Error Bounds and Singularity Degree in Semidefinite Programming

by

Stefan Sremac

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Combinatorics and Optimization

Waterloo, Ontario, Canada, 2019

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Tamás Terlaky Professor, Industrial and Systems Engineering Lehigh University
Supervisor:	Henry Wolkowicz Professor, Combinatorics and Optimization University of Waterloo
Internal Member:	Levent Tunçel Professor, Combinatorics and Optimization University of Waterloo
Internal Member:	Stephen Vavasis Professor, Combinatorics and Optimization University of Waterloo
Internal-External Member:	Patrick Mitran Associate Professor, Electrical and Computer Engineering University of Waterloo

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

An important process in optimization is to determine the quality of a proposed solution. This usually entails calculation of the distance of a proposed solution to the optimal set and is referred to as *forward error*. Since the optimal set is not known, we generally view forward error as intractable. An alternative to forward error is to measure the violation in the constraints or optimality conditions. This is referred to as *backward error* and it is generally easy to compute. A major issue in optimization occurs when a proposed solution has small backward error, i.e., looks good to the user, but has large forward error, i.e., is far from the optimal set.

In [77], Jos Sturm developed a remarkable upper bound on forward error for spectrahedra (optimal sets of semidefinite programs) in terms of backward error. His bound creates a hierarchy among spectrahedra that is based on *singularity degree*, an integer between 0 and $n-1$, derived from *facial reduction*. For problems with small singularity degree, forward error is similar to backward error, but this may not be true for problems with large singularity degree.

In this thesis we provide a method to obtain numerical lower bounds on forward error, thereby complimenting the bounds of Sturm. While the bounds of Sturm identify good convergence, our bounds allow us to detect poor convergence. Our approach may also be used to provide lower bounds on singularity degree, a measure that is difficult to compute in some instances. We show that large singularity degree leads to some undesirable convergence properties for a specific family of *central paths*.

We apply our results in a theoretical sense to some Toeplitz matrix completion problems and in a numerical sense to several test spectrahedra.

Acknowledgements

I would like to thank the committee members Tamás Terlaky, Levent Tunçel, Stephen Vavasis, and Patrick Mitran for making the time to take part in my examination, especially considering the short timeframe that I provided.

The four and a half years I have spent at Waterloo have been great in many ways and I would like to acknowledge the people who have been a part of my experience.

Working with my supervisor, Henry, has been an experience I will cherish for a long time. His insight into optimization and his approach to problem solving are inspiring, to say the least. What has left the greatest impression on me, though, is his unwavering optimism and work ethic. I am thankful for Henry's willingness to send me to conferences, arrange talks for me, and introduce me to many wonderful members of the mathematical community, some of whom I have had the opportunity to collaborate with. I have enjoyed the many conversations we have had, both on mathematical topics and otherwise. In particular I appreciate his value of family and his understanding during the pregnancy and birth of my two children.

I am greatly indebted to my collaborators Fei, Hao, Henry, and Hugo. I am very thankful to Hugo for his incredible breadth of mathematical knowledge and for introducing me to Bezoutians, the Gohberg-Semencul formula, and other fascinating concepts and results.

I am thankful to the many great people of the Combinatorics and Optimization department. In particular, I learned a great deal from the classes of Henry, Levent, Steve, and Bill. I can definitely say that I looked forward to their lectures. My experiences with the department and graduate chairs – Jochen, Swamy, and Jim – were very positive and on the few occasions that I made a request of them, they went out of their way to accommodate me. While several staff members have come and gone, Melissa and Carol have been here the entire time and they have made administrative tasks painless. I especially appreciate a very encouraging conversation I had with Melissa toward the end of a difficult first term in the department!

To the many friends I have shared an office with, thank you for some great memories and exciting crossword sessions! I greatly value the many conversations Jimit and I have had over the years and the winter camping expeditions I undertook with Matt.

I am very grateful to have been a part of the Waterloo Seventh-Day Adventist Church during my time here. Thank you for welcoming our family so wonderfully. I looked forward to worshipping with you every Sabbath!

I am thankful to my family for supporting me throughout my life. I thank my parents, Sima and Brankica, for having instilled in me so many important values and for introducing me to God. They have provided crucial guidance and helped in so many ways throughout my time here. I thank my sisters, Lidija and Ana, and their families for being so loving, for visiting us, and for providing us many reasons to frequently visit the greatest place

on earth, B.C. I would also like to thank my new family, the Boyds, for their help and support. In particular I am thankful to Bruce and Loma for assisting so much with Jodi's pregnancies and our babies.

I could not have asked for a better partner in life than my wife Jodi. You have patiently endured these four years of low-income-living with a very busy husband! Thank you for your selflessness and for allowing me to pursue this goal! The greatest 'accomplishment' during our time here has, without a doubt, been the birth of our two children, Jerra and Malachi. Thank you for being an amazing mother!

I thank my Heavenly Father for all that I have been given in life. Thank you for the many doors that have opened for me and for realizing this journey that I never even planned to embark on! "Praise God from whom all blessing flow!"

Let us not grow weary of doing good,
for in due season we will reap, if we do not give up.
- Galatians 6:9

To my wife Jodi and my father Sima.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction and Overview of Contributions

In this thesis we make contributions to the understanding of error bounds and *singularity degree* in semidefinite programming. Our study is motivated by well-known, hard instances of semidefinite programming on which state-of-the-art algorithms converge very slowly.

In order to state our contributions explicitly, we need an understanding of the type of slow convergence we are concerned with. To this end, let $\mathcal{F} \subset \mathbb{S}^n$ be the solution set of a *semidefinite program (SDP)*. Throughout this thesis we refer to $\mathcal{F}$ as a *spectrahedron*. Here $\mathbb{S}^n$ denotes the Euclidean space of $n \times n$ symmetric matrices. It is always possible to express $\mathcal{F}$ as the intersection of an affine subspace, $\mathcal{L}$, and the set of positive semidefinite matrices, $\mathbb{S}^n_+$. Given a matrix $X \in \mathbb{S}^n$, the *forward error* of X with respect to $\mathcal{F}$ is defined as,

$$\epsilon^f(X, \mathcal{F}) := \operatorname{dist}(X, \mathcal{F}). \tag{1.0.1}$$

Here, $\operatorname{dist}(X, \mathcal{F})$ denotes the distance from X to $\mathcal{F}$. See (2.1.1) for a definition specific to our setting. We cannot expect to measure forward error accurately without substantial knowledge of the solution set $\mathcal{F}$. For this reason forward error is generally unknown. What is readily available to users is the *backward error* of X with respect to $\mathcal{F}$,

$$\epsilon^b(X, \mathcal{F}) := \operatorname{dist}(X, \mathcal{L}) + \operatorname{dist}(X, \mathbb{S}^n_+). \tag{1.0.2}$$

In backward error it is recognized that $\mathcal{F}$ is the intersection of two sets with easily computable forward errors. For this reason, backward error is used as a proxy for forward error. Backward error also measures how much $\mathcal{L}$ or $\mathbb{S}^n_+$ need to be perturbed in order for X to be feasible for $\mathcal{F}$.

The type of slow convergence we are concerned with is when *backward error is sufficiently small but forward error is much larger*. The problem with this scenario is not just the poor quality of the proposed solution. More than this, it is the lack of awareness of a poor solution.

To demonstrate how severe the discrepancy between forward error and backward error can be, we consider an SDP, with $n = 5$, from the family introduced in [80] and stated in Example 4.2.6 of this thesis. The output of `cvx`, a package for specifying and solving convex optimization problems [28, 29], using the solver `SDPT3`, [81], is,

$$X \approx \begin{bmatrix} 0.94 & 0 & 0.028 & 0.001 & 2.3 \times 10^{-6} \\ 0 & 0.057 & 0 & 0 & 0 \\ 0.028 & 0 & 0.028 & 4.1 \times 10^{-5} & 6.5 \times 10^{-8} \\ 0.001 & 0 & 4.1 \times 10^{-5} & 4.5 \times 10^{-6} & 3.1 \times 10^{-9} \\ 2.3 \times 10^{-6} & 0 & 6.5 \times 10^{-8} & 3.1 \times 10^{-9} & 0 \end{bmatrix}.$$

Similar results were obtained with the solvers `SeDuMi`, [76], and `MOSEK`, [1]. The backward error for X is quite small at 5.46×10^{-12} and `cvx` output states that the problem is "solved". All indicators point to a 'good' solution. However, the solution set of the SDP is a singleton consisting of the matrix with 1 in the upper left entry and zeros everywhere else. Given this information, X does *not* look like a very good solution. Indeed, forward error is 9.15×10^{-2}.

Much of the literature devoted to addressing this disparity focuses on upper bounds of the form,

$$\epsilon^f(X, \mathcal{F}) = \mathcal{O}\left(\epsilon^b(X, \mathcal{F})^\gamma\right), \tag{1.0.3}$$

for $\gamma \in (0, 1]$ and backward error that is sufficiently small. Here $\mathcal{O}$ denotes the usual 'big-O' notation, although our usage in (1.0.3) is inconsistent with the definition in Section 2.7. What we mean to say by (1.0.3) is that $\epsilon^f(X, \mathcal{F})$ is in 'the order of $\epsilon^b(X, \mathcal{F})^\gamma$' or smaller, when the backward error is sufficiently small. Precise statements about error bounds will be developed throughout the thesis. For the example above, the exponent is $\gamma \approx 1/6$.

Bounds of the type in (1.0.3) are referred to as *Hölderian error bound*s. Large values of γ indicate small discrepancies between forward error and backward error, while small values of γ indicate large discrepancies.

The utility of this type of bound lies in detecting families of spectrahedra where forward error and backward error are similar. Specifically, if for a given family of spectrahedra, a bound such as (1.0.3) exists with $\gamma \approx 1$, then backward error is a valid approximation of forward error for any member of the family. The limitation of this type of bound is that it cannot be used to detect the scenario of the above example where forward error is much larger than backward error. If γ is much smaller than 1, we may only conclude that there *may* be a large discrepancy between the two measures, but it is also possible that there is no distinction between them. To detect scenarios where forward error is much larger than backward error we need to complement the upper bound of (1.0.3) with a lower bound. Hence our first contribution.

Contribution 1: A method to provide a numerical lower bound on forward error for spectrahedra, (Chapter 5).

In [77], Sturm introduced a bound of the type in (1.0.3), that is perhaps the most intriguing and informative among such bounds in the literature of SDPs. To state his bound, Sturm defined *singularity degree*, a positive integer that is derived from the *facial reduction algorithm* of Borwein and Wolkowicz, [6, 7, 8]. The singularity degree of a spectrahedron, $\mathcal{F}$, is denoted by $\mathrm{sd}(\mathcal{F})$ and the corresponding bound is,

$$\epsilon^f(X,\mathcal{F}) = \mathcal{O}\left(\epsilon^b(X,\mathcal{F})^{2^{-\mathrm{sd}(\mathcal{F})}}\right). \tag{1.0.4}$$

Singularity degree zero corresponds to no discrepancy between forward error and backward error, while larger values of singularity degree imply the possibility of larger discrepancies.

The challenge with singularity degree is that, like forward error, it is generally unknown. In [10], a facial reduction algorithm is presented that is backwards stable when singularity degree is 0 or 1. However, we are not aware of a stable algorithm for performing facial reduction for problems with larger singularity degree. Moreover, the empirical evidence we have obtained indicates a lack of stability of the algorithm when singularity degree is greater than 1. For this reason, we view finding singularity degree as intractable for general instances of SDP. The following contributions address this problem.

Contribution 2: Theoretical bounds on $\mathrm{sd}(\widehat{\mathcal{F}})$*, where* $\widehat{\mathcal{F}}$ *is obtained by certain transformations of a given* $\mathcal{F}$*, where* $\mathrm{sd}(\mathcal{F})$ *is known, (Chapter 4).*

Contribution 3: A method to provide a numerical lower bound for the singularity degree of a spectrahedron, (Chapter 5).

An immediate consequence of the bound of Sturm, is that large singularity degree is a necessary property of spectrahedra that exhibit large discrepancy between forward error and backward error. We show that the converse is, in some sense, also true, thereby providing evidence that singularity degree may be viewed as a measure-of-hardness for solving SDPs.

Contribution 4: Large singularity degree is a sufficient condition for high irregularity in the convergence of a family of external-type central paths, (Chapter 6).

Having established the aforementioned theoretical results, we study the singularity degree of spectrahedra formed by solution sets to certain *Toeplitz* matrix completion problems. In a somewhat peripheral endeavor we make the following contribution. We opt for a vague description of the result in an attempt to avoid excessive terminology at this point.

Contribution 5: A characterization of partial Toeplitz patterns that admit a certain type of 'nice' positive definite completion, (Chapter 7).

The thesis is structured as follows. In Chapter 2, we introduce notation and highlight elementary results pertaining to positive semidefinite matrices, SDPs, and convex analysis.

The notions of facial reduction and singularity degree are discussed in Chapter 3 and in Chapter 4, respectively. The main results are presented in Chapter 4 through Chapter 7, as outlined in the overview of contributions, above. We demonstrate the quality and utility of bounds developed throughout the thesis by numerically analyzing several test spectrahedra in Chapter 8. Concluding remarks are made in Chapter 9.

Some of the results in Chapter 5 and in Chapter 6 are based on the preprint [74] and some of the results of Chapter 7 are based on the published article [75].

We forego a review of the literature at this point, choosing instead to insert the material in the more immediate context of the chapters to which the literature is relevant.

Chapter 2

Background and Notation

In this chapter we present brief introductions to some of the topics of the thesis and present elementary and classical results without proof. We assume a basic knowledge of convex analysis, optimization, and undergraduate-level mathematics. For further reading or proofs of some of the unreferenced claims we suggest Wolkowicz, Saigal, and Vandenberghe [86] and Tunçel [80] for SDPs, Rockafellar [71] and Rockafellar and Wets [72] for convex analysis, and Nocedal and Wright [59] and Boyd and Vandenberghe [9] for optimization.

2.1 Symmetric and Positive Semidefinite Matrices

Our ambient space is the Euclidean space of $n \times n$ real symmetric matrices, denoted $\mathbb{S}^n$, with the standard trace inner product, $\langle X, Y \rangle := \text{trace}(XY)$, and the induced Frobenius norm, $\|X\|_F := \sqrt{\langle X, X \rangle}$.

The eigenvalues of any $X \in \mathbb{S}^n$ are real and ordered so as to satisfy,

$$\lambda_1(X) \geq \cdots \geq \lambda_n(X),$$

and $\lambda(X) \in \mathbb{R}^n$ is the vector consisting of all the eigenvalues. In terms of this notation we have $\|X\|_F = \|\lambda(X)\|_2$, where $\|\cdot\|_2$ is the Euclidian norm when the argument is a vector in $\mathbb{R}^n$. When the argument to $\|\cdot\|_2$ is a symmetric matrix, then we mean the operator 2-norm, defined as $\|X\|_2 := \max_i |\lambda_i(X)|$. In some of our discussion we use the notation $\lambda_{\max}(X) = \lambda_1(X)$ and $\lambda_{\min}(X) = \lambda_n(X)$ if we are not concerned with the dimensions of the matrix, or wish to stress the minimality and maximality of the values. The distance from $X \in \mathbb{S}^n$ to a set $\mathcal{S} \subseteq \mathbb{S}^n$ is defined as,

$$\text{dist}(X, \mathcal{S}) := \inf_{Y \in \mathcal{S}} \|X - Y\|_F. \tag{2.1.1}$$

The set of positive semidefinite matrices, $\mathbb{S}^n_+$, is a closed convex cone in $\mathbb{S}^n$, with interior consisting of the positive definite matrices, $\mathbb{S}^n_{++}$. The cone $\mathbb{S}^n_+$ induces the *Löwner partial*

order on $\mathbb{S}^n$. That is, for $X, Y \in \mathbb{S}^n$ we write $X \succeq Y$ when $X - Y \in \mathbb{S}^n_+$ and similarly $X \succ Y$ when $X - Y \in \mathbb{S}^n_{++}$.

We collect some elementary results regarding positive semidefinite matrices in the following.

Fact 2.1.1. *Let $X, Y \in \mathbb{S}^n_+$ and $u \in \mathbb{R}^n$. Then the following hold.*

(*i*) $u^T X u = 0$ *if, and only if,* $u \in \text{null}(X)$.

(*ii*) $\langle X, Y \rangle = 0$ *if, and only if,* $XY = 0$.

(*iii*) $\text{null}(X + Y) = \text{null}(X) \cap \text{null}(Y)$.

(*iv*) $\text{range}(X + Y) = \text{range}(X) + \text{range}(Y)$.

(*v*) $\text{rank}(X + Y) \geq \max\{\text{rank}(X), \text{rank}(Y)\}$.

The '+' in Fact 2.1.1 (*iv*) denotes the usual *Minkowski sum*. A nice property of symmetric matrices is the relationship between eigenvalues of the matrix itself and the eigenvalues of any of its symmetric submatrices.

Fact 2.1.2 (Interlacing Eigenvalues, Corollary 1.21, [80])**.** *Let $X \in \mathbb{S}^n$ and let $\bar{X}$ be a symmetric submatrix of X of order r. Then,*

$$\lambda_{n-(r-k)}(X) \leq \lambda_k(\bar{X}) \leq \lambda_k(X), \ \forall k \in \{1, \ldots, r\}.$$

A classical result for symmetric matrices is the following bound on the trace inner product in terms of eigenvalue vectors. It may not be obvious at first, but this bound is equivalent to the Hoffman-Wielandt inequality, [36].

Fact 2.1.3 (Lemma 1, [25])**.** *For $X, Y \in \mathbb{S}^n$ it holds that,*

$$\sum_{i=1}^{n} \lambda_i(X)\lambda_{n+1-i}(Y) \leq \langle X, Y \rangle \leq \lambda(X)^T \lambda(Y).$$

2.2 Linear Maps and Spectrahedra

For every linear map $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$, there exist $A_1, \ldots, A_m \in \mathbb{S}^n$ such that,

$$\big(\mathcal{A}(X)\big)_i = \langle X, A_i \rangle, \quad \forall i \in \{1, \ldots, m\}. \tag{2.2.1}$$

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

$$\langle \mathcal{A}(X), y \rangle = \langle X, \mathcal{A}^*(y) \rangle, \quad \forall X \in \mathbb{S}^n, \, y \in \mathbb{R}^m.$$

When the inner product on $\mathbb{R}^m$ is the usual $\mathcal{A}(X)^T y$, then $\mathcal{A}^*(y) = \sum_{i=1}^m y_i A_i$. It follows that $\text{range}(\mathcal{A}^*) = \text{span}\{A_1, \ldots, A_m\}$. For $M \in \mathbb{R}^{n\times p}$ we denote the composition of the maps $\mathcal{A}$ and $M \cdot M^T$ as,

$$\mathcal{A}_M(X) := \mathcal{A}(MXM^T). \tag{2.2.2}$$

It is a simple observation that the map $M \cdot M^T$ is an automorphism of $\mathbb{S}^n_+$ when M is square and non-singular.

For a given linear map $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$ and a vector $b \in \mathbb{R}^m$, we define

$$\mathcal{L}(\mathcal{A}, b) := \{X \in \mathbb{S}^n : \mathcal{A}(X) = b\}.$$

When $b = 0$, then $\mathcal{L}(\mathcal{A}, b) = \text{null}(\mathcal{A})$, a linear subspace of $\mathbb{S}^n$. The following Farkas Lemma type of result about linear subspaces of $\mathbb{S}^n$ is used throughout the thesis. For a proof see, e.g., Corollary 2 of [51].

Fact 2.2.1. *Let $\mathcal{L} \subset \mathbb{S}^n$ be a linear subspace. Then it holds that,*

$$\mathcal{L} \cap \mathbb{S}^n_+ = \{0\} \iff \mathcal{L}^\perp \cap \mathbb{S}^n_{++} \neq \emptyset.$$

In Fact 2.2.1, $\mathcal{L}^\perp$ denotes the usual orthogonal complement of $\mathcal{L}$. Our main object of study is the spectrahedron, defined in the following.

Definition 2.2.2. *Given a linear map $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$ and a vector $b \in \mathbb{R}^m$, we define the* spectrahedron $\mathcal{F}(\mathcal{A}, b)$ *as the intersection of $\mathbb{S}^n_+$ with the affine manifold defined by $\mathcal{A}$ and b. That is,*

$$\mathcal{F}(\mathcal{A}, b) := \mathcal{L}(\mathcal{A}, b) \cap \mathbb{S}^n_+ = \{X \in \mathbb{S}^n_+ : \mathcal{A}(X) = b\}.$$

Bounded spectrahedra play an important role in this thesis. To derive a characterization of such spectrahedra we recall the notion of recession cone. For a convex set $C \subseteq \mathbb{S}^n$ the *recession cone of C*, denoted C^∞, is defined as,

$$C^\infty := \{X \in \mathbb{S}^n : Y + \tau X \in C,\ \forall Y \in C,\ \forall \tau \geq 0\}.$$

Lemma 2.2.3. *For a non-empty spectrahedron $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$, the following are equivalent:*

(i) $\mathcal{F}$ *is bounded,*

(ii) $\text{null}(\mathcal{A}) \cap \mathbb{S}^n_+ = \{0\}$,

(iii) $\text{range}(\mathcal{A}^*) \cap \mathbb{S}^n_{++} \neq \emptyset$.

Proof. First note that *(ii)* and *(iii)* are equivalent by Fact 2.2.1. Therefore, it suffices to show that *(i)* is equivalent to *(ii)*. By Theorem 8.4 of [71] and the convexity of $\mathcal{F}$ it holds that,

$$\mathcal{F} \text{ is bounded} \iff \mathcal{F}^\infty = \{0\}. \tag{2.2.3}$$

Next, the definition of $\mathcal{F}$ and Corollary 8.3.3 of [71] yield,

$$\mathcal{F}^\infty = \mathcal{L}(\mathcal{A}, b)^\infty \cap \left(\mathbb{S}^n_+\right)^\infty. \tag{2.2.4}$$

Now the recession cone of an affine subspace is the linear subspace that is parallel to it. Hence $\mathcal{L}(\mathcal{A}, b)^\infty = \text{null}(\mathcal{A})$. Moreover, the recession cone of a closed convex cone is the cone itself. Thus $(\mathbb{S}^n_+)^\infty = \mathbb{S}^n_+$. Combining these observations with (2.2.3) and (2.2.4) gives us,

$$\mathcal{F} \text{ is bounded} \iff \text{null}(\mathcal{A}) \cap \mathbb{S}^n_+ = \{0\},$$

as desired. □

2.3 Convexity in $\mathbb{S}^n_+$

Given a convex set $C \subset \mathbb{S}^n$, we denote by $\text{cl}(C)$ and $\text{relint}(C)$, the *closure* and *relative interior* of C. We define the *boundary* of a set C as those elements of C that do not belong to the relative interior:

$$\text{bd}(C) = \text{cl}(C) \setminus \text{relint}(C).$$

This definition of boundary is referred to as *relative boundary* in some texts. The convex subsets of $\mathbb{S}^n_+$ exhibit some remarkable properties when viewed through the rank function. Namely that the rank function is constant over, and maximized by, the relative interior. To state this relationship more clearly, we introduce the following.

Definition 2.3.1. *The rank of a non-empty set $C \in \mathbb{S}^n_+$ is defined as,*

$$\text{rank}(C) := \max_{X \in C} \text{rank}(X),$$

where $\text{rank}(\cdot)$ *in the objective of the optimization problem denotes the usual rank of a matrix.*

Lemma 2.3.2. *Let $C \subseteq \mathbb{S}^n_+$ be convex and non-empty and let* $\text{rank}(C)$ *be as in Definition 2.3.1. Then,* $\text{rank}(C) = \text{rank}(X)$ *for any* $X \in \text{relint}(C)$.

Proof. Since the image of C under the matrix rank function consists of finitely many elements, it follows that $\text{rank}(C)$ is attained by some $\bar{X} \in C$. Now let $X \in \text{relint}(C)$. By the convexity of C and Theorem 6.4 of [71], there exists $Y \in C$ and $\theta \in (0, 1]$ such that,

$$X = \theta \bar{X} + (1 - \theta) Y.$$

By definition of $\bar{X}$ we have, $\text{rank}(\bar{X}) \geq \text{rank}(X)$ and by Fact 2.1.1 (*i*) we have,

$$\text{rank}(X) \geq \max\{\text{rank}(\theta \bar{X}), \text{rank}((1 - \theta) Y)\} = \text{rank}(\bar{X}).$$

The equality is due to the maximality of $\text{rank}(\bar{X})$ and the assumption that $\theta \neq 0$. Therefore $\text{rank}(X) = \text{rank}(\bar{X})$, implying the desired result. □

In fact, the relative interior elements of a convex set $C \subseteq \mathbb{S}^n_+$ not only have the same rank, but even have the same range and null space.

Lemma 2.3.3. *Let $C \subseteq \mathbb{S}^n_+$ be convex and non-empty. Then,*

(i) $X \in \text{relint}(C),\ Y \in C \implies \text{range}(X) \supseteq \text{range}(Y)$,

(ii) $X \in \text{relint}(C),\ Y \in \text{relint}(C) \implies \text{range}(X) = \text{range}(Y)$.

Proof. First, note that (ii) is implied by (i) and Lemma 2.3.2. To prove (i) let X and Y be as in the hypothesis. By convexity, $\frac{1}{2}(X+Y) \in C$ and by Fact 2.1.1 (iv),

$$\text{range}\left(\frac{1}{2}(X+Y)\right) = \text{range}(X+Y) = \text{range}(X) + \text{range}(Y).$$

Now, by Lemma 2.3.2 it holds that $\text{rank}\left(\frac{1}{2}(X+Y)\right) \leq \text{rank}(X)$. Thus we may conclude that $\text{range}(Y) \subseteq \text{range}(X)$, as desired. □

This result motivates the definition of *range of a convex set* in $\mathbb{S}^n_+$.

Definition 2.3.4. *Let $C \subseteq \mathbb{S}^n_+$ be non-empty and convex and let $X \in \text{relint}(C)$. Then the range of C is defined as,*

$$\text{range}(C) := \text{range}(X).$$

2.4 The Faces of $\mathbb{S}^n_+$

In the previous section we saw that convex subsets of $\mathbb{S}^n_+$ exhibit some remarkable properties with respect to rank and range. In this section we focus on special types of convex subsets of $\mathbb{S}^n_+$, referred to as *faces*. Recall that $\mathcal{K} \subseteq \mathbb{S}^n$ is a *convex cone* if $\mathcal{K} + \mathcal{K} = \mathcal{K}$, where '+' denotes the usual Minkowski sum.

Definition 2.4.1. *A convex cone $f \subseteq \mathbb{S}^n_+$ is a* face*, denoted $f \trianglelefteq \mathbb{S}^n_+$, if the following implication holds:*

$$X, Y \in \mathbb{S}^n_+,\ X + Y \in f \implies X, Y \in f.$$

From this definition it is easy to verify that $\mathbb{S}^n_+$ and $\{0\}$ are two faces of $\mathbb{S}^n_+$. We say that $f \trianglelefteq \mathbb{S}^n_+$ is a *proper* face of $\mathbb{S}^n_+$, denoted $f \triangleleft \mathbb{S}^n_+$, if f is not the empty set and $f \neq \mathbb{S}^n_+$.

In Lemmas 2.3.2 and 2.3.3 we saw that the relative interior elements of a convex set $C \in \mathbb{S}^n_+$ have the same rank and range. The faces of $\mathbb{S}^n_+$ exhibit the stronger property that the relative interior consists of *all* matrices that share a common range.

Fact 2.4.2. *Let f be a non-empty face of $\mathbb{S}^n_+$ and let* range(f) *be as in Definition 2.3.4. Then,*

(*i*) $f = \{X \in \mathbb{S}^n_+ : \text{range}(X) \subseteq \text{range}(f)\}$,

(*ii*) $\text{relint}(f) = \{X \in \mathbb{S}^n_+ : \text{range}(X) = \text{range}(f)\}$.

One implication of this characterization of the faces of $\mathbb{S}^n_+$ is that there exists a bijection between subspaces of $\mathbb{R}^n$ and faces of $\mathbb{S}^n_+$. To gain further insight into the faces of $\mathbb{S}^n_+$ consider the following characterization, that is readily obtained from Fact 2.4.2.

Fact 2.4.3. *Let f be a non-empty face of $\mathbb{S}^n_+$, with $f \neq \{0\}$, and let $r := \text{rank}(f)$. Let $V \in \mathbb{R}^{n\times r}$ be such that its columns form a basis for* range(f) *and let $U \in \mathbb{R}^{n\times n-r}$ be chosen so that its columns complete the columns of V to a basis of $\mathbb{R}^n$. Then,*

(*i*) $f = V\mathbb{S}^r_+V^T = \mathbb{S}^n_+ \cap \left(UU^T\right)^\perp$,

(*ii*) $\text{relint}(f) = V\mathbb{S}^r_{++}V^T$.

There are two important implications of this characterization. First, every face is isomorphic to a smaller dimensional positive semidefinite cone. This is an idea that is central to *facial reduction* and will be expanded upon in Chapter 3. Secondly, every face of $\mathbb{S}^n_+$ can be expressed as the intersection of $\mathbb{S}^n_+$ and a hyperplane. Such faces are referred to as *exposed*. In this context, the matrix UU^T in Fact 2.4.3 is called an *exposing vector*. It is important to note that not all convex cones exhibit the property that every face is exposed.

An important notion related to exposing vectors is that of the conjugate face.

Definition 2.4.4. *Let $f \trianglelefteq \mathbb{S}^n_+$ be non-empty. The* conjugate face *of f is,*

$$f^c := f^\perp \cap \mathbb{S}^n_+.$$

We record several important properties of the conjugate face in the following.

Fact 2.4.5. *Let $f \trianglelefteq \mathbb{S}^n_+$ be non-empty, with $f \neq \{0\}$, and let $r := \text{rank}(f)$. Let V and U be the matrices of Fact 2.4.3 and let f^c denote the conjugate face as in Definition 2.4.4. Then,*

(*i*) $f^c \trianglelefteq \mathbb{S}^n_+$,

(*ii*) $f^c = U\mathbb{S}^{n-r}_+U^T$,

(*iii*) $W \in \text{relint}(f^c) \implies f = \mathbb{S}^n_+ \cap W^\perp$,

(*iv*) $X \in \text{relint}(f) \implies f^c = \mathbb{S}^n_+ \cap X^\perp$.

The property of 'face' is preserved under set intersection.

Fact 2.4.6. *If $f_1, \ldots, f_m$ are faces of $\mathbb{S}^n_+$, then $f_1 \cap \cdots \cap f_m \trianglelefteq \mathbb{S}^n_+$.*

We conclude this preliminary discussion of the faces of $\mathbb{S}^n_+$ with an expression for $\mathbb{S}^n_+$ and $\mathrm{bd}(\mathbb{S}^n_+)$. Here $\dot{\bigcup}$ denotes a disjoint union.

Lemma 2.4.7. *It holds that,*

(*i*) $\mathbb{S}^n_+ = \dot{\bigcup}_{f \trianglelefteq \mathbb{S}^n_+} \mathrm{relint}(f)$,

(*ii*) $\mathrm{bd}(\mathbb{S}^n_+) = \dot{\bigcup}_{f \triangleleft \mathbb{S}^n_+} \mathrm{relint}(f)$.

Proof. For (*i*), it is clear that $\mathbb{S}^n_+ \supseteq \dot{\bigcup}_{f \trianglelefteq \mathbb{S}^n_+} \mathrm{relint}(f)$. For the reverse inclusion, let $X \in \mathbb{S}^n_+$. Then X belongs to the relative interior of the face f satisfying,

$$\mathrm{range}(f) = \mathrm{range}(X),$$

as desired. All that remains is to show that the union of the relative interiors of the faces of $\mathbb{S}^n_+$ is disjoint. But this is a direct implication of Fact 2.4.2 (*ii*), as the relative interior of each face is completely characterized by its range.

Now for (*ii*), the only face that is not included in $\dot{\bigcup}_{f \triangleleft \mathbb{S}^n_+} \mathrm{relint}(f)$ is the set $\mathbb{S}^n_+$ itself, with relative interior $\mathbb{S}^n_{++}$. Thus from (*i*) and the definition of boundary we have,

$$\dot{\bigcup}_{f \triangleleft \mathbb{S}^n_+} \mathrm{relint}(f) = \mathbb{S}^n_+ \setminus \mathbb{S}^n_{++} = \mathrm{bd}(\mathbb{S}^n_+),$$

as desired. □

2.5 Convex Sets and the Minimal Face

The expression of $\mathbb{S}^n_+$ as a disjoint union in Lemma 2.4.7 allows us to, loosely speaking, 'place' every convex set in exactly one of the partitions.

Lemma 2.5.1. *Let $C \subseteq \mathbb{S}^n_+$ be convex and non-empty. Let f be the face of $\mathbb{S}^n_+$ satisfying* $\mathrm{range}(f) = \mathrm{range}(C)$. *Then f is the unique face such that,*

$$\mathrm{relint}(C) \subseteq \mathrm{relint}(f).$$

Proof. The desired result is a direct implication of the definition of the range of a set in Definition 2.3.4 and Fact 2.4.2 (*ii*). □

In fact, the face of Lemma 2.5.1 is the 'smallest' face that contains the convex set C. To be more specific about our use of the word 'smallest', we define the minimal face.

Definition 2.5.2. *The* minimal face *of a convex set $C \subseteq \mathbb{S}^n_+$ is the intersection of all faces containing C:*

$$\text{face}(C) = \bigcap_{\substack{f \trianglelefteq \mathbb{S}^n_+ \\ C \subseteq f}} f.$$

Lemma 2.5.3. *Let $C \subseteq \mathbb{S}^n_+$ be convex and non-empty. Let* face(C) *be the minimal face containing C, as in Definition 2.5.2, and let $f \trianglelefteq \mathbb{S}^n_+$ satisfy* range$(f) =$ range(C)*. Then,*

$$\text{face}(C) = f.$$

Proof. By Lemma 2.3.3 (i) we have $C \subseteq f$. Hence f is one of the faces in the intersection defining face(C) (see Definition 2.5.2). Moreover, we claim that for any other face f', for which $C \subseteq f'$, it holds that $f \subseteq f'$. Indeed, if $C \subseteq f'$ then Fact 2.4.2 (i) implies that,

$$\text{range}(C) \subseteq \text{range}(f').$$

By construction we have range$(f) =$ range(C) and thus range$(f) \subseteq$ range(f'). Applying Fact 2.4.2 (i) yields that $f \subseteq f'$. The desired result is now immediate. □

The result of Fact 2.4.6 is extended to the minimal face as follows.

Lemma 2.5.4. *Let $C_1, \ldots, C_m \subseteq \mathbb{S}^n_+$ be convex and non-empty. Then,*

$$\text{face}(C_1 \cap \cdots \cap C_m) \subseteq \text{face}(C_1) \cap \cdots \cap \text{face}(C_m).$$

Proof. The result certainly holds when $C_1 \cap \cdots \cap C_m$ is empty. For the remaining case, let $\bar{X} \in \text{face}(C_1 \cap \cdots \cap C_m)$. Then by the range characterization of the minimal face it holds that,

$$\text{range}(\bar{X}) = \text{range}(C_1 \cap \cdots \cap C_m) \subseteq \text{range}(C_1) \cap \cdots \cap \text{range}(C_m),$$

implying the desired result. □

Another useful result concerns linear maps of the form $M \cdot M^T$.

Lemma 2.5.5. *Let $C \subseteq \mathbb{S}^n_+$ be convex and let $M \in \mathbb{R}^{p \times n}$. Then,*

$$\text{face}\left(MCM^T\right) = M \,\text{face}(C) M^T.$$

Proof. By Theorem 6.6 of [71] it holds that,

$$\text{relint}\left(MCM^T\right) = M \,\text{relint}(C) M^T, \quad \text{relint}\left(M \,\text{face}(C) M^T\right) = M \,\text{relint}(\text{face}(C)) M^T.$$

Thus if $\bar{X} \in \text{relint}(C)$ we have,

$$\text{range}\left(MCM^T\right) = \text{range}\left(M \,\text{face}(C) M^T\right) = \text{range}\left(M\bar{X}M^T\right).$$

Since the two sets have the same range all that remains is to show that $M \,\text{face}(C) M^T$ is a face. Indeed, if $\text{face}(C) = V\mathbb{S}^r_+ V^T$ for some $V \in \mathbb{R}^{n \times r}$ then,

$$M \,\text{face}(C) M^T = MV\mathbb{S}^r_+ (MV)^T = \left\{Y \in \mathbb{S}^p_+ : \text{range}(Y) \subseteq \text{range}(MV)\right\},$$

a face by Fact 2.4.2. □

2.6 Semidefinite Programming

SDPs have become popular over the last several decades due, in part, to their modelling power in areas as diverse as combinatorial optimization, electric power systems, molecular conformation, distance geometry, sensor network localization, etc. An SDP is a linear optimization problem over a spectrahedron:

$$(SDP) \qquad \begin{aligned} p^\star := \inf\ & \langle C, X\rangle, \\ \text{s.t. } & \mathcal{A}(X) = b, \\ & X \succeq 0. \end{aligned} \tag{2.6.1}$$

Here $C \in \mathbb{S}^n$, $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$ is a linear map, and $b \in \mathbb{R}^m$. The feasible set of (SDP) is the spectrahedron $\mathcal{F}(\mathcal{A}, b)$ and the Lagrangian dual of (SDP) is,

$$(DSDP) \qquad \begin{aligned} d^\star := \sup\ & b^T y, \\ \text{s.t. } & \mathcal{A}^*(y) + Z = C, \\ & Z \succeq 0. \end{aligned} \tag{2.6.2}$$

We refer to the difference $p^\star - d^\star$ as the *duality gap*. By virtue of the Lagrangian dual, *weak duality* ensures that the duality gap is non-negative for every pair (SDP) and $(DSDP)$.

While SDPs may appear to be rather benign (a linear function is optimized over a relatively simple convex set), there exist instances with remarkably undesirable properties. For example, we can construct an SDP so that both (SDP) and $(DSDP)$ are feasible, but the duality gap is positive and neither $p^\star$ nor $d^\star$ is attained. Such instances are very challenging to solve algorithmically since it is hard to determine the quality of a proposed solution. To eliminate such classes of SDPs we place restrictions on the constraint sets of (SDP) and $(DSDP)$.

Definition 2.6.1. *We say that the Slater condition holds for (SDP) if there exists $X \succ 0$ with $\mathcal{A}(X) = b$. Equivalently, in the notation of spectrahedra, the Slater condition holds if,*

$$\mathcal{F}(\mathcal{A}, b) \cap \mathbb{S}^n_{++} \neq \emptyset.$$

Then X is referred to as a Slater point. Similarly, the Slater condition holds for $(DSDP)$ if there exists $y \in \mathbb{R}^m$ and $Z \succ 0$ such that $\mathcal{A}^*(y) + Z = C$.

We now state several *strong duality* results for SDPs.

Fact 2.6.2. *Let (SDP) and $(DSDP)$ be as in (2.6.1) and (2.6.2) and suppose that both (SDP) and $(DSDP)$ are feasible. Then the following hold.*

(i) If the Slater condition holds for (SDP), then $p^\star = d^\star$ and $d^\star$ is attained.

(ii) If the Slater condition holds for $(DSDP)$, then $p^\star = d^\star$ and $p^\star$ is attained.

(*iii*) *If the Slater condition holds for both* (SDP) *and* $(DSDP)$*, then* $p^\star = d^\star$ *and both* $p^\star$ *and* $d^\star$ *are attained.*

The following optimality conditions form the basis of many SDP algorithms.

Fact 2.6.3. *Let* (SDP) *and* $(DSDP)$ *be as in* (2.6.1) *and* (2.6.2)*. Suppose that* $p^\star$ *and* $d^\star$ *are equal and attained. Then* $X^\star$ *is optimal for* (SDP) *and* $(y^\star, Z^\star)$ *are optimal for* $(DSDP)$ *if, and only if,*

$$\begin{bmatrix} \mathcal{A}^*(y^\star) + Z^\star - C \\ \mathcal{A}(X^\star) - b \\ Z^\star X^\star \end{bmatrix} = 0, \;\; X^\star \succeq 0, \;\; Z^\star \succeq 0.$$

It is important to note that the classes of SDPs identified in Fact 2.6.2 are not the only ones for which a zero duality gap exists. There is also the class of unbounded-infeasible SDPs. Suppose, for instance, that (SDP) is unbounded from below. Then $p^\star = -\infty$ and weak duality implies that $d^\star = -\infty$. This corresponds to an infeasible instance of $(DSDP)$. A similar result may be obtained when $(DSDP)$ is unbounded from above. The challenge with such instances is that they may be hard to detect due to the phenomenon of *weak infeasibility*. In this work we focus predominantly on instances where both (SDP) and $(DSDP)$ are feasible.

