xi)I$. Moreover, by Remark 66 we have $A(\gamma)\leq 2\zeta I\ \forall\gamma\in\operatorname{int}(\Gamma)$ and hence $A(\gamma)P^2A(\gamma)P^2A(\gamma)\preceq 8\zeta^3\xi^{-2}I$. We conclude,

$$\frac{d}{d\gamma}\nu(\gamma)\leq-\frac{\delta^2\xi^2}{4\zeta^3}.$$

The final assertion follows as $\frac{\epsilon}{\zeta}\cdot\frac{4\zeta^3}{\delta^2\xi^2}=\frac{4\epsilon\zeta^2}{\delta^2\xi^2}$. ■

Remark 69. As in the proof of Proposition 18, we will assume that Line 1.(b) of CRLeft (Algorithm 6) succeeds in every iteration. Suppose that CRLeft outputs "maybe regular" on iteration t. Recall that in this case we have $\nu(\gamma_t), \nu(\gamma') \in [\pm\epsilon/2\zeta]$ and $\mu_t \geq \mu^*/2$. By construction, $\gamma' = \gamma_t - \mu_t/4$. By Lemma 64 we deduce that the coherence parameter δ is bounded by

$$\delta \leq \frac{2\sqrt{2}\zeta}{\xi}\sqrt{\frac{\epsilon}{\mu^*}}.$$

Momentarily treating ξ, ζ as constant, we deduce that CRLeft can only output "maybe regular" if the coherence parameter is sufficiently small, i.e., $\delta = O(\sqrt{\epsilon/\mu^*})$ (assuming that line 1.(b) succeeds in every iteration). □

5.4 Numerical Experiments

In this section, we study the numerical performance of our approach (Section 5.3) for solving the GTRS. We compare our proposed approach with other algorithms [2, 21, 94, 180] suggested in the literature. In the following, we will refer to our algorithm as WLK21 and the algorithms in [2, 21, 94, 180] as AN19, BTH14, JL19, and WK20 respectively. Recall that WK20 [180] builds a convex reformulation of the GTRS (see Remark 63) and applies Nesterov's accelerated gradient descent method. JL19 [94] builds the same convex reformulation and applies a saddle-point-based first-order algorithm to solve it. AN19 [2] computes the minimum generalized eigenvalue (and an associated eigenvector) of an indefinite $(2n+1) \times (2n+1)$ matrix pencil and recovers γ^* and x^* from these quantities. BTH14 [21] notes that the SDP relaxation of (5.1) (which is known to be exact) can be reformulated as a second-order cone program (SOCP) after computing an appropriate diagonalizing basis. The corresponding SOCP reformulation can then be solved via interior-point method solvers such as MOSEK.

In our experiments, we have implemented slight modifications to WK20, WLK21, JL19, and AN19. First, we have replaced the eigenvalue calls within WK20 and WLK21 with generalized eigenvalue calls. Indeed, in both algorithms a series of eigenvalue calls are used to simulate a single generalized eigenvalue call. While the theoretical analysis using eigenvalue calls is simpler, the practical running time using generalized eigenvalue calls is faster due to the availability of efficient generalized eigenvalue solvers. Second, in view of practical applications where ϵ-feasibility may be unacceptable or undesirable, we also implement a "rounding" step at the ends of WLK21, WK20, and JL19 to ensure feasibility, i.e., $q_1(\tilde{x}) \leq 0$. As suggested in [2], AN19 implements a Newton refinement process to ensure $q_1(\tilde{x}) \leq 0$. The feasibility in BTH14 depends on MOSEK and is often slightly violated. Further implementation details are described in Section 5.4.1.

All experiments were performed in MATLAB R2021a and MOSEK 9.3.6 on a machine with an AMD Opteron 4184 processor and 70GB of RAM. Our MATLAB code is available at:

https://github.com/alexlihengwang/linear-time-gtrs

5.4.1 Implementation

We discuss some implementation details.

Eigenvalue solvers We replace ApproxGammaLeft (Algorithm 17) of CRLeft (Algorithm 6) using a generalized eigenvalue solver as follows. Recall that ApproxGammaLeft finds $\gamma_t \leq \hat{\gamma}$ and unit vector $v_t \in \mathbb{R}^n$ such that $\mu_t/2 \leq \mu(\gamma_t) \leq v_t^\intercal A(\gamma_t) v_t \leq \mu_t$. We can achieve the same guarantee using a generalized eigenvalue solver: Approximate the minimum generalized eigenvalue λ_t of $-A_1 v_t = \lambda_t (A(\hat{\gamma}) - \frac{3\mu_t}{4} I) v_t$ to some tolerance ϵ and set $\gamma_t = \hat{\gamma} + \frac{1}{\lambda_t}$. Then, as long as $\epsilon > 0$ is small enough, we can show that γ_t, v_t satisfy the same guarantees as ApproxGammaLeft. Detailed proofs can be found in Section E.4. In our implementations, we use the generalized eigenvalue solver eigifp [75] for WLK21, WK20 and JL19. In contrast, as AN19 requires the minimum eigenvalue to an *indefinite* matrix pencil, we use the generalized eigenvalue solver eigs for AN19.

Rounding At the end of WLK21, WK20 and JL19, we implement the following rounding procedure. Given the output $\bar{x}$ of one of these algorithms, we will construct $\tilde{x} := \bar{x} + \delta$ where $\delta = \alpha v$. The direction v is picked so that $x^\intercal A_1 x$ is either positive or negative depending on the sign of $q_1(\bar{x})$. Then, we pick α by solving the quadratic equation $q_1(\bar{x} + \alpha v) = 0$. For WK20 and JL19, we may set v to be an approximate eigenvector of γ_- or γ_+ as we have already computed these quantities while constructing the convex reformulation. For WLK21, we compute an (inaccurate) eigenvalue corresponding to either $\lambda_{\min}(A_1)$ or $\lambda_{\max}(A_1)$.

5.4.2 Random instances

We evaluate the numerical performance of the different algorithms on random instances with dimension n, number of nonzero entries $N \approx \bar{N}$, regularity $\mu^* \approx \bar{\mu}^*$, and $\xi = 0.1$. Our random generation process is similar to that of [2] and allows us to generate instances with known optimizers.

First, sample a sparse symmetric matrix $\hat{A}$ using the MATLAB command sprandsym(n,N/(n*n)). This matrix is then scaled so that $0 \prec \xi I \preceq \hat{A} \preceq (1+\xi)I$. We will define A_0, A_1, and $\hat{\gamma}$ in such a way that $A(\hat{\gamma}) = \hat{A}$. Generate A_0 using the same function call and scale it so that $\|A_0\| = 1$. We then set $\hat{\gamma} := \lambda_{\max}(\hat{A} - A_0)$ and $A_1 := (\hat{A} - A_0)/\hat{\gamma}$. Note that $\|A_1\| = 1$. We sample b_0 and b_1 uniformly from the unit sphere.

We have the option to choose γ^* to lie to either the left or right of $\hat{\gamma}$. In the former case, we set $\gamma^* := \hat{\gamma} + 1/\lambda_{\min}(-A_1, A(\hat{\gamma}) - \bar{\mu} I)$. In the latter, we set $\gamma^* := \hat{\gamma} - 1/\lambda_{\min}(A_1, A(\hat{\gamma}) - \bar{\mu} I)$. Here, the notation $\lambda_{\min}(X, Y)$ denotes the minimum generalized eigenvalue of X with respect to Y. To ensure that γ^* is indeed the dual optimizer, we set $c_0 = 0$ and c_1 such that $\nu(\gamma^*) = 0$. The exact optimizer is then given by $x^* := -A(\gamma^*)^{-1} b(\gamma^*)$. Finally, we normalize b_0, b_1, c_1 and x^* to ensure Assumption 14.

To summarize, the output of this method is a random GTRS instance satisfying Assumption 14 with $N \approx \bar{N}$, $\mu^* \approx \bar{\mu}^*$ and known Opt and x^* (up to machine precision).

5.4.3 Experimental setup

The numerical experiments were performed with $n \in \{10^3, 10^4, 10^5\}$, $\bar{N} \in \{10n, 100n\}$ and $\bar{\mu}^* \in \{10^{-2}, 10^{-4}, 10^{-6}\}$. We generated 100 random instances for $n = 10^3$ and 10^4 and five random instances for $n = 10^5$ due to large running times. BTH14 was only reported for $n = 10^3$ as for $n \geq 10^4$ it was unable to return a solution within five times the average running time of

WLK21 or WK20. The dominant cost in BTH14 for (5.1) is in computing the diagonalizing basis, which requires computing a full set of generalized eigenvalues and is unlikely to scale favorably with n and N. AN19 was not reported for $n = 10^5$ because of numerical issues and large running times associated with eigs applied to the *indefinite* generalized eigenvalue problem.

For each algorithm and each random instance, we record the error,

$$\texttt{Error} = q_0(\tilde{x}) - \mathrm{Opt},$$

of the output. For the three "convex-reformulation and gradient-descent" algorithms WLK21, WK20, and JL19, we additionally record the error *within* the corresponding convex reformulations, i.e.,

$$\begin{aligned}
&\texttt{ErrorCR} = \max\Big(q(\gamma^{(1)}, \bar{x}), q(\gamma^{(2)}, \bar{x})\Big) - \mathrm{Opt}, \quad \text{for WLK21}, \quad \text{and}\\
&\texttt{ErrorCR} = \max(q(\gamma_-, \bar{x}), q(\gamma_+, \bar{x})) - \mathrm{Opt}, \quad \text{for WK20 and JL19}.
\end{aligned}$$

See (5.2) and Proposition 17 for definitions of $\gamma_-, \gamma_+, \gamma^{(1)}$ and $\gamma^{(2)}$. Here, $\bar{x}$ is an iterate within the gradient descent method for the corresponding convex reformulation and $\tilde{x}$ is a "rounded" solution satisfying $q_1(\tilde{x}) \leq 0$.

5.4.4 Results

Our numerical results are illustrated in Figures 5.2 to 5.4 which display ErrorCR for WLK21, WK20, and JL19 and Error for AN19 and BTH14 over time (in seconds) for each $n \in \{10^3, 10^4, 10^5\}$, respectively. Tables containing detailed statistics are given in Section E.5.

Remark 70. We decide to plot ErrorCR for WLK21, WK20, and JL19 as that is the error that the respective algorithms are designed to drive to zero. We observe empirically (see Section E.5) that ErrorCR and Error track quite closely for WLK21. □

We make a number of observations:

- The lines plotted in Figures 5.2 to 5.4 begin after time zero. For WLK21, WK20, and JL19 this gap corresponds to the time required to construct the corresponding convex reformulations of (5.1). For AN19, this corresponds to the time required to compute $x(\hat{\gamma})$ exactly, which is required to set up the appropriate $(2n+1) \times (2n+1)$ generalized eigenvalue problem [2]. For BTH14, this gap corresponds to the time required to compute a diagonalizing basis of (5.1).

- WLK21 constructs its reformulation faster than WK20 and JL19 when $\bar{\mu}^* = 10^{-2}$. The situation is reversed for $\bar{\mu}^* \in \{10^{-4}, 10^{-6}\}$. Nevertheless, WLK21 outperforms both WK20 and JL19 due to its significantly improved performance in solving the resulting convex reformulation. See Section E.5.

- As expected from Theorem 26, WLK21 exhibits a *linear* convergence rate in terms of ϵ. This is most apparent in the plots corresponding to $\bar{\mu}^* = 10^{-2}$ and $\bar{\mu}^* = 10^{-4}$.

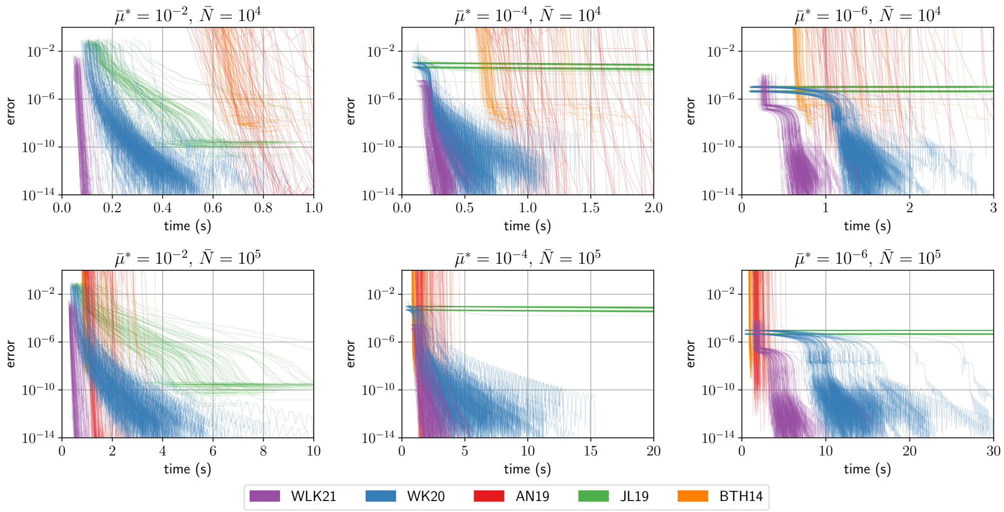

Figure 5.2: Comparison of algorithms for $n = 10^3$.

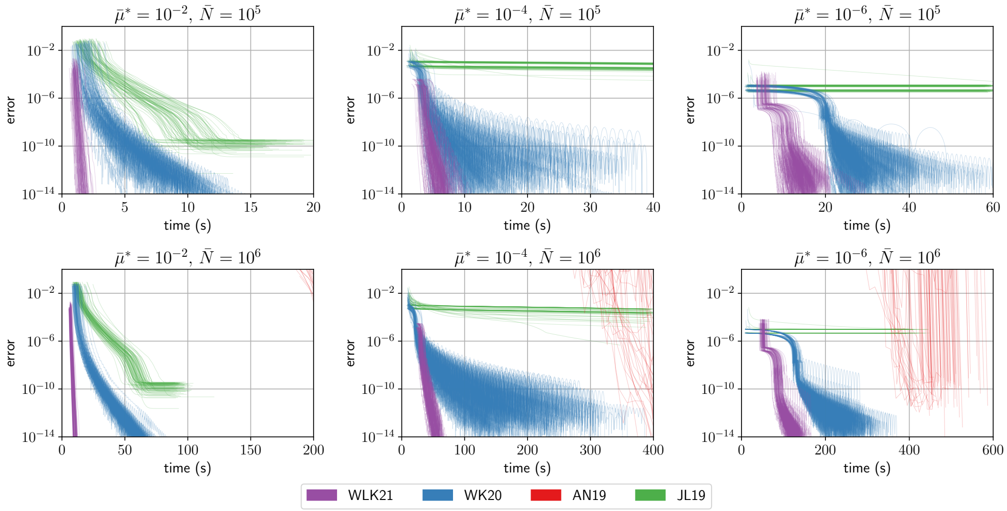

Figure 5.3: Comparison of algorithms for $n = 10^4$.

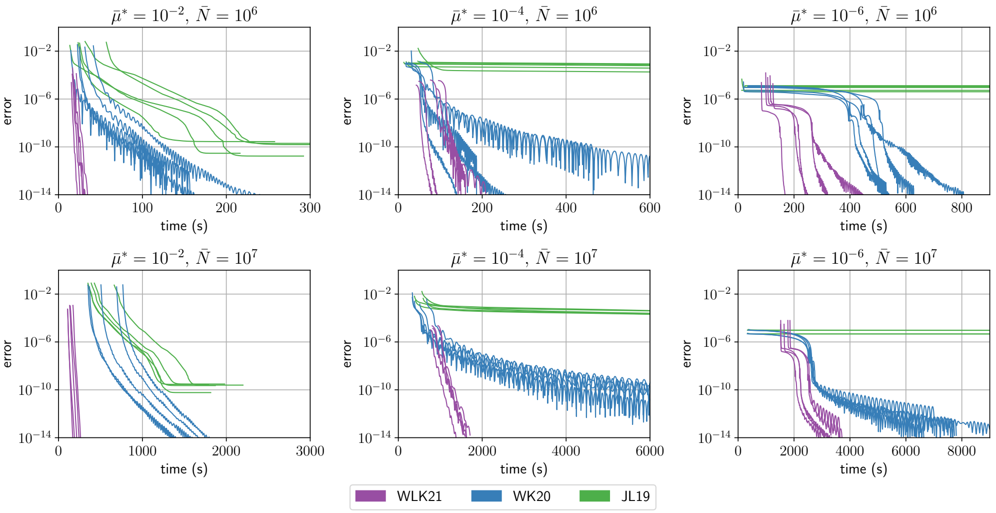

Figure 5.4: Comparison of algorithms for $n = 10^5$.

- Although the convergence guarantees established for WK20 [180] and JL19 [94] do not depend on μ^*, our results show empirically that these algorithms in fact perform better when μ^* is large. The degree to which the running times of these algorithms vary with μ^* is less than that of WLK21.

- The convergence rates of AN19 and BTH14 do not vary significantly with either N or μ^*, but they exhibit heavy dependence on n. Specifically, the convergence rate of AN19 empirically varies in n as $\approx n^2$. This is consistent with the results reported in [2]. Similarly, due to the complete eigenbasis computation embedded in BTH14, we expect BTH14 to vary in n as $\approx n^3$. Thus, as can be seen in Figures 5.2 to 5.4, although AN19 outperforms WLK21 and WK20 for $(n, \bar{N}, \bar{\mu}^*) = (10^3, 10^5, 10^{-6})$, AN19 and BTH14 become impractical for $n = 10^4$ and $n = 10^5$.

- The saddle-point based first-order algorithm employed in JL19 is unable to decrease the error below $\approx 10^{-4}$ for $\bar{\mu}^* = 10^{-4}$ and $\bar{\mu}^* = 10^{-6}$.

6 Accelerated first-order methods for a class of semidefinite programs

This chapter is based on joint work [182] with Fatma Kılınç-Karzan.

This chapter introduces a new storage-optimal first-order method (FOM), CertSDP, for solving a special class of semidefinite programs (SDPs) to high accuracy. The class of SDPs that we consider, the *exact QMP-like SDPs*, is characterized by low-rank solutions, *a priori* knowledge of the restriction of the SDP solution to a small subspace, and standard regularity assumptions such as strict complementarity. This class is inspired by structural assumptions that hold for exact SDP relaxations of quadratically constrained quadratic programs (QCQPs) and quadratic matrix programs (QMPs). Crucially, we show how to use a *certificate of strict complementarity* to construct a low-dimensional strongly convex minimax problem whose optimizer coincides with a factorization of the SDP optimizer. From an algorithmic standpoint, we show how to construct the necessary certificate and how to solve the minimax problem efficiently. We accompany our theoretical results with preliminary numerical experiments suggesting that CertSDP significantly outperforms current state-of-the-art methods on large sparse exact QMP-like SDPs.

6.1 Introduction

Semidefinite programs (SDPs) are among the most powerful tools that optimizers have for tackling both convex *and* nonconvex problems. In the former direction, SDPs are routinely used to model convex optimization problems that arise in a variety of applications such as robust optimization, engineering, and robotics [22, 174]. In the latter direction, many results over the last thirty years have shown that SDPs perform provably well as convex relaxations of certain nonconvex optimization problems; see [22, 40, 74, 149] and references therein. As examples, exciting results in phase retrieval [40] and clustering [1, 122, 153] show that these nonconvex problems have exact SDP relaxations with high probability under certain random models. More abstractly, a line of recent work [7, 17, 20, 35, 37, 38, 92, 100, 113, 167, 179, 181] has investigated general conditions under which *exactness* holds between nonconvex quadratically constrained quadratic programs (QCQPs) or quadratic matrix programs (QMPs) and their standard SDP relaxations.

Despite the expressiveness and strong theoretical guarantees of SDPs, they have seen limited application in practice and have a reputation of being "prohibitively expensive," especially for large-scale applications. Indeed, standard methods for solving SDPs, such as the interior point methods (IPMs) [4, 133], scale poorly with problem dimension due to both their expensive iterations and also significant memory needs. See [198, Section 8.1] for a more thorough discussion.

In this chapter, we show how to derive highly efficient (in iteration complexity, per-iteration-cost, and memory usage) first-order methods (FOMs) for solving general SDPs that admit a desirable *exactness* property. Our developments are inspired by recent results on linearly convergent FOMs for the trust-region subproblem (TRS) and the generalized trust-region subproblem (GTRS) [41, 183] that operate in the original problem space. We briefly discuss these problems now to motivate our assumptions and our problem class. We will discuss this literature in further detail in Section 6.1.3.

The TRS [125] seeks to minimize a general quadratic objective over the unit ball. The GTRS [124] then replaces the unit ball constraint with a general quadratic equality or inequality constraint:

$$\inf_{x\in\mathbb{R}^{n-1}}\left\{q_{\text{obj}}(x):\ q_1(x)=0\right\}$$

(presented as an equality constraint). Here, both q_{obj} and q_1 may be nonconvex, but it is standard to assume that there exists $\hat{\gamma}\in\mathbb{R}$ such that $q_{\text{obj}}+\hat{\gamma}q_1$ is a *strongly convex* quadratic function. Under this assumption, the S-lemma [67] guarantees that the GTRS has an exact SDP relaxation in the following sense: Let $M_{\text{obj}},\ M_1$ be symmetric matrices such that $q_{\text{obj}}(x)=\left(\begin{smallmatrix}x\\1\end{smallmatrix}\right)^\intercal M_{\text{obj}}\left(\begin{smallmatrix}x\\1\end{smallmatrix}\right)$ and $q_1(x)=\left(\begin{smallmatrix}x\\1\end{smallmatrix}\right)^\intercal M_1\left(\begin{smallmatrix}x\\1\end{smallmatrix}\right)$. Then, equality holds between the GTRS, its SDP relaxation, and the dual of the SDP relaxation:

$$\begin{aligned}
&\min_{x\in\mathbb{R}^{n-1}}\left\{q_{\text{obj}}(x):\ q_1(x)=0\right\}\\
&=\min_{Y\in\mathbb{S}^n}\left\{\left\langle M_{\text{obj}},Y\right\rangle:\ \begin{array}{c}\langle M_1,Y\rangle=0\\ Y=\begin{pmatrix}* & *\\ * & 1\end{pmatrix}\succeq 0\end{array}\right\}\\
&=\sup_{\gamma\in\mathbb{R},\,t\in\mathbb{R}}\left\{t:\ M_{\text{obj}}+\gamma M_1-t\begin{pmatrix}0_{n-1} & \\ & 1\end{pmatrix}\succeq 0\right\}.
\end{aligned}$$

Here, $\mathbb{S}^n$ is the vector space of $n\times n$ symmetric matrices, the inner product $\langle M,Y\rangle$ is defined as $\langle M,Y\rangle:=\operatorname{tr}(M^\intercal Y)$ and $Y\succeq 0$ indicates that Y is positive semidefinite (PSD).

In particular, the SDP relaxation of the GTRS has an optimal solution Y^* with rank one. Furthermore, we know the value of $(Y^*)_{n,n}=1$ before we even solve the SDP relaxation. We will think of this as *a priori* knowledge of the restriction of Y^* to a subspace of dimension $\operatorname{rank}(Y^*)$.

Despite the fact that the SDP relaxation solves the GTRS exactly, the large computational cost of solving SDPs has spurred an extensive line of work developing new algorithms for the GTRS (that avoid explicitly solving large SDPs). Most relatedly, Wang et al. [183] (Chapter 5) assume that the dual SDP is solvable and that there exists an optimal dual solution (γ^*,t^*) such that $M_{\text{obj}}+\gamma^*M_1-t^*\left(\begin{smallmatrix}0_{n-1} & \\ & 1\end{smallmatrix}\right)$ has rank $n-1$. This assumption holds generically for GTRS problems conditioned on strict feasibility of the dual SDP and can be phrased as assuming strict complementarity [5] between the dual SDP and the desired rank-one solution Y^*. Wang et al. [183] then showed that it is possible to construct a strongly convex reformulation of the GTRS in the original space using a sufficiently accurate estimate of γ^*.

In our study, we will examine general SDPs satisfying similar structural assumptions and design an efficient *storage-optimal* FOM to solve them. In this respect, our approach also extends a recent line of work [60, 70, 160, 198] towards developing storage-optimal FOMs for SDPs possessing low-rank solutions. We discuss storage optimality in SDP algorithms in Section 6.1.3.

6.1.1 Problem setup and assumptions

Consider an SDP in standard form and its dual:

$$\begin{aligned}\inf_{Y\in\mathbb{S}^n}&\left\{\left\langle M_{\text{obj}}, Y\right\rangle : \begin{array}{l}\langle M_i, Y\rangle + d_i = 0,\ \forall i\in[m]\\ Y\succeq 0\end{array}\right\} &&\text{(SDP)}\\ &\geq \sup_{\gamma\in\mathbb{R}^m}\left\{d^\intercal\gamma :\ M_{\text{obj}} + \textstyle\sum_{i=1}^m \gamma_i M_i \succeq 0\right\}.\end{aligned}$$

For notational convenience, we let $d_{\text{obj}} := 0$ and define $M(\gamma) := M_{\text{obj}} + \sum_{i=1}^m \gamma_i M_i$ and $d(\gamma) := d_{\text{obj}} + \sum_{i=1}^m \gamma_i d_i$.

In this chapter, inspired by the structural properties of the GTRS that make it amenable to highly efficient FOMs, we will work under two major assumptions. First, we will assume (Assumption 15) that the primal and dual SDPs are both solvable, strong duality holds, and there exist primal and dual optimal solutions $Y^*\in\mathbb{S}^n$ and $\gamma^*\in\mathbb{R}^m$ such that $\operatorname{rank}(Y^*) = k$ and $\operatorname{rank}(M(\gamma^*)) = n-k$. The assumption that $\operatorname{rank}(Y^*) + \operatorname{rank}(M(\gamma^*)) = n$ is referred to as *strict complementarity* and is known to hold generically conditioned on primal and dual attainability [5].

Second, we will assume (Assumption 16) that the optimal primal solution Y^* is known *a priori* on some k-dimensional subspace $W^\perp$, on which it is positive definite. This assumption is inspired by QCQP and QMP applications [17, 161, 181]: Recall that the standard SDP relaxation [161] of an equality-constrained QCQP (in the variable $x\in\mathbb{R}^{n-1}$) is given by

$$\begin{aligned}\inf_{x\in\mathbb{R}^{n-1}}&\left\{\begin{pmatrix}x\\1\end{pmatrix}^\intercal M_{\text{obj}}\begin{pmatrix}x\\1\end{pmatrix} :\ \begin{pmatrix}x\\1\end{pmatrix}^\intercal M_i\begin{pmatrix}x\\1\end{pmatrix} = 0,\ \forall i\in[m]\right\}\\ &\geq \inf_{Y\in\mathbb{S}^n}\left\{\left\langle M_{\text{obj}}, Y\right\rangle : \begin{array}{l}\langle M_i, Y\rangle = 0,\ \forall i\in[m]\\ Y = \begin{pmatrix}* & *\\ * & 1\end{pmatrix}\succeq 0\end{array}\right\}.\end{aligned}$$

Thus, the optimal solution (in fact, any feasible solution) to the SDP will have a 1 in the bottom-right corner. Taking W to be the subspace corresponding to the first $(n-1)$-coordinates of $\mathbb{R}^n$, we have that the restriction of Y^* to $W^\perp$ is known *a priori* and is positive definite. Similarly, the standard SDP relaxation [17] of an equality-constrained QMP (in the variable $X\in\mathbb{R}^{(n-k)\times k}$) is given by

$$\begin{aligned}\inf_{X\in\mathbb{R}^{(n-k)\times k}}&\left\{\operatorname{tr}\left(\begin{pmatrix}X\\I_k\end{pmatrix}^\intercal M_{\text{obj}}\begin{pmatrix}X\\I_k\end{pmatrix}\right) :\ \operatorname{tr}\left(\begin{pmatrix}X\\I_k\end{pmatrix}^\intercal M_i\begin{pmatrix}X\\I_k\end{pmatrix}\right) = 0,\ \forall i\in[m]\right\}\\ &\geq \inf_{Y\in\mathbb{S}^n}\left\{\left\langle M_{\text{obj}}, Y\right\rangle : \begin{array}{l}\langle M_i, Y\rangle = 0,\ \forall i\in[m]\\ Y = \begin{pmatrix}* & *\\ * & I_k\end{pmatrix}\succeq 0\end{array}\right\}.\end{aligned}$$

Taking W to be the subspace corresponding to the first $(n-k)$ coordinates of $\mathbb{R}^n$, we have that the restriction of Y^* to $W^\perp$ is known *a priori* and positive definite.

We will refer to SDPs where Assumptions 15 and 16 hold as *rank-k exact QMP-like SDPs* or *k-exact SDPs* for short.

6.1.2 Overview and outline of the chapter

In this chapter, we develop a new FOM for rank-k exact QMP-like SDPs. This FOM enjoys low iteration complexity, simple iterative subprocedures, storage optimality, and strong numerical performance. A summary of our contributions, along with an outline of the remainder of this chapter, is as follows. For the sake of presentation, we will assume that W corresponds to the first $n-k$ coordinates of $\mathbb{R}^n$ in the following outline.

- We close this section by discussing thematically related work in storage-optimal or storage-efficient FOMs for solving SDPs and FOMs for solving the GTRS. We then discuss some work on acceleration within FOMs with inexact prox oracles and FOMs for saddle-point problems as these are related to our techniques.

- In Section 6.2, we show how to reformulate a k-exact SDP as a *strongly convex* quadratic matrix minimax problem (QMMP) using a *certificate of strict complementarity* (see Definition 23). There are two key ideas here: First, in the setting of k-exact SDPs, we may parameterize the rank-k matrices in $\mathbb{S}^n_+$ which agree with the restriction of Y^* to $W^\perp$ as

$$Y(X) \coloneqq \begin{pmatrix} XX^\intercal & X(Z^*)^{1/2} \\ (Z^*)^{1/2}X^\intercal & Z^* \end{pmatrix},$$

  where $Z^* \succ 0$ is the known restriction of Y^* to $W^\perp$ and $X \in \mathbb{R}^{(n-k)\times k}$ is unknown. The task of recovering Y^* then reduces to the task of recovering X^*. We replace the variable $Y \in \mathbb{S}^n_+$ with the parameterization $Y(X)$ in the primal SDP to derive a *nonconvex* QMP in the variable X whose optimizer is X^*. This first step can be compared to the Burer–Monteiro reformulation (see Remark 72). The second key idea then shows that this nonconvex QMP can be further reformulated into a strongly convex QMMP ($\text{QMMP}_{\mathcal{U}}$) given a certificate of strict complementarity $\mathcal{U} \subseteq \mathbb{R}^m$. Theorem 27 verifies that the minimax problem

$$\min_{X\in\mathbb{R}^{(n-k)\times k}} \max_{\gamma\in\mathcal{U}} \Big(\langle M(\gamma), Y(X)\rangle + d(\gamma) \Big) \tag{6.1}$$

  has X^* as its unique optimizer and $\text{Opt}_{(\text{SDP})}$ as its optimal value.

- In Section 6.3, we derive a two-level accelerated FOM for solving strongly convex QMMPs of the form (6.1). Due to the minimax structure of (6.1), we focus on Nesterov's optimal method for strongly convex minimax problems [132, Algorithm 2.3.13]. This algorithm relies on a prox-map (see Definition 24) computation in each iteration, and its analysis assumes that prox-map is given by an explicit expression or can be computed exactly. In our setting, the prox-map will not admit a closed-form expression in general. Instead, we will treat the prox-map as an optimization problem in its own right and solve it via an inner FOM. Therefore, we suggest CautiousAGD (Algorithm 9), a new variant of [132, Algorithm 2.3.13] that handles inexact computations in the prox-map procedure. We extend the original estimating sequences analysis of [132, Algorithm 2.3.13] to prove bounds on the accuracy required in each individual prox-map computation to recover an accelerated linear convergence rate in terms of outer iterations (see Theorem 29). In our case, the prox-map can be computed efficiently using an inner loop via the strongly convex excessive gap technique [132, Chapter 6.2]. In all, CautiousAGD computes an ϵ-optimal solution of a QMMP after $O(\log(\epsilon^{-1}))$ outer iterations and $O\left(\epsilon^{-1/2}\right)$ total inner iterations.

- In Section 6.4, we show how to combine any method for producing iterates $\gamma^{(i)} \to \gamma^*$ with CautiousAGD to construct a certificate of strict complementarity. Combined with Algorithm 9, this completes the description of our new FOM, CertSDP (Algorithm 10), for rank-k exact QMP-like SDPs. Informally, we show that CertSDP returns an ϵ-optimal solution to the underlying SDP after performing a fixed (i.e., independent of ϵ) number of iterations of $\gamma^{(i)} \to \gamma^*$ plus either $O(\log(\epsilon^{-1}))$ outer iterations or $O(\epsilon^{-1})$ inner iterations in CautiousAGD. See Theorem 32 for a formal statement.

- In Section 6.5, we present numerical experiments comparing an implementation of CertSDP with similar algorithms from the literature [60, 136, 164, 198], on random sparse k-exact SDP instances with $n \approx 10^3$, 10^4, and 10^5. Our code outperforms previous state of the art and is the only algorithm among those we tested that was able to solve our largest instances to high accuracy.

6.1.3 Related work

Storage-optimal/efficient FOMs. A growing body of literature, itself containing multiple research strands, has explored FOMs for SDPs [9, 23, 51, 60, 70, 104, 115, 118, 136, 160, 164, 192, 197, 198]. Below, we recount some recent developments in this direction with a particular view towards storage-efficient or *storage-optimal* FOMs for SDPs admitting low-rank solutions. Storage-optimality alludes to the fact that a rank-k PSD matrix $Y \in \mathbb{S}^n_+$ can be represented as the outer product of an $n \times k$ *factor matrix* with itself, i.e., $Y = XX^\intercal$ for some $X \in \mathbb{R}^{n\times k}$, so that a primal iterate with rank k can be implicitly stored using only $O(nk)$ memory. Similarly, a dual iterate may be stored using only $O(m)$ memory. Then, a storage-optimal FOM is allowed to use only $O(m + nk)$ storage where k is the rank of the *true* primal SDP solution.

Low-storage and storage-optimal FOMs are particularly attractive for SDPs where $M_{\text{obj}},\ M_1, \dots,\ M_m$ are either structured or sparse, so that it is possible to not only store the instance efficiently, but also to compute matrix-vector products efficiently [60]. The algorithm that we develop in this chapter follows this pattern and similarly interacts with $M_{\text{obj}},\ M_1, \dots,\ M_m$ via *only* matrix-vector products.

One paradigm towards developing storage-optimal FOMs leverages duality to construct surrogate primal SDPs that can be solved with optimal storage. In this paradigm, the variable $Y \in \mathbb{S}^n_+$ is *compressed*, i.e., replaced with $U\overline{Y}U^\intercal$ for some matrix $U \in \mathbb{R}^{n\times k}$ and $\overline{Y} \in \mathbb{S}^k_+$. Ding et al. [60] give rigorous guarantees for such a method assuming strict complementarity. Specifically, they show that if $U \in \mathbb{R}^{n\times k}$ corresponds to a minimum eigenspace of an approximate dual solution, then the optimal solution $\overline{Y}$ of the compressed SDP (in penalty form) is a good approximation of the true primal solution. Then, combining their bounds with existing FOMs for solving the dual SDP approximately, Ding et al. [60] show that $\left\|U\overline{Y}U^\intercal - Y^*\right\|_F \leq \epsilon$ after $O(\epsilon^{-2})$-many minimum eigenvector computations. It is unclear how this convergence guarantee changes when only approximate eigenvector computations (which are the only practical option) are allowed. Friedlander and Macêdo [70] explore a similar idea for trace-minimization SDPs (i.e., SDPs where $M_{\text{obj}} = I$) from the viewpoint of *gauge duality*. Specifically, they show that if U corresponds to a minimum eigenspace associated with the true solution to the gauge dual, then the optimal solution of the compressed SDP exactly recovers the true primal SDP solution. Unfortunately, they do not analyze the accuracy of the recovered primal solution when the gauge dual is solved only approximately, which is the case in practice.

A second paradigm towards developing storage efficient/optimal FOMs works simultaneously in both the primal and dual spaces by employing *linear sketches*. Yurtsever et al. [198] apply the Nyström sketch to the conditional gradient–augmented Lagrangian (CGAL) technique [197] to derive SketchyCGAL. They show that it is possible to reconstruct a $(1+\zeta)$-optimal rank-k approximation of an ϵ-optimal solution to the primal SDP[1] by tracking only the dual iterates as well as a $O(nk/\zeta)$-sized sketch of the primal iterates. When the true solution is unique and has rank-k, it is appropriate to take $\zeta = O(1)$ so that the total storage is $O(m + nk)$. Furthermore, Yurtsever et al. [198] bound the required accuracy in the approximate eigenvector computations within SketchyCGAL. In all, they show that it is possible to implement their algorithm in $O(\epsilon^{-2})$ iterations where each iteration involves computing an eigenvector via $\tilde{O}\left(\epsilon^{-1/2}\right)$ matrix-vector products. In follow-up work, Shinde et al. [160] combine the algorithmic architecture of SketchyCGAL with the additional observation that in specific applications (e.g., max-cut), the goal is simply to *sample* from a Gaussian distribution with variance given by an approximate solution Y_ϵ to the SDP. Under this alternate goal, it is possible to further reduce the storage requirements to $O(n + m)$.

One may compare these storage-optimal FOMs for SDPs with the Burer–Monteiro method [36]. In the Burer–Monteiro method, the convex SDP in the variable $Y \in \mathbb{S}^n_+$ is explicitly replaced with an outer product term involving an $n \times k'$ factor matrix where $k' \geq k$. The resulting *nonconvex* problem is then tackled via local optimization methods. While results [31, 46, 47] have shown that non-global local minima cannot exist when $k' = \Omega(\sqrt{m})$ (so that local optimization methods are certifiably correct), more recent work [176] has shown that such spurious local minima can in fact exist even if $k = 1$ and $k' = \Theta(\sqrt{m})$. In other words, the Burer–Monteiro approach provably cannot achieve storage-optimality.

FOMs for the GTRS. The algorithms developed in the current chapter are inspired by recent developments in FOMs for the TRS and the GTRS. There has been extensive work [41, 82, 87, 94, 124, 148, 165, 180, 183] towards developing customized algorithms for the TRS and GTRS that circumvent solving large SDPs; see Chapter 5 and references therein for a more thorough

[1] In [198], a rank-k matrix $\tilde{Y} \in \mathbb{S}^n_+$ is a $(1+\zeta)$-optimal rank-k approximation of an ϵ-optimal solution $Y_\epsilon \in \mathbb{S}^n_+$ if $\left\|Y_\epsilon - \tilde{Y}\right\|_* \leq (1+\zeta)\|Y_\epsilon - [Y_\epsilon]_k\|_*$ where $\|\cdot\|_*$ is the nuclear norm and $[Y_\epsilon]_k$ is the best rank-k approximation of Y_ϵ.

account of algorithmic ideas for solving large-scale GTRS instances. We highlight only the two most relevant results from this area.

Carmon and Duchi [41] consider iterative methods that produce Krylov subspace solutions to the TRS, i.e., solutions to the TRS restricted to a Krylov subspace generated by the objective function. They show that these solutions converge to the true TRS solution *linearly* as long as the linear term in q_{obj} is not orthogonal to the minimum eigenspace of the Hessian in q_{obj}. One may interpret this condition as requiring strict complementarity between the SDP relaxation of the TRS and its dual.

More recently, Wang et al. [183] make a connection between the GTRS and optimal FOMs for strongly convex minimax problems [132]. In the language of the current chapter, Wang et al. [183] assume strict complementarity between the SDP relaxation of the GTRS and its dual, and show how to construct a strongly convex reformulation of the GTRS using low-accuracy eigenvalue computations. More concretely, they show how to construct $\tilde{\gamma}_-$ and $\tilde{\gamma}_+$ such that the minimax problem

$$\min_{x\in\mathbb{R}^n} \max_{\gamma\in[\tilde{\gamma}_-,\tilde{\gamma}_+]} \left(q_{\text{obj}}(x) + \gamma \cdot q_1(x)\right)$$

is strongly convex and has as its unique optimizer the optimizer of the underlying GTRS. The resulting strongly convex minimax problem is then solved via [132, Algorithm 2.3.13] to achieve a linear convergence rate. One may compare the strongly convex reformulation of the GTRS in [183] with the more natural Lagrangian reformulation (through S-lemma):

$$\min_{x\in\mathbb{R}^n} \sup_{\gamma\in\Gamma} \left(q_{\text{obj}}(x) + \gamma \cdot q_1(x)\right)$$

where $\Gamma = \left\{\gamma \in \mathbb{R}_+ : \, q_{\text{obj}} + \gamma q_1 \text{ is convex}\right\}$. Specialized FOMs have also been developed for the GTRS using this Lagrangian reformulation [180]. Unfortunately, since the Lagrangian reformulation may not be strongly convex in general, the resulting algorithms can only achieve sublinear (in terms of ϵ) convergence rates—specifically, rates of the form $O\left(\epsilon^{-1/2}\right)$ as opposed to rates of the form $O\left(\log(\epsilon^{-1})\right)$.

Accelerated FOMs for non-smooth problems via saddle-point problems. One may treat the QMMP reformulation (6.1) of the SDP as a saddle-point problem in the variables $(X, \gamma) \in \mathbb{R}^{(n-k)\times k} \times \mathbb{R}^m$ as opposed to a non-smooth problem in just $X \in \mathbb{R}^{(n-k)\times k}$. There is a vast body of work developing accelerated FOMs for non-smooth problems that leverages saddle-point structure [97, 128, 132, 139]. Both Nesterov [131] and Nemirovski [128] achieve an accelerated convergence rate of $O(\epsilon^{-1})$ for general convex–concave saddle point problems (see also [172]). This rate can be further improved for the special case of strongly convex–concave saddle-point problems [43, 97, 130]: Nesterov's excessive gap technique [130, 132] achieves an $O(\epsilon^{-1/2})$ convergence for strongly convex–concave saddle-point problems where the coupling term is linear. This is generalized in [43] to allow nonlinear proximal operators. Hamedani and Aybat [81], Juditsky and Nemirovski [97] generalize this convergence rate to the setting where the gradient of the coupling term is only assumed to be Lipschitz. These rates match the known [138] lower bound of $O(\epsilon^{-1/2})$ for any FOM on the general class of strongly convex-concave saddle-point problems. Note that the assumption that the gradient of the coupling term is Lipschitz does not hold for our setting. Indeed, the saddle point function we are interested in, $\langle M(\gamma),\, Y(X)\rangle$, is jointly *cubic* in the variables (X, γ) (so that the gradients vary quadratically). Nonetheless, we will show that it is possible achieve the optimal $O(\epsilon^{-1/2})$ iteration complexity in our setting.

Accelerated FOMs with inexact first-order information. A related line of work [57, 58] has analyzed the convergence rate of (accelerated) FOMs in the presence of inexact first-order information. Devolder et al. [58] analyzes FOMs for smooth convex functions. In [57], the same authors extend these results to FOMs for smooth and strongly convex functions. Our algorithm (Algorithm 9) continues this line of work by considering an inexact *prox-map* for strongly convex *max-type functions*.

6.1.4 Additional notation

Given $X \in \mathbb{R}^{n\times m}$, let $\|X\|_F$ denote the Frobenius norm of X. Given $X \in \mathbb{S}^n$, let $\|X\|_2$ denote the spectral norm of X. Let W be a subspace of $\mathbb{R}^n$. Abusing notation, we write $\mathbb{S}^W$ to denote the vector space of self-adjoint operators on W and $\mathbb{R}^{W,W^\perp}$ for the vector space of linear maps from $W^\perp$ to W. Given $M \in \mathbb{S}^n$ let $M_W \in \mathbb{S}^W$, $M_{W,W^\perp} \in \mathbb{R}^{W,W^\perp}$, and $M_{W^\perp} \in \mathbb{S}^{W^\perp}$ denote the restrictions of M to the corresponding subspaces. Given $x \in \mathbb{R}^n$ and $r \geq 0$, let $\mathbb{B}(x, r)$ denote the closed ℓ_2-ball centered at x with radius r. Given a function in multiple arguments $f(x_1, \dots, x_m)$, we write $\nabla_k f(x_1, \dots, x_m)$ to denote the gradient of f in the kth argument evaluated at $x_1, \dots, x_m$.

6.2 Strongly convex reformulations of k-exact SDPs

In this section, we describe how to construct a strongly convex reformulation of a *rank-k exact QMP-like SDP* using a *certificate of strict complementarity* (see Definitions 22 and 23). The following sections will expand on these ideas and show how these properties can be exploited to achieve algorithmic efficiency.

6.2.1 Definitions and problem setup

We make the following two assumptions on (SDP).

Assumption 15. Assume in (SDP) that the primal and dual problems are both solvable, strong duality holds, and there exist primal and dual optimal solutions $Y^* \in \mathbb{S}^n$ and $\gamma^* \in \mathbb{R}^m$ such that $\operatorname{rank}(Y^*) = k$ and $\operatorname{rank}(M(\gamma^*)) = n - k$. □

We fix Y^* and γ^* to be solutions to (SDP) satisfying $\operatorname{rank}(Y^*) = k$ and $\operatorname{rank}(M(\gamma^*)) = n - k$.

Assumption 16. Let $W \subseteq \mathbb{R}^n$ be a k-dimensional subspace such that the restriction of Y^* to $W^\perp$ is known and positive definite. □

Definition 22. We say that an instance of (SDP) is a *rank-k exact QMP-like SDP* or a *k-exact SDP* for short if both Assumptions 15 and 16 hold. □

Definition 23. We say that a compact subset $\mathcal{U} \subseteq \mathbb{R}^m$ *certifies strict complementarity* if $\gamma^* \in \mathcal{U}$ and, for all $\gamma \in \mathcal{U}$, it holds that $M(\gamma)_W \succ 0$. □

Remark 71. Suppose we are given a certificate of strict complementarity $\mathcal{U}$, i.e., $\gamma^* \in \mathcal{U}$ and $M(\gamma)_W \succ 0$ for all $\gamma \in \mathcal{U}$. We immediately deduce that $\operatorname{rank}(M(\gamma^*)) \geq \operatorname{rank}(M(\gamma^*)_W) = n - k$. On the other hand, $\operatorname{rank}(Y^*) \geq \operatorname{rank}(Y^*_{W^\perp}) = k$. This is the sense in which $\mathcal{U}$ *certifies strict complementarity*. □

6.2.2 Identifying $\mathbb{S}^n$ with quadratic matrix functions

Suppose (SDP) is a k-exact SDP and that $\mathcal{U}$ certifies strict complementarity. For ease of presentation, we will assume in this subsection that W is the $(n-k)$-dimensional subspace corresponding to the first $n-k$ coordinates of $\mathbb{R}^n$. This is without loss of generality and our results extend in the natural way to the setting where W is general.

Our strongly convex reformulation of (SDP) will regard the $M_i \in \mathbb{S}^n$ as inducing *quadratic matrix functions* on the space $\mathbb{R}^{W \times W^\perp} \simeq \mathbb{R}^{(n-k)\times k}$. We begin by writing each M_i, for $i \in \{\mathrm{obj}\} \cup [m]$, as a block matrix

$$M_i = \begin{pmatrix} A_i/2 & \tilde{B}_i/2 \\ \tilde{B}_i^\intercal/2 & C_i \end{pmatrix},$$

where $A_i \in \mathbb{S}^{n-k}$, $\tilde{B}_i \in \mathbb{R}^{(n-k)\times n}$ and $C_i \in \mathbb{S}^k$.

We will partition Y^* as a block matrix with compatible block structure: Define $Z^* \coloneqq Y^*_{W^\perp}$ and $X^* \coloneqq Y^*_{W,W^\perp}(Z^*)^{-1/2}$. Note here that Z^* is known *a priori* due to Assumption 16. Next, by the assumption that $\operatorname{rank}(Y^*) = k$ (Assumption 15), we have that

$$Y^* = \begin{pmatrix} X^* X^{*\intercal} & X^*(Z^*)^{1/2} \\ (Z^*)^{1/2}(X^*)^\intercal & Z^* \end{pmatrix}.$$

Finally, given $X \in \mathbb{R}^{(n-k)\times k}$, define

$$Y(X) \coloneqq \begin{pmatrix} XX^\intercal & X(Z^*)^{1/2} \\ (Z^*)^{1/2}X^\intercal & Z^* \end{pmatrix}$$

and note that $Y(X^*) = Y^*$.

One of our key ideas in building a strongly convex reformulation of (SDP) is that $Y(X)$ is a matrix whose entries are *quadratic* in X. We can thus identify each M_i with a quadratic matrix function. For each $i \in \{\mathrm{obj}\} \cup [m]$, define

$$\begin{aligned} q_i(X) \coloneqq \langle M_i, Y(X)\rangle + d_i &= \frac{\operatorname{tr}(X^\intercal A_i X)}{2} + \left\langle \tilde{B}_i(Z^*)^{1/2}, X \right\rangle + \langle C_i, Z^*\rangle + d_i \\ &= \frac{\operatorname{tr}(X^\intercal A_i X)}{2} + \langle B_i, X\rangle + c_i, \end{aligned}$$

where we let $B_i := \tilde{B}_i(Z^*)^{1/2}$ and $c_i := \langle C_i, Z^*\rangle + d_i$. Finally, given $\gamma \in \mathbb{R}^m$, define $A(\gamma) := A_{\text{obj}} + \sum_{i=1}^m \gamma_i A_i$. We define $B(\gamma),\ \tilde{B}(\gamma),\ c(\gamma),\ d(\gamma)$, and $q(\gamma, X)$ analogously.

Remark 72. One may compare our parameterization of rank-k matrices in $\mathbb{S}^n_+$ with the Burer-Monteiro approach [36]. In the Burer-Monteiro approach, one replaces the matrix variable $Y \in \mathbb{S}^n_+$ with a rank-k matrix variable parameterized by

$$Y = \begin{pmatrix} X \\ X' \end{pmatrix}\begin{pmatrix} X \\ X' \end{pmatrix}^\intercal$$

where $X \in \mathbb{R}^{(n-k)\times k}$ and $X' \in \mathbb{R}^{k\times k}$. This transformation replaces the $O(n^2)$-dimensional variable $Y \in \mathbb{S}^n_+$ with the nk-dimensional variable $(X; X') \in \mathbb{R}^{n\times k}$. Unfortunately, this approach also transforms the SDP from a convex problem into a *nonconvex* problem. Our parameterization will allow us to remedy this nonconvexity. As we will see in the next subsection, when Z^* is known (so that we may fix X') *a priori*, we may further reformulate the remaining *nonconvex* problem in $X \in \mathbb{R}^{(n-k)\times k}$ into a *strongly convex* problem. □

6.2.3 A strongly convex reformulation of (SDP)

The following theorem states that if $\mathcal{U}$ certifies strict complementarity, then X^* is the unique minimizer of a strongly convex *quadratic matrix minimax problem* (QMMP) that can be constructed from $\mathcal{U}$.

Theorem 27. *Suppose* (SDP) *is a rank-k exact QMP-like SDP and that $\mathcal{U}$ certifies strict complementarity. Then, X^* is the unique minimizer of the strongly convex QMMP*

$$\min_{X\in\mathbb{R}^{W\times W^\perp}} \max_{\gamma\in\mathcal{U}} q(\gamma, X). \tag{QMMP$_\mathcal{U}$}$$

Furthermore, $X^* = -A(\gamma^*)^{-1}B(\gamma^*)$ and $\text{Opt}_{(\text{QMMP}_\mathcal{U})} = \text{Opt}_{(\text{SDP})}$.

Proof. Without loss of generality, we work in the basis where W is the first $n-k$ coordinates of $\mathbb{R}^n$. Note that the assumption that $\mathcal{U}$ certifies strict complementarity implies that $A(\gamma^*) = M(\gamma^*)_W \succ 0$.

We begin by verifying that $X^* = -A(\gamma^*)^{-1}B(\gamma^*)$. By complementary slackness, we have

$$\begin{aligned}
0 &= \langle M(\gamma^*), Y(X^*)\rangle \\
&= \operatorname{tr}\left(\begin{pmatrix} X^* \\ (Z^*)^{1/2} \end{pmatrix}^\intercal \begin{pmatrix} A(\gamma^*)/2 & \tilde{B}(\gamma^*)/2 \\ \tilde{B}(\gamma^*)^\intercal/2 & C(\gamma^*) \end{pmatrix}\begin{pmatrix} X^* \\ (Z^*)^{1/2} \end{pmatrix}\right) \\
&= \operatorname{tr}\left(\frac{(X^* + A(\gamma^*)^{-1}B(\gamma^*))^\intercal A(\gamma^*)(X^* + A(\gamma^*)^{-1}B(\gamma^*))}{2}\right) \\
&\quad + \left[\langle C(\gamma^*), Z^*\rangle - \operatorname{tr}\left(\frac{B(\gamma^*)^\intercal A(\gamma^*)^{-1}B(\gamma^*)}{2}\right)\right].
\end{aligned}$$

Here, the second line follows by the definitions of $M(\gamma^*)$ and $Y(X^*)$, and the third line follows from the definition $B(\gamma) := \tilde{B}(\gamma)(Z^*)^{1/2}$. We claim that the square-bracketed term on the final line is zero: By the assumption that $\operatorname{rank}(M(\gamma^*)) = n - k$ and the fact that $A(\gamma^*) \succ 0$, we have that $C(\gamma^*) = \frac{\tilde{B}(\gamma^*)^\intercal A(\gamma^*)^{-1}\tilde{B}(\gamma^*)}{2}$. Pre- and post-multiplying $C(\gamma^*)$ by $(Z^*)^{1/2}$ and taking the trace of this identity gives

$$\langle C(\gamma^*), Z^* \rangle = \operatorname{tr}\left(\frac{(Z^*)^{1/2}\tilde{B}(\gamma^*)^\intercal A(\gamma^*)^{-1}\tilde{B}(\gamma^*)(Z^*)^{1/2}}{2}\right) = \operatorname{tr}\left(\frac{B(\gamma^*)^\intercal A(\gamma^*)^{-1}B(\gamma^*)}{2}\right).$$

Thus, we have that

$$0 = \operatorname{tr}\left(\left(X^* + A(\gamma^*)^{-1}B(\gamma^*)\right)^\intercal A(\gamma^*)\left(X^* + A(\gamma^*)^{-1}B(\gamma^*)\right)\right),$$

so that $X^* = -A(\gamma^*)^{-1}B(\gamma^*)$ by the positive definiteness of $A(\gamma^*)$.