2.7 Function Bounds and Asymptotic Notation

In developing error bounds we need notation that captures the asymptotic behaviour of real valued functions over the positive reals as 0 is approached. To this end, let ϕ and ψ be real valued functions on $(0, \bar{t}]$ for some $\bar{t} > 0$ where $\psi(t) > 0$ for all $t \in (0, \bar{t}]$. We say that ϕ *is big-O of* ψ, denoted,

$$\phi(t) = \mathcal{O}(\psi(t)),$$

if there exists $M > 0$ such that $\phi(t) \leq M\psi(t)$ for all $t \in (0, \bar{t}]$. Up to a constant multiple, ψ is an upper bound on ϕ. A lower bound for ϕ is expressed in a similar way. We say that ϕ *is omega of* ψ, denoted,

$$\phi(t) = \Omega(\psi(t)),$$

if there exists $M > 0$ such that $\phi(t) \geq M\psi(t)$ for all $t \in (0, \bar{t}]$. When ϕ is both big-O of ψ and omega of ψ we say that ϕ *is theta of* ψ denoted,

$$\phi(t) = \Theta(\psi(t)).$$

In some instances we may avoid specifying $\bar{t}$ by considering ϕ and ψ on $\mathbb{R}_{++}$. In such cases we say that ϕ is big-O of ψ as $t \searrow 0$ to mean that there exists $\bar{t} > 0$ such that ϕ restricted to $(0, \bar{t}]$ is big-O of ψ restricted to $(0, \bar{t}]$. Analogous notation is used with Ω and Θ.

Chapter 3

Facial Reduction in Semidefinite Programming

Generically, SDPs possess very nice properties such as, the Slater condition for the primal and the dual, unique optimal solutions, and *strict complementarity*. See for instance [19, 62, 64]. However, it is not difficult to construct instances where these properties fail, e.g., Section 2.4 of [80]. We have already seen that absence of the Slater condition in SDPs can have undesirable implications, such as a positive duality gap and lack of optimal solutions. One way to remedy such SDPs is by *facial reduction*. The term refers both to a technique and an algorithm for converting an SDP without the Slater condition into an equivalent one for which the Slater condition does hold. Both the technique and an algorithm were introduced by Borwein and Wolkowicz in the papers [6, 7, 8] for the abstract convex optimization problem.

3.1 The Facial Reduction Technique

To introduce the facial reduction technique let us restate the primal SDP problem,

$$(SDP) \qquad \begin{array}{l} \inf \langle C, X \rangle, \\ \text{s.t. } \mathcal{A}(X) = b, \\ \quad X \succeq 0. \end{array}$$

Let us assume that the convex feasible set, $\mathcal{F} := \mathcal{F}(\mathcal{A}, b)$, is non-empty. The Slater condition fails for (SDP) when $\mathcal{F}$ does not contain a positive definite matrix. In this case $\mathcal{F}$ is contained in the boundary of $\mathbb{S}^n_+$. Now Lemma 2.4.7 *(ii)* shows that the boundary of $\mathbb{S}^n_+$ is the disjoint union of the relative interiors of the proper faces of $\mathbb{S}^n_+$. Moreover, Lemma 2.5.1 states that our convex set $\mathcal{F}$ can be 'placed' into exactly one of these partitions of $\text{bd}(\mathbb{S}^n_+)$. This face is identified in Lemma 2.5.3 as the minimal face, $\text{face}(\mathcal{F})$. With this in mind, let

us replace the positive semidefinite constraint in (SDP) with the constraint $X \in \text{face}(\mathcal{F})$,

$$\begin{aligned} &\inf \langle C, X \rangle, \\ &\text{s.t. } \mathcal{A}(X) = b, \\ &\quad X \in \text{face}(\mathcal{F}). \end{aligned} \tag{3.1.1}$$

By the arguments above, (3.1.1) is equivalent to (SDP), since the objective functions and feasible sets are the same. Moreover, $\text{face}(\mathcal{F})$ is isomorphic to a smaller dimensional positive semidefinite cone. Hence the new optimization problem is in fact an instance of SDP. To see this, we utilize results established in Chapter 2.

Suppose, as above, that $\mathcal{F}$ is non-empty, with $\mathcal{F} \neq \{0\}$, and define $r := \text{rank}(\mathcal{F})$. Since $\mathcal{F}$ is convex, we may define $V \in \mathbb{R}^{n \times r}$ such that its columns form a basis for $\text{range}(\mathcal{F})$. By Lemma 2.5.3 and Fact 2.4.3 we have that $\text{face}(\mathcal{F}) = V\mathbb{S}^r_+V^T$. Now observe the following transformation of the problem obtained in (3.1.1):

$$\inf \{\langle C, X \rangle : \mathcal{A}(X) = b,\ X \in \text{face}(\mathcal{F})\} \tag{3.1.2}$$

$$\downarrow$$

$$\inf \{\langle C, X \rangle : \mathcal{A}(X) = b,\ X \in V\mathbb{S}^r_+V^T\} \tag{3.1.3}$$

$$\downarrow$$

$$\inf \{\langle C, X \rangle : \mathcal{A}(X) = b,\ X = VRV^T,\ R \in \mathbb{S}^r_+\} \tag{3.1.4}$$

$$\downarrow$$

$$\inf \{\langle C, VRV^T \rangle : \mathcal{A}(VRV^T) = b,\ R \in \mathbb{S}^r_+\} \tag{3.1.5}$$

$$\downarrow$$

$$(RSDP) \qquad \inf \{\langle V^TCV, R \rangle : \mathcal{A}_V(R) = b,\ R \in \mathbb{S}^r_+\}, \tag{3.1.6}$$

where $\mathcal{A}_V$ is defined as in (2.2.2). Every step in the transformation has the same feasible set (up to isomorphism) and the same objective. Therefore, these optimization problems are all equivalent to (SDP). Moreover, the last optimization problem, $(RSDP)$, satisfies the Slater condition. While we motivated the derivation of $(RSDP)$ by assuming that $\mathcal{F}$ does not have a Slater point, this need not be assumed. In fact, when the Slater condition holds for (SDP) the transformation to $(RSDP)$ is trivial since $\text{face}(\mathcal{F}) = \mathbb{S}^n_+$ and V may be taken to be I.

The other special case is $\mathcal{F} = \{0\}$. Here $r = 0$ and, therefore, the matrix $V \in \mathbb{R}^{n \times r}$ is not well defined. Consequently, facial reduction, in the form above, is not well defined. However, this case is somewhat trivial, since knowledge of the minimal face immediately yields the optimal set.

Lemma 3.1.1. *Let (SDP) and $(RSDP)$ be as in (2.6.1) and (3.1.6), respectively, and suppose that the feasible set, $\mathcal{F}(\mathcal{A}, b)$, of (SDP) is non-empty. Then the Slater condition holds for $(RSDP)$.*

Proof. Let $\mathcal{F}(\mathcal{A}, b)$ be as in the hypothesis and let r and V be as above so that,

$$\text{face}(\mathcal{F}) = V\mathbb{S}^r_+V^T. \tag{3.1.7}$$

By Lemma 2.5.3 and Lemma 2.5.1 there exists $\bar{X} \in \mathcal{F}(\mathcal{A}, b)$ such that $\text{rank}(\bar{X}) = r$. Then by (3.1.7), there exists $\bar{R} \in \mathbb{S}^r_{++}$ such that,

$$\bar{X} = V\bar{R}V^T.$$

It follows that $\bar{R}$ is a Slater point for $(RSDP)$, as desired. □

Observe that $(RSDP)$ is an instance of SDP in the primal form where C is replaced by V^TCV and $\mathcal{A}$ is replaced by $\mathcal{A}_V$. Therefore the dual of $(RSDP)$ may be obtained as $(DSDP)$ by replacing C and $\mathcal{A}$ to get,

$$(DRSDP) \qquad \begin{aligned} &\sup\ b^Ty, \\ &\text{s.t. } (\mathcal{A}_V)^*(y) + Z = V^TCV, \\ &\quad Z \succeq 0. \end{aligned} \tag{3.1.8}$$

We now obtain the following strong duality results for the pair $(RSDP)$ and $(DRSDP)$.

Theorem 3.1.2. *Let (SDP) and $(DSDP)$ be as in (2.6.1) and (2.6.2), respectively, and assume that both are feasible. Then the following hold.*

(i) The optimal values of $(RSDP)$ and $(DRSDP)$ are equal and finite and the optimal value of $(DRSDP)$ is attained.

(ii) If $(DSDP)$ satisfies the Slater condition, then $(DRSDP)$ satisfies the Slater condition and the optimal value of $(RSDP)$ is attained.

Proof. By the hypothesis and Lemma 3.1.1, the Slater condition holds for $(RSDP)$. Then *(i)* follows from Fact 2.6.2 *(i)*. For *(ii)*, the hypothesis implies the existence of $\bar{y} \in \mathbb{R}^m$ and $\bar{Z} \succ 0$ such that,

$$\mathcal{A}^*(\bar{y}) + \bar{Z} = C.$$

Let $V \in \mathbb{R}^{n\times r}$ be as in $(DRSDP)$. Then,

$$V^T\bar{Z}V = V^TCV - V^T\mathcal{A}^*(\bar{y})V = V^TCV - (\mathcal{A}_V)^*(\bar{y}).$$

Since $\bar{Z} \succ 0$ and V is full rank, it follows that $V^T\bar{Z}V \succ 0$. Therefore, the pair $\bar{y}$ and $V^T\bar{Z}V$ is a Slater point for $(DRSDP)$. The desired result now follows from Fact 2.6.2 *(ii)*. □

Aside from the desirable consequences of the Slater condition, there is an additional benefit to facial reduction. When the Slater condition does not hold for (SDP), the new problem $(RSDP)$ lives in a smaller space than the original SDP. In applications such as sensor network localization, e.g., [41], facial reduction leads to a substantial decrease in the dimension of the ambient space. Since current SDP algorithms are limited to problems of only several thousand variables, this reduction can lead to non-trivial improvements in computation.

3.2 Partial Facial Reduction

Suppose that our feasible set $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ does not satisfy the Slater condition. Furthermore, suppose that we do not know the matrix $V \in \mathbb{R}^{n\times r}$ that characterizes face$(\mathcal{F})$, as in Section 3.1, but we do have access to a matrix $U \in \mathbb{R}^{n\times p}$ with $p > r$ and,

$$\text{range}(V) \subset \text{range}(U).$$

Then we may perform *partial facial reduction*. By arguments analogous to those of Section 3.1, the original problem (SDP) is equivalent to,

$$\inf \{\langle U^T C U, P\rangle : \mathcal{A}_U(P) = b,\ P \in \mathbb{S}^p_+\}, \tag{3.2.1}$$

where $\mathcal{A}_U$ is defined as in (2.2.2). While the Slater condition does not hold for (3.2.1), it may still exhibit nicer properties than the original SDP. For one, the dimension of (3.2.1) could be substantially smaller. This property has been successfully exploited in applications such as Euclidean distance matrix completion. See [73], for instance. Secondly, (3.2.1) may behave better than the original SDP. We mean to say that not all SDPs for which the Slater condition fails are equally ill-behaved. This is a notion we will explore further when we introduce *singularity degree* in Chapter 4.

The following result gives an expression for the minimal face of the partially reduced spectrahedron.

Lemma 3.2.1. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a non-empty spectrahedron. Let $r :=$ rank$(\mathcal{F})$ and let $V \in \mathbb{R}^{n\times r}$ be such that its columns form a basis for* range$(\mathcal{F})$. *Let $p > r$ and let $U \in \mathbb{R}^{n\times p}$ be a full rank matrix satisfying* range$(V) \subset$ range(U). *Then,*

$$\text{face}\left(\mathcal{F}(\mathcal{A}_U, b)\right) = M\mathbb{S}^r_+ M^T,$$

where $M \in \mathbb{R}^{p\times r}$ is full rank and satisfies $UM = V$.

Proof. First of all, M is well defined since each column of V is spanned by the columns of U as implied by the assumption that range$(V) \subset$ range(U). Now we show that,

$$\mathcal{F}(\mathcal{A}, b) = U\mathcal{F}\left(\mathcal{A}_U, b\right)U^T. \tag{3.2.2}$$

The inclusion '$\supseteq$' is trivial. For the converse inclusion, let $X \in \mathcal{F}(\mathcal{A}, b)$. Then by the definition of V it holds that,

$$X = VRV^T,\ R \succeq 0,\ \mathcal{A}(VRV^T) = b.$$

Replacing V with UM yields,

$$X = UMRM^TU^T,\ R \succeq 0,\ \mathcal{A}(UMRM^TU^T) = b.$$

Now we replace MRM^T with P to get,

$$X = UPU^T,\ P \succeq 0,\ \mathcal{A}(UPU^T) = b.$$

The '$\subseteq$' inclusion is now immediate giving us (3.2.2). The assumption that U is full rank also gives us,

$$\operatorname{relint}\left(\mathcal{F}(\mathcal{A}, b)\right) = U \operatorname{relint}\left(\mathcal{F}\left(\mathcal{A}_U, b\right)\right) U^T. \tag{3.2.3}$$

By (3.2.3) for $P \in \operatorname{relint}\left(\mathcal{F}\left(\mathcal{A}(U \cdot U^T), b\right)\right)$ we have,

$$\operatorname{range}(UPU^T) = \operatorname{range}(V) = \operatorname{range}(UM) = U \operatorname{range}(M).$$

On the other hand,

$$\operatorname{range}(UPU^T) = \operatorname{range}(UP) = U \operatorname{range}(P).$$

Combining these two observations with the assumption that U is full rank, gives us,

$$\operatorname{range}(P) = \operatorname{range}(M).$$

The desired result is now implied by Lemma 2.5.3. □

3.3 Facial Reduction as an Algorithm

The key ingredient to facial reduction, from a practical perspective, is an algebraic expression for face($\mathcal{F}$). Specifically, the matrix V with columns forming a basis for range($\mathcal{F}$). In special cases, V may be obtained analytically by taking advantage of underlying structure of the SDP. When such information is not available or is difficult to analyze, Borwein and Wolkowicz introduced a facial reduction algorithm. The output of the algorithm is exactly the matrix $V \in \mathbb{R}^{n \times r}$ and consequently the face $V\mathbb{S}^r_+V^T$. This expression for the minimal face is the terminus of a finite sequence of successively smaller faces. Specifically, for some positive integer d, the facial reduction algorithm generates a sequence of faces $f^1, \cdots, f^d$ satisfying,

$$f^1 \supsetneq \cdots \supsetneq f^d, \quad f^d = \operatorname{face}(\mathcal{F}).$$

Equivalently, from the primal characterization of Fact 2.4.3, the algorithm generates a sequence of positive integers $r_1, r_2, \ldots, r_d$ and matrices $V^i \in \mathbb{R}^{n \times r_i}$ for each $i \in \{1, \ldots, d\}$ such that,

$$\begin{aligned} &r_1 > \cdots > r_d, \\ &\operatorname{range}(V^1) \supsetneq \cdots \supsetneq \operatorname{range}(V^d), \\ &f^i = V^i\mathbb{S}^{r_i}_+\left(V^i\right)^T,\ \forall i \in \{1, \ldots, d\}. \end{aligned} \tag{3.3.1}$$

In terms of exposing vectors, also from Fact 2.4.3, the facial reduction algorithm generates a sequence of exposing vectors, $W^1, \ldots, W^d \in \mathbb{S}^n$, such that,

$$\begin{aligned} &W^i \in \mathbb{S}^n_+ \setminus \{0\}, \ \forall i \in \{1, \ldots, d\}, \\ &\text{range}(W^1) \subsetneq \cdots \subsetneq \text{range}(W^d), \\ &f^i = \left(W^i\right)^\perp \cap \mathbb{S}^n_+, \ \forall i \in \{1, \ldots, d\}. \end{aligned} \tag{3.3.2}$$

Clearly such a sequence of faces exists. For instance we may choose $d = 1$ and $f^1 = \text{face}(\mathcal{F})$. Moreover, if $\text{face}(\mathcal{F})$ is proper, the sequence is not unique. The sequences generated by the facial reduction algorithm of Borwein and Wolkowicz rely on a specific class of exposing vectors: those contained in $\text{range}(\mathcal{A}^*)$. We denote this set by $\mathcal{E}(\mathcal{A}, b)$,

$$\mathcal{E}(\mathcal{A}, b) := (\text{face}(\mathcal{F})^c \cap \text{range}(\mathcal{A}^*)) \setminus \{0\}. \tag{3.3.3}$$

Fact 3.3.1. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a non-empty spectrahedron and let $\mathcal{E}(\mathcal{A}, b)$ be as in (3.3.3). Then exactly one of the following holds:*

(*i*) $\mathcal{F} \cap \mathbb{S}^n_{++} \neq \emptyset$, *i.e., the Slater condition holds for* $\mathcal{F}$,

(*ii*) $\mathcal{E}(\mathcal{A}, b) \neq \emptyset$, *i.e., there exists* $y \in \mathbb{R}^m$ *such that* $0 \neq \mathcal{A}^*(y) \succeq 0$ *and* $y^T b = 0$.

A proof of this theorem of the alternative may be found in [18], for instance. If (*i*) holds, then $\mathcal{F}$ has a Slater point and we have obtained the facially reduced problem $(RSDP)$, trivially. On the other hand if (*ii*) holds, then there exists $y \in \mathbb{R}^m$ such that $\mathcal{A}^*(y) \in \mathcal{E}(\mathcal{A}, b)$. By definition of $\mathcal{E}(\mathcal{A}, b)$ it follows that $\mathcal{A}^*(y)$ is an exposing vector for a face containing $\text{face}(\mathcal{F})$. In other words,

$$\text{face}(\mathcal{F}) \subseteq \mathbb{S}^n_+ \cap (\mathcal{A}^*(y))^\perp .$$

If this inclusion is an equality, then we have obtained $(RSDP)$ and we are done. Otherwise we may perform partial facial reduction and then apply Fact 3.3.1 to the new reduced problem. The algorithm of Borwein and Wolkowicz continues in this fashion until for some reduced problem Fact 3.3.1 (*i*) does hold. At each step, the dimension of the reduced problem decreases by at least 1 and therefore at most n steps are required. In fact, the tighter bound of $n - 1$ is shown to hold in Section 4.2. For a rigorous description of the facial reduction algorithm refer to Algorithm 1.

3.4 Correctness of the Facial Reduction Algorithm

In this section we prove that Algorithm 1 outputs an expression for $\text{face}(\mathcal{F}(\mathcal{A}, b))$ when the spectrahedron $\mathcal{F}(\mathcal{A}, b)$ is not empty. We include a proof, partly for completeness, and partly to derive results that will be used throughout the thesis.

Algorithm 1 Facial Reduction for the Spectrahedron $\mathcal{F}(\mathcal{A}, b)$

INPUT: Linear map $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$ and $b \in \mathbb{R}^m$.
initialize: $k = 0$, $\mathcal{A}^k = \mathcal{A}$, $V^k = I$, $W^k = 0$, $r_k = n$, $q_k = 0$.
while $\mathcal{E}(\mathcal{A}^k, b) \neq \emptyset$ **do**
  Choose $Z^{k+1} = \left(\mathcal{A}^k\right)^* (y^{k+1}) \in \mathcal{E}(\mathcal{A}^k, b)$ and orthogonal $\begin{bmatrix} Q_1^{k+1} & Q_2^{k+1} \end{bmatrix}$ such that,

$$Z^{k+1} = \begin{bmatrix} Q_1^{k+1} & Q_2^{k+1} \end{bmatrix} \begin{bmatrix} \Lambda^{k+1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Q_1^{k+1} & Q_2^{k+1} \end{bmatrix},$$

  where $\Lambda^{k+1} \succ 0$ and $Q_1^{k+1} \in \mathbb{R}^{r_k \times q_{k+1}}$.
  if $Z^{k+1} \succ 0$ **then**
    $Q_2^{k+1} = 0 \in \mathbb{S}^{r_k}$, $V^{k+1} = 0 \in \mathbb{S}^n$, and $r_{k+1} = 0$
  else
    $Q_2^{k+1} \in \mathbb{R}^{r_k \times r_{k+1}}$ and $V^{k+1} = V^k Q_2^{k+1} \in \mathbb{R}^{n \times r_{k+1}}$
  end if
  $W^{k+1} = W^k + V^k Z^{k+1} \left(V^k\right)^T \in \mathbb{S}^n_+$
  $\mathcal{A}^{k+1} = \mathcal{A}_{V^{k+1}}$
  $k = k + 1$
end while
OUTPUT: $d = k$, $V = V^k$, $W = W^k$, $r = r_d$.

Lemma 3.4.1. *Let $\mathcal{F}(\mathcal{A}, b)$ be a non-empty spectrahedron. Then each step of the while loop of Algorithm 1 is well defined (but not uniquely) whenever the while loop is called.*

Proof. Suppose the while loop is called for some $k \geq 0$. To show that the steps of the while loop are well defined, it suffices to show that Z^{k+1}, Q_1^{k+1}, and Q_2^{k+1} exist. Since the while loop is called it holds that $\mathcal{E}(\mathcal{A}^k, b)$ is not empty. Thus Z^{k+1} exists and has rank at least 1. Consequently Q_1^{k+1} exists. The existence of Q_2^{k+1} follows immediately. □

Lemma 3.4.2. *Let $\mathcal{F}(\mathcal{A}, b)$ be a non-empty spectrahedron. If there exists k such that $Z^k \succ 0$, then define $\bar{k}$ to be the smallest such integer. Otherwise $\bar{k} = +\infty$. Then for every $k < \bar{k}$,*

(i) the columns of V^k form a basis for $\text{null}(W^k)$

(ii) $\mathcal{F}(\mathcal{A}, b) = V^k \mathcal{F}(\mathcal{A}^k, b) \left(V^k\right)^T$.

Proof. Both claims hold for $k = 0$ and we proceed by induction. Suppose the claims hold for some $k < \bar{k} - 1$ so that $k + 1 < \bar{k}$.

For (i), note that since $k < \bar{k}$ we have $V^{k+1} = IQ_2^1 \cdots Q_2^{k+1}$. Thus V^{k+1} has rank r_{k+1} and $(V^{k+1})^T V^{k+1} = I$. Clearly, these observations hold for smaller values of k as well. Then,

$$W^{k+1} V^{k+1} = \left(W^k + V^k Z^{k+1} \left(V^k\right)^T\right) V^k Q_2^{k+1} = V^k Z^{k+1} Q_2^{k+1} = 0,$$

where the first equality is due to the inductive hypothesis and the conclusion follows from the fact that Q_2^{k+1} is chosen so that its columns form a basis for $\text{null}(Z^{k+1})$. We have shown that the columns of V^{k+1} are elements of $\text{null}(W^{k+1})$. To show that they form a basis for the null space, we show that

$$\text{rank}(W^{k+1}) = n - \text{rank}(V^{k+1}) = n - r_{k+1}.$$

By induction we conclude that W^k is orthogonal to $V^k Z^{k+1} \left(V^k\right)^T$ thus,

$$\begin{aligned}\text{rank}(W^{k+1}) &= \text{rank}(W^k) + \text{rank}\left(V^k Z^{k+1} \left(V^k\right)^T\right) \\ &= n - \text{rank}(V^k) + \text{rank}(Q_1^{k+1}) \\ &= n - r_k + q_{k+1} \\ &= n - r_{k+1},\end{aligned}$$

as desired.

For *(ii)* we have that Z^{k+1} is an exposing vector for a face containing $\text{face}(\mathcal{F}(\mathcal{A}^k, b))$ and thus,

$$\mathcal{F}(\mathcal{A}^k, b) \subset \left(Z^{k+1}\right)^\perp \cap \mathbb{S}_+^{r^k} = Q_2^{k+1} \mathbb{S}_+^{r^{k+1}} \left(Q_2^{k+1}\right)^T.$$

Then invoking the inductive hypothesis we have,

$$\begin{aligned}\mathcal{F}(\mathcal{A}, b) &= V^k \mathcal{F}(\mathcal{A}^k, b) \left(V^k\right)^T \\ &= V^k \left\{R \succeq 0 : \mathcal{A}^k(R) = b,\ R = Q_2^{k+1} P \left(Q_2^{k+1}\right)^T\right\} \left(V^k\right)^T \\ &= V^k \left\{Q_2^{k+1} P \left(Q_2^{k+1}\right)^T \succeq 0 : \mathcal{A}_{V^k Q_2^{k+1}}(P) = b\right\} \left(V^k\right)^T \\ &= V^{k+1} \mathcal{F}(\mathcal{A}^{k+1}, b) \left(V^{k+1}\right)^T.\end{aligned}$$

In the last line we used the assumption that $k < \bar{k}$ and therefore $V^{k+1} = V^k Q_2^{k+1}$ □

Theorem 3.4.3. *Let $\mathcal{F}(\mathcal{A}, b)$ be a non-empty spectrahedron. Then Algorithm 1, with input $\mathcal{A}, b$, terminates finitely and the outputs d, V, W, and r are well defined (but may depend on the choice of Z^{k+1}). Moreover, $d \leq n$ and*

$$\text{face}(\mathcal{F}(\mathcal{A}, b)) = V \mathbb{S}_+^r V^T = W^\perp \cap \mathbb{S}_+^n.$$

Proof. If the while loop is never called, then $\mathcal{F}(\mathcal{A}, b)$ satisfies the Slater condition and we have $\text{face}(\mathcal{F}(\mathcal{A}, b)) = \mathbb{S}_+^n$. The algorithm terminates with $d = 0$, $V = I$, $W = 0$ and $r = n$, yielding the desired result.

Now suppose the while loop is called at least once. By Lemma 3.4.1 each iteration of the while loop is well defined. Let $\bar{k}$ be as in Lemma 3.4.2 and suppose $\bar{k} < +\infty$. By definition, $Z^{\bar{k}} \succ 0$ is an exposing vector for $\text{face}(\mathcal{F}(\mathcal{A}^{\bar{k}-1}, b))$. Therefore $\mathcal{F}(\mathcal{A}^{\bar{k}-1}, b) = \{0\}$. Then by Lemma 3.4.2 *(ii)* we get that $\text{face}(\mathcal{F}(\mathcal{A}, b)) = \{0\}$ and therefore $\mathcal{F}(\mathcal{A}, b) = \{0\}$.

Moreover, $V^{\bar{k}} = 0 \in \mathbb{S}^n$ and $r_{\bar{k}} = 0$. Then $\mathcal{F}(\mathcal{A}^{\bar{k}}, 0) = \mathbb{S}^n_+$, satisfying the Slater condition, and implying that $\mathcal{E}(\mathcal{A}^{\bar{k}}, b) = \emptyset$. Thus the algorithm terminates with $d = \bar{k}$, $V = 0$, and $r = 0$. Since $r_k < r_{k-1}$ for each $k < \bar{k}$. we conclude that $d \leq n$. Moreover we have,

$$\text{face}(\mathcal{F}(\mathcal{A}, b)) = V\mathbb{S}^r_+V^T = \{0\}.$$

All that remains is to show that $W = W^{\bar{k}}$ is positive definite. By construction,

$$W = W^{\bar{k}-1} + V^{\bar{k}-1}Z^{\bar{k}}\left(V^{\bar{k}-1}\right)^T,$$

and by Lemma 3.4.2 *(i)* we have $W^{\bar{k}-1}V^{\bar{k}-1} = 0$ and,

$$\text{rank}\left(W^{\bar{k}-1}\right) + \text{rank}\left(V^{\bar{k}-1}\right) = n. \tag{3.4.1}$$

Thus the matrices $W^{\bar{k}-1}$ and $V^{\bar{k}-1}Z^{\bar{k}}\left(V^{\bar{k}-1}\right)^T$ are positive semidefinite and orthogonal. Moreover, since $Z^{\bar{k}} \succ 0$ it holds that,

$$\text{rank}\left(V^{\bar{k}-1}\right) = \text{rank}\left(V^{\bar{k}-1}Z^{\bar{k}}\left(V^{\bar{k}-1}\right)^T\right). \tag{3.4.2}$$

Combining (3.4.1) with (3.4.2) gives us,

$$\text{rank}(W) = \text{rank}\left(W^{\bar{k}-1}\right) + \text{rank}\left(V^{\bar{k}-1}Z^{\bar{k}}\left(V^{\bar{k}-1}\right)^T\right) = n.$$

Hence $W \succ 0$ implying the desired result.

Now we may assume that $\bar{k} = +\infty$. Then for each k generated by the algorithm, the matrix Z^k is a non-zero, rank deficient matrix. It follows that $r^k < r^{k-1}$ and thus the sequence of integers $r^0, r^1, \ldots$ is strictly decreasing and bounded above by n. Moreover the sequence is bounded below by 1. Otherwise we have $Z^k \succ 0$ for some k, contradicting the assumption that $\bar{k} = +\infty$. Thus Algorithm 1 terminates with $d \leq n$. By Lemma 3.4.2 *(ii)* we have $\mathcal{F}(\mathcal{A}, b) = V\mathcal{F}(\mathcal{A}^d, b)V^T$. Since the Slater condition holds for $\mathcal{F}(\mathcal{A}^d, b)$ we conclude that $\text{range}(\mathcal{F}(\mathcal{A}, b)) = \text{range}(V)$ and thus by Lemma 2.5.3 it holds that $\text{face}(\mathcal{F}(\mathcal{A}, b)) = V\mathbb{S}^r_+V^T$. The second of the desired equalities now follows from Lemma 3.4.2 *(i)*. □

3.5 Some Properties of the Facial Reduction Algorithm

In Theorem 3.4.3, we obtained an upper bound of n on the number of iterations required by Algorithm 1. In this section we derive results that imply a bound in terms of the number of matrices, m, defining the map $\mathcal{A}$. First we show that the iterates of the algorithm are linearly independent. To better facilitate the discussion we introduce a definition.

Definition 3.5.1. *We denote the vectors $y^1, \dots, y^d$, generated by Algorithm 1 with input $(\mathcal{A}, b)$, as a facial reduction sequence.*

Lemma 3.5.2. *Let $d \geq 1$ and let $y^1, \dots, y^d$ be a facial reduction sequence for input $(\mathcal{A}, b)$. Then the following hold.*

(i) The matrices $\mathcal{A}^*(y^1), \dots, \mathcal{A}^*(y^d)$ are linearly independent.

(ii) If $\mathcal{A}$ is surjective then $y^1, \dots, y^d$ are linearly independent.

Proof. Both statements hold vacuously when $d = 1$, since $\mathcal{A}^*(y^1) \neq 0$. Thus we may assume that $d \geq 2$. For *(i)* let $\hat{k} \in [2, d]$ be an integer and observe that by construction,

$$\left(V^{\hat{k}-1}\right)^T \mathcal{A}^*(y^{\hat{k}}) V^{\hat{k}-1} = Z^{\hat{k}} \neq 0. \tag{3.5.1}$$

An implication of the proof of Theorem 3.4.3 is that $Z^k \succ 0$ if, and only if, $k = d$. Thus Q_2^k is non-zero for all $k < d$. Now for every integer $k \in [1, \hat{k} - 1]$ we have,

$$\begin{aligned}
\left(V^{\hat{k}-1}\right)^T \mathcal{A}^*(y^k) V^{\hat{k}-1} &= \left(Q_2^{\hat{k}-1}\right)^T \cdots \left(Q_2^k\right)^T \left(V^{k-1}\right)^T \mathcal{A}^*(y^k) V^{k-1} Q_2^k \cdots Q_2^{\hat{k}-1} \\
&= \left(Q_2^{\hat{k}-1}\right)^T \cdots \left(Q_2^k\right)^T Z^k Q_2^k \cdots Q_2^{\hat{k}-1} \\
&= 0,
\end{aligned}$$

since the columns of Q_2^k form a basis for $\operatorname{null}(Z^k)$, by construction. Combining this observation with (3.5.1) we conclude that,

$$\mathcal{A}^*(y^{\hat{k}}) \notin \operatorname{span}\left\{\mathcal{A}^*(y^1), \dots, \mathcal{A}^*(y^{\hat{k}-1})\right\}.$$

Since this statement holds for every integer $\hat{k} \in [2, d]$, it follows that $\mathcal{A}^*(y^1), \dots, \mathcal{A}^*(y^d)$ are linearly independent, as desired. The proof of *(ii)* follows immediately. □

Next, we show that the nullspace of the maps $(\mathcal{A}^k)^*$, generated by Algorithm 1, grows with k.

Lemma 3.5.3. *Suppose the linear maps $\mathcal{A}^1, \dots, \mathcal{A}^d$ are generated by Algorithm 1 for input $(\mathcal{A}, b)$ with $d \geq 1$. Then for $k, \hat{k} \in \{1, \dots, d\}$ with $k < \hat{k}$ it holds that,*

$$\operatorname{null}\left(\left(\mathcal{A}^k\right)^*\right) \subsetneq \operatorname{null}\left(\left(\mathcal{A}^{\hat{k}}\right)^*\right).$$

Proof. First we show that the inclusion holds. To this end, let $y \in \operatorname{null}\left((\mathcal{A}^k)^*\right)$. This implies that,

$$0 = \left(\mathcal{A}^k\right)^*(y) = \left(V^{k-1}\right)^T \mathcal{A}^*(y) V^{k-1}.$$

Then,

$$\begin{aligned}\left(\mathcal{A}^{\hat{k}}\right)^*(y) &= \left(V^{\hat{k}-1}\right)^T \mathcal{A}^*(y) V^{\hat{k}-1} \\ &= \left(Q_2^{\hat{k}-1}\right)^T \cdots \left(Q_2^k\right)^T \left[\left(V^{k-1}\right)^T \mathcal{A}^*(y) V^{k-1}\right] Q_2^k \cdots Q_2^{\hat{k}-1} \\ &= 0.\end{aligned}$$

Now to show that the inclusion is strict, let us consider y^k. By construction it holds that $(\mathcal{A}^k)^*(y^k) = Z^k$ and $Z^k \neq 0$. On the other hand, replacing y with y^k in the above derivation gives us,

$$\left(\mathcal{A}^{\hat{k}}\right)^*(y^k) = \left(Q_2^{\hat{k}-1}\right)^T \cdots \left(Q_2^k\right)^T Z^k Q_2^k \cdots Q_2^{\hat{k}-1} = 0,$$

by definition of Q_2^k, completing the proof. □

Generally speaking, it is desirable that the map $\mathcal{A}$, defining the spectrahedron $\mathcal{F}(\mathcal{A}, b)$, is surjective. However, Lemma 3.5.3 implies that surjectivity is lost after the first iteration of Algorithm 1. It may be desirable to modify the map $\mathcal{A}^k$, at each iteration, so that it is surjective. This may be accomplished by expressing $\mathcal{A}^k$ as in (2.2.1) and removing redundant matrices in the expression. Corresponding elements of b are also removed. The result is a reduction in the number of constraints, m, defining $\mathcal{F}(\mathcal{A}^k, b)$. Thus the facial reduction algorithm reduces the number of variables and the number of constraints.

Now each of these two results, Lemma 3.5.2 and Lemma 3.5.3, implies the following result first observed in [77].

Theorem 3.5.4. *Let d be generated by Algorithm 1 for input data $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$ and $b \in \mathbb{R}^m$. Then $d \leq m$.*

3.6 Facial Reduction in the Literature

We introduced facial reduction as a way of guaranteeing strong duality in SDPs. While there are other techniques designed to achieve the same goal, it is remarkable that they are essentially equivalent to facial reduction. The two other techniques are conic expansion, developed by Luo, Sturm, and Zhang in [51, 52], and the extended Lagrange-Slater dual of Ramana [69].

While in facial reduction the primal constraint $X \succeq 0$ is replaced with successively smaller cones, in the conic expansion approach, the dual constraint $Z \succeq 0$ is replaced with successively larger cones. After finitely many iterations, the cone in the dual constraint is sufficiently large so as to guarantee a zero duality gap. Waki and Muramatsu [83] show that the faces obtained in the facial reduction algorithm are *duals* of the cones obtained

in the conic expansion approach. Thus conic expansion may be viewed as an equivalent dual procedure to facial reduction. Waki and Muramatsu also present a version of the facial reduction algorithm of Borwein and Wolkowicz that can detect infeasibility. Their algorithm is modified by Liu and Pataki [45] by introducing elementary reformulations of the constraints from which certificates of infeasibility and strong duality are easily obtained.

The extended Lagrange-Slater dual is a different dual for (SDP) than the Lagrangian dual $(DSDP)$ presented in this thesis. Strong duality holds for the extended Lagrange-Slater dual without any assumptions on the feasible set of (SDP), such as the Slater assumption required in Fact 2.6.2. In [70], Ramana, Tunçel, and Wolkowicz show that the extended Lagrange-Slater dual is equivalent to the dual of (3.1.1). The extended Lagrange-Slater dual is generalized to *nice* cones by Pataki in [61] and to *homogenous* cones by Truong and Tunçel in [79], Chua and Tunçel in [12], and by Pólik and Terlaky in [67].

From an algorithmic perspective, facial reduction has seen increased attention in recent years. We highlight a few of the more notable contributions. In [10], Cheung, Schurr, and Wolkowicz formulate an optimization problem to find Z^k and prove that the problem is an instance of conic optimization where the Slater condition holds. The resulting facial reduction algorithm is provably "backward stable" when singularity degree is at most 1. Permenter, Friberg, and Andersen [65], combine facial reduction with the self-dual embedding approach to solve SDPs. The authors derive a method for performing a step of facial reduction that is based on solutions to the self-dual embedding formulation of (SDP). The facial reduction step is taken only if the solution for the self-dual embedding problem is not satisfactory. When facial reduction is necessary, fewer steps may be taken than required by other facial reduction algorithms. Lourenço, Muramatsu, and Tsuchiya [49], introduce a facial reduction algorithm that takes advantage of polyhedral faces. Worst case iteration bounds for their algorithm are better than the corresponding bounds for classical facial reduction when applied to several well-known classes of closed convex cones. For SDPs, however, the algorithm coincides with classical facial reduction.

Facial reduction, as presented in this chapter, ensures that the reduced problem $(RSDP)$ satisfies the Slater condition, but it need not be the case that the dual reduced problem, $(DRSDP)$, also satisfies the Slater condition. In [47], Lourenço, Muramatsu, and Tsuchiya introduce the notion of *double facial reduction*, where a second facial reduction procedure is performed on $(DRSDP)$. The authors prove that both the primal and dual problems after the second round of facial reduction, satisfy the Slater condition. This result can also be obtained by applying a dual version of Theorem 3.1.2 to $(DRSDP)$.

In facial reduction algorithms, the most involved step is in obtaining the exposing vector Z^k. The corresponding subproblem is generally computationally expensive. This difficulty has led to the development of heuristic facial reduction algorithms. The algorithm of Permenter and Parillo [66] utilizes approximations of the cone $\mathbb{S}^n_+$ to obtain exposing vectors more efficiently. This approach is not proven to obtain the minimal face and may require more iterations than other approaches. Another heuristic approach is proposed

by Zhu, Pataki, and Tran-Dinh [87]. Their algorithm consists of identifying exposing vectors among the matrices A_i, and requires only Cholesky decompositions at each iteration.

Problems that are not facially reduced may be prone to non-trivial numerical inaccuracies when an interior point algorithm is applied. In [84], Waki, Nakata, and Muramatsu present an instance of SDP where the optimal value is provably 0 but standard interior point algorithms return the value 1. It is subsequently shown, by Waki and Muramatsu [83], that the SDP in question has large singularity degree and that the correct optimal value may be obtained after facial reduction.

Lourenço, Muramatsu, and Tsuchiya [48] use facial reduction to derive a finite certificate for weak infeasibility, one of the hardest scenarios for SDPs. For additional reading on facial reduction, we suggest the survey of Drusvyatskiy and Wolkowicz [18].

Chapter 4

Singularity Degree in Semidefinite Programming

The singularity degree of a spectrahedron is a measure introduced by Sturm in [77]. It is defined in terms of the facial reduction algorithm as follows.