Next, note that by the feasibility of Y^*, we have $q_i(X^*) = \langle M_i, Y^* \rangle + d_i = 0$ for all $i \in [m]$. Similarly, by the optimality of Y^*, we have $q_{\text{obj}}(X^*) = \left\langle M_{\text{obj}}, Y^* \right\rangle = \text{Opt}_{(\text{SDP})}$. In particular, $\max_{\gamma\in\mathcal{U}} q(\gamma, X^*) = q(\gamma^*, X^*) = \text{Opt}_{(\text{SDP})}$. On the other hand, for any $X \in \mathbb{R}^{W\times W^\perp}$,

$$\max_{\gamma\in\mathcal{U}} q(\gamma, X) \geq q(\gamma^*, X) = \text{Opt}_{(\text{SDP})} + \operatorname{tr}\left(\frac{(X - X^*)^\intercal A(\gamma^*)(X - X^*)}{2}\right).$$

As $A(\gamma^*) \succ 0$, we conclude that X^* is the unique minimizer of ($\text{QMMP}_{\mathcal{U}}$) with optimal value $\text{Opt}_{(\text{QMMP}_{\mathcal{U}})} = \text{Opt}_{(\text{SDP})}$.

Finally, strong convexity of ($\text{QMMP}_{\mathcal{U}}$) follows from compactness of $\mathcal{U}$ and the assumption that $\mathcal{U}$ certifies strict complementarity (so that $A(\gamma) = M(\gamma)_W$ is positive definite over $\mathcal{U}$). ■

Remark 73. One may compare ($\text{QMMP}_{\mathcal{U}}$) with the more natural Lagrangian formulation of (SDP), which results in a QMMP in the same space:

$$\min_{X\in\mathbb{R}^{W\times W^\perp}} \sup_{\gamma\in\mathbb{R}^m:\, A(\gamma)\succeq 0} q(\gamma, X). \tag{6.2}$$

Indeed, it is possible to show that X^* is also the unique minimizer of (6.2). Nevertheless, the formulation (6.2), in contrast to ($\text{QMMP}_{\mathcal{U}}$), has two major downsides: First, it may be the case that $\sup_{\gamma\in\mathbb{R}^m:\, A(\gamma)\succeq 0} q(\gamma, X)$ is a convex function in X that is *not strongly convex*. Second, the domain of the supremum, $\{\gamma \in \mathbb{R}^m : A(\gamma) \succeq 0\}$, is itself a spectrahedron so that *even evaluating* $\sup_{\gamma\in\mathbb{R}^m:\, A(\gamma)\succeq 0} q(\gamma, X)$ (that is, evaluating *zeroth-order* information in the X variable) requires solving an SDP. In contrast, ($\text{QMMP}_{\mathcal{U}}$) is strongly convex by construction. Furthermore, we may pick $\mathcal{U}$ to have efficient projection and linear maximization oracles (e.g., by taking $\mathcal{U}$ to be an ℓ_2 ball). From this viewpoint, ($\text{QMMP}_{\mathcal{U}}$) will be much more amenable than (6.2) to first-order methods. □

6.3 Algorithms for strongly convex QMMPs

In this section, we describe and analyze an accelerated first-order method (FOM) for solving strongly convex QMMPs. While we will apply this algorithm to problems arising from the application of Theorem 27, the algorithms from this section can handle *general* strongly convex QMMPs.

We state explicitly the setup and assumptions of this section. Let $q_{\text{obj}}, q_1, \dots, q_m : \mathbb{R}^{(n-k)\times k} \to \mathbb{R}$ be quadratic matrix functions of the form

$$q_i(X) = \frac{\operatorname{tr}(X^\intercal A_i X)}{2} + \langle B_i, X\rangle + c_i.$$

Given $\gamma \in \mathbb{R}^m$, let $A(\gamma) := A_{\text{obj}} + \sum_{i=1}^m \gamma_i A_i$. Define $B(\gamma),\ c(\gamma),\ q(\gamma, X)$ analogously.

Let $\mathcal{U} \subseteq \mathbb{R}^m$ be a compact convex set with exact projection and linear maximization oracles. Our goal is to find an ϵ-optimal solution to

$$\min_{X\in\mathbb{R}^{(n-k)\times k}} \max_{\gamma\in\mathcal{U}} q(\gamma, X). \tag{QMMP}$$

That is, our goal is to find some $\tilde{X} \in \mathbb{R}^{(n-k)\times k}$ satisfying $\max_{\gamma\in\mathcal{U}} q(\gamma, \tilde{X}) \leq \text{Opt}_{(\text{QMMP})} + \epsilon$. For notational convenience, we will define

$$Q(X) := \max_{\gamma\in\mathcal{U}} q(\gamma, X).$$

While we will treat $\mathcal{U}$ as fixed in this section, in future sections, we will explicitly call attention to the dependence of the function Q on the set $\mathcal{U}$ and write $Q_{\mathcal{U}}$ instead.

We present a FOM for (QMMP) under two assumptions. The first assumption (Assumption 17) requires uniform strong convexity and smoothness of $q(\gamma, X)$ over $\mathcal{U}$.

Assumption 17. We will assume algorithmic access to parameters $0 < \mu \leq L$ such that $\mu I \preceq A(\gamma) \preceq LI$ for all $\gamma \in \mathcal{U}$. □

When Assumption 17 holds, we define the *condition number* of (QMMP) as $\kappa := L/\mu$. We will state our second assumption (which bounds the norms of various quantities) when needed in Section 6.3.3.

Our FOM will closely follow Nesterov's accelerated gradient descent scheme for strongly convex minimax functions [132, Algorithm 2.3.13] (henceforth AGD-MM) with one major difference. In contrast to the presentation in [132] and its application in [180], the necessary *prox-map* (see Definition 24 below) in the QMMP setting cannot be computed explicitly or exactly.

We break our FOM for strongly convex QMMPs into two levels, presented as the first two subsections in this section. In Section 6.3.1, we give a convergence analysis for a modified version of AGD-MM using an *inexact* prox-map oracle. In particular, we will bound the necessary accuracy of the prox-map to recover accelerated convergence rates. In Section 6.3.2, we show how to implement the approximate prox-map oracle efficiently for each iteration using the strongly convex excessive gap technique [132, Algorithm 6.2.37]. Finally, in Section 6.3.3, we state an assumption (Assumption 18) that allows us to bound the iteration cost of the prox-map oracle uniformly across iterations. Taken together with the results from the previous subsections, this will give a rigorous guarantee for the overall FOM.

6.3.1 An FOM for strongly convex QMMPs using an inexact prox-map oracle

This subsection generalizes AGD-MM by allowing *inexact* prox-map computations. We first recall the definition of the prox-map and the fundamental relation (6.3) that is used in the convergence rate analysis of AGD-MM. Next, we show how to recover a similar inequality (6.5) when the prox-map is computed only approximately. Finally, we show how to modify the step-sizes in AGD-MM to prevent error accumulation that may otherwise build up from inexact prox-map computations. These step-sizes allow us to recover the accelerated linear convergence rates of AGD-MM even with inexact prox-map computations.

The prox-map

AGD-MM requires computing the prox-map $X_L(\Xi)$ (defined in Definition 24) *exactly* in every iteration (adapted from [132, Definition 2.3.2]).

Definition 24. Let $\Xi \in \mathbb{R}^{(n-k)\times k}$. Define

$$\begin{aligned} Q(\Xi; X) &\coloneqq \max_{\gamma\in\mathcal{U}} \left(q(\gamma,\Xi) + \langle \nabla_2\, q(\gamma,\Xi), X-\Xi\rangle\right)\\ Q_L(\Xi; X) &\coloneqq Q(\Xi; X) + \frac{L}{2}\|X-\Xi\|_F^2\\ Q_L^*(\Xi) &\coloneqq \min_{X\in\mathbb{R}^{(n-k)\times k}} Q_L(\Xi; X)\\ X_L(\Xi) &\coloneqq \operatorname*{arg\,min}_{X\in\mathbb{R}^{(n-k)\times k}} Q_L(\Xi; X)\\ g_L(\Xi) &\coloneqq L(\Xi - X_L(\Xi)). \end{aligned}$$

Here, $\nabla_2\, q(\gamma,\Xi)$ is the gradient of $q(\gamma, X)$ in X at Ξ and is an affine function of γ (more explicitly, $\nabla_2\, q(\gamma,\Xi) = A(\gamma)\Xi + B(\gamma)$). Note that the function $Q(\Xi; X)$ simply replaces the inside function $q(\gamma, X)$ in the definition of $Q(X)$ with its *linearization* around Ξ. The quantities X_L and g_L are the *prox-map* and the *grad-map*. □

Recall also the main property of the prox-map and grad-map that is used in the analysis of the convergence rate of AGD-MM as given in the following lemma (adapted from [132, Theorem 2.3.2]).

Lemma 65. *Let $\Xi \in \mathbb{R}^{(n-k)\times k}$. Then, for all $X \in \mathbb{R}^{(n-k)\times k}$,*

$$Q(X) \geq Q(X_L(\Xi)) + \frac{1}{2L}\|g_L(\Xi)\|_F^2 + \langle g_L(\Xi), X-\Xi\rangle + \frac{\mu}{2}\|X-\Xi\|_F^2. \qquad (6.3)$$

An approximate prox-map inequality

In the setting of general QMMPs, it is not possible to compute the prox-map exactly. Instead, we will apply an inner FOM to solve the prox-map $X_L(\Xi)$ to some prescribed accuracy. This necessitates an analysis of (a variant of) AGD-MM that works with inexact prox-map computations. To this end, we show how to recover a version of (6.3) where $X_L(\Xi)$ is computed only approximately.

Define

$$\tilde{\mu} := \mu/2, \qquad \tilde{L} := L - \mu/2, \quad \text{and} \quad \tilde{\kappa} := \tilde{L}/\tilde{\mu}. \tag{6.4}$$

We will need the following geometric fact.

Lemma 66. *Let $\tilde{X}$, $X_L \in \mathbb{R}^{(n-k)\times k}$ be such that $\left\|\tilde{X} - X_L\right\|_F \leq \delta$. Then, for all $X \in \mathbb{R}^{(n-k)\times k}$,*

$$\frac{L}{2}\|X - X_L\|_F^2 \geq \frac{\tilde{L}}{2}\left\|X - \tilde{X}\right\|_F^2 - \frac{L\delta^2}{2}(2\kappa - 1).$$

Proof of Lemma 66. Let $\Delta := \tilde{X} - X_L$. Then,

$$\begin{aligned}\frac{L}{2}\|X - X_L\|_F^2 &= \frac{L}{2}\left\|X - \tilde{X} + \Delta\right\|_F^2 \\ &= \frac{\tilde{L}}{2}\left\|X - \tilde{X}\right\|_F^2 + \frac{\tilde{\mu}}{2}\left\|X - \tilde{X}\right\|_F^2 + L\left\langle X - \tilde{X}, \Delta\right\rangle + \frac{L}{2}\|\Delta\|_F^2,\end{aligned}$$

where the second equality follows from expanding the square and the fact that $L = \tilde{L} + \tilde{\mu}$. Moreover,

$$0 \leq \frac{L}{2}\left\|\sqrt{\frac{\tilde{\mu}}{L}}(X - \tilde{X}) + \sqrt{\frac{L}{\tilde{\mu}}}\Delta\right\|_F^2 = \frac{\tilde{\mu}}{2}\left\|X - \tilde{X}\right\|_F^2 + L\left\langle X - \tilde{X}, \Delta\right\rangle + L\kappa\|\Delta\|_F^2.$$

Combining these two inequalities gives

$$\frac{L}{2}\|X - X_L\|_F^2 \geq \frac{\tilde{L}}{2}\left\|X - \tilde{X}\right\|_F^2 - \frac{L\delta^2}{2}(2\kappa - 1). \qquad \blacksquare$$

We may now derive a variant of (6.3) which only uses an approximate prox-map.

Theorem 28. *Let $\Xi \in \mathbb{R}^{(n-k)\times k}$. Suppose $\tilde{X}$ satisfies*

$$Q_L(\Xi; \tilde{X}) \leq Q_L^*(\Xi) + \epsilon.$$

Set $\tilde{g} := \tilde{L}(\Xi - \tilde{X})$. Then, for all $X \in \mathbb{R}^{(n-k)\times k}$,

$$Q(X) \geq Q(\tilde{X}) + \frac{1}{2\tilde{L}}\|\tilde{g}\|_F^2 + \langle \tilde{g}, X - \Xi\rangle + \frac{\tilde{\mu}}{2}\|X - \Xi\|_F^2 - 2\kappa\epsilon. \tag{6.5}$$

Proof. As $Q_L(\Xi; X)$ is L-strongly convex, from the premise of the lemma we have $\left\|\tilde{X} - X_L(\Xi)\right\|_F \leq \sqrt{2\epsilon/L}$.

We bound

$$\begin{aligned}
Q(X) &\geq Q(\Xi; X) + \frac{\mu}{2}\|X - \Xi\|_F^2 \\
&= Q_L(\Xi; X) - \frac{\tilde{L}}{2}\|X - \Xi\|_F^2 + \frac{\tilde{\mu}}{2}\|X - \Xi\|_F^2 \\
&\geq Q_L^*(\Xi) + \frac{L}{2}\|X - X_L(\Xi)\|_F^2 - \frac{\tilde{L}}{2}\|X - \Xi\|_F^2 + \frac{\tilde{\mu}}{2}\|X - \Xi\|_F^2 \\
&\geq Q(\tilde{X}) + \frac{\tilde{L}}{2}\left\|X - \tilde{X}\right\|_F^2 - \frac{\tilde{L}}{2}\|X - \Xi\|_F^2 + \frac{\tilde{\mu}}{2}\|X - \Xi\|_F^2 - 2\kappa\epsilon \\
&= Q(\tilde{X}) + \frac{\tilde{L}}{2}\left(2\left\langle X - \Xi, \Xi - \tilde{X}\right\rangle + \left\|\Xi - \tilde{X}\right\|_F^2\right) + \frac{\tilde{\mu}}{2}\|X - \Xi\|_F^2 - 2\kappa\epsilon \\
&= Q(\tilde{X}) + \langle \tilde{g}, X - \Xi\rangle + \frac{1}{2\tilde{L}}\|\tilde{g}\|_F^2 + \frac{\tilde{\mu}}{2}\|X - \Xi\|_F^2 - 2\kappa\epsilon.
\end{aligned}$$

Here, the first inequality follows from μ-strong convexity of Q, the first equation follows from the definitions of $Q_L(\Xi; X)$, $\tilde{L}$ and $\tilde{\mu}$, the second inequality follows from optimality of $X_L(\Xi)$, the third inequality follows from Lemma 66 applied with $\delta = \sqrt{2\epsilon/L}$ and the L-smoothness of $q(\gamma, X)$ for each $\gamma \in \mathcal{U}$, and the last two equations follow from expanding the squares and the definition of $\tilde{g}$. ■

Estimating sequences

We now modify the estimating sequences analysis of AGD-MM to use (6.5) instead of (6.3): Fix $X_0 \in \mathbb{R}^{(n-k)\times k}$ and let $\{\epsilon_t\} \subseteq \mathbb{R}_{++}$ and $\{\Xi_t\} \subseteq \mathbb{R}^{(n-k)\times k}$ to be fixed later. Define

$$\phi_0(X) \coloneqq Q(X_0) + \frac{\tilde{\mu}}{2}\|X - X_0\|_F^2.$$

For $t \geq 0$, let X_{t+1} be an ϵ-approximate prox-map, i.e., X_{t+1} satisfies

$$Q_L(\Xi_t; X_{t+1}) \leq Q_L^*(\Xi_t) + \epsilon_t,$$

and set $\tilde{g}_t \coloneqq \tilde{L}(\Xi_t - X_{t+1})$. Let $\alpha \coloneqq \tilde{\kappa}^{-1/2}$ and recursively define

$$\begin{aligned}
\phi_{t+1}(X) &\coloneqq (1-\alpha)\phi_t(X) \\
&\qquad + \alpha\left(Q(X_{t+1}) + \frac{1}{2\tilde{L}}\|\tilde{g}_t\|_F^2 + \langle \tilde{g}_t, X - \Xi_t\rangle + \frac{\tilde{\mu}}{2}\|X - \Xi_t\|_F^2\right).
\end{aligned}$$

The following lemma shows how $\phi_t(X)$ evolves. Its proof follows verbatim from the standard proof [132, Lemma 2.3.3] and is deferred to Section F.1. Indeed, the standard proof makes no assumption on how Ξ_t and X_t are related.

Lemma 67. *For all $t \geq 0$, $\phi_t(X)$ is a quadratic matrix function in X of the form*

$$\phi_t(X) = \phi_t^* + \frac{\tilde{\mu}}{2}\|X - V_t\|_F^2. \tag{6.6}$$

The sequences $\{\phi_t^*\}$, $\{V_t\}$ *are given by* $V_0 = X_0$, $\phi_0^* = Q(X_0)$ *and the recurrences*

$$\begin{aligned} V_{t+1} &= (1-\alpha)V_t + \alpha\left(\Xi_t - \frac{1}{\tilde{\mu}}\tilde{g}_t\right), \textit{ and} \\ \phi_{t+1}^* &= (1-\alpha)\phi_t^* + \alpha\left(Q(X_{t+1}) + \frac{1}{2\tilde{L}}\|\tilde{g}_t\|_F^2\right) - \frac{\alpha^2}{2\tilde{\mu}}\|\tilde{g}_t\|_F^2 \\ &\qquad + \alpha(1-\alpha)\left(\frac{\tilde{\mu}}{2}\|\Xi_t - V_t\|_F^2 + \langle \tilde{g}_t, V_t - \Xi_t\rangle\right). \end{aligned}$$

For all $t \geq 0$, we will henceforth set

$$\Xi_t \coloneqq \frac{X_t + \alpha V_t}{1+\alpha}.$$

The following lemma shows that Ξ_{t+1} can be written as an *extragradient* step from X_t towards X_{t+1}. Its proof follows verbatim from the standard proof [132, Page 92] and is deferred to Section F.1. Indeed, the standard proof only needs the relation $X_{t+1} = \Xi_t - \tilde{g}_t/\tilde{L}$, which continues to hold in our setting by construction.

Lemma 68. *It holds that* $\Xi_0 = X_0$ *and* $\Xi_{t+1} = X_{t+1} + \frac{1-\alpha}{1+\alpha}(X_{t+1} - X_t)$ *for all* $t \geq 0$.

The following two lemmas bound the two types of errors that result from inexact prox-map computations. Define $E_0^{(1)} \coloneqq 0$, $E_0^{(2)} \coloneqq 0$, and for all $t \geq 0$ inductively set

$$E_{t+1}^{(1)} \coloneqq (1-\alpha)E_t^{(1)} + (1-\alpha)\epsilon_t \qquad \text{and} \qquad E_{t+1}^{(2)} \coloneqq (1-\alpha)E_t^{(2)} + \alpha\epsilon_t.$$

Let $E_t \coloneqq E_t^{(1)} + E_t^{(2)}$ be the sum of the two types of errors. Equivalently, let $E_t \coloneqq 0$ and inductively set $E_{t+1} = (1-\alpha)E_t + \epsilon_t$ for all $t \geq 0$.

Lemma 69. *It holds that* $Q(X_t) \leq \phi_t^* + 2\kappa E_t^{(1)}$ *for all* $t \geq 0$.

Proof. It is clear that $Q(X_0) \leq \phi_0^*$. Thus, consider X_{t+1} with $t \geq 0$. By induction and Lemma 67,

$$\begin{aligned} \phi_{t+1}^* &\geq (1-\alpha)Q(X_t) + \alpha Q(X_{t+1}) + \left(\frac{\alpha}{2\tilde{L}} - \frac{\alpha^2}{2\tilde{\mu}}\right)\|\tilde{g}_t\|_F^2 \\ &\qquad + \alpha(1-\alpha)\langle \tilde{g}_t, V_t - \Xi_t\rangle - (1-\alpha)\left(2\kappa E_t^{(1)}\right). \end{aligned}$$

As X_{t+1} satisfies $Q_L(\Xi_t; X_{t+1}) \leq Q^*(\Xi_t) + \epsilon_t$, we deduce (see Theorem 28) that

$$Q(X_t) \geq Q(X_{t+1}) + \frac{1}{2\tilde{L}}\|\tilde{g}_t\|_F^2 + \langle \tilde{g}_t, X_t - \Xi_t\rangle + \frac{\tilde{\mu}}{2}\|X_t - \Xi_t\|_F^2 - 2\kappa\epsilon_t.$$

These two inequalities together lead to

$$\begin{aligned}\phi_{t+1}^* &\geq Q(X_{t+1}) - 2\kappa(1-\alpha)(E_t^{(1)} + \epsilon_t)\\ &\quad + \left(\frac{\alpha}{2\tilde{L}} - \frac{\alpha^2}{2\tilde{\mu}} + \frac{1-\alpha}{2\tilde{L}}\right)\|\tilde{g}_t\|_F^2 + (1-\alpha)\langle \tilde{g}_t, \alpha(V_t - \Xi_t) + (X_t - \Xi_t)\rangle.\end{aligned}$$

It is straightforward to show that the two quantities on the final line are identically zero using the relations $\alpha^2 = \tilde{\mu}/\tilde{L}$ and $\Xi_t = \frac{X_t + \alpha V_t}{1+\alpha}$ (see Lemma 68). ■

Lemma 70. *For all* $t \geq 0$, *it holds that*

$$\phi_t(X) \leq (1-(1-\alpha)^t)Q(X) + (1-\alpha)^t\phi_0(X) + 2\kappa E_t^{(2)}, \quad \forall X \in \mathbb{R}^{(n-k)\times k}.$$

Proof. The statement holds holds for $t = 0$. Thus, consider ϕ_{t+1} for $t \geq 0$. By definition

$$\phi_{t+1}(X) = (1-\alpha)\phi_t(X) + \alpha\Big(Q(X_{t+1}) + \frac{1}{2\tilde{L}}\|\tilde{g}_t\|_F^2 + \langle \tilde{g}_t, X - \Xi_t\rangle + \frac{\tilde{\mu}}{2}\|X - \Xi_t\|_F^2\Big).$$

As X_{t+1} satisfies $Q_L(\Xi_t; X_{t+1}) \leq Q^*(\Xi_t) + \epsilon_t$, we deduce (see Theorem 28) that

$$Q(X) \geq Q(X_{t+1}) + \frac{1}{2\tilde{L}}\|\tilde{g}_t\|_F^2 + \langle \tilde{g}_t, X - \Xi_t\rangle + \frac{\tilde{\mu}}{2}\|X - \Xi_t\|_F^2 - 2\kappa\epsilon.$$

Then, these inequalities combined with the inductive hypothesis give

$$\begin{aligned}\phi_{t+1}(X) &\leq (1-\alpha)\phi_t(X) + \alpha Q(X) + 2\kappa\alpha\epsilon_t\\ &= (1-(1-\alpha)^{t+1})Q(X) + (1-\alpha)(\phi_t(X) - (1-(1-\alpha)^t)Q(X)) + 2\kappa\alpha\epsilon_t\\ &\leq (1-(1-\alpha)^{t+1})Q(X) + (1-\alpha)^{t+1}\phi_0(X) + 2\kappa\Big((1-\alpha)E_t^{(2)} + \alpha\epsilon_t\Big).\end{aligned}$$ ■

Combining Lemmas 69 and 70, we get a bound on the total error due to inexact prox-maps as a function of the accuracy of each individual prox-map.

Corollary 26. *For all* $t \geq 0$, *it holds that*

$$Q(X_t) - \mathrm{Opt}_{(\mathrm{QMMP})} \leq (1-\alpha)^t\Big[2\Big(Q(X_0) - \mathrm{Opt}_{(\mathrm{QMMP})}\Big)\Big] + 2\kappa E_t.$$

Proof. Let $X_{\mathcal{U}}^*$ denote the optimizer of (QMMP) so that $Q(X_{\mathcal{U}}^*) = \mathrm{Opt}_{(\mathrm{QMMP})}$. Then, Lemmas 69 and 70 give

$$\begin{aligned}Q(X_t) - \mathrm{Opt}_{(\mathrm{QMMP})} &\leq \phi_t^* + 2\kappa E_t^{(1)} - Q(X_{\mathcal{U}}^*)\\ &\leq \phi_t(X_{\mathcal{U}}^*) + 2\kappa E_t^{(1)} - Q(X_{\mathcal{U}}^*)\\ &\leq (1-(1-\alpha)^t)Q(X_{\mathcal{U}}^*) + (1-\alpha)^t\phi_0(X_{\mathcal{U}}^*) + 2\kappa E_t - Q(X_{\mathcal{U}}^*)\\ &= (1-\alpha)^t\Big(\phi_0(X_{\mathcal{U}}^*) - \mathrm{Opt}_{(\mathrm{QMMP})}\Big) + 2\kappa E_t.\end{aligned}$$

Note also that by the definition of $\phi_0(\cdot)$ and the μ-strong convexity of Q, we have

$$\phi_0(X^*_{\mathcal{U}}) - \mathrm{Opt}_{(\text{QMMP})} = Q(X_0) - \mathrm{Opt}_{(\text{QMMP})} + \frac{\tilde{\mu}}{2}\|X^*_{\mathcal{U}} - X_0\|_F^2 \leq 2\Big(Q(X_0) - \mathrm{Opt}_{(\text{QMMP})}\Big).$$

Combining the two inequalities completes the proof. ■

We are now ready to present CautiousAGD (Algorithm 9) and its guarantee.

Algorithm 9 CautiousAGD

```
Given $q(\gamma, X)$ and $\mathcal{U}$ satisfying Assumption 17; $X_0 \in \mathbb{R}^{(n-k)\times k}$, and a bound $\mathrm{gap}_0 \in \mathbb{R}$ such
that $Q(X_0) - \mathrm{Opt}_{(\text{QMMP})} \leq \mathrm{gap}_0$
  1. Set $\tilde{\mu}, \tilde{L}, \tilde{\kappa}$ as in (6.4) and $\alpha := \tilde{\kappa}^{-1/2}$. Set $\Xi_0 := X_0$.
  2. For $t \geq 0$
       a) Compute an inexact prox-map $X_{t+1}$ satisfying
            $Q_L(\Xi_t; X_{t+1}) \leq Q_L^*(\Xi_t) + \epsilon_t$,   where   $\epsilon_t = \begin{cases} \frac{\mathrm{gap}_0}{\kappa}\left(1 - \frac{\alpha}{2}\right), & \text{if } t = 0, \text{ and} \\ \frac{\mathrm{gap}_0}{\kappa}\left(1 - \frac{\alpha}{2}\right)^t \frac{\alpha}{2}, & \text{else.} \end{cases}$
                                                                                     (6.7)
       b) Set $\Xi_{t+1} := X_{t+1} + \frac{1-\alpha}{1+\alpha}(X_{t+1} - X_t)$
```

Theorem 29. *Let $q(\gamma, X)$ and $\mathcal{U}$ satisfy Assumption 17. Let gap_0 be a known upper bound on $Q(X_0) - \mathrm{Opt}_{(\text{QMMP})}$ and let X_t denote the iterates produced by Algorithm 9 with starting point X_0. Then, for all $t \geq 1$, the iterate X_t satisfies*

$$Q(X_t) - \mathrm{Opt}_{(\text{QMMP})} \leq \left(1 - \frac{\alpha}{2}\right)^t (4 \cdot \mathrm{gap}_0).$$

In particular, $Q(X_T) - \mathrm{Opt}_{(\text{QMMP})} \leq \epsilon$ after at most

$$T = O\left(\sqrt{\kappa}\log\left(\frac{\mathrm{gap}_0}{\epsilon}\right)\right)$$

iterations. The t-th iteration of Algorithm 9 requires computing a prox-map X_{t+1} satisfying (6.7).

Proof. We first claim that $E_t = (\mathrm{gap}_0/\kappa)(1 - \alpha/2)^t$ for all $t \geq 1$. Indeed, this claim holds for $t = 1$ as $E_1 = \epsilon_0 = (\mathrm{gap}_0/\kappa)(1 - \alpha/2)$ by construction (see (6.7)). Then, by induction

$$E_{t+1} = (1-\alpha)E_t + \epsilon_t = (1-\alpha)\frac{\mathrm{gap}_0}{\kappa}\left(1 - \frac{\alpha}{2}\right)^t + \frac{\mathrm{gap}_0}{\kappa}\left(1 - \frac{\alpha}{2}\right)^t \frac{\alpha}{2} = \frac{\mathrm{gap}_0}{\kappa}\left(1 - \frac{\alpha}{2}\right)^{t+1}.$$

Then, the bound on $Q(X_t) - \mathrm{Opt}_{(\text{QMMP})}$ follows from Corollary 26 and the starting condition $Q(X_0) - \mathrm{Opt}_{(\text{QMMP})} \leq \mathrm{gap}_0$. ■

Remark 74. We refer to Algorithm 9 as CautiousAGD to allude to the fact that Algorithm 9 is simply AGD-MM with inexact prox-maps and smaller extra-gradient steps. Specifically, AGD-MM sets

$$\Xi_{t+1} = X_{t+1} + \left(\frac{1-\kappa^{-1/2}}{1+\kappa^{-1/2}}\right)(X_{t+1} - X_t)$$

whereas CautiousAGD sets

$$\Xi_{t+1} = X_{t+1} + \left(\frac{1-\tilde{\kappa}^{-1/2}}{1+\tilde{\kappa}^{-1/2}}\right)(X_{t+1} - X_t).$$

Note that $\kappa \leq \tilde{\kappa} \leq 2\kappa$. □

6.3.2 Approximating the prox-map

Recall that the prox-map $X_L(\Xi)$ is the minimizer of $Q_L(\Xi; X)$:

$$\min_{X\in\mathbb{R}^{(n-k)\times}} Q_L(\Xi; X) = \min_{X\in\mathbb{R}^{(n-k)\times k}} \max_{\gamma\in\mathcal{U}} \left(\frac{L}{2}\|X-\Xi\|_F^2 + \langle \nabla_2\, q(\gamma,\Xi), X-\Xi\rangle + q(\gamma,\Xi)\right).$$

There are a number of ways to solve for $X_L(\Xi)$. For example, when m is small, one may apply an interior point method to solve for γ in the dual problem:

$$\begin{aligned}\max_{\gamma\in\mathcal{U}} &\left[\min_{X\in\mathbb{R}^{(n-k)\times k}} \left(\frac{L}{2}\|X-\Xi\|_F^2 + \langle \nabla_2\, q(\gamma,\Xi), X-\Xi\rangle + q(\gamma,\Xi)\right)\right]\\ &= \max_{\gamma\in\mathcal{U}}\left(-\frac{1}{2L}\|\nabla_2\, q(\gamma,\Xi)\|_F^2 + q(\gamma,\Xi)\right). \end{aligned} \tag{6.8}$$

Note here that strong duality holds as the term inside the parenthesis is linear in γ and convex quadratic in X and $\mathcal{U}$ is a compact convex set so we can apply Sion's Minimax Theorem [162]. An approximate primal solution $\tilde{X}$ can then be reconstructed from an approximate solution $\tilde{\gamma}$ of the dual problem by setting $\tilde{X} = \Xi - \frac{\nabla_2\, q(\tilde{\gamma},\Xi)}{L}$.

Sticking with FOMs, one may apply the strongly convex excessive gap technique [132, Chapter 6.2] to compute the prox-map $X_L(\Xi)$ as well. We will rewrite $Q_L(\Xi; X)$ in a form that is more natural for applying the excessive gap technique [132, Algorithm 6.2.37]. Note that $\nabla_2\, q(\gamma,\Xi) = A(\gamma)\Xi + B(\gamma)$. Thus, defining the matrix $G_{\text{obj}} := A_{\text{obj}}\Xi + B_{\text{obj}}$ and the linear operator $\mathcal{G} : \gamma \mapsto \sum_{i=1}^m \gamma_i(A_i\Xi + B_i)$, we have $\nabla_2\, q(\gamma,\Xi) = A(\gamma)\Xi + B(\gamma) = G_{\text{obj}} + \mathcal{G}\gamma$. Hence, we arrive at

$$Q_L(\Xi; X) = \frac{L}{2}\|X-\Xi\|_F^2 + \left\langle G_{\text{obj}}, X-\Xi\right\rangle + \max_{\gamma\in\mathcal{U}}\{\langle\mathcal{G}\gamma, X\rangle + (q(\gamma,\Xi) - \langle\mathcal{G}\gamma,\Xi\rangle)\}. \tag{6.9}$$

The inner saddle-point function is strongly convex in X and linear in γ so that we may approximate the prox-map by approximately solving a strongly convex–concave saddle point problem. Thus, applying [132, Theorem 6.2.4] to $Q_L(\Xi; X)$ in the form (6.9) gives the following result.

Theorem 30. *Initialize [132, Theorem 6.2.4] with initial iterate $\gamma_0 \in \mathcal{U}$. Let $(\tilde{\gamma}, \tilde{X})$ denote the output of [132, Algorithm 6.2.37] after*

$$O\left(\frac{\max_{\gamma\in\mathbf{S}^{m-1}}\left\|\mathcal{G}\gamma\right\|_F \cdot \max_{\gamma\in\mathcal{U}}\left\|\gamma-\gamma_0\right\|_2}{\sqrt{L\epsilon}}\right)$$

iterations. Here, each iteration may require two exact projections onto $\mathcal{U}$. Then,

$$Q_L(\Xi;\tilde{X}) - Q_L^*(\Xi) \leq Q_L(\Xi;\tilde{X}) - \left(q(\bar{\gamma}_k,\Xi) - \frac{\left\|\nabla_2\, q(\bar{\gamma}_k,\Xi)\right\|_F^2}{2L}\right) \leq \epsilon. \tag{6.10}$$

Remark 75. For simplicity, in our numerical implementation of CertSDP, we opt to run the accelerated gradient descent method for simple sets [132, Algorithm 2.2.63] on the dual problem (6.8). □

6.3.3 Putting the pieces together

We conclude this section by showing how to combine Theorems 29 and 30 to get a guarantee on the total iteration count (including iterations *within* the inexact prox-map calls). To this end, we will need an additional assumption on the norms of various quantities.

Assumption 18. Suppose Algorithm 9 starts at $X_0 = 0_{(n-k)\times k}$. Suppose $R > 0$ satisfies

$$\frac{\left\|\nabla_2\, q(\tilde{\gamma}, X_0)\right\|_F}{L} = \frac{\left\|B(\tilde{\gamma})\right\|_F}{L} \leq R$$

where $\tilde{\gamma} \in \arg\max_{\gamma\in\mathcal{U}} q(\gamma, X_0)$. Let D denote the diameter of $\mathcal{U}$; this is the natural scale parameter for the dual iterates. We will see soon that R is a natural scale parameter for the primal iterates X_t, $\Xi_t \in \mathbb{R}^{(n-k)\times k}$. Suppose $H \geq 1$ bounds

$$\frac{D\left\|\sum_{i=1}^m \gamma_i A_i\right\|_2}{\mu}, \quad \text{and} \quad \frac{D\left\|\sum_{i=1}^m \gamma_i B_i\right\|_F}{\mu\kappa R}$$

for all $\gamma \in \mathbf{S}^{m-1}$. We will assume algorithmic access to D, H, and R. □

Lemma 71. *Under Assumption 18, it holds that $Q(X_0) - \mathrm{Opt}_{(\mathrm{QMMP})} \leq \frac{\mu\kappa^2R^2}{2}$. In particular, we may take $\mathrm{gap}_0 = \frac{\mu\kappa^2R^2}{2}$ in Algorithm 9. Then, for every $t \geq 0$, the iterate Ξ_t computed by Algorithm 9 satisfies*

$$\left\|\Xi_t\right\|_F \leq 10\kappa R.$$

Proof. Let $\tilde{\gamma} \in \arg\max_{\gamma \in \mathcal{U}} q(\gamma, X_0)$. By μ-strong convexity of $Q(X)$, we have that

$$\begin{aligned} Q(X) &\geq q(\tilde{\gamma}, X) \\ &\geq q(\tilde{\gamma}, X_0) + \langle \nabla_2\, q(\tilde{\gamma}, X_0), X - X_0 \rangle + \frac{\mu}{2} \|X - X_0\|_F^2 \\ &= Q(X_0) - \frac{1}{2\mu} \|\nabla_2\, q(\tilde{\gamma}, X_0)\|_F^2 + \frac{\mu}{2} \left\| X - X_0 + \frac{\nabla_2\, q(\tilde{\gamma}, X_0)}{\mu} \right\|_F^2. \end{aligned}$$

In particular, taking $X = \arg\min_{X \in \mathbb{R}^{(n-k)\times k}} Q(X)$ gives

$$Q(X_0) - \mathrm{Opt}_{(\text{QMMP})} \leq \frac{\|\nabla_2\, q(\tilde{\gamma}, X_0)\|_F^2}{2\mu} \leq \frac{\mu\kappa^2 R^2}{2},$$

where the last inequality follows from Assumption 18. This proves the first claim. Next, by Theorem 29, we have that for all $t \geq 0$, that $Q(X_t) - Q(X_0) \leq Q(X_t) - \mathrm{Opt}_{(\text{QMMP})} \leq 2\mu\kappa^2 R^2$ and hence

$$\frac{\mu}{2} \left\| X_t - X_0 + \frac{\nabla_2\, q(\tilde{\gamma}, X_0)}{\mu} \right\|_F^2 \leq Q(X_t) - Q(X_0) + \frac{\|\nabla_2\, q(\tilde{\gamma}, X_0)\|_F^2}{2\mu} \leq \frac{5\mu\kappa^2 R^2}{2}.$$

Using the assumption $X_0 = 0_{(n-k)\times k}$ in Assumption 18 and applying triangle inequality together with the bound $\|\nabla_2\, q(\tilde{\gamma}, X_0)\|_F^2 \leq \mu^2\kappa^2 R^2$ derived from Assumption 18, we deduce that for all $t \geq 0$,

$$\|X_t\|_F \leq \left(1 + \sqrt{5}\right)\kappa R.$$

Then, as $\Xi_{t+1} = X_{t+1} + \frac{1-\alpha}{1+\alpha}(X_{t+1} - X_t)$, we have

$$\|\Xi_{t+1}\|_F \leq 3\left(1 + \sqrt{5}\right)\kappa R \leq 10\kappa R. \qquad \blacksquare$$

With this bound on $\|\Xi_t\|_F$ we are now able to bound the operator norm $\max_{\gamma \in \mathbf{S}^{m-1}} \|\mathcal{G}\gamma\|_F$ in Theorem 30.

Lemma 72. *Suppose Assumption 18 holds and we set* $\mathrm{gap}_0 = \frac{\mu\kappa^2 R^2}{2}$ *in Algorithm 9. Then, for every iterate* $t \geq 0$*, we have*

$$\max_{\gamma \in \mathbf{S}^{m-1}} \|\mathcal{G}\gamma\|_F \leq 11 \frac{\mu\kappa H R}{D}.$$

Proof. Recall that by definition, the linear operator $\mathcal{G}$ maps γ to $\sum_{i=1}^m \gamma_i(A_i\Xi_t + B_i)$. Thus, for any $\gamma \in \mathbf{S}^{m-1}$,

$$\begin{aligned} \|\mathcal{G}\gamma\|_F &= \left\| \sum_{i=1}^m \gamma_i(A_i\Xi_t + B_i) \right\|_F \\ &\leq \left\| \sum_{i=1}^m \gamma_i A_i \right\|_2 \|\Xi_t\|_F + \left\| \sum_{i=1}^m \gamma_i B_i \right\|_F \\ &\leq 11\frac{\mu\kappa HR}{D}. \end{aligned}$$ ■

The following theorem gives the iteration complexity of Algorithm 9 instantiated with the excessive gap technique to compute the prox-map. It follows as a corollary to Theorems 29 and 30 and Lemma 72.

Theorem 31. *Let $q(\gamma, X)$ and $\mathcal{U}$ satisfy Assumptions 17 and 18. Suppose* gap_0 *is set to $\frac{\mu\kappa^2R^2}{2}$ in Algorithm 9. Let X_t denote the iterates produced by Algorithm 9 with starting point $X_0 = 0_{(n-k)\times k}$. Then, for all $t \geq 1$, the iterate X_t satisfies*

$$Q(X_t) - \mathrm{Opt}_{\text{(QMMP)}} \leq \left(1 - \frac{\alpha}{2}\right)^t \left(2\mu\kappa^2R^2\right).$$

In particular, $Q(X_T) - \mathrm{Opt}_{\text{(QMMP)}} \leq \epsilon$ after at most

$$T = O\left(\sqrt{\kappa}\log\left(\frac{\mu\kappa^2R^2}{\epsilon}\right)\right)$$

outer iterations of Algorithm 9. The iterate X_T is computed after a total (including iterations within the inexact prox-map computations) of $O\left(\frac{\kappa^{5/4}HR\sqrt{L}}{\sqrt{\epsilon}}\right)$ iterations.

Proof. We will take T to be the first positive integer such that

$$\left(1 - \frac{\alpha}{2}\right)^T \left(2\mu\kappa^2R^2\right) \leq \epsilon.$$

Clearly, $T = O(\sqrt{\kappa}\log(\kappa LR^2/\epsilon))$. Next, if $T > 1$, then by the maximality of T we have

$$\left(1 - \frac{\alpha}{2}\right)^T \geq \left(1 - \frac{\alpha}{2}\right)\left(\frac{\epsilon}{2\mu\kappa^2R^2}\right) \geq \frac{\epsilon}{4\mu\kappa^2R^2}.$$

From (6.7) and $\mathrm{gap}_0 = \frac{\mu\kappa^2R^2}{2}$, we deduce that $\epsilon_t \geq \frac{\mu\kappa R^2}{2}(1 - \frac{\alpha}{2})^t\frac{\alpha}{2}$. By Lemma 72 and Assumption 18, we may bound

$$\max_{\gamma\in\mathbf{S}^{m-1}}\|\mathcal{G}\gamma\|_F \cdot \max_{\gamma\in\mathcal{U}}\|\gamma - \gamma_0\|_2 \leq 11\mu\kappa HR.$$

Thus, X_t can be computed in

$$O\left(\frac{\mu\kappa H R}{\sqrt{L\epsilon_t}}\right) = O\left(\kappa^{1/4} H (1-\alpha/2)^{-t/2}\right)$$

iterations. Summing over the first T outer iterations and observing our lower bound on $(1-\frac{\alpha}{2})^T$, we have that

$$\sum_{t=0}^{T}\left(1-\frac{\alpha}{2}\right)^{-t/2} \leq O\left(\frac{(1-\frac{\alpha}{2})^{-T/2}}{\alpha}\right) = O\left(\frac{\kappa R\sqrt{L}}{\sqrt{\epsilon}}\right). \qquad \blacksquare$$

6.4 Solving k-exact SDPs via strongly convex QMMP algorithms

In this section, we show how to combine Theorems 27 and 31 to develop first-order methods for approximately solving rank-k exact QMP-like SDPs. We will use the following notion of an approximate solution to (SDP).

Definition 25. We will say that $\tilde{Y} \in \mathbb{S}^n$ is *ϵ-optimal and δ-feasible* for (SDP) if

$$\left\langle M_{\text{obj}}, \tilde{Y} \right\rangle \leq \text{Opt}_{(\text{SDP})} + \epsilon, \qquad \left(\sum_{i=1}^{m}\left(\left\langle M_i, \tilde{Y}\right\rangle + d_i\right)^2\right)^{1/2} \leq \delta,$$

and $\tilde{Y} \succeq 0$. □

The final piece towards this goal is developing algorithms for constructing a certificate of strict complementarity $\mathcal{U}$.

By Definition 23, the properties we need to ensure for $\mathcal{U}$ are that $\gamma^* \in \mathcal{U}$ and $A(\gamma) \succ 0$ for all $\gamma \in \mathcal{U}$. We will construct $\mathcal{U}$ by taking it to be an ℓ_2-ball centered at a sufficiently accurate estimate $\tilde{\gamma}$ of γ^*.

We begin by verifying that $A(\gamma^*) \succ 0$.

Lemma 73. *Suppose M^*, $Y^* \in \mathbb{S}^n_+$ have rank $n-k$ and k respectively and that $\langle M^*, Y^*\rangle = 0$. Let W be an $(n-k)$-dimensional subspace. Then, $M^*_W \succ 0$ if and only if $Y^*_{W^\perp} \succ 0$.*

Proof. It suffices to prove the forward direction as we may interchange the roles of Y^* and M^*.

We prove the contrapositive. Suppose $Y^*_{W^\perp} \not\succ 0$ so that $\ker(Y^*_{W^\perp})$ is nontrivial. As $Y^* \succeq 0$, we have that in fact $\ker(Y^*) \cap W^\perp$ is nontrivial. Then, $\text{range}(Y^*)$ is a k-dimensional subspace contained in $(\ker(Y^*) \cap W^\perp)^\perp$. Similarly, W is an $(n-k)$-dimensional subspace contained in $(\ker(Y^*) \cap W^\perp)^\perp$. Then, as $(\ker(Y^*) \cap W^\perp)^\perp$ has dimension at most $n-1$, we deduce that $\text{range}(Y^*) \cap W$ is nontrivial and $\langle Y^*, M^*\rangle > 0$, a contradiction. ■

Clearly then, for all $\tilde{\gamma}$ close enough to γ^*, we have that $A(\tilde{\gamma}) \succ 0$ and there exists some $r > 0$ such that $\tilde{\mathcal{U}} := \mathbb{B}(\tilde{\gamma}, r)$ satisfies $A(\gamma) \succ 0$ for all $\gamma \in \tilde{\mathcal{U}}$. We consider one setting for r below. It remains to ask, does the condition that $\gamma^* \in \tilde{\mathcal{U}}$ hold? Below, we show that this condition

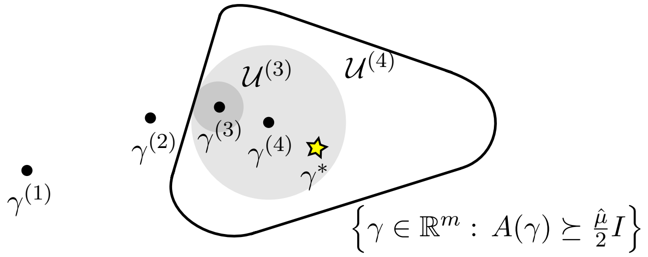

Figure 6.1: CertSDP (Algorithm 10) produces a series of iterates $\gamma^{(i)} \to \gamma^*$. For each $\gamma^{(i)}$, CertSDP constructs a ball $\mathcal{U}^{(i)}$ around $\gamma^{(i)}$. Intuitively, we want to pick $\mathcal{U}^{(i)}$ to be the largest ball around $\gamma^{(i)}$ for which we can solve the associated QMMP efficiently, in hopes of enclosing γ^*. We will thus choose $\mathcal{U}^{(i)}$ to satisfy certain regularity estimates (see (6.11) and Lemma 75). At the minimum, we will ensure $A(\gamma) \succeq \hat{\mu}/2$ for all $\gamma \in \mathcal{U}^{(i)}$.

indeed holds when $\tilde{\gamma}$ is a sufficiently accurate estimate of γ^* and that we can effectively check this condition using CautiousAGD.

Assumption 19. Suppose we have algorithmic access to

- parameters $0 < \hat{\mu} \leq \hat{L}$ such that $\hat{\mu} I \preceq A(\gamma^*) \preceq \hat{L} I$,
- parameters $\hat{R}_p,\ \hat{R}_d > 0$ such that $\|X^*\|_F \leq \hat{R}_p$ and $\|\gamma^*\|_2 \leq \hat{R}_d$, and
- a parameter $\hat{\rho} > 0$ upper bounding

$$\frac{\hat{\mu}}{\hat{R}_d}, \qquad \left\|\sum_{i=1}^m \gamma_i A_i\right\|_2, \qquad \text{and} \qquad \frac{\|\sum_{i=1}^m \gamma_i B_i\|_F}{\hat{R}_p} \qquad \forall \gamma \in \mathbf{S}^{m-1}.$$

For notational simplicity, we will additionally assume $\hat{R}_d \geq 1$. This is not strictly necessary and simply allows us to write $O(\hat{R}_d)$ in place of $O(1 + \hat{R}_d)$. □

Note from the identity $X^* = -A(\gamma^*)^{-1} B(\gamma^*)$ that $\|B(\gamma^*)\|_F \leq \hat{L}\hat{R}_p$. Now, suppose $\gamma^{(1)}, \gamma^{(2)}, \ldots$ is a sequence converging to γ^* (such a sequence can be constructed via subgradient methods [108]; see also [60, Section 6.2.2]). Given $\gamma^{(i)}$, define

$$\begin{aligned} r^{(i)} := \min\Bigg(&\frac{\hat{\mu}}{2\hat{\rho}},\ 2\hat{R}_d - \left\|\gamma^{(i)}\right\|_2,\ \frac{\lambda_{\min}\left(A\left(\gamma^{(i)}\right)\right) - \hat{\mu}/2}{\hat{\rho}}, \\ &\frac{2\hat{L} - \lambda_{\max}\left(A\left(\gamma^{(i)}\right)\right)}{\hat{\rho}},\ \frac{2\hat{L}\hat{R}_p - \left\|B\left(\gamma^{(i)}\right)\right\|_F}{\hat{\rho}\hat{R}_p}\Bigg). \end{aligned} \tag{6.11}$$

If $r^{(i)}$ is positive, define $\mathcal{U}^{(i)} := \mathbb{B}(\gamma^{(i)}, r^{(i)})$.

We present three lemmas below. The first lemma states that $r^{(i)}$ is positive and $\gamma^* \in \mathcal{U}^{(i)}$ for all $\gamma^{(i)}$ sufficiently close to γ^*. The second lemma establishes parameters for which the regularity conditions of Assumption 18 hold for $q(\gamma, X)$ along with $\mathcal{U}^{(i)}$. Finally, the third lemma shows that for each $\mathcal{U}^{(i)}$, an approximate solution of the corresponding strongly convex QMMP (which can be computed using Algorithm 9) can be used to either produce an approximate optimizer of the underlying SDP or declare $\gamma^* \notin \mathcal{U}^{(i)}$.

Lemma 74. *Suppose Assumption 19 holds. Then, $r^{(i)}$ is positive and $\gamma^* \in \mathcal{U}^{(i)}$ if*

$$\left\|\gamma^{(i)} - \gamma^*\right\|_2 \leq \frac{\hat{\mu}}{4\hat{\rho}}.$$

Proof. Let $r := \left\|\gamma^{(i)} - \gamma^*\right\|_2$. Using Assumption 19, we may bound the individual terms within the definition of $r^{(i)}$ as

$$\begin{aligned}
2\hat{R}_d - \left\|\gamma^{(i)}\right\|_2 &\geq \hat{R}_d - r \geq \frac{\hat{\mu}}{\hat{\rho}} - r,\\
\frac{\lambda_{\min}\left(A\left(\gamma^{(i)}\right)\right) - \hat{\mu}/2}{\hat{\rho}} &\geq \frac{\hat{\mu}/2 - \hat{\rho}r}{\hat{\rho}} = \frac{\hat{\mu}}{2\hat{\rho}} - r,\\
\frac{2\hat{L} - \lambda_{\max}\left(A\left(\gamma^{(i)}\right)\right)}{\hat{\rho}} &\geq \frac{\hat{L} - \hat{\rho}r}{\hat{\rho}} = \frac{\hat{L}}{\hat{\rho}} - r, \text{ and}\\
\frac{2\hat{L}\hat{R}_p - \left\|B\left(\gamma^{(i)}\right)\right\|_F}{\hat{\rho}\hat{R}_p} &\geq \frac{\hat{L} - \hat{\rho}r}{\hat{\rho}} = \frac{\hat{L}}{\hat{\rho}} - r.
\end{aligned}$$

Thus, $r^{(i)} \geq \min\left(\frac{\hat{\mu}}{2\hat{\rho}}, \frac{\hat{\mu}}{2\hat{\rho}} - r\right) = \frac{\hat{\mu}}{2\hat{\rho}} - r$. Then, when $r \leq \frac{\hat{\mu}}{4\hat{\rho}}$, we have $r^{(i)} > 0$ and furthermore, $r^{(i)} \geq r = \left\|\gamma^{(i)} - \gamma^*\right\|_2$. ■

Lemma 75. *Suppose Assumption 19 holds and $r^{(i)}$ is positive. Then, $q(\gamma, X)$ and $\mathcal{U}^{(i)}$ satisfy Assumption 18 with $\mu = \frac{\hat{\mu}}{2}$, $L = 2\hat{L}$, $R = \hat{R}_p$, $D = 2r^{(i)}$, and $H = 2$.*

Proof. Begin by noting that for all $\gamma \in \mathcal{U}^{(i)}$,

$$\frac{\hat{\mu}}{2}I \preceq A\left(\gamma^{(i)}\right) - r^{(i)}\hat{\rho}I \preceq A(\gamma) \preceq A\left(\gamma^{(i)}\right) + r^{(i)}\hat{\rho}I \preceq 2\hat{L}I.$$

Let $\tilde{\gamma} \in \arg\max_{\gamma\in\mathcal{U}^{(i)}} q(\gamma, 0_{(n-k)\times k})$. Then,

$$\|B(\tilde{\gamma})\|_F \leq \left\|B\left(\gamma^{(i)}\right)\right\|_F + \hat{\rho}r^{(i)}\hat{R}_p \leq 2\hat{L}\hat{R}_p = LR.$$

Next, for $\gamma \in \mathbf{S}^{m-1}$

$$\begin{aligned}
\frac{D\|\sum_{i=1}^m \gamma_i A_i\|_2}{\mu} &\leq \frac{4r^{(i)}\hat{\rho}}{\hat{\mu}} \leq 2\\
\frac{D\|\sum_{i=1}^m \gamma_i B_i\|_F}{LR} &\leq \frac{r^{(i)}\hat{\rho}}{\hat{L}} \leq 1/2.
\end{aligned}$$ ■

Lemma 76. *Suppose Assumption 19 holds, $r^{(i)}$ is positive, and $0 < \epsilon \leq 9\hat{\rho}\hat{R}_d\hat{R}_p^2$. Set $\delta := \frac{\hat{\mu}\epsilon^2}{\left(9\hat{\rho}\hat{R}_d\hat{R}_p\right)^2}$ and $\eta := \frac{4\epsilon}{9\hat{R}_d}$. Suppose $\tilde{X} \in \mathbb{R}^{(n-k)\times k}$ satisfies*

$$Q_{\mathcal{U}^{(i)}}\left(\tilde{X}\right) \leq \min_{X\in\mathbb{R}^{(n-k)\times k}} Q_{\mathcal{U}^{(i)}}(X) + \delta.$$

Then,

- *If $\gamma^* \in \mathcal{U}^{(i)}$, then $Y(\tilde{X})$ is η-feasible.*
- *If $Y(\tilde{X})$ is η-feasible, then $Y(\tilde{X})$ is ϵ-optimal and ϵ-feasible.*

Proof. Suppose $\gamma^* \in \mathcal{U}^{(i)}$ and define $\Delta := \tilde{X} - X^*$. By strong convexity and Theorem 27, we have that $\frac{\hat{\mu}}{2}\|\Delta\|_F^2 \leq \delta$. Next, recalling that $q_i(X^*) = 0$ for all $i \in [m]$, we deduce

$$\begin{aligned}\left(\sum_{i=1}^m\left(\left\langle M_i, Y(\tilde{X})\right\rangle + d_i\right)^2\right)^{1/2} &= \left(\sum_{i=1}^m q_i(\tilde{X})^2\right)^{1/2}\\ &= \max_{\|\gamma\|_2=1}\sum_{i=1}^m \gamma_i\left(\frac{\operatorname{tr}(\Delta^\intercal A_i\Delta)}{2} + \langle A_iX^* + B_i, \Delta\rangle\right)\\ &\leq \hat{\rho}\left(\frac{\delta}{\hat{\mu}}\right) + \sqrt{8}\hat{\rho}\hat{R}_p\sqrt{\frac{\delta}{\hat{\mu}}}\\ &\leq \frac{4\hat{\rho}\hat{R}_p}{\sqrt{\hat{\mu}}}\sqrt{\delta} = \eta.\end{aligned}$$

Here, the first inequality follows from $\|\Delta\|_F^2 \leq \frac{2\delta}{\hat{\mu}}$ and Assumption 19, and the last inequality follows as $\delta = \frac{\hat{\mu}\epsilon^2}{\left(9\hat{\rho}\hat{R}_d\hat{R}_p\right)^2} \leq \hat{\mu}\hat{R}_p^2$ since $0 < \epsilon \leq 9\hat{\rho}\hat{R}_d\hat{R}_p^2$.

Now, suppose $Y(\tilde{X})$ is η-feasible. Note that $\eta \leq \epsilon$ (as $\hat{R}_d \geq 1$) and thus $Y(\tilde{X})$ is immediately ϵ-feasible. Let $\tilde{\gamma} \in \arg\max_{\gamma\in\mathcal{U}^{(i)}} q(\gamma, \tilde{X})$ so that $Q_{\mathcal{U}^{(i)}}(\tilde{X}) = q(\tilde{\gamma}, \tilde{X})$. Then,

$$\begin{aligned}\left\langle M_{\text{obj}}, Y(\tilde{X})\right\rangle = q_{\text{obj}}(\tilde{X}) = q(\tilde{\gamma}, \tilde{X}) &- \sum_{i=1}^m \tilde{\gamma}_i q_i(\tilde{X})\\ &\leq Q_{\mathcal{U}^{(i)}}(\tilde{X}) + \|\tilde{\gamma}\|_2\eta\\ &\leq \left(\min_{X\in\mathbb{R}^{(n-k)\times k}} Q_{\mathcal{U}^{(i)}}(X) + \delta\right) + 2\hat{R}_d\eta\\ &\leq \operatorname{Opt}_{(\text{SDP})} + \left(\delta + 2\hat{R}_d\eta\right),\end{aligned}$$

where the first inequality follows from the η-feasibility of $Y(\tilde{X})$, the second inequality from the premise of the lemma on $\tilde{X}$ and the fact that $\|\tilde{\gamma}\|_2 \leq \left\|\tilde{\gamma} - \gamma^{(i)}\right\|_2 + \left\|\gamma^{(i)}\right\|_2 \leq 2\hat{R}_d$ (this holds because $\tilde{\gamma} \in \mathcal{U}^{(i)}, \mathcal{U}^{(i)}$ is the ℓ_2-ball of radius $r^{(i)}$ centered at $\gamma^{(i)}$, and by definition of $r^{(i)}$ we have $r^{(i)} \leq 2\hat{R}_d - \left\|\gamma^{(i)}\right\|_2$). We may then use the definitions of δ and η to bound

$$\delta + 2\hat{R}_d\eta = \frac{\hat{\mu}\epsilon^2}{\left(9\hat{\rho}\hat{R}_d\hat{R}_p\right)^2} + \frac{8\epsilon}{9} \leq \frac{\hat{\mu}\epsilon}{9\hat{\rho}\hat{R}_d} + \frac{8\epsilon}{9} \leq \epsilon.$$

Here, the first inequality follows from the upper bound on ϵ and the second inequality follows from $\frac{\hat{\mu}}{\hat{R}_d} \leq \hat{\rho}$ (Assumption 19). This then shows that $Y(\tilde{X})$ is ϵ-optimal. ■

We are now ready to present our full algorithm for computing approximate solutions to (SDP). CertSDP (Algorithm 10) assumes access to a sequence $\gamma^{(i)} \to \gamma^*$ and applies a guess-and-double scheme to guess when $\left\|\gamma^{(i)} - \gamma^*\right\|_2$ is sufficiently small. It then applies Algorithm 9 to either compute an ϵ-optimal ϵ-feasible solution $Y(\tilde{X})$ or to declare that $\gamma^* \notin \mathcal{U}^{(i)}$.

Algorithm 10 CertSDP

```
Given a rank-$k$ exact QMP-like SDP satisfying Assumption 19, a sequence $\gamma^{(1)}, \gamma^{(2)}, \dots \to \gamma^*$,
and $0 < \epsilon \leq 9\hat{\rho}\hat{R}_d\hat{R}_p^2$
   1. Set $\delta$ and $\eta$ as in Lemma 76
   2. For each $i = 2^0, 2^1, 2^2, \dots$
        • If $r^{(i)} > 0$
             a) Let $\mathcal{U}^{(i)} := \mathbb{B}(\gamma^{(i)}, r^{(i)})$ and compute $\tilde{X}$ satisfying

                  $Q_{\mathcal{U}^{(i)}}(\tilde{X}) \leq \min_{X \in \mathbb{R}^{(n-k)\times k}} Q_{\mathcal{U}^{(i)}}(X) + \delta$

                using Algorithm 9
             b) If $Y(\tilde{X})$ is $\eta$-feasible, output $Y(\tilde{X})$
```

The next theorem gives rigorous guarantees on CertSDP and follows from Lemmas 74 to 76 and Theorem 31.

Theorem 32. *Suppose* (SDP) *is a rank-k exact QMP-like SDP satisfying Assumption 19,* $\gamma^{(1)}, \gamma^{(2)}, \dots \to \gamma^*$ *and* $0 < \epsilon \leq 9\hat{\rho}\hat{R}_d\hat{R}_p^2$*. Let* T *be such that* $\left\|\gamma^{(t)} - \gamma^*\right\|_2 \leq \frac{\hat{\mu}}{4\hat{\rho}}$ *for all* $t \geq T$*. Then, CertSDP (Algorithm 10) accesses at most* $2T$ *iterates of the sequence* $\gamma^{(i)}$ *and outputs an* ϵ*-optimal and* ϵ*-feasible solution in*

$$O\left(\sqrt{\hat{\kappa}}\log\left(\frac{\hat{\kappa}\hat{\rho}\hat{R}_p\hat{R}_d}{\epsilon}\right)\cdot\log(T)\right) \textit{ prox-map calls, and}$$

$$O\left(\frac{\hat{\kappa}^{7/4}\hat{\rho}\hat{R}_p^2\hat{R}_d}{\epsilon}\cdot\log(T)\right) \textit{ iterations within all prox-map calls.}$$

6.5 Numerical experiments

In this section, we investigate the numerical performance of our new FOM, CertSDP, on rank-k exact QMP-like SDPs that are both large and sparse. Specifically, we consider random instances of distance-minimization QMPs and their primal and dual SDP relaxations of the form

$$\begin{aligned}
&\inf_{X\in\mathbb{R}^{(n-k)\times k}} \left\{ \frac{\|X\|_F^2}{2} :\; \operatorname{tr}\left(\frac{X^\intercal A_i X}{2}\right) + \langle B_i, X\rangle + c_i = 0,\ \forall i\in[m] \right\} \qquad (6.12)\\
&\geq \inf_{Y\in\mathbb{S}^n} \left\{ \left\langle \begin{pmatrix} I_{n-k}/2 & \\ & 0_k \end{pmatrix}, Y \right\rangle :\; \begin{array}{l} \left\langle \begin{pmatrix} A_i/2 & B_i/2 \\ B_i^\intercal/2 & \frac{c_i}{k} I_k \end{pmatrix}, Y \right\rangle = 0,\ \forall i\in[m] \\ Y = \begin{pmatrix} * & * \\ * & I_k \end{pmatrix} \succeq 0 \end{array} \right\}\\
&\geq \sup_{\gamma\in\mathbb{R}^m,\, T\in\mathbb{S}^k} \left\{ \operatorname{tr}(T) : \begin{pmatrix} A(\gamma)/2 & B(\gamma)/2 \\ B(\gamma)^\intercal/2 & \frac{c(\gamma)}{k} I_k - T \end{pmatrix} \succeq 0 \right\}.
\end{aligned}$$

In our instance generation procedure, we ensure that equality holds throughout this chain of inequalities.

We will compare the performance of CertSDP on instances of (6.12) to that of several first-order methods from the literature: the complementary slackness SDP algorithm (CSSDP) [60], ProxSDP [164], and the splitting cone solver (SCS) [136]. We discuss these algorithms and relevant implementation details in Section 6.5.1 and the instance generation procedure in Section 6.5.2 before presenting the numerical results in Section 6.5.3.

All algorithms and experiments are implemented in Julia and run on a machine with an AMD Opteron 4184 processor with 12 CPUs and 70GB of RAM. Our code is publicly available at:

https://github.com/alexlihengwang/CertSDP

We additionally implement a variant of SketchyCGAL [198] adapted to our setting (see Section 6.5.1). We choose to leave this algorithm off of our large-format experiments as its performance was very similar to that of CSSDP in preliminary experimentation (see Section F.2).

6.5.1 Implementation details

CertSDP. We implemented CertSDP (Algorithm 10) as presented in this chapter except a few modifications. In addition to simplifying the overall algorithm, these modifications enable CertSDP to be run *without* knowledge of the parameters $\hat\mu$ and $\hat L$. While the convergence guarantees of Theorem 32 may no longer hold, we find empirically that CertSDP continues to perform very effectively with these modifications.

- We instantiate CertSDP with Accelegrad [108] as the iterative method for producing iterates $\gamma^{(i)}$. As in [60], we apply Accelegrad to the penalized dual problem

$$\max_{\gamma\in\mathbb{R}^m,\, T\in\mathbb{S}^k} \operatorname{tr}(T) + \text{penalty}\cdot\min\left(0,\ \lambda_{\min}\begin{pmatrix} A(\gamma)/2 & B(\gamma)/2 \\ B(\gamma)^\intercal/2 & \frac{c(\gamma)}{k} I_k - T \end{pmatrix}\right)$$

for some large value for the penalty parameter. It can be shown that the optimal value and optimizers of this penalized dual problem coincide with that of the dual SDP whenever the penalty parameter is larger than $\operatorname{tr}(Y^*)$; see [60]. In our experiments, we set the penalty parameter to be $20 \cdot \operatorname{tr}(Y^*)$.

- In practice, it is extremely cheap to solve ($\text{QMMP}_{\mathcal{U}}$) even to high accuracy. Thus, we replace the guess-and-double scheme in Algorithm 10 with a linear schedule, i.e., we solve ($\text{QMMP}_{\mathcal{U}}$) once every ≈ 250 iterations. Additionally, we replace the excessive gap technique used in Theorem 31 with accelerated gradient descent (see Remark 75).

- We set

$$r^{(i)} = \frac{1}{\hat{\rho}} \cdot \frac{\lambda_{\max}(A(\gamma^{(i)}))\lambda_{\min}(A(\gamma^{(i)}))}{\lambda_{\max}(A(\gamma^{(i)})) + \lambda_{\min}(A(\gamma^{(i)}))}$$

  if $A(\gamma^{(i)})) \succ 0$, and $r^{(i)} = 0$ else. Equivalently, $\mathcal{U}^{(i)} := \mathbb{B}(\gamma^{(i)}, r^{(i)})$ is the largest ball centered at $\gamma^{(i)}$ for which the condition number of $A(\gamma)$ for any $\gamma \in \mathbb{B}(\gamma^{(i)}, r^{(i)})$ is guaranteed to be at most twice the condition number of $A(\gamma^{(i)})$. Note that it still holds that $A(\gamma) \succ 0$ for all $\gamma \in \mathcal{U}^{(i)}$ (as long as $r^{(i)}$ is positive) and that $\gamma^* \in \mathcal{U}^{(i)}$ for all $\gamma^{(i)}$ close enough to γ^*.

- In CautiousAGD (Algorithm 9), we terminate early if $\max_{i\in[m]} |q_i(X_t)|$ does not decrease to zero geometrically. Indeed, this can only happen if $\gamma^* \notin \mathcal{U}^{(i)}$.

- Theorem 30 gives an *a priori* guarantee on the number of inner iterations required for solving each prox-map. Instead of using this number of iterations, in our code, we will monitor the saddle point gap, i.e., the second term in (6.10), and break as soon as the saddle point gap is small enough.

- We warm-start the iterate X in CautiousAGD using the last iterate of the previous run of CautiousAGD and warm-start γ in the prox-map computation using the last iterate of the previous run of the prox-map computation.

- Unless the time limit is met first, the overall algorithm is terminated once CautiousAGD produces a (10^{-13})-optimal solution of ($\text{QMMP}_{\mathcal{U}}$) that satisfies $\max_{i\in[m]} |q_i(X_t)| \leq 10^{-13}$.

CSSDP. The complementary slackness SDP algorithm (CSSDP) [60] similarly constructs a sequence of iterates $\gamma^{(i)} \to \gamma^*$ and occasionally solves a compressed k-dimensional SDP [60, MinFeasSDP] in the vector space corresponding to the k-many minimum eigenvalues of the slack matrix $M(\gamma^{(i)})$. As in our implementation of CertSDP, we instantiate CSSDP with Accelegrad [108] as the iterative method for producing iterates $\gamma^{(i)}$ and solve the compressed SDP once every ≈ 250 iterations. The compressed SDPs are solved using SCS solver with all error parameters set to 10^{-13}. Since CSSDP needs to solve the compressed SDP frequently, we make sure to instantiate the optimization problem just once in order to amortize the cost of allocating the $k \times k$ symmetric matrix variable.

SketchyCGAL. Yurtsever et al. [198] observe that one may track any *linear image* of the primal matrix iterates (as opposed to the matrix iterate itself) in the CGAL [197] algorithm. Combining this observation with the Nyström sketch gives SketchyCGAL. In our code, we implement a variant of this idea, where we replace the Nyström sketch with the linear map sending a matrix in $\mathbb{S}^n$ to its top-right $(n-k)\times k$ submatrix. We omit this algorithm in our large-format experiments as its performance was very similar to that of CSSDP in preliminary experiments (see Section F.2).

ProxSDP and SCS. ProxSDP [164] and the splitting cone solver (SCS) [136] are first-order methods that can be used to tackle large-scale SDPs. ProxSDP combines the primal-dual hybrid gradient method with an approximate projection operation that allows it to replace a full eigendecomposition with a partial one whenever the rank of the true SDP solution is small. SCS employs a first-order method to tackle the homogeneous self-dual embedding but does not explicitly take advantage of possible low rank solutions.

In our experiments, we pass the SDP relaxations of our QMPs to the corresponding Julia interfaces `ProxSDP.jl` and `SCS.jl` with all error parameters set to 10^{-13}. In contrast to CertSDP and CSSDP, which achieve storage optimality, ProxSDP and SCS both store matrix iterates and thus require substantially more memory.

6.5.2 Random instance generation

We generate random sparse instances of distance-minimization QMPs (6.12) as follows: Let $(n, k, m, \mu^*, \mathsf{nnz})$ be input parameters. Here, (n, k, m) control the size of (6.12), μ^* is the desired value of $\lambda_{\min}(A(\gamma^*))$ and nnz approximately controls the number of nonzero entries in each $A_1, \dots, A_m$.

- Let $A_1, \dots, A_m \in \mathbb{S}^{n-k}$ be sparse symmetric matrices each with $\approx \mathsf{nnz}$ nonzero entries that are i.i.d. normal. We scale $A_1, \dots, A_m$ such that $\|A_i\|_2 = 1$ for all $i \in [m]$.
- Let $B_1, \dots, B_m \in \mathbb{R}^{(n-k)\times k}$ be matrices where all entries are i.i.d. normal. We scale $B_1, \dots, B_m$ such that $\|B_i\|_F = 1$ for all $i \in [m]$.
- Pick a direction $\hat{\gamma}$ uniformly from the surface of the sphere $\mathbf{S}^{m-1}$, then set $\gamma^* := r\hat{\gamma}$ where $r > 0$ solves $\lambda_{\min}(A(\gamma^*)) = 1 + r\lambda_{\min}(\sum_{i=1}^m \hat{\gamma}_i A_i) = \mu^*$. Let $X^* := -A(\gamma^*)^{-1}B(\gamma^*)$.
- Finally, for each $i \in [m]$, set c_i such that $\operatorname{tr}\left(\frac{X^{*\intercal}A_iX^*}{2}\right) + \langle B_i, X^*\rangle + c_i = 0$.

Exactness is guaranteed to hold throughout (6.12) as $(\gamma^*,\ T^*)$, where

$$T^* := \frac{c(\gamma^*)}{k} I_k - \frac{B(\gamma^*)^\intercal A(\gamma^*)^{-1}B(\gamma^*)}{2},$$

achieves the value $\frac{\|X^*\|_F^2}{2}$ in the third line of (6.12) (see Section F.1).

6.5.3 Numerical results

To investigate the scalability of CertSDP in terms of n, we fix $k = 10$, $m = 10$, $\mu^* = 0.1$ and $\mathsf{nnz} = n$. Note that in this regime, the A_i matrices are each individually very sparse with approximately one nonzero entry per row or column. We then vary n such that the height of the matrix variable $X \in \mathbb{R}^{(n-k)\times k}$, i.e., $n-k$, takes the values 10^3, 10^4, 10^5. For each value of $n-k$, we generate 10 random instances of (6.12) according to Section 6.5.2 and measure the time, error, and memory consumption of the tested algorithms.

Remark 76. We measure the memory consumption of each algorithm by monitoring the virtual memory size (`vsz`) of the process throughout the run of the algorithm and report the difference between the maximum value and the starting value. This is the same measurement that is performed in [198]. We caution that this number should only be treated as a *very rough* estimate of the storage requirements. Indeed, virtual memory need not be allocated at all for small enough programs (so that some algorithms register as using no memory at all for small enough values of $n-k$) and furthermore, when it is allocated, it is not always fully used. Experimentally, we found that on our machine, storage of up to ≈ 1.0 MB was often measured as not using any memory at all. We report such measurements as 0.0 MB in our tables (Tables 6.1 to 6.3) and as 1.0 MB in our log-scale plot Figure 6.3. □

We ran each algorithm with time limits of 3×10^3, 10^4, and 5×10^4 seconds for $n-k=10^3$, 10^4, 10^5 respectively. SCS is not tested for $n-k=10^4$ as it was unable to complete a single iteration within the time limits and utilized over 70GB of memory. Similarly, ProxSDP and SCS were not tested for $n-k=10^5$ as both came to complete failures due to excessive memory allocation.

Detailed numerical results are reported in Tables 6.1 to 6.3 for $n-k=10^3$, 10^4, and 10^5 respectively. Additional plots show the time and accuracy of each algorithm (Figure 6.2), the average memory usage of each algorithm (Figure 6.3), and the convergence *behavior* of CertSDP versus CSSDP on a single instance of each size (Figure 6.4). The plots on the left of Figure 6.4 show the primal squared distance $\|X-X^*\|_F^2$ and the dual suboptimality

$$\mathrm{Opt}_{(6.12)} - \left(\mathrm{tr}(T) + \mathrm{penalty}\cdot\min\left(0,\ \lambda_{\min}\begin{pmatrix} A(\gamma)/2 & B(\gamma)/2 \\ B(\gamma)^\intercal/2 & \frac{c(\gamma)}{k}I_k - T\end{pmatrix}\right)\right)$$

for the iterates produced by CertSDP and CSSDP as a function of time. The plots on the right of Figure 6.4 show the primal squared distance for the iterates produced by CertSDP within the final call to CautiousAGD.

Algorithm	time (s)	std.	$\|X-X^*\|_F^2$	std.	memory (MB)	std.
CertSDP	1.3×10^3	7.6×10^2	1.9×10^{-22}	4.2×10^{-23}	0.0	0.0
CSSDP	3.0×10^3	5.8×10^{-1}	7.3×10^{-2}	3.4×10^{-2}	0.0	0.0
ProxSDP	2.1×10^2	1.1×10^1	1.2×10^{-19}	3.2×10^{-19}	4.8×10^1	1.9×10^1
SCS	3.1×10^3	2.5×10^1	5.1×10^{-5}	9.5×10^{-5}	5.3×10^2	4.3×10^1

Table 6.1: Experimental results for $(n-k)=10^3$ (10 instances) with time limit 3×10^3 seconds.

We make a few observations:

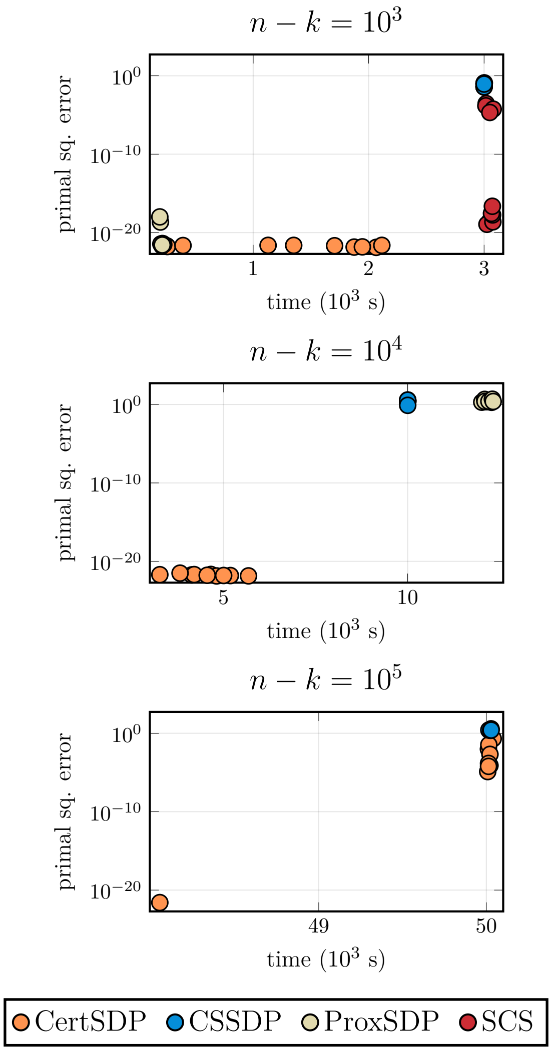

Figure 6.2: Convergence plots comparing CertSDP, CSSDP, ProxSDP, and SCS for $n-k = 10^3,\ 10^4,\ 10^5$. At each setting of $n-k$, we generate 10 instances of (6.12) and plot the time and error of the solution returned by each algorithm.

Algorithm	time (s)	std.	$\|X - X^*\|_F^2$	std.	memory (MB)	std.
CertSDP	4.5×10^3	7.0×10^2	1.9×10^{-22}	5.2×10^{-23}	8.5	1.2×10^1
CSSDP	1.0×10^4	6.6×10^{-1}	2.7	9.4×10^{-1}	6.2	1.5×10^1
ProxSDP	1.2×10^4	1.1×10^2	2.9	9.9×10^{-1}	1.9×10^4	1.2×10^2

Table 6.2: Experimental results for $(n - k) = 10^4$ (10 instances) with time limit 10^4 seconds. SCS was unable to complete a single iteration within the time limit and utilized over 70GB of memory.

Algorithm	time (s)	std.	$\|X - X^*\|_F^2$	std.	memory (MB)	std.
CertSDP	5.0×10^4	6.2×10^2	2.5×10^{-2}	6.5×10^{-2}	2.3×10^2	2.0×10^2
CSSDP†	5.0×10^4	4.7	2.8	5.1×10^{-1}	2.0×10^2	2.5×10^2

Table 6.3: Experimental results for $(n - k) = 10^5$ (10 instances) with time limit 5×10^4 seconds. SCS and ProxSDP are not tested as they both come to complete failure due to memory allocation. †CSSDP failed due to numerical issues within the eigenvalue computation on three instances.

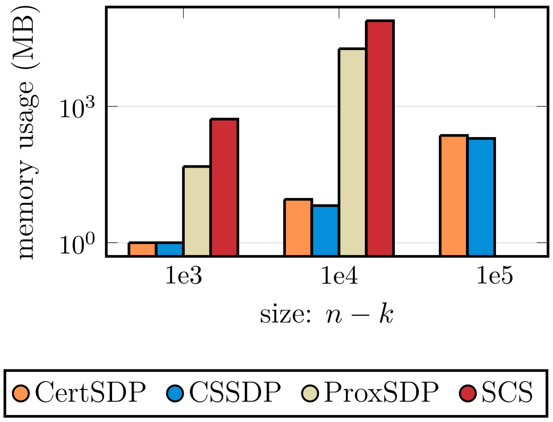

Figure 6.3: Memory usage of different algorithms as a function of the size $n - k$. In this chart, we plot 0.0 MB at 1.0 MB (see Remark 76 for a discussion on measuring memory usage).

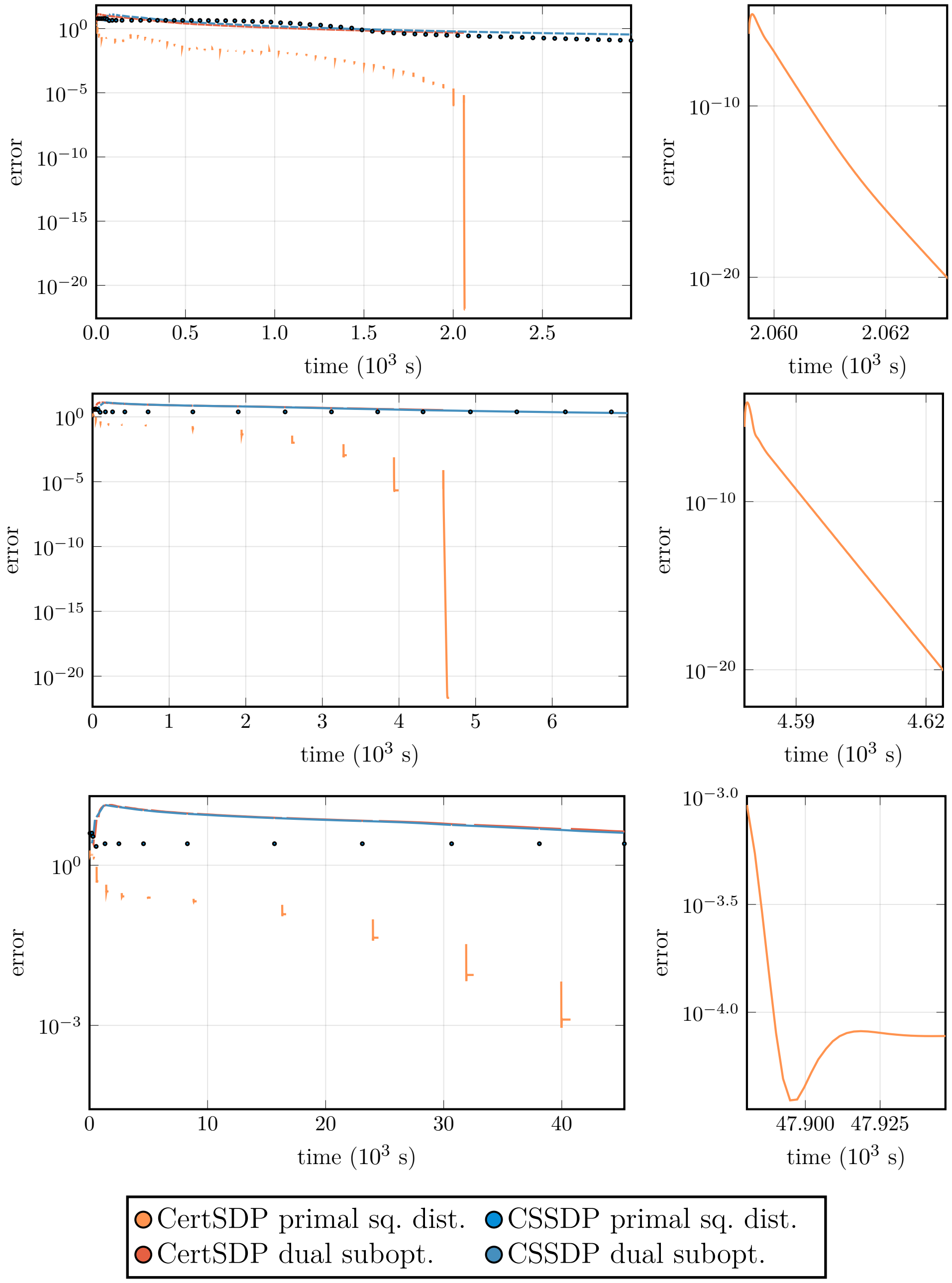

Figure 6.4: Comparison of convergence behavior between CertSDP (Algorithm 10) and CSSDP. The first, second, and third rows show experiments with $n - k = 10^3$, 10^4, and 10^5 respectively. The right subplots give zoomed-in views of the primal squared distance in CertSDP on the final call to Algorithm 9.

- For $n - k = 10^3$ (see Table 6.1), both CertSDP and ProxSDP were able to achieve high accuracy within the time limit, while CSSDP and SCS could not. ProxSDP was faster than CertSDP while CertSDP used significantly less memory.
- For $n - k = 10^4$ (see Table 6.2), CertSDP was the only algorithm that was able to achieve high accuracy within the time limit. The measured memory usage of CertSDP and CSSDP both had high variance, however it is clear that these algorithms use much less memory than ProxSDP and SCS. As previously mentioned, SCS used over 70GB of memory at this size.
- For $n - k = 10^5$ (see Table 6.3), CertSDP and CSSDP were the only algorithms that could be run without memory allocation errors. While neither algorithm was able to achieve the desired accuracy within the time limit, CertSDP (average primal squared distance of 2.5×10^{-2}) significantly outperformed CSSDP (average primal squared distance of 2.8).
- The dual suboptimality for CertSDP and CSSDP behave identically. This is expected as we employ Accelegrad to generate both sequences.
- The primal squared distance and the dual suboptimality for CSSDP track quite closely. This is expected from [60, Theorem 4.1, Table 3], which bounds the primal squared distance by a constant factor of the dual suboptimality for CSSDP.
- The convergence behavior of CautiousAGD depends on whether $\mathcal{U}^{(i)}$ in CertSDP is a certificate of strict complementarity.

  When $\mathcal{U}^{(i)}$ is *not* a certificate of strict complementarity, CautiousAGD behaves as in the bottom-right plot of Figure 6.4: It briefly converges linearly before plateauing. This makes sense as the iterates in CautiousAGD should converge linearly to $\arg\min_X Q_{\mathcal{U}^{(i)}}(X) \neq X^*$.

  When $\mathcal{U}^{(i)}$ is a certificate of strict complementarity, the iterates of CautiousAGD converge linearly to X^* (see the top-right and middle-right plots of Figure 6.4).

7 Variants of simultaneous diagonalizability of quadratic forms

This chapter is based on joint work [177] with Rujun Jiang.

A set of quadratic forms is simultaneously diagonalizable via congruence (SDC) if there exists a basis under which each of the quadratic forms is diagonal. This property appears naturally when analyzing quadratically constrained quadratic programs (QCQPs) and has important implications in globally solving such problems using branch-and-bound methods. This chapter extends the reach of the SDC property by studying two new weaker notions of simultaneous diagonalizability. Specifically, we say that a set of quadratic forms is almost SDC (ASDC) if it is the limit of SDC sets and d-restricted SDC (d-RSDC) if it is the restriction of an SDC set in up to d-many additional dimensions. In the context of QCQPs, these properties correspond to problems that may be diagonalized after arbitrarily small perturbations or after the introduction of d additional variables. Our main contributions are complete characterizations of the ASDC pairs and nonsingular triples of symmetric matrices, as well as a sufficient condition for the 1-RSDC property for pairs of symmetric matrices. Surprisingly, we show that *every* singular symmetric pair is ASDC and that *almost every* symmetric pair is 1-RSDC. We accompany our theoretical results with preliminary numerical experiments applying these constructions to solve QCQPs within branch-and-bound schemes.

7.1 Introduction

This chapter investigates two new notions of *simultaneous diagonalizability* of quadratic forms and their applications in solving quadratically constrained quadratic programs (QCQPs).

Let $\mathbb{S}^n$ denote the real vector space of $n \times n$ symmetric matrices.[1] Recall that a set of matrices $\mathcal{A} \subseteq \mathbb{S}^n$ is said to be *simultaneously diagonalizable via congruence* (SDC) if there exists an invertible $P \in \mathbb{R}^{n\times n}$ such that $P^\intercal AP$ is diagonal for every $A \in \mathcal{A}$. This property has attracted significant interest in the optimization community in recent years in the context of solving subclasses of QCQPs and their relaxations [93, 107, 116, 134, 181, 199, 200]. Specifically, the SDC property corresponds to the ability to rewrite a given QCQP as a *diagonal QCQP* (see Section 7.1.1 below). The SDC property also finds applications in areas such as signal processing, multivariate statistics, medical imaging analysis, and genetics; see [39, 175] and references therein.

[1] While all of our results hold with only minor modifications over both $\mathbb{C}^n$ and Hermitian matrices and $\mathbb{R}^n$ and symmetric matrices, we will simplify our presentation in the main body by discussing only the real setting; see Section G.3 for a discussion of our results in the complex setting.

In this chapter, we take a step towards increasing the practical importance of the SDC property in the context of globally solving QCQPs by investigating two weaker notions of simultaneous diagonalizability. These weaker notions formalize methods for diagonalizing classes of *a priori* non-diagonalizable QCQPs.

7.1.1 Motivation

A general QCQP can be written as

$$\mathrm{Opt} \coloneqq \inf_{x\in\mathbb{R}^n} \left\{ x^\intercal A_1 x + 2b_1^\intercal x + c_1 : \begin{array}{l} x^\intercal A_i x + 2b_i^\intercal x + c_i \;\Box_i\; 0,\ \forall i\in[2,m] \\ x\in\mathcal{L} \end{array} \right\}, \tag{7.1}$$

where for every $i\in[m]$, we have $A_i\in\mathbb{S}^n$, $b_i\in\mathbb{R}^n$, $c_i\in\mathbb{R}$, and $\Box_i\in\{\leq,=\}$; and $\mathcal{L}\subseteq\mathbb{R}^n$ is a polyhedron. In words, the objective is to minimize a quadratic function subject to quadratic (in)equality constraints and linear (in)equality constraints. QCQPs are highly expressive and capture numerous hard problems of both applied and theoretical interest; see [11, 161, 181] and references therein. In fact, this class of problems is NP-hard even if $\mathcal{L}=[-1,1]^n$ and there are no quadratic constraints (e.g., via max-cut).

We will refer to a QCQP in which the set of symmetric matrices $\mathcal{A}=\{A_1,\dots,A_m\}$ is SDC as a *diagonalizable QCQP*. By definition, a diagonalizable QCQP can be rewritten as a *diagonal QCQP* (one in which $\mathcal{A}$ is a set of *diagonal* matrices) upon a linear change of variables. Indeed, letting $y=P^{-1}x$ and $D_i=P^\intercal A_i P$ gives

$$\inf_{y\in\mathbb{R}^n} \left\{ y^\intercal D_1 y + 2(P^\intercal b_1)^\intercal y + c_1 : \begin{array}{l} y^\intercal D_i y + 2(P^\intercal b_i)^\intercal y + c_i \;\Box_i\; 0,\ \forall i\in[2,m] \\ y\in P^{-1}\mathcal{L} \end{array} \right\}.$$

While diagonal QCQPs are still NP-hard in general, they benefit from a number of advantages over more general QCQPs:

- It is well known that the standard Shor semidefinite program (SDP) relaxation of a diagonal QCQP is equivalent to a second-order cone program (SOCP) [181]. Consequently, the SDP relaxation can be solved substantially faster for diagonal QCQPs than for general QCQPs. Similar ideas have be used to build cheap but strong convex relaxations within branch and bound (BB) frameworks for nonconvex QCQPs [199, 200].

  As we will see in Section 7.7, when P is well-conditioned, the computational savings of replacing an SDP with an SOCP within every node of a BB tree can outweigh the computational costs of preprocessing a diagonalizable QCQP into a diagonal QCQP.

- Additionally, qualitative properties of the standard SDP relaxation are often easier to analyze in the context of diagonal QCQPs. For example, a long line of work has investigated when the SDP relaxations of certain diagonal QCQPs are *exact* (for various definitions of exact) and have given sufficient conditions for these properties [21, 24, 38, 87, 89, 92, 93, 112, 180]. Often, such arguments rely on conditions (such as convexity[2] or polyhedrality) of the quadratic image [147] or the set of convex Lagrange multipliers [181]. In this context, the

[2] The convexity of the quadratic image is sometimes referred to as "hidden convexity."

SDC property ensures that both of these sets are polyhedral. While such conditions have been generalized beyond only diagonal or diagonalizable QCQPs, the sufficient conditions often become much more difficult to verify [179, 181].

As we will see in Section 7.7, the SDP relaxation of a diagonal QCQP with bound constraints (as are encountered within BB schemes) admits low-rank solutions. Heuristically, this suggests that the corresponding SDP relaxations should be strong. We verify this intuition with numerical experiments.

7.1.2 Main contributions and outline

In this chapter, we define and analyze the *almost SDC* (ASDC) and *d-restricted SDC* (d-RSDC) properties; see Sections 7.2 and 7.5 for precise definitions. Informally, $\mathcal{A} \subseteq \mathbb{S}^n$ is ASDC if it is the limit of SDC sets and d-RSDC if it is the restriction of an SDC set in $\mathbb{S}^{n+d}$ to $\mathbb{S}^n$. In the context of QCQPs, if the set $\mathcal{A} = \{A_1, \dots, A_m\}$ is ASDC, then the QCQP can be diagonalized after arbitrarily small perturbations to the A_i matrices. In a similar vein, if $\mathcal{A}$ is d-RSDC, then the QCQP can be diagonalized after the introduction of d additional "dummy" variables.

A summary of our contributions, along with an outline of the chapter, follows:

- We conclude this section in Section 7.1.3 by reviewing related work on BB methods for QCQPs, the SDC property, and the almost simultaneously diagonalizable via similarity property.

- In Section 7.2, we formally define the SDC and ASDC properties and review known characterizations of the SDC property. We additionally highlight a number of behaviors of the SDC property which will later contrast with those of the ASDC property.

- In Section 7.3, we give a complete characterization of the ASDC property for pairs of symmetric matrices. In particular, Theorem 34 states that *every* singular[3] pair $\{A, B\} \subseteq \mathbb{S}^n$ is ASDC. The proof of this statement relies on the canonical form for pairs of symmetric matrices [173] under congruence transformations and the invertibility of a certain matrix related to the eigenvalues of an "arrowhead" matrix.

- In Section 7.4, we give a complete characterization of the ASDC property for *nonsingular* triples of symmetric matrices. Our proof and constructions rely on facts about block matrices with Toeplitz upper triangular blocks. We review the relevant properties of such matrices in Section G.2.

- In Section 7.5, we formally define the d-RSDC property and highlight its relation to the ASDC property. We then show in Theorem 36 that the 1-RSDC property holds for almost every pair of symmetric matrices. We also give a construction for the d-RSDC property for $d \geq 1$ and almost every pair of symmetric matrices. This second construction makes use of additional degrees of freedom and empirically leads to improved performance in the context of globally solving QCQPs (see Section 7.7).

[3] See Definition 28.

- In Section 7.6, we construct obstructions to *a priori* plausible generalizations of our developments in Sections 7.3 to 7.5. Section 7.6.1 shows that, in contrast to Theorem 34, there exist singular triples of symmetric matrices which are *not* ASDC. The same construction can be interpreted as a triple of symmetric matrices which is not d-RSDC for any $d < \lfloor n/2 \rfloor$; this contrasts with Theorem 36. Next, Section 7.6.2 shows that a natural generalization of our characterizations of the ASDC property for pairs and triples of symmetric matrices cannot hold for general m-tuples; specifically this natural generalization fails for $m \geq 7$.

- In Section 7.7, we revisit one of the key motivations for studying the ASDC and d-RSDC properties—solving QCQPs more efficiently. In this context, we begin by deriving a number of theoretical results that give heuristic reasons why one would expect SOCP-based BB methods for diagonal QCQPs to outperform SDP-based BB methods for more general QCQPs. We then present a number of preliminary numerical experiments that corroborate this intuition.

Remark 77. In the main body of this chapter, we will state and prove our results for only the real symmetric setting. Nevertheless, our results and proofs extend almost verbatim to the Hermitian setting by replacing the canonical form of a pair of real symmetric matrices (Proposition 23) by the canonical form for a pair of Hermitian matrices (see [105, Theorem 6.1]). As no new ideas or insights are required for handling the Hermitian setting, we defer formally stating our results in the Hermitian setting and discussing the necessary modifications to our proofs to Section G.3. □

7.1.3 Related work

Branch-and-bound methods for QCQPs Most existing works for globally solving QCQPs are based on spatial BB methods. Audet et al. [8] developed an LP-based branch and cut method for QCQPs using the reformulation-linearization technique (RLT) [158]. Linderoth [109] proposed a triangle-based BB algorithm for solving nonconvex QCQPs, where two-dimensional triangles and rectangles are used to partition the feasible region. Recently, Zhou et al. [200] proposed a BB algorithm for QCQPs with nonconvex objective functions and convex quadratic constraints, based on the SDC property between the objective function and a specific aggregation of the convex quadratic constraints under a positive definiteness assumption. Luo et al. [116] propose a BB algorithm based on the SDC property of two positive semidefinite matrices for solving a nonconvex QCQP arising from optimal portfolio deleveraging problems. Please refer to [27, 44, 45, 63, 114] for other recent developments in globally solving nonconvex QCQPs.

The SDC property for sets of quadratic forms and SDC algorithms. The SDC property for a pair of symmetric matrices (more generally, Hermitian matrices) is well-understood and follows from results due to Weierstrass [187] and Kronecker (see [102]). We review these results in Section 7.2 (see also Proposition 23). More recently, there has been much interest in the optimization literature towards understanding the SDC property for general m-tuples of quadratic forms [93, 107, 134]. In fact, the search for "sensible and "palpable" conditions" for this property appeared as an open question on a short list of 14 open questions in nonlinear analysis and optimization [84]. In the real symmetric setting, Jiang and Li [93] gave a complete characterization of this property under a semidefiniteness assumption. This result was then improved upon by

Nguyen et al. [134] who removed the semidefiniteness assumption. Le and Nguyen [107] additionally extend these characterizations to the case of Hermitian matrices. Bustamante et al. [39] gave a complete characterization of the simultaneous diagonalizability of an m-tuple of *symmetric complex* matrices under $^\intercal$-congruence.[4]

We remark that this line of work is "algorithmic" and gives numerical procedures for deciding if a given set of quadratic forms is SDC. See [107] and references therein.

The almost SDS property. An analogous theory for the *almost* simultaneous diagonalizability of *linear operators* has been studied in the literature. In this setting, the congruence transformation is naturally replaced by a similarity transformation[5] and the SDC property is replaced by simultaneous diagonalizability *via similarity* (SDS). A widely cited theorem due to Motzkin and Taussky [127] shows that every pair of commuting linear operators, i.e., a pair of matrices in $\mathbb{C}^{n\times n}$, is almost SDS. This line of investigation was more recently picked up by O'meara and Vinsonhaler [137] who showed that triples of commuting linear operators are almost SDS under a regularity assumption on the dimensions of eigenspaces associated with the linear operators.

7.1.4 Additional notation

Let $\mathbb{N} = \{1, 2, \dots\}$ and $\mathbb{N}_0 = \{0, 1, \dots\}$. For $m, n \in \mathbb{N}_0$, let $[m, n] = \{m, m+1, \dots, n\}$ and $[n] = \{1, \dots, n\}$. By convention, if $m \geq n+1$ (respectively, $n \leq 0$), then $[m, n] = \varnothing$ (respectively, $[n] = \varnothing$). Given $x \in \mathbb{R}^n$, let $\operatorname{supp}(x) \coloneqq \{i \in [n] : x_i \neq 0\}$ denote the support of x. Let $|I|$ be the cardinality of a set I. For $\alpha_1, \dots, \alpha_k \in \mathbb{R}$, let $\operatorname{Diag}(\alpha_1, \dots, \alpha_k) \in \mathbb{R}^{k\times k}$ denote the diagonal matrix with ith entry α_i. For $A_1, \dots, A_k$ square matrices, let $\operatorname{Diag}(A_1, \dots, A_k)$ denote the block diagonal matrix with ith block A_i. Given $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{m\times m}$, let $A \oplus B \in \mathbb{R}^{(n+m)\times(n+m)}$ and $A \otimes B \in \mathbb{R}^{nm\times nm}$ denote the direct sum and Kronecker product of A and B respectively. Given $A, B \in \mathbb{R}^{n\times n}$, let $[A, B] \coloneqq AB - BA$ denote the commutator of A and B. For $A \in \mathbb{R}^{n\times n}$, let $\|A\|$ denote the spectral norm of A. Given $\alpha \in \mathbb{C}$, let $\operatorname{Re}(\alpha)$, $\operatorname{Im}(\alpha)$, and α^* denote the real and imaginary parts and complex conjugate of α respectively. For $A \in \mathbb{C}^{n\times n}$, let A^* denote the conjugate transpose of A. We will denote the imaginary unit by the symbol i in order to distinguish it from the variable i, which will often be used as an index.

7.2 Preliminaries

In this section, we define our main objects of study and recall some useful results from the literature.

Definition 26. A set $\mathcal{A} \subseteq \mathbb{S}^n$ is *simultaneously diagonalizable via congruence* (SDC) if there exists an invertible $P \in \mathbb{R}^{n\times n}$ such that $P^\intercal AP$ is diagonal for all $A \in \mathcal{A}$. □

Remark 78. The SDC property is the natural notion for simultaneous diagonalization in the context of quadratic forms. Indeed, suppose $\mathcal{A} \subseteq \mathbb{S}^n$ is SDC and let P denote the corresponding invertible matrix. Then, performing the change of variables $y = P^{-1}x$, we have that $x^\intercal Ax = y^\intercal(P^\intercal AP)y$ is separable in y for every $A \in \mathcal{A}$. □

[4] We emphasize that Bustamante et al. [39] consider complex symmetric matrices and adopt $^\intercal$-congruence as their notion of congruence.

[5] Recall that two matrices $A, B \in \mathbb{C}^{n\times n}$ are similar if there exists an invertible $P \in \mathbb{C}^{n\times n}$ such that $A = P^{-1}BP$.

Observation 5. *The SDC property is closed under taking spans and subsets. In particular, $\mathcal{A} \subseteq \mathbb{S}^n$ is SDC if and only if $\{A_1, \dots, A_m\}$ is SDC for some basis $\{A_1, \dots, A_m\}$ of $\operatorname{span}(\mathcal{A})$.*

We begin by studying the following relaxation of the SDC property.

Definition 27. A set $\mathcal{A} \subseteq \mathbb{S}^n$ is *almost simultaneously diagonalizable via congruence* (ASDC) if there exist sequences $A_i \to A$ for every $A \in \mathcal{A}$ such that for every $i \in \mathbb{N}$, the set $\{A_i : A \in \mathcal{A}\}$ is SDC. □

Observation 6. *The ASDC property is closed under taking spans and subsets. In particular, $\mathcal{A} \subseteq \mathbb{S}^n$ is ASDC if and only if $\{A_1, \dots, A_m\}$ is ASDC for some basis $\{A_1, \dots, A_m\}$ of $\operatorname{span}(\mathcal{A})$.*

When $|\mathcal{A}|$ is finite, we will use the following equivalent definition of ASDC.

Observation 7. *A finite set $\{A_1, \dots, A_m\} \subseteq \mathbb{S}^n$ is ASDC if and only if for all $\epsilon > 0$, there exist $\tilde{A}_1, \dots, \tilde{A}_m \in \mathbb{S}^n$ such that*

- *for all $i \in [m]$, the spectral norm $\left\| A_i - \tilde{A}_i \right\| \leq \epsilon$, and*
- *$\left\{ \tilde{A}_1, \dots, \tilde{A}_m \right\}$ is SDC.*

We will additionally need the following two definitions.

Definition 28. A set $\mathcal{A} \subseteq \mathbb{S}^n$ is *nonsingular* if there exists a nonsingular $A \in \operatorname{span}(\mathcal{A})$. Else, it is *singular*. □

Definition 29. Given a set $\mathcal{A} \subseteq \mathbb{S}^n$, we will say that $S \in \mathcal{A}$ is a *max-rank element* of $\operatorname{span}(\mathcal{A})$ if $\operatorname{rank}(S) = \max_{A \in \mathcal{A}} \operatorname{rank}(A)$. □

7.2.1 Characterization of SDC

A number of necessary and/or sufficient conditions for the SDC property have been given in the literature [39, 88, 105]. For our purposes, we will need the following two results. The first result gives a characterization of the SDC property for nonsingular sets of symmetric matrices and is well-known (see [88, Theorem 4.5.17]). The second result, due to Bustamante et al. [39], gives a characterization of the SDC property for singular sets of symmetric matrices by reducing to the nonsingular case. For completeness, we provide a short proof for each of these results in Section G.1.

Proposition 21. *Let $\mathcal{A} \subseteq \mathbb{S}^n$ and suppose $S \in \operatorname{span}(\mathcal{A})$ is nonsingular. Then, $\mathcal{A}$ is SDC if and only if $S^{-1}\mathcal{A}$ is a commuting set of diagonalizable matrices with real eigenvalues.*

Proposition 22. *Let $\mathcal{A} \subseteq \mathbb{S}^n$ and suppose $S \in \operatorname{span}(\mathcal{A})$ is a max-rank element of $\operatorname{span}(\mathcal{A})$. Then, $\mathcal{A}$ is SDC if and only if $\operatorname{range}(A) \subseteq \operatorname{range}(S)$ for every $A \in \mathcal{A}$ and $\left\{ A|_{\operatorname{range}(S)} : A \in \mathcal{A} \right\}$ is SDC.*

We close this section with two lemmas highlighting consequences of the SDC property which we will compare and contrast with consequences of the ASDC property.

Lemma 77. *Let $\mathcal{A} \subseteq \mathbb{S}^n$ and suppose $S \in \operatorname{span}(\mathcal{A})$ is positive definite. Then, $\mathcal{A}$ is SDC if and only if $S^{-1/2}\mathcal{A}S^{-1/2}$ is a commuting set.*

Proof. This follows as an immediate corollary to Proposition 21 and the fact that $S^{-1}A$ has the same eigenvalues as the symmetric matrix $S^{-1/2}AS^{-1/2}$. ■

In particular, when $\operatorname{span}(\mathcal{A})$ contains a positive definite matrix, the SDC and ASDC properties can be shown to be equivalent.

Corollary 27. *Let $\mathcal{A} \subseteq \mathbb{S}^n$ and suppose $S \in \operatorname{span}(\mathcal{A})$ is positive definite. Then, $\mathcal{A}$ is SDC if and only if $\mathcal{A}$ is ASDC.*

Despite Corollary 27, we will see soon that the ASDC property is qualitatively quite different to the SDC property in a number of settings (in particular, for singular pairs of symmetric matrices; see Theorem 34). Specifically, we will contrast the following consequence of the SDC property.

Lemma 78 ([107, Lemma 9])**.** *Let $\mathcal{A} \subseteq \mathbb{S}^n$ and suppose there exists a common block decomposition*

$$A = \begin{pmatrix} \bar{A} & \\ & 0_d \end{pmatrix}$$

for all $A \in \mathcal{A}$. Then $\mathcal{A}$ is SDC if and only if $\left\{\bar{A} :\ A \in \mathcal{A}\right\} \subseteq \mathbb{S}^{n-d}$ is SDC.

7.3 The ASDC property of symmetric pairs

In this section, we will give a complete characterization of the ASDC property for pairs of symmetric matrices (henceforth, *symmetric pairs*). We will switch the notation above and label our matrices $\mathcal{A} = \{A, B\}$. Our analysis will proceed in two cases: when $\{A, B\}$ is nonsingular and singular respectively.

7.3.1 A canonical form for symmetric pairs

In this section and the next, we will make regular use of the canonical form for symmetric pairs [105, 173].

We will need to define the following special matrices. For $n \geq 2$, let $F_n,\ G_n,\ H_n \in \mathbb{S}^n$ denote the matrices of the form

$$F_n = \begin{pmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{pmatrix}, \qquad G_n = \begin{pmatrix} & & & 0 \\ & & \iddots & 1 \\ & 0 & \iddots & \\ 0 & 1 & & \end{pmatrix}, \quad \text{and} \quad H_n = \begin{pmatrix} & & 1 & 0 \\ & \iddots & 0 & \\ 1 & \iddots & & \\ 0 & & & \end{pmatrix}.$$

Set $F_1 = (1)$ and $G_1 = H_1 = (0)$.

The following proposition is adapted[6] from [105, Theorem 9.1].

[6] The original statement of [105, Theorem 9.1] contains one additional type of block: those corresponding to the eigenvalues at infinity. These blocks do not exist in our setting by the assumption that A is a max-rank element of $\operatorname{span}(\{A, B\})$.

Proposition 23. *Let $A, B \in \mathbb{S}^n$ and suppose A is a max-rank element of* $\operatorname{span}(\{A, B\})$*. Then, there exists an invertible $P \in \mathbb{R}^{n\times n}$ such that $P^\intercal AP = \operatorname{Diag}(S_1, \dots, S_m)$ and $P^\intercal BP = \operatorname{Diag}(T_1, \dots, T_m)$ are block diagonal matrices with compatible block structure. Here, $m = m_1 + m_2 + m_3 + m_4$ corresponds to four different types of blocks where each $m_i \in \mathbb{N}_0$ may be zero. Additionally, $m_4 \in \{0, 1\}$.*

The first m_1-many blocks of $P^\intercal AP$ and $P^\intercal BP$ have the form

$$S_i = \sigma_i F_{n_i}, \qquad T_i = \sigma_i(\lambda_i F_{n_i} + G_{n_i}),$$

where $n_i \in \mathbb{N}$, $\sigma_i \in \{\pm 1\}$, and $\lambda_i \in \mathbb{R}$. The next m_2-many blocks of $P^\intercal AP$ and $P^\intercal BP$ have the form

$$S_i = \begin{pmatrix} & F_{n_i} \\ F_{n_i} & \end{pmatrix}, \qquad T_i = F_{n_i} \otimes \begin{pmatrix} \operatorname{Im}(\lambda_i) & \operatorname{Re}(\lambda_i) \\ \operatorname{Re}(\lambda_i) & -\operatorname{Im}(\lambda_i) \end{pmatrix} + G_{n_i} \otimes F_2, \tag{7.2}$$

where $n_i \in \mathbb{N}$ and $\lambda_i \in \mathbb{C} \setminus \mathbb{R}$. The next m_3-many blocks of $P^\intercal AP$ and $P^\intercal BP$ have the form

$$S_i = \begin{pmatrix} & & F_{n_i} \\ & 0 & \\ F_{n_i} & & \end{pmatrix}, \qquad T_i = G_{2n_i+1},$$

where $n_i \in \mathbb{N}$. If $m_4 = 1$, then the last block of $P^\intercal AP$ and $P^\intercal BP$ has the form $S_m = T_m = 0_{n_m}$ for some $n_m \in \mathbb{N}$.