Definition 4.0.1. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a non-empty spectrahedron. The* singularity degree *of $\mathcal{F}$, denoted $\mathrm{sd}(\mathcal{F})$, is the length of a shortest facial reduction sequence generated by Algorithm 1 with input $(\mathcal{A}, b)$.*

As we have already mentioned in Chapter 3, the output of the facial reduction algorithm is not unique. In particular, it depends on the choice of exposing vector Z^k (equivalently the choice of y^k) obtained at each iteration. In Section 4.2 we show how to choose each Z^k so as to obtain $\mathrm{sd}(\mathcal{F})$.

It should be noted that our definition differs from that of Sturm in the treatment of the case $\mathcal{F} = \{0\}$. For this case Sturm defines $\mathrm{sd}(\mathcal{F}) = 0$. This definition coincides with the corresponding error bound. According to our definition, however, it should be clear that $\mathrm{sd}(\mathcal{F}) \geq 1$ in this case.

The original motivation leading to the definition of singularity degree, was to bound forward error for SDPs. We introduce the error bounds, due to Sturm, in Section 4.3. In Section 4.4 we provide theoretical bounds on $\mathrm{sd}(\widehat{\mathcal{F}})$ where $\widehat{\mathcal{F}}$ is derived from $\mathcal{F}$ and $\mathrm{sd}(\mathcal{F})$ is known. A relationship between singularity degree and *complementary slackness* is presented in Section 4.5.

4.1 Extensions to Empty Spectrahedra

The requirement that $\mathcal{F}$ is non-empty in Definition 4.0.1 is necessary due to the same assumption in Algorithm 1 and in Fact 3.3.1. We see two ways to extend the definition to also include certain types of empty spectrahedra. First a definition.

Definition 4.1.1. *Let $\mathcal{F} := \mathcal{F}(\mathcal{A}, b)$ be a spectrahedron and define the displacement of $\mathcal{L}(\mathcal{A}, b)$ and $\mathbb{S}^n_+$ as,*

$$\operatorname{disp}(\mathcal{A}, b) := \inf\{\|X - Y\| : X \in \mathcal{L}(\mathcal{A}, b),\ Y \in \mathbb{S}^n_+\}.$$

When $\operatorname{disp}(\mathcal{A}, b) > 0$ we say that $\mathcal{F}$ is strongly infeasible. When $\operatorname{disp}(\mathcal{A}, b) = 0$ and $\mathcal{F} = \emptyset$ we say that $\mathcal{F}$ is weakly infeasible.

One way to extend singularity degree to empty spectrahedra is by modifying Fact 3.3.1, as in Theorem 3.1.2 of [18]. In this case, strong infeasibility can be detected (with a certificate) in the first step of facial reduction, terminating Algorithm 1. Naturally then we could define $\operatorname{sd}(\mathcal{F}) := 1$ for all strongly infeasible $\mathcal{F}$.

A second way to define $\operatorname{sd}(\mathcal{F})$ for empty spectrahedra relies on the notion of displacement, as in [17]. When $\mathcal{F}$ is strongly infeasible and $\operatorname{disp}(\mathcal{A}, b)$ is attained, there exists a displacement matrix, say D with $\|D\| = \operatorname{disp}(\mathcal{A}, b)$, such that $\mathcal{L}(\mathcal{A}, b) + D$ is feasible, but does not satisfy the Slater condition. The matrix D translates $\mathcal{L}(\mathcal{A}, b)$ so that it intersects the boundary of $\mathbb{S}^n_+$. The new affine manifold $\mathcal{L}(\mathcal{A}, b) + D$ can be expressed as $\mathcal{L}(A, b + \hat{b})$ for some $\hat{b} \in \mathbb{R}^m$. We could now define the singularity degree of $\mathcal{F}$ as,

$$\operatorname{sd}(\mathcal{F}) := \operatorname{sd}\left(\mathcal{F}(\mathcal{A}, b + \hat{b})\right). \tag{4.1.1}$$

Unlike our first definition, this one does not treat all strongly infeasible spectrahedra equally. Algorithms such as alternating projections are interesting even for infeasible problems. When applied to a strongly infeasible $\mathcal{F}$, the algorithm returns the displacement matrix. A topic of future research would be to determine whether the convergence of alternating projections for strongly infeasible spectrahedra is affected by the definition of singularity degree in (4.1.1).

Neither of these definitions extend to weakly infeasible spectrahedra and it is not yet clear to us how to go about such an extension.

Our main interest in this thesis is forward error, which is not defined (or $+\infty$ by convention) for infeasible spectrahedra. Therefore it seems justified to assume that $\mathcal{F}$ is non-empty and stick with Definition 4.0.1.

4.2 Attainment of Singularity Degree

To obtain $\operatorname{sd}(\mathcal{F})$ we need to determine the fewest iterations required by the facial reduction algorithm. It turns out that the greedy approach, of choosing each exposing vector, Z^k, to have maximum possible rank, leads to the fewest iterations. First we address the simple case of $\operatorname{sd}(\mathcal{F}) = 0$.

Lemma 4.2.1. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a non-empty spectrahedron and let d be the number of calls to the while loop in Algorithm 1 with input $\mathcal{A}$ and b. Then the following are equivalent:*

(*i*) $d = 0$,

(*ii*) $d = \mathrm{sd}(\mathcal{F})$,

(*iii*) $\mathcal{F} \cap \mathbb{S}^n_{++} \neq \emptyset$.

Proof. The proof is a trivial implication of the condition for the while loop in Algorithm 1 and of Definition 4.0.1. □

For the more complicated case where at least 1 call to the while loop is made, we show that the number of iterations can not be made larger by choosing each Z^k in the relative interior of $\mathcal{E}(\mathcal{A}^{k-1}, b)$.

Lemma 4.2.2. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a non-empty spectrahedron and suppose that $d \geq 1$ calls to the while loop of Algorithm 1 are made when the input is $(\mathcal{A}, b)$. Let $y^1, \ldots, y^d$ be a facial reduction sequence and let $Z^1, \ldots, Z^d$ be defined as in the algorithm. Let $\hat{k} \in [1, d]$ be the smallest integer such that,*

$$Z^{\hat{k}} \notin \mathrm{relint}\left(\mathcal{E}(\mathcal{A}^{\hat{k}-1}, b)\right). \tag{4.2.1}$$

Consider the alternative sequence $Z^1, \ldots, Z^{\hat{k}-1}, \tilde{Z}^{\hat{k}}, \ldots, \tilde{Z}^{\tilde{d}}$ where,

$$\tilde{Z}^k \in \mathrm{relint}\left(\mathcal{E}(\mathcal{A}^{k-1}, b)\right), \ \forall k \geq \hat{k}. \tag{4.2.2}$$

Then $\tilde{d} \leq d$.

Proof. For $Z^{\hat{k}}$ let us define $\begin{bmatrix} Q_1^{\hat{k}} & Q_2^{\hat{k}} \end{bmatrix}$ as in Algorithm 1, i.e.,

$$\mathrm{range}(Z^{\hat{k}}) = \mathrm{range}(Q_1^{\hat{k}}), \ \mathrm{null}(Z^{\hat{k}}) = \mathrm{range}(Q_2^{\hat{k}}). \tag{4.2.3}$$

By 4.2.1, $Z^{\hat{k}}$ is not positive definite and thus $Q_1^{\hat{k}}$ and $Q_2^{\hat{k}}$ have positive dimension. Now let $\tilde{Z}^{\hat{k}} \in \mathrm{relint}\left(\mathcal{E}(\mathcal{A}^{\hat{k}-1}, b)\right)$. As for $Z^{\hat{k}}$ in (4.2.3), let $\begin{bmatrix} \tilde{Q}_1^{\hat{k}} & \tilde{Q}_2^{\hat{k}} \end{bmatrix}$ capture the range and nullspace of $\tilde{Z}^{\hat{k}}$,

$$\mathrm{range}(\tilde{Z}^{\hat{k}}) = \mathrm{range}(\tilde{Q}_1^{\hat{k}}), \ \mathrm{null}(\tilde{Z}^{\hat{k}}) = \mathrm{range}(\tilde{Q}_2^{\hat{k}}). \tag{4.2.4}$$

By (4.2.1), the convexity of $\mathcal{E}(\mathcal{A}^{\hat{k}-1}, b)$, and Lemma 2.3.3 it holds that,

$$\mathrm{range}(Z^{\hat{k}}) \subsetneq \mathrm{range}(\tilde{Z}^{\hat{k}}). \tag{4.2.5}$$

First, suppose that $\tilde{Z}^{\hat{k}} \succ 0$. Then from the proof of Theorem 3.4.3 it holds that $\mathcal{F} = \{0\}$. Thus the algorithm terminates if $\tilde{Z}^{\hat{k}}$ is chosen instead of $Z^{\hat{k}}$, but it does not terminate with the choice of $Z^{\hat{k}}$. Hence $\tilde{d} \leq d$, as desired.

We may, therefore, assume that $\tilde{Z}^{\hat{k}}$ is not positive definite. Consequently, the matrices $\tilde{Q}_1^{\hat{k}}$ and $\tilde{Q}_2^{\hat{k}}$ have positive dimension. Thus, by (4.2.5) we may write,

$$Q_2^{\hat{k}} = \begin{bmatrix} Q_R & \tilde{Q}_2^{\hat{k}} \end{bmatrix}, \tag{4.2.6}$$

for some Q_R. It follows that,

$$\operatorname{range}(\tilde{Z}^{\hat{k}}) = \operatorname{range}\left(\begin{bmatrix} Q_1^{\hat{k}} & Q_R \end{bmatrix}\right). \tag{4.2.7}$$

Now let $V^{\hat{k}} = V^{\hat{k}-1}Q_2^{\hat{k}}$ be as in Algorithm 1 and let $\tilde{V}^{\hat{k}} := V^{\hat{k}-1}\tilde{Q}_2^{\hat{k}}$ be defined for the alternate sequence where $\tilde{Z}^{\hat{k}}$ is chosen instead of $Z^{\hat{k}}$. By (4.2.6) it holds that,

$$V^{\hat{k}} = \begin{bmatrix} V^{\hat{k}-1}Q_R & V^{\hat{k}-1}\tilde{Q}_2^{\hat{k}} \end{bmatrix} = \begin{bmatrix} V^{\hat{k}-1}Q_R & \tilde{V}^{\hat{k}} \end{bmatrix} \tag{4.2.8}$$

The exposing vector obtained by the original sequence is,

$$W^{\hat{k}} = W^{\hat{k}-1} + V^{\hat{k}-1}Z^{\hat{k}}\left(V^{\hat{k}-1}\right), \tag{4.2.9}$$

and the exposing vector obtained by choosing $\tilde{Z}^{\hat{k}}$ instead of $Z^{\hat{k}}$ is,

$$\tilde{W}^{\hat{k}} := W^{\hat{k}-1} + V^{\hat{k}-1}\tilde{Z}^{\hat{k}}\left(V^{\hat{k}-1}\right). \tag{4.2.10}$$

The following rank relationship holds,

$$\begin{aligned} \operatorname{rank}(\tilde{W}^{\hat{k}}) &= \operatorname{rank}(W^{\hat{k}-1}) + \operatorname{rank}(\tilde{Z}^{\hat{k}}) \\ &= \operatorname{rank}(W^{\hat{k}-1}) + \operatorname{rank}(Q_1^{\hat{k}}) + \operatorname{rank}(Q_R) \\ &> \operatorname{rank}(W^{\hat{k}-1}) + \operatorname{rank}(Q_1^{\hat{k}}) \\ &= \operatorname{rank}(W^{\hat{k}-1}) + \operatorname{rank}(Z^{\hat{k}}) \\ &= \operatorname{rank}(W^{\hat{k}}). \end{aligned} \tag{4.2.11}$$

Moreover, $\operatorname{range}(W^{\hat{k}}) \subsetneq \operatorname{range}(\tilde{W}^{\hat{k}})$. Hence the face exposed by $\tilde{W}^{\hat{k}}$ is strictly contained in the face exposed by $W^{\hat{k}}$. This observation implies that $\hat{k} < d$. As otherwise we have,

$$\operatorname{face}(\mathcal{F}) = \mathbb{S}_+^n \cap \left(W^{\hat{k}}\right)^{\perp} \supsetneq \mathbb{S}_+^n \cap \left(\tilde{W}^{\hat{k}}\right)^{\perp} \supseteq \operatorname{face}(\mathcal{F}), \tag{4.2.12}$$

a contradiction. We may also assume that $\tilde{Z}^{\hat{k}}$ does not complete the facial reduction algorithm. Indeed, in this case it is trivial that $\tilde{d} \leq d$. With these assumptions, for either choice, $Z^{\hat{k}}$ or $\tilde{Z}^{\hat{k}}$, the while loop of the algorithm will be called at least once more.

Now we show that the next exposing vector, $\tilde{W}^{\hat{k}+1}$, for the alternate sequence is not 'worse' than the next exposing vector, $W^{\hat{k}+1}$, for the original sequence. By construction,

$$\tilde{Z}^{\hat{k}+1} \in \operatorname{relint}\left(\mathcal{E}(\tilde{\mathcal{A}}^{\hat{k}}, b)\right) = \operatorname{relint}\left\{\left(\tilde{V}^{\hat{k}}\right)^T \mathcal{A}^*(y)\tilde{V}^{\hat{k}} \succeq 0 : y^T b = 0\right\}, \tag{4.2.13}$$

where $\tilde{\mathcal{A}}^{\hat{k}} = \mathcal{A}_{\tilde{V}^{\hat{k}}}$. The corresponding exposing vector is,

$$\tilde{W}^{\hat{k}+1} = \tilde{W}^{\hat{k}} + \tilde{V}^{\hat{k}} \tilde{Z}^{\hat{k}+1} \left(\tilde{V}^{\hat{k}}\right)^T. \tag{4.2.14}$$

Similarly, for the original sequence we have,

$$Z^{\hat{k}+1} \in \left(\mathcal{E}(\mathcal{A}^{\hat{k}}, b)\right) = \left\{ \left(V^{\hat{k}}\right)^T \mathcal{A}^*(y) V^{\hat{k}} \succeq 0 : y^T b = 0 \right\} \setminus \{0\}, \tag{4.2.15}$$

with corresponding exposing vector,

$$W^{\hat{k}+1} = W^{\hat{k}} + V^{\hat{k}} Z^{\hat{k}+1} \left(V^{\hat{k}}\right)^T. \tag{4.2.16}$$

Now we show that range$(W^{\hat{k}+1}) \subseteq$ range$(\tilde{W}^{\hat{k}+1})$. By Lemma 3.4.2 (i), it holds that

$$\text{range}\left(W^{\hat{k}+1}\right) = \text{range}\left(W^{\hat{k}}\right) + \text{range}\left(V^{\hat{k}} Z^{\hat{k}+1} (V^{\hat{k}})^T\right). \tag{4.2.17}$$

We have already established that range$(W^{\hat{k}}) \subset$ range$(\tilde{W}^{\hat{k}+1})$. For the second term in (4.2.17), the expression of (4.2.8) implies that,

$$\text{range}\left(V^{\hat{k}} Z^{\hat{k}+1} (V^{\hat{k}})^T\right) \subseteq \text{range}\left(\begin{bmatrix} V^{\hat{k}-1} Q_R & \tilde{V}^{\hat{k}} \end{bmatrix}\right). \tag{4.2.18}$$

Now the subset of range $\left(V^{\hat{k}} Z^{\hat{k}+1} (V^{\hat{k}})^T\right)$ that is contained in range$(V^{\hat{k}-1} Q_R)$ is also contained in range$(\tilde{W}^{\hat{k}+1})$ by (4.2.7). All that remains is to show that the subspace of the range of $V^{\hat{k}} Z^{\hat{k}+1} (V^{\hat{k}})^T$ contained in range$(\tilde{V}^{\hat{k}})$ is also contained in the range of $\tilde{W}^{\hat{k}+1}$. To see this, note that (4.2.8) implies the existence of $y^{\hat{k}+1}$ satisfying,

$$Z^{\hat{k}+1} = \begin{bmatrix} \left(V^{\hat{k}-1} Q_R\right)^T \mathcal{A}^*(y^{\hat{k}+1}) V^{\hat{k}-1} Q_R & \left(V^{\hat{k}-1} Q_R\right)^T \mathcal{A}^*(y^{\hat{k}+1}) \tilde{V}^{\hat{k}} \\ \left(\left(V^{\hat{k}-1} Q_R\right)^T \mathcal{A}^*(y^{\hat{k}+1}) \tilde{V}^{\hat{k}}\right)^T & \left(\tilde{V}^{\hat{k}}\right)^T \mathcal{A}^*(y^{\hat{k}+1}) \tilde{V}^{\hat{k}} \end{bmatrix}. \tag{4.2.19}$$

The bottom right block implies that $y^{\hat{k}+1}$ is also feasible for the set in (4.2.13). Then the choice of $\tilde{Z}^{\hat{k}+1} \in$ relint$(\mathcal{E}(\tilde{\mathcal{A}}^{\hat{k}}, b))$ implies that the subspace of the range of $V^{\hat{k}} Z^{\hat{k}+1} (V^{\hat{k}})^T$ contained in range$(\tilde{V}^{\hat{k}})$ is also contained in the range of $\tilde{V}^{\hat{k}} \tilde{Z}^{\hat{k}+1} (\tilde{V}^{\hat{k}})^T$. Finally, (4.2.14) implies that range$(W^{\hat{k}+1}) \subseteq$ range$(\tilde{W}^{\hat{k}+1})$. Continuing in this fashion we see that at every subsequent iteration, the exposing vector obtained by the alternative sequence is at least as good as the one obtained through the original sequence. Hence $\tilde{d} \leq d$, as desired. □

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

Theorem 4.2.3. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a non-empty spectrahedron and suppose that $d \geq 1$ calls to the while loop of Algorithm 1 are made when the input consists of $\mathcal{A}$ and b. Let $y^1, \ldots, y^d$ be a facial reduction sequence. Then,*

$$\left(\mathcal{A}^{k-1}\right)^* (y^k) \in \operatorname{relint}\left(\mathcal{E}(\mathcal{A}^{k-1}, b)\right) \ \forall k \in [1, d] \implies d = \operatorname{sd}(\mathcal{F}).$$

Proof. For a simpler discussion let $Z^k := \left(\mathcal{A}^{k-1}\right)^* (y^k)$ for each $k \in \{1, \ldots, d\}$, as in Algorithm 1. We claim that d is the same regardless of the choice of Z^k, as long as Z^k is chosen in the relative interior of $\mathcal{E}(\mathcal{A}^{k-1}, b)$. Indeed, let $k \in \{1, \ldots d\}$ and let,

$$Z^k, \tilde{Z}^k \in \operatorname{relint}\left(\mathcal{E}(\mathcal{A}^{k-1}, b)\right).$$

Then Lemma 2.3.3 implies that $\operatorname{range}(Z^k) = \operatorname{range}(\tilde{Z}^k)$ and $\operatorname{null}(Z^k) = \operatorname{null}(\tilde{Z}^k)$. Defining V^k as in Alogirthm 1 and defining $\tilde{V}^k$ analogously for the alternate choice of $\tilde{Z}^k$ gives us the expression,

$$\operatorname{range}(V^k) = \operatorname{range}(\tilde{V}^k).$$

Now it is clear that for each $k \in [1, d]$ the choice of matrix in $\operatorname{relint}\left(\mathcal{E}(\mathcal{A}^{k-1}, b)\right)$ does not affect the number of iterations.

To prove the desired result we proceed by contradiction. Suppose that $d > \operatorname{sd}(\mathcal{F})$. Then there exists an alternate sequence $\tilde{Z}^1, \ldots, \tilde{Z}^{\tilde{d}}$ with $\tilde{d} = \operatorname{sd}(\mathcal{F})$. From the above discussion, there exists a smallest integer $\hat{k} \in [1, d]$ such that,

$$\tilde{Z}^{\hat{k}} \notin \operatorname{relint}\left(\mathcal{E}(\mathcal{A}^{\hat{k}-1}, b)\right).$$

Then Lemma 4.2.2 implies the existence of another alternate sequence $\hat{Z}^1, \ldots, \hat{Z}^{\hat{d}}$ with,

$$\hat{d} = \tilde{d} = \operatorname{sd}(\mathcal{F}) < d,$$

where $\hat{Z}^k \in \operatorname{relint}\left(\mathcal{E}(\mathcal{A}^{\hat{k}-1}, b)\right)$ for each $k \in [1, \hat{d}]$. The result we obtained at the beginning of this proof implies that $\hat{d} = d$, a contradiction. □

For the special case where $\mathcal{F} = \{0\}$ we have the following result.

Corollary 4.2.4. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a non-empty spectrahedron. Then,*

$$\mathcal{F} = \{0\} \implies \operatorname{sd}(\mathcal{F}) = 1.$$

Proof. The hypothesis, $\mathcal{F} = \{0\}$, implies that $b = 0$ and thus,

$$\mathcal{F} = \operatorname{null}(\mathcal{A}) \cap \mathbb{S}^n_+. \tag{4.2.20}$$

Since the Slater condition does not hold for $\mathcal{F}$, the while loop is called with $k = 0$. By Theorem 4.2.3, Z^1 is taken from the relative interior of the set,

$$\mathcal{E}(\mathcal{A}, b) = \operatorname{face}(\mathcal{F})^c \cap \operatorname{range}(\mathcal{A}^*) = \mathbb{S}^n_+ \cap \operatorname{range}(\mathcal{A}^*). \tag{4.2.21}$$

By (4.2.20), Fact 2.2.1, and (4.2.21), it holds that $\mathcal{E}(\mathcal{A}, b)$ has a Slater point. Thus the maximum rank exposing vector is obtained in the first iteration of the algorithm and the while loop is not called again, giving us $\mathrm{sd}(\mathcal{F}) = 1$, as desired. □

We conclude this section with a universal upper bound on singularity degree.

Corollary 4.2.5. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a non-empty spectrahedron with $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$. If $n \geq 2$ then,*

$$\mathrm{sd}(\mathcal{F}) \leq \min\{n-1, m\}.$$

Proof. The upper bound of m follows from Theorem 3.5.4. Moreover, from Theorem 3.4.3 it is implied that $\mathrm{sd}(\mathcal{F}) \leq n$. The only scenario for which $\mathrm{sd}(\mathcal{F}) = n$ is that of $\mathcal{F} = \{0\}$. However, in Corollary 4.2.4 we showed that $\mathrm{sd}(\mathcal{F}) = 1$ in this case, implying the desired result. □

The worst case bound of $n-1$ is shown to be attained by an example of Tunçel on p. 43 of [80]. To state the example, let us introduce some notation. For $i \in \{1, \ldots, n\}$ let $e_i \in \mathbb{R}^n$ denote the ith column of the identity matrix and define $E(i,j) \in \mathbb{S}^n$ as the matrix with 1 in the (i,j) and (j,i) positions and 0 everywhere else. That is,

$$E(i,j) = \begin{cases} e_i e_i^T & \text{if } i = j, \\ e_i e_j^T + e_j e_i^T & \text{otherwise.} \end{cases}$$

For instance in $\mathbb{S}^4$,

$$E(1,2) := \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \text{ and } E(3,3) := \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Example 4.2.6. *The example of Tunçel is a spectrahedron that is expressed in the dual form $\mathcal{A}^*(y) \preceq C$. In the notation of this thesis, the spectrahedron is $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ where,*

$$b = e_1 \text{ and } \begin{cases} A_1 := E(1,1), \\ A_i := E(i,i) + E(1,i+1) \text{ for } i \in \{2, \ldots, n-1\}, \\ A_n := E(n,n). \end{cases}$$

To see that $\mathrm{sd}(\mathcal{F}) = n-1$ let us obtain a more explicit description of $\mathcal{F}$. The constraint $\langle A_n, X \rangle = 0$ implies that $X_{nn} = 0$. Since $X \succeq 0$, the entire nth row and column consist of zeros. Then the $(n-1)$th constraint gives us,

$$0 = \langle A_{n-1}, X \rangle = X_{n-1,n-1} + X_{1,n} + X_{n,1} = X_{n-1,n-1}.$$

It follows that the $(n-1)$th row and column of X consists of zeros. We continue in this fashion until we get that the only non-zero entry of X is the $(1,1)$ entry. Then the

constraint $\langle A_1, X\rangle = 1$ *implies that* $\mathcal{F}$ *consists of the matrix with* 1 *in the upper left entry and zeros elsewhere.*

Now let $y \in \mathbb{R}^m$ *be such that* $\mathcal{A}^*(y) \in \mathcal{E}(\mathcal{A}, b)$*. Then by definition,*

$$\mathcal{A}^*(y) \in \mathbb{S}^n_+ \setminus \{0\} \text{ and } y^T b = 0.$$

It follows that $y_1 = 0$ *and that,*

$$\mathcal{A}^*(y) = y_n E(n,n) + \sum_{i=2}^{n-1} y_i \left(E(i,i) + E(1,i+1)\right). \tag{4.2.22}$$

Note that $(\mathcal{A}^*(y))_{11} = 0$*. Therefore, positive semidefiniteness implies that the first row and column of* $\mathcal{A}^*(y)$ *consists of zeros. In particular, for* $i \in \{2, \ldots, n-1\}$ *it holds that*

$$0 = (\mathcal{A}^*(y))_{1,i+1} = y_i.$$

Now (4.2.22) *simplifies to* $\mathcal{A}^*(y) = y_n E(n,n)$*, a rank one matrix. A step of facial reduction reduces the dimension by* 1 *and restricts the matrices* A_i *to their upper left* $(n-1)\times(n-1)$ *blocks for each* $i \in \{1, \ldots, n\}$*. The reduced problem is just an instance of the original problem in* $\mathbb{S}^{n-1}$*. Therefore,* $n-1$ *steps are needed to reduce the problem to the one dimensional minimal face of* $\mathcal{F}$*, implying* $\operatorname{sd}(\mathcal{F}) = n-1$*.*

4.3 Error Bounds and Singularity Degree

In (1.0.4), we stated a version of the error bound of Sturm that is 'morally' correct. The exact bound of Sturm, under our definition of singularity degree, is stated here. For a proof, see Theorem 3.3 of [77].

Fact 4.3.1. *Let* $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ *be a non-empty spectrahedron and let* $\{X(\alpha) : \alpha > 0\}$ *be a sequence where* $\|X(\alpha)\|$ *is bounded for small* α*. Then,*

$$\epsilon^f(X(\alpha), \mathcal{F}) = \begin{cases} \mathcal{O}\left(\epsilon^b(X(\alpha), \mathcal{F})\right) & \text{if } \mathcal{F} = \{0\}, \\ \mathcal{O}\left(\epsilon^b(X(\alpha), \mathcal{F})^{2^{-\operatorname{sd}(\mathcal{F})}}\right) & \text{otherwise.} \end{cases}$$

The two scenarios in which forward error and backward error are of the same order, are when $\operatorname{sd}(\mathcal{F}) = 0$ (the Slater condition holds) or when $\mathcal{F} = \{0\}$. For other cases, the bound on the discrepancy between forward error and backward error grows with singularity degree. In proving Fact 4.3.1, Sturm actually obtained the following more precise statement about the way in which $X(\alpha)$ approaches $\mathcal{F}$.

Fact 4.3.2. *Let* $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ *be a non-empty spectrahedron with* $\operatorname{sd}(\mathcal{F}) \geq 1$ *where* $\mathcal{F} \neq \{0\}$ *and let* $\{X(\alpha) : \alpha > 0\}$ *be a sequence where* $\epsilon^b(X(\alpha), \mathcal{F}) = \mathcal{O}(\alpha)$*. For each* $k \in [1, \operatorname{sd}(\mathcal{F})]$*,*

let Z^k be a maximum rank exposing vector obtained as in Algorithm 1 and let $q_k := \operatorname{rank}(Z^k)$. Let $\bar{\alpha} > 0$ be fixed. Then there exists an orthogonal matrix Q such that,

$$\operatorname{face}(Q\mathcal{F}Q^T) = \begin{bmatrix} \mathbb{S}^r_+ & 0 \\ 0 & 0 \end{bmatrix},$$

and,

$$QX(\alpha)Q^T = \begin{bmatrix} X_0(\alpha) & * & \cdots & * \\ * & X_1(\alpha) & & \\ \vdots & & \ddots & * \\ * & * & * & X_{\operatorname{sd}(\mathcal{F})}(\alpha) \end{bmatrix},$$

where $X_0(\alpha) \in \mathbb{S}^r$ and for all $k \in [1, \operatorname{sd}(\mathcal{F})]$ and $\alpha \in (0, \bar{\alpha})$ it holds that,

$$X_k(\alpha) \in \mathbb{S}^{q_k} \text{ and } \|X_k(\alpha)\| = \mathcal{O}\left(\alpha^{\xi(k)}\right),$$

where $\xi(k) := 2^{-(\operatorname{sd}(\mathcal{F})-k)}$.

Under the correct orthogonal transformation, the diagonal blocks of $X(\alpha)$ that converge to 0 may do so at different rates. Such unbalanced convergence is a real hindrance to algorithms, since only components of the fastest coverging blocks can be approximated to high accuracy.

4.4 Singularity Degree of Transformed Spectrahedra

In this section we consider several transformations of spectrahedra and analyze the effect they have on singularity degree. Suppose we are given a spectrahedron $\mathcal{F}$ with known singularity degree, $\operatorname{sd}(\mathcal{F})$, and we obtain a new spectrahedron $\widehat{\mathcal{F}}$ by somehow modifying $\mathcal{F}$. Our goal here is to say as much as we can about $\operatorname{sd}(\widehat{\mathcal{F}})$ without explicitly computing it using Algorithm 1.

4.4.1 Transformations of the form $M \cdot M^T$

Given a non-empty spectrahedron $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ and a matrix $M \in \mathbb{R}^{n\times p}$ define,

$$\widehat{\mathcal{F}} := \mathcal{F}(\mathcal{A}_M, b), \tag{4.4.1}$$

where $\mathcal{A}_M$ is as in (2.2.2). We begin by obtaining expressions for $\widehat{\mathcal{F}}$ and $\operatorname{face}(\widehat{\mathcal{F}})$.

Lemma 4.4.1. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a non-empty spectrahedron and let $\widehat{\mathcal{F}}$ be as in (4.4.1) for some $M \in \mathbb{R}^{n\times p}$. If $\widehat{\mathcal{F}}$ is non-empty and $M^\dagger$ denotes the Moore-Penrose pseudoinverse of M then,*

$$\widehat{\mathcal{F}} = M^\dagger \left(\mathcal{F} \cap M\mathbb{S}^p_+ M^T\right) \left(M^\dagger\right)^T + \left(\mathbb{S}^p_+ \cap \left(M^T M\right)^\perp\right).$$

Proof. By definition of $\widehat{\mathcal{F}}$ we have,

$$\begin{aligned}\widehat{\mathcal{F}} &= \{Y \in \mathbb{S}^p : \mathcal{A}_M(Y) = b,\ Y \succeq 0\} \\ &= \{Y \in \mathbb{S}^p : \exists X \in \mathcal{F} \text{ with } X = MYM^T,\ Y \succeq 0\} \\ &= \{Y \in \mathbb{S}^p : \exists X \in \mathcal{F} \text{ with } X = MYM^T\} \cap \mathbb{S}^p_+.\end{aligned}$$

The set $\{Y \in \mathbb{S}^p : \exists X \in \mathcal{F} \text{ with } X = MYM^T\}$ is the preimage of $\mathcal{F} \cap M\mathbb{S}^pM^T$ under the map $M \cdot M^T$. Since the set is also contained in the range of $M \cdot M^T$, the preimage is obtained in terms of the Moore-Penrose pseudoinverse of $M \cdot M^T$. It is a simple exercise to verify that this pseudoinverse is $M^\dagger \cdot \left(M^\dagger\right)^T$. Therefore,

$$\widehat{\mathcal{F}} = \left(M^\dagger \left(\mathcal{F} \cap M\mathbb{S}^p_+M^T\right) \left(M^\dagger\right)^T + \text{null}(M \cdot M^T)\right) \cap \mathbb{S}^p_+. \tag{4.4.2}$$

Now $\text{null}(M \cdot M^T)^\perp = \text{range}(M^\dagger \cdot (M^\dagger)^T)$ by the properties of the Moore-Penrose pseudoinverse. Thus,

$$Y^1 \in M^\dagger \left(\mathcal{F} \cap M\mathbb{S}^p_+M^T\right) \left(M^\dagger\right)^T,\ Y^2 \in \text{null}(M \cdot M^T) \implies \langle Y^1, Y^2\rangle = 0.$$

Moreover, Y^1 as above is positive semidefinite. So $Y^1 + Y^2 \in \mathbb{S}^p_+$ if, and only if, $Y^2 \in \mathbb{S}^p_+$. We may, therefore, write $\widehat{\mathcal{F}}$ as,

$$\widehat{\mathcal{F}} = M^\dagger \left(\mathcal{F} \cap M\mathbb{S}^p_+M^T\right) \left(M^\dagger\right)^T + \left(\text{null}(M \cdot M^T) \cap \mathbb{S}^p_+\right). \tag{4.4.3}$$

Finally,

$$\begin{aligned}\text{null}(M \cdot M^T) \cap \mathbb{S}^p_+ &= \{Y \in \mathbb{S}^p_+ : MYM^T = 0\} \\ &= \{Y \in \mathbb{S}^p_+ : \text{trace}(MYM^T) = 0\} \\ &= \{Y \in \mathbb{S}^p_+ : \langle M^TM, Y\rangle = 0\} \\ &= \mathbb{S}^p_+ \cap \left(M^TM\right)^\perp.\end{aligned}$$

Substituting into (4.4.3) yields the desired result. □

How does the singularity degree of $\widehat{\mathcal{F}}$ compare to that of $\mathcal{F}$? Let us first answer this question for matrices M with $p \leq n$ and linearly independent columns. Naively, we may think that $\text{sd}(\widehat{\mathcal{F}}) \leq \text{sd}(\mathcal{F})$, however, this is not true in general as demonstrated by the following example.

Example 4.4.2. *Let $\mathcal{A} : \mathbb{S}^2 \to \mathbb{R}^2$ be defined by the matrices,*

$$A_1 := \begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix},\ A_2 := \begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix},$$

and consider the spectrahedron $\mathcal{F} = \mathcal{F}(\mathcal{A}, 0)$. *Note that* $\langle A_1, X\rangle = 0$ *implies that* X *is diagonal and the second constraint* $\langle A_2, X\rangle = 0$ *implies that the diagonal entries have the same value. Therefore,*

$$\mathcal{F} = \{\mu I : \mu \in \mathbb{R}_+\}.$$

It follows that $\operatorname{sd}(\mathcal{F}) = 0$. *Now let us consider* $\widehat{\mathcal{F}}$ *as in* (4.4.1), *defined in terms of,*

$$M := \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{4.4.4}$$

Then,

$$\begin{aligned} Y \in \widehat{\mathcal{F}} &\iff \mathcal{A}_M(Y) = 0,\ Y \succeq 0 \\ &\iff \left\langle A_1, \begin{bmatrix} Y & 0 \\ 0 & 0 \end{bmatrix} \right\rangle = 0,\ \left\langle A_2, \begin{bmatrix} Y & 0 \\ 0 & 0 \end{bmatrix} \right\rangle = 0,\ Y \geq 0 \\ &\iff Y = 0. \end{aligned}$$

However, 0 *is not a Slater point of* $\widehat{\mathcal{F}}$, *implying that* $\operatorname{sd}(\widehat{\mathcal{F}}) \geq 1 > \operatorname{sd}(\mathcal{F})$.

Why does singularity degree increase after the transformation $M \cdot M^T$ in this example? It turns out that, while the transformation $M \cdot M^T$ takes the problem to a smaller dimension, the problem is such that the minimal face of the new problem is also brought to a lower dimension. In particular $\operatorname{face}(\mathcal{F} \cap M\mathbb{S}^p_+M^T)$ is a proper subset of the face $\operatorname{face}(\mathcal{F}) \cap \left(M\mathbb{S}^p_+M^T\right)$. Eliminating this type of phenomenon actually ensures that singularity degree decreases with the transformation $M \cdot M^T$.

Theorem 4.4.3. *Let* $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ *be a non-empty spectrahedron and let* $\widehat{\mathcal{F}}$ *be as in* (4.4.1) *for some* $M \in \mathbb{R}^{n\times p}$ *with linearly independent columns. Then,*

$$\operatorname{face}(\mathcal{F} \cap M\mathbb{S}^p_+M^T) = \operatorname{face}(\mathcal{F}) \cap M\mathbb{S}^p_+M^T \neq \emptyset \implies \operatorname{sd}(\widehat{\mathcal{F}}) \leq \operatorname{sd}(\mathcal{F}). \tag{4.4.5}$$

Moreover, if $y^1, \ldots, y^d$ *is a facial reduction sequence for Algorithm 1 with input* $(\mathcal{A}, b)$, *then a subset of the* y^k*s is a facial reduction sequence when the input is* $(\mathcal{A}_M, b)$.

Proof. First note that $M\widehat{\mathcal{F}}M^T = \mathcal{F} \cap M\mathbb{S}^p_+M^T$. Thus by the hypothesis we have,

$$\operatorname{face}(M\widehat{\mathcal{F}}M^T) = \operatorname{face}(\mathcal{F}) \cap M\mathbb{S}^p_+M^T.$$

It follows that,

$$M \operatorname{range}\left(\widehat{\mathcal{F}}\right) = \operatorname{range}\left(M\widehat{\mathcal{F}}M^T\right) = \operatorname{range}(\mathcal{F}) \cap \operatorname{range}(M). \tag{4.4.6}$$

Now (4.4.5) certainly holds when $\operatorname{sd}(\widehat{\mathcal{F}}) = 0$ and it also holds when $\operatorname{sd}(\mathcal{F}) = 0$. Indeed in this case $\operatorname{range}(\mathcal{F}) \cap \operatorname{range}(M) = \operatorname{range}(M)$ and (4.4.6) implies that $\operatorname{range}(\widehat{\mathcal{F}}) = \mathbb{R}^p$.

Hence $\mathrm{sd}(\widehat{\mathcal{F}}) = 0$. Thus we may assume that the while loop of Algorithm 1 is called at least once with input $(\mathcal{A}, b)$ and it is called at least once with input $(\mathcal{A}_M, b)$.