We will repeatedly encounter real matrices that represent complex numbers, e.g., the blocks $S_i^{-1}T_i$ for i corresponding to m_2 in the canonical form. We recall some useful facts: Let $J \in \mathbb{C}^{2\times 2}$ be the unitary matrix

$$J := \begin{pmatrix} \frac{\mathrm{i}}{\sqrt{2}} & \frac{-\mathrm{i}}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \in \mathbb{C}^{2\times 2}.$$

Then, a matrix of the form $\begin{pmatrix} \operatorname{Im}(\lambda_i) & \operatorname{Re}(\lambda_i) \\ \operatorname{Re}(\lambda_i) & -\operatorname{Im}(\lambda_i) \end{pmatrix}$ has the same eigenvalues as

$$J^* \begin{pmatrix} \operatorname{Re}(\lambda_i) & -\operatorname{Im}(\lambda_i) \\ \operatorname{Im}(\lambda_i) & \operatorname{Re}(\lambda_i) \end{pmatrix} J = \begin{pmatrix} \lambda_i & \\ & \lambda_i^* \end{pmatrix}.$$

7.3.2 The nonsingular case

In this section, we will show that if A is invertible, then $\{A, B\}$ is ASDC if and only if $A^{-1}B$ has real eigenvalues. We begin by examining two examples that are representative of the situation when A is invertible. Note in this case, that $m_3 = m_4 = 0$ in the canonical form (Proposition 23).

Example 18. Let $\lambda \in \mathbb{R}$ and consider

$$A = \begin{pmatrix} & 1 \\ 1 & \end{pmatrix} = F_2, \qquad B = \begin{pmatrix} 0 & \lambda \\ \lambda & 1 \end{pmatrix} = \lambda F_2 + G_2.$$

Noting that $A^{-1}B$ is not diagonalizable, we conclude via Proposition 21 that $\{A, B\}$ is not SDC. On the other hand, let $\epsilon > 0$ and define

$$\tilde{B} = \begin{pmatrix} \epsilon & \lambda \\ \lambda & 1 \end{pmatrix}.$$

Now, $A^{-1}\tilde{B}$ has eigenvalues $\lambda \pm \sqrt{\epsilon}$, whence by Proposition 21 $\left\{A, \tilde{B}\right\}$ is SDC. □

Example 19. Let $\lambda \in \mathbb{C} \setminus \mathbb{R}$ and consider

$$A = F_2 = \begin{pmatrix} & 1 \\ 1 & \end{pmatrix}, \qquad B = \begin{pmatrix} \operatorname{Im}(\lambda) & \operatorname{Re}(\lambda) \\ \operatorname{Re}(\lambda) & -\operatorname{Im}(\lambda) \end{pmatrix}.$$

Noting that $A^{-1}B$ has non-real eigenvalues, we conclude via Proposition 21 (and the fact that eigenvalues vary continuously) that $\{A, B\}$ is not ASDC. □

The following technical lemma will be useful in proving the main result of this section and shows that it is possible to perturb B to ensure that $A^{-1}B$ has simple eigenvalues while maintaining its number of real/complex eigenvalues.

Lemma 79. *Let $\{A, B\} \subseteq \mathbb{S}^n$ and suppose A is invertible. For all $\epsilon > 0$, there exists $\tilde{B}$ such that*

- $\left\| B - \tilde{B} \right\| \leq \epsilon$,
- *$A^{-1}\tilde{B}$ has simple eigenvalues (whence $A^{-1}\tilde{B}$ is diagonalizable), and*
- *$A^{-1}\tilde{B}$ and $A^{-1}B$ have the same number of real eigenvalues counted with multiplicity.*

Proof. Without loss of generality, we may assume that $A = \operatorname{Diag}(S_1, \dots, S_m)$ and $B = \operatorname{Diag}(T_1, \dots, T_m)$ are in canonical form (Proposition 23). Note that as A is invertible, we will have $m_3 = m_4 = 0$. For notational convenience, let $r = m_1$ and let $\sigma_1, \dots, \sigma_r, n_1 \dots, n_m, \lambda_1, \dots, \lambda_m$ denote the quantities furnished by Proposition 23. We will give a probabilistic construction (summarized in Algorithm 11) for $\tilde{B}$ that satisfies all three conditions with probability one.

Let $\delta = \frac{\epsilon}{2}$ and pick a random η uniformly from $[-\delta, \delta]^m$. Define the blocks $\tilde{T}_i$ as

$$\begin{aligned} \tilde{T}_i &\coloneqq T_i + \sigma_i(\eta_i F_{n_i} + \delta H_{n_i}),\ \forall i \in [r], \\ \tilde{T}_i &\coloneqq T_i + (\eta_i F_{n_i} + \delta H_{n_i}) \otimes F_2,\ \forall i \in [r+1, m], \end{aligned} \tag{7.3}$$

and set $\tilde{B} \coloneqq \operatorname{Diag}(\tilde{T}_1, \dots, \tilde{T}_m)$. Then, $A^{-1}\tilde{B} = \operatorname{Diag}(S_1^{-1}\tilde{T}_1, \dots, S_k^{-1}\tilde{T}_k)$ is again a block diagonal matrix. Note that for $i \in [r]$, the block

$$S_i^{-1}\tilde{T}_i = (\lambda_i + \eta_i)I_{n_i} + F_{n_i}G_{n_i} + \delta F_{n_i}H_{n_i}$$

is a Toeplitz tridiagonal matrix. Next, for $i \in [r+1, m]$, the block $S_i^{-1}\tilde{T}_i$ has the form

$$S_i^{-1}\tilde{T}_i = I_{n_i} \otimes \begin{pmatrix} \operatorname{Re}(\lambda_i) & -\operatorname{Im}(\lambda_i) \\ \operatorname{Im}(\lambda_i) & \operatorname{Re}(\lambda_i) \end{pmatrix} + (\eta_i I_{n_i} + F_{n_i}G_{n_i} + \delta F_{n_i}H_{n_i}) \otimes I_2. \tag{7.4}$$

Algorithm 11 Construction for simple eigenvalues

```
Given $A', B' \in \mathbb{S}^n$ such that $A'$ is invertible and $\epsilon' > 0$
  1. Compute the canonical form [105] for $\{A', B'\}$, i.e.,
         $P^\intercal A'P = A = \mathrm{Diag}(S_1, \dots, S_m)$, and
         $P^\intercal B'P = B = \mathrm{Diag}(T_1, \dots, T_m)$.
  2. Set $\epsilon = \epsilon'/\left\|P^{-1}\right\|^2$ and $\delta = \frac{\epsilon}{2}$
  3. Pick $\eta$ uniformly at random from $[-\delta, \delta]^m$
  4. Return $\left\{A',\ P^{-\intercal}\tilde{B}P^{-1}\right\}$ where $\tilde{B} := \mathrm{Diag}(\tilde{T}_1, \dots, \tilde{T}_m)$ and $\tilde{T}_i$ are defined in (7.3)
```

Note that $S_i^{-1}\tilde{T}_i$ has the same eigenvalues as

$$\begin{aligned}&(I_{n_i} \otimes J)^{-1} S_i^{-1}\tilde{T}_i(I_{n_i} \otimes J)\\ &\qquad = I_{n_i} \otimes \begin{pmatrix} \lambda_i & \\ & \lambda_i^* \end{pmatrix} + (\eta_i I_{n_i} + F_{n_i}G_{n_i} + \delta F_{n_i}H_{n_i}) \otimes I_2.\end{aligned}$$

This is, up to a simultaneous permutation of rows and columns, a direct sum of two Toeplitz tridiagonal matrices.

Using the closed form expression for eigenvalues of Toeplitz tridiagonal matrices [88], we have that $A^{-1}\tilde{B}$ has eigenvalues

$$\begin{aligned}&\bigcup_{i=1}^{r}\left\{\lambda_i + \eta_i + 2\sqrt{\delta}\cos\left(\frac{\pi j}{n_i+1}\right) : j \in [n_i]\right\}\\ &\qquad \cup \bigcup_{i=r+1}^{m}\left\{\lambda + \eta_i + 2\sqrt{\delta}\cos\left(\frac{\pi j}{n_i+1}\right) : j \in [n_i],\ \lambda \in \{\lambda_i, \lambda_i^*\}\right\}.\end{aligned}$$

Note that the quantity $\eta_i + 2\sqrt{\delta}\cos\left(\frac{\pi j}{n_i+1}\right)$ is real so that $A^{-1}B$ and $A^{-1}\tilde{B}$ have the same number of real eigenvalues counted with multiplicity. Furthermore, as $\eta \in [-\delta, \delta]^m$ was picked uniformly at random, we have that $A^{-1}\tilde{B}$ has only simple eigenvalues with probability one.

Finally, $\left\|B - \tilde{B}\right\| = \left\|\mathrm{Diag}(T_1 - \tilde{T}_1, \dots, T_m - \tilde{T}_m)\right\| = \max_i\left\|T_i - \tilde{T}_i\right\| \leq \epsilon$. ■

The following theorem follows as a simple corollary to our developments thus far.

Theorem 33. *Let $A, B \in \mathbb{S}^n$ and suppose A is invertible. Then, $\{A, B\}$ is ASDC if and only if $A^{-1}B$ has real eigenvalues.*

Proof. ($\Rightarrow$) This direction holds trivially by continuity of eigenvalues and the assumption that A is invertible.

($\Leftarrow$) Let $\epsilon > 0$. Then, applying Lemma 79 to $\{A, B\}$, we get $\tilde{B}$ such that $\left\|B - \tilde{B}\right\| \leq \epsilon$ and $A^{-1}\tilde{B}$ is a matrix with real simple eigenvalues. We deduce by Proposition 21 that $\left\{A, \tilde{B}\right\}$ is SDC. ■

Corollary 28. *Let* $\mathcal{A} = \{A_1, \dots, A_m\}$ *in* (7.1) *and suppose* $\operatorname{span}(\mathcal{A}) = \operatorname{span}\{A, B\}$ *where* A *is invertible. Furthermore, suppose* $A^{-1}B$ *has real eigenvalues. Then for any* $\epsilon > 0$, *there exist* $\left\|\tilde{A}_i - A_i\right\| \leq \epsilon$ *such that*

$$\inf_{x\in\mathbb{R}^n} \left\{ x^\intercal \tilde{A}_1 x + 2b_1^\intercal x + c_1 : \begin{array}{l} x^\intercal \tilde{A}_i x + 2b_i^\intercal x + c_i \ \square_i \ 0,\ \forall i \in [2, m] \\ x \in \mathcal{L} \end{array} \right\} \tag{7.5}$$

is a diagonalizable QCQP. The matrices $\tilde{A}_i$ *and the invertible matrix* P *diagonalizing* (7.5) *can be computed via Algorithm 11.*

7.3.3 The singular case

In the remainder of this section, we investigate the ASDC property when $\{A, B\}$ is singular. We will show, surprisingly, that *every* singular symmetric pair is ASDC. We begin with an example and some intuition.

Example 20. In contrast to the SDC property (cf. Lemma 78), the ASDC property of a pair $\{A, B\}$ in the singular case does not reduce to the ASDC property of $\left\{\bar{A}, \bar{B}\right\}$, where $\bar{A}$ and $\bar{B}$ are the restrictions of A and B to the joint range of A and B. For example, let

$$A = \begin{pmatrix} & 1 & \\ 1 & & \\ & & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & & \\ & -1 & \\ & & 0 \end{pmatrix},$$

and let $\bar{A}$ and $\bar{B}$ denote the respective 2×2 leading principal submatrices.

By Theorem 33, $\left\{\bar{A}, \bar{B}\right\}$ is not ASDC (and in particular not SDC). On the other hand, we claim that $\{A, B\}$ is ASDC: For $\epsilon > 0$, consider the matrices

$$\tilde{A} = \begin{pmatrix} & 1 & \\ 1 & & \\ & & \epsilon \end{pmatrix}, \qquad \tilde{B} = \begin{pmatrix} 1 & & \sqrt{\epsilon} \\ & -1 & \sqrt{\epsilon} \\ \sqrt{\epsilon} & \sqrt{\epsilon} & 0 \end{pmatrix}.$$

A straightforward computation shows that $\tilde{A}^{-1}\tilde{B}$ has simple eigenvalues $\{-1, 0, 1\}$ whence $\left\{\tilde{A}, \tilde{B}\right\}$ is SDC.

The fact that $\left\{\bar{A}, \bar{B}\right\}$ is not SDC is equivalent to the statement: there does *not* exist a basis $\{p_1, p_2\} \in \mathbb{R}^2$ such that the quadratic forms $x^\intercal \bar{A} x$ and $x^\intercal \bar{B} x$ can be expressed as

$$\begin{aligned} x^\intercal \bar{A} x &= \alpha_1 (p_1^\intercal x)^2 + \alpha_2 (p_2^\intercal x)^2, \quad \text{and} \\ x^\intercal \bar{B} x &= \beta_1 (p_1^\intercal x)^2 + \beta_2 (p_2^\intercal x)^2, \end{aligned}$$

for some $\alpha_i, \beta_i \in \mathbb{R}$. On the other hand, the fact that $\left\{\tilde{A}, \tilde{B}\right\}$ is SDC shows that there exists a spanning set $\{p_1, p_2, p_3\} \subseteq \mathbb{R}^2$ and $\alpha_i, \beta_i \in \mathbb{R}$ such that

$$\begin{aligned} x^\intercal \bar{A} x &= \alpha_1 (p_1^\intercal x)^2 + \alpha_2 (p_2^\intercal x)^2 + \alpha_3 (p_3^\intercal x)^2, \quad \text{and} \\ x^\intercal \bar{B} x &= \beta_1 (p_1^\intercal x)^2 + \beta_2 (p_2^\intercal x)^2 + \beta_3 (p_3^\intercal x)^2. \end{aligned}$$

Intuitively, the ASDC property asks whether a set of quadratic forms can be (almost) diagonalized using n (the ambient dimension)-many linear forms whereas the SDC property may be forced to use a smaller number of linear forms. □

Theorem 34. *Let $\{A, B\} \subseteq \mathbb{S}^n$. If $\{A, B\}$ is singular, then it is ASDC.*

Proof. We make simplifying assumptions: Without loss of generality, we may assume that A is a max-rank element of $\operatorname{span}(\{A, B\})$ and $A = \operatorname{Diag}(S_1, \dots, S_m)$ and $B = \operatorname{Diag}(T_1, \dots, T_m)$ are in canonical form (Proposition 23). We may assume $m_1 = 0$ (else consider the submatrix of A, B corresponding to the remaining blocks). As A is singular, we have $m_3 + m_4 \geq 1$. In fact, we may assume $m_3 + m_4 = 1$ (else, perturb the submatrix of A corresponding to the first $m - 1$ blocks so that A is nonsingular on those blocks). Similarly, if $m_4 = 1$, we may assume that $n_m = 1$. Finally, assume $\operatorname{Diag}(S_1^{-1}T_1, \dots, S_{m-1}^{-1}T_{m-1})$ has simple eigenvalues (else apply Algorithm 11 to the first $m - 1$ blocks). For notational convenience, let $m_2 = k$.

After the above simplifying assumptions, there are three cases left to consider: where $m \geq 2$ and $m_4 = 1$, where $m \geq 2$ and $m_3 = 1$, and where $m = 1$. In the first two cases, A, B have the form

$$A = \left(\begin{array}{c|c|c|c} \begin{smallmatrix} & 1 \\ 1 & \end{smallmatrix} & & & \\ \hline & \ddots & & \\ \hline & & \begin{smallmatrix} & 1 \\ 1 & \end{smallmatrix} & \\ \hline & & & S_m \end{array}\right), \qquad B = \left(\begin{array}{c|c|c|c} \begin{matrix} \mathrm{Im}(\lambda_1) & \mathrm{Re}(\lambda_1) \\ \mathrm{Re}(\lambda_1) & -\mathrm{Im}(\lambda_1) \end{matrix} & & & \\ \hline & \ddots & & \\ \hline & & \begin{matrix} \mathrm{Im}(\lambda_k) & \mathrm{Re}(\lambda_k) \\ \mathrm{Re}(\lambda_k) & -\mathrm{Im}(\lambda_k) \end{matrix} & \\ \hline & & & T_m \end{array}\right) \tag{7.6}$$

where $\lambda_1, \lambda_1^*, \dots, \lambda_k, \lambda_k^* \in \mathbb{C} \setminus \mathbb{R}$ are distinct.

Case 1. In case 1, $S_m = T_m = 0_1$. Set

$$\tilde{A}_\delta = \left(\begin{array}{c|c|c|c} \begin{smallmatrix} & 1 \\ 1 & \end{smallmatrix} & & & \\ \hline & \ddots & & \\ \hline & & \begin{smallmatrix} & 1 \\ 1 & \end{smallmatrix} & \\ \hline & & & \delta \end{array}\right), \; \tilde{B}_\delta = \left(\begin{array}{c|c|c|c} \begin{matrix} \mathrm{Im}(\lambda_1) & \mathrm{Re}(\lambda_1) \\ \mathrm{Re}(\lambda_1) & -\mathrm{Im}(\lambda_1) \end{matrix} & & & \begin{matrix} \sqrt{\delta}\,\mathrm{Re}(\alpha_1) \\ \sqrt{\delta}\,\mathrm{Im}(\alpha_1) \end{matrix} \\ \hline & \ddots & & \vdots \\ \hline & & \begin{matrix} \mathrm{Im}(\lambda_k) & \mathrm{Re}(\lambda_k) \\ \mathrm{Re}(\lambda_k) & -\mathrm{Im}(\lambda_k) \end{matrix} & \begin{matrix} \sqrt{\delta}\,\mathrm{Re}(\alpha_k) \\ \sqrt{\delta}\,\mathrm{Im}(\alpha_k) \end{matrix} \\ \hline \sqrt{\delta}\,\mathrm{Re}(\alpha_1) \;\; \sqrt{\delta}\,\mathrm{Im}(\alpha_1) & \cdots & \sqrt{\delta}\,\mathrm{Re}(\alpha_k) \;\; \sqrt{\delta}\,\mathrm{Im}(\alpha_k) & \delta z \end{array}\right) \tag{7.7}$$

for some $\alpha \in \mathbb{C}^k$, $z \in \mathbb{R}$, and $\delta > 0$ to be chosen later. The eigenvalues of $\tilde{A}_\delta^{-1}\tilde{B}_\delta$ are equal to the eigenvalues of

$$\left(\begin{array}{c|c|c|c} J & & & \\ \hline & \ddots & & \\ \hline & & J & \\ \hline & & & \frac{1}{\sqrt{\delta}} \end{array}\right)^{-1} \tilde{A}_\delta^{-1}\tilde{B}_\delta \left(\begin{array}{c|c|c|c} J & & & \\ \hline & \ddots & & \\ \hline & & J & \\ \hline & & & \frac{1}{\sqrt{\delta}} \end{array}\right)$$

$$= \left(\begin{array}{cc|c|cc|c} \lambda_1 & & & & & \alpha_1^*/\sqrt{2} \\ & \lambda_1^* & & & & \alpha_1/\sqrt{2} \\ \hline & & \ddots & & & \vdots \\ \hline & & & \lambda_k & & \alpha_k^*/\sqrt{2} \\ & & & & \lambda_k^* & \alpha_k/\sqrt{2} \\ \hline -\alpha_1^*\mathrm{i}^*/\sqrt{2} & -\alpha_1\mathrm{i}/\sqrt{2} & \cdots & -\alpha_k^*\mathrm{i}^*/\sqrt{2} & -\alpha_k\mathrm{i}/\sqrt{2} & z \end{array}\right).$$

The characteristic polynomial (in ξ) of this latter matrix is

$$(z - \xi)\prod_{i=1}^{k}(\lambda_i - \xi)(\lambda_i^* - \xi) + \sum_{i=1}^{k}\left(\operatorname{Im}\left(\alpha_i^2\right)\xi - \operatorname{Im}\left(\alpha_i^2\lambda_i\right)\right)\prod_{j\neq i}(\lambda_j - \xi)(\lambda_j^* - \xi) \tag{7.8}$$

and is independent of $\delta > 0$. As λ_i are all non-real, given any $x, y \in \mathbb{R}^k$, it is possible to construct $\alpha \in \mathbb{C}^k$ such that

$$\operatorname{Im}(\alpha_i^2) = y_i \quad \text{and} \quad -\operatorname{Im}(\alpha_i^2\lambda_i) = x_i, \quad \forall i \in [k]. \tag{7.9}$$

Setting α in this manner reduces the characteristic polynomial to

$$(z - \xi)\prod_{i=1}^{k}(\lambda_i - \xi)(\lambda_i^* - \xi) + \sum_{i=1}^{k}(x_i + y_i\xi)\prod_{j\neq i}(\lambda_j - \xi)(\lambda_j^* - \xi). \tag{7.10}$$

It suffices to show that there exist $x, y \in \mathbb{R}^k$ and $z \in \mathbb{R}$ such that the roots of (7.10) are all real, as we may take $\delta > 0$ to zero independently of our choice of x, y, z.

Define the following polynomials.

$$f_i(\xi) \coloneqq \prod_{j\neq i}(\lambda_j - \xi)(\lambda_j^* - \xi), \quad g_i(\xi) \coloneqq \xi f_i(\xi),\ \forall i \in [k], \quad \text{and}$$

$$h(\xi) \coloneqq \prod_{i=1}^{k}(\lambda_i - \xi)(\lambda_i^* - \xi).$$

As $\{\lambda_1, \lambda_1^*, \ldots, \lambda_k, \lambda_k^*\}$ are distinct values in $\mathbb{C} \setminus \mathbb{R}$, we have that $\{f_1, g_1, \ldots, f_k, g_k, h\}$ are a basis for the degree-$2k$ polynomials in ξ. Now pick $2k+1$ distinct values $\xi_1, \ldots, \xi_{2k+1} \in \mathbb{R}$. Note that $\{\xi_1, \ldots, \xi_{2k+1}\}$ are the roots to (7.10) if and only if $x, y \in \mathbb{R}^n$ and $z \in \mathbb{R}$ satisfy

$$\begin{pmatrix} f_1(\xi_1) & g_1(\xi_1) & \cdots & f_k(\xi_1) & g_k(\xi_1) & h(\xi_1) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ f_1(\xi_{2k+1}) & g_1(\xi_{2k+1}) & \cdots & f_k(\xi_{2k+1}) & g_k(\xi_{2k+1}) & h(\xi_{2k+1}) \end{pmatrix} \begin{pmatrix} x_1 \\ y_1 \\ \vdots \\ x_k \\ y_k \\ z \end{pmatrix} = \begin{pmatrix} \xi_1 h(\xi_1) \\ \vdots \\ \xi_{2k+1} h(\xi_{2k+1}) \end{pmatrix}. \tag{7.11}$$

Note that the matrix on the left is invertible (as $\{f_1, g_1, \ldots, f_k, g_k, h\}$ is independent and the ξ_i are distinct) and real (as the ξ_i are real). Consequently, the matrix on the left has a real inverse. Note also that the vector on the right is real. We deduce that there exist $x, y \in \mathbb{R}^k$ (and thus $\alpha \in \mathbb{C}^k$) and $z \in \mathbb{R}$ such that the eigenvalues of $\tilde{A}_\delta^{-1}\tilde{B}_\delta$ are real and simple.

Case 2. In case 2, $S_m = \begin{pmatrix} & & F_{n_m} \\ & 0 & \\ F_{n_m} & & \end{pmatrix}$ and $T_m = G_{2n_m+1}$. Set

$$\tilde{A}_\delta = \left(\begin{array}{c|c|c|c|c|c} \begin{smallmatrix} & 1 \\ 1 & \end{smallmatrix} & & & & & \\ \hline & \ddots & & & & \\ \hline & & \begin{smallmatrix} & 1 \\ 1 & \end{smallmatrix} & & & \\ \hline & & & & & F_{n_m} \\ \hline & & & & \delta & \\ \hline & & & F_{n_m} & & \end{array}\right), \text{ and}$$

$$\tilde{B}_\delta = \left(\begin{array}{c|c|c|c|c|c} \begin{matrix} \mathrm{Im}(\lambda_1) & \mathrm{Re}(\lambda_1) \\ \mathrm{Re}(\lambda_1) & -\mathrm{Im}(\lambda_1) \end{matrix} & & & & \begin{matrix} \sqrt{\delta}\,\mathrm{Re}(\alpha_1) \\ \sqrt{\delta}\,\mathrm{Im}(\alpha_1) \end{matrix} & \\ \hline & \ddots & & & \vdots & \\ \hline & & \begin{matrix} \mathrm{Im}(\lambda_k) & \mathrm{Re}(\lambda_k) \\ \mathrm{Re}(\lambda_k) & -\mathrm{Im}(\lambda_k) \end{matrix} & & \begin{matrix} \sqrt{\delta}\,\mathrm{Re}(\alpha_k) \\ \sqrt{\delta}\,\mathrm{Im}(\alpha_k) \end{matrix} & \\ \hline & & & & & G_{n_m} \\ \hline \sqrt{\delta}\,\mathrm{Re}(\alpha_1) \; \sqrt{\delta}\,\mathrm{Im}(\alpha_1) & \cdots & \sqrt{\delta}\,\mathrm{Re}(\alpha_k) \; \sqrt{\delta}\,\mathrm{Im}(\alpha_k) & & \delta z & e_1^* \\ \hline & & & G_{n_m} & e_1 & \end{array}\right) \tag{7.12}$$

for some $\alpha \in \mathbb{C}^k$, $z \in \mathbb{R}$, and $\delta > 0$ to be chosen later. The eigenvalues of $\tilde{A}_\delta^{-1}\tilde{B}_\delta$ are equal to the eigenvalues of

$$\left(\begin{array}{c|c|c|ccc} J & & & & & \\ \hline & \ddots & & & & \\ \hline & & J & & & \\ \hline & & & I_{n_m}/\sqrt{\delta} & & \\ & & & & 1/\sqrt{\delta} & \\ & & & & & \sqrt{\delta}I_{n_m} \end{array}\right)^{-1} \tilde{A}_\delta^{-1}\tilde{B}_\delta \left(\begin{array}{c|c|c|ccc} J & & & & & \\ \hline & \ddots & & & & \\ \hline & & J & & & \\ \hline & & & I_{n_m}/\sqrt{\delta} & & \\ & & & & 1/\sqrt{\delta} & \\ & & & & & \sqrt{\delta}I_{n_m} \end{array}\right)$$

$$= \left(\begin{array}{cc|c|cc|c|c|c} \lambda_1 & & & & & & \alpha_1^*/\sqrt{2} & \\ & \lambda_1^* & & & & & \alpha_1/\sqrt{2} & \\ \hline & & \ddots & & & & \vdots & \\ \hline & & & \lambda_k & & & \alpha_k^*/\sqrt{2} & \\ & & & & \lambda_k^* & & \alpha_k/\sqrt{2} & \\ \hline & & & & & F_{n_m}G_{n_m} & e_{n_m} & \\ \hline -(\alpha_1\mathrm{i})^*/\sqrt{2} & -\alpha_1\mathrm{i}/\sqrt{2} & \cdots & -(\alpha_k\mathrm{i})^*/\sqrt{2} & -\alpha_k\mathrm{i}/\sqrt{2} & & z & e_1^\intercal \\ \hline & & & & & & & F_{n_m}G_{n_m} \end{array}\right).$$

The characteristic polynomial (in ξ) of this latter matrix is

$$\xi^{2n_m}\Bigg((z-\xi)\prod_{i=1}^{k}(\lambda_i-\xi)(\lambda_i^*-\xi) \\ + \sum_{i=1}^{k}\Big(\mathrm{Im}(\alpha_i^2)\xi - \mathrm{Im}(\alpha_i^2\lambda_i)\Big)\prod_{j\neq i}(\lambda_j-\xi)(\lambda_j^*-\xi)\Bigg) \tag{7.13}$$

and is independent of $\delta > 0$. As in Case 1 (cf. (7.8)), we may pick $\alpha \in \mathbb{C}^k$ and $z \in \mathbb{R}$ such that $\tilde{A}_\delta^{-1}\tilde{B}_\delta$ has real (but no longer necessarily simple) eigenvalues. Finally, applying Theorem 33, we deduce that for all $\delta > 0$, $\left\{\tilde{A}_\delta, \tilde{B}_\delta\right\}$ is ASDC. We conclude that $\{A, B\}$ is ASDC.

Case 3. In the final case, we have that $m = m_3 + m_4 = 1$. If $m_4 = 1$ (so that $A = B = 0$), it is clear that $\{A, B\}$ is actually SDC. Finally, suppose $m_3 = 1$ so that

$$A = \begin{pmatrix} & & F_{n_m} \\ & 0 & \\ F_{n_m} & & \end{pmatrix}, \qquad B = G_{2n_m+1}.$$

Then for $\delta \neq 0$, set

$$\tilde{A}_\delta = \begin{pmatrix} & & F_{n_m} \\ & \delta & \\ F_{n_m} & & \end{pmatrix}.$$

Note that $\tilde{A}^{-1}B$ is upper triangular with all diagonal entries equal to zero. Then applying Theorem 33, we deduce that for all $\delta \neq 0$, $\left\{\tilde{A}_\delta, B\right\}$ is ASDC. We conclude that $\{A, B\}$ is ASDC. ■

Corollary 29. *Let $\mathcal{A} = \{A_1, \dots, A_m\}$ in (7.1) and suppose $\operatorname{span}(\mathcal{A}) = \operatorname{span}\{A, B\}$ is singular. Then for any $\epsilon > 0$, there exist $\left\|\tilde{A}_i - A_i\right\| \leq \epsilon$ such that*

$$\inf_{x\in\mathbb{R}^n} \left\{ x^\intercal \tilde{A}_1 x + 2b_1^\intercal x + c_1 : \begin{array}{l} x^\intercal \tilde{A}_i x + 2b_i^\intercal x + c_i \;\square_i\; 0,\ \forall i \in [2, m] \\ x \in \mathcal{L} \end{array} \right\} \tag{7.14}$$

is a diagonalizable QCQP. The matrices $\tilde{A}_i$ and an invertible matrix P diagonalizing (7.14) can be computed via the construction in Theorem 34.

7.4 The ASDC property of nonsingular symmetric triples

In this section, we will prove the following characterization of the ASDC property for nonsingular triples of symmetric matrices (henceforth, *symmetric triples*).

Theorem 35. *Let $\{A, B, C\} \subseteq \mathbb{S}^n$ and suppose A is invertible. Then, $\{A, B, C\}$ is ASDC if and only if $\{A^{-1}B, A^{-1}C\}$ are a pair of commuting matrices with real eigenvalues.*

As always, the forward direction follows trivially from Proposition 21 and continuity. For the reverse direction, we will extend an inductive argument due to Motzkin and Taussky [127] to show that we may repeatedly perturb either $A^{-1}B$ or $A^{-1}C$ to increase the number of simple eigenvalues. In contrast to the original argument in [127], which establishes that any commuting pair $\{S, T\} \subseteq \mathbb{C}^{n\times n}$ is almost simultaneously diagonalizable *via similarity* (and thus only needs to inductively maintain commutativity of S and T), for our proof we will further need to maintain that A, B, C are symmetric matrices and that $A^{-1}B$ and $A^{-1}C$ have real eigenvalues.

Our proof will require two technical facts about block matrices consisting of upper triangular Toeplitz blocks. We present these facts below and defer their proofs to Section G.2.

Definition 30. $T \in \mathbb{R}^{n_i \times n_j}$ is an *upper triangular Toeplitz matrix* if T is of the form

$$T = \left(\begin{array}{c|cccc} & t^{(1)} & t^{(2)} & \cdots & t^{(n_i)} \\ 0_{n_i\times(n_j-n_i)} & & t^{(1)} & \ddots & \vdots \\ & & & \ddots & t^{(2)} \\ & & & & t^{(1)} \end{array}\right) \quad \text{or} \quad T = \left(\begin{array}{cccc} t^{(1)} & t^{(2)} & \cdots & t^{(n_j)} \\ & t^{(1)} & \ddots & \vdots \\ & & \ddots & t^{(2)} \\ & & & t^{(1)} \\ \hline \multicolumn{4}{c}{0_{(n_i-n_j)\times n_j}} \end{array}\right)$$

if $n_i \leq n_j$ and $n_j \leq n_i$ respectively. □

Definition 31. Let $(n_1, \dots, n_k)$ such that $\sum_i n_i = n$. Let $\mathbb{T}(n_1, \dots, n_k) \subseteq \mathbb{R}^{n\times n}$ denote the linear subspace of matrices T such that each block $T_{i,j}$ (when the rows and columns of T are partitioned according to $(n_1, \dots, n_k)$) is an upper triangular Toeplitz matrix. When the partition $(n_1, \dots, n_k)$ is clear from context, we will simply write $\mathbb{T}$. □

The following fact characterizes the set of matrices which commute with a nilpotent Jordan chain (see for example [168, Theorem 6]).

Lemma 80. *Let $(n_1, \dots, n_k)$ such that $\sum_i n_i = n$. Let $\mathcal{J} \in \mathbb{R}^{n\times n}$ be a block diagonal matrix with diagonal block $\mathcal{J}_{i,i} = F_{n_i} G_{n_i}$, i.e., a nilpotent Jordan block of size n_i. Then, $T \in \mathbb{R}^{n\times n}$ commutes with $\mathcal{J}$ if and only if $T \in \mathbb{T}$.*

Definition 32. Let $(n_1, \dots, n_k)$ such that $\sum_i n_i = n$. Define the linear map $\Pi_{(n_1,\dots,n_k)} : \mathbb{T}(n_1, \dots, n_k) \to \mathbb{R}^{k\times k}$ by

$$(\Pi_{(n_1,\dots,n_k)}(T))_{i,j} = \begin{cases} T^{(1)}_{i,j} & \text{if } n_i = n_j, \\ 0 & \text{else.} \end{cases}$$

When the partition $(n_1, \dots, n_k)$ is clear from context, we will simply write Π. □

The following fact follows from the observation that the characteristic polynomial of a matrix $T \in \mathbb{T}$ depends on only a few of its entries (see Lemma 103).

Lemma 81. *Let $(n_1, \dots, n_k)$ such that $\sum_i n_i = n$. Then, for any $T \in \mathbb{T}$, the matrices $T \in \mathbb{R}^{n\times n}$ and $\Pi(T) \in \mathbb{R}^{k\times k}$ have the same eigenvalues.*

We are now ready to prove Theorem 35.

Proof of Theorem 35. It suffices to show that if $\{A^{-1}B, A^{-1}C\}$ are a pair of commuting matrices with real eigenvalues, then $\{A, B, C\}$ is ASDC. Note that any set $\{A, B, C\} \subseteq \mathbb{S}^1$ is SDC. Thus, we may assume that $n \geq 2$ and that the statement is true inductively for all smaller n.

We make the following simplifying assumptions: Without loss of generality, we may assume that either $A^{-1}B$ has multiple eigenvalues or that $A^{-1}B$ and $A^{-1}C$ are both nilpotent. Indeed, if $A^{-1}B$ and $A^{-1}C$ both have a single eigenvalue, then we may consider the basis $\{A, B + \lambda_B A, C + \lambda_C A\}$ for span$\{A, B, C\}$ where $A^{-1}(B + \lambda_B A) = A^{-1}B + \lambda_B I$ and $A^{-1}(C + \lambda_C A) = A^{-1}C + \lambda_C I$ are both nilpotent. We will work in the basis for $\mathbb{R}^n$ furnished by Proposition 23 so that $A^{-1}B$ is in Jordan canonical form (note that $m_2 = m_3 = m_4 = 0$ by the assumptions that A is invertible and $A^{-1}B$ has real eigenvalues) and assume that the blocks of $A^{-1}B$ are ordered first according to increasing eigenvalue then increasing block sizes.

We will break our proof into four cases: First, we will consider where $A^{-1}B$ has multiple eigenvalues. The remaining three cases will consider when the Jordan block structure of $A^{-1}B$ has: multiple block sizes, multiple blocks of the same size, and a single block.

Case 1. Suppose $A^{-1}B$ has ℓ-many distinct eigenvalues. Write C as an $\ell \times \ell$ block matrix according to the partition induced by the eigenvalues of $A^{-1}B$. Then, as $A^{-1}C$ and $A^{-1}B$ commute, we have that $A^{-1}C$ (perforce C) is block diagonal. Thus, according to the block structure induced by the eigenvalues of $A^{-1}B$, the matrices A, B, C are jointly block diagonal, with each diagonal block satisfying the conditions of the inductive hypothesis. We conclude that $\{A, B, C\}$ is ASDC.

Case 2. Suppose $A^{-1}B$ and $A^{-1}C$ are nilpotent and that $A^{-1}B$ has distinct block sizes. For concreteness, suppose $A^{-1}B$ has k blocks of size $\eta = n_1 = \cdots = n_k < n_{k+1} \leq \cdots \leq n_m$. By Proposition 23,

$$A = \mathrm{Diag}(\sigma_1 F_\eta, \dots, \sigma_k F_\eta, \sigma_{k+1} F_{n_{k+1}}, \dots, \sigma_m F_{n_m})$$

for some $\sigma_i \in \{\pm 1\}$. Set

$$\tilde{C} = C + \delta\, \mathrm{Diag}(\sigma_1 F_\eta, \ldots, \sigma_k F_\eta, 0_{n_{k+1}}, \ldots, 0_{n_m}). \tag{7.15}$$

Applying Lemma 80, we have that $A^{-1}\tilde{C}$ commutes with $A^{-1}B$ and that $\tilde{C} \in \mathbb{S}^n$. Let Π denote the linear map furnished by Lemma 81. As $n_i \neq n_j$ for all $i \leq k$ and $j \geq k+1$, we have that $\Pi(A^{-1}C)$ can be written as a block diagonal matrix

$$\Pi(A^{-1}C) = \begin{pmatrix} \Pi(A^{-1}C)_{1,1} & \\ & \Pi(A^{-1}C)_{2,2} \end{pmatrix}$$

with blocks of size $k \times k$ and $(m-k) \times (m-k)$ respectively. As Π preserves eigenvalues for inputs in $\mathbb{T}$, we have that $\Pi(A^{-1}C)_{1,1}$ and $\Pi(A^{-1}C)_{2,2}$ are both nilpotent. Then, as $A^{-1}\tilde{C}$ has the same eigenvalues as

$$\Pi(A^{-1}\tilde{C}) = \begin{pmatrix} \Pi(A^{-1}C)_{1,1} + \delta I_k & \\ & \Pi(A^{-1}C)_{2,2} \end{pmatrix},$$

we deduce that $A^{-1}\tilde{C}$ has eigenvalues $\{0, \delta\}$. We have reduced to case (1) whence $\left\{A, B, \tilde{C}\right\}$ is ASDC. We conclude that $\{A, B, C\}$ is ASDC.

Case 3. Suppose $A^{-1}B$ and $A^{-1}C$ are nilpotent and that $A^{-1}B$ has Jordan blocks all of the same dimension. For concreteness, suppose $A^{-1}B$ has $m \geq 2$ Jordan blocks of dimension η. In this case Proposition 23 states that

$$A = \mathrm{Diag}(\sigma_1, \ldots, \sigma_m) \otimes F_\eta \quad \text{and} \quad B = \mathrm{Diag}(\sigma_1, \ldots, \sigma_m) \otimes G_\eta$$

where $\sigma_i \in \{\pm 1\}$. Write C as a $m \times m$ block matrix with blocks $C_{i,j} \in \mathbb{R}^{\eta \times \eta}$. By Lemma 80, $A^{-1}C \in \mathbb{T}$ and we may write

$$C_{i,j} = F_\eta \left(\gamma_{i,j}^{(1)} I_\eta + \sum_{\ell=2}^{\eta} \gamma_{i,j}^{(\ell)} (F_\eta G_\eta)^{\ell-1} \right).$$

Let Π denote the linear map furnished by Lemma 81. Let

$$\bar{A} = \mathrm{Diag}(\sigma_1, \ldots, \sigma_m) \quad \text{and} \quad \bar{C} = \left(\gamma_{i,j}^{(1)}\right). \tag{7.16}$$

Note that as $C \in \mathbb{S}^n$, we have $\gamma_{i,j}^{(1)} = \gamma_{j,i}^{(1)}$, whence $\bar{A}, \bar{C} \in \mathbb{S}^m$. As Π preserves the eigenvalues for inputs in $\mathbb{T}$ and $\bar{A}^{-1}\bar{C} = \Pi(A^{-1}C)$, we deduce that $\bar{A}^{-1}\bar{C}$ has real eigenvalues (in fact, the single eigenvalue 0). Then applying Lemma 79, there exists $\bar{C}' \in \mathbb{S}^m$ such that $\left\|\bar{C} - \bar{C}'\right\| \leq \delta$ and $\bar{A}^{-1}\bar{C}'$ has m-many distinct real eigenvalues. Finally, set

$$\tilde{C} = C + (\bar{C}' - \bar{C}) \otimes F_\eta.$$

Then Lemma 80 implies that $A^{-1}B$ and $A^{-1}\tilde{C}$ commute. Furthermore, by construction, $A^{-1}\tilde{C}$ has upper triangular Toeplitz blocks so that its eigenvalues are the same as the eigenvalues of $\Pi(A^{-1}\tilde{C}) = \bar{A}^{-1}\bar{C}'$. We have reduced to case (1) and $\left\{A, B, \tilde{C}\right\}$ is ASDC. We conclude that $\{A, B, C\}$ is also ASDC.

Case 4. Suppose $A^{-1}B$ and $A^{-1}C$ are nilpotent and that $A^{-1}B$ is a single Jordan block. Then, by Proposition 23, $A = \sigma F_n$ and $B = \sigma G_n$ for some $\sigma \in \{\pm 1\}$. Furthermore, by Lemma 80 and the assumption that $A^{-1}C$ is nilpotent, we may write

$$C = \sigma F_n \left(\sum_{i=2}^{n} c_i (F_n G_n)^{i-1} \right)$$

for some $c_2, \dots, c_n \in \mathbb{R}$.

If $n = 2$, then $C = c_2 \sigma G_2$. We may set

$$\tilde{B} = \sigma(G_2 + \delta H_2) \quad \text{and} \quad \tilde{C} = c_2\sigma(G_2 + \delta H_2). \tag{7.17}$$

Then, $\left\{A^{-1}\tilde{B}, A^{-1}\tilde{C}\right\}$ are a pair of commuting matrices with real simple eigenvalues.

If $n \geq 3$, set

$$\tilde{B} = B + \delta\sigma(e_1 e_n^\intercal + e_n e_1^\intercal) \quad \text{and} \quad \tilde{C} = C + \sigma(e_n \gamma^\intercal + \gamma e_n^\intercal) \tag{7.18}$$

where $\gamma \in \mathbb{R}^n$ is defined recursively as $\gamma_n = \gamma_{n-1} = 0$ and $\gamma_i = \delta(c_{i+1} + \gamma_{i+1})$ for $i \in [n-2]$. A straightforward calculation shows that $A^{-1}\tilde{B}$ and $A^{-1}\tilde{C}$ commute and both have real eigenvalues. Finally, as $A^{-1}\tilde{B}$ has distinct eigenvalues $\{0, \delta\}$, we have reduced to case (1) and $\left\{A, \tilde{B}, \tilde{C}\right\}$ is ASDC. We conclude that $\{A, B, C\}$ is also ASDC. ■

Corollary 30. *Let $\mathcal{A} = \{A_1, \dots, A_m\}$ in (7.1) and suppose* $\operatorname{span}(\mathcal{A}) = \operatorname{span}\{A, B, C\}$ *where A is invertible. Furthermore, suppose $A^{-1}B$ and $A^{-1}C$ commute and have real eigenvalues. Then, for any $\epsilon > 0$, there exist $\left\|\tilde{A}_i - A_i\right\| \leq \epsilon$ such that*

$$\inf_{x \in \mathbb{R}^n} \left\{ x^\intercal \tilde{A}_1 x + 2b_1^\intercal x + c_1 : \begin{array}{l} x^\intercal \tilde{A}_i x + 2b_i^\intercal x + c_i \ \Box_i \ 0, \ \forall i \in [2, m] \\ x \in \mathcal{L} \end{array} \right\} \tag{7.19}$$

is a diagonalizable QCQP. The matrices $\tilde{A}_i$ and an invertible matrix P diagonalizing (7.19) can be computed via the construction in Theorem 35.

7.5 Restricted SDC

In this section, we investigate a second new notion of simultaneous diagonalizability called *restricted* SDC. We will see soon that we have in fact already seen this property before in Section 7.3.

Definition 33. Let $\mathcal{A} \subseteq \mathbb{S}^n$ and $d \in \mathbb{N}$. We will say that $\mathcal{A}$ is *d-restricted SDC* (d-RSDC) if there exist matrices $\bar{A} \in \mathbb{S}^{n+d}$ containing A as its top-left $n \times n$ principal submatrix for every $A \in \mathcal{A}$ such that $\left\{\bar{A} : \ A \in \mathcal{A}\right\}$ is SDC. □

We record some simple consequences of the d-RSDC property that follow from Observation 5 and Lemma 78.

Observation 8. *Let $\mathcal{A} \subseteq \mathbb{S}^n$ and $d \in \mathbb{N}$.*

- *$\mathcal{A}$ is d-RSDC if and only if $\{A_1, \dots, A_m\}$ is d-RSDC for some basis $\{A_1, \dots, A_m\}$ of* $\operatorname{span}(\mathcal{A})$.
- *If $\mathcal{A}$ is d-RSDC, then $\mathcal{A}$ is d′-RSDC for all $d' \geq d$.*

The following lemma explains the connection between the d-RSDC property and the ASDC property.

Lemma 82. *Let $A_1, \dots, A_m \in \mathbb{S}^n$ and let $d \in \mathbb{N}$. If $\mathcal{A} = \{A_1, \dots, A_m\}$ is d-RSDC, then*

$$\mathcal{A} \oplus 0_d := \left\{\begin{pmatrix} A_i & \\ & 0_d \end{pmatrix} : \ i \in [m]\right\}$$

is ASDC. On the other hand, if $\mathcal{A} \oplus 0_d$ is ASDC, then for all $\epsilon > 0$, there exist $\tilde{A}_1, \dots, \tilde{A}_m \in \mathbb{S}^n$ such that

- *for all $i \in [m]$, the spectral norm $\left\|A_i - \tilde{A}_i\right\| \leq \epsilon$, and*
- *$\left\{\tilde{A}_1, \dots, \tilde{A}_m\right\}$ is d-RSDC.*

Proof. First, suppose $\{A_1, \dots, A_m\}$ is d-RSDC and let $\left\{\tilde{A}_1, \dots, \tilde{A}_m\right\} \subseteq \mathbb{S}^{n+d}$ denote the matrices furnished by d-RSDC. Next, let $\epsilon > 0$ and set

$$P = \begin{pmatrix} I_n & \\ & \sqrt{\epsilon} I_d \end{pmatrix}.$$

Clearly, P is invertible so that $\left\{P^\intercal \tilde{A}_i P : \ i \in [m]\right\}$ is also SDC. Then, note that

$$P^\intercal \tilde{A}_i P = P^\intercal \begin{pmatrix} A_i & (\tilde{A}_i)_{1,2} \\ (\tilde{A}_i)_{1,2}^* & (\tilde{A}_i)_{2,2} \end{pmatrix} P = \begin{pmatrix} A_i & \sqrt{\epsilon}(\tilde{A}_i)_{1,2} \\ \sqrt{\epsilon}(\tilde{A}_i)_{1,2}^* & \epsilon(\tilde{A}_i)_{2,2} \end{pmatrix}$$

so that $\mathcal{A} \oplus 0_d$ is ASDC.

Next, suppose $\mathcal{A} \oplus 0_d$ is ASDC and let $\epsilon > 0$. Then, there exist $\bar{A}_1, \dots, \bar{A}_m \in \mathbb{S}^{n+d}$ such that $\left\|\bar{A}_i - A_i \oplus 0_d\right\| \leq \epsilon$ and $\left\{\bar{A}_1, \dots, \bar{A}_m\right\}$ is SDC. Finally, note that $\left\|A_1 - (\bar{A}_1)_{1,1}\right\| \leq \epsilon$. ■

Remark 79. While the restriction of an SDC set does not necessarily result in an SDC set, there is a setting arising naturally when analyzing QCQPs in which the restriction of an SDC set is again SDC.

Algorithm 12 1-RSDC construction

```
Given $A,\ B \in \mathbb{S}^n$ such that $A$ is invertible and $A^{-1}B$ has simple eigenvalues
  1. Let $P \in \mathbb{R}^{n\times n}$ denote the invertible matrix guaranteed by [173]; this can be computed using
     an eigenvalue decomposition of $A^{-1}B$. Then $P^\intercal AP = \mathrm{Diag}(\sigma_1, \dots, \sigma_r, F_2, \dots, F_2)$ and
     $P^\intercal BP = \mathrm{Diag}(\sigma_1\mu_1, \dots, \sigma_r\mu_r, T_1, \dots, T_k)$. Here, $\sigma_1, \dots, \sigma_r \in \{\pm 1\}$, $\mu_1, \dots, \mu_r \in \mathbb{R}$
     and for $i \in [k]$, the matrix $T_i$ has the form
         $$T_i = \begin{pmatrix} \mathrm{Im}(\lambda_i) & \mathrm{Re}(\lambda_i) \\ \mathrm{Re}(\lambda_i) & -\mathrm{Im}(\lambda_i) \end{pmatrix}$$
     for some $\lambda_i \in \mathbb{C} \setminus \mathbb{R}$.
  2. Choose an arbitrary set of $2k+1$ distinct points $\xi_1, \dots, \xi_{2k+1} \in \mathbb{R}$
  3. Solve for $x,\ y \in \mathbb{R}^k$ and $z \in \mathbb{R}$ in the linear system (7.11)
  4. Let $\alpha \in \mathbb{C}^k$ solve (7.9) and define $\gamma \in \mathbb{R}^{r+2k}$ as
         $$\gamma = \begin{pmatrix} 0_{1\times k} & \mathrm{Re}(\alpha_1) & \mathrm{Im}(\alpha_1) & \dots & \mathrm{Re}(\alpha_k) & \mathrm{Im}(\alpha_k) \end{pmatrix}^\intercal.$$
  5. Return
         $$\tilde{A} = \begin{pmatrix} A & \\ & 1 \end{pmatrix}, \qquad \tilde{B} = \begin{pmatrix} B & P^{-\intercal}\gamma \\ \gamma^\intercal P^{-1} & z \end{pmatrix}.$$
```

Specifically, let $Q_1, \dots, Q_m \in \mathbb{S}^{n+1}$ where Q_i has A_i as its top-left $n \times n$ principal submatrix. Furthermore suppose that there exists a positive definite matrix in $\mathrm{span}(\{A_1, \dots, A_m\})$. Then, if $\{Q_1, \dots, Q_m, e_{n+1}e_{n+1}^\intercal\}$ is SDC, so is $\{A_1, \dots, A_m\}$. In words, if the homogenized quadratic forms in a QCQP, along with $e_{n+1}e_{n+1}^\intercal$, are SDC, then so are the original quadratic forms (under a standard "definiteness" assumption). See Section G.4 for details. □

7.5.1 1-restricted SDC

We record a recasting of Theorem 34 in terms of these new definitions.

Theorem 36. *Let $A, B \in \mathbb{S}^n$. Then for every $\epsilon > 0$, there exist $\tilde{A}, \tilde{B} \in \mathbb{S}^n$ such that $\left\| A - \tilde{A} \right\|, \left\| B - \tilde{B} \right\| \leq \epsilon$ and $\left\{ \tilde{A}, \tilde{B} \right\}$ is 1-RSDC. Furthermore, if A is invertible and $A^{-1}B$ has simple eigenvalues, then $\{A, B\}$ is itself 1-RSDC.*

Proof. The first claim follows from Theorem 34 and Lemma 82 applied to $\{A, B\} \oplus 0_1$. The second claim follows from the additional observation that if A is invertible and $A^{-1}B$ has simple eigenvalues, then the construction of Theorem 34 follows case 1 and *does not* perturb either A or B (see also Algorithm 12). ■

Corollary 31. *Let $\mathcal{A} = \{A_1, \dots, A_m\}$ in (7.1) and suppose* $\operatorname{span}(\mathcal{A}) = \{A, B\}$. *Then, for any $\epsilon > 0$, there exist $\bar{A}_i \in \mathbb{S}^{n+1}$ such that $\left\| (\bar{A}_i)_{1,1} - A_i \right\| \leq \epsilon$ and*

$$\inf_{x \in \mathbb{R}^n, w} \left\{ \begin{pmatrix} x \\ w \end{pmatrix}^\intercal \bar{A}_1 \begin{pmatrix} x \\ w \end{pmatrix} + 2b_1^\intercal x + c_1 : \begin{array}{l} \begin{pmatrix} x \\ w \end{pmatrix}^\intercal \bar{A}_i \begin{pmatrix} x \\ w \end{pmatrix} + 2b_i^\intercal x + c_i \ \square_i \ 0, \ \forall i \in [2, m] \\ x \in \mathcal{L} \\ w = 0 \end{array} \right\} \tag{7.20}$$

is a diagonalizable QCQP. If A is invertible and $A^{-1}B$ has simple eigenvalues, then $(\bar{A}_i)_{1,1}$ can be taken to be equal to A_i. The matrices $\bar{A}_i$ and an invertible matrix P diagonalizing (7.19) can be computed via Algorithms 11 and 12.

7.5.2 d-restricted SDC

Let $\{A, B\} \subseteq \mathbb{S}^n$ such that A is invertible and $A^{-1}B$ has simple eigenvalues. By Observation 8 and Theorem 36, we have that $\{A, B\}$ is d-RSDC for any $d \geq 1$. In this section, we record an alternate construction showing that $\{A, B\}$ is d-RSDC for $d \geq 1$. This alternate construction applies Algorithm 12 on smaller blocks of the canonical form and has empirically better numerical performance in QCQP applications (see Section 7.7.2).

Theorem 37. *Let $A, B \in \mathbb{S}^n$ such that A is invertible and $A^{-1}B$ has simple eigenvalues. Then, $\{A, B\}$ is d-RSDC for any $d \geq 1$.*

Proof. Without loss of generality, we may assume that A, B are in canonical form (Proposition 23) and $m_1 = 0$ (else consider the submatrix of A, B corresponding to the remaining blocks).

Partition $[m]$ into d-many contiguous subintervals. Write A and B as diagonal block matrices of diagonal block matrices according to the partition of $[m]$. In other words, write $A = \operatorname{Diag}(A_1, \dots, A_d)$ and $B = \operatorname{Diag}(B_1, \dots, B_d)$ where each A_i is a diagonal block matrix of F_2-matrices and each B_i is a diagonal block matrix with matrices of the form $\begin{pmatrix} \operatorname{Im}(\lambda) & \operatorname{Re}(\lambda) \\ \operatorname{Re}(\lambda) & -\operatorname{Im}(\lambda) \end{pmatrix}$ for $\lambda \in \mathbb{C} \setminus \mathbb{R}$. Set

$$\tilde{A} = \left(\begin{array}{c|c|c|c|c|c} A_1 & & & & & \\ \hline & \ddots & & & & \\ \hline & & A_d & & & \\ \hline & & & 1 & & \\ \hline & & & & \ddots & \\ \hline & & & & & 1 \end{array}\right) \quad \text{and} \quad \tilde{B} = \left(\begin{array}{c|c|c|c|c|c} B_1 & & & x_1 & & \\ \hline & \ddots & & & \ddots & \\ \hline & & B_d & & & x_d \\ \hline x_1^\intercal & & & z_1 & & \\ \hline & \ddots & & & \ddots & \\ \hline & & x_d^\intercal & & & z_d \end{array}\right) \tag{7.21}$$

Algorithm 13 d-RSDC construction

Given $A,\ B \in \mathbb{S}^n$ such that A is invertible and $A^{-1}B$ has simple eigenvalues

1. Let $P \in \mathbb{R}^{n\times n}$ denote the invertible matrix guaranteed by [173]; this can be computed using an eigenvalue decomposition of $A^{-1}B$. Then $P^\intercal AP = \operatorname{Diag}(\sigma_1, \dots, \sigma_r, F_2, \dots, F_2)$ and $P^\intercal BP = \operatorname{Diag}(\sigma_1\mu_1, \dots, \sigma_r\mu_r, T_1, \dots, T_k)$. Here, $\sigma_1, \dots, \sigma_r \in \{\pm 1\}$, $\mu_1, \dots, \mu_r \in \mathbb{R}$ and for $i \in [k]$, the matrix T_i has the form
$$T_i = \begin{pmatrix} \operatorname{Im}(\lambda_i) & \operatorname{Re}(\lambda_i) \\ \operatorname{Re}(\lambda_i) & -\operatorname{Im}(\lambda_i) \end{pmatrix}$$
for some $\lambda_i \in \mathbb{C} \setminus \mathbb{R}$.
2. Partition $[k] = \kappa_1 \cup \dots \kappa_d$ into contiguous subintervals where $\kappa_i = [\text{start}_i, \text{end}_i]$
3. For each $i \in [d]$, apply Algorithm 12 to get $x_i \in \mathbb{R}^{|\kappa_i|}$ and $z_i \in \mathbb{R}$ such that
$$\left(\begin{array}{ccc|c} F_2 & & & \\ & \ddots & & \\ & & F_2 & \\ \hline & & & 1 \end{array}\right) \quad \text{and} \quad \left(\begin{array}{ccc|c} T_{\text{start}_i} & & & \\ & \ddots & & x_i \\ & & T_{\text{end}_i} & \\ \hline & x_i^\intercal & & z_i \end{array}\right)$$
are SDC
4. Let $Q \coloneqq P \oplus I_1$ and return
$$Q^{-\intercal}\tilde{A}Q^{-1} \quad \text{and} \quad Q^{-\intercal}\tilde{B}Q^{-1}$$
where $\tilde{A}$ and $\tilde{B}$ are defined in (7.21).

for $z_1, \dots, z_d \in \mathbb{R}$ and vectors $x_1, \dots, x_d$ of the appropriate dimensions to be chosen later. After a simultaneous permutation of the coordinates, we can write $\tilde{A}$ and $\tilde{B}$ as diagonal block matrices with blocks of the form

$$\begin{pmatrix} A_i & \\ & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} B_i & x_i \\ x_i^\intercal & z_i \end{pmatrix}.$$

By Theorem 36 (summarized in Algorithm 12) and the assumption that $A^{-1}B$ has simple eigenvalues, we may, for each $i \in [d]$, pick $x_i \in \mathbb{R}^n$ and $z_i \in \mathbb{R}$ such that the pair of matrices above is SDC. It remains to note that the diagonal block concatenation of SDC matrices is SDC. ■

Corollary 32. *Let* $\mathcal{A} = \{A_1, \dots, A_m\}$ *in* (7.1) *and suppose* $\operatorname{span}(\mathcal{A}) = \{A, B\}$. *Then, for any* $\epsilon > 0$, *there exist* $\bar{A}_i \in \mathbb{S}^{n+d}$ *such that* $\left\|(\bar{A}_i)_{1,1} - A_i\right\| \leq \epsilon$ *and*

$$\inf_{x\in\mathbb{R}^n, w\in\mathbb{R}^d} \left\{ \begin{pmatrix} x \\ w \end{pmatrix}^\intercal \bar{A}_1 \begin{pmatrix} x \\ w \end{pmatrix} + 2b_1^\intercal x + c_1 : \begin{array}{l} \begin{pmatrix} x \\ w \end{pmatrix}^\intercal \bar{A}_i \begin{pmatrix} x \\ w \end{pmatrix} + 2b_i^\intercal x + c_i \ \square_i \ 0,\ \forall i \in [2, m] \\ x \in \mathcal{L} \\ w = 0 \end{array} \right\} \tag{7.22}$$

is a diagonalizable QCQP. If A is invertible and $A^{-1}B$ has simple eigenvalues, then $(\bar{A}_i)_{1,1}$ can be taken to be equal to A_i. The matrices $\bar{A}_i$ and an invertible matrix P diagonalizing (7.19) can be computed via Algorithms 11 and 13.

7.6 Obstructions to further generalization

In this section, we record explicit counterexamples to *a priori* plausible extensions to Theorems 33 to 35.

7.6.1 Singular symmetric triples

In Theorem 34, we showed that any singular symmetric pair is ASDC. A natural question to ask is whether any singular set of symmetric matrices (regardless of the dimension of its span) is ASDC. The following theorem presents an obstruction to generalizations in this direction. Specifically, in contrast to Theorem 34 (where it was shown that singularity implies ASDC in the context of symmetric pairs), Theorem 38 below shows that even symmetric *triples* with *"large amounts"* of singularity can fail to be ASDC.

Theorem 38. *Let $\{A = I_n, B, C\} \subseteq \mathbb{S}^n$. Then, if $d < \operatorname{rank}([B, C])/2$, the set*

$$\left\{\begin{pmatrix} A & \\ & 0_d \end{pmatrix}, \begin{pmatrix} B & \\ & 0_d \end{pmatrix}, \begin{pmatrix} C & \\ & 0_d \end{pmatrix}\right\}$$

is not ASDC.

Proof. Suppose for the sake of contradiction that this set is ASDC. Let $\epsilon \in (0, 1/2)$ and let $\left\{\tilde{A}, \tilde{B}, \tilde{C}\right\} \subseteq \mathbb{S}^{n+d}$ denote an SDC set furnished by the ASDC assumption. Without loss of generality, $\tilde{A}$ has rank $n + d$. Write

$$\tilde{A} = \begin{pmatrix} \tilde{A}_{1,1} & \tilde{A}_{1,2} \\ \tilde{A}_{1,2}^\intercal & \tilde{A}_{2,2} \end{pmatrix}.$$

Similarly decompose $\tilde{B}$ and $\tilde{C}$. As $\epsilon \in (0, 1/2)$, we have that $\tilde{A}_{1,1}$ is invertible. Let

$$P = \begin{pmatrix} \tilde{A}_{1,1}^{-1/2} & -\tilde{A}_{1,1}^{-1}\tilde{A}_{1,2} \\ 0 & I_d \end{pmatrix}.$$

Then as P is invertible, $\left\{P^\intercal \tilde{A} P, P^\intercal \tilde{B} P, P^\intercal \tilde{C} P\right\}$ is again SDC. Note that $P^\intercal \tilde{A} P$ has the form

$$P^\intercal \tilde{A} P = \begin{pmatrix} I_n & \\ & \tilde{A}_{2,2} - \tilde{A}_{1,2}^\intercal \tilde{A}_{1,1}^{-1} \tilde{A}_{1,2} \end{pmatrix}.$$

Furthermore,

$$\begin{aligned}&\left\|P^\intercal \tilde{B} P - B\right\|\\&\qquad= \left\|(P - I_{n+d})^\intercal \tilde{B}(P - I_{n+d}) + \tilde{B}(P - I_{n+d}) + (P - I_{n+d})^\intercal \tilde{B} + (\tilde{B} - B)\right\|\\&\qquad\leq \left\|\tilde{B}\right\| \|P - I_{n+d}\|^2 + 2\left\|\tilde{B}\right\| \|P - I_{n+d}\| + \epsilon.\end{aligned}$$

We claim that $\|P - I_{n+d}\|$ can be bounded in terms of ϵ:

$$\begin{aligned}\|P - I_{n+d}\| &\leq \left\|\tilde{A}_{1,1}^{-1/2} - I\right\| + \left\|\tilde{A}_{1,1}^{-1}\right\|\left\|\tilde{A}_{1,2}\right\|\\&\leq \max\left\{\frac{1}{\sqrt{1-\epsilon}} - 1,\, 1 - \frac{1}{\sqrt{1+\epsilon}}\right\} + \frac{\epsilon}{1-\epsilon}\\&\leq \frac{2\epsilon}{1-\epsilon}.\end{aligned}$$

Here, we have used the fact that $\left\|\tilde{A} - A \oplus 0_d\right\| \leq \epsilon$, so that $\left\|\tilde{A}_{1,1} - I_n\right\| \leq \epsilon$ and $\left\|\tilde{A}_{1,2}\right\| \leq \epsilon$. Consequently, as we may also bound $\left\|\tilde{B}\right\| \leq \|B\| + \epsilon$, we deduce that for any $\delta > 0$, we can pick $\epsilon \in (0, 1/2)$ small enough such that $\left\|P^\intercal \tilde{B} P - B\right\| \leq \delta$. An identical calculation holds for $\left\|P^\intercal \tilde{C} P - C\right\|$. We conclude that for all $\delta > 0$, there exist $\bar{A}, \bar{B}, \bar{C}$ of the form

$$\bar{A} = \begin{pmatrix} I_n & \\ & \bar{A}_{2,2}\end{pmatrix}, \quad \bar{B} = \begin{pmatrix} \bar{B}_{1,1} & \bar{B}_{1,2}\\ \bar{B}_{1,2}^\intercal & \bar{B}_{2,2}\end{pmatrix}, \quad \bar{C} = \begin{pmatrix} \bar{C}_{1,1} & \bar{C}_{1,2}\\ \bar{C}_{1,2}^\intercal & \bar{C}_{2,2}\end{pmatrix}$$

such that $\left\{\bar{A}, \bar{B}, \bar{C}\right\}$ is SDC, $\left\|A - \bar{A}\right\|, \left\|B - \bar{B}\right\|, \left\|C - \bar{C}\right\| \leq \delta$, and $\bar{A}_{2,2}$ is invertible. Then by Proposition 21, the top-left block of the commutator $[\bar{A}^{-1}\bar{B}, \bar{A}^{-1}\bar{C}]$ is equal to 0_n. Expanding this top-left block, we deduce

$$[\bar{B}_{1,1}, \bar{C}_{1,1}] = \bar{C}_{1,2}\bar{A}_{2,2}^{-1}\bar{B}_{1,2}^\intercal - \bar{B}_{1,2}\bar{A}_{2,2}^{-1}\bar{C}_{1,2}^\intercal. \tag{7.23}$$

Finally, by lower semi-continuity of rank, we have $\operatorname{rank}([\bar{B}_{1,1}, \bar{C}_{1,1}]) \geq \operatorname{rank}([B, C])$ for all $\delta > 0$ small enough. This is a contradiction as the expression on the right of (7.23) has rank at most $2d < \operatorname{rank}([B, C])$. ■

This same construction can be viewed as an obstruction to generalizations of Theorem 36 to symmetric triples with constant d.

Corollary 33. *Let $\{A = I_n, B, C\} \subseteq \mathbb{S}^n$. Then $A^{-1}B$ and $A^{-1}C$ are both diagonalizable with real eigenvalues and $\{A, B, C\}$ is not d-RSDC for any $d < \operatorname{rank}([B, C])/2$.*

Remark 80. Note that for all $n \in \mathbb{N}$, there exist $B,\ C \in \mathbb{S}^{2n}$ such that $\operatorname{rank}([B, C]) = 2n$. For example, set

$$B = \begin{pmatrix} I_n & \\ & -I_n \end{pmatrix}, \qquad C = \begin{pmatrix} & I_n \\ I_n & \end{pmatrix}.$$

Then, $\{A = I_{2n}, B, C\} \subseteq \mathbb{S}^{2n}$ is a nonsingular symmetric triple such that $A^{-1}B$ and $A^{-1}C$ are both diagonalizable. On the other hand, Corollary 33 and Theorem 38 imply that

$$\left\{ \begin{pmatrix} A & \\ & 0_{n-1} \end{pmatrix}, \begin{pmatrix} B & \\ & 0_{n-1} \end{pmatrix}, \begin{pmatrix} C & \\ & 0_{n-1} \end{pmatrix} \right\}$$

is not ASDC and $\{A, B, C\}$ is not $(n-1)$-RSDC. □

7.6.2 Nonsingular septuples

We may reinterpret Theorems 33 and 35 as saying that if $\mathcal{A}$ satisfies $\dim(\operatorname{span}(\mathcal{A})) \leq 3$ and contains an invertible matrix S, then $\mathcal{A}$ is ASDC if and only if $S^{-1}\mathcal{A}$ consists of a set of commuting matrices with real eigenvalues. A natural question to ask is whether the same statement holds without any assumption on the dimension of the span of $\mathcal{A}$. Theorem 39 below presents an obstruction to generalizations in this direction. Specifically, Theorem 39 constructs a non-ASDC set $\mathcal{A} = \{A_1, \dots, A_7\} \subseteq \mathbb{S}^6$ where A_1 is invertible and $A_1^{-1}\mathcal{A}$ consists of a set of commuting matrices with real eigenvalues.

The following lemma adapts a technique introduced by O'meara and Vinsonhaler [137] for studying the almost simultaneously diagonalizable via similarity property of subsets of $\mathbb{C}^{n\times n}$.

Lemma 83. *Let $\mathcal{A} = \{A_1, \dots, A_m\} \subseteq \mathbb{S}^n$ where $A_1 \in \mathcal{A}$ is invertible. If $\mathcal{A}$ is SDC, then $\dim(\mathbb{R}[A_1^{-1}\mathcal{A}]) \leq n$. Here, $\mathbb{R}[A_1^{-1}\mathcal{A}]$ is the real algebra generated by $A_1^{-1}\mathcal{A}$.*

Proof. Let P denote the invertible matrix furnished by SDC and suppose $P^\intercal A_i P = D_i$. Then,

$$\dim\Big(\mathbb{R}\Big[A_1^{-1}\mathcal{A}\Big]\Big) = \dim\Big(\mathbb{R}\Big[\Big\{D_1^{-1}D_i : i \in [m]\Big\}\Big]\Big) \leq n. \qquad \blacksquare$$

The following corollary then follows by lower semi-continuity.

Corollary 34. *Let $\mathcal{A} = \{A_1, \dots, A_m\} \subseteq \mathbb{S}^n$ where $A_1 \in \mathcal{A}$ is invertible. If $\mathcal{A}$ is ASDC, then $\dim(\mathbb{R}[A_1^{-1}\mathcal{A}]) \leq n$. Here, $\mathbb{R}[A_1^{-1}\mathcal{A}]$ is the real algebra generated by $A_1^{-1}\mathcal{A}$.*

Theorem 39. *There exists a set $\mathcal{A} = \{A_1, \dots, A_7\} \subseteq \mathbb{S}^6$ such that A_1 is invertible, $A_1^{-1}\mathcal{A}$ is a set of commuting matrices with real eigenvalues, and $\mathcal{A}$ is* not *ASDC.*

Remark 81. The analogous example in the complex setting states that there exists a set $\mathcal{A} = \{A_1, \dots, A_5\} \subseteq \mathbb{H}^4$ such that A_1 is invertible, $A_1^{-1}\mathcal{A}$ is a set of commuting matrices with real eigenvalues, and $\mathcal{A}$ is not ASDC. See Section G.3. □

Proof. Set

$$A_1 = \begin{pmatrix} &&&&& 1\\ &&&& 1 &\\ &&& 1 &&\\ && 1 &&&\\ & 1 &&&&\\ 1 &&&&& \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 &&&&&\\ & 0 &&&&\\ && 0 &&&\\ &&& 1 &&\\ &&&& 0 &\\ &&&&& 0 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 0 &&&&&\\ & 0 &&&&\\ && 0 &&&\\ &&&& 1 &\\ &&& 1 &&\\ &&&&& 0 \end{pmatrix},$$

$$A_4 = \begin{pmatrix} 0 &&&&&\\ & 0 &&&&\\ && 0 &&&\\ &&&&& 1\\ &&&& 0 &\\ &&& 1 && \end{pmatrix}, \quad A_5 = \begin{pmatrix} 0 &&&&&\\ & 0 &&&&\\ && 0 &&&\\ &&& 0 &&\\ &&&& 1 &\\ &&&&& 0 \end{pmatrix},$$

$$A_6 = \begin{pmatrix} 0 &&&&&\\ & 0 &&&&\\ && 0 &&&\\ &&& 0 &&\\ &&&&& 1\\ &&&& 1 & \end{pmatrix}, \quad A_7 = \begin{pmatrix} 0 &&&&&\\ & 0 &&&&\\ && 0 &&&\\ &&& 0 &&\\ &&&& 0 &\\ &&&&& 1 \end{pmatrix}.$$

Note that A_1 is invertible. It is not hard to verify that $A_1^{-1}\mathcal{A}$ forms a set of commuting matrices with real eigenvalues. On the other hand, note that

$$\mathbb{R}[A_1^{-1}\mathcal{A}] = \left\{ \begin{pmatrix} a &&& d & f & g\\ & a && c & e & f\\ && a & b & c & d\\ &&& a &&\\ &&&& a &\\ &&&&& a \end{pmatrix} : \ a, b, c, d, e, f, g \in \mathbb{R} \right\}$$

so that $\dim(\mathbb{R}[A_1^{-1}\mathcal{A}]) = 7 > 6 = n$. We deduce from Corollary 34 that $\mathcal{A}$ is *not* ASDC. ■

7.7 Applications to QCQPs

In this section, we suggest several applications of diagonalization to optimizing QCQPs. We begin by proving properties regarding the SDP and SOCP relaxations of diagonal QCQPs with bound constraints. Note that QCQPs with bound constraints are the main problems of interest within spatial branch and bound (BB) schemes for QCQPs. These results give heuristic reasons why one might expect the SOCP relaxation to give strong yet efficiently computable lower bounds within BB schemes. These results serve as additional motivation for the ASDC and d-RSDC properties. We then examine these assertions numerically with preliminary computational experiments.

7.7.1 The SOCP relaxation of a diagonal QCQP with bound constraints

Consider solving a QCQP over $\mathbb{R}^n$ of the form (7.1) within a BB scheme. At each node of the BB tree, we encounter the original QCQP with additional bound constraints,

$$\inf_{x\in\mathbb{R}^n} \left\{ x^\intercal A_1 x + 2b_1^\intercal x + c_1 : \begin{array}{l} x^\intercal A_i x + 2b_i^\intercal x + c_i \ \square_i \ 0, \ \forall i \in [2, m]\\ x \in \mathcal{L}\\ x \in [\ell, u] \end{array} \right\}, \tag{7.24}$$

and desire to construct and solve strong convex relaxations of (7.24).

One powerful method for constructing such convex relaxations combines the reformulation-linearization technique with semidefinite programming [6]. We begin by linearizing (7.24) using the variable $Y = xx^\intercal$. Specifically, replace $x^\intercal A_i x$ with $\langle A_i, Y\rangle$ and include the constraint

$Y = xx^\intercal$. Then, relax the pair of constraints $x \in [\ell, u]$ and $Y = xx^\intercal$ to the constraint that (x, Y) belong to the set

$$\mathcal{S}_{\text{SDP}} := \left\{ (x,Y) \in \mathbb{R}^n \times \mathbb{S}^n : \begin{array}{l} Y_{i,j} \geq \ell_j x_i + \ell_i x_j - \ell_j \ell_j,\, \forall i,j \\ Y_{i,j} \leq u_j x_i + \ell_i x_j - u_j \ell_i,\, \forall i,j \\ Y_{i,j} \geq u_j x_i + u_i x_j - u_i u_j,\, \forall i,j \\ Y \succeq xx^\intercal \end{array} \right\}.$$

The SDP+RLT relaxation then reads

$$\inf_{x \in \mathbb{R}^n, Y \in \mathbb{S}^n} \left\{ \langle A_1, Y\rangle + 2b_1^\intercal x + c_1 : \begin{array}{l} \langle A_i, Y\rangle + 2b_i^\intercal x + c_i \ \Box_i\ 0,\, \forall i \in [2,m] \\ x \in \mathcal{L} \\ (x,Y) \in \mathcal{S}_{\text{SDP}} \end{array} \right\}. \quad (7.25)$$

Note that for diagonal QCQPs (i.e., the setting where $A_1, \dots, A_m$ are diagonal) that $\langle A_i, Y\rangle = \operatorname{diag}(A_i)^\intercal \operatorname{diag}(Y)$ so that the SDP+RLT relaxation does not depend on the off-diagonal entries of Y. In particular, we may replace the variable $Y \in \mathbb{S}^n$ with a variable $y \in \mathbb{R}^n$ representing its diagonal entries. Naturally, we then replace the term the term $\langle A_i, Y\rangle$ and the constraint $(x, Y) \in \mathcal{S}_{\text{SDP}}$ with the term $\operatorname{diag}(A_i)^\intercal y$ and the constraint

$$(x,y) \in \mathcal{S}_{\text{SOCP}} := \{(x, \operatorname{diag}(Y)) : (x,Y) \in \mathcal{S}_{\text{SDP}}\}.$$

The following lemma shows that $\mathcal{S}_{\text{SOCP}}$ is SOC-representable so that the resulting relaxation is in fact an SOCP. Thus, the SDP+RLT relaxation of a QCQP with bound constraints can be solved substantially faster when $A_1, \dots, A_m$ are diagonal.

In the remainder of this section, let $\circ$ denote the elementwise product between two vectors.

Lemma 84. *The following identities hold*

$$\begin{aligned} \mathcal{S}_{\text{SOCP}} &:= \{(x, \operatorname{diag}(Y)) : (x,Y) \in \mathcal{S}_{\text{SDP}}\} \\ &= \left\{ (x,y) \in \mathbb{R}^n \times \mathbb{R}^n : \ x \circ x \leq y \leq (u+\ell) \circ x - u \circ \ell \ \right\} \\ &= \operatorname{conv}\left\{ (x,y) \in \mathbb{R}^n \times \mathbb{R}^n : \begin{array}{c} x \in [\ell, u] \\ x \circ x = y \end{array} \right\}. \end{aligned}$$

In particular, $\mathcal{S}_{\text{SOCP}}$ is SOC-representable.

Proof. For notational convenience, let $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$ denote the three sets on the right in order. Note $\mathcal{S}_{\text{SOCP}} = \mathcal{S}_1$ by definition. We will show $\mathcal{S}_1 \subseteq \mathcal{S}_2 = \mathcal{S}_3 \subseteq \mathcal{S}_1$.

The containment $\mathcal{S}_1 \subseteq \mathcal{S}_2$ follows by definition: Given $(x, Y) \in \mathcal{S}_{\text{SDP}}$, we have that $\operatorname{diag}(Y) \geq x \circ x$ by the SDP constraint and $\operatorname{diag}(Y) \leq (u + \ell) \circ x - u \circ \ell$ by the RLT constraints.

The identity $\mathcal{S}_2 = \mathcal{S}_3$ follows the well-known (and easy to verify) fact that in one-dimension

$$\left\{ (x_i, y_i) \in \mathbb{R}^2 : \begin{array}{l} y_i \geq x_i^2 \\ y_i \leq (u_i + \ell_i)x_i - u_i \ell_i \end{array} \right\} = \operatorname{conv}\left\{ (x_i, y_i) \in \mathbb{R}^2 : \begin{array}{c} x_i \in [\ell_i, u_i] \\ x_i^2 = y_i \end{array} \right\}$$

and the fact that direct products commute with convex hulls.

Finally, suppose $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$ satisfies $x \in [\ell, u]$ and $y = x \circ x$. Set $Y = xx^\intercal$ so that $\operatorname{diag}(Y) = y$. It is straightforward to show that $(x, Y) \in \mathcal{S}_{\mathrm{SDP}}$ so that $(x, y) \in \mathcal{S}_1$. Then as $\mathcal{S}_1$ is convex, we have that $\mathcal{S}_3 \subseteq \mathcal{S}_1$. ■

The following corollary shows how to construct optimizers of (7.25) with bounded rank when $A_1, \dots, A_m$ are diagonal. The bound depends only on m and the complexity of $\mathcal{L} \cap [\ell, u]$. This gives a heuristic reason why one would expect the SDP+RLT relaxation (and hence the SOCP+RLT relaxation) of a diagonal QCQP with bound constraints to be stronger than that of a general QCQP with bound constraints, especially when m is small and $\mathcal{L}$ is simple.

Lemma 85. *Suppose $A_1, \dots, A_m$ are diagonal and that $\mathcal{L} \cap [\ell, u]$ can be expressed as the intersection of $[\ell, u]$ with k additional linear (in)equality constraints. If (7.25) has a solution, then it has a solution (x^*, Y^*) such that*

$$\operatorname{rank}\begin{pmatrix} Y^* & x^* \\ x^{*\intercal} & 1 \end{pmatrix} \leq m + k.$$

Proof. By assumption there exists an affine function $L : \mathbb{R}^n \to \mathbb{R}^k$ such that

$$[\ell, u] \cap \mathcal{L} = [\ell, u] \cap \{x \in \mathbb{R}^n : \ L(x)_i \ \square_i \ 0, \ \forall i \in [k]\}$$

where $\square_i \ \in \{\leq, =\}$. Define $Q : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$ by

$$Q(x, y)_i = \langle \operatorname{diag}(A_i), y \rangle + 2b_i^\intercal x + c_i, \ \forall i \in [m]$$

and let $\tilde{Q} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{m+k}$ denote the affine function mapping $(x, y) \mapsto (Q(x, y), L(x))$.

Let (x^*, Y^*) denote an optimizer of (7.25) and set $y^* = \operatorname{diag}(Y^*)$. By Lemma 84, there exist points $x^{(i)} \in [\ell, u]$ and convex combination weights $\alpha_i > 0$ such that $(x^*, y^*) = \sum_i \alpha_i (x^{(i)}, x^{(i)} \circ x^{(i)})$. Then, by linearity, we have $\tilde{Q}(x^*, y^*) = \sum_i \alpha_i \tilde{Q}(x^{(i)}, x^{(i)} \circ x^{(i)})$.

We claim that $\left\{\tilde{Q}(x^{(i)}, x^{(i)} \circ x^{(i)})\right\}_i$ span an affine subspace of dimension $< m + k$. Indeed, supposing otherwise, $\tilde{Q}(x^*, y^*)$ is in the interior of $\operatorname{conv}\left\{\tilde{Q}(x^{(i)}, x^{(i)} \circ x^{(i)})\right\}_i$. Thus, there exists $(x', y') \in \mathcal{S}_{\mathrm{SOCP}}$ achieving $\tilde{Q}(x', y') = \tilde{Q}(x^*, y^*) - \epsilon e_1$ for some $\epsilon > 0$. This contradicts optimality of (x^*, Y^*).

Applying Carathéodory's Theorem, (x^*, y^*) is a convex combination of at most $m + k$ points from $\left\{(x^{(i)}, x^{(i)} \circ x^{(i)})\right\}$, say $(x^*, y^*) = \sum_{i=1}^{m+k} \alpha_i (x^{(i)}, x^{(i)} \circ x^{(i)})$. Then,

$$\sum_{i=1}^{m+k} \alpha_i \begin{pmatrix} x^{(i)} x^{(i)\intercal} & x^{(i)} \\ x^{(i)\intercal} & 1 \end{pmatrix}$$

is an optimal solution to (7.25) with rank $\leq m + k$. ■

7.7.2 Numerical results

In this subsection, we present preliminary numerical results on diagonalization and the d-RSDC property in solving QCQP problems with one quadratic constraint and additional linear constraints. Problems in this form have received recent interest, for example in the area of optimal portfolio deleveraging [116]. Furthermore, this restricted class of QCQPs is still NP-hard in general—as mentioned in the introduction, even the problem of minimizing a general quadratic function over the hypercube is NP-hard.

We will consider random instances of the following problem

$$\min_{x\in\mathbb{R}^n}\left\{x^\intercal A_1 x : \begin{array}{l} x^\intercal A_2 x + 2b_2^\intercal x \leq 1 \\ x \in \mathcal{L} \end{array}\right\} \tag{7.26}$$

where $A_1,\ A_2 \in \mathbb{S}^n$, $b_2 \in \mathbb{R}^n$, and $\mathcal{L} \subseteq \mathbb{R}^n$ is a polytope.

Random model. We will consider a family of distributions over instances of (7.26) parameterized by $n \in \mathbb{N}$ and $k \in \mathbb{N}_0$. Here, k will parameterize the number of (pairs of) complex eigenvalues of $A_1^{-1}A_2$. Specifically, given (n, k) such that $2k \leq n$:

1. Let $r = n - 2k$.

2. Generate a random orthogonal matrix V by taking M to be a random $n \times n$ matrix with entries i.i.d. $N(0,1)$ and then taking V to be a matrix of left singular vectors of M. Let $\sigma_1, \dots, \sigma_r$ be i.i.d. Rademacher random variables. Let $\mu_1, \dots, \mu_r$ be i.i.d. $N(0,1)$ random variables. Let $x_1, \dots, x_k, y_1, \dots, y_k$ be i.i.d. $N(0,1)$ random variables. Then, set

   $$\begin{aligned} A_1 &= V^\intercal \operatorname{Diag}(\sigma_1, \dots, \sigma_r, F_2, \dots, F_2)V \\ A_2 &= V^\intercal \operatorname{Diag}(\sigma_1\mu_1, \dots, \sigma_r\mu_r, T_1, \dots, T_k)V. \end{aligned}$$

   Here, $T_i \in \mathbb{S}^2$ is the random matrix $\left(\begin{smallmatrix} x_i & y_i \\ y_i & -x_i \end{smallmatrix}\right)$.

3. Let the entries of b_2 be i.i.d. $N(0,1)$ random variables, and $L = \begin{pmatrix} I \\ -I \end{pmatrix} N$, where the entries of $N \in \mathbb{R}^{n\times n}$ are i.i.d. $N(0,1)$ random variables. This ensures that the set $\mathcal{L} := \{x : Lx \leq 1\}$ is bounded almost surely.

Note that Theorem 36 (respectively, Theorem 37) implies that $\{A_1,\ A_2\}$ is almost surely 1-RSDC (respectively, k-RSDC) in this random model.

Branch and bound methods. We use BB methods to solve different reformulations of (7.26) with and without diagonalization.

We implement two classes of BB methods. The first class, the SDP-based BB method, uses a *simplified* SDP+RLT relaxation for computing a lower bound at each node. Specifically, we lower bound the value of (7.26) with the additional box constraint $x \in [\ell, u]$ by

$$\min_{x\in\mathbb{R}^n,\, Y\in\mathbb{S}^n} \left\{ \langle A_1, Y\rangle : \begin{array}{l} \langle A_i, Y\rangle + 2b_2^\intercal x - 1 \leq 0 \\ x \in \mathcal{L} \\ x \in [\ell, u] \\ \operatorname{diag}(Y) \leq (u+\ell)\circ x - u \circ \ell \\ Y \succeq xx^\intercal \end{array} \right\}. \tag{7.27}$$

At the root node, we set $[\ell, u]$ to be coordinate-wise lower and upper bounds on $\mathcal{L}$. Note that in contrast to the *full* SDP+RLT relaxation (see (7.25)), we only impose RLT constraints on the diagonal entries of Y to strike a balance between bound quality and computational cost. This method then applies a spatial BB rule for each coordinate x_i and updates the values of $[\ell, u]$.

The second class, the SOCP-based BB methods, first diagonalize (7.26) before applying a BB scheme. The method of diagonalization differs across the different SOCP-based BB methods but the BB part is the same. Suppose we have already diagonalized (7.26) so that A_i is a diagonal matrix for each $i = 1, 2$. Write $A_i = \operatorname{Diag}(a_i^+) + \operatorname{Diag}(a_i^-)$, where a_i^+, $a_i^- \in \mathbb{R}^n$ are nonnegative and nonpositive respectively. Let $I = \operatorname{supp}(a_1^-) \cup \operatorname{supp}(a_2^-)$. The SOCP-based BB method uses the SOCP+RLT relaxation for computing a lower bound at each node. Specifically we lower bound the value of (7.26) (assuming that the A_is are diagonal) with the additional box constraint $x \in [\ell, u]$ by

$$\min_{x\in\mathbb{R}^n,\, y\in\mathbb{R}^{|I|}} \left\{ x^\intercal \operatorname{Diag}(a_1^+)x + (a_1^-)^\intercal y : \begin{array}{l} x^\intercal \operatorname{Diag}(a_2^+)x + (a_2^-)^\intercal y + 2b_2^\intercal x - 1 \leq 0 \\ x \in \mathcal{L} \\ x \in [\ell, u] \\ x_{j_i}^2 \leq y_i \leq (u_{j_i} + \ell_{j_i})x_{j_i} - u_{j_i}\ell_{j_i},\ \forall j_i \in I \end{array} \right\}. \tag{7.28}$$

Again, at the root node, we set $[\ell, u]$ to be coordinate-wise lower and upper bounds on $\mathcal{L}$. This method then applies a spatial BB rule for each coordinate x_{j_i} such that $j_i \in I$ and updates $[\ell, u]$.

In both methods, we use a successive convex approximation [119], which linearizes nonconvex terms in the quadratic objective and constraint, to attempt to construct feasible solutions and good upper bounds.

In more detail, we implemented the following five BB methods for solving instances of (7.26).

- SDPBB solves (7.26) directly using the SDP-based BB method.

- SDCBB is a solution method which can only be applied when $\{A_1, A_2\}$ is already SDC. In this case (letting P denote the corresponding invertible matrix), SDCBB reformulates (7.26) as

$$\min_{x\in\mathbb{R}^n} \left\{ x^\intercal(P^\intercal A_1 P)x : \begin{array}{l} x^\intercal(P^\intercal A_2 P)x + 2(P^\intercal b_2)^\intercal x \leq 1 \\ LPx \leq 1, \end{array} \right\} \tag{7.29}$$

and solves this reformulation using the SOCP-based BB method.

- `1-RSDCBB` applies Algorithm 12 to construct an SDC pair $\left\{\tilde{A}_1,\ \tilde{A}_2\right\} \in \mathbb{S}^{n+1}$ whose top-left $n \times n$ principal submatrices are A_1 and A_2, respectively. Let $P \in \mathbb{R}^{(n+1)\times(n+1)}$ denote the invertible matrix furnished by the SDC property of $\left\{\tilde{A}_1,\ \tilde{A}_2\right\}$. Also, set $\tilde{b}_2 = (b_2^\intercal, 0)^\intercal$ and $\tilde{L} = (L, 0_{m,1})$. Then, `1-RSDCBB` reformulates (7.26) as

$$\inf_{w\in\mathbb{R}^{n+1}} \left\{ w^\intercal (P^\intercal \tilde{A}_1 P) w : \begin{array}{l} w^\intercal (P^\intercal \tilde{A}_2 P) w + 2(P^\intercal \tilde{b}_2)^\intercal w \leq 1 \\ (\tilde{L}P) w \leq 1 \\ (Pw)_{n+1} = 0 \end{array} \right\} \tag{7.30}$$

  and solves this reformulation using the SOCP-based BB method. Note that for this reformulation, $\mathcal{L}$ is the set of $w \in \mathbb{R}^{n+1}$ satisfying both the linear inequality and linear equality constraints.

- `k-RSDCBB` applies Algorithm 13 with $d = k$ to construct an SDC pair $\left\{\tilde{A}_1,\ \tilde{A}_2\right\} \in \mathbb{S}^{n+k}$ whose top-left $n \times n$ principal submatrices are A_1 and A_2, respectively. Let $P \in \mathbb{R}^{(n+k)\times(n+k)}$ denote the invertible matrix furnished by the SDC property of $\left\{\tilde{A}_1,\ \tilde{A}_2\right\}$. Also, set $\tilde{b}_1 = (b_1^\intercal, 0_{k,1})^\intercal$ and $\tilde{L} = (L, 0_{m,k})$. Then, `k-RSDCBB` reformulates (7.26) as

$$\inf_{w\in\mathbb{R}^{n+1}} \left\{ w^\intercal (P^\intercal \tilde{A}_1 P) w : \begin{array}{l} w^\intercal (P^\intercal \tilde{A}_2 P) w + 2(P^\intercal \tilde{b}_2)^\intercal w \leq 1 \\ (\tilde{L}P) w \leq 1 \\ (Pw)_{n+1} = (Pw)_{n+2} = \cdots = (Pw)_{n+k} = 0 \end{array} \right\} \tag{7.31}$$

  and solves this reformulation using the SOCP-based BB method. Note that for this reformulation, $\mathcal{L}$ is the set of $w \in \mathbb{R}^{n+k}$ satisfying both the linear inequality and linear equality constraints.

- `eigBB` first performs an eigenvalue decomposition on A_1 to write $D_1 = P_1^\intercal A_1 P_1$, where D_1 is a diagonal matrix. Then, it performs a second eigenvalue decomposition to write $D_2 = P_2^\intercal (P_1^\intercal A_2 P_1) P_2$, where D_2 is a diagonal matrix. Finally, `eigBB` reformulates (7.26) as

$$\inf_{y,z\in\mathbb{R}^{n}} \left\{ y^\intercal D_1 y : \begin{array}{l} z^\intercal D_2 z + 2(P_1^\intercal b_2)^\intercal y + c_2 \leq 1 \\ (LP_1) y \leq 1 \\ y = P_2 z \end{array} \right\} \tag{7.32}$$

  and solves this reformulation using the SOCP-based BB method. Note that for this reformulation, $\mathcal{L}$ is the set of $(y, z) \in \mathbb{R}^n \times \mathbb{R}^n$ satisfying both the linear inequality and linear equality constraints.

All experiments are implemented using MATLAB R2021a on a PC running Windows 10 Intel(R) Core(TM) i9-10900KF CPU (3.70GHz) and 64GB RAM. All the SDP and SOCP problems in the BB methods are solved by the commercial solver MOSEK [126] through its Matlab interface.

Remark 82. SDCBB, 1-RSDCBB, k-RSDCBB, and eigBB can be thought of as different reformulations within a parameterized family of reformulations of (7.26). Specifically, these four algorithms reformulate (7.26) as diagonal QCQPs with $n, n+1, n+k$, and $2n$ variables respectively. □

Experiment setup. We tested the solution methods on random instances for various settings of (n, k). For $k = 0$, i.e., the case where A_1 and A_2 are guaranteed to be SDC, we compared SDPBB, SDCBB, and eigBB. For $k > 0$, we compared SDPBB, 1-RSDCBB, k-RSDCBB, and eigBB. For each (n, k), we generated 5 random problems and used the command boxplot in MATLAB to present the statistics. Each procedure was terminated when the CPU time reached 1800 seconds or when the relative gap (between the objective value of the current solution and the best lower bound) fell below the default tolerance threshold, 10^{-4}. In all of our figures and tables, we set

$$\text{Gap} = \frac{v_{\text{best}} - v_0}{|v_{\text{best}}|} \times 100,$$

where v_0 is the initial lower bound computed from the corresponding convex relaxation, and v_{best} is the best upper bound computed within the BB method.

Comparison for the SDC case. We first test instances where $\{A_1, A_2\}$ is SDC, i.e., $k = 0$, for $n = 10, 20, 30, 40, 50$. The results on CPU time, relative gap, and number of explored nodes in the search tree are reported in Figure 7.1. Figure 7.1 shows us that SDCBB performs the best in general, i.e., SDCBB achieves the lowest relative gap and smallest CPU time across all tested values of n. Both of the SOCP-based methods are much more efficient than SDPBB. In fact, SDPBB fails to solve any of the instances to relative gap 10^{-4} when $n > 20$ and fails on four of the five instances with $n = 20$. Moreover, for $n = 10$, we observe that the SDP-based BB method explores more nodes than either of the SOCP-based BB methods, even though the SDP lower bounds are computationally more expensive than the SOCP lower bounds. Indeed, we will see soon that the SOCP relaxation experimentally yields tighter lower bounds (resulting in fewer search tree nodes) than the SDP relaxation. We also observe that eigBB is comparable to but slightly less efficient than SDCBB. Specifically, we note that SDCBB and eigBB explore similar numbers of nodes but that SDCBB does so in comparable or less time.

To further understand the performance between the SDP-based and SOCP-based BB methods, we compare initial bound quality and CPU time for SDPBB, SDCBB and eigBB in the case $k = 0$. For Figure 7.2 only, define

$$\text{Gap} = \frac{v_{\text{SDCBB}} - v_0}{|v_{\text{SDCBB}}|} \times 100,$$

where v_0 is the initial lower bound computed by SDPBB, SDCBB or eigBB and v_{SDCBB} is the best upper bound computed by SDCBB after 1800 seconds. Figure 7.2 shows that both SOCP relaxations are faster to compute than the SDP relaxation, as expected. More interestingly, both SOCP relaxations provide a *better* initial lower bound as can be seen by the fact that the gap is significantly smaller for the SOCP relaxations than it is for the SDP relaxation. See Section 7.7.1 for heuristic explanations why we would expect this to hold. Both observations in Figure 7.2 suggest that diagonalization can be used within branch and bound schemes to solve QCQPs more efficiently.

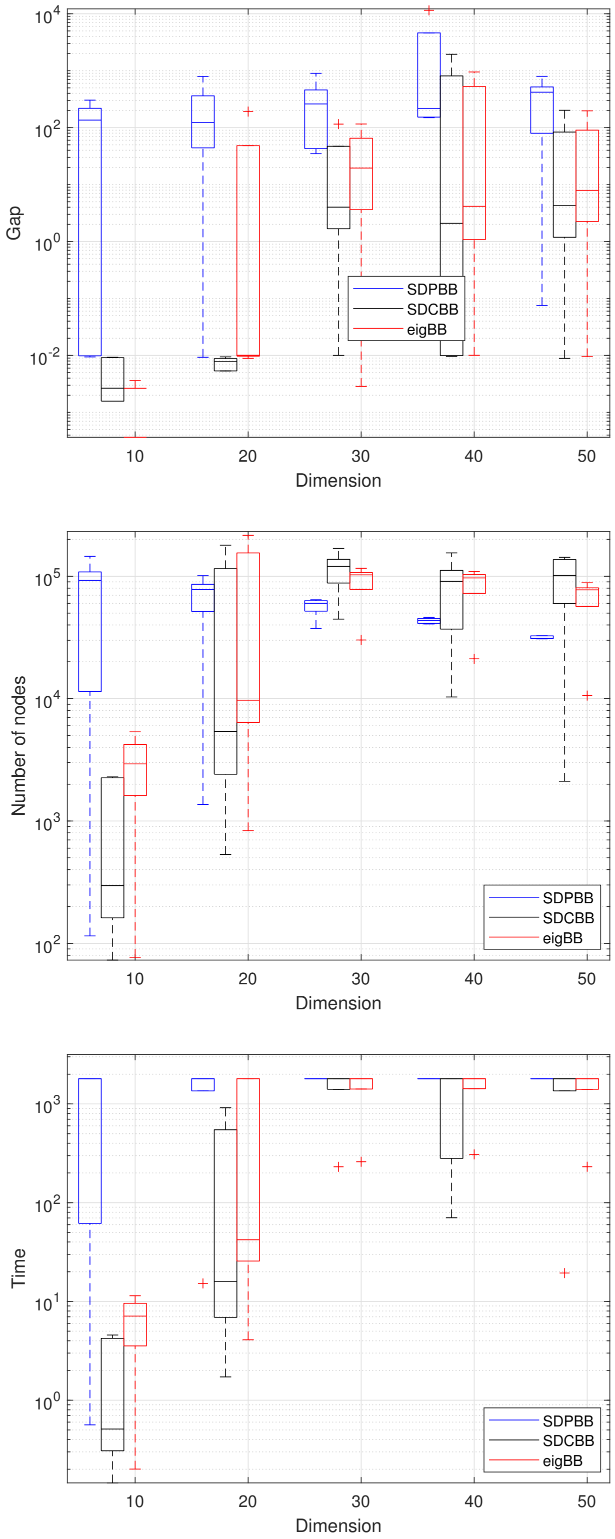

Figure 7.1: Comparison of `SDPBB`, `SDCBB` and `eigBB` for the case with $k = 0$.

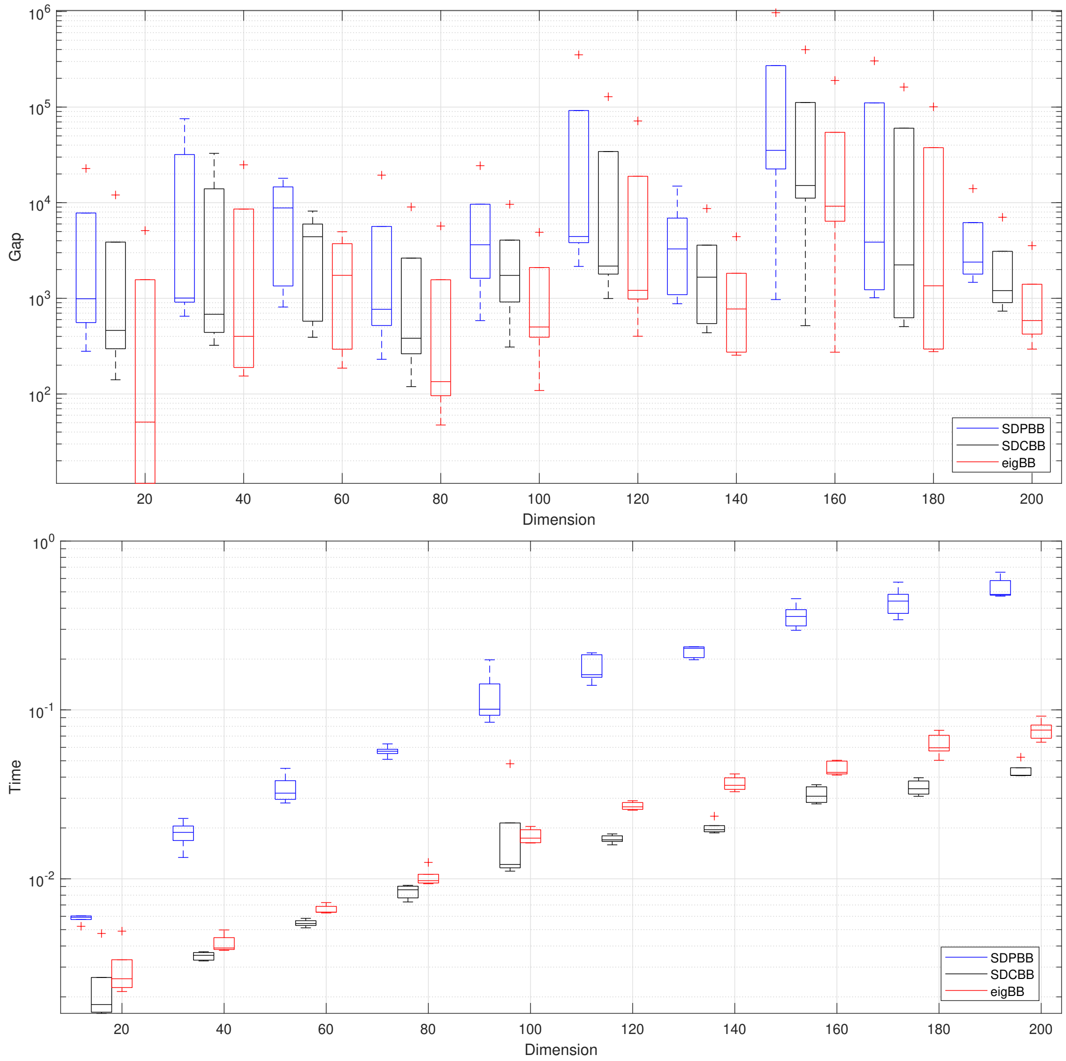

Figure 7.2: Comparison of initial bound and time between SDP and SOCP relaxations for instances of different dimensions.

Comparing SDCBB and eigBB in Figure 7.2, we see that eigBB generally produces tighter lower bounds but SDCBB needs less computation time to solve its relaxation. This parallels the observation in Figure 7.1 that SDCBB is capable of exploring more nodes than eigBB in similar amounts of time. We believe that SDCBB solves its relaxation faster simply because its diagonal reformulation is smaller. Indeed, SDCBB solves an SOCP (7.29) with $\left(n + \left|\mathrm{supp}(a_1^-) \cup \mathrm{supp}(a_2^-)\right|\right)$-many variables while eigBB solves an SOCP (7.32) with roughly twice as many variables: $\left(2n + \left|\mathrm{supp}(a_1^-)\right| + \left|\mathrm{supp}(a_2^-)\right|\right)$-many variables.

Comparison for the non-SDC case. We now consider the case where $\{A_1, A_2\}$ is not SDC, i.e., $k > 0$. We tested SDPBB, 1-RSDCBB, k-RSDCBB and eigBB for $n = 10, 20, 30$ and $k = 1, 1+\frac{n}{10}, 1+\frac{2n}{10}, 1+\frac{3n}{10}, 1+\frac{4n}{10}$. The results on CPU time, relative gap, and number of explored nodes in the search tree for SDPBB, k-RSDCBB and eigBB are reported in Figure 7.3. Figure 7.3 indicates that both k-RSDCBB and eigBB largely outperform SDPBB. Indeed, SDPBB cannot solve most instances in the time limit, evidenced from the left plot in Figure 7.3, while k-RSDCBB and eigBB can solve more instances and have lower relative gaps for unsolved instances in the time limit. In general, k-RSDCBB and eigBB are comparable and do not dominate each other.

It remains to comment on the numerical performance of 1-RSDCBB. Experimentally, we observed that the 1-RSDC construction (Algorithm 12) yields very large condition numbers for the P matrices in (7.30) (e.g., larger than 1e6). This leads to inaccurate solutions or numerical failures in MOSEK when solving the SOCP+RLT relaxation, especially for $k \geq 5$. Note also that 1-RSDCBB coincides with k-RSDCBB for $k = 1$. Thus, we compare the three SOCP-based BB methods, 1-RSDCBB, k-RSDCBB, and eigBB, for values of $1 < k < 5$ in Table 7.1.

One may observe that, for $n = 10$, 1-RSDCBB seems to perform worse (compared to 1-RSDCBB and eigBB) as k increases. This trend can be explained by observing that the condition numbers of the P matrices for (7.30) are likely to “blow up” as k increases (see the two rightmost columns of Table 7.1). In particular, we observed that the lower and upper bounds that we computed for the decision variables (i.e., the values of ℓ and u at the root node) in k-RSDCBB and eigBB were relatively small intervals, while the corresponding bounds for those in 1-RSDCBB were often much larger (e.g., on the order of 1000 times larger for $k = 3$). Comparing the rightmost two columns of Table 7.1, we see that the condition numbers of the invertible matrices P that we construct are often much smaller for k-RSDCBB than for 1-RSDCBB, especially as k gets larger. We believe this explains why k-RSDCBB generally outperforms 1-RSDCBB for larger values of the parameter k. Finally, we observe that for the last instances in (10,4) and (30,4), 1-RSDCBB returned solutions without reaching the prescribed gap or CPU times. We believe that this was caused in both instances by numerical inaccuracies within the interior point solves in MOSEK due to the large condition numbers, i.e., 2.73e6 and 1.15e5. For $k \geq 5$, the condition number of 1-RSDCBB is even worse and 1-RSDCBB fails for almost all instances (not reported here).

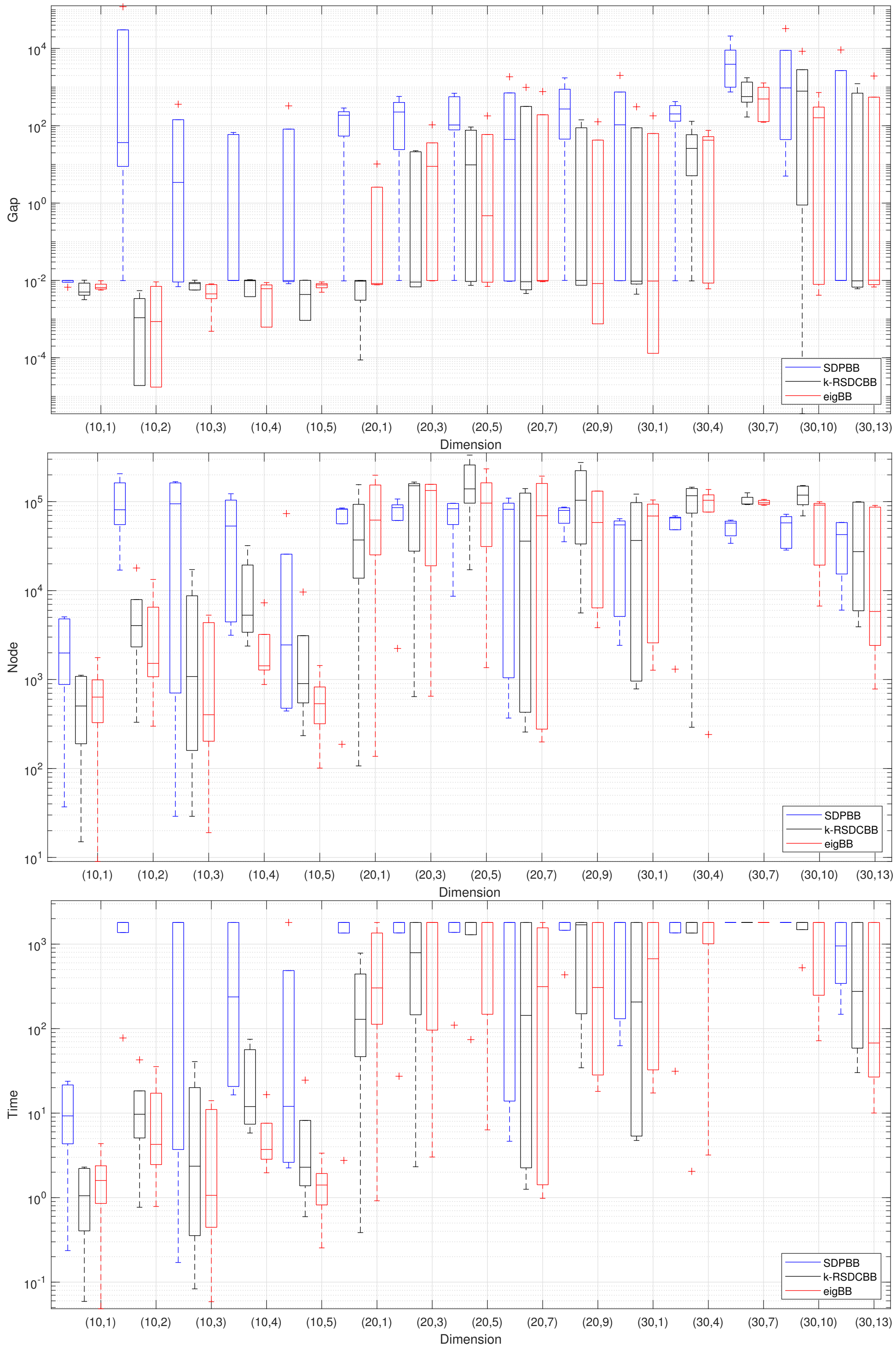

Figure 7.3: Comparison of SDPBB, k-RSDCBB and eigBB for non-SDC instances.