Now let $y^1, \ldots, y^d$ be a facial reduction sequence for input $(\mathcal{A}, b)$ with $d \geq 1$ and let $V^0, V^1, \ldots, V^d$ be as in the algorithm. Let us define $\widehat{V}^0, \widehat{V}^1, \ldots, \widehat{V}^d$ inductively as follows. Let $\widehat{V}^0 = I$ in $\mathbb{S}^p$. Then for each $k \in \{1, \ldots, d\}$ let $\widehat{Q}_2^k$ be defined so that its columns form an orthonormal basis for $\mathrm{null}((V^{k-1})^T M^T \mathcal{A}^*(y^k) M V^{k-1})$ and define,

$$\widehat{V}^k := \widehat{V}^{k-1} \widehat{Q}_2^k. \tag{4.4.7}$$

First we claim that,

$$\mathrm{range}(M\widehat{V}^k) \subseteq \mathrm{range}(V^k), \ \forall k \in \{0, \ldots, d\}. \tag{4.4.8}$$

For the base case let $k = 0$ and recall that $V^0 = I$, as defined in Algorithm 1. Moreover, by construction $\widehat{V}^0 = I$, implying (4.4.8). Now suppose (4.4.8) holds for $k - 1$. Recall that Q_2^k is defined so that,

$$\mathrm{range}(Q_2^k) = \mathrm{null}\left(\left(V^{k-1}\right)^T \mathcal{A}^*(y^k) V^{k-1} \right). \tag{4.4.9}$$

It follows that,

$$\mathrm{range}\left(V^k\right) = \mathrm{range}\left(V^{k-1} Q_2^k\right) = \mathrm{range}\left(V^{k-1}\right) \cap \mathrm{null}\left(\mathcal{A}^*(y^k)\right). \tag{4.4.10}$$

Now we have assumed that (4.4.8) holds for $k - 1$. Thus,

$$\left(V^{k-1}\right)^T \mathcal{A}^*(y^k) V^{k-1} \succeq 0 \implies \left(\widehat{V}^{k-1}\right)^T M^T \mathcal{A}^*(y^k) M \widehat{V}^{k-1} \succeq 0. \tag{4.4.11}$$

Recalling the definition of $\widehat{Q}_2^k$ and the reasoning behind (4.4.9) and (4.4.10) we get that,

$$\mathrm{range}\left(M\widehat{V}^k\right) = \mathrm{range}\left(M\widehat{V}^{k-1}\widehat{Q}_2^k\right) = \mathrm{range}\left(M\widehat{V}^{k-1}\right) \cap \mathrm{null}\left(\mathcal{A}^*(y^k)\right). \tag{4.4.12}$$

Now the inductive hypothesis, (4.4.10), and (4.4.12) imply (4.4.8). We have proven the claim. In particular it holds that,

$$\mathrm{range}\left(M\widehat{V}^d\right) \subseteq \mathrm{range}\left(V^d\right) \cap \mathrm{range}(M) = \mathrm{range}(\mathcal{F}) \cap \mathrm{range}(M). \tag{4.4.13}$$

By construction of the matrices $\widehat{V}^k$ and by (4.4.11) it holds that,

$$\mathrm{range}\left(\widehat{V}^0\right) \supseteq \mathrm{range}\left(\widehat{V}^1\right) \supseteq \cdots \supseteq \mathrm{range}\left(\widehat{V}^d\right) \supseteq \mathrm{range}\left(\widehat{\mathcal{F}}\right). \tag{4.4.14}$$

Note that it is possible that for some k the matrix $(V^{k-1})^T M^T \mathcal{A}^*(y^k) M V^k$ is identically 0. In this case, $\widehat{Q}_2^k$ is full rank and the range of $\widehat{V}^k$ is no different than that of $\widehat{V}^{k-1}$. Therefore let us construct a subsequence of the $\widehat{V}^k$s where $\widehat{V}^{\bar{k}}$ is omitted if it has the same range

as $\widehat{V}^{\bar{k}-1}$. This new sequence contains at most d elements and it can be generated by Algorithm 1 using the corresponding subset of y^ks. All that remains is to show that the final matrix in this sequence captures range($\widehat{\mathcal{F}}$). Let $\widehat{V}$ denote this final matrix in the shortened sequence. Then range($\widehat{V}$) = range($\widehat{V}^d$) and by (4.4.6), (4.4.14), and (4.4.13) it holds that,

$$\operatorname{range}(\mathcal{F}) \cap \operatorname{range}(M) = M \operatorname{range}(\widehat{\mathcal{F}}) \subseteq \operatorname{range}(M\widehat{V}) \subseteq \operatorname{range}(\mathcal{F}) \cap \operatorname{range}(M). \quad (4.4.15)$$

Therefore, $M \operatorname{range}(\widehat{\mathcal{F}}) = M \operatorname{range}(\widehat{V})$ and the assumption that M is full column rank yields that range($\widehat{V}$) = range($\widehat{\mathcal{F}}$). We have shown that for an arbitrary facial reduction sequence $y^1, \ldots, y^d$ generated by Algorithm 1 with input $(\mathcal{A}, b)$, a subset of this sequence is a facial reduction sequence for the algorithm with input $(\mathcal{A}_M, b)$. Using a sequence where $d = \operatorname{sd}(\mathcal{F})$ yields (4.4.5). □

Let us state a few special cases addressed in Theorem 4.4.3. The first is when M induces partial facial reduction as in Section 3.2.

Corollary 4.4.4. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a non-empty spectrahedron and let $\widehat{\mathcal{F}}$ be as in (4.4.1) for some $M \in \mathbb{R}^{n\times p}$ with linearly independent columns. Then,*

$$\operatorname{range}(\mathcal{F}) \subseteq \operatorname{range}(M) \implies \operatorname{sd}(\widehat{\mathcal{F}}) \leq \operatorname{sd}(\mathcal{F}).$$

Proof. Under the hypothesis it holds that $\mathcal{F} \cap M\mathbb{S}^p_+ M^T = \mathcal{F}$. Consequently, the hypothesis of Theorem 4.4.3 holds, as desired. □

The next special case addressed in Theorem 4.4.3 is that singularity degree is invariant under some automorphisms of $\mathbb{S}^n_+$.

Corollary 4.4.5. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a non-empty spectrahedron and let $\widehat{\mathcal{F}}$ be as in (4.4.1) for a square and non-singular M. Then,*

$$\operatorname{sd}(\widehat{\mathcal{F}}) = \operatorname{sd}(\mathcal{F}).$$

Moreover, $y^1, \ldots, y^{\operatorname{sd}(\mathcal{F})}$ is a facial reduction sequence for Algorithm 1 with input $(\mathcal{A}, b)$ if, and only if, it is a facial reduction sequence for Algorithm 1 with input $(\mathcal{A}_M, b)$.

Proof. When M is square and non-singular it satisfies the hypothesis of Corollary 4.4.4. Moreover M^{-1} also satisfies the hypothesis. Hence we have,

$$\operatorname{sd}(\mathcal{F}) \geq \operatorname{sd}(\widehat{\mathcal{F}}) \geq \operatorname{sd}\left(\mathcal{F}\left(\left(\mathcal{A}_M\right)_{M^{-1}}, b\right)\right) = \operatorname{sd}(\mathcal{F}),$$

implying $\operatorname{sd}(\mathcal{F}) = \operatorname{sd}(\widehat{\mathcal{F}})$, as desired. The statement about facial reduction sequences of minimal length follows similarly. □

Possibly the most useful scenario captured by Corollary 4.4.5 is that of an orthogonal matrix M. A suitable orthogonal transformation of a spectrahedron $\mathcal{F}$ allows for exposition and analysis that is less cumbersome. To see this, let $d = \mathrm{sd}(\mathcal{F})$ and let $y^1, \ldots, y^d$ be a facial reduction sequence when the input is $(\mathcal{A}, b)$. Let Q_1^k, Q_2^k and V^k be as in the algorithm for each k. Then we claim that the matrix,

$$Q := \begin{bmatrix} Q_1^1 & V^1 Q_1^2 & \cdots & V^{d-1} Q_1^d & V^d \end{bmatrix},$$

is orthogonal. Indeed, by Lemma 3.4.2 *(i)* it holds that the columns of V^d form a basis for $\mathrm{range}(\mathcal{F})$ and the remaining columns of Q form a basis for $\mathrm{range}(\mathrm{face}(\mathcal{F})^c)$. Hence V^d is orthogonal to the other columns of Q. Moreover, for every integer $k \in \{1, \ldots, d\}$ we have,

$$V^k = Q_2^1 \cdots Q_2^k.$$

Thus, the columns of V^d and the columns of $V^k Q_1^{k+1}$ are orthonormal for every $k \in [1, d-1]$. Finally, if we let $V^0 := I$, as in the algorithm, then for any $0 \leq k < \ell \leq d-1$ it holds that,

$$\begin{aligned} \left(V^k Q_1^{k+1}\right)^T V^\ell Q_1^{\ell+1} &= \left(Q_1^{k+1}\right)^T \left(V^k\right)^T V^k Q_2^{k+1} \cdots Q_2^\ell Q_1^{\ell+1} \\ &= \left(Q_1^{k+1}\right)^T Q_2^{k+1} \cdots Q_2^\ell Q_1^{\ell+1} \\ &= 0, \end{aligned}$$

since $\begin{bmatrix} Q_1^{k+1} & Q_2^{k+1} \end{bmatrix}$ is orthogonal. We have, therefore, shown that Q is orthogonal.

Now suppose that M is orthogonal and that the facial reduction algorithm is run with input $(\mathcal{A}_M, b)$. Suppose furthermore that the same y^ks as those generated with input $(\mathcal{A}, b)$ are used. Then we may construct an orthogonal matrix akin to that of Q and it is exactly the matrix,

$$Q_M := \begin{bmatrix} M^T Q_1^1 & M^T V^1 Q_1^2 & \cdots & M^T V^{d-1} Q_1^d & M^T V^d \end{bmatrix} = M^T Q.$$

This is readily obtained from the proof of Theorem 4.4.3. In particular, choosing $M = Q$ gives us that $Q_M = I$. Thus the range of each exposing vector is captured by a few of the columns of the identity matrix. Consequently, each Z^k may be written as,

$$Z^k = \begin{bmatrix} S^k & 0 \\ 0 & 0 \end{bmatrix}, \; S^k \succ 0.$$

Then $V^k Z^{k+1} (V^k)^T$ is a matrix consisting of S^{k+1} in a diagonal block, and zeros everywhere else. The final exposing vector, denoted W_M, and the minimal face are then written as,

$$W_M = \begin{bmatrix} S^1 & 0 & \cdots & 0 & 0 \\ 0 & S^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & S^d & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathrm{face}(\widehat{\mathcal{F}}) = \begin{bmatrix} 0 & 0 \\ 0 & \mathbb{S}^r_+ \end{bmatrix}, \tag{4.4.16}$$

respectively. When it is beneficial to do so, we will assume that exposing vectors generated by Algorithm 1 have a block structure similar to that of (4.4.16).

Our discussion thus far has been restricted to the case $p \leq n$. We conclude this section with a statement about the case $p > n$.

Theorem 4.4.6. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a non-empty spectrahedron and let $\widehat{\mathcal{F}}$ be as in (4.4.1) for some $M \in \mathbb{R}^{n\times p}$ with linearly independent rows. Then,*

$$\operatorname{sd}(\widehat{\mathcal{F}}) = \operatorname{sd}(\mathcal{F}).$$

Moreover, $y^1, \ldots, y^{\operatorname{sd}(\mathcal{F})}$ is a facial reduction sequence for Algorithm 1 with input $(\mathcal{A}, b)$ if, and only if, it is a facial reduction sequence for Algorithm 1 with input $(\mathcal{A}_M, b)$.

Proof. The result holds when $p = n$ by Corollary 4.4.5. Thus we may assume that $p > n$. Let $N \in \mathbb{R}^{p\times p-n}$ be such that its columns form a basis for $\operatorname{null}(M)$ and define,

$$\tilde{M} := \begin{bmatrix} M^\dagger & N \end{bmatrix}.$$

Since $\tilde{M}$ is square and full rank we have, by Corollary 4.4.5, that,

$$\operatorname{sd}\left(\widehat{\mathcal{F}}\right) = \operatorname{sd}\left(\mathcal{F}\left(\mathcal{A}_{M\tilde{M}}, b\right)\right) = \operatorname{sd}\left(\mathcal{F}\left(\mathcal{A}_{\begin{bmatrix} I & 0 \end{bmatrix}}, b\right)\right).$$

To make the discourse less cumbersome let,

$$\tilde{\mathcal{A}} := \mathcal{A}_{\begin{bmatrix} I & 0 \end{bmatrix}} \text{ and } \tilde{\mathcal{F}} := \mathcal{F}\left(\tilde{\mathcal{A}}, b\right).$$

So it suffices to show that $\operatorname{sd}(\tilde{\mathcal{F}}) = \operatorname{sd}(\mathcal{F})$. To this end observe that,

$$\begin{aligned} \tilde{\mathcal{F}} &= \left\{ Y \in \mathbb{S}^p_+ : Y = \begin{bmatrix} X & * \\ * & * \end{bmatrix},\ \mathcal{A}(X) = b \right\} \\ &= \begin{bmatrix} \mathcal{F} & * \\ * & * \end{bmatrix} \cap \mathbb{S}^p_+. \end{aligned}$$

Thus if $\bar{X} \in \operatorname{relint}(\mathcal{F})$ it holds that,

$$\operatorname{range}\left(\tilde{\mathcal{F}}\right) = \operatorname{range}\left(\begin{bmatrix} \bar{X} & 0 \\ 0 & I \end{bmatrix}\right). \tag{4.4.17}$$

Now we claim that $y^1, \ldots, y^d$ is a facial reduction sequence for the input $(\mathcal{A}, b)$, if and only if, it is a facial reduction sequence for the input $\left(\tilde{\mathcal{A}}, b\right)$. To this end observe that for any $y \in \mathbb{R}^m$ we have,

$$\tilde{\mathcal{A}}^*(y) = \left(\mathcal{A}_{\begin{bmatrix} I & 0 \end{bmatrix}}\right)^*(y) = \begin{bmatrix} \mathcal{A}^*(y) & 0 \\ 0 & 0 \end{bmatrix},$$

and therefore,

$$\mathcal{A}^*(y) \in \mathcal{E}(\mathcal{A}, b) \iff \tilde{\mathcal{A}}^*(y) \in \mathcal{E}\left(\tilde{\mathcal{A}}, b\right).$$

The choice of y at the first iteration of the facial reduction algorithm is the same with input $(\mathcal{A}, b)$ as it is with input $\left(\tilde{\mathcal{A}}, b\right)$. Now let y^1 be chosen so that it is feasible for the first iteration and let V^1 capture the nullspace of $\mathcal{A}^*(y^1)$, keeping in line with the notation of Algorithm 1. Then,

$$\tilde{V}^1 := \begin{bmatrix} V^1 & 0 \\ 0 & I \end{bmatrix},$$

captures the nullspace of $\tilde{\mathcal{A}}^*(y^1)$. Therefore, it holds that,

$$\text{range}(\mathcal{F}) \subseteq \text{range}\left(V^1\right) \text{ and } \text{range}\left(\tilde{\mathcal{F}}\right) \subseteq \text{range}\left(\tilde{V}^1\right).$$

At the next iteration of the algorithm we have,

$$\left(\tilde{\mathcal{A}}_{\tilde{V}^1}\right)^*(y) = \left(\mathcal{A}_{\begin{bmatrix} V^1 & 0 \end{bmatrix}}\right)^*(y) = \begin{bmatrix} (V^1)^T \mathcal{A}^*(y) V^1 & 0 \\ 0 & 0 \end{bmatrix}.$$

It follows that,

$$\left(\mathcal{A}_{V^1}\right)^*(y) \in \mathcal{E}\left(\mathcal{A}_{V^1}, b\right) \iff \left(\tilde{\mathcal{A}}_{\tilde{V}^1}\right)^*(y) \in \mathcal{E}\left(\tilde{\mathcal{A}}_{\tilde{V}^1}, b\right).$$

Having chosen y^1 in the first iteration of the algorithm, the set of choices for y^2 with input $(\mathcal{A}, b)$ is identical to the set of choices for y^2 with input $\left(\tilde{\mathcal{A}}, b\right)$. Let us therefore choose a feasible y^2 and let Q_2^2 (using the notation of Algorithm 1) be such that it captures the nullspace of $(V^1)\,\mathcal{A}^*(y^2)V^1$. Then,

$$\text{range}\left(\begin{bmatrix} Q_2^2 & 0 \\ 0 & I \end{bmatrix}\right) = \text{null}\left(\begin{bmatrix} (V^1)^T \mathcal{A}^*(y^2) V^1 & 0 \\ 0 & 0 \end{bmatrix}\right) = \text{null}\left(\left(\tilde{\mathcal{A}}_{\tilde{V}^1}\right)^*(y)\right).$$

It follows that, if we define,

$$\tilde{V}^2 := \tilde{V}^1 \begin{bmatrix} Q_2^2 & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} V^2 & 0 \\ 0 & I \end{bmatrix},$$

then,

$$\text{range}(\mathcal{F}) \subseteq \text{range}\left(V^2\right) \text{ and } \text{range}\left(\tilde{\mathcal{F}}\right) \subseteq \text{range}\left(\tilde{V}^2\right).$$

Continuing in this fashion we see that for every subsequent iteration k, the set of choices for y^k is the same regardless of the input. Moreover, if V^k and $\tilde{V}^k$ are defined as above we have,

$$\text{range}(\mathcal{F}) \subseteq \text{range}\left(V^k\right) \text{ and } \text{range}\left(\tilde{\mathcal{F}}\right) \subseteq \text{range}\left(\tilde{V}^k\right) = \text{range}\left(\begin{bmatrix} V^k & 0 \\ 0 & I \end{bmatrix}\right).$$

Now (4.4.17) implies that the algorithm terminates at the same iteration for input $(\mathcal{A}, b)$ as it does for input $(\tilde{\mathcal{A}}, b)$. We have therefore shown that $y^1, \ldots, y^d$ is a facial reduction sequence for the input $(\mathcal{A}, b)$ if, and only if, it is a facial reduction sequence for the input $(\tilde{\mathcal{A}}, b)$, proving our claim. The desired results are now immediate. □

4.4.2 Additional Constraints

We now turn our attention to transformations of a spectrahedron that are obtained by introducing additional constraints. Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ and $\tilde{\mathcal{F}} = \mathcal{F}(\tilde{\mathcal{A}}, \tilde{b})$ be two spectrahedra with,

$$\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m\,,\ b \in \mathbb{R}^m \ \text{ and } \tilde{\mathcal{A}} : \mathbb{S}^n \to \mathbb{R}^p,\ \tilde{b} \in \mathbb{R}^p.$$

Define the map $\widehat{\mathcal{A}} : \mathbb{S}^n \to \mathbb{R}^{m+p}$ and $\widehat{b} \in \mathbb{R}^{m+p}$ as,

$$\widehat{\mathcal{A}}(X) := \begin{pmatrix} \mathcal{A}(X) \\ \tilde{\mathcal{A}}(X) \end{pmatrix},\ \widehat{b} := \begin{pmatrix} b \\ \tilde{b} \end{pmatrix}. \tag{4.4.18}$$

Then we define the modified spectrahedron as,

$$\widehat{\mathcal{F}} := \mathcal{F}\left(\widehat{\mathcal{A}}, \widehat{b}\right). \tag{4.4.19}$$

Note that $\widehat{\mathcal{F}}$ is just the intersection of $\mathcal{F}$ and $\tilde{\mathcal{F}}$. This setting has some similarities to that of Section 4.4.1 where $\tilde{\mathcal{F}}$ is assumed to have the form $M\mathbb{S}^p_+M^T$. Our first result addresses several special cases.

Lemma 4.4.7. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ and $\tilde{\mathcal{F}} = \mathcal{F}(\tilde{\mathcal{A}}, \tilde{b})$ be non-empty spectrahedra and let $\widehat{\mathcal{F}}$ be as in (4.4.19). Then,*

(*i*) $\operatorname{face}(\widehat{\mathcal{F}}) = \operatorname{face}(\mathcal{F}) \implies \operatorname{sd}(\widehat{\mathcal{F}}) \leq \operatorname{sd}(\mathcal{F})$,

(*ii*) $\operatorname{face}(\widehat{\mathcal{F}}) = \operatorname{face}(\mathcal{F}) = \operatorname{face}(\tilde{\mathcal{F}}) \implies \operatorname{sd}(\widehat{\mathcal{F}}) \leq \min\{\operatorname{sd}(\mathcal{F}), \operatorname{sd}(\tilde{\mathcal{F}})\}$.

Proof. Let $y^1, \ldots, y^d$ be a minimum length facial reduction sequence for input $(\mathcal{A}, b)$. Then $d = \operatorname{sd}(\mathcal{F})$. Observe that,

$$\widehat{\mathcal{A}}^*\left(\begin{pmatrix} y^1 \\ 0 \end{pmatrix}\right) = \mathcal{A}^*(y^1) \in \mathbb{S}^n_+ \setminus \{0\} \text{ and } \begin{pmatrix} y^1 \\ 0 \end{pmatrix}^T \widehat{b} = \left(y^1\right)^T b = 0.$$

Thus $\begin{pmatrix} (y^1)^T & 0 \end{pmatrix}^T$ is a suitable choice for the first iteration of the facial reduction algorithm with input $(\widehat{\mathcal{A}}, \widehat{b})$. In fact, it is not difficult to see that we may continue in this fashion. That is,

$$\begin{pmatrix} y^1 \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} y^d \\ 0 \end{pmatrix}$$

is a facial reduction sequence for input $(\widehat{\mathcal{A}}, \widehat{b})$, implying (*i*). The proof of (*ii*) is obtained by applying (*i*) to $\tilde{\mathcal{F}}$ as well. □

We may assume a hypothesis similar to that of Theorem 4.4.3, in order to obtain singularity degree bounds for a larger class of spectrahedra.

Theorem 4.4.8. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ and $\tilde{\mathcal{F}} = \mathcal{F}(\tilde{\mathcal{A}}, \tilde{b})$ be non-empty spectrahedra and let $\widehat{\mathcal{F}}$ be as in (4.4.19). Then,*

$$\operatorname{face}\left(\widehat{\mathcal{F}}\right) = \operatorname{face}\left(\mathcal{F}\right) \cap \operatorname{face}\left(\tilde{\mathcal{F}}\right) \neq \emptyset \implies \operatorname{sd}\left(\widehat{\mathcal{F}}\right) \leq \max\left\{\operatorname{sd}(\mathcal{F}), \operatorname{sd}\left(\tilde{\mathcal{F}}\right)\right\}.$$

Proof. Let us address three cases involving the Slater condition. In the first case, $\operatorname{sd}(\widehat{\mathcal{F}}) = 0$, the result is trivially true. In the second case, suppose that $\operatorname{sd}(\mathcal{F}) = 0$ and $\operatorname{sd}(\tilde{\mathcal{F}}) = 0$. By the hypothesis we have,

$$\operatorname{face}\left(\widehat{\mathcal{F}}\right) = \operatorname{face}(\mathcal{F}) \cap \operatorname{face}\left(\tilde{\mathcal{F}}\right) = \mathbb{S}^n_+ \cap \mathbb{S}^n_+ = \mathbb{S}^n_+,$$

and therefore $\operatorname{sd}(\widehat{\mathcal{F}}) = 0$, as desired. For the third, and final, case suppose that exactly one of $\mathcal{F}$ and $\tilde{\mathcal{F}}$ satisfies the Slater condition. Without loss of generality we may assume that $\operatorname{sd}(\mathcal{F}) > 0$ and $\operatorname{sd}(\tilde{\mathcal{F}}) = 0$. Then,

$$\operatorname{face}\left(\widehat{\mathcal{F}}\right) = \operatorname{face}(\mathcal{F}) \cap \operatorname{face}\left(\tilde{\mathcal{F}}\right) = \operatorname{face}(\mathcal{F}) \cap \mathbb{S}^n_+ = \operatorname{face}(\mathcal{F}). \tag{4.4.20}$$

Now let $y^1, \ldots, y^d$ be a facial reduction sequence for input $(\mathcal{A}, b)$ with $d = \operatorname{sd}(\mathcal{F})$. Then it is not difficult to see that,

$$\widehat{y}^k := \begin{pmatrix} y^k \\ 0 \end{pmatrix}, \ k \in \{1, \ldots, d\}, \tag{4.4.21}$$

is a facial reduction sequence for input $(\widehat{\mathcal{A}}, \widehat{b})$. To see this let us just look at the first iteration of Algorithm 1 for input $(\widehat{\mathcal{A}}, \widehat{b})$. We have,

$$\widehat{\mathcal{A}}^*\left(\widehat{y}^1\right) = \mathcal{A}^*(y^1) + \tilde{\mathcal{A}}^*(0) = \mathcal{A}^*(y^1) \in \mathbb{S}^n_+ \setminus \{0\}, \tag{4.4.22}$$

and,

$$\left(\widehat{y}^1\right)^T \widehat{b} = \left(y^1\right)^T b = 0. \tag{4.4.23}$$

It follows that each $\widehat{y}^k$ is a suitable choice at each iteration of Algorithm 1 with input $(\widehat{\mathcal{A}}, \widehat{b})$. After the final iteration we obtain an exposing vector for $\operatorname{face}(\mathcal{F})$ which, by (4.4.20), is also an exposing vector for $\operatorname{face}(\widehat{\mathcal{F}})$, as desired.

Now that we have addressed the cases where at least one of $\mathcal{F}$, $\tilde{\mathcal{F}}$, and $\widehat{\mathcal{F}}$ satisfies the Slater condition, we may assume that $\operatorname{sd}(\mathcal{F}) > 0$ and $\operatorname{sd}(\tilde{\mathcal{F}}) > 0$. Let $y^1, \ldots, y^d$ be a facial reduction sequence for input $(\mathcal{A}, b)$ with $d = \operatorname{sd}(\mathcal{F})$ and let $\tilde{y}^1, \ldots, \tilde{y}^{\tilde{d}}$ be a facial reduction sequence for input $(\tilde{\mathcal{A}}, \tilde{b})$ with $\tilde{d} = \operatorname{sd}(\tilde{\mathcal{F}})$. We may assume, without loss of generality, that $d \geq \tilde{d}$. Now let $\widehat{y}^1, \ldots, \widehat{y}^d$ be defined as,

$$\widehat{y}^k := \begin{cases} \begin{pmatrix} y^k \\ \tilde{y}^k \end{pmatrix} & \text{if } k \leq \tilde{d}, \\ \begin{pmatrix} y^k \\ 0 \end{pmatrix} & \text{otherwise,} \end{cases} \qquad \forall k \in \{1, \ldots, d\}. \tag{4.4.24}$$

We claim that a subset of $\widehat{y}^1, \ldots, \widehat{y}^d$ is a facial reduction sequence for input $(\widehat{\mathcal{A}}, \widehat{b})$, implying the desired result. To see this, let $V^1, \ldots, V^d$ be generated by Algorithm 1 with input $(\mathcal{A}, b)$ and the facial reduction sequence $y^1, \ldots, y^d$ and let $\tilde{V}^1, \ldots, \tilde{V}^{\tilde{d}}$ be the analogous matrices for the facial reduction sequence $\tilde{y}^1, \ldots, \tilde{y}^{\tilde{d}}$. Let us consider the first call to the while loop of Algorithm 1 with input $(\widehat{\mathcal{A}}, \widehat{b})$. We have,

$$\widehat{\mathcal{A}}^*(\widehat{y}^1) = \mathcal{A}^*(y^1) + \tilde{\mathcal{A}}^*(\tilde{y}^1) \in \mathbb{S}^n_+ \setminus \{0\}, \tag{4.4.25}$$

and

$$\left(\widehat{y}^1\right)^T \widehat{b} = \left(y^1\right)^T b + \left(\tilde{y}^1\right)^T \tilde{b} = 0.$$

So $\widehat{y}^1$ is a suitable choice for the first iteration of the facial reduction algorithm. Now let $\widehat{V}^1$ capture the nullspace of $\widehat{\mathcal{A}}^*(\widehat{y}^1)$. Since $\mathcal{A}^*(y^1)$ and $\tilde{\mathcal{A}}^*(\tilde{y}^1)$ are both positive semidefinite, we may invoke Fact 2.1.1 *(iii)* and the definitions of V^1 and $\tilde{V}^1$ to get,

$$\begin{aligned}
\operatorname{range}\left(\widehat{V}^1\right) &= \operatorname{null}\left(\widehat{\mathcal{A}}^*(\widehat{y}^1)\right) \\
&= \operatorname{null}\left(\mathcal{A}^*(y^1)\right) \cap \operatorname{null}\left(\tilde{\mathcal{A}}^*(\tilde{y}^1)\right) \\
&= \operatorname{range}\left(V^1\right) \cap \operatorname{range}\left(\tilde{V}^1\right).
\end{aligned}$$

Consequently,

$$\operatorname{range}\left(\widehat{\mathcal{F}}\right) \subseteq \operatorname{range}\left(\widehat{V}^1\right) = \operatorname{range}\left(V^1\right) \cap \operatorname{range}\left(\tilde{V}^1\right). \tag{4.4.26}$$

The fact that $(V^1)^T \mathcal{A}^*(y^2) V^1$ and $(\tilde{V}^1)^T \tilde{\mathcal{A}}^*(\tilde{y}^2) \tilde{V}^1$ are both positive semidefinite, together with (4.4.26) gives us that,

$$\left(\widehat{V}^1\right)^T \mathcal{A}^*(y^2) \widehat{V}^1 \succeq 0 \text{ and } \left(\widehat{V}^1\right)^T \mathcal{A}^*(\tilde{y}^2) \widehat{V}^1 \succeq 0.$$

Therefore,

$$\left(\widehat{V}^1\right)^T \widehat{\mathcal{A}}^*(\widehat{y}^2) \widehat{V}^1 \succeq 0 \text{ and } \left(\widehat{y}^2\right)^T \widehat{b} = 0. \tag{4.4.27}$$

Now if we let $\widehat{Q}^2_2$ capture the nullspace of $(\widehat{V}^1)^T \widehat{\mathcal{A}}^*(\widehat{y}^2) \widehat{V}^1$ and define,

$$\widehat{V}^2 := \widehat{V}^1 \widehat{Q}^2_2.$$

As for the first iteration of the algorithm, we apply Fact 2.1.1 *(iii)* and recall the definitions of V^2 and $\tilde{V}^2$ to see that,

$$\begin{aligned}
\operatorname{range}\left(\widehat{V}^2\right) &= \operatorname{null}\left(\left(\widehat{V}^1\right)^T \widehat{\mathcal{A}}^*(\widehat{y}^2) \widehat{V}^1\right) \\
&= \operatorname{null}\left(\left(\widehat{V}^1\right)^T \mathcal{A}^*(y^2) \widehat{V}^1\right) \cap \operatorname{null}\left(\left(\widehat{V}^1\right)^T \tilde{\mathcal{A}}^*(\tilde{y}^2) \widehat{V}^1\right) \\
&\subseteq \operatorname{null}\left(\left(V^1\right)^T \mathcal{A}^*(y^2) V^1\right) \cap \operatorname{null}\left(\left(\tilde{V}^1\right)^T \tilde{\mathcal{A}}^*(\tilde{y}^2) \tilde{V}^1\right) \\
&= \operatorname{range}\left(V^2\right) \cap \operatorname{range}\left(\tilde{V}^2\right).
\end{aligned}$$

Consequently,

$$\operatorname{range}\left(\widehat{\mathcal{F}}\right) \subseteq \operatorname{range}\left(\widehat{V}^2\right) \subseteq \operatorname{range}\left(V^2\right) \cap \operatorname{range}\left(\tilde{V}^2\right). \tag{4.4.28}$$

We may continue in this fashion until we obtain $\widehat{V}^d$ satisfying,

$$\operatorname{range}\left(\widehat{\mathcal{F}}\right) \subseteq \operatorname{range}\left(\widehat{V}^d\right) \subseteq \operatorname{range}\left(V^d\right) \cap \operatorname{range}\left(\tilde{V}^{\tilde{d}}\right) = \operatorname{range}(\mathcal{F}) \cap \operatorname{range}\left(\tilde{\mathcal{F}}\right). \tag{4.4.29}$$

Now the hypothesis, $\operatorname{face}(\widehat{\mathcal{F}}) = \operatorname{face}(\mathcal{F}) \cap \operatorname{face}(\tilde{\mathcal{F}})$, implies that,

$$\operatorname{range}\left(\widehat{\mathcal{F}}\right) = \operatorname{range}(\mathcal{F}) \cap \operatorname{range}\left(\tilde{\mathcal{F}}\right). \tag{4.4.30}$$

Thus (4.4.29) and (4.4.30) together imply the desired result. □

We conclude this section by addressing a special case that proves useful later in this thesis. Some algorithms used to solve SDPs are designed for instances where the optimal set is bounded. To this end we show that an unbounded spectrahedron can be made bounded without altering the singularity degree.

Theorem 4.4.9. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a non-empty and unbounded spectrahedron and let $\bar{X} \in \operatorname{relint}(\mathcal{F})$. Define $\widehat{\mathcal{F}} := \mathcal{F}(\widehat{\mathcal{A}}, \widehat{b})$ where,*

$$\widehat{\mathcal{A}}(X) := \begin{pmatrix} \mathcal{A}(X) \\ \langle I, X \rangle \end{pmatrix}, \ \widehat{b} := \begin{pmatrix} b \\ \operatorname{trace}\left(\bar{X}\right) \end{pmatrix}.$$

Then $\widehat{\mathcal{F}}$ is bounded and $\operatorname{sd}(\widehat{\mathcal{F}}) = \operatorname{sd}(\mathcal{F})$.

Proof. Clearly $\widehat{\mathcal{F}}$ is bounded, as it is the restriction of $\mathcal{F}$ to those matrices X that satisfy $\operatorname{trace}(X) = \operatorname{trace}(\bar{X})$. In particular, $\bar{X} \in \operatorname{relint}(\mathcal{F}) \cap \operatorname{relint}(\widehat{\mathcal{F}})$ and therefore,

$$\operatorname{face}\left(\widehat{\mathcal{F}}\right) = \operatorname{face}(\mathcal{F}).$$

By Lemma 4.4.7 it holds that $\operatorname{sd}(\widehat{\mathcal{F}}) \leq \operatorname{sd}(\mathcal{F})$. Taking a similar approach to that of (4.4.16) we may assume that,

$$\operatorname{face}(\mathcal{F}) = \begin{bmatrix} \mathbb{S}^r_+ & 0 \\ 0 & 0 \end{bmatrix}.$$

Now let $y^1, \ldots, y^d$ be a facial reduction sequence for input $(\mathcal{A}, b)$ such that,

$$\left(\mathcal{A}^{k-1}\right)^*\left(y^k\right) \in \operatorname{relint}\left(\mathcal{E}\left(\mathcal{A}^{k-1}, b\right)\right), \ \forall k \in \{1, \ldots, d\}.$$

Then by Theorem 4.2.3 this is a minimum length facial reduction sequence. Hence $d = \operatorname{sd}(\mathcal{F})$. We claim that $\widehat{y}^1, \ldots, \widehat{y}^d$ where,

$$\widehat{y}^k := \begin{pmatrix} y^k \\ 0 \end{pmatrix}, \ \forall k \in \{1, \ldots, d\},$$

is a minimum length facial reduction sequence for input $(\widehat{\mathcal{A}}, \widehat{b})$. To this end, we show that,

$$\left(\widehat{\mathcal{A}}^{k-1}\right)^* (\widehat{y}^k) \in \operatorname{relint}\left(\mathcal{E}\left(\widehat{\mathcal{A}}^{k-1}, \widehat{b}\right)\right), \ \forall k \in \{1, \ldots, d\}. \tag{4.4.31}$$

Suppose for the sake of contradiction that there exists a smallest integer $\bar{k}$ such that (4.4.31) fails. We may assume, as in (4.4.16), that,

$$V^{\bar{k}-1} = \begin{bmatrix} I \\ 0 \end{bmatrix}.$$

Then,

$$\widehat{\mathcal{A}}^{\bar{k}-1} = \widehat{\mathcal{A}}_{V^{\bar{k}-1}} = \begin{pmatrix} \widehat{\mathcal{A}}_{V^{\bar{k}-1}} \\ \langle \left(V^{\bar{k}-1}\right)^T V^{\bar{k}-1}, \cdot \rangle \end{pmatrix} = \begin{pmatrix} \widehat{\mathcal{A}}^{\bar{k}-1} \\ \langle I, \cdot \rangle \end{pmatrix}.$$

Now suppose $\widehat{y} \in \operatorname{relint}(\mathcal{E}(\widehat{\mathcal{A}}^{\bar{k}-1}, \widehat{b}))$ where $\widehat{y} = \begin{pmatrix} y^T & \gamma \end{pmatrix}^T$. Then it holds that,

$$\begin{cases} \left(\mathcal{A}^{\bar{k}-1}\right)^* (y) + \gamma I \succeq 0, \\ y^T b + \gamma \operatorname{trace}(\bar{X}) = 0, \\ \gamma \neq 0. \end{cases}$$

Note that if $\gamma = 0$ then $y \in \operatorname{relint}(\mathcal{E}(\mathcal{A}^{\bar{k}-1}, b))$ and therefore,

$$\begin{aligned} \operatorname{rank}\left(\left(\widehat{\mathcal{A}}^{\bar{k}-1}\right)^* (\widehat{y})\right) &= \operatorname{rank}\left(\left(\mathcal{A}^{\bar{k}-1}\right)^* (y)\right) \\ &\leq \operatorname{rank}\left(\left(\mathcal{A}^{\bar{k}-1}\right)^* (y^{\bar{k}})\right) \\ &= \operatorname{rank}\left(\left(\widehat{\mathcal{A}}^{\bar{k}-1}\right)^* (\widehat{y}^{\bar{k}})\right), \end{aligned} \tag{4.4.32}$$

a contradiction of the assumption that (4.4.31) does not hold for $\bar{k}$. Now we claim that $\gamma > 0$. Indeed, if $\gamma < 0$ then $(\mathcal{A}^{\bar{k}-1})^*(y) \succ 0$ and we arrive at a contradiction as in (4.4.32). Moreover, we may rescale, if necessary, so that $\gamma = 1$. It follows that,

$$y^T b = -\operatorname{trace}(\bar{X}). \tag{4.4.33}$$

Now we may assume that,

$$\operatorname{range}\left(\left(\widehat{\mathcal{A}}^{\bar{k}-1}\right)^* (\widehat{y})\right) = \operatorname{range}\left(\begin{bmatrix} 0 \\ I \end{bmatrix}\right).$$

Then,

$$\left(\widehat{\mathcal{A}}^{\bar{k}-1}\right)^* (\widehat{y}) = \left(\mathcal{A}^{\bar{k}-1}\right)^* (y) + I = \begin{bmatrix} 0 & 0 \\ 0 & * \end{bmatrix},$$

where the upper left block of zeros is $r \times r$. It follows that,

$$\left(\mathcal{A}^{\bar{k}-1}\right)^*(y) = \begin{bmatrix} -I_r & 0 \\ 0 & * \end{bmatrix} \text{ and } \mathcal{A}^*(y) = \begin{bmatrix} -I_r & 0 & * \\ 0 & * & * \\ * & * & * \end{bmatrix}. \tag{4.4.34}$$

Here I_r denotes the $r \times r$ identity. Since we have assumed that $\mathcal{F}$ is unbounded, Lemma 2.2.3 ensures the existence of non-zero $X \in \text{null}(\mathcal{A}) \cap \mathbb{S}^n_+$. By our assumption on the facial structure of $\mathcal{F}$ and by (4.4.34) it holds that,

$$0 = \langle X, \mathcal{A}^*(y) \rangle = \left\langle \begin{bmatrix} X_{11} & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} -I_r & 0 & * \\ 0 & * & * \\ * & * & * \end{bmatrix} \right\rangle = \langle X_{11}, -I \rangle < 0,$$

a contradiction. Thus (4.4.31) holds for all $k \in \{1, \ldots, d\}$. Consequently $\widehat{y}^1, \ldots, \widehat{y}^d$ is a minimum length facial reduction sequence for input $(\widehat{\mathcal{A}}, \widehat{b})$. Therefore,

$$\text{sd}(\widehat{\mathcal{F}}) = d = \text{sd}(\mathcal{F}),$$

as desired. □

4.5 Singularity Degree and Complementary Slackness

We have seen that $\text{sd}(\mathcal{F}) = 0$ is quite a special case as it corresponds to the Slater condition. Another special case, pointed out in [10], is that of $\text{sd}(\mathcal{F}) = 1$. It is shown that instances of SDP where the feasible set has singularity degree 1 are *backward stable*. That is, an approximate solution to the problem is in fact an exact solution to a perturbation of the original problem. Here we show that the case $\text{sd}(\mathcal{F}) = 1$ corresponds to another special case in the SDP literature, that of *strict complementarity*.

Recall the SDP problem from Section 2.6,

$$(SDP) \qquad \begin{aligned} p^\star := \inf\ & \langle C, X \rangle, \\ \text{s.t. } & \mathcal{A}(X) = b, \\ & X \succeq 0, \end{aligned}$$

and its dual,

$$(DSDP) \qquad \begin{aligned} d^\star := \sup\ & b^T y, \\ \text{s.t. } & \mathcal{A}^*(y) + Z = C, \\ & Z \succeq 0. \end{aligned}$$

Let us denote the optimal sets of (SDP) and $(DSDP)$ as $\mathcal{P}$ and $\mathcal{D}$, respectively. We write the sets out explicitly as,

$$\mathcal{P} = \{X \in \mathbb{S}^n_+ : \mathcal{A}(X) = b,\ \langle C, X \rangle = p^\star\},$$

and,

$$\mathcal{D} = \{ Z \in \mathbb{S}^n_+ : Z = C - \mathcal{A}^*(y),\ b^T y = d^\star,\ y \in \mathbb{R}^m \}.$$

Note that $\mathcal{P}$ is a spectrahedron, when $p^\star$ is finite, and it is defined by $(\widehat{\mathcal{A}}, \widehat{b})$ where,

$$\widehat{\mathcal{A}}(X) := \begin{pmatrix} \mathcal{A}(X) \\ \langle C, X \rangle \end{pmatrix} \text{ and } \widehat{b} := \begin{pmatrix} b \\ p^\star \end{pmatrix}.$$

We begin with an elementary result regarding primal-dual optimality.