(n,k)	`1-RSDCBB`			`k-RSDCBB`			`eigBB`			cond num	
	time	node	gap (%)	time	node	gap (%)	time	node	gap (%)	`1-RSDC`	`2-RSDC`
(10,2)	5.73	2830	0.00	9.68	4582	0.00	**3.02**	1335	0.01	5.14e+01	3.79e+00
(10,2)	**27.87**	11462	0.00	42.68	17944	0.00	35.49	13335	0.00	2.78e+01	5.09e+00
(10,2)	30.82	13764	0.00	**6.52**	2995	0.00	11.11	4234	0.00	4.70e+02	4.54e+00
(10,2)	2.55	972	0.00	**0.77**	331	0.00	0.79	299	0.00	4.22e+02	2.25e+00
(10,2)	15.84	4423	0.00	10.23	4045	0.01	**4.27**	1521	0.01	1.59e+02	2.37e+00
(10,3)	2.71	1264	0.01	**0.45**	203	0.01	0.57	264	0.01	2.29e+02	2.72e+00
(10,3)	16.67	6848	0.00	**13.15**	5899	0.00	14.04	5295	0.00	1.89e+02	4.87e+00
(10,3)	19.55	8176	0.01	40.75	17257	0.01	**10.04**	4056	0.00	5.36e+01	3.42e+00
(10,3)	1.91	789	0.00	0.08	29	0.01	**0.06**	19	0.00	1.68e+03	2.24e+00
(10,3)	54.33	20000	0.01	2.36	1080	0.01	**1.06**	402	0.01	2.28e+03	1.44e+01
(10,4)	259.95	69602	0.01	11.95	5289	0.01	**1.97**	879	0.01	4.37e+03	3.31e+00
(10,4)	1800.05	147765	23.56	7.93	3746	0.00	**3.13**	1414	0.00	1.17e+04	8.04e+00
(10,4)	46.22	19976	0.01	74.85	32075	0.01	**16.55**	7295	0.01	3.63e+02	7.57e+01
(10,4)	1800.08	130796	158.72	5.81	2381	0.01	**4.61**	1858	0.00	2.10e+04	6.55e+00
(10,4)	77.54	16565	1.61	50.20	15150	0.01	**3.71**	1427	0.01	2.73e+06	2.49e+01
(20,3)	1800.07	120343	169.36	193.58	36815	0.01	**126.92**	25152	0.01	3.64e+05	3.20e+01
(20,3)	1800.05	107481	216.00	1800.05	150828	22.65	1800.04	156611	**8.94**	8.99e+03	1.44e+01
(20,3)	1800.05	162012	49.30	**790.61**	166079	0.00	1800.05	156891	13.02	2.35e+02	9.71e+00
(20,3)	1800.07	115944	331.43	1800.07	156808	**20.78**	1800.07	133551	106.07	1.06e+03	3.76e+00
(20,3)	6.74	1866	0.01	**2.32**	643	0.01	3.02	650	0.01	6.00e+02	1.01e+01
(30,4)	1800.08	102100	100.73	1800.08	116527	**25.97**	1800.07	103676	42.39	2.85e+03	5.81e+00
(30,4)	1800.06	117590	205.94	1800.05	138837	**34.78**	1800.07	113383	44.58	1.26e+04	6.50e+00
(30,4)	1800.07	95644	838.24	1800.04	145488	6.80	**1345.35**	136907	0.01	2.27e+05	1.41e+01
(30,4)	1800.03	110507	1463.26	1800.08	99003	130.64	1800.08	101895	**75.89**	2.57e+05	6.43e+00
(30,4)	66.06	5380	0.02	**2.05**	291	0.01	3.19	241	0.01	1.15e+05	1.04e+01

Table 7.1: Comparison of different SOCP-based BB methods for $1 < k < 5$. In each row, the solution method with the lowest solution time is highlighted. For instances where all three methods time out (1800 seconds) before reaching optimality, the solution method with the lowest objective value is highlighted. Two outliers are highlighted in blue.

Bibliography

[1] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *IEEE Trans. Inform. Theory*, 62(1):471–487, 2015.

[2] S. Adachi and Y. Nakatsukasa. Eigenvalue-based algorithm and analysis for nonconvex QCQP with one constraint. *Math. Program.*, 173:79–116, 2019.

[3] J. Agler, W. Helton, S. McCullough, and L. Rodman. Positive semidefinite matrices with a given sparsity pattern. *Linear Algebra Appl.*, 107:101–149, 1988.

[4] F. Alizadeh. Interior point methods in semidefinite programming with applications to combinatorial optimization. *SIAM J. Optim.*, 5(1):13–51, 1995.

[5] F. Alizadeh, J. A. Haeberly, and M. L. Overton. Complementarity and nondegeneracy in semidefinite programming. *Math. Program.*, 77:111–128, 1997.

[6] K. M. Anstreicher. Semidefinite programming versus the reformulation-linearization technique for nonconvex quadratically constrained quadratic programming. *J. Global Optim.*, 43(2):471–484, 2009.

[7] C.J. Argue, F. Kılınç-Karzan, and A. L. Wang. Necessary and sufficient conditions for rank-one generated cones. *Math. Oper. Res.*, 2022. Forthcoming, *arXiv preprint*, 2007.07433.

[8] C. Audet, P. Hansen, B. Jaumard, and G. Savard. A branch and cut algorithm for nonconvex quadratically constrained quadratic programming. *Math. Program.*, 87(1):131–152, 2000.

[9] M. Baes, M. Burgisser, and A. Nemirovski. A randomized mirror-prox method for solving structured large-scale matrix saddle-point problems. *SIAM J. Optim.*, 23(2):934–962, 2013.

[10] A. S. Bandeira, N. Boumal, and A. Singer. Tightness of the maximum likelihood semidefinite relaxation for angular synchronization. *Math. Program.*, 163:145–167, 2017.

[11] X. Bao, N. V. Sahinidis, and M. Tawarmalani. Semidefinite relaxations for quadratically constrained quadratic programming: A review and comparisons. *Math. Program.*, 129: 129, 2011.

[12] G. P. Barker. Faces and duality in convex cones. *Linear Multilinear Algebra*, 6(3):161–169, 1978.

[13] G. P. Barker. Theory of cones. *Linear Algebra Appl.*, 39:263–291, 1981.

[14] G. P. Barker and D. Carlson. Cones of diagonally dominant matrices. *Pacific J. Math.*, 57: 15–32, 1975.

[15] A. Barvinok. Feasibility testing for systems of real quadratic equations. *Discrete Comput. Geom.*, 10:1–13, 1993.

[16] A. Barvinok. *A Course in Convexity*, volume 54 of *Graduate Studies in Mathematics*. 2002.

[17] A. Beck. Quadratic matrix programming. *SIAM J. Optim.*, 17(4):1224–1238, 2007.

[18] A. Beck and Y. C. Eldar. Strong duality in nonconvex quadratic optimization with two quadratic constraints. *SIAM J. Optim.*, 17(3):844–860, 2006.

[19] A. Beck and M. Teboulle. A convex optimization approach for minimizing the ratio of indefinite quadratic functions over an ellipsoid. *Math. Program.*, 118:13–35, 2009.

[20] A. Beck, Y. Drori, and M. Teboulle. A new semidefinite programming relaxation scheme for a class of quadratic matrix problems. *Oper. Res. Lett.*, 40(4):298–302, 2012.

[21] A. Ben-Tal and D. den Hertog. Hidden conic quadratic representation of some nonconvex quadratic optimization problems. *Math. Program.*, 143:1–29, 2014.

[22] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization*, volume 2 of *MPS-SIAM Ser. Optim.* 2001.

[23] A. Ben-Tal and A. Nemirovski. Solving large scale polynomial convex problems on ℓ_1/nuclear norm balls by randomized first-order algorithms. *CoRR*, 2012.

[24] A. Ben-Tal and M. Teboulle. Hidden convexity in some nonconvex quadratically constrained quadratic programming. *Math. Program.*, 72:51–63, 1996.

[25] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*, volume 28 of *Princeton Ser. Appl. Math.* 2009.

[26] D. Bienstock and A. Michalka. Polynomial solvability of variants of the Trust-Region Subproblem. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 380–390, 2014.

[27] A. Billionnet, S. Elloumi, and A. Lambert. Exact quadratic convex reformulations of mixed-integer quadratically constrained problems. *Math. Program.*, 158(1):235–266, 2016.

[28] N. Bishop, L. Tran-Thanh, G. Long, and E. Gerding. Optimal learning from verified training data. *Advances in Neural Information Processing Systems*, 33:9520–9529, 2020.

[29] G. Blekherman, R. Sinn, and M. Velasco. Do sums of squares dream of free resolutions? *SIAM J. Appl. Algebra Geom.*, 1:175–199, 2017.

[30] J. Borwein and H. Wolkowicz. Regularizing the abstract convex program. *J. Math. Anal. Appl.*, 83(2):495–530, 1981.

[31] N. Boumal, V. Voroninski, and A. Bandeira. The non-convex Burer–Monteiro approach works on smooth semidefinite programs. volume 29, 2016.

[32] C. Buchheim, M. De Santis, L. Palagi, and M. Piacentini. An exact algorithm for nonconvex quadratic integer minimization using ellipsoidal relaxations. *SIAM J. Optim.*, 23(3):1867–1889, 2013.

[33] S. Burer. A gentle, geometric introduction to copositive optimization. *Math. Program.*, 151:89–116, 2015.

[34] S. Burer and K. M. Anstreicher. Second-order-cone constraints for Extended Trust-Region Subproblems. *SIAM J. Optim.*, 23(1):432–451, 2013.

[35] S. Burer and F. Kılınç-Karzan. How to convexify the intersection of a second order cone and a nonconvex quadratic. *Math. Program.*, 162:393–429, 2017.

[36] S. Burer and R. D.C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Math. Program.*, 95:329–357, 2003.

[37] S. Burer and B. Yang. The trust region subproblem with non-intersecting linear constraints. *Math. Program.*, 149:253–264, 2014.

[38] S. Burer and Y. Ye. Exact semidefinite formulations for a class of (random and non-random) nonconvex quadratic programs. *Math. Program.*, 181:1–17, 2019.

[39] M. D. Bustamante, P. Mellon, and M. V. Velasco. Solving the problem of simultaneous diagonalization of complex symmetric matrices via congruence. *SIAM J. Matrix Anal. Appl.*, 41(4):1616–1629, 2020.

[40] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion. *SIAM Rev.*, 57(2):225–251, 2015.

[41] Y. Carmon and J. C. Duchi. Analysis of Krylov subspace solutions of regularized nonconvex quadratic problems. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 10728–10738, 2018.

[42] S. Ceria and J. Soares. Convex programming for disjunctive convex optimization. *Math. Program.*, 86:595–614, 1999.

[43] A. Chambolle and T. Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Math. Program.*, 159:253–287, 2016.

[44] C. Chen, A. Atamtürk, and S. S. Oren. A spatial branch-and-cut method for nonconvex qcqp with bounded complex variables. *Math. Program.*, 165(2):549–577, 2017.

[45] J. Chen and S. Burer. Globally solving nonconvex quadratic programming problems via completely positive programming. *Math. Prog. Comput.*, 4(1):33–52, 2012.

[46] D. Cifuentes. On the Burer–Monteiro method for general semidefinite programs. *Opt. Lett.*, 15(6):2299–2309, 2021.

[47] D. Cifuentes and A. Moitra. Polynomial time guarantees for the Burer-Monteiro method. *arXiv preprint*, 1912.01745, 2019.

[48] D. Cifuentes, C. Harris, and B. Sturmfels. The geometry of SDP-exactness in quadratic optimization. *Math. Program.*, 182:399–428, 2020.

[49] M. Conforti, G. Cornuéjols, and G. Zambelli. *Integer Programming*, volume 271 of *Grad. Texts in Math.* 2014.

[50] A. R. Conn, N. I. Gould, and P. L. Toint. *Trust Region Methods*, volume 1 of *MPS-SIAM Ser. Optim.* 2000.

[51] A. d'Aspremont and N. El Karoui. A stochastic smoothing algorithm for semidefinite programming. *SIAM J. Optim.*, 24(3):1138–1177, 2014.

[52] M. K. de Carli Silva and L. Tunçel. A notion of Total Dual Integrality for convex, semidefinite, and extended formulations. *SIAM J. Discrete Math.*, 34(1):470–496, 2020.

[53] E. de Klerk and R. Sotirov. Exploiting group symmetry in semidefinite programming relaxations of the quadratic assignment problem. *Math. Program.*, 122:225–246, 2010.

[54] E. de Klerk, D. V. Pasechnik, and A. Schrijver. Reduction of symmetric semidefinite programs using the regular $*$-representation. *Math. Program.*, 109:613–624, 2007.

[55] E. de Klerk, C. Dobre, and D. V. Pasechnik. Numerical block diagonalization of matrix *-algebras with application to semidefinite programming. *Math. Program.*, 129:91–111, 2011.

[56] A. P. Dempster. Covariance selection. *Biometrics*, 28:157–175, 1972.

[57] O. Devolder, F. Glineur, and Y. Nesterov. First-order methods with inexact oracle: the strongly convex case. Technical Report 2013016, 2013.

[58] O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Math. Program.*, 146(1):37–75, 2014.

[59] L. L. Dines. On the mapping of quadratic forms. *Bull. Amer. Math. Soc.*, 47(6):494–498, 1941.

[60] L. Ding, A. Yurtsever, V. Cevher, J. A. Tropp, and M. Udell. An optimal-storage approach to semidefinite programming using approximate complementarity. *SIAM J. Optim.*, 31(4): 2695–2725, 2021.

[61] H. Dong and J. Linderoth. On valid inequalities for quadratic programming with continuous variables and binary indicators. In *Integer Programming and Combinatorial Optimization (IPCO 2013)*, pages 169–180, 2013.

[62] I. Ekeland and R. Temam. *Convex analysis and variational problems*, volume 28 of *Classics Appl. Math.* 1999.

[63] A. Eltved and S. Burer. Strengthened sdp relaxation for an extended trust region subproblem with an application to optimal power flow. *Math. Program.*, pages 1–26, 2022.

[64] S. Fallahi, M. Salahi, and T. Terlaky. Minimizing an indefinite quadratic function subject to a single indefinite quadratic constraint. *Optimization*, 67(1):55–65, 2018.

[65] J.M. Feng, G.X. Xuan, R.L. Sheu, and Y. Xia. Duality and solutions for quadratic programming over single non-homogeneous quadratic constraint. *J. Global Optim.*, 54(2): 275–293, 2012.

[66] C. Fortin and H. Wolkowicz. The Trust Region Subproblem and semidefinite programming. *Optim. Methods Softw.*, 19(1):41–67, 2004.

[67] A. L. Fradkov and V. A. Yakubovich. The S-procedure and duality relations in nonconvex problems of quadratic programming. *Vestnik Leningrad Univ. Math.*, 6:101–109, 1979.

[68] A. Frangioni and C. Gentile. Perspective cuts for a class of convex 0–1 mixed integer programs. *Math. Program.*, 106:225–236, 2006.

[69] S. Friedland and R. Loewy. Subspaces of symmetric matrices containing matrices with a multiple first eigenvalue. *Pacific J. Math.*, 62(2):389–399, 1976.

[70] M. P. Friedlander and I. Macêdo. Low-rank spectral optimization via gauge duality. *SIAM Journal on Scientific Computing*, 38(3):A1616–A1638, 2016.

[71] T. Fujie and M. Kojima. Semidefinite programming relaxation for nonconvex quadratic programs. *J. Global Optim.*, 10(4):367–380, 1997.

[72] K. Gatermann and P. A. Parrilo. Symmetry groups, semidefinite programs, and sums of squares. *J. Pure Appl. Algebra*, 192(1-3):95–128, 2004.

[73] D. Gijswijt. Matrix algebras and semidefinite programming techniques for codes. *arXiv preprint*, 1007.0906, 2010.

[74] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM*, 42(6):1115–1145, 1995.

[75] G. H. Golub and Q. Ye. An inverse free preconditioned krylov subspace method for symmetric generalized eigenvalue problems. *SIAM J. Sci. Comput.*, 24(1):312–334, 2002.

[76] G. H. Golub, P. C. Hansen, and D. P. O'Leary. Tikhonov regularization and total least squares. *SIAM J. Matrix Anal. Appl.*, 21(1):185–194, 1999.

[77] N. I. M. Gould, S. Lucidi, M. Roma, and P. L. Toint. Solving the Trust-Region Subproblem using the Lanczos method. *SIAM J. Optim.*, 9(2):504–525, 1999.

[78] R. Grone, C. R. Johnson, E. M. Sá, and H. Wolkowicz. Positive definite completions of partial Hermitian matrices. *Linear Algebra Appl.*, 58:109–124, 1984.

[79] O. Günlük and J. Linderoth. Perspective reformulations of mixed integer nonlinear programs with indicator variables. *Math. Program.*, 124:183–205, 2010.

[80] C. Guo, N. J. Higham, and F. Tisseur. An improved arc algorithm for detecting definite Hermitian pairs. *SIAM J. Matrix Anal. Appl.*, 31(3):1131–1151, 2010.

[81] E. Y. Hamedani and N. C. Aybat. A primal-dual algorithm with line search for general convex-concave saddle point problems. *SIAM J. Optim.*, 31(2):1299–1329, 2021.

[82] E. Hazan and T. Koren. A linear-time algorithm for trust region problems. *Math. Program.*, 158:363–381, 2016.

[83] R. Hildebrand. Spectrahedral cones generated by rank 1 matrices. *J. Global Optim.*, 64: 349–397, 2016.

[84] J. Hiriart-Urruty. Potpourri of conjectures and open questions in nonlinear analysis and optimization. *SIAM Rev.*, 49(2):255–273, 2007.

[85] J. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of convex analysis*, volume 1 of *Grundlehren Text Editions*. 2004.

[86] H. Hmam. Quadratic optimisation with one quadratic equality constraint. Technical report, Defence Science and Technology Organisation Edinburgh (Australia) Electronic Warfare and Radar Division, 2010.

[87] N. Ho-Nguyen and F. Kılınç-Karzan. A second-order cone based approach for solving the Trust Region Subproblem and its variants. *SIAM J. Optim.*, 27(3):1485–1512, 2017.

[88] R. A. Horn and C. R. Johnson. *Matrix analysis*. 2012.

[89] Y. Hsia and R. Sheu. Trust region subproblem with a fixed number of additional linear inequality constraints has polynomial complexity. *arXiv preprint*, (1312.1398), 2013.

[90] K. Huang and N. D. Sidiropoulos. Consensus-ADMM for general quadratically constrained quadratic programming. *IEEE Trans. Signal Process.*, 64(20):5297–5310, 2016.

[91] E. Phan huy Hao. Quadratically constrained quadratic programming: Some applications and a method for solution. *Z. Oper. Res.*, 26:105–119, 1982.

[92] V. Jeyakumar and G. Y. Li. Trust-region problems with linear inequality constraints: Exact SDP relaxation, global optimality and robust optimization. *Math. Program.*, 147:171–206, 2014.

[93] R. Jiang and D. Li. Simultaneous diagonalization of matrices and its applications in quadratically constrained quadratic programming. *SIAM J. Optim.*, 26(3):1649–1668, 2016.

[94] R. Jiang and D. Li. Novel reformulations and efficient algorithms for the Generalized Trust Region Subproblem. *SIAM J. Optim.*, 29(2):1603–1633, 2019.

[95] R. Jiang and D. Li. A linear-time algorithm for generalized trust region problems. *SIAM J. Optim.*, 30(1):915–932, 2020.

[96] R. Jiang, D. Li, and B. Wu. SOCP reformulation for the Generalized Trust Region Subproblem via a canonical form of two symmetric matrices. *Math. Program.*, 169:531–563, 2018.

[97] A. Juditsky and A. Nemirovski. First order methods for nonsmooth convex large-scale optimization, ii: utilizing problems structure. *Optimization for Machine Learning*, 30(9): 149–183, 2011.

[98] N. Karmarkar, M. G. Resende, and K. G. Ramakrishnan. An interior point algorithm to solve computationally difficult set covering problems. *Math. Program.*, 52:597–618, 1991.

[99] R. M. Karp. Reducibility among combinatorial problems. In *Complexity of computer computations*, pages 85–103, 1972.

[100] F. Kılınç-Karzan and A. L. Wang. Exactness in SDP relaxations of QCQPs: Theory and applications. Tut. in Oper. Res. 2021.

[101] F. Kılınç-Karzan and S. Yıldız. Two-term disjunctions on the Second-Order Cone. *Math. Program.*, 154:463–491, 2015.

[102] L. Kronecker. *Collected works*. 1968.

[103] J. Kuczynski and H. Wozniakowski. Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start. *SIAM J. Matrix Anal. Appl.*, 13(4):1094–1122, 1992.

[104] G. Lan, Z. Lu, and R. D.C. Monteiro. Primal-dual first-order methods with $o(1/\epsilon)$ iteration-complexity for cone programming. *Math. Program.*, 126:1–29, 2011.

[105] P. Lancaster and L. Rodman. Canonical forms for Hermitian matrix pairs under strict equivalence and congruence. *SIAM Rev.*, 47(3):407–443, 2005.

[106] M. Laurent and S. Poljak. On a positive semidefinite relaxation of the cut polytope. *Linear Algebra Appl.*, 223-224:439–461, 1995.

[107] T. H. Le and T. N. Nguyen. Simultaneous diagonalization via congruence of hermitian matrices: some equivalent conditions and a numerical solution. *arXiv preprint*, (2007.14034), 2020.

[108] K. Y. Levy, A. Yurtsever, and V. Cevher. Online adaptive methods, universality and acceleration. In *Advances in Neural Information Processing Systems*, 2018.

[109] J. Linderoth. A simplicial branch-and-bound algorithm for solving quadratically constrained quadratic programs. *Math. Program.*, 103(2):251–282, 2005.

[110] M. Liu and G. Pataki. Exact duals and short certificates of infeasibility and weak infeasibility in conic linear programming. *Math. Program.*, 167:435–480, 2018.

[111] M. Locatelli. Some results for quadratic problems with one or two quadratic constraints. *Oper. Res. Lett.*, 43(2):126–131, 2015.

[112] M. Locatelli. Exactness conditions for an SDP relaxation of the extended trust region problem. *Oper. Res. Lett.*, 10(6):1141–1151, 2016.

[113] M. Locatelli. KKT-based primal-dual exactness conditions for the Shor relaxation. *arXiv preprint*, 2011.05033, 2020.

[114] C. Lu, Z. Deng, and Q. Jin. An eigenvalue decomposition based branch-and-bound algorithm for nonconvex quadratic programming problems with convex quadratic constraints. *J. Global Optim.*, 67(3):475–493, 2017.

[115] Z. Lu, A. Nemirovski, and R. D.C. Monteiro. Large-scale semidefinite programming via a saddle point mirror-prox algorithm. *Math. Program.*, 109:211–237, 2007.

[116] H. Luo, Y. Chen, X. Zhang, and D. Li. Effective algorithms for optimal portfolio deleveraging problem with cross impact. *arXiv preprint*, (2012.07368), 2020.

[117] Z. Luo, W. Ma, A. M. So, Y. Ye, and S. Zhang. Semidefinite relaxation of quadratic optimization problems. *IEEE Signal Process. Mag.*, 27(3):20–34, 2010.

[118] A. Majumdar, G. Hall, and A. A. Ahmadi. Recent scalability improvements for semidefinite programming with applications in machine learning, control, and robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:331–360, 2020.

[119] B. R. Marks and G. P. Wright. A general inner approximation algorithm for nonconvex mathematical programs. *Oper. Res.*, 26(4):681–683, 1978.

[120] N. Megiddo. Linear-time algorithms for linear programming in r^3 and related problems. *SIAM journal on computing*, 12(4):759–776, 1983.

[121] A. Megretski. Relaxations of quadratic programs in operator theory and system analysis. In *Systems, Approximation, Singular Integral Operators, and Related Topics*, pages 365–392, 2001.

[122] D. G. Mixon, S. Villar, and R. Ward. Clustering subgaussian mixtures by semidefinite programming. *arXiv preprint*, 1602.06612, 2016.

[123] S. Modaresi and J. P. Vielma. Convex hull of two quadratic or a conic quadratic and a quadratic inequality. *Math. Program.*, 164:383–409, 2017.

[124] J. J. Moré. Generalizations of the trust region problem. *Optim. Methods Softw.*, 2(3-4): 189–209, 1993.

[125] J. J. Moré and D. C. Sorensen. Computing a trust region step. *SIAM J. on Sci. and Stat. Comput.*, 4(3):553–572, 1983.

[126] MOSEK ApS. *The MOSEK optimization toolbox for MATLAB manual. Version 9.10.*, 2021. URL http://docs.mosek.com/9.0/toolbox/index.html.

[127] T. S. Motzkin and O. Taussky. Pairs of matrices with property L. II. *Trans. Amer. Math. Soc.*, 80(2):387–401, 1955.

[128] A. Nemirovski. Prox-method with rate of convergence o(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM J. Optim.*, 15(1):229–251, 2004.

[129] Y. Nesterov. Quality of semidefinite relaxation for nonconvex quadratic optimization. Technical Report 1997019, Université Catholique de Louvain, Center for Operations Research and Econometrics(CORE), 1997.

[130] Y. Nesterov. Excessive gap technique in nonsmooth convex minimization. *SIAM J. Optim.*, 16(1):235–249, 2005.

[131] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103: 127–152, 2005.

[132] Y. Nesterov. *Lectures on convex optimization*. Number 137 in Springer Optim. and its Appl. 2 edition, 2018.

[133] Y. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

[134] T. Nguyen, V. Nguyen, T. Le, and R. Sheu. On simultaneous diagonalization via congruence of real symmetric matrices. *arXiv preprint*, (2004.06360), 2020.

[135] J. Nocedal and S. Wright. *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. 2 edition, 2006.

[136] 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.

[137] K. O'meara and C. Vinsonhaler. On approximately simultaneously diagonalizable matrices. *Linear Algebra Appl.*, 412:39–74, 2006.

[138] Y. Ouyang and Y. Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Math. Program.*, 185:1–35, 2021.

[139] B. Palaniappan and F. Bach. Stochastic variance reduction methods for saddle-point problems. volume 29, 2016.

[140] P. M. Pardalos, Y. Ye, and CG Han. Algorithms for the solution of quadratic knapsack problems. *Linear Algebra Appl.*, 152:69–91, 1991.

[141] G. Pataki. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Math. Oper. Res.*, 23(2):339–358, 1998.

[142] G. Pataki. The geometry of semidefinite programming. In *Handbook of semidefinite programming*, pages 29–65. 2000.

[143] G. Pataki. On the connection of facially exposed and nice cones. *J. Math. Anal. Appl.*, 400 (1):211–221, 2013.

[144] G. Pataki. Strong duality in conic linear programming: facial reduction and extended duals. In *Computational and analytical mathematics*, volume 50 of *Springer Proceedings in Mathematics & Statistics*, pages 613–634, 2013.

[145] V. I. Paulsen, S. C. Power, and R. R. Smith. Schur products and matrix completions. *J. Funct. Anal.*, 85(1):151–178, 1989.

[146] I. Pólik and T. Terlaky. A survey of the S-lemma. *SIAM Rev.*, 49(3):371–418, 2007.

[147] B. T. Polyak. Convexity of quadratic transformations and its use in control and optimization. *J. Optim. Theory Appl.*, 99(3):553–583, 1998.

[148] T. K. Pong and H. Wolkowicz. The generalized trust region subproblem. *Comput. Optim. Appl.*, 58(2):273–322, 2014.

[149] P. Raghavendra. Optimal algorithms and inapproximability results for every CSP? In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 245–254, 2008.

[150] M. V. Ramana. Polyhedra, spectrahedra, and semidefinite programming. In *Topics in semidefinite and interior-point methods*, volume 18 of *Fields Inst. Commun.*, pages 27–38, 1997.

[151] F. Rendl and H. Wolkowicz. A semidefinite framework for trust region subproblems with applications to large scale minimization. *Math. Program.*, 77:273–299, 1997.

[152] R. T. Rockafellar. *Convex Analysis*. Number 28 in Princeton Mathematical Series. 1970.

[153] N. Rujeerapaiboon, K. Schindler, D. Kuhn, and W. Wiesemann. Size matters: Cardinality-constrained clustering and outlier detection via conic optimization. *SIAM J. Optim.*, 29 (2):1211–1239, 2019.

[154] M. Salahi and S. Fallahi. Trust region subproblem with an additional linear inequality constraint. *Optim. Lett.*, 10(4):821–832, 2016.

[155] M. Salahi and A. Taati. An efficient algorithm for solving the generalized trust region subproblem. *Computational and Applied Mathematics*, 37(1):395–413, 2018.

[156] A. Santana and S. S. Dey. The convex hull of a quadratic constraint over a polytope. *SIAM J. Optim.*, 30(4):2983–2997, 2020.

[157] D. Shamsi, N. Taheri, Z. Zhu, and Y. Ye. Conditions for correct sensor network localization using sdp relaxation. In *Discrete geometry and optimization*, volume 69 of *Fields Inst. Commun.*, pages 279–301, 2013.

[158] H. D. Sherali and W. P. Adams. *A reformulation-linearization technique for solving discrete and continuous nonconvex problems*, volume 31 of *Nonconvex Optim. Appl.* 2013.

[159] J. L. Sheriff. *The convexity of quadratic maps and the controllability of coupled systems*. PhD thesis, Harvard University, 2013.

[160] N. Shinde, V. Narayanan, and J. Saunderson. Memory-efficient structured convex optimization via extreme point sampling. *SIAM Journal on Mathematics of Data Science*, 3(3): 787–814, 2021.

[161] N. Z. Shor. Dual quadratic estimates in polynomial and boolean programming. *Ann. Oper. Res.*, 25:163–168, 1990.

[162] M. Sion. On general minimax theorems. *Pacific J. Math.*, 8(1):171–176, 1958.

[163] S. Sojoudi and J. Lavaei. Exactness of semidefinite relaxations for nonlinear optimization problems with underlying graph structure. *SIAM J. Optim.*, 24(4):1746–1778, 2014.

[164] M. Souto, J. D. Garcia, and Á. Veiga. Exploiting low-rank structure in semidefinite programming by approximate operator splitting. *Optimization*, pages 1–28, 2020.

[165] R. J. Stern and H. Wolkowicz. Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations. *SIAM J. Optim.*, 5(2):286–313, 1995.

[166] J. F. Sturm. Error bounds for linear matrix inequalities. *SIAM J. Optim.*, 10(4):1228–1248, 2000.

[167] J. F. Sturm and S. Zhang. On cones of nonnegative quadratic functions. *Math. Oper. Res.*, 28(2):246–267, 2003.

[168] D. A. Suprunenko and R. I. Tyshkevich. *Commutative matrices*. 1968.

[169] B. S. Tam. A note on polyhedral cones. *J. Aust. Math. Soc.*, 22(4):456–461, 1976.

[170] T. Tao. *Topics in random matrix theory*, volume 132 of *Grad. Stud. Math.* 2012.

[171] M. Tawarmalani and N. Sahinidis. *Convexification and global optimization in continuous and mixed-integer nonlinear programming: Theory, algorithms, software, and applications*, volume 65 of *Nonconvex Optim. Appl.* 2002.

[172] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization, 2008.

[173] F. Uhlig. A canonical form for a pair of real symmetric matrices that generate a nonsingular pencil. *Linear Algebra Appl.*, 14(3):189–209, 1976.

[174] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Rev.*, 38(1):49–95, 1996.

[175] R. Vollgraf and K. Obermayer. Quadratic optimization for simultaneous matrix diagonalization. *IEEE Trans. Signal Process.*, 54(9):3270–3278, 2006.

[176] I. Waldspurger and A. Waters. Rank optimality for the Burer–Monteiro factorization. *SIAM J. Optim.*, 30(3):2577–2602, 2020.

[177] A. L. Wang and R. Jiang. New notions of simultaneous diagonalizability of quadratic forms with applications to QCQPs. *arXiv preprint*, 2101.12141, 2021.

[178] A. L. Wang and F. Kılınç-Karzan. On convex hulls of epigraphs of QCQPs. In *Integer Programming and Combinatorial Optimization (IPCO 2020)*, pages 419–432, 2020.

[179] A. L. Wang and F. Kılınç-Karzan. A geometric view of SDP exactness in QCQPs and its applications. *arXiv preprint*, 2011.07155, 2020.

[180] A. L. Wang and F. Kılınç-Karzan. The generalized trust region subproblem: solution complexity and convex hull results. *Math. Program.*, 191:445–486, 2022.

[181] A. L. Wang and F. Kılınç-Karzan. On the tightness of SDP relaxations of QCQPs. *Math. Program.*, 193:33–73, 2022.

[182] A. L. Wang and F. Kılınç-Karzan. Accelerated first-order methods for a class of semidefinite programs. *arXiv preprint*, 2206.00224, 2022.

[183] A. L. Wang, Y. Lu, and F. Kılınç-Karzan. Implicit regularity and linear convergence rates for the generalized trust-region subproblem. *arXiv preprint*, 2112.13821, 2021.

[184] J. Wang and Y. Xia. A linear-time algorithm for the Trust Region Subproblem based on hidden convexity. *Optim. Lett.*, 11(8):1639–1646, 2017.

[185] J. Wang, H. Chen, R. Jiang, X. Li, and Z. Li. Fast algorithms for stackelberg prediction game with least squares loss. In *International Conference on Machine Learning*, pages 10708–10716, 2021.

[186] J. Wang, W. Huang, R. Jiang, X. Li, and A. L. Wang. Solving stackelberg prediction game with least squares loss via spherically constrained least squares reformulation. In *International Conference on Machine Learning*, 2022. Forthcoming.

[187] K. Weierstrass. Zur Theorie der quadratischen und bilinearen Formen. *Monatsber. Akad. Wiss., Berlin*, pages 310–338, 1868.

[188] H. Wolkowicz, R. Saigal, and L. Vandenberghe. *Handbook of semidefinite programming: Theory, algorithms, and applications*, volume 27 of *Internat. Ser. Oper. Res. Management Sci.* 2012.

[189] Y. Xia. On minimizing the ratio of quadratic functions over an ellipsoid. *Optimization*, 64 (5):1097–1106, 2015.

[190] V. A. Yakubovich. S-procedure in nolinear control theory. *Vestnik Leningrad Univ. Math.*, pages 73–93, 1971.

[191] B. Yang, K. Anstreicher, and S. Burer. Quadratic programs with hollows. *Math. Program.*, 170:541–553, 2018.

[192] H. Yang, L. Liang, L. Carlone, and K. Toh. An inexact projected gradient method with rounding and lifting by nonlinear programming for solving rank-one semidefinite relaxation of polynomial optimization. *arXiv preprint*, 2105.14033, 2021.

[193] Y. Ye. Approximating quadratic programming with bound and quadratic constraints. *Math. Program.*, 84:219–226, 1999.

[194] Y. Ye and S. Zhang. New results on quadratic minimization. *SIAM J. Optim.*, 14(1): 245–267, 2003.

[195] U. Yıldıran. Convex hull of two quadratic constraints is an LMI set. *IMA J. Math. Control Inform.*, 26(4):417–450, 2009.

[196] S. Yıldız and G. Cornuéjols. Disjunctive cuts for cross-sections of the Second-Order Cone. *Oper. Res. Lett.*, 43(4):432—-437, 2015.

[197] A. Yurtsever, O. Fercoq, and V. Cevher. A conditional-gradient-based augmented Lagrangian framework. In *International Conference on Machine Learning*, pages 7272–7281, 2019.

[198] 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.

[199] J. Zhou and Z. Xu. A simultaneous diagonalization based SOCP relaxation for convex quadratic programs with linear complementarity constraints. *Optim. Lett.*, 13(7):1615–1630, 2019.

[200] J. Zhou, S. Chen, S. Yu, and Y. Tian. A simultaneous diagonalization-based quadratic convex reformulation for nonconvex quadratically constrained quadratic program. *Optimization*, pages 1–17, 2020.

A Appendices for Chapter 1

A.1 Proof of Proposition 1

Proposition 1. *For any SD-QCQP, we have*

$$\mathcal{D}_{\text{SOCP}} = \mathcal{D}_{\text{SDP}} \quad \textit{and} \quad \text{Opt}_{\text{SOCP}} = \text{Opt}_{\text{SDP}} .$$

Proof. The second identity follows immediately from the first identity, thus it suffices to prove only the former.

Let $(x,t) \in \mathcal{D}_{\text{SDP}}$. By definition, there exists $X \in \mathbb{S}^N$ such that the following system is satisfied

$$\begin{cases} Y := \begin{pmatrix} 1 & x^\intercal \\ x & X \end{pmatrix} \\ \langle Q_0, Y\rangle \leq 2t \\ \langle Q_i, Y\rangle \leq 0, \ \forall i \in [m_I] \\ \langle Q_i, Y\rangle = 0, \ \forall i \in [m_I+1, m] \\ Y \succeq 0. \end{cases}$$

Taking a Schur complement of 1 in the matrix Y, we see that $X \succeq xx^\intercal$. In particular, we have that $X_{j,j} \geq x_j^2$ for all $j \in [N]$. Define the vector y by $y_j = X_{j,j} \geq x_j^2$. Then, noting that $\langle \text{Diag}(a_i), X\rangle = \langle a_i, y\rangle$ for all $i \in [0,m]$, we conclude that $(x,t) \in \mathcal{D}_{\text{SOCP}}$.

Let $(x,t) \in \mathcal{D}_{\text{SOCP}}$. By definition, there exists $y \in \mathbb{R}^N$ such that the following system is satisfied

$$\begin{cases} \langle a_0, y\rangle + 2\langle b_0, x\rangle + c_0 \leq 2t \\ \langle a_i, y\rangle + 2\langle b_i, x\rangle + c_i \leq 0, \ \forall i \in [m_I] \\ \langle a_i, y\rangle + 2\langle b_i, x\rangle + c_i = 0, \ \forall i \in [m_I+1, m] \\ y_j \geq x_j^2, \ \forall j \in [N]. \end{cases}$$

Define $X \in \mathbb{S}^N$ such that $X_{j,j} = y_j$ for all $j \in [N]$ and $X_{j,k} = x_j x_k$ for $j \neq k$. From the definition of $\mathcal{D}_{\text{SOCP}}$, the relation $y_j \geq x_j^2$ holds for all $j \in [N]$, therefore

$$\begin{pmatrix} 1 & x^\intercal \\ x & X \end{pmatrix} \succeq \begin{pmatrix} 1 & x^\intercal \\ x & xx^\intercal \end{pmatrix} \succeq 0.$$

Finally, noting that $\langle \text{Diag}(a_i), X\rangle = \langle a_i, y\rangle$ for all $i \in [0,m]$, we conclude that $(x,t) \in \mathcal{D}_{\text{SDP}}$. ■

A.2 Proof of Theorem 8

Theorem 8. *Suppose Assumption 1 holds. Define the hyperplane* $H = \left\{(x,t) \in \mathbb{R}^{N+1} : 2t = \text{Opt}_{\text{SDP}}\right\}$. *If the quadratic eigenvalue multiplicity k satisfies $k \geq m+1$, then* $\text{conv}(\mathcal{D} \cap H) = \mathcal{D}_{\text{SDP}} \cap H$. *In particular,* $\text{Opt} = \text{Opt}_{\text{SDP}}$.

Proof. Suppose $(\hat{x},\hat{t})\in\mathcal{D}_{\text{SDP}}\cap H$. Then by Lemma 1 and optimality of $\hat{t}$, we have that $2\hat{t}=\sup_{\gamma\in\Gamma}q(\gamma,\hat{x})$, i.e.,

$$\begin{aligned}2\hat{t}&=\sup_{\gamma\in\mathbb{R}^m}\left\{q(\gamma,\hat{x}):\ \begin{array}{l}A(\gamma)\succeq 0\\ \gamma_i\geq 0,\ \forall i\in[m_I]\end{array}\right\}\\&=\sup_{\gamma\in\mathbb{R}^m}\left\{q(\gamma,\hat{x}):\ \begin{array}{l}\mathcal{A}(\gamma)\succeq 0\\ \gamma_i\geq 0,\ \forall i\in[m_I]\end{array}\right\}.\end{aligned}$$

The second line follows as $A(\gamma)\succeq 0$ if and only if $\mathcal{A}(\gamma)\succeq 0$. Note that Assumption 1 allows us to apply strong conic duality to the program on the second line. Furthermore, this dual SDP achieves its optimal value, i.e., there exists $Z\in\mathbb{S}^n$ such that $(\hat{x},\hat{t},Z)$ satisfies

$$\begin{cases}q_0(\hat{x})+\langle\mathcal{A}_0,Z\rangle=2\hat{t}\\ q_i(\hat{x})+\langle\mathcal{A}_i,Z\rangle\leq 0,\ \forall i\in[m_I]\\ q_i(\hat{x})+\langle\mathcal{A}_i,Z\rangle=0,\ \forall i\in[m_I+1,m]\\ Z\succeq 0.\end{cases}\tag{1}$$

We will show by induction on $\operatorname{rank}(Z)$ that for any $(\hat{x},\hat{t},Z)$ satisfying (1), we have $(\hat{x},\hat{t})\in\operatorname{conv}(\mathcal{D}\cap H)$. The claim clearly holds when $\operatorname{rank}(Z)=0$.

Now suppose $r\coloneqq\operatorname{rank}(Z)\geq 1$. Let $(\hat{x},\hat{t},Z)$ satisfy (1). Write $Z=\sum_{i=1}^r z_iz_i^\intercal$ where each z_i is nonzero. Fix $z\coloneqq z_1$.

We claim that the following system in y is feasible:

$$\begin{cases}\langle A_i\hat{x}+b_i,y\otimes z\rangle=0,\ \forall i\in[m]\\ y\in\mathbf{S}^{k-1}.\end{cases}\tag{2}$$

Indeed, the linear constraints impose at most m homogeneous linear equalities in $k\geq m+1$ variables. In particular, there exists a nonzero solution y to the linear constraints. This y may then be scaled to satisfy $y\in\mathbf{S}^{k-1}$.

Note then that for all $i\in[m]$,

$$\begin{aligned}q_i(\hat{x}\pm y\otimes z)+\langle\mathcal{A}_i,Z-zz^\intercal\rangle&=(\hat{x}\pm y\otimes z)^\intercal A_i(\hat{x}\pm y\otimes z)+2b_i^\intercal(\hat{x}\pm y\otimes z)+c_i+\langle\mathcal{A}_i,Z-zz^\intercal\rangle\\&=q_i(\hat{x})\pm 2\langle A_i\hat{x}+b_i,y\otimes z\rangle+\langle\mathcal{A}_i,Z\rangle\\&=q_i(\hat{x})+\langle\mathcal{A}_i,Z\rangle.\end{aligned}$$

Consequently, $(\hat{x}\pm y\otimes z,\hat{t},Z-zz^\intercal)$ satisfies all of the constraints in (1) except possibly the first. We now verify that the first constraint is also satisfied: From

$$\begin{aligned}q_0(\hat{x}\pm y\otimes z)+\langle\mathcal{A}_0,Z-zz^\intercal\rangle&=q_0(\hat{x})\pm 2\langle A_0\hat{x}+b_0,y\otimes z\rangle+\langle\mathcal{A}_0,zz^\intercal\rangle+\langle\mathcal{A}_0,Z-zz^\intercal\rangle\\&=q_0(\hat{x})+\langle\mathcal{A}_0,Z\rangle\pm 2\langle A_0\hat{x}+b_0,y\otimes z\rangle\\&=2\hat{t}\pm 2\langle A_0\hat{x}+b_0,y\otimes z\rangle,\end{aligned}$$

we deduce that $(\hat{x} \pm y \otimes z, 2\hat{t} \pm 2\langle A_0\hat{x} + b_0, y \otimes z\rangle) \in \mathcal{D}_{\text{SDP}}$. Then, by minimality of $\hat{t}$ in $\mathcal{D}_{\text{SDP}}$, we infer that $\langle A_0\hat{x} + b_0, y \otimes z\rangle = 0$.

We deduce that $(\hat{x} \pm y \otimes z, \hat{t}, Z - zz^\intercal)$ satisfies (1). Furthermore, we have $\operatorname{rank}(Z - zz^\intercal) = r - 1$. By induction, $(\hat{x} \pm y \otimes z, \hat{t}) \in \operatorname{conv}(\mathcal{D} \cap H)$. We conclude that $(\hat{x}, \hat{t}) \in \operatorname{conv}(\mathcal{D} \cap H)$. ■

B Appendices for Chapter 2

B.1 Deferred proofs from Section 2.4

Deferred proofs from Section 2.4.1

We compute

$$\begin{aligned}
\Gamma &= \left\{(\gamma_{\text{obj}}, \gamma_1, \gamma_2) \in \mathbb{R}_+ \times \mathbb{R}^2 : \begin{pmatrix} \gamma_1 & \gamma_2/\sqrt{2} \\ \gamma_2/\sqrt{2} & \gamma_{\text{obj}} \end{pmatrix} \succeq 0\right\} \\
&= \left\{(\gamma_{\text{obj}}, \gamma_1, \gamma_2) \in \mathbb{R}^3 : \begin{array}{l} \gamma_{\text{obj}} + \gamma_1 \geq 0 \\ 2\gamma_{\text{obj}}\gamma_1 \geq \gamma_2^2 \end{array}\right\} \\
&= \left\{(\gamma_{\text{obj}}, \gamma_1, \gamma_2) \in \mathbb{R}^3 : \gamma_{\text{obj}} + \gamma_1 \geq \sqrt{(\gamma_{\text{obj}} - \gamma_1)^2 + (\sqrt{2}\gamma_2)^2}\right\}.
\end{aligned}$$

The expression for Γ° follows from Γ.

Proof of (2.5). Let $(x,t) \in \mathcal{S}_{\text{SDP}} \setminus \mathcal{S}$ such that $\mathcal{G}(x,t)$ is a one-dimensional face of Γ°. For notational convenience, let $\ell_{\text{obj}} = q_{\text{obj}}(x) - 2t$, $\ell_1 = q_1(x)$ and $\ell_2 = q_2(x)$. Note that $\mathcal{G}(x,t) = \mathbb{R}_+(\ell_{\text{obj}}, \ell_1, \ell_2)$ so that $\mathcal{F}(x,t) = \mathbb{R}_+(-\ell_1, -\ell_{\text{obj}}, \ell_2)$. Furthermore, by the assumption that $(\ell_{\text{obj}}, \ell_1, \ell_2)$ is nonzero and on the boundary of Γ°, we have

$$\mathcal{G}(x,t)^\perp = \operatorname{span}\left\{\begin{pmatrix} -\ell_1 \\ -\ell_{\text{obj}} \\ \ell_2 \end{pmatrix}, \begin{pmatrix} \ell_2 \\ -\ell_2 \\ \ell_1 - \ell_{\text{obj}} \end{pmatrix}\right\}.$$

We deduce that

$$\mathcal{R}'(x,t) = \left\{\begin{pmatrix} -\ell_{\text{obj}} \\ \ell_2/\sqrt{2} \\ 0 \end{pmatrix}, \begin{pmatrix} \ell_2/\sqrt{2} \\ -\ell_1 \\ 0 \end{pmatrix},\right. \tag{3}$$

$$\left.\begin{pmatrix} -\ell_{\text{obj}}(2x_1 - 1) + \ell_2(\sqrt{2}x_2) \\ -\ell_1(2x_2) + \sqrt{2}\ell_2(x_1 - 1) \\ 2\ell_1 \end{pmatrix}, \begin{pmatrix} -\ell_2(2x_1 - 1) + (\ell_1 - \ell_{\text{obj}})(\sqrt{2}x_2) \\ \ell_2(2x_2) + \sqrt{2}(\ell_1 - \ell_{\text{obj}})(x_1 - 1) \\ -2\ell_2 \end{pmatrix}\right\}^\perp. \tag{4}$$

Here, the first two vectors span $\operatorname{span}(A(f_{\text{obj}}, f))$. The second two vectors correspond to the constraints $\left\langle A(\gamma_{\text{obj}}, \gamma)x, x'\right\rangle - \gamma_{\text{obj}}t' = 0$ for $(\gamma_{\text{obj}}, \gamma) \in \mathcal{G}(x,t)^\perp$.

Below, we will simplify this expression. By the assumption that $(\ell_{\text{obj}}, \ell_1, \ell_2)$ is nonzero and on the boundary of Γ°, we have

$$-\ell_{\text{obj}} - \ell_1 = \sqrt{(\ell_{\text{obj}} - \ell_1)^2 + (\sqrt{2}\ell_2)^2}$$

where the term within the radical is nonzero. Expanding, we deduce that

$$\begin{cases} \ell_{\text{obj}} + \ell_1 < 0 \\ \ell_2^2 - 2\ell_{\text{obj}}\ell_1 = 0 \end{cases} = \begin{cases} x_2^2 - 2t + x_1(x_1 - 1) < 0 \\ (x_2^2 - 2tx_1)(x_1 - 1) = 0 \end{cases}.$$

Note that $(0, 1, 0) \in \Gamma$ so that $x_1 \in [0, 1]$. If $x_1 = 1$, then $\ell_{\text{obj}} < 0$, $\ell_1 = 0$ and $\ell_2 = 0$ so that $(x_1, x_2, t) \in \mathcal{S}$, a contradiction. We deduce $1 - x_1 > 0$ and $x_2^2 - 2tx_1 = 0$ and that $(\ell_{\text{obj}}, \ell_1, \ell_2) = (x_1 - 1)(2t, x_1, \sqrt{2}x_2)$. Plugging this into (3) gives

$$\mathcal{R}'(x,t) = \left\{ \begin{pmatrix} -2t \\ x_2 \\ 0 \end{pmatrix}, \begin{pmatrix} x_2 \\ -x_1 \\ 0 \end{pmatrix}, \begin{pmatrix} t \\ -x_2 \\ x_1 \end{pmatrix}, \begin{pmatrix} x_2(x_1 - 1 + 2t) \\ -x_1^2 + x_1 - 2tx_1 - 2t \\ 2x_2 \end{pmatrix} \right\}^\perp. \qquad \blacksquare$$

Deferred proofs from Section 2.4.3

We will prove Proposition 8 in the following series of lemmas. Note that the first identity of Proposition 8 follows from definition. To prove the second identity of Proposition 8, we will partition Γ_1 into $n+1$ pieces depending on the sign pattern of $\gamma \in \Gamma_1$.

Note that $\gamma \in \Gamma_1$ if and only if $aa^\top + \operatorname{Diag}(\gamma) \succeq 0$. In particular, $\gamma \in \Gamma_1$ if γ is nonnegative. On the other hand, by the Eigenvalue Interlacing Theorem, $\gamma \notin \Gamma_1$ if it has at least two negative coordinates. It remains to understand $\mathcal{N}_i := \Gamma_1 \cap \{\gamma \in \mathbb{R}^n : \gamma_i < 0,\ \gamma_i \geq 0,\ \forall j \neq i\}$. The next lemma follows from a straightforward application of the Schur Complement Lemma and the Sherman–Morrison Formula.

Lemma 86. *Suppose Assumption 6 holds. Then, for any $i \in [n]$,*

$$\mathcal{N}_i = \left\{ \gamma \in \mathbb{R}^n : \begin{array}{l} 0 > \gamma_i \geq \frac{-a_i^2}{1 + \sum_{j \neq i} a_j^2/\gamma_j} \\ \gamma_j > 0,\ \forall j \neq i \end{array} \right\}.$$

Proof. Without loss of generality we assume $i = n$. For convenience, let $\bar{\gamma}$ and $\bar{a}$ denote the first $n-1$ entries of γ and a respectively. By Assumption 6, we have that $\gamma_j > 0$ for all $j < n$ (as otherwise the 2×2 minor of $aa^\top + \operatorname{Diag}(\gamma)$ corresponding to (j, n) is not positive semidefinite). The Schur Complement Lemma and the Sherman–Morrison Formula then imply that $\gamma \in \mathcal{N}_n$ if and only if $\gamma_n < 0$, $\bar{\gamma} > 0$ and

$$\begin{aligned} \gamma_n + a_n^2 &\geq a_n^2 \bar{a}^\top (\bar{a}\bar{a}^\top + \operatorname{Diag}(\bar{\gamma}))^{-1} \bar{a} \\ &= a_n^2 \bar{a}^\top \left(\operatorname{Diag}(\bar{\gamma})^{-1} - \tfrac{\operatorname{Diag}(\bar{\gamma})^{-1} \bar{a}\bar{a}^\top \operatorname{Diag}(\bar{\gamma})^{-1}}{1 + \bar{a}^\top \operatorname{Diag}(\bar{\gamma})^{-1} \bar{a}} \right) \bar{a} \\ &= a_n^2 \tfrac{\bar{a}^\top \operatorname{Diag}(\bar{\gamma})^{-1} \bar{a}}{1 + \bar{a}^\top \operatorname{Diag}(\bar{\gamma})^{-1} \bar{a}}. \end{aligned}$$

Rearranging terms completes the proof. ■

Then decomposing $\Gamma_1 = \mathbb{R}^n_+ \cup \bigcup_{i\in[n]} \mathcal{N}_i$, we get

$$\mathcal{S}_{\text{SDP}} = \left\{ (x,t) \in \mathbb{R}^n : \begin{array}{l} 2t \geq \max_{i\in[n]} \sup_{\gamma\in\mathcal{N}_i} [\gamma, q(x)] \\ x \in [\pm 1]^n \end{array} \right\}$$

It remains to prove the following lemma.