Lemma 4.5.1. *Let $\mathcal{P}$ and $\mathcal{D}$ be the non-empty optimal sets of (SDP) and $(DSDP)$, respectively. Then*

$$p^\star = d^\star \iff \langle X^\star, Z^\star \rangle = 0,\ \forall X^\star \in \mathcal{P},\ \forall Z^\star \in \mathcal{D}.$$

Proof. Let $X^\star \in \mathcal{P}$ and $Z^\star \in \mathcal{D}$. Then there exists $y^\star \in \mathbb{R}^m$ such that $\mathcal{A}^*(y^\star) + Z^\star = C$. It follows that,

$$\begin{aligned} p^\star = d^\star &\iff \langle C, X^\star \rangle = (y^\star)^T b \\ &\iff \langle C, X^\star \rangle = (y^\star)^T \mathcal{A}(X^\star) \\ &\iff \langle C, X^\star \rangle = \langle \mathcal{A}^*(y^\star), X^\star \rangle \\ &\iff \langle C, X^\star \rangle = \langle C - Z^\star, X^\star \rangle \\ &\iff \langle X^\star, Z^\star \rangle = 0, \end{aligned}$$

as desired. □

The property $\langle X^\star, Z^\star \rangle = 0$ is referred to as *complementary slackness*. We define a special case of complementary slackness in the following.

Definition 4.5.2. *Let $\mathcal{P}$ and $\mathcal{D}$ be the optimal sets of (SDP) and $(DSDP)$, respectively. We say that strict complementarity holds for (SDP) and $(DSDP)$ if $p^\star = d^\star$ and there exists $X^\star \in \mathcal{P}$ and $Z^\star \in \mathcal{D}$ such that* $\operatorname{rank}(X^\star) + \operatorname{rank}(Z^\star) = n$.

Now we describe the relationship between strict complementarity of (SDP) and $(DSDP)$ and the singularity degree of $\mathcal{P}$.

Theorem 4.5.3. *Let $\mathcal{P}$ and $\mathcal{D}$ be the optimal sets of (SDP) and $(DSDP)$, respectively. Suppose that $\mathcal{P}$ and $\mathcal{D}$ are both non-empty, $\mathcal{P}$ does not have a Slater point, and $p^\star = d^\star$. Then strict complementarity holds for (SDP) and $(DSDP)$ if, and only if,* $\operatorname{sd}(\mathcal{P}) = 1$.

Proof. By hypothesis there exist $X^\star \in \operatorname{relint}(\mathcal{P})$ and $Z^\star \in \operatorname{relint}(\mathcal{D})$. Let $y^\star \in \mathbb{R}^m$ such that $Z^\star = C - \mathcal{A}^*(y^\star)$. Suppose strict complementarity holds for (SDP) and $(DSDP)$. By Definition 4.5.2 it follows that $\operatorname{rank}(X^\star) + \operatorname{rank}(Z^\star) = n$. Since we have assumed that $\mathcal{P}$

does not have a Slater point, the while loop of the facial reduction algorithm is called at least once. We show that,

$$\widehat{y} := \begin{pmatrix} -y^\star \\ 1 \end{pmatrix}$$

is feasible for the first iteration of the facial reduction algorithm with input $(\widehat{\mathcal{A}}, \widehat{b})$ and that $\mathcal{A}^*(\widehat{y})$ exposes face$(\mathcal{P})$. First of all,

$$\widehat{\mathcal{A}}^*(\widehat{y}) = -\mathcal{A}^*(y^\star) + C = Z^\star \succeq 0. \tag{4.5.1}$$

Moreover, $Z^\star \neq 0$ since we have assumed that $X^\star$ and $Z^\star$ are strictly complementary and $X^\star$ is not positive definite. Secondly, applying Lemma 4.5.1 we have,

$$(\widehat{y})^T \widehat{b} = -(y^\star)^T b + p^\star = -\langle \mathcal{A}^*(y^\star), X^\star \rangle + \langle X^\star, C \rangle = \langle X^\star, Z^\star \rangle = 0.$$

We have, thus shown that $\widehat{y}$ is a suitable choice for the first iteration of the facial reduction algorithm with input $(\widehat{\mathcal{A}}, \widehat{b})$. Moreover, the strict complementarity assumption and Lemma 4.5.1 imply that $Z^\star$ is an exposing vector for face$(\mathcal{P})$. It follows, by (4.5.1) that $\widehat{\mathcal{A}}^*(\widehat{y})$ exposes face$(\mathcal{P})$ and therefore $\mathrm{sd}(\mathcal{P}) = 1$, as desired.

Now for the converse, suppose that $\mathrm{sd}(\mathcal{P}) = 1$. By definition, there exists $\widehat{y}$ such that,

$$\widehat{\mathcal{A}}^*(\widehat{y}) \in \mathbb{S}^n_+ \setminus \{0\} \text{ and } \widehat{y}^T \widehat{b} = 0, \tag{4.5.2}$$

and $\widehat{\mathcal{A}}^*(\widehat{y})$ exposes face$(\mathcal{P})$. This implies that for $X^\star \in \mathrm{relint}(\mathcal{P})$ we have,

$$\mathrm{rank}\left(\widehat{\mathcal{A}}^*(\widehat{y})\right) + \mathrm{rank}(X^\star) = n \text{ and } \langle \widehat{\mathcal{A}}^*(\widehat{y}), X^\star \rangle = 0. \tag{4.5.3}$$

If we write $\widehat{y} = \begin{pmatrix} y^T & \gamma \end{pmatrix}^T$ for $y \in \mathbb{R}^m$ and $\gamma \in \mathbb{R}$ we may summarize (4.5.2) and (4.5.3) as,

$$\begin{cases} \mathcal{A}^*(y) + \gamma C \in \mathbb{S}^n_+ \setminus \{0\}, \\ y^T b + \gamma p^\star = 0, \\ \mathrm{rank}\left(\mathcal{A}^*(y) + \gamma C\right) + \mathrm{rank}(X^\star) = n, \\ \langle \mathcal{A}^*(y) + \gamma C, X^\star \rangle = 0. \end{cases} \tag{4.5.4}$$

Now we consider three cases. First, suppose $\gamma > 0$. Then let us define,

$$Z := \frac{1}{\gamma}\left(\mathcal{A}^*(y) + \gamma C\right) = C - \mathcal{A}^*\left(-\frac{1}{\gamma} y\right).$$

Then the first equation of (4.5.4) implies that $Z \succeq 0$ and therefore Z is feasible for $(DSDP)$. Moreover, by the second equation in (4.5.4) we have,

$$\left(-\frac{1}{\gamma} y\right)^T b = -\frac{1}{\gamma} y^T b = -\frac{1}{\gamma} \gamma p^\star = p^\star.$$

Therefore $Z \in \mathcal{D}$. Now the third and fourth equations together imply that strict complementarity holds for (SDP) and $(DSDP)$, as desired.

For the second case suppose $\gamma = 0$. In this case (4.5.4) simplifies to

$$\begin{cases} \mathcal{A}^*(y) \in \mathbb{S}^n_+ \setminus \{0\}, \\ y^T b = 0, \\ \operatorname{rank}(\mathcal{A}^*(y)) + \operatorname{rank}(X^\star) = n, \\ \langle \mathcal{A}^*(y), X^\star \rangle = 0. \end{cases} \tag{4.5.5}$$

Now let $Z^\star = C - \mathcal{A}^*(y^\star) \in \mathcal{D}$ and define,

$$Z := Z^\star + \mathcal{A}^*(y) = C - \mathcal{A}^*(y^\star - y).$$

Note that $Z \succeq 0$ as it is the sum of two positive semidefinite matrices. Moreover $Z \in \mathcal{D}$ since,

$$(y^\star - y)^T b = (y^\star)^T b = d^\star.$$

It follows by Lemma 4.5.1 that complementary slackness holds for Z and $X^\star$. Thus in particular,

$$\operatorname{rank}(Z) \leq n - \operatorname{rank}(X^\star). \tag{4.5.6}$$

On the other hand, invoking Fact 2.1.1 and (4.5.5) we have,

$$\operatorname{rank}(Z) \geq \operatorname{rank}(\mathcal{A}^*(y)) = n - \operatorname{rank}(X^\star). \tag{4.5.7}$$

Now (4.5.6) and (4.5.7) yield the desired result.

For the third, and final case, we assume that $\gamma < 0$. As above, let $Z^\star = C - \mathcal{A}^*(y^\star) \in \mathcal{D}$ and define,

$$Z := 2Z^\star + \frac{-1}{\gamma}\left(\mathcal{A}^*(y) + \gamma C\right) = C - \mathcal{A}^*\left(2y^\star + \frac{1}{\gamma} y\right).$$

Then by arguments similar to those used in the case $\gamma = 0$, it holds that $Z \in \mathcal{D}$ and that strict complementarity holds for (SDP) and $(DSDP)$, as desired. □

One application of this theorem is in constructing instances of SDP with specified singularity degree. Let us briefly describe the approach. In [85], the authors present an approach for constructing SDPs with specified *complementarity gap*, defined as,

$$g(SDP) := n - \operatorname{rank}(\mathcal{P}) - \operatorname{rank}(\mathcal{D}).$$

Assuming that the slater condition fails for SDP, the case $g = 0$ corresponds to strict complementarity and Theorem 4.5.3 tells us that $\operatorname{sd}(\mathcal{P}) = 1$. Alternatively setting $g(SDP) > 0$ yields $\operatorname{sd}(\mathcal{P}) > 0$. The code produced by [85], may therefore be modified to produce spectrahedra that satisfy $\operatorname{sd}(\mathcal{F}) = 1$, as well as, spectrahedra that satisfy $\operatorname{sd}(\mathcal{F}) \geq 2$.

4.6 Singularity Degree and Error Bounds in the Literature

Due to the connection between singularity degree and error bounds, we find it natural to discuss the literature on these topics at the same time. For a more general treatment of error bounds than presented thus far, let us consider a set $\mathcal{S} \in \mathbb{S}^n$ and a residual function $R : \mathbb{S}^n \to \mathbb{R}_+$. It is assumed that $R(X)$ is easily computable and we view this function as a proxy for the forward error of X relative to $\mathcal{S}$. A Hölder error bound for $\mathcal{S}$ in terms of R holds if there exists $\gamma \in (0, 1]$ such that,

$$\operatorname{dist}(X, \mathcal{S}) = \mathcal{O}\left(R(X)^\gamma\right), \quad \forall X \in \mathbb{S}^n.$$

Our ambient space is $\mathbb{S}^n$, but more general spaces are considered in the literature. The bound is referred to as *Lipschitzian* when $\gamma = 1$.

A classical result of Hoffman [35] states that when $\mathcal{S}$ is a polyhedron, i.e., a spectrahedron where the positive semidefinite constraint is replaced by non-negativity, and when R is the non-negativity analogue of ϵ^b, then a Lipschitzian error bound exists. Various extensions to general convex $\mathcal{S}$ and even non-convex $\mathcal{S}$ have subsequently been derived. We suggest Lewis and Pang [44] and Pang [60], for an overview of such results.

For spectrahedra, Fact 4.3.1 guarantees a Lipschitzian error bound along the parametric curve $\{X(\alpha) : \alpha > 0\}$ whenever the Slater condition holds. Deng and Hu [15] and Azé and Hiriart-Urruty [2] show that Lipschitzian error bounds exist for spectrahedra that satisfy the Slater condition independent of a curve. In each of these papers an additional assumption, such as boundedness, is made.

Luo, Sturm, and Zhang [50] show that along the classical central path of SDP a Lipschitzian error bound exists when strict complementarity holds. By Theorem 4.5.3 this corresponds to spectrahedra with singularity degree 1. This special case is better than the bound of Sturm, Fact 4.3.1, for general spectrahedra with singularity degree 1. Error bounds along the central path are also studied by Chua [11], where a Lipschitzian error bound is shown to exist under the assumption of strict complementarity. This result differs from that of Luo, Sturm, and Zhang in that Chua takes R to be the duality gap between the primal and dual central paths. More recently, Mohammad-Nezhad and Terlaky [55] use the bounds of Sturm to derive Hölder bounds for the *optimal partition* along the central path. Their approach bares some resemblance to our approach and we address this more specifically in Chapter 6.

In [63], Pataki introduces two families of SDPs that have positive duality gaps and proves some interesting results. He shows that the singularity degree of the feasible set of these SDPs is $m - 1$, where m denotes the number of constraints defining $\mathcal{A}$, as in our setting. For the problems he considers, $m \in \{n - 2, n/2\}$. He also proves, for general SDPs, that if the singularity degree of the feasible set is m, then the duality gap is zero. Drusvyastkiy, Li, and Wolkowicz [17] use the bounds of Sturm to derive error bounds

for iterates of the alternating projections algorithm applied to spectrahedra. They also prove that problems with singularity degree 1 are generic. In [46], Lourenço introduces a remarkable generalization of the error bounds of Sturm to a new class of closed convex cones, that he calls *amenable cones*.

Chapter 5

Bounds on Forward Error and Singularity Degree

The theoretical upper bound on forward error, as stated in Fact 4.3.1, is useful in identifying classes of spectrahedra for which forward error is not too much larger than backward error. For instance, when $\mathrm{sd}(\mathcal{F}) = 0$ or $\mathcal{F} = \{0\}$ we can be sure that forward error is of the same magnitude as backward error. However, upper bounds alone can not be used to detect instances where forward error is much larger than backward error. In this chapter we present an algorithm to obtain a lower bound for forward error. The results of Section 5.1.1 and Section 5.2 are based largely on the preprint [74], coauthored by the author of this thesis.

To avoid well-behaved scenarios, we make the following assumption.

Assumption 5.0.1. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a non-empty spectrahedron with $\mathrm{sd}(\mathcal{F}) \geq 1$ and $\mathcal{F} \neq \{0\}$.*

Our analysis in this section is based on *path-following* algorithms. The foundation of such algorithms is the *central path*, a smooth parametric curve, say $\{X(\alpha) : \alpha > 0\}$, that is known to converge to an element of $\mathcal{F}$. Specifically we mean that,

$$\lim_{\alpha \searrow 0} X(\alpha) = \bar{X} \in \mathcal{F}. \tag{5.0.1}$$

A path-following algorithm is used to produce a sequence of positive numbers $\{\alpha_k\}$ and matrices $\{X^k\}$ such that α_k is successively closer to 0 and X^k is a successively better approximation of $X(\alpha_k)$. In other words, the iterates $\{X^k\}$ approach $\mathcal{F}$ along a trajectory that approximates the central path. We make the following minimal assumptions on the central path.

Assumption 5.0.2. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a spectrahedron satisfying Assumption 5.0.1 and let $\{X(\alpha) : \alpha > 0\}$ be a central path with limit point $\bar{X}$. We assume that,*

(*i*) $X(\alpha) \succ 0$ *for all* $\alpha > 0$,

(*ii*) $\bar{X} \in \text{relint}(\mathcal{F})$.

Many of the well-known algorithms for SDP are based on central paths that satisfy this assumption. A key component to our approach is estimating the rank of the limit point $\bar{X}$. We have devoted Section 5.1 to this discussion. In Section 5.2, we use the bound on rank to obtain lower bounds on forward error and singularity degree.

5.1 A Bound on Maximum Rank

Smoothness of the central path $\{X(\alpha) : \alpha > 0\}$ implies that smooth functions of the central path are also smooth. In particular the eigenvalue function $\lambda_i(X(\alpha))$, with $i \in \{1, \ldots, n\}$, is contiuous with limit point $\lambda_i(\bar{X})$. Therefore, in order to determine the rank of $\bar{X}$ it suffices to find all indices i for which $\lambda_i(X(\alpha))$ is bounded away from 0. The challenge with this approach is that when $\lambda_i(X(\alpha))$ is small, it is not clear whether the limit point is a small positive number or 0. An upper bound on rank may be obtained by the largest index i for which it is not clear that $\lambda_i(X(\alpha))$ converges to 0. This bound may be quite poor. For this reason we consider compositions of eigenvalue functions that amplify the difference between those eigenvalues that converge to 0 and those that do not.

Our first approach is the *ratio of subsequent eigenvalues*,

$$R_i(\alpha) := \frac{\lambda_i(X(\alpha))}{\lambda_{i+1}(X(\alpha))}, \ i \in \{1, \ldots, n-1\}. \tag{5.1.1}$$

Assumption 5.0.2 ensures that $R_i(\alpha)$ is well-defined for every i and $\alpha > 0$. The ratios that blow up indicate one of two scenarios. The first scenario is that both eigenvalues converge to 0, but $\lambda_{i+1}(X(\alpha))$ does so much more quickly. The second is that $\lambda_i(X(\alpha))$ converges to a positive value and $\lambda_{i+1}(X(\alpha))$ vanishes. This only happens when i corresponds to the rank of the limit point $\bar{X}$. We state this observation formally in the following.

Lemma 5.1.1. *Let* $\{X(\alpha) : \alpha > 0\}$ *be a central path satisfying Assumption 5.0.2 for a spectrahedron* $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ *satisfying Assumption 5.0.1. Let* $i \in \{1, \ldots, n-1\}$ *be the smallest integer for which* $R_i(\alpha) \to +\infty$. *Then* $\text{rank}(\bar{X}) = i$.

Analysis of $R_i(\alpha)$ presents similar challenges to analysis of the eigenvalue functions, in that it may be difficult to determine whether $R_i(\alpha)$ converges to a large number or blows up. We may also encounter the case that $R_i(\alpha)$ blows up slowly so that it appears bounded. As for the eigenvalue functions, an upper bound on $\text{rank}(\bar{X})$ may be obtained by the smallest index i for which it is clear that the ratio blows up.

Having addressed the challenges of differentiating between 0 and small numbers, as well as infinity and large numbers, it would be desirable to construct a measure, say $r_i(\alpha)$,

for which there is a positive difference (for α sufficiently small) between $r_i(\alpha)$ and $r_j(\alpha)$ whenever $i \leq \operatorname{rank}(\bar{X})$ and $j > \operatorname{rank}(\bar{X})$. In other words, there exist real numbers τ_1 and τ_2 such that τ_1 is 'discernibly larger' than τ_2 and for α sufficiently small it holds that,

$$r_i(\alpha) = \begin{cases} \geq \tau_1 & \text{if } i \leq \operatorname{rank}(\bar{X}), \\ \leq \tau_2 & \text{otherwise.} \end{cases} \tag{5.1.2}$$

The numerical challenge is lessened under such a measure since we are asked to differentiate between two finite numbers that are sufficiently different from each other. In Section 5.1.1 and in Section 5.1.2 we present two measures that satisfy (5.1.2) morally. Our approach is motivated, in part, by the Tapia indicator [22, 78], used to identify zero variables in linear program.

5.1.1 Eigenvalue Q-Convergence Ratio

Let $\sigma \in (0,1)$ and let $\{\sigma^k\}_{k\in\mathbb{N}}$ be a sequence of powers of σ. Then we define the *eigenvalue Q-convergence ratio* as,

$$Q_{i,\sigma}(k) := \frac{\lambda_i\left(X\left(\sigma^{k+1}\right)\right)}{\lambda_i\left(X\left(\sigma^k\right)\right)}, \ i \in \{1, \ldots, n\}. \tag{5.1.3}$$

Referring to $Q_{i,\sigma}$ as the Q-convergence ratio is somewhat abusive of the term. When $\lambda_i(\alpha)$ converges to 0, $Q_{i,\sigma}$ is the ratio that defines the so-called Q-convergence rate, see for instance [59]. But when $\lambda_i(\alpha)$ converges to a positive number, $Q_{i,\sigma}$ is not the ratio for Q-convergence. Since $Q_{i,\sigma}$ is not a function of α, it does not posses all of the properties of (5.1.2). However, we will show that it satisfies (for the most part) the sequence analogue of (5.1.2). To analyze $Q_{i,\sigma}$, we begin by translating Fact 4.3.2 into a statement about the eigenvalues of $X(\alpha)$.

Lemma 5.1.2. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a spectrahedron that satisfies Assumption 5.0.1 and let $\{X(\alpha) : \alpha > 0\}$ be a central path for $\mathcal{F}$ that satisfies Assumption 5.0.2 and suppose that $\epsilon^b(X(\alpha), \mathcal{F}) = \mathcal{O}(\alpha)$. Let $Z^1, \ldots, Z^{\operatorname{sd}(\mathcal{F})}$ be generated by Algorithm 1 for input $(\mathcal{A}, b)$ such that each Z^i is chosen to have maximum rank. Let $q_i := \operatorname{rank}(Z^i)$ and let r denote the rank of $\mathcal{F}$. Let $\mathcal{I}^0, \mathcal{I}^1, \ldots, \mathcal{I}^{\operatorname{sd}(\mathcal{F})}$ form a partition of $\{1, \ldots, n\}$ such that,*

$$\mathcal{I}^0 = \{1, \ldots, r\}, \ \mathcal{I}^1 = r + \{1, \ldots, q^1\}, \ \mathcal{I}^2 = r + q^1 + \{1, \ldots, q^2\}, \ldots .$$

Then, for $j \in \{1, \ldots, n\}$ it holds that for sufficiently small $\alpha > 0$,

$$j \in \mathcal{I}^i \implies \lambda_j(X(\alpha)) = \begin{cases} \Theta(1) & \text{if } i = 0, \\ \mathcal{O}\left(\alpha^{\xi(i)}\right) & \text{otherwise,} \end{cases}$$

where $\xi(i) := 2^{-(\operatorname{sd}(\mathcal{F})-i)}$.

Proof. By assumption, $X(\alpha) \to \bar{X} \in \operatorname{relint}(\mathcal{F})$ and $\operatorname{rank}(\bar{X}) = r$. Therefore, the r largest eigenvalues of $X(\alpha)$ converge to positive numbers. It follows that for sufficiently small $\alpha > 0$,

$$j \in \mathcal{I}^0 \implies \lambda_j(X(\alpha)) = \Theta(1),$$

proving one part of the desired result.

Next, by Fact 4.3.2 there exists an orthogonal Q such that,

$$\operatorname{face}(Q\mathcal{F}Q^T) = \begin{bmatrix} \mathbb{S}^r_+ & 0 \\ 0 & 0 \end{bmatrix} \text{ and } QX(\alpha)Q^T = \begin{bmatrix} X_0(\alpha) & * & \cdots & * \\ * & X_1(\alpha) & & \\ \vdots & & \ddots & * \\ * & * & * & X_{\operatorname{sd}(\mathcal{F})}(\alpha) \end{bmatrix} \forall \alpha > 0, \tag{5.1.4}$$

where $X_0(\alpha) \in \mathbb{S}^r$ and for all $i \in \{1, \ldots, \operatorname{sd}(\mathcal{F})\}$ it holds that,

$$X_i(\alpha) \in \mathbb{S}^{q_i} \text{ and } \|X_i(\alpha)\| = \mathcal{O}\left(\alpha^{\xi(i)}\right). \tag{5.1.5}$$

Now let $i \in \{1, \ldots, \operatorname{sd}(\mathcal{F})\}$ and let $j \in \mathcal{I}^i$. Consider the principal submatrix of $QX(\alpha)Q^T$,

$$S(\alpha) := \begin{bmatrix} X_i(\alpha) & \cdots & * \\ \vdots & \ddots & \vdots \\ * & \cdots & X_{\operatorname{sd}(\mathcal{F})}(\alpha) \end{bmatrix}. \tag{5.1.6}$$

By Assumption 5.0.2 it holds that $X(\alpha) \succ 0$ and therefore $S(\alpha) \succ 0$. Thus by (5.1.5) we have,

$$\|S(\alpha)\| = \mathcal{O}\left(\max_{\ell \in \{i, \ldots, \operatorname{sd}(\mathcal{F})} \|X_\ell(\alpha)\|\right) = \mathcal{O}\left(\alpha^{\xi(i)}\right). \tag{5.1.7}$$

Then by interlacing eigenvalues (Fact 2.1.2) and by (5.1.7) we have,

$$\lambda_j(X(\alpha)) \leq \lambda_1(S(\alpha)) = \mathcal{O}\left(\|S(\alpha)\|\right) = \mathcal{O}\left(\alpha^{\xi(i)}\right),$$

as desired. □

In this result we have obtained, for each eigenvalue function $\lambda_j(X(\alpha))$, a corresponding upper bound function $\alpha^{\xi(i)}$. This upper bound function is simple in form and it is easy to compute the rate of Q-convergence of this function along the sequence $\{\sigma^k\}$. Indeed, the rate is constant and evaluated as,

$$\frac{\left(\sigma^{k+1}\right)^{\xi(i)}}{\left(\sigma^k\right)^{\xi(i)}} = \sigma^{\xi(i)}.$$

While the Q-convergence rate of the eigenvalue function itself, captured by the limit of $Q_{i,\sigma}(k)$, may not be easy to compute, the following technical lemma allows us to relate the two.

Lemma 5.1.3. *Let $\{a_k\}_{k\in\mathbb{N}}$ and $\{b_k\}_{k\in\mathbb{N}}$ be sequences of positive reals such that $a_k \to 0$ and $b_k \to 0$. If $a_k \leq b_k$ for all $k \in \mathbb{N}$ then,*

$$\liminf_{k\to\infty} \frac{a_{k+1}}{a_k} \leq \limsup_{k\to\infty} \frac{b_{k+1}}{b_k}. \tag{5.1.8}$$

Proof. Let L_a and L_b denote the limit inferior and limit superior of (5.1.8), respectively. For simplicity we assume that L_a and L_b are finite, but the arguments extend to the general case trivially. Suppose for the sake of contradiction that there exists $\tau > 0$ such that $L_a - \tau \geq L_b$. Then there exists $\bar{k} \in \mathbb{N}$ such that for all $k \geq \bar{k}$,

$$\frac{a_{k+1}}{a_k} \geq L_a - \frac{\tau}{3} \text{ and } \frac{b_{k+1}}{b_k} \leq L_a - \frac{\tau}{2} \tag{5.1.9}$$

Rearranging the first equation in (5.1.9) gives us,

$$a_{k+1} \geq a_k \left(L_a - \frac{\tau}{3}\right), \quad \forall k \geq \bar{k}. \tag{5.1.10}$$

Replacing k with $k-1$ we get that,

$$a_k \geq a_{k-1} \left(L_a - \frac{\tau}{3}\right), \quad \forall k \geq \bar{k}+1. \tag{5.1.11}$$

Combining (5.1.10) with (5.1.11) yields,

$$a_{k+1} \geq a_{k-1} \left(L_a - \frac{\tau}{3}\right)^2, \quad \forall k \geq \bar{k}+1.$$

Continuing in this fashion we get,

$$a_k \geq a_{\bar{k}} \left(L_a - \frac{\tau}{3}\right)^{k-\bar{k}} = \frac{a_{\bar{k}}}{\left(L_a - \frac{\tau}{3}\right)^{\bar{k}}} \left(L_a - \frac{\tau}{3}\right)^k, \quad \forall k \geq \bar{k}. \tag{5.1.12}$$

Through an analogous approach applied to the second equation of (5.1.9) we get,

$$b_k \leq \frac{b_{\bar{k}}}{\left(L_a - \frac{\tau}{2}\right)^{\bar{k}}} \left(L_a - \frac{\tau}{2}\right)^k, \quad \forall k \geq \bar{k}. \tag{5.1.13}$$

Combining the hypothesis that b_k dominates a_k with (5.1.12) and (5.1.13) we get,

$$\frac{b_{\bar{k}}}{\left(L_a - \frac{\tau}{2}\right)^{\bar{k}}} \left(L_a - \frac{\tau}{2}\right)^k \geq \frac{a_{\bar{k}}}{\left(L_a - \frac{\tau}{3}\right)^{\bar{k}}} \left(L_a - \frac{\tau}{3}\right)^k, \quad \forall k \geq \bar{k}. \tag{5.1.14}$$

Observe that $L_b \geq 0$ since $b_k \geq 0$ for every $k \in \mathbb{N}$. Therefore, $L_a - \tau \geq 0$ and we have,

$$L_a - \frac{\tau}{3} > L_a - \frac{\tau}{2} > 0.$$

It follows that for sufficiently large k, the inequality in (5.1.14) is violated, giving us the desired contradiction. □

Now we are ready to state the main result of this section.

Theorem 5.1.4. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a spectrahedron that satisfies Assumption 5.0.1 and let $\{X(\alpha) : \alpha > 0\}$ be a central path for $\mathcal{F}$ that satisfies Assumption 5.0.2 and suppose that $\epsilon^b(X(\alpha), \mathcal{F}) = \mathcal{O}(\alpha)$. Let $\mathcal{I}^0, \mathcal{I}^1, \ldots, \mathcal{I}^{\mathrm{sd}(\mathcal{F})}$ be a partition of $\{1, \ldots, n\}$ as constructed in Lemma 5.1.2. Let $\sigma \in (0,1)$ and let $Q_{j,\sigma}(k)$ be as in (5.1.3). Then the following hold.*

(i) If $j \in \mathcal{I}^0$ then,

$$\lim_{k\to\infty} Q_{j,\sigma}(k) = 1.$$

(ii) If $j \in \mathcal{I}^i$ with $i \in \{1, \ldots, \mathrm{sd}(\mathcal{F})\}$ then,

$$\liminf_{k\to\infty} Q_{j,\sigma}(k) \leq \sigma^{\xi(i)} < 1,$$

where $\xi(i) := 2^{-(\mathrm{sd}(\mathcal{F})-i)}$.

Proof. By Assumption 5.0.2 and the definition of $\mathcal{I}^0$ we have that $\lambda_j(X(\sigma^k))$ converges, in k, to a positive number whenever $j \in \mathcal{I}^0$. The proof of (i) follows immediately.

Now let $j \in \mathcal{I}^i$ with $i \in \{1, \ldots, \mathrm{sd}(\mathcal{F})\}$. By Lemma 5.1.2 we have,

$$\lambda_j(X(\sigma^k)) = \mathcal{O}\left(\left(\sigma^k\right)^{\xi(i)}\right), \ \forall k \in \mathbb{N}. \tag{5.1.15}$$

Thus there exists $M > 0$ such that $\lambda_j(X(\sigma^k)) \leq M\sigma^{k\xi(i)}$. Now the sequences $\{\lambda_j(X(\sigma^k))\}_{k\in\mathbb{N}}$ and $\{M\sigma^{k\xi(i)}\}_{k\in\mathbb{N}}$ satisfy the assumptions of Lemma 5.1.3. Therefore,

$$\liminf_{k\to\infty} Q_{j,\sigma}(k) = \liminf_{k\to\infty} \frac{\lambda_j(X(\sigma^{k+1}))}{\lambda_j(X(\sigma^k))} \leq \limsup_{k\to\infty} \frac{M\sigma^{(k+1)\xi(i)}}{M\sigma^{k\xi(i)}} = \sigma^{\xi(i)}.$$

Lastly $\sigma^{\xi(i)} < 1$ holds since, $\sigma \in (0,1)$ and $\xi(i) > 0$. □

Theorem 5.1.4 does not completely have the desired form of (5.1.2) due to the limit inferior. Nonetheless, the theorem provides a way to distinguish between eigenvalue functions that converge to 0 and those that do not, as emphasized in the following.

Corollary 5.1.5. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a spectrahedron that satisfies Assumption 5.0.1 and let $\{X(\alpha) : \alpha > 0\}$ be a central path for $\mathcal{F}$ that satisfies Assumption 5.0.2 and suppose that $\epsilon^b(X(\alpha), \mathcal{F}) = \mathcal{O}(\alpha)$. Let r denote the maximum rank over $\mathcal{F}$ and let $\sigma \in (0,1)$. Then,*

$$\liminf_{k\to\infty} Q_{i,\sigma}(k) = \begin{cases} 1 & \text{if } i \leq r, \\ \leq \sigma^{\xi(1)} & \text{otherwise.} \end{cases}$$

The number $\sigma^{\xi(1)}$ serves as a threshold so that limit inferiors of the eigenvalue ratios lie below this number if, and only if, those eigenvalues converge to 0. For large singularity degree, it may be difficult to distinguish $\sigma^{\xi(1)}$ from 1, numerically. However, if we can identify another number, say $\tau \in (0,1)$, that is numerically distinguishable from 1 and there exists a positive integer $\overline{r}$ such that,

$$\liminf_{k\to\infty} Q_{i,\sigma}(k) \leq \tau \iff i > \overline{r},$$

then $\overline{r}$ is an upper bound on the maximum rank, r, over $\mathcal{F}$. We state this result formally in the following.

Corollary 5.1.6. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a spectrahedron that satisfies Assumption 5.0.1 and let $\{X(\alpha) : \alpha > 0\}$ be a central path for $\mathcal{F}$ that satisfies Assumption 5.0.2 and suppose that $\epsilon^b(X(\alpha), \mathcal{F}) = \mathcal{O}(\alpha)$. Let r denote the maximum rank over $\mathcal{F}$ and let $\sigma \in (0,1)$. Suppose there exists $\tau \in (0,1)$ and $\overline{r} \in \{1, \ldots, n\}$ such that*

$$\liminf_{k\to\infty} Q_{i,\sigma}(k) \leq \tau \iff i > \overline{r}.$$

Then $\overline{r} \geq r$.

Remark 5.1.7. *In the results of this section we require knowledge of $\epsilon^b(X(\alpha), \mathcal{F})$. When $\mathcal{F}$ is the solution set of an SDP, as $\mathcal{P}$ is in Section 4.5, and C in the objective function is not the zero matrix, computing backward error requires us to know the optimal value $p^\star$. Since we can not expect to know $p^\star$, backward error may be intractable. Therefore, it may be challenging to determine whether a central path satisfies the hypothesis of Theorem 5.1.4. However, primal-dual algorithms actually construct a central path that consists of both primal and dual variables and has the form,*

$$\{(X(\alpha), y(\alpha), Z(\alpha)) \in \mathbb{S}^n_{++} \times \mathbb{R}^m \times \mathbb{S}^n_{++} : \alpha > 0\}.$$

Using this path, there is a tractable way to ensure that the backward error for $X(\alpha)$ is sufficiently small as required in the hypothesis of the results of this section. Indeed, in Section 4 of [77], Sturm showed that if,

$$\begin{aligned}
&\operatorname{dist}\left(X(\alpha), \{X : \mathcal{A}(X) = b\}\right) \\
&\quad + \operatorname{dist}\left((y(\alpha), Z(\alpha)), \{(y,Z) : \mathcal{A}^*(y) + Z = C\}\right) \\
&\quad\quad + \operatorname{dist}\left((X(\alpha), Z(\alpha)), \mathbb{S}^n_+ \times \mathbb{S}^n_+\right) \\
&\quad\quad\quad + \langle X(\alpha), Z(\alpha)\rangle = \mathcal{O}(\alpha),
\end{aligned}$$

as $\alpha \searrow 0$, then $\epsilon^b(X(\alpha), \mathcal{F}) = \mathcal{O}(\alpha)$ as $\alpha \searrow 0$.

5.1.2 Sum of Eigenvalues Q-Convergence Ratio

Is it possible to strengthen the results of the previous section by making additional assumptions on the eigenvalue functions? What if we assume that each $\lambda_i(X(\alpha))$ is concave

or convex? To this end we consider a relative of the eigenvalue function: the sum of the smallest eigenvalues,

$$\mu_i(X(\alpha)) := \sum_{j=i}^{n} \lambda_j(X(\alpha)). \tag{5.1.16}$$

For each $i \in \{1, \dots, n\}$ the function $X \mapsto \mu_i(X)$ is known to be concave by the min-max principle of Fan, [24]. Concavity of $\alpha \mapsto \mu_i(X(\alpha))$ depends on the parametrization $\alpha \mapsto X(\alpha)$. While it is possible to construct parametrizations such that $\mu_i(X(\alpha))$ is not concave, we have observed that these functions tend to be concave when $X(\alpha)$ is the central path of Chapter 6.

The function $\mu_i(X(\alpha))$ possesses two properties of $\lambda_i(X(\alpha))$ that lead to the results of the previous section. First, $\mu_i(X(\alpha))$ has similar convergence to $\lambda_i(X(\alpha))$ since,

$$\lim_{\alpha \searrow 0} \mu_i(X(\alpha)) = 0 \iff \lim_{\alpha \searrow 0} \lambda_i(X(\alpha)) = 0. \tag{5.1.17}$$

Secondly, when $j \in \mathcal{I}^i$ with $i \in \{1, \dots, \mathrm{sd}(\mathcal{F})\}$ it holds that,

$$\mu_j(X(\alpha)) = \mathcal{O}\left(\alpha^{\xi(i)}\right). \tag{5.1.18}$$

With the above reasons in mind, let us define the *sum of eigenvalues Q-convergence ratio*, for $\sigma \in (0,1)$ and $k \in \mathbb{N}$ as,

$$S_{i,\sigma}(k) := \frac{\mu_i\left(X\left(\sigma^{k+1}\right)\right)}{\mu_i\left(X\left(\sigma^{k}\right)\right)}, \ i \in \{1, \dots, n\}. \tag{5.1.19}$$

The properties of (5.1.17) and (5.1.18) imply that Theorem 5.1.4 may be restated with $Q_{i,\sigma}$ replaced by $S_{i,\sigma}$. In this section we strengthen that result by assuming that $\mu_i(X(\alpha))$ is concave. We begin with a technical lemma.

Lemma 5.1.8. *Let $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ be concave with $\psi(0) = 0$ and $\psi(\mathbb{R}_{++}) \subseteq \mathbb{R}_{++}$. Suppose there exists $M > 0$ such that $\psi(x) \leq Mx$ for all $x \in \mathbb{R}_+$. Then there exists $\widehat{M} \in \mathbb{R}$ such that,*

(i) $\lim_{x \searrow 0} \frac{\psi(x)}{x} = \widehat{M}$,

(ii) $0 \leq \widehat{M} \leq M$,

(iii) $\psi(x) \leq \widehat{M}x$ *for all* $x \in \mathbb{R}_+$.

Proof. Let y and x be positive numbers satisfying $y < x$. Since ψ is concave and $\psi(0) = 0$, it holds that $(y, \psi(y))$ lies above (or on) the line connecting the origin and $(x, \psi(x))$. Therefore,

$$\frac{\psi(y)}{y} \geq \frac{\psi(x)}{x}. \tag{5.1.20}$$

It follows that the function $\psi(x)/x$ is monotonically non-increasing over $\mathbb{R}_{++}$. Moreover, $\psi(x)/x$ is bounded from above since,

$$\frac{\psi(x)}{x} \leq \frac{Mx}{x} = M.$$

Therefore, there exists a positive number $\widehat{M}$ such that,

$$\lim_{x \searrow 0} \frac{\psi(x)}{x} = \widehat{M}, \tag{5.1.21}$$

proving *(i)*.

Next, observe that M and $\widehat{M}$ are both upper bounds of the image of $\mathbb{R}_{++}$ under the function $\psi(x)/x$. By the monotonicity of $\psi(x)/x$ and (5.1.21) it follows that $\widehat{M}$ is the least upper bound, implying *(ii)*. This observation also yields *(iii)*, since for any $x > 0$ it holds that,

$$\frac{\psi(x)}{x} \leq \widehat{M} \implies \psi(x) \leq \widehat{M}x.$$

□

The main result of this section is the following.

Theorem 5.1.9. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a spectrahedron that satisfies Assumption 5.0.1 and let $\{X(\alpha) : \alpha > 0\}$ be a central path for $\mathcal{F}$ that satisfies Assumption 5.0.2 and suppose that $\epsilon^b(X(\alpha), \mathcal{F}) = \mathcal{O}(\alpha)$. Assume, furthermore, that $\mu_j(X(\alpha))$ is concave in α for all j. Let $\mathcal{I}^0, \mathcal{I}^1, \ldots, \mathcal{I}^{\mathrm{sd}(\mathcal{F})}$ be a partition of $\{1, \ldots, n\}$ as constructed in Lemma 5.1.2. Let $\sigma \in (0, 1)$ and let $S_{j,\sigma}(k)$ be as in (5.1.19). Then the following hold.*

(i) If $j \in \mathcal{I}^0$ then,

$$\lim_{k \to \infty} S_{j,\sigma}(k) = 1.$$

(ii) If $j \in \mathcal{I}^i$ with $i \in \{1, \ldots, \mathrm{sd}(\mathcal{F})\}$ then,

$$\liminf_{k \to \infty} S_{j,\sigma}(k) \in [\sigma, \sigma^{\xi(i)}].$$

(iii) If $j \in \mathcal{I}^{\mathrm{sd}(\mathcal{F})}$ then,

$$\lim_{k \to \infty} S_{j,\sigma}(k) = \sigma.$$

Proof. First note that *(i)* is a direct implication of Theorem 5.1.4 *(i)*. For *(ii)*, let i and j be as in the hypothesis. We have already argued that $Q_{i,\sigma}$ may be replaced by $S_{i,\sigma}$ in Theorem 5.1.4. Therefore it holds that,

$$\liminf_{k\to\infty} S_{j,\sigma}(k) \leq \sigma^{\xi(i)}. \tag{5.1.22}$$

Now the concavity of $\mu_j(X(\cdot))$ over $\mathbb{R}_+$ implies that $(\sigma^{k+1}, \mu_j(X(\sigma^{k+1})))$ lies above the line connecting the origin to $(\sigma^k, \mu_j(X(\sigma^k)))$, for every $k \in \mathbb{N}$. Therefore for all $k \in \mathbb{N}$ it holds that,

$$\frac{\mu_j(X(\sigma^{k+1}))}{\sigma^{k+1}} \geq \frac{\mu_j X(\sigma^k))}{\sigma^k} \implies \mu_j(X(\sigma^{k+1})) \geq \mu_j(X(\sigma^k))\sigma \implies S_{j,\sigma}(k) \geq \sigma. \tag{5.1.23}$$

Then (5.1.22) and (5.1.23) imply *(ii)*.