Lemma 87. *Suppose Assumption 6 holds and let $i \in [n]$. For any $x \in [\pm 1]^n$, we have*

$$\sup_{\gamma\in\mathcal{N}_i} [\gamma, q(x)] = (a^\top x)^2 + \left(a_i\sqrt{1-x_i^2} - \sum_{j\neq i} a_j \sqrt{1-x_j^2} \right)_+^2. \tag{5}$$

We will need the following two useful facts.

Lemma 88. *Let $\xi \in \mathbb{R}^k_-$ and $\alpha > 0$, then*

$$\sup_{\zeta\in\mathbb{R}^k_{++}} \left\{ \sum_{i=1}^k \frac{\xi_i}{\zeta_i^2} : \sum_{i=1}^k \zeta_i^2 = \alpha \right\} = -\frac{1}{\alpha}\left(\sum_{i=1}^k \sqrt{-\xi_i} \right)^2. \tag{6}$$

Proof. Without loss of generality, we may assume $\xi \in \mathbb{R}^k_{--}$. Then by Cauchy-Schwarz, we have $-\sum_{i=1}^k \xi_i/\zeta_i^2 = -\frac{1}{\alpha}\left(\sum_{i=1}^k \xi_i/\zeta_i^2\right)\left(\sum_{i=1}^k \zeta_i^2\right) \geq \frac{1}{\alpha}\left(\sum_{i=1}^k \sqrt{-\xi_i}\right)^2$. Furthermore, equality holds when $\zeta_i^2 \propto \sqrt{\xi_i}$. ■

Lemma 89. *Let $\alpha, \beta \geq 0$, then*

$$\sup_{x>0} \left(\frac{\alpha}{1+x} - \frac{\beta}{x} \right) = \left(\sqrt{\alpha} - \sqrt{\beta} \right)_+^2.$$

Proof. Let $f(x) \coloneqq \alpha(1+x)^{-1} - \beta x^{-1}$. Note that $\frac{d}{dx}f(x) = -\alpha(1+x)^{-2} + \beta x^{-2}$. There are three cases to consider. If $\beta = 0$, then $f(x) = \alpha(1+x)^{-1}$ and $\sup_{x>0} \alpha(1+x)^{-1} = \alpha$. Next, suppose $0 \leq \alpha \leq \beta$, then $\frac{d}{dx}f(x) = -\alpha(1+x)^{-2} + \beta x^{-2} \geq \beta(x^{-2} - (1+x)^{-2}) \geq 0$ so that $\sup_{x>0} f(x) = \lim_{x\to\infty} f(x) = 0$. Finally, suppose $0 < \beta < \alpha$. Note that $f'(x) > 0$ for all x small enough. Similarly, $f'(x) < 0$ for all x large enough. We deduce that $\sup_{x>0} f(x)$ is achieved. Computing the first-order-necessary conditions, we see that $f(x)$ is maximized at $\frac{\sqrt{\beta}}{\sqrt{\alpha}-\sqrt{\beta}}$ with value $\left(\sqrt{\alpha} - \sqrt{\beta}\right)^2$. ■

Proof of Lemma 87. Without loss of generality, $i = n$. Let $b \in \mathbb{R}^n_-$ where $b_j = x_j^2 - 1$. Let $\bar{\gamma}$ denote the first $n-1$ entries of γ. Then,

$$\begin{aligned}\sup_{\gamma\in\mathcal{N}_n} [\gamma, q(x)] - q_{\mathrm{obj}}(x) = \sup_{\gamma\in\mathcal{N}_n} \langle \gamma, b\rangle &= \sup_{\bar{\gamma}\in\mathbb{R}^{n-1}_{++}} \sum_{i=1}^{n-1} \gamma_i b_i - \frac{a_n^2 b_n}{1+\sum_{i=1}^{n-1} a_i^2/\gamma_i}\\ &= \sup_{\alpha>0}\left(-\frac{a_n^2 b_n}{1+\alpha} + \sup_{\zeta\in\mathbb{R}^{n-1}_{++}}\left\{\sum_{i=1}^{n-1}\frac{a_i^2 b_i}{\zeta_i} : \sum_{i=1}^{n-1}\zeta_i = \alpha\right\}\right)\\ &= \sup_{\alpha>0}\left(\frac{\left(a_n\sqrt{-b_n}\right)^2}{1+\alpha} - \frac{1}{\alpha}\left(\sum_{i=1}^{n-1} a_i\sqrt{-b_i}\right)^2\right)\\ &= \left(a_n\sqrt{-b_n} - \sum_{i=1}^{n-1} a_i\sqrt{-b_i}\right)_+^2.\end{aligned}$$

Here, the second line follows from a change of variables of $\zeta_i := a_i^2/\gamma_i$ and $\alpha := \sum_{i=1}^{n-1}\zeta_i$. The third line follows from Lemma 88 and the fourth line follows from Lemma 89. ■

Proof of Corollary 7. Let $\gamma \in \Gamma_1$ and $x \in \mathbb{R}^n$. By convexity of $[\gamma, q(x)]$ in x and the fact that $q(x) = q(-x)$, we deduce that $[\gamma, q(0)] \leq [\gamma, q(x)]$. We deduce that $\mathrm{Opt}_{\mathrm{SDP}} = \inf_x \sup_{\gamma\in\Gamma_1}[\gamma, q(x)] = \sup_{\gamma\in\Gamma_1}[\gamma, q(0)]$. By Proposition 8, we conclude that

$$\mathrm{Opt}_{\mathrm{SDP}} = \max_{i\in[n]}\left(a_i - \sum_{j\neq i} a_j\right)_+^2. \qquad ■$$

Proof of Corollary 8. Pick an open set $U \subseteq [\pm 1]^n$ such that

$$a_1(1-x_1^2) > \sum_{j>1} a_j(1-x_j^2),\ \forall x \in U.$$

Then by Proposition 8, for any $x \in U$, we have $(x, t) \in \mathcal{S}_{\mathrm{SDP}}$ if and only if

$$2t \geq f(x) := (a^\top x)^2 + \left(a_1\sqrt{1-x_1^2} - \sum_{j>1} a_j\sqrt{1-x_j^2}\right)^2.$$

Note that $f(x)$ is smooth on U and nonlinear (for example note $\frac{\partial^2 f(x)}{\partial x^2} \neq 0$). We conclude that $\mathcal{S}_{\mathrm{SDP}} \neq \mathrm{conv}(\mathcal{S})$ as $\mathrm{conv}(\mathcal{S})$ is polyhedral. ■

B.2 Deferred proofs from Section 2.6

Useful lemmas

We first recall that under some minor conditions, pointwise convergence implies uniform convergence for convex functions. We extend this statement to show that pointwise *a.a.s.* convergence implies *a.a.s.* uniform convergence.

Lemma 90. *Let $\Omega \subseteq \mathbb{R}^n$ be an open set and let $f : \Omega \to \mathbb{R}$ be a convex function. Suppose $g_1, g_2, \ldots : \Omega \to \mathbb{R}$ is a sequence of random convex functions such that for all $x \in \Omega$ and $\epsilon > 0$, we have that a.a.s.,*

$$|g_i(x) - f(x)| \leq \epsilon.$$

Then, for any compact $C \subseteq \Omega$ and $\epsilon > 0$, we have that a.a.s.,

$$|g_i(x) - f(x)| \leq \epsilon, \, \forall x \in C.$$

Proof. Fix $C \subseteq \Omega$ compact. Without loss of generality, we will assume that $\epsilon > 0$ satisfies $C + B(0, 3\epsilon) \subseteq \Omega$ and that f is 1-Lipschitz on $C + B(0, 3\epsilon)$.

Fix a finite net $\mathcal{N} \subseteq C + B(0, 3\epsilon)$ such that for all $x \in C + B(0, 2\epsilon)$, we have $x \in \operatorname{conv}(\mathcal{N} \cap B(x, \epsilon))$. By our assumption and the fact that $\mathcal{N}$ is finite, we have that *a.a.s.*, $|f(x) - g_i(x)| \leq \epsilon$ for all $x \in \mathcal{N}$. We condition on this event in the remainder of the proof.

For any $x \in C$, let $x = \sum_j \lambda_j x_j$ denote the convex decomposition guaranteed by $x \in \operatorname{conv}(\mathcal{N} \cap B(x, \epsilon))$. Then,

$$g_i(x) \leq \sum_j \lambda_j g_i(x_j) \leq \sum_j \lambda_j (f(x_j) + \epsilon) \leq f(x) + 2\epsilon.$$

Here, the last inequality follows from $f(x_j) \leq f(x) + \|x - x_j\|_2 \leq f(x) + \epsilon$.

Let $x \in C$ and $x' \in \mathcal{N} \cap B(x, \epsilon)$. Note that $y := x' + (x' - x) \in C + B(0, 2\epsilon)$. By construction, there exists $y' \in \mathcal{N} \cap B(y, \epsilon)$ such that $g_i(y') \geq g_i(y)$. Finally,

$$f(x) + 4\epsilon \geq f(y') + \epsilon \geq g_i(y') \geq g_i(y) \geq 2g_i(x') - g_i(x) \geq 2(f(x') - \epsilon) - g_i(x) \geq 2f(x) - g_i(x) - 4\epsilon.$$

Therefore, by rearranging and combining, we deduce that *a.a.s.*, $|g_i(x) - f(x)| \leq 8\epsilon, \, \forall x \in C$. ■

Lemma 91. *Let $r \in [-1, 1]$, then*

$$-\int_{\sigma=-1}^{1} \frac{r^2}{1 + r\sigma} \, d\mu_{\text{nsc}}(\sigma) = 2(\sqrt{1 - r^2} - 1) = \phi(r).$$

Proof. We begin by expanding the definition of μ_{nsc} and substituting $\sigma = -\cos\theta$:

$$\begin{aligned}
-\int_{\sigma=-1}^{1} \frac{r^2}{1+r\sigma}\, d\mu_{\text{nsc}}(\sigma) &= -\frac{2}{\pi}\int_{\sigma=-1}^{1} \frac{r^2\sqrt{1-\sigma^2}}{1+r\sigma}\, d\sigma \\
&= -\frac{2}{\pi}\int_{\theta=0}^{\pi} \frac{r^2\sin^2\theta}{1-r\cos\theta}\, d\theta \\
&= -\frac{2}{\pi}\int_{\theta=0}^{\pi} \frac{r^2 - r^2\cos^2\theta}{1-r\cos\theta}\, d\theta \\
&= -\frac{2}{\pi}\int_{\theta=0}^{\pi} \frac{r^2-1}{1-r\cos\theta}\, d\theta - \frac{2}{\pi}\int_{\theta=0}^{\pi} (1+r\cos\theta)\, d\theta \\
&= \frac{2(1-r^2)}{\pi}\left(\int_{\theta=0}^{\pi} \frac{1}{1-r\cos\theta}\, d\theta\right) - 2. \qquad (7)
\end{aligned}$$

We now focus on the bracketed integral. Perform the change of variables $\theta = 2\eta$ to get

$$\int_{\theta=0}^{\pi} \frac{1}{1-r\cos\theta}\, d\theta = 2\int_{\eta=0}^{\pi/2} \frac{1}{1-r\cos(2\eta)}\, d\eta. \qquad (8)$$

Recalling the identities $\cos(2\eta) = 2\cos^2(\eta) - 1$ and $\cos^{-2}\eta = \sec^2\eta = \tan^2\eta + 1 = \frac{d}{d\eta}\tan(\eta)$, we then have

$$\frac{1}{1-r\cos(2\eta)} = \frac{1}{1+r-2r\cos^2\eta} = \frac{\frac{d}{d\eta}\tan\eta}{(1+r)\tan^2\eta + (1-r)}.$$

Performing one last change of variables $t = \tan\eta$ gives

$$\begin{aligned}
2\int_{\eta=0}^{\pi/2} \frac{1}{1-r\cos(2\eta)}\, d\eta &= 2\int_{\eta=0}^{\pi/2} \frac{\frac{d}{d\eta}\tan\eta}{(1+r)\tan^2\eta + (1-r)}\, d\eta \\
&= 2\int_{t=0}^{\infty} \frac{1}{(1+r)t^2 + (1-r)}\, dt \\
&= 2\left.\frac{\arctan\left(t\sqrt{\frac{1+r}{1-r}}\right)}{\sqrt{1-r^2}}\right|_{t=0}^{\infty} \\
&= \frac{\pi}{\sqrt{1-r^2}}. \qquad (9)
\end{aligned}$$

Combining (7), (8), and (9) gives the desired identity. ■

Deferred proofs from Section 2.6.2

Proof of Lemma 15. Let $\Omega = \mathbb{R}^m$ and set $f(\gamma) \coloneqq 1 - \|\gamma\|_2$. Note that f and $\lambda_{\min}(A[\gamma])$ are both concave functions on Ω. We have $\lambda_{\min}(A[0]) = 1 = f(0)$. Furthermore, for any nonzero $\gamma \in \mathbb{R}^m$ and $\epsilon > 0$,

$$\lambda_{\min}(A[\gamma]) = 1 + \|\gamma\|_2 \lambda_{\min}\left(\sum_{i=1}^m \frac{\gamma_i}{\|\gamma\|_2} A_i\right) \in 1 + \|\gamma\|_2[-1 \pm \epsilon] = [f(\gamma) \pm \|\gamma\|_2 \epsilon], \text{ *a.a.s.*.}$$

Here, the inclusion holds by Facts 4 and 6. Taking $C = r\mathbf{S}^{m-1}$ and applying Lemma 90. ■

Proof of Lemma 16. Fix $r \in (0,1)$. Without loss of generality, $r + 2\epsilon < 1$. Set $\Omega \coloneqq \{\gamma \in \mathbb{R}^m : \|\gamma\|_2 < r + 2\epsilon\}$. Let $\hat\gamma \in \Omega$. Note that we may generate $A[\hat\gamma]$ and $b[\hat\gamma]$ via the following equivalent process: Sample $\bar A \sim \mathrm{NGOE}(n)$ and $\bar b \sim N(0, I_n/n)$ independently and set $A[\hat\gamma] \coloneqq I + r\bar A$ and $b[\hat\gamma] \coloneqq r\bar b$. With this notation, $-b[\hat\gamma]A[\hat\gamma]^{-1}b[\hat\gamma] = -r^2 \bar b^\top (I + r\bar A)^{-1}\bar b$. Let $\bar A = \sum_{i=1}^n \sigma_i v_i v_i^\top$ be the eigenvalue decomposition of $\bar A$ and let $\mu_{\bar A}$ denote its Empirical Spectral Distribution. By Lemma 91, we have

$$\begin{aligned}
&\frac{1}{r^2}\left|-b[\hat\gamma]^\top A[\hat\gamma]^{-1} b[\hat\gamma] - \phi(r)\right| \\
&\quad = \frac{1}{r^2}\left|-b[\hat\gamma]^\top A[\hat\gamma]^{-1} b[\hat\gamma] + \int_{\sigma=-1}^1 \frac{r^2}{1 + r\sigma} d\mu_{\text{nsc}}\right| \\
&\quad = \left|\bar b^\top (I + r\bar A)^{-1} \bar b - \int_{\sigma=-1}^1 \frac{1}{1 + r\sigma} d\mu_{\text{nsc}}\right| \\
&\quad \leq \left|\sum_{i=1}^n \frac{\left(v_i^\top \bar b\right)^2 - 1/n}{1 + r\sigma_i}\right| + \left|\int \frac{1}{1 + r\sigma} d\mu_{\bar A}(\sigma) - \int \frac{1}{1 + r\sigma} d\mu_{\text{nsc}}(\sigma)\right|,
\end{aligned}$$

where the last inequality follows from the identity $(I + r\bar A)^{-1} = \sum_{i=1}^n \frac{1}{1 + r\sigma_i} v_i v_i^\top$ and Cauchy-Schwartz inequality. Note that by Fact 6, for all $i \in [n]$ we have that $1 + r\sigma_i \geq 1 - r - r\epsilon \geq 1 - r - \epsilon > \epsilon$ *a.a.s.*. We will compute the mean and variance of the first term conditioned on this event. By independence of $\bar b$ and $\bar A$,

$$\mathop{\mathbb{E}}_{\bar b}\left[\sum_{i=1}^n \frac{\left(v_i^\top \bar b\right)^2 - 1/n}{1 + r\sigma_i} \,\middle|\, 1 + r\sigma_i \geq \epsilon,\ \forall i\right] = \sum_{i=1}^n \left(\frac{1}{1 + r\sigma_i}\right) \mathop{\mathbb{E}}_{\bar b}\left[\left(v_i^\top \bar b\right)^2 - \frac{1}{n} \,\middle|\, 1 + r\sigma_i \geq \epsilon,\ \forall i\right] = 0, \text{ and}$$

$$\mathop{\mathbb{E}}_{\bar b}\left[\left(\sum_{i=1}^n \frac{\left(v_i^\top \bar b\right)^2 - 1/n}{1 + r\sigma_i}\right)^2 \,\middle|\, 1 + r\sigma_i \geq \epsilon,\ \forall i\right] \leq \left(\frac{1}{\epsilon}\right) \mathop{\mathbb{E}}_{\bar b}\left[\left(\sum_{i=1}^n \left(v_i^\top \bar b\right)^2 - 1/n\right)^2 \,\middle|\, 1 + r\sigma_i \geq \epsilon,\ \forall i\right] = \frac{2}{\epsilon n}.$$

In particular, the first term can be bounded by $\epsilon/(2r^2)$ *a.a.s.*.

For the second term, define the C_c^∞ function

$$\psi(x) := \begin{cases} \frac{1}{1+rx}, & \text{if } |x| \le 1+\delta \\ 0, & \text{if } |x| \ge 1+2\delta \\ C_c^\infty, & \text{else.} \end{cases}$$

By Fact 6, we have that *a.a.s.* $\int \frac{1}{1+r\sigma} d\mu_{\bar{A}}(\sigma) = \int \psi(\sigma) d\mu_{\bar{A}}(\sigma)$. Applying Fact 5, we conclude that the second term can be bounded by $\epsilon/(2r^2)$ *a.a.s.*.

Combining the two bounds shows that for any $\gamma \in \Omega$ and $\epsilon > 0$, $\left|-b[\gamma]^\top A[\gamma]^{-1} b[\gamma] - \phi(\gamma)\right| \le \epsilon$ *a.a.s.*. Applying Lemma 90 with $C = r\mathbf{S}^{m-1}$ concludes the proof. ■

Deferred proofs from Section 2.6.3

Lemma 92. *Fix $\epsilon > 0$ and $N \in \mathbb{N}$. Let $A \sim \mathrm{NGOE}(n)$. Then, a.a.s. there exists a N-dimensional vector space $W \subseteq \mathbb{R}^n$ such that*

$$w^\top A w \in [1 \pm \epsilon]\|w\|^2, \; \forall w \in W.$$

Proof. Let ψ denote a C_c^∞ function from $\mathbb{R}$ to $[0,1]$ that takes the value one on $[1 \pm \epsilon/2]$ and the value zero outside of $[1 \pm \epsilon]$. Note that $\theta := \int \psi d\mu_{\text{nsc}}$ is some positive constant independent of n. Let W denote the vector space corresponding to the eigenvalues of A in the range $[1 \pm \epsilon]$. Clearly $w^\top A w \in [1 \pm \epsilon]\|w\|_2^2$ for all $w \in W$. It remains to note that by Fact 5, we have *a.a.s.*

$$\frac{\dim(W)}{n} = \frac{|\{i \in [n] : \lambda_i(A) \in [1 \pm \epsilon]\}|}{n} \ge \int \psi d\mu_{\bar{A}} \ge \int \psi d\mu_{\text{nsc}} - \theta/2 = \theta/2$$

so that $\dim(W) \ge N$ *a.a.s.*. ■

Proof of Lemma 18. Let $\mathcal{N}$ denote a finite ϵ-net on $\mathbf{S}^m \subseteq \mathbb{R}^{1+m}$. By Lemma 92, *a.a.s.*, for every $(\gamma_{\text{obj}}, \gamma) \in \mathcal{N}$, there exists an N dimensional subspace W such that

$$w^\top A(\gamma_{\text{obj}}, \gamma, 1) w \in [\pm\epsilon]\|w\|^2, \; \forall w \in W.$$

Furthermore, by Lemma 15, we have that *a.a.s.* $\left\|A(\gamma_{\text{obj}}, \gamma, 0)\right\|_2 \in \left\|(\gamma_{\text{obj}}, \gamma)\right\|[1 \pm \epsilon]$ for all $(\gamma_{\text{obj}}, \gamma) \in \mathbb{R}^m$. We condition on these two events.

Now, let $(\gamma_{\text{obj}}, \gamma) \in \mathbf{S}^m$ and let $(\gamma'_{\text{obj}}, \gamma') \in \mathcal{N} \cap B((\gamma_{\text{obj}}, \gamma), \epsilon)$. Let W denote the N-dimensional subspace guaranteed for $(\gamma'_{\text{obj}}, \gamma')$. Then for all $w \in W$,

$$w^\top A(\gamma_{\text{obj}}, \gamma, 1) w = w^\top A(\gamma'_{\text{obj}}, \gamma', 1) w + w^\top A(\gamma_{\text{obj}} - \gamma'_{\text{obj}}, \gamma - \gamma') w \in [\pm 3\epsilon]\|w\|^2. \blacksquare$$

C Appendices for Chapter 3

C.1 Proof of Lemma 35

For completeness we restate Lemma 35.

Lemma 35. *Let $\mathcal{M} = \{M_1, M_2\}$. Suppose Assumption 7 holds and $n = 3$. If neither conditions (i) nor (ii) of Theorem 17 hold, then $\mathcal{N}(\mathcal{M})$ is the union of at most four one-dimensional subspaces of $\mathbb{R}^3$.*

Proof. As $\alpha_1 M_1 + \alpha_2 M_2 \notin \mathbb{S}^3_+$ for any $(\alpha_1, \alpha_2) \neq (0,0)$, we have that M_1 and M_2 must each have rank either two or three. We will break the proof into two cases.

Suppose first that $\operatorname{rank}(M_1) = \operatorname{rank}(M_2) = 2$. As $M_1, M_2 \notin \mathbb{S}^3_+$, each M_i has exactly one positive and one negative eigenvalue. We can then write $M_1 = \operatorname{Sym}(ab^\intercal)$ and $M_2 = \operatorname{Sym}(cd^\intercal)$. Then

$$\begin{aligned}\mathcal{N}(\mathcal{M}) &= \{x : \ x^\intercal(ab^\intercal)x = x^\intercal(cd^\intercal)x = 0\}\\ &= (a^\perp \cup b^\perp) \cap (c^\perp \cup d^\perp)\\ &= (a^\perp \cap c^\perp) \cup (a^\perp \cap d^\perp) \cup (b^\perp \cap c^\perp) \cup (b^\perp \cap d^\perp).\end{aligned}$$

As condition (ii) does not hold, each of the four spaces on the final line have dimension one. Thus $\mathcal{N}(\mathcal{M})$ is the union of at most four distinct lines.

Next suppose without loss of generality that $\operatorname{rank}(M_1) = 3$. As $M_1 \notin \mathbb{S}^3_+$, we may assume that it has two positive eigenvalues and one negative eigenvalue. Performing a change of basis, it suffices to consider when

$$M_1 = \begin{pmatrix} 1 & & \\ & 1 & \\ & & -1 \end{pmatrix} \qquad \text{and} \qquad M_2 = \begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix}.$$

We will consider the intersection $\mathcal{N}(\mathcal{M}) \cap \{x \in \mathbb{R}^3 : \ x_3 = 1\}$. Note that if $x \in \mathcal{N}(\mathcal{M})$ has x_3 coordinate equal to zero, then $x = 0$. Thus, the number of distinct lines in $\mathcal{N}(\mathcal{M})$ is equal to the number of distinct points in

$$\mathcal{P} := \left\{ (x_1, x_2) \in \mathbb{R}^2 : \ \begin{array}{l} x_1^2 + x_2^2 - 1 = 0 \\ (ax_1^2 + dx_2^2 + 2cx_1 + f) + x_2(2bx_1 + 2e) = 0 \end{array} \right\}.$$

Suppose that $\mathcal{N}(\mathcal{M})$ contains at least five lines so that $\mathcal{P}$ contains at least five points. Without loss of generality, we may assume that the x_1 coordinates of these five points are distinct (else, perform an orthonormal change of basis on the first two dimensions). Let the x_1 coordinates of these five points be $\xi_1, \xi_2, \ldots, \xi_5$. For each ξ_i, by the first constraint in the definition of $\mathcal{P}$, we have that the corresponding x_2 coordinate must be either $\sqrt{1-\xi_i^2}$ or $-\sqrt{1-\xi_i^2}$. Hence,

$$\begin{aligned}&\left[\left(a\xi^2 + d(1-\xi^2) + 2c\xi + f\right) + \sqrt{1-\xi^2}(2b\xi + 2e)\right]\left[\left(a\xi^2 + d(1-\xi^2) + 2c\xi + f\right) - \sqrt{1-\xi^2}(2b\xi + 2e)\right]\\ &\quad = \left[(a-d)^2 + 4b^2\right]\xi^4 + [4(a-d)c + 8be]\xi^3 + \left[2(a-d)(d+f) + 4c^2 + 4e^2 - 4b^2\right]\xi^2 +\\ &\qquad [4c(d+f) - 8be]\xi + \left[(d+f)^2 - 4e^2\right]\end{aligned}$$

is a degree-4 polynomial in ξ which is zero on five distinct points $\xi_1, \ldots, \xi_5$. We conclude that this polynomial is identically zero. The coefficient of ξ^4 implies that $a = d$ and $b = 0$. The coefficient of ξ^2 implies that $c = e = 0$. The constant term implies that $f = -d$. We conclude that M_2 has the form

$$M_2 = \begin{pmatrix} a & & \\ & a & \\ & & -a \end{pmatrix}.$$

This contradicts the assumption that there does not exist an $(\alpha_1, \alpha_2) \neq (0,0)$ such that $\alpha_1 M_1 + \alpha_2 M_2 \in \mathbb{S}^n_+$. ■

D Appendices for Chapter 4

D.1 Proofs of Theorems 21 and 22

In this appendix, we outline how to modify the proofs of Theorems 19 and 20 to prove Theorems 21 and 22.

Theorem 21. *Suppose there exists $\gamma^* \geq 0$ such that $A(\gamma^*) \succ 0$. Consider the closed nonempty interval $\Gamma := \{\gamma \in \mathbb{R}_+ : A(\gamma) \succeq 0\}$. Let γ_- denote its leftmost endpoint.*

- *If Γ is bounded above, let γ_+ denote its rightmost endpoint. Then,*

$$\operatorname{conv}(\mathcal{S}) = \mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+).$$

In particular, we have $\min_{x\in\mathbb{R}^n}\{q_0(x) : q_1(x) \leq 0\} = \min_{x\in\mathbb{R}^n} \max\{q(\gamma_-, x),\ q(\gamma_+, x)\}$.

- *If Γ is not bounded above, then $q_1(x)$ is convex and*

$$\operatorname{conv}(\mathcal{S}) = \mathcal{S}(\gamma_-) \cap \left\{(x,t) \in \mathbb{R}^{n+1} : q_1(x) \leq 0\right\}.$$

In particular, we have $\min_{x\in\mathbb{R}^n}\{q_0(x) : q_1(x) \leq 0\} = \min_{x\in\mathbb{R}^n}\{q(\gamma_-, x) : q_1(x) \leq 0\}$.

Proof. The "$\subseteq$" inclusions follow from a trivial modification of Lemma 43. It suffices to prove the "$\supseteq$" inclusions. The case where A_0 and A_1 are both nonconvex is covered by Theorem 19. We consider the four remaining cases:

- Suppose A_0 and A_1 are both convex. In this case, $\Gamma = [0, \infty)$ and it suffices to show that $\operatorname{conv}(\mathcal{S}) = \{(x,t) : q_0(x) \leq t,\ q_1(x) \leq 0\} = \mathcal{S}$. This holds as $\mathcal{S}$ is convex.
- Suppose A_0 is nonconvex and A_1 is convex. In this case, $\Gamma = [\gamma_-, \infty)$ is unbounded above. Furthermore, γ_- is positive and $A(\gamma_-)$ has a zero eigenvalue. Suppose $(\hat{x}, \hat{t})$ satisfies $q(\gamma_-, \hat{x}) \leq \hat{t}$ and $q_1(\hat{x}) \leq 0$. If $q_1(\hat{x}) = 0$, then we also have $q_0(\hat{x}) = q(\gamma_-, \hat{x}) \leq \hat{t}$, whence $(\hat{x}, \hat{t}) \in \mathcal{S}$. On the other hand, if $q_1(\hat{x}) < 0$, we may apply the argument in case (iii) in the proof of Lemma 44 verbatim (after replacing all occurrences of γ_+ by γ^*) to conclude that $(\hat{x}, \hat{t}) \in \operatorname{conv}(\mathcal{S})$.
- Suppose A_0 is convex and A_1 is nonconvex. In this case, $\Gamma = [0, \gamma_+]$ is bounded above and γ_- is defined to be $\gamma_- = 0$. Furthermore, $A(\gamma_+)$ has a zero eigenvalue. Suppose

$(\hat{x}, \hat{t}) \in \mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+)$. If $q_1(\hat{x}) \leq 0$, then we also have $q_0(\hat{x}) = q(\gamma_-, \hat{x}) \leq \hat{t}$, whence $(\hat{x}, \hat{t}) \in \mathcal{S}$. On the other hand, if $q_1(\hat{x}) > 0$, we may apply the argument in case (ii) in the proof of Lemma 44 verbatim to conclude that $(\hat{x}, \hat{t}) \in \operatorname{conv}(\mathcal{S})$. ■

We will prove Theorem 22 using a limiting argument and reducing it to Theorem 21. The proof follows that of Lemma 46 almost verbatim.

Theorem 22. *Suppose there exists $\gamma^* \geq 0$ such that $A(\gamma^*) \succeq 0$. Consider the closed nonempty interval $\Gamma := \{\gamma \in \mathbb{R}_+ :\ A(\gamma) \succeq 0\}$. Let γ_- denote its leftmost endpoint.*

- *If Γ is bounded above, let γ_+ denote its rightmost endpoint. Then,*

$$\overline{\operatorname{conv}}(\mathcal{S}) = \mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+).$$

In particular, $\inf_{x\in\mathbb{R}^n} \{q_0(x) :\ q_1(x) \leq 0\} = \inf_{x\in\mathbb{R}^n} \max\{q(\gamma_-, x),\ q(\gamma_+, x)\}$.

- *If Γ is not bounded above, then $q_1(x)$ is convex and*

$$\overline{\operatorname{conv}}(\mathcal{S}) = \mathcal{S}(\gamma_-) \cap \left\{(x, t) \in \mathbb{R}^{n+1} :\ q_1(x) \leq 0\right\}.$$

In particular, $\inf_{x\in\mathbb{R}^n} \{q_0(x) :\ q_1(x) \leq 0\} = \inf_{x\in\mathbb{R}^n} \{q(\gamma_-, x) :\ q_1(x) \leq 0\}$.

Proof. The "$\subseteq$" inclusions follow from a trivial modification of Lemma 45. It suffices to prove the "$\supseteq$" inclusions.

Denote the set on the right hand side by $\mathcal{R}$, i.e., $\mathcal{R} := \mathcal{S}(\gamma_-) \cap \mathcal{S}(\gamma_+)$ when Γ is bounded and $\mathcal{R} := \mathcal{S}(\gamma_-) \cap \{(x, t) :\ q_1(x) \leq 0\}$ when Γ is unbounded.

Let $(\hat{x}, \hat{t}) \in \mathcal{R}$. It suffices to show that $(\hat{x}, \hat{t} + \epsilon) \in \operatorname{conv}(\mathcal{S})$ for all $\epsilon > 0$.

We will perturb A_0 slightly to create a new instance of the problem. Let $\delta > 0$ to be picked later. Define $A_0' = A_0 + \delta I_n$ and let all remaining data be unchanged, i.e.,

$$\begin{aligned} q_0'(x) &:= x^\intercal A_0' x + 2b_0'^\intercal x + c_0' := x^\intercal (A_0 + \delta I_n) x + 2b_0^\intercal x + c_0 \\ q_1'(x) &:= x^\intercal A_1' x + 2b_1'^\intercal x + c_1' := x^\intercal A_1 x + 2b_1^\intercal x + c_1. \end{aligned}$$

We will denote all quantities related to the perturbed system with an apostrophe.

We claim it suffices to show that there exists $\delta > 0$ small enough such that $(\hat{x}, \hat{t} + \epsilon) \in \mathcal{R}'$. Indeed, suppose this is the case. Note that for any $x \in \mathbb{R}^n$, we have $q_1(x) = q_1'(x)$ and $q_0(x) \leq q_0'(x)$. Hence, $\operatorname{conv}(\mathcal{S}') \subseteq \operatorname{conv}(\mathcal{S})$. Then, noting that $A'(\gamma^*) = A(\gamma^*) + \delta I_n \succ 0$, we may apply Theorem 21 to the perturbed system to get $(\hat{x}, \hat{t} + \epsilon) \in \mathcal{R}' = \operatorname{conv}(\mathcal{S}') \subseteq \operatorname{conv}(\mathcal{S})$ as desired.

First note that $A_1 = A_1'$ so that Γ is bounded if and only if Γ' is bounded. We will then pick $\delta > 0$ small enough such that

$$\delta \|\hat{x}\|^2 \leq \frac{\epsilon}{2}, \qquad |\gamma_-' - \gamma_-||q_1(\hat{x})| \leq \frac{\epsilon}{2}, \qquad |\gamma_+' - \gamma_+||q_1(\hat{x})| \leq \frac{\epsilon}{2},$$

where the last condition is only required when γ_+ and γ_+' both exist. This is possible as the expression on the left of each inequality is continuous in δ and is strictly satisfied when $\delta = 0$.

The following computation shows that $q'(\gamma'_-, \hat{x}) \leq \hat{t} + \epsilon$.

$$\begin{aligned} q'(\gamma'_-, \hat{x}) - (\hat{t} + \epsilon) &= q'(\gamma_-, \hat{x}) - (\hat{t} + \epsilon) + (\gamma'_- - \gamma_-)q_1(\hat{x}) \\ &\leq q(\gamma_-, \hat{x}) + \delta\|\hat{x}\|^2 - (\hat{t} + \epsilon) + |\gamma'_- - \gamma_-||q_1(\hat{x})| \\ &\leq q(\gamma_-, \hat{x}) - \hat{t} \\ &\leq 0 \end{aligned}$$

The first inequality follows by noting $q'(\gamma, x) = q(\gamma, x) + \delta\|x\|^2$, the second inequality follows from our assumptions on δ, and the third inequality follows from the assumption that $(\hat{x}, \hat{t}) \in \mathcal{S}(\gamma_-)$. Thus $(\hat{x}, \hat{t} + \epsilon) \in \mathcal{S}'(\gamma'_-)$. When Γ is bounded (or equivalently, when γ'_+ and γ_+ exist), a similar calculation shows that $q'(\gamma'_+, \hat{x}) - (\hat{t} + \epsilon) \leq 0$ so that $(\hat{x}, \hat{t} + \epsilon) \in \mathcal{S}'(\gamma'_+)$. Finally, when Γ is unbounded we have $q'_1(\hat{x}) = q_1(\hat{x}) \leq 0$ so that $(\hat{x}, \hat{t} + \epsilon) \in \{(x, t) : q'_1(x) \leq 0\}$. Thus, $(\hat{x}, \hat{t} + \epsilon)$ is in $\mathcal{R}'$, concluding the proof. ■

D.2 Estimation of the regularity parameters

In Section 4.4 we gave algorithms to solve the GTRS assuming that we had access to (ξ, ζ) and $\hat{\gamma}$ satisfying Assumption 12. In this appendix, we show how to compute these quantities.

Let q_0, q_1 satisfy Assumption 11. Recall the definitions

$$\xi^* := \min\left\{1, \max_{\gamma \geq 0} \lambda_{\min}(A(\gamma))\right\}, \qquad \zeta^* := \max\{1, \gamma_+\}.$$

We will find (ξ, ζ) satisfying

$$\xi^*/4 \leq \xi \leq \xi^*, \qquad \zeta^* \leq \zeta \leq 4\zeta^*$$

and a $\hat{\gamma}$ such that $\lambda_{\min}(A(\hat{\gamma})) \geq \xi$.

We will accomplish this in two stages. We begin by estimating ξ^* using only an upper bound $\bar{\zeta}$ of ζ^*. Then using our estimate ξ we will compute ζ.

Computing ξ and $\hat{\gamma}$

We start with the following guarantee for the algorithm TestXi (Algorithm 14).

Lemma 93. *Given q_0, q_1 satisfying Assumption 11, an arbitrary $0 < \xi \leq 1$, an upper bound $\bar{\zeta} \geq \zeta^*$, and a failure probability $p_\xi > 0$, TestXi (Algorithm 14) will output*

$$\begin{cases} \hat{\gamma} \text{ such that } \lambda_{\min}(A(\hat{\gamma})) \geq \xi/2 & \text{if } \xi \leq \xi^* \\ \hat{\gamma} \text{ such that } \lambda_{\min}(A(\hat{\gamma})) \geq \xi/2 \text{ or “Fail”} & \text{if } \xi^* < \xi \leq 2\xi^* \\ \text{“Fail”} & \text{if } 2\xi^* < \xi \end{cases}$$

Algorithm 14 TestXi$(q_0, q_1, \xi, \bar{\zeta}, p_\xi)$

```
Given $q_0, q_1$ satisfying Assumption 11, a guess $\xi$, an upper bound $\bar{\zeta} \geq \zeta^*$, and a failure probability $p_\xi > 0$
  1. Let $s_0 = 0$ and $t_0 = \bar{\zeta}$
  2. Let $T = \lceil \log \kappa \rceil + 2$ where $\kappa = \bar{\zeta}/\xi$
  3. For $k = 0, \ldots, T-1$
      a) Let $x = \text{ApproxEig}(A(s_k), 2\bar{\zeta}, \xi/4, \frac{p_\xi}{3T})$. If $x^\intercal A(s_k) x \geq 3\xi/4$, then return $\hat{\gamma} = s_k$.
      b) Let $x = \text{ApproxEig}(A(t_k), 2\bar{\zeta}, \xi/4, \frac{p_\xi}{3T})$. If $x^\intercal A(t_k) x \geq 3\xi/4$, then return $\hat{\gamma} = t_k$.
      c) Let $\bar{\gamma} = (s_k + t_k)/2$
      d) Let $x = \text{ApproxEig}(A(\bar{\gamma}), 2\bar{\zeta}, \xi/4, \frac{p_\xi}{3T})$. If $x^\intercal A(\bar{\gamma}) x \geq 3\xi/4$, then return $\hat{\gamma} = \bar{\gamma}$.
      e) If $x^\intercal A_1 x \geq 0$, let $s_{k+1} = \bar{\gamma}$ and $t_{k+1} = t_k$. Else, let $s_{k+1} = s_k$ and $t_{k+1} = \bar{\gamma}$.
  4. Return "Fail"
```

with probability $1 - p_\xi$. *This algorithm runs in time*

$$\tilde{O}\left(N\sqrt{\frac{\bar{\zeta}}{\xi}}\log\left(\frac{n}{p_\xi}\right)\log\left(\frac{\bar{\zeta}}{\xi}\right)\right).$$

Proof. We condition on the event that ApproxEig succeeds every time it is called. By the union bound, this happens with probability at least $1 - p_\xi$.

As we have conditioned on ApproxEig succeeding, any $\hat{\gamma}$ which is output by TestXi will satisfy

$$\lambda_{\min}(A(\hat{\gamma})) \geq 3\xi/4 - \xi/4 = \xi/2.$$

It is clear that TestXi will output "Fail" if $\xi > 2\xi^*$ as there does not exist any $\hat{\gamma}$ such that $\lambda_{\min}(A(\hat{\gamma})) \geq \xi^*$. It remains to show that, given $\xi \leq \xi^*$, TestXi will output some $\hat{\gamma}$.

For the sake of contradiction, assume that the algorithm fails to output in each of the T rounds. Let $P := \{\gamma : \lambda_{\min}(A(\gamma)) \geq 3\xi^*/4\}$. Recall that $\lambda_{\min}(A(\gamma))$ is 1-Lipschitz in γ. As there exists some γ such that $\lambda_{\min}(A(\gamma)) \geq \xi^*$ (see Definition 17), we conclude that P is an interval of length at least $\xi^*/2$.

Note that $P \subseteq [s_0, t_0]$. We will inductively show that $P \subseteq [s_k, t_k]$ for each $k \in \{1, \ldots, T\}$. Let $k \in \{0, \ldots, T-1\}$ and let $s_k, \bar{\gamma}, t_k$ be defined as in the algorithm and let x be the unit vector found in step 3.(d). We claim that $x^\intercal A_1 x \neq 0$. Indeed suppose $x^\intercal A_1 x = 0$, then $x^\intercal A(\gamma) x = x^\intercal A(\bar{\gamma}) x \leq 3\xi/4$ for all γ. This contradicts the assumption that there exists some γ such that $\lambda_{\min}(A(\gamma)) \geq \xi$. Now suppose $\gamma \in P$, then

$$\frac{3\xi^*}{4} \leq x^\intercal A(\gamma) x = x^\intercal A(\bar{\gamma}) x + (\gamma - \bar{\gamma}) x^\intercal A_1 x \leq \frac{3\xi^*}{4} + (\gamma - \bar{\gamma}) x^\intercal A_1 x,$$

where the first inequality follows from $\gamma \in P$, and the last one from the fact that the algorithm did not output in iteration k (and thus the if statement in step 3.(d) did not hold). Thus, if $x^\intercal A_1 x > 0$, then we have the implication $\gamma \in P \implies \gamma \geq \bar{\gamma}$. Similarly, if $x^\intercal A_1 x < 0$, then we have the implication $\gamma \in P \implies \gamma \leq \bar{\gamma}$. Then by induction, we have $P \subseteq [s_{k+1}, t_{k+1}]$.

We conclude that P, an interval of length at least $\xi^*/2$, is contained in $[s_T, t_T]$ an interval of length

$$t_T - s_T = \frac{t_0 - s_0}{2^T} \leq \xi/4.$$

Noting that $\xi \leq \xi^*$ gives us the desired contradiction.

The running time of this algorithm follows from Lemma 52. ■

Given a lower bound $\xi \leq \xi^*$, Lemma 93 guarantees that TestXi will find a $\hat{\gamma}$ satisfying $\lambda_{\min}(A(\hat{\gamma})) \geq \xi/2$ with high probability. In order to make use of this lemma *without* a lower bound on ξ^*, we will simply repeatedly call TestXi with decreasing guesses for ξ. Consider Algorithm 15.

Algorithm 15 ApproxXi$(q_0, q_1, \bar{\zeta}, p)$

```
Given $q_0, q_1$ satisfying Assumption 11, an upper bound $\bar{\zeta} \geq \zeta^*$, and failure probability $p > 0$
  1. For $k = 1, 2, \ldots$
       a) Run TestXi$(q_0, q_1, 2^{-(k-1)}, \bar{\zeta}, 2^{-k}p)$.
       b) If TestXi outputs "Fail" then continue.
       c) Else, let $\hat{\gamma}$ be the output of TestXi and let $\xi = 2^{-k}$; return $\xi$ and $\hat{\gamma}$.
```

Theorem 40. *Given q_0, q_1 satisfying Assumption 11, an upper bound $\bar{\zeta} \geq \zeta^*$, and a failure probability $p > 0$, ApproxXi (Algorithm 15) will output ξ and $\hat{\gamma}$ such that*

$$\xi^*/4 \leq \xi \leq \xi^*, \qquad \lambda_{\min}(A(\hat{\gamma})) \geq \xi$$

and run in time

$$\tilde{O}\left(N\sqrt{\frac{\bar{\zeta}}{\xi^*}}\log\left(\frac{n}{p}\right)\log\left(\bar{\zeta}\right)\log\left(\frac{1}{\xi^*}\right)^3\right)$$

with probability $1 - p$.

Proof. We condition on the event that TestXi succeeds every time it is called. By the union bound, this happens with probability at least $1 - p$.

Let $k^* \in \{1, 2, \ldots\}$ be such that $\xi^*/2 \leq 2^{-k^*} < \xi^*$. Then, as we have conditioned on TestXi succeeding, Lemma 93 guarantees that TestXi$(q_0, q_1, 2^{-k}, \bar{\zeta}, 2^{-(k+1)}p)$ outputs

$$\begin{cases} \hat{\gamma} \text{ such that } \lambda_{\min}(A(\hat{\gamma})) \geq 2^{-k} & \text{if } 2^{-k} \leq \xi^*/2 \\ \hat{\gamma} \text{ such that } \lambda_{\min}(A(\hat{\gamma})) \geq 2^{-k} \text{ or "Fail"} & \text{if } \xi^*/2 < 2^{-k} \leq \xi^* \\ \text{"Fail"} & \text{if } \xi^* < 2^{-k}. \end{cases}$$

Thus, TestXi will output "Fail" for every $k < k^*$ and will output $\hat{\gamma}$ either on round k^* or $k^* + 1$. We can then bound

$$\lambda_{\min}(A(\hat{\gamma})) \geq 2^{-(k^*+1)} \geq \frac{\xi^*}{4}.$$

We bound the run time of the algorithm as follows.

$$\begin{aligned} &\sum_{k=1}^{k^*+1} \tilde{O}\left(N\sqrt{\frac{\bar{\zeta}}{2^{-(k-1)}}} \log\left(\frac{n}{2^{-k}p}\right) \log\left(\frac{\bar{\zeta}}{2^{-(k-1)}}\right)\right) \\ &= \tilde{O}\left(k^{*3} N\sqrt{\frac{\bar{\zeta}}{2^{-k^*}}} \log\left(\frac{n}{p}\right) \log\left(\bar{\zeta}\right)\right) \\ &= \tilde{O}\left(N\sqrt{\frac{\bar{\zeta}}{\xi^*}} \log\left(\frac{n}{p}\right) \log\left(\bar{\zeta}\right) \log\left(\frac{1}{\xi^*}\right)^3\right). \end{aligned}$$

■

Computing ζ

Recall the guarantee of the algorithm ApproxGammaPlus.

Lemma 53. *Given q_0, q_1 satisfying Assumption 11, (ξ, ζ) and $\hat{\gamma}$ satisfying Assumption 12, $\delta > 0$, and $p_{\tilde{\gamma}_+}$, ApproxGammaPlus (Algorithm 2) outputs $\tilde{\gamma}_+$ satisfying*

$$\tilde{\gamma}_+ \in [\gamma_+ - \delta,\ \gamma_+], \qquad \lambda_{\min}(A(\tilde{\gamma}_+)) \leq \delta/\kappa$$

with probability $1 - p_{\tilde{\gamma}_+}$. This algorithm runs in time

$$\tilde{O}\left(\frac{N\sqrt{\kappa\zeta}}{\sqrt{\delta}} \log\left(\frac{n}{p_{\tilde{\gamma}_+}}\right) \log\left(\frac{\kappa}{\delta}\right)\right).$$

We will repeatedly call ApproxGammaPlus with different choices of δ. Consider the algorithm ApproxZeta.

Algorithm 16 ApproxZeta$(q_0, q_1, \xi, \bar{\zeta}, \hat{\gamma}, p)$

```
Given $q_0, q_1$ satisfying Assumption 11, $(\xi, \bar{\zeta})$ and $\hat{\gamma}$ satisfying Assumption 12, and failure probability $p > 0$
    1. For $k = 1, 2, \ldots$
        a) Let $\hat{\zeta}_k$ be the output of ApproxGammaPlus$(q_0, q_1, \xi, 2^{-(k-1)}\bar{\zeta}, \hat{\gamma}, 2^{-(k+1)}\bar{\zeta}, 2^{-k}p)$
        b) If $\hat{\zeta}_k \leq 2^{-(k+1)}\bar{\zeta}$ then continue
        c) Else set $\zeta := 2^{-(k-1)}\bar{\zeta}$; return $\zeta$
```

Theorem 41. *Given q_0, q_1 satisfying Assumption 11, $(\xi, \bar{\zeta})$ and $\hat{\gamma}$ satisfying Assumption 12, and failure probability $p > 0$, ApproxZeta (Algorithm 16) will output ζ such that*

$$\zeta^* \leq \zeta \leq 4\zeta^*$$

and run in time

$$\tilde{O}\left(\frac{N\sqrt{\zeta^*}}{\sqrt{\xi}}\log\left(\frac{n}{p}\right)\log\left(\frac{1}{\xi}\right)\log\left(\frac{\bar{\zeta}}{\zeta^*}\right)^2\right)$$

with probability $1-p$.

Proof. We condition on the event that ApproxGammaPlus succeeds every time it is called. By the union bound, this happens with probability at least $1-p$.

We first check that the assumptions of Lemma 53 hold. For $k=1$, we have $2^{-(k-1)}\bar{\zeta}=\bar{\zeta}\geq\zeta^*$. Then by induction, and conditioning on ApproxGammaPlus succeeding, Lemma 53 guarantees

$$\zeta^*\leq\hat{\zeta}_k+2^{-(k+1)}\bar{\zeta}.$$

If ApproxZeta fails to terminate in round k, then 1.(b) ensures $\hat{\zeta}_k\leq 2^{-(k+1)}\bar{\zeta}$. This in turn implies that $\zeta^*\leq 2^{-((k+1)-1)}\bar{\zeta}$ and, by induction, the assumptions of Lemma 53 hold in every round that ApproxGammaPlus is called.

Let k be the round in which the algorithm terminates. If $k=1$, then the guarantee of Lemma 53 implies $\zeta^*\geq\hat{\zeta}_1$, whence

$$\bar{\zeta}\geq\zeta^*\geq\hat{\zeta}_1>\frac{1}{4}\bar{\zeta}.$$

Thus, we may assume $k\geq 2$. The condition of step 1.(b) then guarantees the two inequalities

$$\hat{\zeta}_{k-1}\leq 2^{-k}\bar{\zeta},\ \text{ and }\ \hat{\zeta}_k>2^{-(k+1)}\bar{\zeta}. \tag{10}$$

Then, we have

$$\zeta^*\geq\hat{\zeta}_k>2^{-(k+1)}\bar{\zeta}=\frac{1}{4}\Big(2^{-k}\bar{\zeta}+2^{-k}\bar{\zeta}\Big)\geq\frac{1}{4}\Big(\hat{\zeta}_{k-1}+2^{-((k-1)+1)}\bar{\zeta}\Big)\geq\zeta^*/4$$

where the first and fifth relations follow from Lemma 53 and the second and fourth relations follow from (10) above.

It remains to bound the run time of ApproxZeta. Let $k^*\in\{1,2,\dots\}$ be such that $\zeta^*\leq 2^{-(k^*-1)}\bar{\zeta}<2\zeta^*$. We show that ApproxZeta terminates within k^* rounds. Suppose ApproxZeta reaches the k^*th round. Then, we have

$$\hat{\zeta}_{k^*}\geq\zeta^*-2^{-(k^*+1)}\bar{\zeta}>2^{-k^*}\bar{\zeta}-2^{-(k^*+1)}\bar{\zeta}=2^{-(k^*+1)}\bar{\zeta},$$

where we used Lemma 53 in the first relation, and the definition of k^* in the second relation. Therefore, ApproxZeta terminates in round k^* at the latest and we can bound the run time of this algorithm as

$$
\begin{aligned}
&\sum_{k=1}^{k^*} \tilde{O}\left(\frac{2^{-(k-1)} N \bar{\zeta}}{\sqrt{2^{-(k+1)} \xi \bar{\zeta}}} \log\left(\frac{n}{2^{-k} p} \right) \log\left(\frac{2^{-(k-1)} \bar{\zeta}}{2^{-(k+1)} \xi \bar{\zeta}} \right) \right) \\
&= \tilde{O}\left(k^{*2} \frac{N \sqrt{2^{-k^*} \bar{\zeta}}}{\sqrt{\xi}} \log\left(\frac{n}{p} \right) \log\left(\frac{1}{\xi} \right) \right) \\
&= \tilde{O}\left(\frac{N \sqrt{\zeta^*}}{\sqrt{\xi}} \log\left(\frac{n}{p} \right) \log\left(\frac{1}{\xi} \right) \log\left(\frac{\bar{\zeta}}{\zeta^*} \right)^2 \right). \qquad \blacksquare
\end{aligned}
$$

E Appendices for Chapter 5

E.1 Useful lemmas regarding quadratic functions

The following two basic bounds will be useful in our error analysis.

Lemma 94. *Let* $q(x) = x^\intercal A x + 2b^\intercal x + c$ *for* $A \in \mathbb{S}^n$, $b \in \mathbb{R}^n$, *and* $c \in \mathbb{R}$. *Then, for all* $x, y \in \mathbb{R}^n$, $|q(x) - q(y)| \leq \|A\| \|y - x\|^2 + 2(\|A\|\|x\| + \|b\|)\|y - x\|$. *In particular, if* $\|A\|, \|b\| \leq 1$, $\|x\| \leq \rho$ *and* $\|x - y\| \leq \delta$ *for some* $\delta \leq 1 \leq \rho$, *then* $|q(x) - q(y)| \leq 5\delta\rho$.

Proof. Writing $y = (y - x) + x$ and expanding the formula for $q(y)$, we obtain

$$
\begin{aligned}
q(y) &= (y - x)^\intercal A (y - x) + 2x^\intercal A (y - x) + x^\intercal A x + 2b^\intercal (y - x) + 2b^\intercal x + c \\
&= q(x) + ((y - x)^\intercal A (y - x) + 2\langle Ax + b, y - x \rangle). \qquad \blacksquare
\end{aligned}
$$

Lemma 95. *Let* $\alpha, \beta, \gamma \in \mathbb{R}$ *where* $\alpha \neq 0$ *and* $\gamma/\alpha \leq 0$. *Then the roots of* $\alpha z^2 + 2\beta z + \gamma = 0$ *satisfy* $|z| \leq 2\left|\frac{\beta}{\alpha}\right| + \sqrt{\frac{-\gamma}{\alpha}}$.

Proof. Let $\{z_-, z_+\}$ denote the roots (possibly with multiplicity). We bound

$$
\begin{aligned}
\{z_-,\, z_+\} &= \left\{ -\tfrac{\beta}{\alpha} \pm \sqrt{\left(\tfrac{\beta}{\alpha}\right)^2 - \tfrac{\gamma}{\alpha}} \right\} \\
&\subseteq \left[-\tfrac{\beta}{\alpha} - \left(\left|\tfrac{\beta}{\alpha}\right| + \sqrt{\tfrac{-\gamma}{\alpha}} \right),\ -\tfrac{\beta}{\alpha} + \left(\left|\tfrac{\beta}{\alpha}\right| + \sqrt{\tfrac{-\gamma}{\alpha}} \right) \right] \\
&\subseteq \left[-\left(2\left|\tfrac{\beta}{\alpha}\right| + \sqrt{\tfrac{-\gamma}{\alpha}} \right),\ \left(2\left|\tfrac{\beta}{\alpha}\right| + \sqrt{\tfrac{-\gamma}{\alpha}} \right) \right]. \qquad \blacksquare
\end{aligned}
$$

E.2 Useful procedures

This appendix contains running time guarantees for well-known algorithms that we will utilize as building blocks in Algorithm 5.

The Lanczos method

The following lemma characterizes the running time for approximating the minimum eigenvalue of a symmetric matrix.

Lemma 96 ([103]). *There exists an algorithm,* ApproxEig(A, ρ, δ, p), *which given a symmetric matrix $A \in \mathbb{S}^n$, ρ such that $\|A\|_2 \leq \rho$, and parameters $\delta, p > 0$, will, with probability at least $1 - p$, return a unit vector $x \in \mathbb{R}^n$ such that $x^\intercal Ax \leq \lambda_{\min}(A) + \delta$. This algorithm runs in time*

$$O\left(\frac{N\sqrt{\rho}}{\sqrt{\delta}}\log\left(\frac{n}{p}\right)\right),$$

where N is the number of nonzero entries in A.

ApproxGamma

The following algorithm extends [180, Algorithm 2] to find a $\gamma \leq \hat{\gamma}$ such that $\mu(\gamma)$ falls in a prescribed range. An analogous algorithm can be used to find a $\gamma \geq \hat{\gamma}$ such that $\mu(\gamma)$ falls in a prescribed range.

Algorithm 17 `ApproxGammaLeft`

```
Given $(A_0, A_1)$, $(\xi, \zeta, \hat{\gamma})$, $p \in (0,1)$, and $\mu \in (0, \xi)$
    1. Set $\ell_1 = 0, r_1 = \hat{\gamma}$
    2. For $t = 1, \dots, T = \left\lceil \log\left(\frac{5\zeta}{\mu}\right)\right\rceil$
        a) $\gamma_t = (\ell_t + r_t)/2$
        b) Let $x_t = ApproxEig(A(\gamma_t), 2\zeta, \mu/8, p/T)$ and $\hat{\mu}_t = x_t^\intercal A(\gamma_t)x_t$
        c) If $\hat{\mu}_t > \mu$, set $\ell_{t+1} = \ell_t,\ r_{t+1} = \gamma_t$
        d) Else if $\hat{\mu}_t < \frac{5}{8}\mu$, set $\ell_{t+1} = \gamma_t,\ r_{t+1} = r_t$
        e) Else, output $\gamma_t, x_t$
```

Lemma 61. *Suppose Assumption 14 holds, $\mu \in (0, \xi)$ and $p \in (0,1)$. Then, with probability at least $1 - p$,* `ApproxGammaLeft`(μ, p) *(Algorithm 17) returns (γ, v) such that $\gamma \leq \hat{\gamma}$ and v is a unit vector satisfying $\mu/2 \leq \mu(\gamma) \leq v^\intercal A(\gamma)v \leq \mu$ in time*

$$\tilde{O}\left(\frac{N\sqrt{\zeta}}{\sqrt{\mu}}\log\left(\frac{n}{p}\right)\log\left(\frac{\zeta}{\mu}\right)\right).$$

Proof. We condition on ApproxEig succeeding in each call. This happens with probability at least $1 - p$.

Suppose `ApproxGammaLeft` outputs on iteration t. On this iteration, we have $\mu(\gamma_t) \geq \hat{\mu}_t - \mu/8 \geq \mu/2$. Similarly note $x^\intercal A(\gamma_t)x = \hat{\mu}_t \leq \mu$.

Next, we show that `ApproxGammaLeft` is guaranteed to output within T iterations. Suppose otherwise and consider the interval

$$\mathcal{I} := \left\{\gamma \in \mathbb{R}_+ : \begin{array}{c} \gamma \leq \hat{\gamma} \\ \mu(\gamma) \in \left[\frac{5}{8}\mu, \frac{7}{8}\mu\right] \end{array}\right\}.$$

Note that if $\gamma_t \in \mathcal{I}$ for some t then `ApproxGammaLeft` will output at step t. Indeed, at iteration t, we will have $\hat{\mu}_t \in \left[\mu(\gamma_t), \mu(\gamma_t) + \frac{\mu}{8}\right] \subseteq \left[\frac{5}{8}\mu, \mu\right]$. In particular, we deduce that $\gamma_t \notin \mathcal{I}$ for any $t = 1, \ldots, T$. Next, by construction, the interval $[\ell_t, r_t]$ contains $\mathcal{I}$ for every t. On the other hand, $|[\ell_T, r_T]| \leq 2^{-T}\zeta < \frac{\mu}{4} \leq |\mathcal{I}|$, a contradiction.

It remains to bound the running time of `ApproxGammaLeft`. By Lemma 96, each iteration of step 2.(b) runs in time

$$\tilde{O}\left(\frac{N\sqrt{\zeta}}{\sqrt{\mu}} \log\left(\frac{n}{p}\right)\right).$$

Finally, note that the number of iterations of step 2 is bounded by $T = O\left(\log\left(\frac{\zeta}{\mu}\right)\right)$. ■

Conjugate gradient

The following lemma characterizes the running time for approximately minimizing a strongly convex quadratic function using the conjugate gradient algorithm.

Lemma 97. *There exists an algorithm,* ConjGrad$(A, b, \rho, \mu, \delta)$*, which given symmetric matrix* $A \in \mathbb{S}^n$ *with* $\mu I \preceq A \preceq \rho I$ *and* $b \in \mathbb{R}^n$*, returns* $x \in \mathbb{R}^n$ *such that* $\|x + A^{-1}b\| \leq \delta$*. This algorithm runs in time*

$$O\left(\frac{N\sqrt{\rho}}{\sqrt{\mu}} \log\left(\frac{\|b\|}{\mu\delta}\right)\right).$$

ApproxNu

The following algorithm uses the conjugate gradient algorithm to approximate $\nu(\gamma)$ for a given value of γ.

Algorithm 18 `ApproxNu`

```
Given $(A_0, A_1, b_0, b_1, c_0, c_1)$, $(\xi, \zeta, \hat{\gamma})$ satisfying Assumption 14, $\gamma, \mu$ such that $\mu \in (0,1)$ and
$A(\gamma) \succeq \mu I$, and $\delta > 0$
    • Apply the conjugate gradient method to find $\tilde{x}$ such that $\|\tilde{x} - x(\gamma)\| \leq \frac{\mu\delta}{10\zeta}$
    • Return $\tilde{x}, q_1(\tilde{x})$
```

Lemma 62. *Suppose Assumption 14 holds,* $\mu \in (0, \xi]$*,* $\delta \in (0, 1)$*, and* $A(\gamma) \succeq \mu I$*. Then* `ApproxNu`(μ, δ, γ) *(Algorithm 18) returns* $(\tilde{x}, \tilde{\nu})$ *such that* $\|\tilde{x} - x(\gamma)\| \leq \mu\delta/10\zeta$*, and* $\tilde{\nu} = q_1(\tilde{x}) \in [\nu(\gamma) \pm \delta]$ *in time*

$$O\left(\frac{N\sqrt{\zeta}}{\sqrt{\mu}} \log\left(\frac{\zeta}{\mu\delta}\right)\right).$$

Proof. The running time follows from Lemma 97. Note that Assumption 14 and $A(\gamma) \succeq \mu I$ together imply $\|x(\gamma)\| \leq \frac{2\zeta}{\mu}$. Then, from the definition of $\nu(\gamma)$ and $x(\gamma)$ and applying Lemma 94, we arrive at

$$|q_1(\hat{x}) - \nu(\gamma)| \leq 5\left(\frac{2\zeta}{\mu}\right)\left(\frac{\mu\delta}{10\zeta}\right) \leq \delta.$$ ■

Nesterov's accelerated minimax scheme

The following lemma characterizes the running time for finding an approximate optimizer of the maximum of two strongly convex smooth quadratic functions.

Lemma 98. *There exists an algorithm,* AccMinimax, *which given* $A^{(1)}, A^{(2)} \in \mathbb{S}^n$*,* $b^{(1)}, b^{(2)} \in \mathbb{R}^n$*,* $c^{(1)}, c^{(2)} \in \mathbb{R}$*, and* $(\mu, \rho, \delta) > 0$ *satisfying* $\mu I \preceq A^{(i)} \preceq \rho I$ *and* $\left\|b^{(i)}\right\| \leq \rho$*, will return* $\bar{x}$ *such that*

$$\max_i \; \bar{x}^\intercal A^{(i)} \bar{x} + 2b^{(i)\intercal}\bar{x} + c_i \leq \left(\min_{x \in \mathbb{R}^n} \max_i \; x^\intercal A^{(i)} x + 2b^{(i)\intercal} x + c_i \right) + \delta,$$

in time

$$O\left(\tfrac{N\sqrt{\rho}}{\sqrt{\mu}} \log\left(\tfrac{\rho}{\delta\mu} \right) \right).$$

Proof. For notational convenience, define $q^{(i)}(x) := x^\intercal A^{(i)} x + 2b^{(i)\intercal} x + c^{(i)}$ and $f(x) := \max_i q^{(i)}(x)$. We may take $x_0 = 0$ in [132, Algorithm 2.3.12] and bound

$$\begin{aligned} f(0) - \min_x f(x) &\leq f(0) - \max_i \min_x q^{(i)}(x) \\ &\leq \max_i \left(q^{(i)}(0) - \min_x q^{(i)}(x) \right) \\ &= \max_i \; b^{(i)\intercal} \left(A^{(i)} \right)^{-1} b^{(i)} \\ &\leq \tfrac{\rho^2}{\mu}. \end{aligned}$$

The running time then follows from [132, Theorem 2.3.5] and [180, Lemma 14]. ■

E.3 Deferred proofs from Section 5.2

Lemma 99. *Suppose Assumption 13 holds. Then*

$$\mathrm{Opt} = \inf_{x \in \mathbb{R}^n} \sup_{\gamma \in \Gamma} q(\gamma, x).$$

Proof. ($\geq$) Let $x \in \mathbb{R}^n$ such that $q_1(x) \leq 0$. Then, as $\Gamma \subseteq \mathbb{R}_+$, we have $q_0(x) \geq \sup_{\gamma \in \Gamma} q(\gamma, x)$. Taking the infimum in x concludes this direction.

($\leq$) Let $x \in \mathbb{R}^n$. We split into three cases depending on the sign of $q_1(x)$.

If $q_1(x) = 0$, then $\mathrm{Opt} \leq q_0(x) = \sup_{\gamma \in \Gamma} q(\gamma, x)$.

Next, suppose $q_1(x) < 0$ so that $\sup_{\gamma \in \Gamma} q(\gamma, x) = q(\gamma_-, x)$. If $\gamma_- = 0$, then again $\mathrm{Opt} \leq q_0(x) = \sup_{\gamma \in \Gamma} q(\gamma, x)$. On the other hand, if $\gamma_- > 0$, then $A(\gamma_-)$ is positive semidefinite but not positive definite and there exists nonzero $v \in \ker(A(\gamma_-))$. Without loss of generality, $\langle v, b(\gamma_-) \rangle \leq 0$. Let $\alpha > 0$ such that $q_1(x + \alpha v) = 0$ (this exists as $v^\intercal A_1 v = v^\intercal \frac{A(\bar{\gamma}) - A(\gamma_-)}{\bar{\gamma} - \gamma_-} v > 0$). We deduce $\mathrm{Opt} \leq q_0(x + \alpha v) = q(\gamma_-, x + \alpha v) \leq q(\gamma_-, x) = \sup_{\gamma \in \Gamma} q(\gamma, x)$.

Finally, suppose $q_1(x) > 0$. If Γ is unbounded, then $\sup_{\gamma \in \Gamma} q(\gamma, x) = +\infty$ and $\mathrm{Opt} \leq \sup_{\gamma \in \Gamma} q(\gamma, x)$. Else, we have that $A(\gamma_+)$ is positive semidefinite but not positive definite and there exists nonzero $v \in \ker(A(\gamma_+))$. An argument identical to the one in the previous paragraph shows $\mathrm{Opt} \leq \sup_{\gamma\in\Gamma} q(\gamma, x)$.

Taking the infimum over all $x \in \mathbb{R}^n$ completes the proof. ■

E.4 Deferred proofs from Section 5.4.1

In this appendix, we motivate a generalized-eigenvalue-based replacement for `ApproxGammaLeft` (Algorithm 17) of `CRLeft` (Algorithm 6). Given $\mu \in (0, \xi)$, our goal is to compute $\gamma \leq \hat{\gamma}$ and v such that $\mu/2 \leq \mu(\gamma) \leq v^\intercal A(\gamma) v \leq \mu$. We will do so by approximating the minimum eigenvalue $\tilde{\lambda}$ (and a corresponding eigenvector) for

$$-A_1 v = \lambda\Big(A(\hat{\gamma}) - \tfrac{3\mu}{4} I\Big) v \tag{11}$$

and setting $\tilde{\gamma} := \hat{\gamma} + \frac{1}{\tilde{\lambda}}$. Note that defining $\gamma := \hat{\gamma} + \frac{1}{\lambda}$, where λ is the true minimum eigenvalue to (11), gives

$$\mu(\gamma) = \lambda_{\min}\Big(A(\hat{\gamma}) - \tfrac{3\mu}{4} I + \tfrac{1}{\lambda} A_1\Big) + 3\mu/4 = 3\mu/4.$$

In the following, we abbreviate $\hat{A} := A(\hat{\gamma}) - \frac{3\mu}{4} I$. As in Lemma 61, we will assume Assumption 14 throughout this appendix. We will take $\tilde{\lambda}, \tilde{v}$ to be the output of `eigifp` on the input $(-A_1, \hat{A}, \delta)$ where $\delta > 0$ will be fixed later.

Recall [75] that $\tilde{\lambda}, \tilde{v}$ satisfies

$$(-A_1 + B)\tilde{v} = \tilde{\lambda}(\hat{A} + C)\tilde{v} \tag{12}$$

for some $\|B\| \leq \delta\|A_1\|$ and $\|C\| \leq \delta\Big\|\hat{A}\Big\|$. We will assume that $\tilde{\lambda}$ is in fact the *minimum eigenvalue* of (12).

Lemma 100. *Suppose* $\Big|\lambda - \tilde{\lambda}\Big| \leq \mu/5\zeta^2$*, then* $\mu(\tilde{\gamma}) \geq \mu/2$*.*

Proof. As $\mu(\tilde{\gamma})$ is 1-Lipschitz, it suffices to show that $|\tilde{\gamma} - \gamma| \leq \mu/4$. Note that $\frac{1}{\lambda} = \gamma - \hat{\gamma}$ so that $|\lambda| \geq 1/\zeta$. We deduce that $\Big|\tilde{\lambda}\Big| \geq |\lambda| - \Big|\lambda - \tilde{\lambda}\Big|$. Combining,

$$|\tilde{\gamma} - \gamma| = \frac{\Big|\lambda - \tilde{\lambda}\Big|}{|\lambda|\Big|\tilde{\lambda}\Big|} \leq \frac{\frac{\mu}{5\zeta^2}}{\Big(\frac{1}{\zeta}\Big)\Big(\frac{1}{\zeta} - \frac{\mu}{5\zeta^2}\Big)} \leq \mu/4.$$

■

Lemma 101. *Suppose* $\tilde{\lambda}$ *is a minimum eigenvalue of* (12) *and* $2\delta\zeta \leq \xi/8$*. Then,*

$$\Big|\lambda - \tilde{\lambda}\Big| \leq \delta\frac{72\zeta}{\xi^2}.$$

Proof. Note that

$$\lambda = \max\Big\{\lambda :\ -A_1 - \lambda\hat{A} \succeq 0\Big\}, \quad \text{and} \quad \tilde{\lambda} = \max\Big\{\tilde{\lambda} :\ (-A_1 + B) - \tilde{\lambda}(\hat{A} + C) \succeq 0\Big\}.$$

We compute

$$\begin{aligned} -A_1 - (\tilde{\lambda} - \alpha)\hat{A} &= (-A_1 + B) - \tilde{\lambda}(\hat{A} + C) - B + \tilde{\lambda}C + \alpha\hat{A} \\ &\succeq -\delta(1 + 2\zeta|\tilde{\lambda}|) + \alpha\hat{A}. \end{aligned}$$

We may thus deduce that $-A_1 - (\tilde{\lambda} - \alpha)\hat{A} \succeq 0$ whenever $\alpha \geq \delta\frac{4(1+2\zeta|\tilde{\lambda}|)}{\xi}$. Hence,

$$\tilde{\lambda} - \lambda \leq \delta\frac{4(1 + 2\zeta|\tilde{\lambda}|)}{\xi}.$$

Similarly,

$$\begin{aligned} (-A_1 + B) - (\lambda - \alpha)(\hat{A} + C) &= -A_1 - \lambda\hat{A} + B - \lambda C + \alpha(\hat{A} + C) \\ &\succeq -\delta(1 + 2\zeta|\lambda|) + \alpha(\hat{A} + C). \end{aligned}$$

We may thus deduce that $(-A_1 + B) - (\lambda - \alpha)(\hat{A} + C) \succeq 0$ whenever $\alpha \geq \delta\frac{8(1+2\zeta|\lambda|)}{\xi}$. Hence,

$$\tilde{\lambda} - \lambda \geq -\delta\frac{8(1 + 2\zeta|\lambda|)}{\xi}.$$

Finally, we may estimate $\left|\frac{1}{\lambda}\right| \geq \frac{\xi}{4}$ and $\left|\frac{1}{\tilde{\lambda}}\right| \geq \frac{2\xi}{17}$. We conclude

$$-\delta\frac{8(1 + 8\zeta/\xi)}{\xi} \leq \tilde{\lambda} - \lambda \leq \delta\frac{4(1 + 17\zeta/\xi)}{\xi}. \qquad \blacksquare$$

Proposition 24. *Let* $\delta = \frac{\mu\xi^2}{360\zeta^3}$ *and suppose* $\tilde{\lambda}$ *is the minimum eigenvalue of* (12). *Then,*

$$\mu/2 \leq \mu(\tilde{\gamma}) \leq \tilde{v}^\intercal A(\tilde{\gamma})\tilde{v} \leq \mu.$$

Proof. The first inequality follows from Lemmas 100 and 101. The second inequality follows from the definition of μ. The third inequality follows as

$$\begin{aligned} \tilde{v}^\intercal A(\tilde{\gamma})\tilde{v} &= \tilde{v}^\intercal\left(\hat{A} + \tfrac{1}{\tilde{\lambda}}A_1\right)\tilde{v} + 3\mu/4 \\ &= \tilde{v}^\intercal\left((\hat{A} + C) + \tfrac{1}{\tilde{\lambda}}(A_1 - B) - C + \tfrac{1}{\tilde{\lambda}}B\right)\tilde{v} + 3\mu/4 \\ &\leq \|C\| + \tfrac{1}{|\tilde{\lambda}|}\|B\| + 3\mu/4 \\ &\leq 4\delta\zeta + 3\mu/4. \end{aligned}$$

Here, the first inequality holds as $(-A_1 + B)\tilde{v} = \tilde{\lambda}(\hat{A} + C)\tilde{v}$. The second inequality follows as $\|C\| \leq 2\delta\zeta$ and $\left|\tilde{\lambda}\right| \geq |\lambda| - \left|\lambda - \tilde{\lambda}\right| \geq 1/2\zeta$. $\blacksquare$

		$\bar{N} = 10^4$					$\bar{N} = 10^5$				
					Time					Time	
$\bar{\mu}^*$	Alg.	Error	ErrorCR	Time	Ref.	Solve	Error	ErrorCR	Time	Ref.	Solve
1e-2	WLK21	4.8	6.1	**0.1**	0.05	0.05	5.1	5.4	**0.8**	0.3	0.4
	WK20	5.7	6.7	0.5	0.1	0.3	4.8	5.3	4.4	0.5	3.8
	JL19	1.5e+03	1.8e+06	0.7	0.1	0.6	5.1e+01	2.1e+06	8.2	0.6	7.6
	AN19	6.7e+02	-	1.5	-	-	6.4e+02	-	2.2	-	-
	BTH14	4.2e+08	-	1.1	-	-	7.5e+08	-	1.5	-	-
1e-4	WLK21	6.7	7.2	**0.4**	0.2	0.2	8.5	8.4	2.9	1.0	1.8
	WK20	8.1	7.1	0.7	0.1	0.6	7.0	7.1	7.0	0.6	6.4
	JL19	2.3e+09	4.1e+12	3.0	0.1	2.8	1.0e+09	3.6e+12	49.9	0.6	49.3
	AN19	4.9	-	1.6	-	-	5.0	-	2.4	-	-
	BTH14	4.0e+08	-	1.2	-	-	4.4e+08	-	**1.7**	-	-
1e-6	WLK21	6.5	6.1	**0.8**	0.3	0.5	8.3	8.2	6.3	1.8	4.4
	WK20	6.4	6.4	1.6	0.1	1.5	7.6	8.2	15.5	0.5	15.0
	JL19	7.9e+04	7.5e+10	3.1	0.1	3.0	8.4e+04	7.1e+10	40.4	0.5	39.9
	AN19	1.4e+06	-	1.7	-	-	1.3e+06	-	2.4	-	-
	BTH14	1.3e+09	-	1.4	-	-	1.0e+09	-	**1.7**	-	-

Table E.1: Average errors and solution times for $n = 10^3$ over 100 random instances for each parameter combination. Note that errors are reported in units of 10^{-16}. We call attention to the setting $(\bar{N}, \bar{\mu}^*) = (10^5, 10^{-6})$. In this setting, the fastest algorithm is BTH14. On the other hand, BTH14 also reports the highest error of $\approx 10^{-7}$. BTH14 is followed by AN19 which achieves slightly smaller error of $\approx 10^{-10}$. While WLK21 is slightly slower than both of these algorithms it achieves significantly smaller errors of $\approx 10^{-16}$. The results are similar for $(\bar{N}, \bar{\mu}^*) = (10^5, 10^{-4})$ as well.

E.5 Numerical Experiment Tables

We provide additional statistics for the numerical results plotted in Figures 5.2 to 5.4 for $n = 10^3$, 10^4, 10^5, respectively. In Tables E.1 and E.2, we present the averages for $n = 10^3$, 10^4 respectively over 100 random instances each, and in Table E.3 the averages for $n = 10^5$ are given over 5 random instances. In these tables, Error and ErrorCR correspond to the error of $q_0(\tilde{x})$ and the error of $\bar{x}$ within the convex reformulation respectively as defined in Section 5.4.3. For WLK21, WK20 and JL19, we also report time for constructing the convex reformulation and solving the reformulation as Ref. and Solve. For each parameter combination, we highlight the algorithm with the smallest running time.

F Appendices for Chapter 6

F.1 Deferred proofs

The following proof is adapted from [132].

Proof of Lemma 67. It is evident that $\phi_t(X)$ are quadratic matrix functions of the form (6.6) with $V_0 = X_0$ and $\phi_0^* = Q(X_0)$. The remainder of the proof verifies the recurrences on V_{t+1} and

		$\bar{N} = 10^4$					$\bar{N} = 10^5$				
					Time					Time	
$\bar{\mu}^*$	Alg.	Error	ErrorCR	Time	Ref.	Solve	Error	ErrorCR	Time	Ref.	Solve
1e-2	WLK21	4.9	6.4	**1.8**	0.8	0.9	4.7	5.4	**11.1**	4.8	4.8
	WK20	4.9	5.7	9.8	1.6	8.1	5.3	6.0	67.5	10.5	56.8
	JL19	1.4e+02	1.7e+06	15.3	1.6	13.6	6.3e+02	1.8e+06	93.8	10.7	82.8
	AN19	6.8e+02	-	184.1	-	-	1.2e+03	-	324.5	-	-
1e-4	WLK21	1.5e+01	1.6e+01	**6.6**	2.6	3.7	4.1e+01	4.2e+01	**57.0**	24.0	30.3
	WK20	1.0e+01	1.1e+01	16.6	1.5	15.1	2.9e+01	3.0e+01	207.0	11.0	195.8
	JL19	6.7e+09	4.2e+12	57.9	1.5	56.4	2.1e+10	3.1e+12	393.1	11.3	381.6
	AN19	4.3	-	205.7	-	-	4.5	-	476.4	-	-
1e-6	WLK21	9.1e+01	9.2e+01	**15.1**	5.1	9.8	2.7e+01	2.8e+01	**130.7**	49.1	79.0
	WK20	6.1e+01	6.1e+01	33.0	1.5	31.5	3.1e+01	3.1e+01	264.0	10.6	253.2
	JL19	2.5e+09	7.8e+10	59.7	1.5	58.1	1.6e+08	7.1e+10	402.7	11.0	391.4
	AN19	8.0e+06	-	206.6	-	-	4.4e+06	-	475.5	-	-

Table E.2: Average errors and solution times for $n = 10^4$ over 100 random instances for each parameter combination. Note that errors are reported in units of 10^{-16}.

		$\bar{N} = 10^4$					$\bar{N} = 10^5$				
					Time					Time	
$\bar{\mu}^*$	Alg.	Error	ErrorCR	Time	Ref.	Solve	Error	ErrorCR	Time	Ref.	Solve
1e-2	WLK21	3.3	9.9	**30.1**	12.6	13.6	5.3	2.7	**229.2**	100.8	101.7
	WK20	4.7	7.8	162.9	24.7	137.0	3.1	4.9	1748.4	527.9	1216.3
	JL19	4.9	1.4e+06	287.3	27.4	259.1	1.6e+02	2.3e+06	1930.7	419.0	1507.5
1e-4	WLK21	1.5e+01	1.6e+01	**141.6**	65.1	70.8	9.5e+01	9.5e+01	**1586.3**	767.0	728.5
	WK20	1.6e+01	1.6e+01	334.3	25.7	307.9	1.4e+02	1.4e+02	10622.9	437.7	10180.8
	JL19	2.5e+09	4.3e+12	1044.3	26.8	1016.5	9.2e+10	8.7e+11	11526.9	514.5	11007.9
1e-6	WLK21	2.2e+01	2.0e+01	**294.2**	97.8	190.0	6.2e+01	6.4e+01	**3361.1**	1569.5	1701.7
	WK20	1.5e+01	1.6e+01	612.3	25.7	585.6	1.4e+02	1.4e+02	7781.5	367.8	7409.8
	JL19	7.6e+04	8.5e+10	1081.4	19.5	1061.2	2.1e+06	7.5e+10	10960.0	355.3	10600.8

Table E.3: Average errors and solution times for $n = 10^5$ over 5 random instances for each parameter combination. Note that errors are reported in units of 10^{-16}.

ϕ^*_{t+1}. We suppose that the stated form holds for some t, and we will show that it will hold for $t+1$ as well. We compute

$$\frac{1}{\tilde{\mu}}\nabla\phi_{t+1}(X) = (1-\alpha)(X-V_t) + \alpha\Big(X - \Big(\Xi_t - \frac{1}{\tilde{\mu}}\tilde{g}_t\Big)\Big).$$

We deduce that $V_{t+1} = (1-\alpha)V_t + \alpha\Big(\Xi_t - \frac{1}{\tilde{\mu}}\tilde{g}_t\Big)$. Noting that $\phi^*_{t+1} = \phi_{t+1}(V_{t+1})$, and applying the recursive definition of $\phi_{t+1}(X)$ gives us

$$\begin{aligned}
\phi^*_{t+1} &= (1-\alpha)\Big(\phi^*_t + \frac{\tilde{\mu}}{2}\|V_{t+1}-V_t\|_F^2\Big) \\
&\quad + \alpha\Big(Q(X_{t+1}) + \frac{1}{2\tilde{L}}\|\tilde{g}_t\|_F^2 + \langle \tilde{g}_t, V_{t+1}-\Xi_t\rangle + \frac{\tilde{\mu}}{2}\|V_{t+1}-\Xi_t\|_F^2\Big) \\
&= (1-\alpha)\phi^*_t + \alpha\Big(Q(X_{t+1}) + \frac{1}{2\tilde{L}}\|\tilde{g}_t\|_F^2\Big) \\
&\quad + (1-\alpha)\frac{\tilde{\mu}}{2}\|V_{t+1}-V_t\|_F^2 + \frac{\alpha\tilde{\mu}}{2}\Big\|V_{t+1} - (\Xi_t - \tfrac{1}{\tilde{\mu}}\tilde{g}_t)\Big\|_F^2 - \frac{\alpha}{2\tilde{\mu}}\|\tilde{g}_t\|_F^2 \\
&= (1-\alpha)\phi^*_t + \alpha\Big(Q(X_{t+1}) + \frac{1}{2\tilde{L}}\|\tilde{g}_t\|_F^2\Big) \\
&\quad + \frac{\tilde{\mu}(1-\alpha)\alpha^2}{2}\Big\|V_t - (\Xi_t - \tfrac{1}{\tilde{\mu}}\tilde{g}_t)\Big\|_F^2 + \frac{\tilde{\mu}\alpha(1-\alpha)^2}{2}\Big\|V_t - (\Xi_t - \tfrac{1}{\tilde{\mu}}\tilde{g}_t)\Big\|_F^2 - \frac{\alpha}{2\tilde{\mu}}\|\tilde{g}_t\|_F^2 \\
&= (1-\alpha)\phi^*_t + \alpha\Big(Q(X_{t+1}) + \frac{1}{2\tilde{L}}\|\tilde{g}_t\|_F^2\Big) \\
&\quad + \alpha(1-\alpha)\Big(\frac{\tilde{\mu}}{2}\|\Xi_t - V_t\|_F^2 + \langle \tilde{g}_t, V_t - \Xi_t\rangle\Big) - \frac{\alpha^2}{2\tilde{\mu}}\|\tilde{g}_t\|_F^2,
\end{aligned}$$

where the third equation follows from substituting the expression for V_{t+1}, and the last one from regrouping the terms. ■

The following proof is adapted from [132, Page 92].

Proof of Lemma 68. Note that

$$\begin{aligned}
\Xi_t &= \frac{X_t + \alpha V_t}{1+\alpha} \\
X_{t+1} &= \Xi_t - \frac{\tilde{g}_t}{\tilde{L}} \\
V_{t+1} &= (1-\alpha)V_t + \alpha\Big(\Xi_t - \frac{1}{\tilde{\mu}}\tilde{g}_t\Big).
\end{aligned}$$

Therefore,

$$\begin{aligned} V_{t+1} &= (1-\alpha)\frac{(1+\alpha)\Xi_t - X_t}{\alpha} + \alpha\left(\Xi_t - \frac{1}{\tilde{\mu}}\tilde{g}_t\right) \\ &= X_t + \frac{1}{\alpha}\left(\Xi_t - X_t - \frac{1}{\tilde{L}}\tilde{g}_t\right) \\ &= X_t + \frac{1}{\alpha}(X_{t+1} - X_t). \end{aligned}$$

Then,

$$\begin{aligned} \Xi_{t+1} &= X_{t+1} + \frac{\alpha}{1+\alpha}(V_{t+1} - X_{t+1}) \\ &= X_{t+1} + \frac{1-\alpha}{1+\alpha}(X_{t+1} - X_t). \end{aligned} \qquad \blacksquare$$

Lemma 102. *Consider an instance of* (6.12) *generated by the random procedure in Section 6.5.2. Then equality holds throughout* (6.12).

Proof. It suffices to show that γ^* and T^* are feasible and achieve value $\|X^*\|_F^2$ in the dual SDP (i.e., the third line of (6.12)).

Note that by Schur Complement Theorem,

$$\begin{pmatrix} A(\gamma^*)/2 & B(\gamma^*)/2 \\ B(\gamma^*)^\intercal/2 & \frac{c(\gamma^*)}{k}I_k - T^* \end{pmatrix} \sim \begin{pmatrix} I_{n-k} & \\ & \frac{c(\gamma^*)}{k}I_k - T^* - \frac{B(\gamma^*)^\intercal A(\gamma^*)^{-1}B(\gamma^*)}{2} \end{pmatrix} = \begin{pmatrix} I_{n-k} & \\ & 0_k \end{pmatrix}.$$

Here, $\sim$ indicates matrix similarity. Thus γ^* and T^* are feasible in the dual SDP.

Next,

$$\begin{aligned} \operatorname{tr}(T^*) &= \operatorname{tr}\left(\frac{c(\gamma^*)}{k}I_k - \frac{B(\gamma^*)^\intercal A(\gamma^*)^{-1}B(\gamma^*)}{2}\right) \\ &= \frac{\operatorname{tr}((X^*)^\intercal A(\gamma^*)X^*)}{2} + \langle B(\gamma^*), X^*\rangle + c(\gamma^*) \\ &= \frac{\|X^*\|_F^2}{2} + \sum_{i=1}^m \gamma_i^*\left(\operatorname{tr}\left(\frac{(X^*)^\intercal A_i X^*}{2}\right) + \langle B_i, X^*\rangle + c_i\right) = \frac{\|X^*\|_F^2}{2}. \end{aligned} \qquad \blacksquare$$

F.2 Additional numerical results

Table F.4 displays numerical results for a variant of SketchyCGAL (see Section 6.5.1 for implementation details) on one random instance of (6.12) for each of $n-k = 10^3$, 10^4, and 10^5.

G Appendices for Chapter 7

G.1 Proof of Propositions 21 and 22

Proposition 21. *Let $\mathcal{A} \subseteq \mathbb{S}^n$ and suppose $S \in \operatorname{span}(\mathcal{A})$ is nonsingular. Then, $\mathcal{A}$ is SDC if and only if $S^{-1}\mathcal{A}$ is a commuting set of diagonalizable matrices with real eigenvalues.*

$n-k$	time (s)	$\|X - X^*\|_F^2$	memory (MB)
10^3	3.0×10^3	2.1	2.2×10^1
10^4	1.0×10^4	2.5	6.0×10^1
10^5	5.6×10^4	3.0	2.9×10^2

Table F.4: Preliminary experimental results for $(n - k) = 10^3,\ 10^4,\ 10^5$ (1 instance) with time limit $3 \times 10^3,\ 10^4,\ 5 \times 10^4$ seconds for a variant of SketchyCGAL.

Proof. ($\Rightarrow$) Let $P \in \mathbb{R}^{n\times n}$ furnished by SDC. For $A \in \mathcal{A}$, note that

$$P^{-1}S^{-1}AP = (P^\intercal SP)^{-1}(P^\intercal AP).$$

Then, as $P^\intercal SP$ and $P^\intercal AP$ are both diagonal matrices with real entries, we deduce that $S^{-1}A$ is diagonalizable with real eigenvalues. The fact that $S^{-1}\mathcal{A}$ is a set of commuting matrices follows similarly.

($\Leftarrow$) Recall that a commuting set of diagonalizable matrices can be simultaneously diagonalized via a similarity transformation, i.e., there exists an invertible $P \in \mathbb{R}^{n\times n}$ such that $P^{-1}S^{-1}AP$ is diagonal for each $A \in \mathcal{A}$ [88]. The diagonal entries of $P^{-1}S^{-1}AP$ are furthermore real by the assumption that $S^{-1}A$ has a real spectrum. For each $A \in \mathcal{A}$, define

$$\bar{A} \coloneqq P^\intercal AP, \qquad D_A \coloneqq P^{-1}S^{-1}AP.$$

Next, note that the identity $P^{-1}S^{-1}AP = (P^\intercal SP)^{-1}(P^\intercal AP)$ can be expressed as $D_A = \bar{S}^{-1}\bar{A}$. Or, equivalently, $\bar{S}D_A = \bar{A}$ for all $A \in \mathcal{A}$. For $i, j \in [n]$, we have the identity

$$\bar{S}_{i,j}(D_A)_{j,j} = \bar{A}_{i,j} = \bar{A}_{j,i} = \bar{S}_{j,i}(D_A)_{i,i} = \bar{S}_{i,j}(D_A)_{i,i}.$$

Here, we have used that $\bar{S}$ and $\bar{A}$ are symmetric and D_A is real diagonal. In particular, if there exists some $A \in \mathcal{A}$ such that $(D_A)_{i,i} \neq (D_A)_{j,j}$, then $\bar{S}_{i,j} = \bar{A}_{i,j} = 0$. Furthermore, by the relation $\bar{S}D_B = \bar{B}$, we also have that $\bar{B}_{i,j} = 0$ for all other $B \in \mathcal{A}$.

We conclude that by permuting the columns of P if necessary (so that $[n]$ is grouped according to the equivalence relation: $i \sim j$ if and only if $(D_A)_{i,i} = (D_A)_{j,j}$ for all $A \in \mathcal{A}$), we can write $\bar{S}$ as a block diagonal matrix $\bar{S} = \operatorname{Diag}(S^{(1)}, \dots, S^{(k)})$. Furthermore, for every $A \in \mathcal{A}$, there exists $\lambda_1, \dots, \lambda_k \in \mathbb{R}$ such that $\bar{A} = \operatorname{Diag}(\lambda_1 S^{(1)}, \dots, \lambda_k S^{(k)})$. It remains to note that each block $S^{(i)}$ can be diagonalized separately. ■

Proposition 22. *Let $\mathcal{A} \subseteq \mathbb{S}^n$ and suppose $S \in \operatorname{span}(\mathcal{A})$ is a max-rank element of $\operatorname{span}(\mathcal{A})$. Then, $\mathcal{A}$ is SDC if and only if $\operatorname{range}(A) \subseteq \operatorname{range}(S)$ for every $A \in \mathcal{A}$ and $\left\{A|_{\operatorname{range}(S)} : A \in \mathcal{A}\right\}$ is SDC.*

Proof. It suffices to show that if $\mathcal{A}$ is SDC then $\operatorname{range}(A) \subseteq \operatorname{range}(S)$ for every $A \in \mathcal{A}$ as then applying Lemma 78 completes the proof.

Let $r = \operatorname{rank}(S)$. Let $P \in \mathbb{R}^{n\times n}$ furnished by SDC. Note that by permuting the columns of P if necessary, we may assume that $P^\intercal SP$ is a diagonal matrix with support contained in its first

r-many diagonal entries. As S is a max-rank element of $\mathrm{span}(\mathcal{A})$, we similarly have that for every $A \in \mathcal{A}$, the matrix $P^{\intercal}AP$ is a diagonal matrix with support contained in its first r-many diagonal entries. For $A \in \mathcal{A}$, write $P^{\intercal}AP = \mathrm{Diag}(\bar{A}, 0_{(n-r)\times(n-r)})$ where $\bar{A}$ is a diagonal $r \times r$ matrix. Then,

$$\mathrm{range}(A) = \mathrm{range}(P^{-\intercal}P^{\intercal}APP^{-1}) \subseteq \mathrm{span}\{q_1, \dots, q_r\}.$$

Here, $q_i \in \mathbb{R}^n$ is the ith column of $P^{-\intercal}$. On the other hand, as $\bar{S}$ has full rank, $\mathrm{range}(S) = \mathrm{span}\{q_1, \dots, q_r\}$. ■

G.2 Facts about matrices with upper triangular Toeplitz blocks

Lemma 103. *Let $(n_1, \dots, n_k)$ with $\sum_i n_i = n$. Suppose $T \in \mathbb{T}$. Then, the characteristic polynomial of T depends only on the entries $\left\{t^{(1)}_{i,j} \,:\, n_i = n_j\right\}$.*

Proof. In this proof, we will use $a, b \in [n]$ to index entries in T (specifically, $T_{a,b} \in \mathbb{R}$ is a scalar, not a matrix block). For each $a \in [n]$, let $i_a \in [k]$ denote the block containing a, and let $\ell_a \in [n_k]$ denote the position of a within block i_a. By the assumption that $T \in \mathbb{T}$, we have

$$T_{a,b} \neq 0 \implies \min\{n_{i_a},\, n_{i_b}\} - n_{i_a} + (\ell_a - \ell_b) \geq 0.$$

Now, for each $a \in [n]$, assign the weight $w_a \coloneqq \ell_a - \frac{n_{i_a}}{2}$. Note that by construction, if $T_{a,b} \neq 0$, then

$$w_a - w_b = \frac{n_{i_b}}{2} - \frac{n_{i_a}}{2} + (\ell_a - \ell_b) \geq 0.$$

Furthermore, note that if $T_{a,b} \neq 0$ and $w_a - w_b = 0$, then $n_{i_a} = n_{i_b}$ and $\ell_a = \ell_b$.

Next, consider a permutation $\sigma \in S_n$ such that $\prod_{a=1}^n T_{a,\sigma(a)} \neq 0$. Note that

$$\sum_{a=1}^n w_a - w_{\sigma(a)} = \sum_{a=1}^n w_a - \sum_{a=1}^n w_{\sigma(a)} = 0.$$

Then, by the above paragraph, we conclude that σ satisfies $n_{i_a} = n_{i_{\sigma(a)}}$ and $\ell_a = \ell_\sigma(a)$ for all $a \in [n]$.

Returning to the previous notation, the characteristic polynomial of T depends only on the entries $\left\{t^{(1)}_{i,j} \,:\, n_i = n_j\right\}$. ■

Lemma 81. *Let $(n_1, \dots, n_k)$ such that $\sum_i n_i = n$. Then, for any $T \in \mathbb{T}$, the matrices $T \in \mathbb{R}^{n\times n}$ and $\Pi(T) \in \mathbb{R}^{k\times k}$ have the same eigenvalues.*

Proof. Without loss of generality, suppose $n_1 \leq \cdots \leq n_k$ and let $T \in \mathbb{T}$. By Lemma 103, T has the same eigenvalues as the matrix $\hat{T} \in \mathbb{T}$ with entries

$$\hat{T}^{(\ell)}_{i,j} = \begin{cases} T^{(\ell)}_{i,j} & \text{if } n_i = n_j,\ \ell = 1, \\ 0 & \text{else.} \end{cases}$$

Now, suppose that there are m distinct block sizes $s_1, \dots, s_m$. Partitioning both $\Pi(T)$ and $\hat{T}$ according to $s_1, \dots, s_m$, we have that

$$\Pi(T) = \operatorname{Diag}(\tilde{T}_1, \dots, \tilde{T}_m) \quad \text{and} \quad \bar{T} = \operatorname{Diag}(\tilde{T}_1 \otimes I_{s_1}, \dots, \tilde{T}_m \otimes I_{s_m}).$$

We conclude that $\Pi(T)$ and $\bar{T}$ have the same eigenvalues. ■

G.3 Details for the Hermitian case

Let $\mathbb{H}^n$ denote the real vector space of $n \times n$ Hermitian matrices. For $v \in \mathbb{C}^n$ and $A \in \mathbb{C}^{n\times n}$ let v^* and A^* denote the conjugate transpose of v and A respectively.

Definitions and theorem statements

Almost all of our results extend verbatim to the Hermitian setting. For brevity, we only state our more interesting definitions and results as adapted to this setting.

Definition 34. A set $\mathcal{A} \subseteq \mathbb{H}^n$ is *simultaneously diagonalizable via congruence* (SDC) if there exists an invertible $P \in \mathbb{C}^{n\times n}$ such that P^*AP is diagonal for all $A \in \mathcal{A}$. □

Definition 35. A set $\mathcal{A} \subseteq \mathbb{H}^n$ is *almost simultaneously diagonalizable via congruence* (ASDC) if there exist sequences $A_i \to A$ for every $A \in \mathcal{A}$ such that for every $i \in \mathbb{N}$, the set $\{A_i : A \in \mathcal{A}\}$ is SDC. □

Definition 36. A set $\mathcal{A} \subseteq \mathbb{H}^n$ is *nonsingular* if there exists a nonsingular $A \in \operatorname{span}(\mathcal{A})$. Else, it is *singular*. □

Definition 37. Given a set $\mathcal{A} \subseteq \mathbb{H}^n$, we will say that $S \in \mathcal{A}$ is a *max-rank element* of $\operatorname{span}(\mathcal{A})$ if $\operatorname{rank}(S) = \max_{A\in\mathcal{A}} \operatorname{rank}(A)$. □

Theorem 42. *Let $A, B \in \mathbb{H}^n$ and suppose A is invertible. Then, $\{A, B\}$ is ASDC if and only if $A^{-1}B$ has real eigenvalues.*

Theorem 43. *Let $\{A, B\} \subseteq \mathbb{H}^n$. If $\{A, B\}$ is singular, then it is ASDC.*

Theorem 44. *Let $\{A, B, C\} \subseteq \mathbb{H}^n$ and suppose A is invertible. Then, $\{A, B, C\}$ is ASDC if and only if $\{A^{-1}B, A^{-1}C\}$ are a pair of commuting matrices with real eigenvalues.*

Definition 38. Let $\mathcal{A} \subseteq \mathbb{H}^n$ and $d \in \mathbb{N}$. We will say that $\mathcal{A}$ is *d-restricted SDC* (d-RSDC) if there exist matrices $\bar{A} \in \mathbb{H}^{n+d}$ containing A as its top-left $n \times n$ principal submatrix for every $A \in \mathcal{A}$ such that $\left\{\bar{A} : A \in \mathcal{A}\right\}$ is SDC. □

Theorem 45. *Let $A, B \in \mathbb{H}^n$. Then for every $\epsilon > 0$, there exist $\tilde{A}, \tilde{B} \in \mathbb{H}^n$ such that $\left\|A - \tilde{A}\right\|, \left\|B - \tilde{B}\right\| \leq \epsilon$ and $\left\{\tilde{A}, \tilde{B}\right\}$ is 1-RSDC. Furthermore, if A is invertible and $A^{-1}B$ has simple eigenvalues, then $\{A, B\}$ is itself 1-RSDC.*

Theorem 46. *Let $\{A = I_n, B, C\} \subseteq \mathbb{H}^n$. Then, if $d < \operatorname{rank}([B, C])/2$, the set*

$$\left\{\begin{pmatrix} A & \\ & 0_d \end{pmatrix}, \begin{pmatrix} B & \\ & 0_d \end{pmatrix}, \begin{pmatrix} C & \\ & 0_d \end{pmatrix}\right\}$$

is not ASDC.

Theorem 47. *There exists a set $\mathcal{A} = \{A_1, \dots, A_5\} \subseteq \mathbb{H}^4$ such that A_1 is invertible, $A_1^{-1}\mathcal{A}$ is a set of commuting matrices with real eigenvalues, and $\mathcal{A}$ is not ASDC.*

In the Hermitian setting, the statement in Theorem 39 should be changed to: “There exists a set $\mathcal{A} = \{A_1, \dots, A_5\} \subseteq \mathbb{H}^4$ such that A_1 is invertible, $A_1^{-1}\mathcal{A}$ is a set of commuting matrices with real eigenvalues and $\mathcal{A}$ is not ASDC.” The proof is unchanged after setting

$$A_1 = \begin{pmatrix} &&&1\\ &&1&\\ &1&&\\ 1&&& \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0&&&\\ &0&&\\ &&1&\\ &&&0 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 0&&&\\ &0&&\\ &&0&1\\ &&1&0 \end{pmatrix},$$
$$A_4 = \begin{pmatrix} 0&&&\\ &0&&\\ &&0&\mathrm{i}\\ &&-\mathrm{i}&0 \end{pmatrix}, \quad A_5 = \begin{pmatrix} 0&&&\\ &0&&\\ &&0&\\ &&&1 \end{pmatrix}.$$

Necessary modifications

Next, we discuss technical changes that need to be made to adapt our proofs from the real symmetric setting to the Hermitian setting. For brevity, we only list changes beyond the trivial changes, e.g., replacing $\mathbb{S}^n$ by $\mathbb{H}^n$, $\mathbb{R}^{n\times n}$ by $\mathbb{C}^{n\times n}$, and $^\intercal$ by *.

- In the Hermitian version of Proposition 23, the m_2-many blocks corresponding to non-real eigenvalues (previously (7.2)) will have the form

  $$S_i = F_{2n_i}, \qquad T_i = F_{n_i} \otimes \begin{pmatrix} & \lambda_i^* \\ \lambda_i & \end{pmatrix} + G_{n_i} \otimes F_2$$

  where $n_i \in \mathbb{N}$ and $\lambda_i \in \mathbb{C} \setminus \mathbb{R}$. See [105, Theorem 9.2] for further details.

- In the proof of Lemma 79, note that for all $i \in [r+1, m]$, the block

  $$S_i^{-1}\tilde{T}_i = I_{n_i} \otimes \begin{pmatrix} \lambda_i & \\ & \lambda_i^* \end{pmatrix} + (\eta_i I_{n_i} + F_{n_i} G_{n_i} + \delta F_{n_i} H_{n_i}) \otimes I_2.$$

  The remainder of the proof is unchanged.

- In the proof of Theorem 34, we will work in the basis furnished by the Hermitian version of Proposition 23 for $\mathbb{C}^{2k}$. That is, we may assume in the first two cases that A and B (previously (7.6)) have the form

  $$A = \left(\begin{array}{c|c|c|c} \begin{smallmatrix} & 1 \\ 1 & \end{smallmatrix} & & & \\ \hline & \ddots & & \\ \hline & & \begin{smallmatrix} & 1 \\ 1 & \end{smallmatrix} & \\ \hline & & & S_m \end{array}\right), \qquad B = \left(\begin{array}{c|c|c|c} \begin{smallmatrix} & \lambda_1^* \\ \lambda_1 & \end{smallmatrix} & & & \\ \hline & \ddots & & \\ \hline & & \begin{smallmatrix} & \lambda_k^* \\ \lambda_k & \end{smallmatrix} & \\ \hline & & & T_m \end{array}\right).$$

We will set $\tilde{A}_\delta$ as in the Hermitian case for both Cases 1 and 2. We will set $\tilde{B}_\delta$ to be

$$\tilde{B}_\delta = \left(\begin{array}{cc|c|cc|c}
 & \lambda_1^* & & & & \alpha_1\sqrt{-\delta\mathrm{i}/2} \\
\lambda_1 & & & & & \left(\alpha_1\sqrt{-\delta\mathrm{i}/2}\right)^* \\ \hline
 & & \ddots & & & \vdots \\ \hline
 & & & & \lambda_k^* & \alpha_k\sqrt{-\delta\mathrm{i}/2} \\
 & & & \lambda_k & & \left(\alpha_k\sqrt{-\delta\mathrm{i}/2}\right)^* \\ \hline
\left(\alpha_1\sqrt{\frac{-\delta\mathrm{i}}{2}}\right)^* & \alpha_1\sqrt{\frac{-\delta\mathrm{i}}{2}} & \cdots & \left(\alpha_k\sqrt{\frac{-\delta\mathrm{i}}{2}}\right)^* & \alpha_k\sqrt{\frac{-\delta\mathrm{i}}{2}} & \delta z
\end{array}\right)$$

and

$$\tilde{B}_\delta = \left(\begin{array}{cc|c|cc|c|c|c}
 & \lambda_1^* & & & & & \alpha_1\sqrt{-\delta\mathrm{i}/2} & \\
\lambda_1 & & & & & & \left(\alpha_1\sqrt{-\delta\mathrm{i}/2}\right)^* & \\ \hline
 & & \ddots & & & & \vdots & \\ \hline
 & & & & \lambda_k^* & & \alpha_k\sqrt{-\delta\mathrm{i}/2} & \\
 & & & \lambda_k & & & \left(\alpha_k\sqrt{-\delta\mathrm{i}/2}\right)^* & \\ \hline
 & & & & & & & G_{n_m} \\ \hline
\left(\alpha_1\sqrt{\frac{-\delta\mathrm{i}}{2}}\right)^* & \alpha_1\sqrt{\frac{-\delta\mathrm{i}}{2}} & \cdots & \left(\alpha_k\sqrt{\frac{-\delta\mathrm{i}}{2}}\right)^* & \alpha_k\sqrt{\frac{-\delta\mathrm{i}}{2}} & & \delta z & e_1^\intercal \\ \hline
 & & & & & G_{n_m} & e_1 &
\end{array}\right)$$

for Cases 1 and 2, respectively. Here, $\alpha \in \mathbb{C}^k$, $z \in \mathbb{R}$, and $\delta > 0$. The characteristic polynomials of $\tilde{A}_\delta^{-1}\tilde{B}_\delta$ are given by (7.8) and (7.13) in Cases 1 and 2 respectively. The remainder of the proof remains unchanged.

G.4 An example where the SDC property is preserved under restriction

In this section, we give an example of a setting in which the restriction of an SDC set to one of its principal submatrices results in another SDC set. This setting arises for example in QCQPs [93].

Proposition 25. *Let* $A_1, \dots, A_m \in \mathbb{S}^n$ *such that* $\operatorname{span}(\{A_1, \dots, A_m\})$ *contains a positive definite matrix. Let* $b_1, \dots, b_m \in \mathbb{R}^n$ *and* $c_1, \dots, c_m \in \mathbb{R}$*, and define*

$$Q_i = \begin{pmatrix} A_i & b_i \\ b_i^\intercal & c_i \end{pmatrix} \in \mathbb{S}^{n+1}.$$

If $\{Q_1, \dots, Q_m, e_{n+1}e_{n+1}^\intercal\}$ *is SDC, then so is* $\{A_1, \dots, A_m\}$*.*

Proof. Without loss of generality, let $A_1 \succ 0$. Note that for all $\lambda \in \mathbb{R}$ large enough, the matrix $S_\lambda \coloneqq Q_1 + \lambda e_{n+1}e_{n+1}^\intercal \succ 0$. By the inverse formula for a block matrix [88], we have that for all λ large enough,

$$S_\lambda^{-1} = \begin{pmatrix} A_1^{-1} + \frac{A_1^{-1}b_1b_1^\intercal A_1^{-1}}{\lambda+(c_1-b_1^\intercal A_1 b_1)} & \frac{-A_1^{-1}b_1}{\lambda+(c_1-b_1A_1^{-1}b_1)} \\ \frac{-b_1^\intercal A_1^{-1}}{\lambda+(c_1-b_1A_1^{-1}b_1)} & \frac{1}{\lambda+(c_1-b_1A_1^{-1}b_1)} \end{pmatrix}.$$

In particular,

$$\lim_{\lambda\to\infty} S_\lambda^{-1} = \begin{pmatrix} A_1^{-1} & \\ & 0 \end{pmatrix}.$$

On the other hand, by Lemma 77, we have that for all $i, j \in [m]$,

$$0 = \left[S_\lambda^{-1} Q_i,\, S_\lambda^{-1} Q_j\right].$$

Finally, by continuity we have that

$$0 = \lim_{\lambda\to\infty} \left[S_\lambda^{-1} Q_i,\, S_\lambda^{-1} Q_j\right] = \begin{pmatrix} \left[A_1^{-1} A_i,\, A_1^{-1} A_j\right] & \\ & 0 \end{pmatrix}.$$

We conclude that $A_1^{-1}\{A_1, \ldots, A_m\}$ commute, whence by Lemma 77 this set is SDC. ■