Lastly, we prove *(iii)*. Let $j \in \mathcal{I}^{\mathrm{sd}(\mathcal{F})}$. From *(ii)* it holds that,

$$\liminf_{k\to\infty} S_{j,\sigma}(k) = \sigma.$$

For the sake of contradiction, suppose that the limit superior of the sequence differs from σ. Then there exists $\varepsilon > 0$ and a subsequence $\{k_\ell\}_{\ell\in\mathbb{N}}$ such that,

$$S_{j,\sigma}(k_\ell) \geq \sigma + \varepsilon, \ \forall \ell \in \mathbb{N}. \tag{5.1.24}$$

Rearranging (5.1.24) and recalling that $\mu_j(X(\alpha)) = \mathcal{O}(\alpha)$ it holds that there exists $M > 0$ such that,

$$(\sigma+\varepsilon)\mu_j(X(\sigma^{k_\ell})) \leq \mu_j(X(\sigma^{k_\ell+1})) \leq M\sigma^{k_\ell+1}, \ \forall \ell \in \mathbb{N}. \tag{5.1.25}$$

Moreover, by Lemma 5.1.8 we may assume M is such that,

$$\lim_{k\to\infty} \frac{\mu_j(X(\sigma^k))}{\sigma^k} = M. \tag{5.1.26}$$

Rearranging (5.1.25) yields,

$$\frac{\mu_j(X(\sigma^{k_\ell}))}{\sigma^{k_\ell}} \leq \frac{\sigma}{\sigma+\varepsilon} M < M, \ \forall \ell \in \mathbb{N},$$

a contradiction of (5.1.26). □

When every sequence $S_{j,\sigma}(k)$ has a limit in k, we get a cleaner version of Theorem 5.1.9.

Corollary 5.1.10. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a spectrahedron that satisfies Assumption 5.0.1 and let $\{X(\alpha) : \alpha > 0\}$ be a central path for $\mathcal{F}$ that satisfies Assumption 5.0.2 and suppose that $\epsilon^b(X(\alpha), \mathcal{F}) = \mathcal{O}(\alpha)$. Assume, furthermore, that $\mu_j(X(\alpha))$ is concave in α for all j. Let $\sigma \in (0,1)$, let $S_{j,\sigma}(k)$ be as in (5.1.19), and let $r := \mathrm{rank}(\mathcal{F})$. Suppose that for each $j \in \{1, \dots, n\}$ there exists $L_j \in \mathbb{R}$ such that,*

$$\lim_{k\to\infty} S_{j,\sigma}(k) = L_j.$$

Then,

$$\sigma = L_n \leq \cdots \leq L_{r+1} \leq \sigma^{\xi(1)} < L_r = \cdots = L_1 = 1.$$

In general we cannot guarantee that every sequence $S_{j,\sigma}(k)$ converges in k. For this reason 'lim inf' cannot be replaced by 'lim' in Theorem 5.1.9. We address this further in Section 5.1.3, where we construct a family of functions that possesses properties similar to that of $\mu_i(X(\alpha))$, but the ratio of successive function values along the sequence $\{\sigma^k\}$, does not converge.

A special case to consider is that of $\mathrm{sd}(\mathcal{F}) = 1$. Here every $S_{j,\sigma}(k)$ belongs to case *(i)* or *(iii)* of Theorem 5.1.9, yielding the following result.

Corollary 5.1.11. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a spectrahedron that satisfies Assumption 5.0.1 and let $\{X(\alpha) : \alpha > 0\}$ be a central path for $\mathcal{F}$ that satisfies Assumption 5.0.2 and suppose that $\epsilon^b(X(\alpha), \mathcal{F}) = \mathcal{O}(\alpha)$. Assume, furthermore, that $\mu_j(X(\alpha))$ is concave in α for all j. Let $\sigma \in (0, 1)$, let $S_{j,\sigma}(k)$ be as in (5.1.19), and let $r := \mathrm{rank}(\mathcal{F})$. Then,*

$$\mathrm{sd}(\mathcal{F}) = 1 \implies \lim_{k\to\infty} S_{j,\sigma}(k) = \begin{cases} 1, & \text{if } i \le r, \\ \sigma, & \text{if } i > r. \end{cases}$$

5.1.3 An Interesting Family of Functions

In this section we construct a function $\psi(\alpha)$ that behaves like $\mu_i(X(\alpha))$ in case *(ii)* of Theorem 5.1.9, but the ratio $\psi(\sigma^{k+1})/\psi(\sigma^k)$ does not converge. The relevant properties of $\mu_i(X(\alpha))$ are that it is concave, it maps positive values to positive values and 0 to 0, it is bounded above by α^p for some $p \in (0, 1)$, and it is not bounded above by a linear function, otherwise we could obtain the result of case *(iii)* of Theorem 5.1.9.

Now to construct our function, let $\sigma, p, \theta \in (0, 1)$. We construct a piecewise linear function,

$$\psi : [0, \sigma] \to \mathbb{R}_+,$$

with the properties,

(i) $\psi(0) = 0$ and $\psi((0, \sigma]) \subset \mathbb{R}_{++}$,

(ii) ψ is concave and non-decreasing,

(iii) $\psi(x) \le x^p$ over $[0, \sigma]$,

(iv) there does not exist $\varepsilon \in (0, \sigma]$ and $M \ge 0$ such that $\psi(x) \le Mx$ over $[0, \varepsilon]$,

(v) the sequence $\psi(\sigma^{k+1})/\psi(\sigma^k)$ for odd k has cluster points in $(\sigma^p, 1]$,

(vi) the same sequence for even k has cluster points in $[0, \sigma^p)$.

Note that *(i)*-*(iv)* are exactly the properties we identified for $\mu_i(X(\alpha))$, with the addition of the non-decreasing assumption. But this is not really an assumption as it necessarily

exists in some neighbourhood of 0 by the assumption of concavity. The last two properties ensure that the limit of $\psi(\sigma^{k+1})/\psi(\sigma^k)$ does not exist.

It suffices to describe ψ at the 'breaking points', which we choose to be $\sigma, \sigma^2, \sigma^3, \ldots$. We begin with

$$\psi(\sigma) = \theta\sigma^p, \ \psi(\sigma^2) = \min\{\psi(\sigma^1), (\sigma^2)^p\}.$$

There may be many values that $\psi(\sigma^3)$ can take and still satisfy the first two of the desired properties. For ψ to be concave and non-decreasing over $[\sigma^3, \sigma]$ we need $(\sigma^3, \psi(\sigma^3))$ to lie below the line that passes through $(\sigma, \psi(\sigma))$ and $(\sigma^2, \psi(\sigma^2))$. We also need $\psi(\sigma^3) \leq (\sigma^3)^p$ in order for *(iii)* to hold. These two conditions give the upper bound,

$$\psi(\sigma^3) \leq \overline{u}_3 := \min\left\{\frac{\psi(\sigma^2) - \sigma\psi(\sigma)}{1-\sigma}, \sigma^{3p}\right\}. \tag{5.1.27}$$

We also need a lower bound so that *(iv)* holds. In particular, $(\sigma^3, \psi(\sigma^3))$ should lie strictly above the line connecting $(0,0)$ and $(\sigma^2, \psi(\sigma^2))$, i.e., $\psi(\sigma^3) > \sigma\psi(\sigma^2)$. We actually impose the more restrictive lower bound,

$$\psi(\sigma^3) \geq \underline{u}_3 := \theta\overline{u}_3 + (1-\theta)\sigma\psi(\sigma^2). \tag{5.1.28}$$

Now any choice of $\psi(\sigma^3) \in [\underline{u}_3, \overline{u}_3]$ ensures that properties *(ii)* and *(iii)* hold on the interval $[\sigma^3, \sigma]$. These bounds easily extend to $\psi(\sigma^k)$ for any $k \geq 3$:

$$\begin{aligned} \overline{u}_k &:= \min\left\{\frac{\psi(\sigma^{k-1}) - \sigma\psi(\sigma^{k-2})}{1-\sigma}, \sigma^{kp}\right\}, \\ \underline{u}_k &:= \theta\overline{u}_k + (1-\theta)\sigma\psi(\sigma^{k-1}). \end{aligned} \tag{5.1.29}$$

In order to satisfy *(v)* and *(vi)* we alternate between $\underline{u}_k$ and $\overline{u}_k$ for the value of $\psi(\sigma^k)$. Specifically, for $k \geq 3$,

$$\psi(\sigma^k) := \begin{cases} \underline{u}_k & \text{if } k \text{ is odd,} \\ \overline{u}_k & \text{if } k \text{ is even.} \end{cases} \tag{5.1.30}$$

Having defined ψ for all positive elements of the domain, we conclude the construction by setting $\psi(0) := 0$. With this construction it is not difficult to verify that conditions *(i)* through *(iv)* hold. Since it may not be obvious that ψ is concave, we have included the following result.

Proposition 5.1.12. *Let $[a,b] \subset \mathbb{R}$ with $a < b$ and let $\phi : [a,b] \to \mathbb{R}$ be a continuous piecewise linear function with breaking points,*

$$a = c_1 < c_2 < \cdots < c_N = b,$$

for some positive integer $N > 1$. Then ϕ is convex if, and only if,

$$\phi(c_{k+1}) \leq t\phi(c_k) + (1-t)\phi(c_{k+2}), \ \forall k \in \{1, \ldots, N-2\}, \tag{5.1.31}$$

where $t \in (0,1)$ satisfies $c_{k+1} = tc_k + (1-t)c_{k+2}$.

Proof. The forward direction is trivial. For the converse, we show that ϕ is the maximum of finitely many affine functions, and therefore convex. For each $k \in \{1, \dots, N-1\}$, let $\phi_k : \mathbb{R} \to \mathbb{R}$ denote the affine function such that $\phi(x) = \phi_k(x)$ on $[c_k, c_{k+1}]$. We now prove the claim that,

$$\phi(x) = \max_{i \in \{1,\dots,N-1\}} \phi_i(x), \tag{5.1.32}$$

on $[a, b]$. Convexity of ϕ is then implied by Theorem 5.5 of [71]. First we show that for any $k \in \{1, \dots, N-2\}$ the claim of (5.1.32) holds on the interval $[c_k, c_{k+2}]$ with i restricted to $\{k, k+1\}$. Since ϕ_k and ϕ_{k+1} are affine, there exist real numbers $m_k, m_{k+1}, b_k, b_{k+1}$ such that $\phi_k(x) = m_k x + b_k$ and $\phi_{k+1}(x) = m_{k+1}x + b_{k+1}$. Then rearranging (5.1.31) we have,

$$\begin{aligned}
\phi(c_{k+2}) &\geq \frac{1}{1-t}\left(\phi(c_{k+1}) - t\phi(c_k)\right) \\
&= \frac{1}{1-t}\left(m_k c_{k+1} + b_k - t(m_k c_k + b_k)\right) \\
&= \frac{1}{1-t}\left(m_k(tc_k + (1-t)c_{k+2}) + b_k - t(m_k c_k + b_k)\right) \\
&= m_k c_{k+2} + b_k \\
&= \phi_k(c_{k+2}).
\end{aligned}$$

Since $\phi(c_{k+2}) = \phi_{k+1}(c_{k+2})$ and ϕ_{k+1} coincides with ϕ_k at c_{k+1}, and $c_{k+1} < c_{k+2}$, we conclude that $m_{k+1} \geq m_k$. Therefore, $\phi_k \geq \phi_{k+1}$ over $[c_k, c_{k+1}]$ and $\phi_{k+1} \geq \phi_k$ over $[c_{k+1}, c_{k+2}]$. It follows that,

$$\phi(x) = \max_{i \in \{k,k+1\}} \phi_i(x), \tag{5.1.33}$$

for $x \in [c_k, c_{k+2}]$. By induction we have that $m_1 \leq m_2 \leq \cdots \leq m_{N-1}$. Then (5.1.33) easily extends to (5.1.32), as desired. □

Since convexity and concavity are interchangeable for our purposes, this result ensures that ψ is concave as long as it is concave on every interval of the form $[\sigma^{k+2}, \sigma^k]$. We know this property holds for ψ, by construction.

While Proposition 5.1.12 only addresses the case of finitely many break points, it easily extends to the case of ψ, where there are infinitely many break points. To see this, note that the proposition ensures that ψ is concave on $[\sigma^k, \sigma]$ for every $k \in \mathbb{N}$. Now let $\alpha_1, \alpha_2 \in [0, \sigma]$ and let $t \in (0, 1)$. For ψ to be concave we need,

$$\psi(t\alpha_1 + (1-t)\alpha_2) \geq t\psi(\alpha_1) + (1-t)\psi(\alpha_2). \tag{5.1.34}$$

Without loss of generality, we may assume that $\alpha_1 < \alpha_2$. Moreover, we may assume $\alpha_1 = 0$, otherwise the line segment $t\alpha_1 + (1-t)\alpha_2$ belongs to $[\sigma^{\bar{k}}, \sigma]$ for some $\bar{k} \in \mathbb{N}$, implying (5.1.34). Then (5.1.34) reduces to,

$$\psi((1-t)\alpha_2) \geq (1-t)\psi(\alpha_2). \tag{5.1.35}$$

For every $k \in \mathbb{N}$ it holds that,

$$\psi(t\sigma^k + (1-t)\alpha_2) \geq t\psi(\sigma^k) + (1-t)\psi(\alpha_2). \tag{5.1.36}$$

Taking the limit of both sides as $k \to \infty$ yields (5.1.34), as desired.

We have now shown that ψ possesses all the relevant properties of $\mu_i(X(\alpha))$ in case *(ii)* of Theorem 5.1.9. All that remains is to prove that *(v)* and *(vi)* hold. We break the argument into two claims.

Proposition 5.1.13. *Let $\sigma, p, \theta \in (0,1)$ be fixed. If $\overline{u}_k = \sigma^{kp}$ for every $k \geq 3$ then,*

$$\frac{\psi(\sigma^{k+1})}{\psi(\sigma^k)} = \begin{cases} \frac{\sigma^p}{\theta + (1-\theta)\sigma^{1-p}} & \text{if } k \text{ is odd,} \\ \theta\sigma^p + (1-\theta)\sigma & \text{if } k \text{ is even.} \end{cases}$$

In particular, for even k, the sequence $\psi(\sigma^{k+1})/\psi(\sigma^k)$ is constant and strictly smaller than σ^p. While for odd k, the sequence is constant and strictly greater than σ^p.

Proof. When k is odd,

$$\begin{aligned} \frac{\psi(\sigma^{k+1})}{\psi(\sigma^k)} &= \frac{\overline{u}_{k+1}}{\underline{u}_k} \\ &= \frac{\sigma^{(k+1)p}}{\theta\overline{u}_k + (1-\theta)\sigma\psi(\sigma^{k-1})} \\ &= \frac{\sigma^{(k+1)p}}{\theta\sigma^{kp} + (1-\theta)\sigma\sigma^{(k-1)p}} \\ &= \frac{\sigma^p}{\theta + (1-\theta)\sigma^{1-p}}, \end{aligned}$$

as desired. The denominator is strictly smaller than 1, therefore the ratio is strictly greater than σ^p. For even k,

$$\begin{aligned} \frac{\psi(\sigma^{k+1})}{\psi(\sigma^k)} &= \frac{\underline{u}_{k+1}}{\overline{u}_k} \\ &= \frac{\theta\overline{u}_{k+1} + (1-\theta)\sigma\psi(\sigma^k)}{\sigma^{kp}} \\ &= \frac{\theta\sigma^{(k+1)p} + (1-\theta)\sigma\sigma^{kp}}{\sigma^{kp}} \\ &= \theta\sigma^p + (1-\theta)\sigma. \end{aligned}$$

Since $\theta \in (0,1)$ the ratio lies strictly between σ and σ^p. Moreover, $p \in (0,1)$ so $\sigma < \sigma^p$ and thus the ratio is constant and strictly less than σ^p. □

To conclude the example, it suffices to show the existence of $\sigma, \theta, p \in (0,1)$ such that σ^{kp} is the choice for $\overline{u}_k$, for every $k \geq 3$.

Proposition 5.1.14. *Let $\sigma, \theta, p \in (0,1)$. Then,*

$$\theta = \sigma^p \left(\frac{1-\sigma}{1-\sigma^{1-p}} \right), \ \sigma^{1-p} + \sigma^p - \sigma^{1+p} < 1 \implies \overline{u}_k = \sigma^{kp}, \ \forall k \geq 3.$$

Proof. By (5.1.29) it suffices to show that,

$$\sigma^{kp} \leq \frac{\psi(\sigma^{k-1}) - \sigma\psi(\sigma^{k-2})}{1-\sigma},$$

whenever $k \geq 3$. We proceed by induction on k. Let θ and σ be as in the hypothesis. By construction and the choice of θ we have,

$$\psi(\sigma) = \sigma^{2p} \left(\frac{1-\sigma}{1-\sigma^{1-p}} \right) \text{ and } \psi(\sigma^2) = \sigma^{2p}.$$

For the base case, $k = 3$, we have,

$$\begin{aligned} \frac{\psi(\sigma^2) - \sigma\psi(\sigma)}{1-\sigma} &= \frac{\sigma^{2p} - \sigma\sigma^{2p}\left(\frac{1-\sigma}{1-\sigma^{1-p}}\right)}{1-\sigma} \\ &= \sigma^{2p} \left(\frac{1 - \sigma\left(\frac{1-\sigma}{1-\sigma^{1-p}}\right)}{1-\sigma} \right). \end{aligned}$$

Then,

$$\begin{aligned} \frac{\psi(\sigma^2) - \sigma\psi(\sigma)}{1-\sigma} \geq \sigma^{3p} &\iff \sigma^{2p} \left(\frac{1 - \sigma\left(\frac{1-\sigma}{1-\sigma^{1-p}}\right)}{1-\sigma} \right) \geq \sigma^{3p} \\ &\iff \frac{1 - \sigma\left(\frac{1-\sigma}{1-\sigma^{1-p}}\right)}{1-\sigma} \geq \sigma^p \\ &\iff 1 - \sigma\left(\frac{1-\sigma}{1-\sigma^{1-p}}\right) \geq \sigma^p(1-\sigma) \\ &\iff 1 \geq (1-\sigma)\left(\sigma^p + \frac{\sigma}{1-\sigma^{1-p}}\right) \\ &\iff 1 \geq \sigma^p \left(\frac{1-\sigma}{1-\sigma^{1-p}}\right) \\ &\iff 1 \geq \theta. \end{aligned}$$

Since $\theta \in (0,1)$ it holds that $\overline{u}_k = \sigma^{kp}$ when $k = 3$. Now we address the even and odd cases separately. Let $k \geq 4$ be even and assume that the claim holds for all smaller integers.

Then,

$$\begin{aligned}
\frac{\psi(\sigma^{k-1}) - \sigma\psi(\sigma^{k-2})}{1-\sigma} &= \frac{\underline{u}_{k-1} - \sigma\overline{u}_{k-2}}{1-\sigma} \\
&= \frac{\theta\overline{u}_{k-1} + (1-\theta)\sigma\psi(\sigma^{k-2}) - \sigma\sigma^{(k-2)p}}{1-\sigma} \\
&= \theta\left(\frac{\sigma^{(k-1)p} - \sigma\sigma^{(k-2)p}}{1-\sigma}\right) \\
&= \sigma^{kp}\theta\sigma^{-p}\left(\frac{1-\sigma^{1-p}}{1-\sigma}\right) \\
&= \sigma^{kp}\theta\theta^{-1} \\
&= \sigma^{kp}.
\end{aligned}$$

When $k \geq 5$ is odd then,

$$\begin{aligned}
\frac{\psi(\sigma^{k-1}) - \sigma\psi(\sigma^{k-2})}{1-\sigma} &= \frac{\overline{u}_{k-1} - \sigma\underline{u}_{k-2}}{1-\sigma} \\
&= \frac{\sigma^{(k-1)p} - \sigma\left(\theta\overline{u}_{k-2} + (1-\theta)\sigma\psi(\sigma^{k-3})\right)}{1-\sigma} \\
&= \frac{\sigma^{(k-1)p} - \sigma\left(\theta\sigma^{(k-2)p} + (1-\theta)\sigma\sigma^{(k-3)p}\right)}{1-\sigma} \\
&= \sigma^{kp}\sigma^{-p}\left(\frac{1-\theta\sigma^{1-p} - (1-\theta)\sigma^{2(1-p)}}{1-\sigma}\right) \\
&= \sigma^{kp}\sigma^{-p}\left(\frac{1-\sigma^{2(1-p)} + \theta\sigma^{1-p}(\sigma^{1-p}-1)}{1-\sigma}\right) \\
&= \sigma^{kp}\sigma^{-p}\left(\frac{1-\sigma^{2(1-p)} + \sigma(\sigma-1)}{1-\sigma}\right) \\
&= \sigma^{kp}\sigma^{-p}\left(\frac{1-\sigma^{1-p}\left(\sigma^{1-p} + \sigma^{p} - \sigma^{1+p}\right)}{1-\sigma}\right).
\end{aligned}$$

Recalling that $\sigma^{1-p} + \sigma^p - \sigma^{1+p} < 1$ and $\theta \in (0,1)$ by hypothesis, we get,

$$\begin{aligned}
\frac{\psi(\sigma^{k-1}) - \sigma\psi(\sigma^{k-2})}{1-\sigma} &> \sigma^{kp}\sigma^{-p}\left(\frac{1-\sigma^{1-p}}{1-\sigma}\right) \\
&= \sigma^{kp}\theta^{-1} \\
&\geq \sigma^{kp},
\end{aligned}$$

as desired. □

The set of admissible values for σ and p is far from empty. For instance, when $p = 1/2$, any choice of $\sigma \in (0, 1/4]$ satisfies the hypothesis of Proposition 5.1.14.

5.2 Bounds on Forward Error and Singularity Degree

Any bound on maximum rank, such as the bounds of Section 5.1, may be used to provide a lower bound on forward error and singularity degree.

Theorem 5.2.1. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a spectrahedron that satisfies Assumption 5.0.1 and let $\{X(\alpha) : \alpha > 0\}$ be a central path for $\mathcal{F}$ that satisfies Assumption 5.0.2. Then,*

$$\bar{r} \geq \operatorname{rank}(\mathcal{F}) \implies \epsilon^f(X(\alpha), \mathcal{F}) \geq \|\begin{pmatrix}\lambda_{\bar{r}+1}(X(\alpha)) & \cdots & \lambda_n(X(\alpha))\end{pmatrix}^T\|_2, \ \forall \alpha > 0.$$

Proof. Let $\bar{r}$ be as in the hypothesis and let $\alpha > 0$. Since $\mathcal{F}$ is a closed convex set, there exists $X \in \mathcal{F}$ such that,

$$\epsilon^f(X(\alpha), \mathcal{F}) = \|X(\alpha) - X\|_F.$$

Then observing that $\|S\|_F = \|\lambda(S)\|_2$ for any $S \in \mathbb{S}^n$ we have,

$$\begin{aligned}\epsilon^f(X(\alpha), \mathcal{F})^2 &= \|X(\alpha) - X\|_F^2 \\ &= \|X(\alpha)\|_F^2 + \|X\|_F^2 - 2\langle X(\alpha), X\rangle \\ &= \|\lambda(X(\alpha))\|_2^2 + \|\lambda(X)\|_2^2 - 2\langle X(\alpha), X\rangle.\end{aligned}$$

Applying Fact 2.1.3 we get,

$$\begin{aligned}\epsilon^f(X(\alpha), \mathcal{F})^2 &\geq \|\lambda(X(\alpha))\|_2^2 + \|\lambda(X)\|_2^2 - 2\lambda(X(\alpha))^T\lambda(X) \\ &= \|\lambda(X(\alpha)) - \lambda(X)\|_2^2 \\ &\geq \|\begin{pmatrix}\lambda_{\bar{r}+1}(X(\alpha)) & \cdots & \lambda_n(X(\alpha))\end{pmatrix}^T\|_2^2.\end{aligned}$$

Taking the square root of both sides yields the desired result. □

Theorem 5.2.2. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a spectrahedron that satisfies Assumption 5.0.1 and let $\{X(\alpha) : \alpha > 0\}$ be a central path for $\mathcal{F}$ that satisfies Assumption 5.0.2 and suppose that $\epsilon^b(X(\alpha), \mathcal{F}) = \mathcal{O}(\alpha)$. Let $\bar{r} \geq \operatorname{rank}(\mathcal{F})$ and let $\sigma \in (0, 1)$. Suppose $\underline{d}$ is the smallest positive integer such that,*

$$\liminf_{k\to\infty} Q_{i,\sigma}(k) \leq \sigma^{2^{-(\underline{d}-1)}} \iff i > \bar{r}. \tag{5.2.1}$$

Then $\underline{d} \leq \operatorname{sd}(\mathcal{F})$.

Proof. Let r denote the maximum rank over $\mathcal{F}$. Suppose $\bar{r} > r$. Then by Corollary 5.1.5 and by (5.2.1) we have,

$$\sigma^{2^{-(\underline{d}-1)}} < \liminf_{k\to\infty} Q_{(r+1),\sigma}(k) \leq \sigma^{2^{-(\operatorname{sd}(\mathcal{F})-1)}}.$$

It follows that $\underline{d} \leq \operatorname{sd}(\mathcal{F})$. Now suppose that $\bar{r} = r$. Then by Corollary 5.1.5 we have,

$$\liminf_{k\to\infty} Q_{i,\sigma}(k) \leq \sigma^{2^{-(\operatorname{sd}(\mathcal{F})-1)}} \iff i > r = \bar{r}. \tag{5.2.2}$$

Out of all positive integers that could replace $\operatorname{sd}(\mathcal{F})$ in (5.2.2), we chose $\underline{d}$ to be the smallest. Hence $\underline{d} \leq \operatorname{sd}(\mathcal{F})$, as desired. □

Chapter 6

Singularity Degree as a Measure of Hardness

It is certainly possible to construct a parametric curve $\{X(\alpha) : \alpha > 0\}$ with the properties of Assumption 5.0.2, for which singularity degree is large, but forward error is small. For instance, the path defined as $X(\alpha) := \bar{X} + \alpha I$, where $\bar{X} \in \text{relint}(\mathcal{F})$, exhibits fast convergence and low forward error irrespective of the singularity degree. This demonstrates that large singularity degree is not a sufficient condition for slow convergence for *all* parametric curves. However, for many of the central paths constructed by state of the art algorithms, empirical evidence indicates otherwise. In this section we present several results that give credence to the notion that singularity degree is a measure of hardness for a family of central paths. In Section 6.1 we introduce the family of central paths and in Section 6.2 we present the main results of the chapter. The results of Section 6.2 are based on the preprint [74], coauthored by the author of this thesis.

6.1 Analysis of a Family of Central Paths

The classical *interior point* method for SDP is based on a central path that is constructed by assuming that both the primal and the dual satisfy the Slater condition. As this assumption is quite restrictive, *infeasible* central paths or those constructed by the *self-dual embedding* assuming weaker conditions have been subsequently proposed. Among these are [13, 14, 52, 58, 68].

In [68], Potra and Sheng propose a family of infeasible central paths that are based on perturbing the feasible region so as to satisfy the Slater condition, and then decreasing the perturbation. From the family of paths proposed by Potra and Sheng, we choose the path

defined by,

$$\begin{cases} X(\alpha) := \arg\max\{\alpha \log\det(X) : X \in \mathcal{F}(\alpha)\}, \\ \mathcal{F}(\alpha) := \{X \in \mathbb{S}^n_+ : \mathcal{A}(X) = b(\alpha)\}, \\ b(\alpha) := b + \alpha\mathcal{A}(B), \end{cases} \tag{6.1.1}$$

where $B \succ 0$ is fixed. The matrix $X(\alpha)$ exists for each $\alpha > 0$ if, and only if, $\mathcal{F}(\alpha) \cap \mathbb{S}^n_{++}$ is non-empty and bounded. The existence of $X(\alpha)$ is also guaranteed by the following assumptions on $\mathcal{F}$.

Assumption 6.1.1. *We assume that the spectrahedron $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ satisfies,*

(i) $\mathcal{F}$ is non-empty, bounded, $\mathrm{sd}(\mathcal{F}) \geq 1$, *and $\mathcal{F} \neq \{0\}$,*

(ii) $\mathcal{A}$ is surjective.

Assumption 6.1.1 differs from Assumption 5.0.1 in the additional requirements that $\mathcal{F}$ is bounded and that $\mathcal{A}$ is surjective. When $\mathcal{F}$ is bounded, Lemma 2.2.3 implies that $\mathcal{F}(\alpha)$ is also bounded for every $\alpha > 0$. Moreover, this assumption is not very restrictive due to Theorem 4.4.9. In other words, we may bound an unbounded spectrahedron by introducing a suitable trace constraint. Under such a transformation, important properties such as range and singularity degree remain the same. The restriction on $\mathcal{A}$ ensures a unique $y \in \mathbb{R}^m$ for every $Z \in \mathrm{range}(\mathcal{A}^*)$, another property that is not restrictive and will prove convenient in the subsequent discussion.

It is not hard to see that if $\mathcal{F} \neq \emptyset$ then $\mathcal{F}(\alpha)$ has a Slater point for every $\alpha > 0$. For instance, the set $\mathcal{F} + \alpha B$ has positive definite elements and is contained in $\mathcal{F}(\alpha)$. Since $X(\alpha)$ is chosen to be the determinant maximizer over $\mathcal{F}(\alpha)$ it follows that $X(\alpha) \succ 0$ and $X(\alpha) \in \mathrm{relint}(\mathcal{F})$ for each $\alpha > 0$. We have thus shown that this central path satisfies Assumption 5.0.2 *(i)*. In the remainder of this section we show that it also possesses the other properties of Assumption 5.0.2. Namely, that $\{X(\alpha) : \alpha > 0\}$ is smooth and converges to a matrix in $\mathrm{relint}(\mathcal{F})$ as $\alpha \searrow 0$.

6.1.1 Optimality Conditions and the Central Path

Let us derive the optimality conditions for the optimization problem defining $X(\alpha)$ in (6.1.1). Similar problems have been thoroughly studied throughout the literature in matrix completions and SDP, e.g., [3, 30, 82, 86]. Nonetheless, we include a proof for completeness and to emphasize its simplicity.

Theorem 6.1.2. *For every $\alpha > 0$ there exists unique $X(\alpha) \in \mathcal{F}(\alpha) \cap \mathbb{S}^n_{++}$ such that,*

$$X(\alpha) = \arg\max\{\alpha \log\det(X) : X \in \mathcal{F}(\alpha)\}. \tag{6.1.2}$$

Moreover, $X(\alpha)$ satisfies (6.1.2) if, and only if, there exists unique $y(\alpha) \in \mathbb{R}^m$ and unique $Z(\alpha) \succ 0$ such that,

$$\begin{bmatrix} \mathcal{A}^*(y(\alpha)) - Z(\alpha) \\ \mathcal{A}(X(\alpha)) - b(\alpha) \\ Z(\alpha)X(\alpha) - \alpha I \end{bmatrix} = 0. \tag{6.1.3}$$

Proof. By Assumption 6.1.1, $\mathcal{F} \neq \emptyset$ and bounded. Therefore, Lemma 2.2.3 implies that,

$$\text{null}(\mathcal{A}) \cap \mathbb{S}^n_+ = \{0\}. \tag{6.1.4}$$

Let $\alpha > 0$. Since $\mathcal{F}(\alpha) = \mathcal{F}(\mathcal{A}, b(\alpha))$, Lemma 2.2.3 and (6.1.4) also imply that $\mathcal{F}(\alpha)$ is bounded. Next, $\log\det(\cdot)$ is a strictly concave function over $\mathcal{F}(\alpha) \cap \mathbb{S}^n_{++}$ (a so-called barrier function) and,

$$\lim_{\det(X) \searrow 0} \alpha \log\det(X) = -\infty.$$

Thus, we conclude that the optimum $X(\alpha) \in \mathcal{F}(\alpha) \cap \mathbb{S}^n_{++}$ exists and is unique. The Lagrangian of problem (6.1.2) is,

$$\begin{aligned} L(X, y) &= \alpha \log\det(X) - \langle y, \mathcal{A}(X) - b(\alpha) \rangle \\ &= \alpha \log\det(X) - \langle \mathcal{A}^*(y), X \rangle + \langle y, b(\alpha) \rangle. \end{aligned} \tag{6.1.5}$$

Since the constraints are linear, stationarity of the Lagrangian holds at $X(\alpha)$. Hence there exists $y(\alpha) \in \mathbb{R}^m$ such that $\alpha(X(\alpha))^{-1} = \mathcal{A}^*(y(\alpha)) =: Z(\alpha)$. Clearly $Z(\alpha)$ is unique, and since $\mathcal{A}$ is surjective, we conclude in addition that $y(\alpha)$ is unique as well. □

The optimality conditions for (6.1.1) yield the primal-dual central path,

$$\left\{ (X(\alpha), y(\alpha), Z(\alpha)) \in \mathbb{S}^n_{++} \times \mathbb{R}^m \times \mathbb{S}^n_{++} : \begin{bmatrix} \mathcal{A}^*(y(\alpha)) - Z(\alpha) \\ \mathcal{A}(X(\alpha)) - b(\alpha) \\ Z(\alpha)X(\alpha) - \alpha I \end{bmatrix} = 0, \ \alpha > 0 \right\}. \tag{6.1.6}$$

6.1.2 Convergence to the Relative Interior

In this section we show that the central path of (6.1.6) has cluster points and that any cluster point, say $(\bar{X}, \bar{y}, \bar{Z})$, satisfies,

$$\left(\bar{X}, \bar{y}, \bar{Z}\right) \in \text{relint}(\mathcal{F}) \times \mathbb{R}^m \times \text{relint}\left(\mathcal{E}(\mathcal{A}, b)\right).$$

Recalling the definition of $\mathcal{E}(\mathcal{A}, b)$, we see that $\bar{Z}$ is a suitable choice for the first step of the facial reduction algorithm. Moreover, it has maximum rank over all suitable choices in the first step. The proof of this result was alluded to by Goldfarb and Scheinberg in [27]. Here we present the proof in its entirety by developing useful bounds on the eigenvalue functions for $X(\alpha)$ and $Z(\alpha)$. Some of the convergence bounds on eigenvalues are also derived by Mohammad-Nezhad and Terlaky in [55] for a different central path.

We begin by showing that the X component of the parametric path has cluster points and that they lie in $\mathcal{F}$.

Lemma 6.1.3. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a spectrahedron satisfying Assumption 6.1.1 and let $(X(\alpha), y(\alpha), Z(\alpha))$ be the central path of (6.1.6). Let $\{\alpha_k\}_{k\in\mathbb{N}}$ be a sequence with $\alpha_k \searrow 0$. Then there exists a subsequence $\{\alpha_{k_\ell}\}_{\ell\in\mathbb{N}}$ such that $\{X(\alpha_{k_\ell})\}$ is convergent. Moreover, for every such subsequence, there exists $\bar{X} \in \mathcal{F}$ such that $X(\alpha_{k_\ell}) \to \bar{X}$.*

Proof. Since $\alpha_k \searrow 0$, there exists $\bar{\alpha} > 0$ such that $\{\alpha_k\} \subset (0, \bar{\alpha}]$. First we show that the sequence $X(\alpha_k)$ is bounded. For any $k \in \mathbb{N}$ we have

$$\|X(\alpha_k)\|_2 \leq \|X(\alpha_k) + (\bar{\alpha} - \alpha_k)B\|_2 \leq \max_{X\in\mathcal{F}(\bar{\alpha})} \|X\|_2 < +\infty.$$

The second inequality is due to $X(\alpha_k) + (\bar{\alpha} - \alpha_k)B \in \mathcal{F}(\bar{\alpha})$ and the third inequality holds since $\mathcal{F}(\bar{\alpha})$ is bounded. Thus there exists a subsequence $\{\alpha_{k_\ell}\}_{\ell\in\mathbb{N}}$ and $\bar{X} \in \mathbb{S}^n$ such that $X(\alpha_{k_\ell}) \to \bar{X}$. That $\bar{X} \in \mathcal{F}$ is clear. □

For the dual variables we need only prove that $Z(\alpha)$ converges (for a subseqence) since this implies that $y(\alpha)$ also converges, by the assumption that $\mathcal{A}$ is surjective. As for $X(\alpha)$, we show that the tail of the parametric path corresponding to $Z(\alpha)$ is bounded. To this end, we first prove the following technical lemma.

Lemma 6.1.4. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a spectrahedron satisfying Assumption 6.1.1 and let $(X(\alpha), y(\alpha), Z(\alpha))$ be the central path of (6.1.6). Recall that $B \succ 0$, let $X_0 \in \operatorname{relint}(\mathcal{F})$, and let $Z_0 \in \operatorname{relint}(\mathcal{E}(\mathcal{A}, b))$. Then,*

(i) $\langle X(\alpha)^{-1}, X_0 + \alpha B\rangle = \mathcal{O}(1)$ *as* $\alpha \searrow 0$,

(ii) $\langle X(\alpha), Z_0 + \alpha B\rangle = \mathcal{O}(\alpha)$ *as* $\alpha \searrow 0$.

Proof. Let $\bar{\alpha} > 0$ be fixed and consider $\alpha \in (0, \bar{\alpha}]$. For *(i)* we have,

$$\begin{aligned}
\langle X(\bar{\alpha})^{-1} - X(\alpha)^{-1}, X_0 + \bar{\alpha}B - X(\alpha)\rangle &= \langle \frac{1}{\bar{\alpha}}\mathcal{A}^*(y(\bar{\alpha})) - \frac{1}{\alpha}\mathcal{A}^*(y(\alpha)), X_0 + \bar{\alpha}I - X(\alpha)\rangle, \\
&= \langle \frac{1}{\bar{\alpha}}y(\bar{\alpha}) - \frac{1}{\alpha}y(\alpha), \mathcal{A}(X_0 + \bar{\alpha}B) - \mathcal{A}(X(\alpha))\rangle, \\
&= \langle \frac{1}{\bar{\alpha}}y(\bar{\alpha}) - \frac{1}{\alpha}y(\alpha), (\bar{\alpha} - \alpha)\mathcal{A}(B)\rangle, \\
&= \langle X(\bar{\alpha})^{-1} - X(\alpha)^{-1}, (\bar{\alpha} - \alpha)B\rangle.
\end{aligned}$$

Rearranging we get,

$$\langle X(\bar{\alpha})^{-1} - X(\alpha)^{-1}, X_0 + \alpha B - X(\alpha)\rangle = 0.$$

Then,

$$\begin{aligned}
\langle X(\alpha)^{-1}, X_0 + \alpha B\rangle &= \langle X(\alpha)^{-1}, X(\alpha)\rangle + \langle X(\bar{\alpha})^{-1}, X_0 + \alpha B - X(\alpha)\rangle \\
&= n + \langle X(\bar{\alpha})^{-1}, X_0\rangle + \alpha\langle X(\bar{\alpha})^{-1}, B\rangle - \langle X(\bar{\alpha})^{-1}, X(\alpha)\rangle \\
&\leq n + \langle X(\bar{\alpha})^{-1}, X_0\rangle + \bar{\alpha}\langle X(\bar{\alpha})^{-1}, B\rangle.
\end{aligned}$$

We obtain (*i*) by observing that the right hand side is a positive constant.

For (*ii*) let y_0 be such that $Z_0 = \mathcal{A}^*(y_0)$. Recall that $\langle b, y_0\rangle = 0$. Then,

$$\begin{aligned}\langle X(\alpha), Z_0 + \alpha B\rangle &= \langle X(\alpha), \mathcal{A}^*(y_0)\rangle + \alpha\langle X(\alpha), B\rangle \\ &= \langle b(\alpha), y_0\rangle + \alpha\langle X(\alpha), B\rangle \\ &= \alpha\langle \mathcal{A}(B), y_0\rangle + \alpha\langle X(\alpha), B\rangle \\ &= \alpha\langle Z_0 + X(\alpha), B\rangle \\ &\leq \alpha \sup_{\alpha\in(0,\bar{\alpha}]} \langle Z_0 + X(\alpha), B\rangle.\end{aligned}$$

Since $\langle Z_0 + X(\bar{\alpha}), B\rangle$ is positive and $X(\alpha)$ is bounded over $(0, \bar{\alpha}]$, it follows that,

$$\sup_{\alpha\in(0,\bar{\alpha}]} \langle Z_0 + X(\alpha), B\rangle \in \mathbb{R}_{++},$$

implying (*ii*), as desired. □

Next we provide bounds on the eigenvalues of $X(\alpha)$.

Theorem 6.1.5. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a spectrahedron satisfying Assumption 6.1.1 and let $(X(\alpha), y(\alpha), Z(\alpha))$ be the central path of (6.1.6). Let r and q denote the rank of $\mathcal{F}$ and $\mathcal{E}(\mathcal{A}, b)$, respectivley. Then as $\alpha \searrow 0$,*

$$\lambda_i(X(\alpha)) = \begin{cases} \Theta(1) & i \leq r, \\ \Omega(\alpha) & i \in \{r+1, n-q\}, \\ \Theta(\alpha) & i \geq n-q+1. \end{cases} \tag{6.1.7}$$

Proof. Let X_0 and Z_0 be as in Lemma 6.1.4. Combining Lemma 6.1.4 (*i*) with Fact 2.1.3 we get that

$$\sum_{i=1}^{n} \lambda_{n+1-i}(X(\alpha)^{-1})\lambda_i(X_0 + \alpha B) = \mathcal{O}(1) \text{ as } \alpha \searrow 0.$$

The left hand side is a sum of positive terms, hence the bound applies to each term. Therefore, there exists $M > 0$ such that,

$$\lambda_{n+1-i}(X(\alpha)^{-1})\lambda_i(X_0 + \alpha B) < M, \quad \forall i \in \{1, \ldots, n\}. \tag{6.1.8}$$

Now $\lambda_{n+1-i}(X(\alpha)^{-1}) = \lambda_i(X(\alpha))^{-1}$ and thus (6.1.8) becomes,

$$\lambda_i(X(\alpha)) > \frac{1}{M}\lambda_i(X_0 + \alpha B) \geq \frac{1}{M}\left(\lambda_i(X_0) + \alpha\lambda_n(B)\right), \quad \forall i \in \{1, \ldots, n\}. \tag{6.1.9}$$

The second inequality in (6.1.9) is due to Corollary 1.21 of [80]. Now the r largest eigenvalues of X_0 are positive. Thus (6.1.9) yields,

$$\lambda_i(X(\alpha)) > \frac{1}{M}\lambda_i(X_0), \quad \forall i \in \{1, \ldots, r\}.$$

Equivalently $\lambda_i(X(\alpha)) = \Omega(1)$ when $i \leq r$. Moreover, in Lemma 6.1.3 we proved that $X(\alpha)$ is bounded on any interval of the form $(0, \bar{\alpha}]$ with $\bar{\alpha} > 0$. It follows that $\lambda_i(X(\alpha)) = \mathcal{O}(1)$ for all $i \in \{1, \dots, n\}$, proving the first case of (6.1.7). Now the $n - r$ smallest eigenvalues of X_0 are exactly 0, hence from (6.1.9) we get,

$$\lambda_i(X(\alpha)) > \frac{\lambda_n(B)}{M}\alpha, \quad \forall i \in \{r+1, \dots, n\}.$$

It follows that $\lambda_i(X(\alpha)) = \Omega(\alpha)$ when $i \in \{r+1, \dots, n\}$, proving the second case of (6.1.7). All that remains is to show that $\lambda_i(X(\alpha)) = \mathcal{O}(\alpha)$ when $i \geq n - q + 1$. To see this we combine Lemma 6.1.4 *(ii)* with Fact 2.1.3 to obtain,

$$\sum_{i=1}^{n} \lambda_i(X(\alpha))\lambda_{n+1-i}(Z_0 + \alpha B) = \mathcal{O}(\alpha) \text{ as } \alpha \searrow 0.$$

Then there exists $M' > 0$ such that for all $i \in \{1, \dots, n\}$,

$$\lambda_i(X(\alpha)) < \frac{M'}{\lambda_{n+1-i}(Z_0 + \alpha B)}\alpha \leq \frac{M'}{\lambda_{n+1-i}(Z_0) + \alpha\lambda_n(B)}\alpha. \tag{6.1.10}$$

The second inequality holds by the same reasoning used in (6.1.9). Now when $i \geq n-q+1$ then $n + 1 - i \leq q$ and $\lambda_{n+1-i}(Z_0) > 0$. It follows from (6.1.10) that $\lambda_i(X(\alpha)) = \mathcal{O}(\alpha)$ whenever $i \geq n-q+1$. Together with the lower bound $\lambda_i(X(\alpha)) = \Omega(\alpha)$ for $i \geq n-q+1$, we get $\lambda_i(X(\alpha)) = \Theta(\alpha)$ when $i \geq n - q + 1$, as desired. □

We now have the necessary tools to prove a result about limit points of the parametric path.

Theorem 6.1.6. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a spectrahedron satisfying Assumption 6.1.1 and let $(X(\alpha), y(\alpha), Z(\alpha))$ be the central path of (6.1.6). Let $\{\alpha_k\}_{k\in\mathbb{N}}$ be a sequence such that $\alpha_k \searrow 0$. Then there exists a subsequence $\{\alpha_{k_\ell}\}_{\ell\in\mathbb{N}}$ such that $(X(\alpha_{k_\ell}), y(\alpha_{k_\ell}), Z(\alpha_{k_\ell}))$ converges. Moreover, for every such subsequence the limit point, say $(\bar{X}, \bar{y}, \bar{Z})$, satisfies,*

$$\bar{X} \in \text{relint}(\mathcal{F}), \;\; \bar{Z} \in \text{relint}(\mathcal{E}(\mathcal{A}, b)), \;\; \bar{Z} = \mathcal{A}^*(\bar{y}).$$

Proof. We may, without loss of generality, assume that $X(\alpha_k) \to \bar{X} \in \mathcal{F}$ due to Lemma 6.1.3. By Theorem 6.1.5 the r largest eigenvalues of $X(\alpha)$ are bounded below, hence $\text{rank}(\bar{X}) = r$ and we conclude that $\bar{X} \in \text{relint}(\mathcal{F})$.

Next we look at the sequence $\{Z(\alpha_k)\}_{k\in\mathbb{N}}$. By Theorem 6.1.5 we have for $i \in \{1, \dots, q\}$ that,

$$\lambda_i(Z(\alpha)) = \alpha\lambda_i(X(\alpha)^{-1}) = \frac{\alpha}{\lambda_{n+1-q}(X(\alpha))} = \frac{\alpha}{\Theta(\alpha)} = \Theta(1). \tag{6.1.11}$$

In particular, the largest eigenvalue of $Z(\alpha_k)$ is bounded in magnitude, hence $\|Z(\alpha_k)\|_2$ is bounded. Then there exists a subsequence $\{\alpha_{k_\ell}\}_{\ell\in\mathbb{N}}$ such that $Z(\alpha_{k_\ell}) \to \bar{Z}$, for some $\bar{Z} \in \mathbb{S}^n$.

Clearly $\bar{Z} \in \mathbb{S}^n_+ \cap \operatorname{range}(\mathcal{A}^*)$ by closure and the fact that $Z(\alpha) \in \mathbb{S}^n_+ \cap \operatorname{range}(\mathcal{A}^*)$ for every $\alpha > 0$. Moreover, $\langle X(\alpha), Z(\alpha)\rangle \to \langle \bar{X}, \bar{Z}\rangle$ and,

$$\langle X(\alpha), Z(\alpha)\rangle = \langle X(\alpha), \alpha X(\alpha)^{-1}\rangle = \alpha n \to 0.$$

Hence $\langle \bar{X}, \bar{Z}\rangle = 0$ and it follows that $\bar{Z} \in \mathcal{E}(\mathcal{A}, b)$. Now, by (6.1.11) the q largest eigenvalues of $Z(\alpha_{k_\ell})$ are bounded below. Therefore, $\operatorname{rank}(\bar{Z}) = q$ and consequently $\bar{Z} \in \operatorname{relint}(\mathcal{E}(\mathcal{A}, b))$. Convergence of $\{y(\alpha_{k_\ell})\}_{\ell\in\mathbb{N}}$ follows from the assumption that $\mathcal{A}$ is surjective. □

An immediate consequence of Theorem 6.1.6 is an extension of Theorem 6.1.5.

Corollary 6.1.7. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a spectrahedron satisfying Assumption 6.1.1 and let $(X(\alpha), y(\alpha), Z(\alpha))$ be the central path of (6.1.6). Let $(\bar{X}, \bar{y}, \bar{Z})$ be a limit point of the central path and let r and q denote the rank of $\mathcal{F}$ and $\mathcal{E}(\mathcal{A}, b)$, respectively. Then as $\alpha \searrow 0$ it holds that,*

$$\lambda_i(X(\alpha)) = \begin{cases} \Theta(1) & i \le r, \\ \Omega(\alpha) \text{ and } \neq \mathcal{O}(\alpha) & i \in \{r+1, \dots, n-q\}, \\ \Theta(\alpha) & i \ge n-q+1, \end{cases} \tag{6.1.12}$$

and

$$\lambda_i(Z(\alpha)) = \begin{cases} \Theta(1) & i \le q, \\ \mathcal{O}(1) \text{ and } \neq \Omega(1) & i \in \{q+1, \dots, n-r\}, \\ \Theta(\alpha) & i \ge n-r+1. \end{cases} \tag{6.1.13}$$

Proof. By construction of the central path it holds that for every $i \in \{1, \dots, n\}$,

$$\lambda_i(Z(\alpha)) = \lambda_i\left(\alpha X(\alpha)^{-1}\right) = \frac{\alpha}{\lambda_{n+1-i}(X(\alpha))}. \tag{6.1.14}$$

Applying the bounds of Theorem 6.1.5 to (6.1.14) gives us,

$$\lambda_i(Z(\alpha)) = \begin{cases} \Theta(1) & i \le q, \\ \mathcal{O}(1) & i \in \{q+1, \dots, n-r\}, \\ \Theta(\alpha) & i \ge n-r+1. \end{cases} \tag{6.1.15}$$

Moreover, when $i \in \{q+1, \dots, n-r\}$ it holds that $\lambda_i(Z(\alpha)) \neq \Omega(1)$ otherwise every limit point of $Z(\alpha)$ would have $\lambda_i(\bar{Z}) > 0$ implying that $\operatorname{rank}(\bar{Z}) > \operatorname{rank}(\mathcal{E}(\mathcal{A}, b))$, a contradiction. It follows by (6.1.14) that $\lambda_i(X(\alpha)) \neq \mathcal{O}(\alpha)$ when $i \in \{r+1, \dots, n-q\}$, as desired. □

6.1.3 Smoothness

We conclude this section by proving that the parametric path is smooth and has a limit point as $\alpha \searrow 0$. Our proof relies on a result of Milnor and is motivated by a similar proof for the classical central path of SDP in [31, 32]. Recall that an *algebraic set* is the solution set of a system of finitely many polynomial equations. We denote set closure by $\mathrm{cl}(\cdot)$.

Fact 6.1.8 (Milnor's Lemma [54])**.** *Let $\mathcal{V} \subseteq \mathbb{R}^k$ be an algebraic set and let $\mathcal{U} \subseteq \mathbb{R}^k$ be an open set defined by finitely many polynomial inequalities. If $0 \in \mathrm{cl}(\mathcal{U} \cap \mathcal{V})$, then there exists $\varepsilon > 0$ and a real analytic curve $p : [0, \varepsilon) \to \mathbb{R}^k$ such that $p(0) = 0$ and $p(t) \in \mathcal{U} \cap \mathcal{V}$ whenever $t > 0$.*

Theorem 6.1.9. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a spectrahedron satisfying Assumption 6.1.1 and let $(X(\alpha), y(\alpha), Z(\alpha))$ be the central path of (6.1.6). Then there exists $(\bar{X}, \bar{y}, \bar{Z})$ such that,*

$$\lim_{\alpha \searrow 0} (X(\alpha), y(\alpha), Z(\alpha)) = (\bar{X}, \bar{y}, \bar{Z}) \in \mathrm{relint}(\mathcal{F}) \times \mathbb{R}^m \times \mathrm{relint}\left(\mathcal{E}(\mathcal{A}, b)\right),$$

where $\bar{Z} = \mathcal{A}^*(\bar{y})$.

Proof. Let $(\bar{X}, \bar{y}, \bar{Z})$ be a cluster point of the parametric path as in Theorem 6.1.6. We define the set $\mathcal{U}$ as

$$\mathcal{U} := \{(X, y, Z, \alpha) \in \mathbb{S}^n \times \mathbb{R}^m \times \mathbb{S}^n \times \mathbb{R} : \bar{X} + X \succ 0,\ \bar{Z} + Z \succ 0,\ Z = \mathcal{A}^*(y),\ \alpha > 0\}.$$

Note that each of the positive definite constraints is equivalent to n strict determinant (polynomial) inequalities. Therefore, $\mathcal{U}$ satisfies the assumptions of Fact 6.1.8. Next, let us define the set $\mathcal{V}$ as,

$$\mathcal{V} := \left\{ (X, y, Z, \alpha) \in \mathbb{S}^n \times \mathbb{R}^m \times \mathbb{S}^n \times \mathbb{R} : \begin{bmatrix} \mathcal{A}^*(y) - Z \\ \mathcal{A}(X) - \alpha \mathcal{A}(B) \\ (\bar{Z} + Z)(\bar{X} + X) - \alpha I \end{bmatrix} = 0 \right\},$$

and note that $\mathcal{V}$ is indeed a real algebraic set. Next we show that there is a one-to-one correspondance between $\mathcal{U} \cap \mathcal{V}$ and the parametric path without any of its cluster points. Consider $(\tilde{X}, \tilde{y}, \tilde{Z}, \tilde{\alpha}) \in \mathcal{U} \cap \mathcal{V}$ and let $(X(\tilde{\alpha}), y(\tilde{\alpha}), Z(\tilde{\alpha}))$ be a point on the parametric path. We show that

$$(\bar{X} + \tilde{X}, \bar{y} + \tilde{y}, \bar{Z} + \tilde{Z}) = (X(\tilde{\alpha}), y(\tilde{\alpha}), Z(\tilde{\alpha})). \tag{6.1.16}$$

First of all $\bar{X} + \tilde{X} \succ 0$ and $\bar{Z} + \tilde{Z} \succ 0$ by inclusion in $\mathcal{U}$. Secondly, $(\bar{X} + \tilde{X}, \bar{y} + \tilde{y}, \bar{Z} + \tilde{Z})$ solves the system of (6.1.6) when $\alpha = \tilde{\alpha}$. Indeed,

$$\begin{bmatrix} \mathcal{A}^*(\bar{y} + \tilde{y}) - (\bar{Z} + \tilde{Z}) \\ \mathcal{A}(\bar{X} + \tilde{X}) - b(\tilde{\alpha}) \\ (\bar{Z} + \tilde{Z})(\bar{X} + \tilde{X}) - \tilde{\alpha} I \end{bmatrix} = \begin{bmatrix} \mathcal{A}^*(\bar{y}) - \bar{Z} + (\mathcal{A}^*(\tilde{y}) - \tilde{Z}) \\ b + \tilde{\alpha}\mathcal{A}(B) - b(\tilde{\alpha}) \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Since the system defining the parametric path has a unique solution, (6.1.16) holds. Thus,

$$(\tilde{X}, \tilde{y}, \tilde{Z}) = (X(\alpha) - \bar{X}, y(\alpha) - \bar{y}, Z(\alpha) - \bar{Z}),$$

and it follows that $\mathcal{U} \cap \mathcal{V}$ is a translation of the parametric path (without its cluster points):

$$\mathcal{U} \cap \mathcal{V} = \{(X, y, Z, \alpha) : (X, y, Z) = (X(\alpha) - \bar{X}, y(\alpha) - \bar{y}, Z(\alpha) - \bar{Z}),\ \alpha > 0\}. \quad (6.1.17)$$

Next, we show that $0 \in \mathrm{cl}(\mathcal{U} \cap \mathcal{V})$. To see this, note that

$$(X(\alpha), y(\alpha), Z(\alpha)) \to (\bar{X}, \bar{y}, \bar{Z}),$$

as $\alpha \searrow 0$ along a subsequence. Therefore, along the same subsequence, we have

$$(X(\alpha) - \bar{X}, y(\alpha) - \bar{y}, Z(\alpha) - \bar{Z}, \alpha) \to 0.$$

Each of the elements of this subsequence belongs to $\mathcal{U} \cap \mathcal{V}$ by (6.1.17) and it follows that $0 \in \mathrm{cl}(\mathcal{U} \cap \mathcal{V})$.

We have shown that $\mathcal{U}$ and $\mathcal{V}$ satisfy the assumptions of Fact 6.1.8. Therefore, there exists $\varepsilon > 0$ and an analytic curve $p : [0, \varepsilon) \to \mathbb{S}^n \times \mathbb{R}^m \times \mathbb{S}^n \times \mathbb{R}$ such that $p(0) = 0$ and $p(t) \in \mathcal{U} \cap \mathcal{V}$ whenever $t > 0$. Let

$$p(t) = (X_{(t)}, y_{(t)}, Z_{(t)}, \alpha_{(t)}),$$

and observe that by (6.1.17), we have

$$(X_{(t)}, y_{(t)}, Z_{(t)}, \alpha_{(t)}) = (X(\alpha_{(t)}) - \bar{X}, y(\alpha_{(t)}) - \bar{y}, Z(\alpha_{(t)}) - \bar{Z}). \quad (6.1.18)$$

Since p is a real analytic curve, the map $g : [0, \varepsilon) \to \mathbb{R}$ defined as $g(t) = \alpha_{(t)}$, is a differentiable function on the open interval $(0, \varepsilon)$ with

$$\lim_{t \searrow 0} g(t) = 0.$$

In particular, this implies that there is an interval $[0, \bar{\varepsilon}) \subseteq [0, \varepsilon)$ where g is monotonic. It follows that on $[0, \bar{\varepsilon})$, g^{-1} is a well defined continuous function that converges to 0 from the right. Note that for any $t > 0$, $(X(t), y(t), Z(t))$ is on the parametric path. Therefore,

$$\lim_{t \searrow 0} X(t) = \lim_{t \searrow 0} X(g(g^{-1}(t))) = \lim_{t \searrow 0} X(\alpha_{(g^{-1}(t))}).$$

Substituting with (6.1.18), we have

$$\lim_{t \searrow 0} X(t) = \lim_{t \searrow 0} X_{(g^{-1}(t))} + \bar{X} = \bar{X}.$$

Similarly, $y(t)$ and $Z(t)$ converge to $\bar{y}$ and $\bar{Z}$, respectively. Thus every cluster point of the parametric path is identical to $(\bar{X}, \bar{y}, \bar{Z})$. □

We have shown that the tail of the parametric path is smooth and it has a limit point. Smoothness of the entire path follows from Example 5.22 of [72] or the classical Maximum Theorem of [5].

6.2 Singularity Degree and Irregular Convergence

As the main result of this chapter, we show that large singularity degree leads to undesirable convergence properties for the central path of (6.1.6). An immediate implication of Corollary 6.1.7 is that 'fast' convergence of vanishing eigenvalues does not occur when singularity degree is greater than 1.

Theorem 6.2.1. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a spectrahedron satisfying Assumption 6.1.1 and let $(X(\alpha), y(\alpha), Z(\alpha))$ be the central path of (6.1.6). Let r denote the rank of $\mathcal{F}$. Then,*

$$\lambda_i(X(\alpha)) = \mathcal{O}(\alpha), \ \forall i \in \{r+1, \dots, n\} \iff \operatorname{sd}(\mathcal{F}) = 1.$$

In the main result of this section we show that larger singularity degree leads to greater irregularity in the way that components of $Z(\alpha)$ converge. To simplify the proof we introduce three lemmas. First, we show that backward error is 'small' for the central path of (6.1.6), allowing us to use results from Chapter 5.

Lemma 6.2.2. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a spectrahedron satisfying Assumption 6.1.1 and let $(X(\alpha), y(\alpha), Z(\alpha))$ be the central path of (6.1.6). Then on $\mathbb{R}_{++}$ it holds that,*

$$\epsilon^b(X(\alpha), \mathcal{F}) = \mathcal{O}(\alpha).$$

Proof. By definition of backward error,

$$\epsilon^b(X(\alpha), \mathcal{F}) = \operatorname{dist}(X(\alpha), \mathbb{S}^n_+) + \operatorname{dist}(X(\alpha), \mathcal{L}(\mathcal{A}, b)) = \operatorname{dist}(X(\alpha), \mathcal{L}(\mathcal{A}, b)),$$

since $X(\alpha) \in \mathbb{S}^n_+$ for all $\alpha > 0$ by construction. Furthermore, by definition of $\operatorname{dist}(\cdot)$ we have,

$$\epsilon^b(X(\alpha), \mathcal{F}) = \inf_{Y \in \mathcal{L}(\mathcal{A}, b)} \|X(\alpha) - Y\| = \|X(\alpha) - \bar{Y}\|,$$

where $\bar{Y}$ is the orthogonal projection of $X(\alpha)$ onto $\mathcal{L}(\mathcal{A}, b)$. In other words, backward error is the norm of a matrix $\bar{P} \in \mathbb{S}^n$ such that $\bar{P}$ has minimum norm over all matrices P satisfying,

$$X(\alpha) - P \in \mathcal{L}(\mathcal{A}, b). \tag{6.2.1}$$

We claim that $\alpha \mathcal{A}^\dagger \mathcal{A}(B)$ is a possible choice for P in (6.2.1). Indeed,

$$\mathcal{A}\left(X(\alpha) - \alpha \mathcal{A}^\dagger \mathcal{A}(B)\right) = b(\alpha) - \alpha \mathcal{A}(B) = b.$$

Now B is positive definite. So by the assumption that $\mathcal{F}$ is bounded and Lemma 2.2.3 it holds that $B \notin \operatorname{null}(\mathcal{A})$ and therefore,

$$\|\mathcal{A}^\dagger \mathcal{A}(B)\| > 0. \tag{6.2.2}$$

By the characterization of backward error and by (6.2.2) we have,

$$\epsilon^b(X(\alpha), \mathcal{F}) = \|\bar{P}\| \leq \|\alpha \mathcal{A}^\dagger \mathcal{A}(B)\| = \mathcal{O}(\alpha),$$

as desired. □

Next we bound the magnitude of the projection of $y(\alpha)$ onto $\text{span}(b)$.

Lemma 6.2.3. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a spectrahedron satisfying Assumption 6.1.1 and let $(X(\alpha), y(\alpha), Z(\alpha))$ be the central path of (6.1.6). Suppose that $b \neq 0$ and let,*

$$v^1, \ldots, v^{m-1} \in \mathbb{R}^m,$$

form a basis for $b^\perp$. Let $\bar{\alpha} > 0$ and for all $\alpha \in (0, \bar{\alpha})$ let $\beta(\alpha)$ and $\nu_1(\alpha), \ldots, \nu_{m-1}(\alpha)$ be real coefficients such that,

$$y(\alpha) = \beta(\alpha) b + \sum_{i=1}^{m-1} \nu_i(\alpha) v^i.$$

Then for all $\alpha \in (0, \bar{\alpha})$, it holds that $|\beta(\alpha)| = \mathcal{O}(\alpha)$.

Proof. By definition of $b(\alpha)$ in (6.1.1) and by (6.1.6) we have,

$$\begin{aligned} y(\alpha)^T b &= y(\alpha)^T (b(\alpha) - \alpha \mathcal{A}(B)) \\ &= y(\alpha)^T (\mathcal{A}(X(\alpha)) - \alpha \mathcal{A}(B)) \\ &= \langle \mathcal{A}^*(y(\alpha)), X(\alpha) \rangle - \alpha \langle \mathcal{A}^*(y(\alpha)), B \rangle \\ &= \alpha \langle X(\alpha)^{-1}, X(\alpha) \rangle - \alpha \langle Z(\alpha), B \rangle \\ &= \alpha (n - \langle Z(\alpha), B \rangle). \end{aligned}$$

Now $n - \langle Z(\alpha), B \rangle$ is bounded in magnitude on $\alpha \in (0, \bar{\alpha})$, implying that $|y(\alpha)^T b| = \mathcal{O}(\alpha)$. On the other hand, $\{v^1, v^2, \ldots, v^{m-1}\} \in b^\perp$ by construction. Therefore $y(\alpha)^T b = \beta(\alpha) \|b\|^2$, yielding,

$$|\beta(\alpha)| = \frac{|y(\alpha)^T b|}{\|b\|^2} = \mathcal{O}(\alpha),$$

as desired. □

In the final lemma preceding the main result we bound the norm of principal submatrices of $Z(\alpha)$ that consist of at least $r + 1$ rows.

Lemma 6.2.4. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a spectrahedron satisfying Assumption 6.1.1 and let $(X(\alpha), y(\alpha), Z(\alpha))$ be the central path of (6.1.6). Let $r := \text{rank}(\mathcal{F})$. Let $\widehat{Z}(\alpha) \in \mathbb{S}^{\hat{n}}$ be a principal submatrix of $Z(\alpha)$ for a positive integer $\hat{n} \leq n$. Then it holds that,*

(i) $\|\widehat{Z}(\alpha)\| = \Omega(\alpha)$ *as* $\alpha \searrow 0$,

(ii) if $\hat{n} \geq r + 1$ *then* $\|\widehat{Z}(\alpha)\| = \Omega(\alpha^{1-\xi(1)})$ *as* $\alpha \searrow 0$.

Proof. By interlacing eigenvalues, Fact 2.1.2, and by Corollary 6.1.7 we have,

$$\|\widehat{Z}(\alpha)\| = \Theta\left(\lambda_1\left(\widehat{Z}(\alpha)\right)\right) \geq \Theta\left(\lambda_{\min}\left(\widehat{Z}(\alpha)\right)\right) \geq \Theta\left(\lambda_n\left(Z(\alpha)\right)\right) = \Theta(\alpha),$$

implying (*i*). For (*ii*) we apply interlacing eigenvalues with the additional assumption to get,

$$\|\widehat{Z}(\alpha)\|_2 = \lambda_1\left(\widehat{Z}(\alpha)\right) \geq \lambda_{n-r}(Z(\alpha)).$$

Now Lemma 6.2.2 allows us to apply a bound of Lemma 5.1.2 to get,

$$\begin{aligned}\|\widehat{Z}(\alpha)\|_2 &\geq \lambda_{n-r}\left(\alpha X(\alpha)^{-1}\right)\\ &= \alpha\lambda_{r+1}(X(\alpha))^{-1}\\ &= \alpha\Omega\left(\alpha^{-\xi(1)}\right)\\ &= \Omega\left(\alpha^{1-\xi(1)}\right).\end{aligned}$$

The bound readily extends to any other norm as well. □

After a suitable orthogonal transformation, $Z(\alpha)$ admits a block partition where at least $\mathrm{sd}(\mathcal{F})$ of these blocks converge to 0, each at a different rate.

Theorem 6.2.5. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a spectrahedron satisfying Assumption 6.1.1 and let $(X(\alpha), y(\alpha), Z(\alpha))$ be the central path of (6.1.6). Let $\bar{\alpha} > 0$ be fixed. Then there exists an integer $d \in [\mathrm{sd}(\mathcal{F}), \bar{m}]$ where,*

$$\bar{m} = \begin{cases} m & \text{if } b = 0,\\ m-1 & \text{otherwise,}\end{cases}$$

and a suitable orthogonal transformation of $\mathcal{F}$, such that,

$$Z(\alpha) = \begin{bmatrix} Z_{d+1}(\alpha) & * & \cdots & * \\ * & Z_d(\alpha) & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & Z_1(\alpha)\end{bmatrix},$$

where for all $\alpha \in (0, \bar{\alpha})$ it holds that,

(*i*) $Z_1(\alpha) \to S_1 \succ 0$ *and* $Z_i(\alpha) \to 0$ *for all* $i \in \{2, \ldots, d+1\}$,

(*ii*) $\frac{\lambda_{\min}(Z_i(\alpha))}{\lambda_{\max}(Z_{i+1}(\alpha))} \to \infty$ *for all* $i \in \{1, \ldots, d\}$,

(*iii*) $\lambda_{\min}(Z_i(\alpha)) = \Theta\left(\lambda_{\max}(Z_i(\alpha))\right)$ *for all* $i \in \{1, \ldots, d+1\}$,

(*iv*) $\|Z_{d+1}(\alpha)\| = \Theta(\alpha)$.

Proof. We assume, without loss of generality, that,

$$\mathrm{face}(\mathcal{F}) = \begin{bmatrix}\mathbb{S}^r_+ & 0\\ 0 & 0\end{bmatrix}. \tag{6.2.3}$$

As above, let $(\bar{X}, \bar{y}, \bar{Z})$ be the limit point of the primal-dual central path. We know from Theorem 6.1.6 that $\bar{Z}$ is an exposing vector for a face containing face$(\mathcal{F})$. Therefore $\bar{y}^T b = 0$. Without loss of generality we may assume that,

$$\bar{Z} =: \begin{bmatrix} 0 & 0 \\ 0 & S_1 \end{bmatrix}, \; S_1 \succ 0. \tag{6.2.4}$$

Now we define $y^1 := \bar{y}$ and choose $v^1, \ldots, v^{m_v} \in \text{span}\{b, y^1\}^\perp$ for some $m_v \leq m$ so that the collection of vectors $\{y^1, v^1, \ldots, v^{m_v}\}$ is a basis for $b^\perp$. Note that when $b \neq 0$, then $\{b, y^1, v^1, \ldots, v^{m_v}\}$ is also a basis for $\mathbb{R}^m$.

Now for each $\alpha > 0$ there exist coefficients $\beta(\alpha)$ and $\nu_1(\alpha), \ldots, \nu_{m_v}(\alpha)$ and $\gamma_1(\alpha)$ such that,

$$y(\alpha) = \gamma_1(\alpha) y^1 + \beta(\alpha) b + \sum_{i=1}^{m_v} \nu_i(\alpha) v^i. \tag{6.2.5}$$

Since $y(\alpha) \to y^1$ we have $\gamma_1(\alpha) \to 1$ and $\gamma_1(\alpha)$ dominates the other coefficients. That is,

$$\lim_{\alpha \searrow 0} \frac{\sum_{i=1}^{m_v} |\nu_i(\alpha)| + |\beta(\alpha)|}{\gamma_1(\alpha)} = 0. \tag{6.2.6}$$

Now let us consider a block partition of $Z(\alpha)$ according to the block partition of $\bar{Z}$ in (6.2.4). We have,

$$Z(\alpha) = \begin{bmatrix} 0 & 0 \\ 0 & \gamma_1(\alpha) S_1 \end{bmatrix} + \mathcal{A}^* \left(\beta(\alpha) b + \sum_{i=1}^{m_v} \nu_i(\alpha) v^i \right). \tag{6.2.7}$$

Note that the two diagonal blocks of $Z(\alpha)$ in (6.2.7) possess properties *(i)* and *(ii)*.

Let $Z_{11}(\alpha)$ denote the upper left block of $Z(\alpha)$. We consider two possibilities. First, suppose that $\|Z_{11}(\alpha)\| = \mathcal{O}(\alpha)$. Then the lower bound of Lemma 6.2.4 *(i)* implies that $\|Z_{11}(\alpha)\| = \Theta(\alpha)$. Now the two diagonal blocks also satisfy properties *(iii)* and *(iv)*. Setting $d = 1$ it is easy to see that $d \leq \bar{m}$. Since $\|Z_{11}(\alpha)\| = \mathcal{O}(\alpha)$ and $\xi(1) \in (0, 1)$, it holds that $\|Z_{11}(\alpha)\| \neq \Omega\left(\alpha^{1-\xi(1)}\right)$. Then Lemma 6.2.4 *(ii)* implies that $Z_{11}(\alpha)$ has at most r rows. Moreover, by our assumption on the facial structure of face$(\mathcal{F})$ in (6.2.3) we conclude that $Z_{11}(\alpha)$ has exactly r rows, otherwise $\bar{Z}$ exposes a face that is strictly smaller than face$(\mathcal{F})$. Thus we have $\text{sd}(\mathcal{F}) = 1 \leq d$, as desired.

The second possibility is that $\|Z_{11}(\alpha)\| \neq \mathcal{O}(\alpha)$. In this case, at least one of the coefficients other than $\gamma_1(\alpha)$ converges to 0 at a rate not equal to $\mathcal{O}(\alpha)$. This coefficient is not $\beta(\alpha)$, since Lemma 6.2.3 implies that $|\beta(\alpha)| = \mathcal{O}(\alpha)$ when $b \neq 0$. When $b = 0$ we may set $\beta(\alpha) = 0$ as it is irrelevant. Thus we conclude that $|\nu_i(\alpha)| \neq \mathcal{O}(\alpha)$ for some $i \in \{1, \ldots, m_v\}$.

Now let, y^2 be a limit point of $\beta(\alpha) b + \sum_{i=1}^{m_v} \nu_i(\alpha) v^i$, after normalizing. By the arguments above, $y^2 \in \text{span}\{v^1, \ldots, v^{m_v}\}$ and thus $(y^2)^T b = 0$. Secondly, $(\mathcal{A}^*(\beta(\alpha) b + \sum_{i=1}^{m_v} \nu_i(\alpha) v^i))_{11}$ is positive definite for every $\alpha > 0$ by (6.2.7). This implies that $(\mathcal{A}^*(y^2))_{11} \succeq 0$. Therefore,

if $(\mathcal{A}^*(y^2))_{11}$ is not the zero matrix it is an exposing vector in the second step of facial reduction.

Let us first address the case $(\mathcal{A}^*(y^2))_{11} = 0$. Here we let $w^1 := y^2$ and choose

$$v^1, \dots, v^{m_v} \in \operatorname{span}\{b, y^1, w^1\}^\perp$$

for some m_v, different than the previously used m_v, so that $\{y^1, w^1, v^1, \dots, v^{m_v}\}$ is a basis for $b^\perp$. Now we repeat the above process until we obtain a new y^2 such that $(\mathcal{A}^*(y^2))_{11} \neq 0$. This brings us to the second case.

Now we may assume that we have obtained y^2 as above and $(\mathcal{A}^*(y^2))_{11} \neq 0$. We also assume, without loss of generality that,

$$(\mathcal{A}^*(y^2))_{11} = \begin{bmatrix} 0 & 0 \\ 0 & S_2 \end{bmatrix}, \ S_2 \succ 0. \tag{6.2.8}$$

Then the matrix,

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & S_2 & 0 \\ 0 & 0 & S_1 \end{bmatrix}, \tag{6.2.9}$$

exposes a face containing face$(\mathcal{F})$ and this face is smaller than the one exposed by $\bar{Z}$. In other, words, we have obtained a better exposing vector. Let us assume that we have accumulated m_w vectors of the type w^1 obtained in the case $(\mathcal{A}^*(y^2))_{11} = 0$. Then we choose,

$$v^1, \dots, v^{m_v} \in \operatorname{span}\{b, y^1, y^2, w^1, \dots, w^{m_w}\}^\perp, \tag{6.2.10}$$

so that $\{y^1, y^2, w^1, \dots, w^{m_w}, v^1, \dots, v^{m_v}\}$ is a basis for $b^\perp$. As above, there exist coefficients,

$$\beta(\alpha),\ \gamma_1(\alpha), \gamma_2(\alpha),\ \omega_1(\alpha), \dots, \omega_{m_w}(\alpha),\ \nu_1(\alpha), \dots, \nu_{m_v}(\alpha), \tag{6.2.11}$$

such that,

$$y(\alpha) = \beta(\alpha)b + \sum_{i=1}^{2} \gamma_i(\alpha) y^i + \sum_{i=1}^{m_w} \omega_i(\alpha) w^i + \sum_{i=1}^{m_v} \nu_i(\alpha) v^i.$$

Then,

$$Z(\alpha) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \gamma_2(\alpha) S_2 & 0 \\ 0 & 0 & \gamma_1(\alpha) S_1 \end{bmatrix} + \mathcal{A}^* \left(\beta(\alpha) b + \sum_{i=1}^{m_v} \nu_i(\alpha) v^i + \sum_{i=1}^{m_w} \omega_i(\alpha) w^i \right), \tag{6.2.12}$$

where the upper left block is $\mathcal{A}^* \left(\beta(\alpha) b + \sum_{i=1}^{m_v} \nu_i(\alpha) v^i \right)_{11}$. By construction we have

$$\frac{\gamma_1(\alpha)}{\gamma_2(\alpha)} \to \infty \text{ and } \frac{\gamma_2(\alpha)}{b(\alpha) + \sum_{i=1}^{m_v} |\nu_i(\alpha)|} \to \infty. \tag{6.2.13}$$

Thus we conclude that the diagonal blocks of $Z(\alpha)$ satisfy properties *(i)* and *(ii)*. In addition, the blocks containing $\gamma_1(S_1)$ and $\gamma_2(S_2)$ satisfy property *(iii)*.

By Lemma 6.2.4 *(ii)* we may continue in this fashion until,

$$Z(\alpha) = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & \gamma_d(\alpha)S_d & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \gamma_1(\alpha)S_1 \end{bmatrix} + \mathcal{A}^* \left(\beta(\alpha)b + \sum_{i=1}^{m_v} \nu_i(\alpha)v^i + \sum_{i=1}^{m_w} \omega_i(\alpha)w^i \right), \quad (6.2.14)$$

for some positive integer d and the upper left block has norm that is $\mathcal{O}(\alpha)$. By reasoning analogous to that of the discussion following (6.2.7), we conclude that the blocks of $Z(\alpha)$, according to the block partition of (6.2.14), satisfy properties *(i)* - *(iv)* and that $d \in [\mathrm{sd}(\mathcal{F}), \bar{m}]$, as desired. □

The issue with different rates of convergence among the blocks of $Z(\alpha)$ that vanish, is one of a practical nature. Suppose a path-following algorithm is applied and accurately follows the primal-dual central path. Then, once the block of $Z(\alpha)$ that converges to 0 at the fastest rate, $Z_{d+1}(\alpha)$, reaches machine precision, the remaining blocks cannot be made smaller. Hence the forward error is a function of the difference between the rate of convergence of the slowest block, $Z_2(\alpha)$, and the fastest block, $Z_{d+1}(\alpha)$.

The integer d of Theorem 6.2.5 actually provides an upper bound on $\mathrm{sd}(\mathcal{F})$ that complements the lower bound of Theorem 5.2.2. However, this upper bound is generally intractable due to its reliance on an unknown orthogonal transformation. If the statement of the theorem can be translated to a statement about convergence rates of blocks of eigenvalues, then we would have a tractable upper bound on $\mathrm{sd}(\mathcal{F})$. However this may not be true as illustrated by the parametric sequence,

$$S(\alpha) := \begin{bmatrix} 3 & \alpha^{1/2} & 0 \\ \alpha^{1/2} & \frac{\alpha}{3-\alpha^2} & 0 \\ 0 & 0 & \alpha^3 \end{bmatrix}, \ \alpha > 0. \quad (6.2.15)$$

Here $S(\alpha)$ is positive definite for every $\alpha > 0$ and has different rates of convergence among the diagonal. However, the two eigenvalues that vanish, do so at the same rate, $\Theta(\alpha^3)$. One way to guarantee that the diagonal blocks correspond to blocks of eigenvalues is the following.

Corollary 6.2.6. *Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be a spectrahedron satisfying Assumption 6.1.1 and let $(X(\alpha), y(\alpha), Z(\alpha))$ be the central path of (6.1.6). Let $\bar{\alpha} > 0$ be fixed and let d be as in Theorem 6.2.5. For every $\alpha \in (0, \bar{\alpha})$ we assume that $Z(\alpha)$ has the block structure of Theorem 6.2.5. If the principal submatrices of $Z(\alpha)$ of the form,*

$$S_i(\alpha) = \begin{bmatrix} Z_i(\alpha) & \cdots & * \\ \vdots & \ddots & \vdots \\ * & \cdots & Z_1(\alpha) \end{bmatrix}, \ i \in \{2, \ldots, d\},$$

satisfy $\lambda_{\min}(S_i(\alpha)) = \Theta\left(\lambda_{\min}(Z_i(\alpha))\right)$, then there are exactly d different rates of convergence among the eigenvalues of $X(\alpha)$ that vanish.

Proof. Applying interlacing eigenvalues, Fact 2.1.2, to the submatrices,

$$\begin{bmatrix} Z_{d+1}(\alpha) & \cdots & * \\ \vdots & \ddots & \vdots \\ * & \cdots & Z_i(\alpha) \end{bmatrix} \text{ and } \begin{bmatrix} Z_i(\alpha) & \cdots & * \\ \vdots & \ddots & \vdots \\ * & \cdots & Z_1(\alpha) \end{bmatrix},$$

gives us the upper bound of $\lambda_{\max}(Z_i(\alpha))$ and the lower bound $\Theta(\lambda_{\min}(Z_i(\alpha)))$ on a block of $\lambda(Z(\alpha))$ having the same size as the number of rows $Z_i(\alpha)$. Since,

$$\lambda_{\min}(Z_i(\alpha)) = \Theta\left(\lambda_{\max}(Z_i(\alpha))\right),$$

by assumption *(iii)* of Theorem 6.2.5, the upper and lower bounds on the block of $\lambda(Z(\alpha))$ are of the same order. Thus we conclude that for each $i \in \{2, \ldots, d+1\}$ there is a block of eigenvalues that converges to 0 at the same rate as $Z_i(\alpha)$ does. The desired result follows from the relation $X(\alpha) = \alpha Z(\alpha)^{-1}$. □

The challenge with this result is that the hypothesis is unverifiable just as the block structure of $Z(\alpha)$ is unobservable without knowledge of the appropriate orthogonal transformation. However, our numerical observations indicate that the conclusion of the corollary holds for those test cases for which we have prior knowledge of singularity degree.

Chapter 7

Singularity Degree of Some Toeplitz Matrix Completions

In this chapter we show how structural properties of spectrahedra may be exploited to theoretically analyze some of the notions developed in the previous chapters. One class of such spectrahedra arises from *matrix completion problems*. In this type of problem we know the value of some entries of a matrix and are tasked with filling in the remaining entries so as to satisfy a known desired property. Let us introduce the material of this chapter by considering the following problem.

Problem 7.0.1. *Given $n \geq 4$ and $\theta, \phi \in [0, \pi]$ find $X \in \mathbb{S}^n_+$ such that,*

$$X_{ij} = \begin{cases} 1 & \text{if } i = j, \\ \cos(\theta) & \text{if } |j - i| = 1, \\ \cos(\phi) & \text{if } |j - i| = n - 1. \end{cases}$$

A matrix X is a solution to Problem 7.0.1 if it is positive semidefinite and has the form,

$$X = \begin{bmatrix} 1 & \cos(\theta) & & & \cos(\phi) \\ \cos(\theta) & 1 & \cos(\theta) & ? & \\ & \cos(\theta) & 1 & \ddots & \\ & ? & \ddots & \ddots & \cos(\theta) \\ \cos(\phi) & & & \cos(\theta) & 1 \end{bmatrix},$$

where '?' denotes 'portions' of the matrix that have unspecified values.

To uncover the underlying structure of the solution set of Problem 7.0.1, we begin by showing that it is a spectrahedron. Recall that $E(i, j) \in \mathbb{S}^n$ is the matrix with 1 in the (i, j) and (j, i) positions and zeros everywhere else. Then the solution set to Problem 7.0.1 is

the set of all matrices X that are positive semidefinite and satisfy,

$$\langle E(i,j), X\rangle = \begin{cases} 1 & \text{if } i = j, \\ 2\cos(\theta) & \text{if } j - i = 1, \\ 2\cos(\phi) & \text{if } j - i = n - 1. \end{cases} \tag{7.0.1}$$

Note that $E(i,j) = E(j,i)$ and by symmetry, the value of X_{ij} is implied by the value of X_{ji}. To avoid redundancy, therefore, in (7.0.1) we only consider indices (i,j) where $j \geq i$. Taking into account (2.2.1), equation 7.0.1 implies the existence of a linear map $\mathcal{A}$ comprised of the matrices $E(i,j)$ where $j-i \in \{0,1,3\}$ and a vector b with elements consisting of 1, $2\cos(\theta)$, and $2\sin(\theta)$ such that the solutions to Problem 7.0.1 are exactly the elements of $\mathcal{F}(\mathcal{A}, b)$.

Now if we assume that the solution set to Problem 7.0.1 satisfies Assumption 6.1.1, then we may construct the central path of (6.1.6). Taking $Z(\alpha)$ from the central path we have,

$$Z(\alpha) \in \text{range}(\mathcal{A}^*) = \text{span}\left\{E(i,j) : j - i \in \{0,1,3\}\right\} = \begin{bmatrix} * & * & 0 & \cdots & 0 & * \\ * & * & * & \ddots & & 0 \\ 0 & * & * & \ddots & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & * & 0 \\ 0 & & 0 & * & * & * \\ * & 0 & \cdots & 0 & * & * \end{bmatrix}.$$

It is not hard to see that $Z(\alpha)$ is sparse and that this sparsity is more pronounced for larger n. In fact, $Z(\alpha)_{ij}$ is 0 for every pair (i,j) for which the value of X_{ij} in Problem 7.0.1 is unspecified. Without much effort we have revealed some structural properties of $Z(\alpha)$ that allow for simpler analysis. It requires more effort, see Section 7.1, to show that $X(\alpha)$ also admits a special structure. Specifically, if we let $B = I$ in the construction of the central path, then $X(\alpha)$ is *Toeplitz* (see Definition 7.1.5).

The Toeplitz structure of $X(\alpha)$ allows us to use the Gohberg-Semencul formula for the inverse of a non-singular Toeplitz matrix, to obtain an expression for $Z(\alpha)$ in terms of the entries of $X(\alpha)$. In Section 7.2, we use this formula, the relation $Z(\alpha) = \alpha X(\alpha)^{-1}$, and the sparsity observation to theoretically determine the singularity degree of the feasible set to Problem 7.0.1.

Note that the Toplitz structure is partly enforced in the statement of Problem 7.0.1. In Section 7.1 we study *positive definite Toeplitz completion problems* in general to determine when a partial Toeplitz structure, such as the one in Problem 7.0.1, extends to $X(\alpha)$.

Many of the results of this chapter and the corresponding proofs are based on the article [75], coauthored by the author of this thesis.

7.1 Maximum Determinant Positive Definite Toeplitz Completions

7.1.1 Partial Matrices

We begin by formalizing some of the concepts from the introduction to this chapter.

Definition 7.1.1. *A partial matrix $\mathcal{S} = \mathcal{S}(P, D)$ of order n, is defined by a pattern P and data D where,*

$$\emptyset \neq P \subseteq \{1, \ldots, n\} \times \{1, \ldots, n\},$$

with $(i, j) \in P$ if, and only if, $(j, i) \in P$ and $D = \{a_{ij} \in \mathbb{R} : (i, j) \in P\}$ where $a_{ij} = a_{ji}$. A matrix $S \in \mathbb{S}^n$ is said to be a completion of $\mathcal{S}$ if $S_{ij} = a_{ij}$ for all $(i, j) \in P$.

In general partial matrices need not be symmetric. However, our interest in positive (semi) definite completions immediately rules out non-symmetric partial matrices, thereby justifying Definition 7.1.1.

Definition 7.1.2. *A partial matrix $\mathcal{S}$ is said to be positive (semi) definite completable if there exists a completion of $\mathcal{S}$ that is positive (semi) definite.*

It may be difficult to determine whether a partial matrix is positive (semi) definite completable, especially if we wish to do so analytically. The following condition is easier to verify and may imply that a partial matrix is positive (semi) definite completable in some cases.

Definition 7.1.3. *A partial matrix $\mathcal{S} = \mathcal{S}(P, D)$ is said to be partially positive (semi) definite if every principal submatrix of $\mathcal{S}$ consisting entirely of entries in P is positive (semi) definite.*

Let us now state some elementary observations about partial matrices and the set of positive semidefinite completions of it.

Lemma 7.1.4. *Let $\mathcal{S} = \mathcal{S}(P, D)$ be a partial matrix. The following hold.*

(*i*) *The set of positive semidefinite completions of $\mathcal{S}$ is a spectrahedron.*

(*ii*) *Suppose the set of positive semidefinite completions of $\mathcal{S}$ is non-empty and bounded. If $\mathcal{S}$ is positive definite completable, then there exists a unique matrix that maximizes the determinant over all positive semidefinite completions of $\mathcal{S}$. Moreover $S^\star \in \mathbb{S}^n$ is this determinant maximizer if, and only if, $S^\star$ is a positive definite completion of $\mathcal{S}$ and $(S^\star)^{-1}_{ij} = 0$ whenever $(i, j) \notin P$.*

Proof. The proof of (*i*) follows from arguments similar to those around (7.0.1). Item (*ii*) is proved in [30] and is also implied by Theorem 6.1.2. □

7.1.2 Toeplitz and Bezoutian Matrices

Our main result in this section regards the maximum determinant positive definite completions of a partial Toeplitz matrix. Recall the definition of a Toeplitz matrix.

Definition 7.1.5. *A matrix $T \in \mathbb{S}^n$ is said to be Toeplitz if there exist $t_0, t_1, \dots, t_{n-1} \in \mathbb{R}$ such that $T_{ij} = t_{|i-j|}$ for all $i, j \in \{1, \dots, n\}$.*

Toeplitz matrices are characterized by invariance over 'left-to-right' diagonals. Let us be more specific by introducing the following notation.

Definition 7.1.6. *Let $S \in \mathbb{S}^n$ and let $k \in [-(n-1), n-1]$ be an integer. Then the kth diagonal of S denotes the entries S_{ij} such that $j - i = k$.*

In light of this terminology, a matrix $T \in \mathbb{S}^n$ is Toeplitz if for each $k \in \{0, \dots, n-1\}$ it holds that every entry of the kth diagonal of T has the same value. Toeplitz matrices admit clean inversion formulas that may be expressed in terms of *Bezoutian matrices*. For $a = \begin{pmatrix} a_0 & a_1 & \cdots & a_n \end{pmatrix}^T \in \mathbb{R}^{n+1}$ we define two lower triangular matrices of order n:

$$A(a) = \begin{bmatrix} a_0 & & & & \\ a_1 & a_0 & & & \\ a_2 & a_1 & \ddots & & \\ \vdots & \ddots & \ddots & a_0 & \\ a_{n-1} & \cdots & a_2 & a_1 & a_0 \end{bmatrix} \text{ and } B(a) = \begin{bmatrix} a_n & & & & \\ a_{n-1} & a_n & & & \\ a_{n-2} & a_{n-1} & \ddots & & \\ \vdots & \ddots & \ddots & a_n & \\ a_1 & \cdots & a_{n-2} & a_{n-1} & a_n \end{bmatrix}. \quad (7.1.1)$$

Definition 7.1.7. *The Bezoutian matrix of $a = \begin{pmatrix} a_0 & a_1 & \cdots & a_n \end{pmatrix}^T \in \mathbb{R}^{n+1}$ is defined as,*

$$\mathrm{Bez}(a) := A(a)A(a)^T - B(a)B(a)^T.$$

This definition is in fact a very special case of *Toeplitz Bezoutians*, which are defined in more general settings than this thesis. See, for instance, the survey [34] for a thorough discussion of the subject. Bezoutians are related to Toeplitz marices through inversion.

Fact 7.1.8 ([43],[21])**.** *The inverse of a non-singular Bezoutian matrix is a Toeplitz matrix.*

The converse is also true and we state it with more detail.

Fact 7.1.9 (Gohberg-Semencul, [26], [39])**.** *Let $T \in \mathbb{S}^n$ be Toeplitz and non-singular. Let us denote the first column of T^{-1} by $t = \begin{pmatrix} t_0 & \dots & t_{n-1} \end{pmatrix}^T$. Then,*

$$T^{-1} = \frac{1}{t_0}\mathrm{Bez}\left(\begin{pmatrix} t \\ 0 \end{pmatrix}\right).$$

Bezoutian matrices are classically defined in the language of polynomials. While we have opted to ignore this context in our definition, we do find utility for the following connection between Bezoutians and algebraic geometry.

Fact 7.1.10 (Schur-Cohn Criterion, Theorem XVa [40]). *Let $f(z) = a_0 + a_1 z + \cdots + a_n z^n$ be a polynomial with real coefficients and let $a := (a_0, \ldots, a_n)^T$. Then every root of $f(z)$ satisfies $|z| > 1$ if, and only if,* $\mathrm{Bez}(a) \succ 0$.

The Schur-Cohn criterion is usually stated for the case where the roots are contained within the interior of the unit disk, but a simple reversal of the coefficients, as described in Chapter X of [53], gives us Fact 7.1.10. Although we are not interested in the geometry of the roots of polynomials in this thesis, the positive definite requirement on the Bezoutian of Fact 7.1.10 proves useful in Section 7.1.5.

7.1.3 Partial Toeplitz Matrices

A partial Toeplitz matrix is a partial matrix where the pattern consists of one or more entire diagonals and the data is constant over each specified diagonal. Note that we need not specify the entire pattern of the partial matrix. It suffices to indicate which diagonals are specified. Similarly, we specify the data for each diagonal as opposed to each specified entry. For this reason we adopt notation that is more compact.

Definition 7.1.11. *A partial Toeplitz matrix $\mathcal{T} = \mathcal{T}(P, D)$ of order n, is defined by its non-empty pattern $P \subseteq \{0, \ldots, n-1\}$ and its data $D := \{t_k \in \mathbb{R} : k \in P\}$. A completion of $\mathcal{T}$ is a matrix $T \in \mathbb{S}^n$ such that $T_{ij} = t_{|i-j|}$ whenever $|i-j| \in P$.*

In Definition 7.1.11 we have abandoned the notation of Section 7.1.1 pertaining to partial matrices. However, a partial Toeplitz matrix is still a partial matrix. Therefore, the notions of positive (semi) definite completability and partial positive (semi) definiteness of Definition 7.1.2 and Definition 7.1.3, respectively, naturally apply to partial Toeplitz matrices. We can also state a result analogous to that of Lemma 7.1.4, but first, a simple observation.

Lemma 7.1.12. *Let $\mathcal{T} = \mathcal{T}(P, D)$ be a partial Toeplitz matrix. Then $0 \notin P$ implies that the set of positive definite completions of $\mathcal{T}$ is non-empty and unbounded.*

Proof. Since $0 \notin P$, there are no restrictions on the diagonal elements of any completion. Thus, if $T \in \mathbb{S}^n$ is a completion of $\mathcal{T}$ then $T + \gamma I$ is also a completion of $\mathcal{T}$ for every $\gamma \in \mathbb{R}$. We may choose γ arbitrarily large, yielding the desired result. □

Lemma 7.1.13. *Let $\mathcal{T} = \mathcal{T}(P, D)$ be a positive definite completable partial Toeplitz matrix with $0 \in P$. Then there exists a unique matrix that maximizes the determinant over all positive semidefinite completions of $\mathcal{T}$. Moreover $T^\star \in \mathbb{S}^n$ is this determinant maximizer if, and only if, $T^\star$ is a positive definite completion of $\mathcal{T}$ and $(T^\star)^{-1}_{ij} = 0$ whenever $|i-j| \notin P$.*

Proof. By the hypothesis and Lemma 7.1.12 the set of positive definite completions of $\mathcal{T}$ is non-empty and bounded. The result now follows from Lemma 7.1.4. □

Throughout the remainder of Section 7.1, we use $T^\star$ to denote the determinant maximizer of the positive definite completions of a partial Toeplitz matrix.

7.1.4 Toeplitz Determinant Maximizers

The main result of Section 7.1 is a characterization of the patterns of partial Toeplitz matrices that ensure that $T^\star$ is Toeplitz.

Theorem 7.1.14. *Let $P \subseteq \{0, \dots, n-1\}$ such that $0 \in P$. Then the following are equivalent.*

(*i*) *For every positive definite completable partial Toeplitz matrix $\mathcal{T}(P, D)$, it holds that $T^\star$ is Toeplitz.*

(*ii*) *There exist $r, k \in \mathbb{N}$ such that P has one of the following forms:*

- $P_1 := \{0, k, 2k, \dots, rk\}$,
- $P_2 := \{0, k, 2k, \dots, (r-2)k, rk\}$, *where* $n = (r+1)k$,
- $P_3 := \{0, k, n-k\}$.

One of the directions of Theorem 7.1.14 states that for any pattern P, not having one of the prescribed forms, there exists data D such that $\mathcal{T}(P, D)$ is positive definite completable but the determinant is not Toeplitz. For instance take the pattern $\{0, 1, 3, 4\}$ for $n = 5$ with data $\{6, 1, 1, 1\}$. It can be verified, up to four decimal places, that,

$$T^\star = \begin{bmatrix} 6 & 1 & 0.3113 & 1 & 1 \\ 1 & 6 & 1 & 0.4247 & 1 \\ 0.3113 & 1 & 6 & 1 & 0.3113 \\ 1 & 0.4247 & 1 & 6 & 1 \\ 1 & 1 & 0.3113 & 1 & 6 \end{bmatrix}.$$

The pattern of this partial Toeplitz matrix is not among the patterns of Theorem 7.1.14 and the determinant maximizer is not Toeplitz.

Positive definite Toeplitz matrices play an important role throughout the mathematical sciences. Correlation matrices of data arising from time series, [56], and solutions to the trigonometric moment problem, [38], are two such examples. Among the early contributions to this area is the following sufficient condition and characterization, for a special case of pattern P_1.

Fact 7.1.15 ([20]). *If a partially positive definite partial Toeplitz matrix has pattern P_1 with $k = 1$, then it is positive definite completable and $T^\star$ is Toeplitz.*

The following result characterizes patterns such that partial positive definiteness implies positive definite completability, independent of data.

Fact 7.1.16 (Theorem 1.1,[38]). *All partial Toeplitz matrices with pattern P, satisfying $0 \in P$, that are partially positive definite, admit a positive definite Toeplitz completion if, and only if, P is of the form P_1.*

In Theorem 7.1.14 we make the stronger assumption, that our partial Toeplitz matrices are positive definite completable, giving us patterns of the form P_2 and P_3. These two patterns are not entirely new. In [57], sufficient and necessary conditions are provided for a partially Toeplitz matrix with pattern P_2 and $k = 1$ to have a positive semidefinite completion. A special case of pattern P_3, with $k = 1$, is considered in [4], where the authors characterize the data for which the pattern is positive definite completable. In [33] the result is extended to arbitrary k and sufficient conditions for Toeplitz completions are provided. Moreover, the authors conjecture that whenever a partially positive definite Toeplitz matrix with pattern P_3 is positive definite completable then it admits a Toeplitz completion. This conjecture is confirmed in our main result, Theorem 7.1.14, and more specifically in Proposition 7.1.22.

7.1.5 Proof of Theorem 7.1.14

We have broken the proof up into several smaller results that ought to be be easer to digest. To avoid wordy statements we make the following assumption throughout Section 7.1.5.

Assumption 7.1.17. *We assume that $\mathcal{T} = \mathcal{T}(P, D)$ is a positive definite completable partial Toeplitz matrix with $0 \in P$ and that $T^\star$ is the maximum determinant positive definite completion of $\mathcal{T}$.*

Let K be the symmetric $n \times n$ anti-diagonal matrix defined as,

$$K_{ij} := \begin{cases} 1 & \text{if } i + j = n + 1, \\ 0 & \text{otherwise.} \end{cases} \tag{7.1.2}$$

Note that K is the permutation matrix that reverses the order of the sequence $\{1, 2, \ldots, n\}$. For $n = 4$,

$$K = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

Lemma 7.1.18. *For K defined as in (7.1.2) the following hold.*

(i) $T^\star = KT^\star K$.

(ii) If P is of the form P_2 with $k = 1$, then $T^\star$ is Toeplitz.

Proof. For *(i)* it is a simple exercise to verify that the permutation reverses the order of the rows and columns. Thus we have,

$$[KT^\star K]_{ij} = {T^\star}_{n+1-i,n+1-j},\ \forall i,j \in \{1,\dots,n\}.$$

Moreover,

$$|n+1-i-(n+1-j)| = |i-j|.$$

Therefore, if $\mathcal{T}$ has data $\{t_k \in \mathbb{R} : k \in P\}$, it follows that,

$$[KT^\star K]_{ij} = {T^\star}_{n+1-i,n+1-j} = {T^\star}_{ij} = t_{|i-j|},\ \forall |i-j| \in P.$$

Hence $KT^\star K$ is a completion of $\mathcal{T}$. Moreover, $K \cdot K$ is an automorphism of the cone of positive definite matrices. Hence $KT^\star K$ is a positive definite completion of $\mathcal{T}$, and since K is a permutation matrix, we conclude that $\det(KT^\star K) = \det(T^\star)$. By Lemma 7.1.13, $T^\star$ is the unique maximizer of the determinant. Therefore $T^\star = KT^\star K$, as desired.

For *(ii)* we let P be as in the hypothesis and note that the only unspecified entries are $(1, n-1)$ and $(2, n)$, and their symmetric counterparts. Therefore it suffices to show that ${T^\star}_{1,n-1} = {T^\star}_{2,n}$. By applying *(i)* we get

$${T^\star}_{1,n-1} = [KT^\star K]_{1,n-1} = {T^\star}_{n+1-1,n+1-(n-1)} = {T^\star}_{n,2} = {T^\star}_{2,n},$$

as desired. □

Now we show that a general pattern P_2 may always be reduced to the special case considered in Lemma 7.1.18. Moreover, patterns of the form P_2 differ from those of the form P_1 only in the specification of diagonal $(r-1)k$. It turns out that this similarity allows us to analyze the patterns in, more or less, the same way. We now state a useful lemma for proving that Theorem 7.1.14 *(ii)* implies Theorem 7.1.14 *(i)*, when P is of the form P_1 or P_2.

Lemma 7.1.19. *Let $\mathcal{S}$ be a partial symmetric matrix such that the set of positive definite completions of $\mathcal{S}$ is non-empty and bounded. Let Q be a permutation matrix such that,*

$$Q^T \mathcal{S} Q = \begin{bmatrix} \mathcal{S}_1 & & & \\ & \mathcal{S}_2 & & \\ & & \ddots & \\ & & & \mathcal{S}_\ell \end{bmatrix},$$

for some $\ell \in \mathbb{N}$. Here $\mathcal{S}_i$ is a partial symmetric matrix for each $i \in \{1, \dots, \ell\}$, and the elements outside of the blocks are all unspecified. Then the maximum determinant completion of $\mathcal{S}_i$, denoted $S_i^\star$, exists and is unique for each i. Moreover, the unique maximum determinant completion of $\mathcal{S}$ exists and is given by,

$$S^\star = Q \begin{bmatrix} S_1^\star & 0 & \cdots & 0 \\ 0 & S_2^\star & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & S_\ell^\star \end{bmatrix} Q^T.$$

Proof. The assumption that the set of positive definite completions of $\mathcal{S}$ is non-empty and bounded implies, by Lemma 7.1.4, that there exists a unique maximum determinant positive definite completion of $\mathcal{S}$, say $S^\star$. Since $Q^T \cdot Q$ is an automorphism of the positive definite matrices, with inverse $Q \cdot Q^T$, it follows that $Q^T \mathcal{S} Q$ is a partial symmetric matrix that is positive definite completable and admits a unique maximum determinant completion, say $\widehat{S}$. Under the map $Q \cdot Q^T$, every completion of $Q^T \mathcal{S} Q$ corresponds to a unique completion of $\mathcal{S}$, with the same determinant, since the determinant is invariant under the transformation $Q \cdot Q^T$. Therefore, it holds that $S^\star = Q \widehat{S} Q^T$. All that remains is to show that $\widehat{S}$ has the desired block diagonal form. It is a trivial observation that each $\mathcal{S}_i$ is positive definite completable and admits a maximum determinant positive definite completion. Thus $S_i^\star$ is well defined for each $i \in \{1, \dots, \ell\}$. Then by the determinant Fischer inequality, e.g., Theorem 7.8.3 of [37], it holds that,

$$\widehat{S} = \begin{bmatrix} S_1^\star & 0 & \cdots & 0 \\ 0 & S_2^\star & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & S_\ell^\star \end{bmatrix},$$

as desired. □

In [38] it is shown that a partial Toeplitz matrix of the form P_1 with $rk = n - 1$ can be permuted into a block diagonal matrix as in Lemma 7.1.19. We use this observation and extend it to all patterns of the form P_1, as well as patterns of the form P_2, in the following.

Proposition 7.1.20. *If P is of the form P_1 or P_2, then $T^\star$ is Toeplitz.*

Proof. First, suppose that P is of the form P_1 with data $\{t_0, t_k, t_{2k}, \dots, t_{rk}\}$ for positive integers r and k. Let $p \geq r$ be the largest integer so that $pk \leq n - 1$. As in [38], there

exists a permutation matrix Q of order n such that

$$Q^T\mathcal{T}Q = \begin{bmatrix} \mathcal{T}_0 & & & & & \\ & \ddots & & & & \\ & & \mathcal{T}_0 & & & \\ & & & \mathcal{T}_1 & & \\ & & & & \ddots & \\ & & & & & \mathcal{T}_1 \end{bmatrix},$$

where $\mathcal{T}_0$ is a $(p+1)\times(p+1)$ partial Toeplitz matrix occuring $n-pk$ times and $\mathcal{T}_1$ is a $p\times p$ partial Toeplitz matrix. Moreover, $\mathcal{T}_0$ and $\mathcal{T}_1$ are both partially positive definite. Let us first consider the case $p=r$. Then $\mathcal{T}_0$ and $\mathcal{T}_1$ are actually fully specified, and the maximum determinant completion of $Q^T\mathcal{T}Q$, as in Lemma 7.1.19, is obtained by fixing the elements outside of the blocks to 0. After permuting back to the original form, $T^\star$ has zeros in every unspecified entry. Hence it is Toeplitz. Now suppose $p>r$. Then $\mathcal{T}_0$ is a partial Toeplitz matrix with pattern $\{1,2,\ldots,r\}$ and data $\{t_0,t_k,t_{2k},\ldots,t_{rk}\}$ and $\mathcal{T}_1$ is a partial Toeplitz matrix having the same pattern and data as $\mathcal{T}_0$, but one dimension smaller. That is, $\mathcal{T}_1$ is a partial principal submatrix of $\mathcal{T}_0$. By Fact 7.1.15 both $\mathcal{T}_0$ and $\mathcal{T}_1$ are positive definite completable and their maximum determinant completions, $T_0^\star$ and $T_1^\star$, are Toeplitz. Let $\{a_{(r+1)k},a_{(r+2)k},\ldots,a_{pk}\}$ be the data of $T_0^\star$ corresponding to the unspecified diagonals of $\mathcal{T}_0$ and let $\{b_{(r+1)k},b_{(r+2)k},\ldots,b_{(p-1)k}\}$ be the data of $T_1^\star$ corresponding to the unspecified diagonals of $\mathcal{T}_1$. By the permanence principle of [23], $T_1^\star$ is a principal submatrix of $T_0^\star$ and therefore $b_i=a_i$, for all $i\in\{(r+1)k,(r+2)k,\ldots,(p-1)k\}$. By Lemma 7.1.19, the maximum determinant completion of $Q^T\mathcal{T}Q$ is obtained by completing $\mathcal{T}_0$ and $\mathcal{T}_1$ to $T_0^\star$ and $T_1^\star$ respectively, and setting the entries outside of the blocks to zero. After permuting back to the original form we get that $T^\star$ is Toeplitz with data $a_{(r+1)k},a_{(r+2)k},\ldots,a_{pk}$ in the diagonals $(r+1)k,(r+2)k,\ldots,pk$ and zeros in all other unspecified diagonals, as desired.

Now suppose that $\mathcal{T}$ is of the form P_2. By applying the same permutation as above, and by using the fact that $n=(r+1)k$ and each block $\mathcal{T}_0$ is of size $r+1$, we see that the submatrix consisting only of blocks $\mathcal{T}_0$ is of size

$$(n-rk)(r+1)=((r+1)k-rk)(r+1)=k(r+1)=n.$$

Hence,

$$Q^T\mathcal{T}Q = \begin{bmatrix} \mathcal{T}_0 & & \\ & \ddots & \\ & & \mathcal{T}_0 \end{bmatrix},$$

where $\mathcal{T}_0$ is a partial matrix with pattern $\{1,2,\ldots,r-2,r\}$ and data

$$\{t_0,t_k,t_{2k},\ldots,t_{(r-2)k},t_{rk}\}.$$

The unspecified elements of diagonal $(r-1)k$ of $\mathcal{T}$ are contained in the unspecified elements of diagonal $r-1$ of the partial matrices $\mathcal{T}_0$. By Lemma 7.1.18, the maximum

determinant completion of $\mathcal{T}_0$ is Toeplitz with value $t_{(r-1)k}$ in the unspecified diagonal. As in the above, after completing $Q^T\mathcal{T}Q$ to its maximum determinant positive definite completion and permuting back to the original form, we obtain the maximum determinant Toeplitz completion of $\mathcal{T}$ with value $t_{(r-1)k}$ in the diagonal $(r-1)k$ and zeros in every other unspecified diagonal, as desired. □

We have proved one direction of Theorem 7.1.14 for the patterns P_1 and P_2. To complete the proof of this direction, we turn our attention to patterns of the form P_3. Let J denote the $n \times n$ lower triangular matrix with 1s on diagonal -1 and zeros everywhere else. That is,

$$J := \begin{bmatrix} 0 & & & & \\ 1 & 0 & & & \\ 0 & 1 & \ddots & & \\ \vdots & \ddots & \ddots & 0 & \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix}.$$

We may also define J in terms of the canonical basis of $\mathbb{R}^n$. Recall that $e_i \in \mathbb{R}^n$ denotes the ith column of I. Then,

$$J = \sum_{j=1}^{n-1} e_{j+1}e_j^T. \tag{7.1.3}$$

We state several technical results regarding J in the following lemma.

Lemma 7.1.21. *Let J be defined as in (7.1.3) and let $k, l \in \{0, 1, \ldots, n-1\}$. Then the following hold.*

(*i*) $J^k = \sum_{j=1}^{n-k} e_{j+k}e_j^T$.

(*ii*) *If $\ell = n$ and $k = 0$, then $J^k(J^\ell)^T = 0$, otherwise $J^k(J^T)^\ell$ has non-zero elements only in the diagonal $\ell - k$.*

(*iii*) $J^k(J^\ell)^T - J^{n-\ell}(J^{n-k})^T = 0$ *if, and only if,* $\ell = n - k$.

Proof. For (*i*) the result clearly holds when $k \in \{0, 1\}$. Now observe that $(e_k e_\ell^T)(e_i e_j^T) \neq 0$ if, and only if, $\ell = i$ in which case the product is $e_k e_j^T$. Thus we have,

$$J^2 = \left(\sum_{j=1}^{n-1} e_{j+1}e_j^T\right)\left(\sum_{j=1}^{n-1} e_{j+1}e_j^T\right) = \sum_{j=2}^{n-1}(e_{j+1}e_j^T)(e_j e_{j-1}^T) = \sum_{j=1}^{n-2} e_{j+2}e_j^T.$$

Applying an inductive argument yields the desired expression for arbitrary k.

For (*ii*), we use (*i*) to get,

$$\begin{aligned} J^k \left(J^\ell\right)^T &= \left(\sum_{j=1}^{n-k} e_{j+k} e_j^T\right) \left(\sum_{j=1}^{n-\ell} e_j e_{j+\ell}^T\right), \\ &= \sum_{j=1}^{n-\max\ \{k,\ell\}} (e_{j+k} e_j^T)(e_j e_{j+\ell}^T), \\ &= \sum_{j=1}^{n-\max\ \{k,\ell\}} e_{j+k} e_{j+\ell}^T. \end{aligned}$$

The non-zero elements of this matrix are contained in the diagonal $j+\ell-(j+k) = \ell - k$.

Finally, for (*iii*) we have,

$$J^k \left(J^\ell\right)^T - J^{n-\ell} \left(J^{n-k}\right)^T = \sum_{j=1}^{n-\max\{k,\ell\}} e_{j+k} e_{j+\ell}^T - \sum_{j=1}^{n-\max\{k,\ell\}} e_{j+n-\ell} e_{j+n-k}^T.$$

This matrix is the zero matrix if, and only if, $\ell = n-k$. □

The matrices $A(a)$ and $B(a)$ of (7.1.1) may be expressed in terms of J as,

$$A(a) = \sum_{j=0}^{n-1} a_j J^j \text{ and } B(a) = \sum_{j=0}^{n-1} a_{n-j} J^j. \tag{7.1.4}$$

Now we recall the results of Section 7.1.2 on Toeplitz and Bezoutian matrices to obtain the following.

Proposition 7.1.22. *If P is of the form P_3, then $T^\star$ is Toeplitz.*

Proof. Let k be a positive integer such that $P = \{0, k, n-k\}$. We may assume, without loss of generality, that $k \leq n-k$. Let $\mathcal{V} \subset \mathbb{R}_{++} \times \mathbb{R}^2$ consist of all triples (t_0, t_k, t_{n-k}) so that the partial Toeplitz matrix with pattern P and data $\{t_0, t_k, t_{n-k}\}$ is positive definite completable. Note that the positive definite completions of $\mathcal{V}$ consist of the positive definite matrices that are constant along the 0th, kth, and $(n-k)$th diagonals. Since this is a convex set and $\mathcal{V}$ is a projection of this set onto the $(1,1)$, $(1,k+1)$, and $(1,n-k+1)$ entries, it follows that $\mathcal{V}$ is also a convex set. Moreover, $\mathcal{V}$ is open and connected. We let $\mathcal{U} \subseteq \mathcal{V}$ consist of those triples (t_0, t_k, t_{n-k}) for which the corresponding maximum determinant completion is Toeplitz. We are done if we show that $\mathcal{U} = \mathcal{V}$. Note that $\mathcal{U} \neq \emptyset$ since it contains the triples $(t_k, 0, 0)$ for all $t_k > 0$. Therefore, to prove $\mathcal{U} = \mathcal{V}$, it suffices to show that $\mathcal{U}$ is both open and closed in $\mathcal{V}$.

First let us show that $\mathcal{U}$ is closed in $\mathcal{V}$. Observe that the map $F : \mathcal{V} \to \mathbb{S}^n_{++}$ taking (t_0, t_k, t_{n-k}) to its corresponding positive definite maximum determinant completion, is

continuous and injective. This follows from the classical Maximum Theorem of Berge, [5]. Therefore, by the continuity of F, it suffices to show that $F(\mathcal{U})$ is closed in $F(\mathcal{V})$. To this end let $\{X^k\}_{k\in\mathbb{N}}$ be a sequence in $F(\mathcal{U})$ with limit point $\bar{X} \in F(\mathcal{V})$. Since $\bar{X} \in F(\mathcal{V})$, it is a maximum determinant completion and since X^k is Toeplitz for each k and the Toeplitz matrices are closed, it follows that $\bar{X}$ is Toeplitz. Hence $\bar{X} \in F(\mathcal{U})$, as desired.

To show that $\mathcal{U}$ is also open in $\mathcal{V}$, we introduce the set,

$$C := \{(p,q,r) \in \mathbb{R}_{++} \times \mathbb{R}^2 : p + qz^k + rz^{n-k} \text{ has all roots satisfy } |z| > 1\}.$$

Since the region defined by the inequality $|z| > 1$ is an open subset of the complex plane, C is an open set. Now for each $(p,q,r) \in C$ we define,

$$a(p,q,r) = \begin{pmatrix} a_0 & a_1 & \cdots & a_n \end{pmatrix}^T \in \mathbb{R}^{n+1}, \tag{7.1.5}$$

such that $a_0 = p$, $a_k = q$, and $a_{n-k} = r$ and the other entries are zeros. Then consider the map $G : C \to \mathbb{R}^3$ defined as,

$$G(p,q,r) = \left([\text{Bez}(a(p,q,r))^{-1}]_{11}, [\text{Bez}(a(p,q,r))^{-1}]_{k+1,1}, [\text{Bez}(a(p,q,r))^{-1}]_{n-k+1,1}\right).$$

The map G is well-defined since Fact 7.1.10 and the construction of C guarantee that the matrix $\text{Bez}(a(p,q,r))$ is non-singular. Now we show that $G(C) = \mathcal{U}$, implying by the openness of C and by the continuity of G, that $\mathcal{U}$ is open, as desired.

By Fact 7.1.8, it holds that $\text{Bez}(a(p,q,r))^{-1}$ is positive definite and Toeplitz whenever $(p,q,r) \in C$. Thus $\text{Bez}(p,q,r)^{-1}$ is a positive definite completion of the partial Toeplitz matrix having pattern P and data $G(p,q,r)$. Therefore $G(C) \subseteq \mathcal{V}$. Recalling Definition 7.1.7 and (7.1.4) we have,

$$\begin{aligned} \text{Bez}(p,q,r) = &(pJ^0 + qJ^k + rJ^{n-k})(pJ^0 + qJ^k + rJ^{n-k})^T \\ &- (rJ^k + qJ^{n-k})(rJ^k + qJ^{n-k})^T. \end{aligned} \tag{7.1.6}$$

In expanding $\text{Bez}(p,q,r)$, terms with $J^0(J^0)^T$, $J^k(J^k)^T$, and $J^{n-k}(J^{n-k})^T$ have non-zero entries only on the diagonal, by Lemma 7.1.21 *(ii)*. Similarly, terms with $J^0(J^k)^T$ and with $J^0(J^{n-k})^T$ have non-zero entries only on diagonals k and $n-k$, respectively. The remaining terms, with $J^k(J^{n-k})^T$ and $J^{n-k}(J^k)^T$, vanish by Lemma 7.1.21 *(ii)*. Therefore, $\text{Bez}(p,q,r)$ has non-zero values only on the diagonals 0, k, and $n-k$. In other words, $\text{Bez}(p,q,r)$ has zeros in the entries not specified by the partial Toeplitz matrix having pattern P and data $G(p,q,r)$. Thus by Lemma 7.1.13, $\text{Bez}(p,q,r)^{-1}$ is a maximum determinant completion of this partial Toeplitz matrix and it follows that $G(C) \subseteq \mathcal{U}$.

All that remains is to show that $\mathcal{U} \subseteq G(C)$. To this end, let $(t_0, t_k, t_{n-k}) \in \mathcal{U}$ and let $F(t_0, t_k, t_{n-k})$, as above, be the maximum determinant positive definite completion of the partial matrix with pattern P and data $\{t_0, t_k, t_{n-k}\}$. By definition of $\mathcal{U}$, it holds that $F(t_0, t_k, t_{n-k})$ is Toeplitz. Let f_0, f_k, and f_{n-k} be the $(1,1)$, $(k+1,1)$

and $(n-k+1,1)$ elements of $F(t_0,t_kt_{n-k})^{-1}$ respectively. Then by the Gohberg-Semencul formula, Fact 7.1.9, we have,

$$\begin{aligned} F(t_0,t_k,t_{n-k})^{-1} &= \frac{1}{f_0}\left(f_0J^0+f_kJ^k+f_{n-k}J^{n-k}\right)\left(f_0J^0+f_kJ^k+f_{n-k}J^{n-k}\right)^T \\ &\quad -\frac{1}{f_0}\left(f_{n-k}J^k+f_kJ^{n-k}\right)\left(f_{n-k}J^k+f_kJ^{n-k}\right)^T, \\ &= \mathrm{Bez}\left(a\left(\sqrt{f_0},\frac{f_k}{\sqrt{f_0}},\frac{f_{n-k}}{\sqrt{f_0}}\right)\right), \end{aligned}$$

where a is defined as in (7.1.5). Since $F(t_0,t_k,t_{n-k})^{-1}\succ 0$, it follows, by Fact 7.1.10, that $\left(\sqrt{f_0},\frac{f_k}{\sqrt{f_0}},\frac{f_{n-k}}{\sqrt{f_0}}\right)\in C$ and therefore,

$$(t_0,t_k,t_{n-k}) = G\left(\sqrt{f_0},\frac{f_k}{\sqrt{f_0}},\frac{f_{n-k}}{\sqrt{f_0}}\right)\in G(C),$$

as desired. □

In Proposition 7.1.20 and Proposition 7.1.22 we have proved direction *(ii)* $\implies$ *(i)* of Theorem 7.1.14. The converse is proved in the following.

Proposition 7.1.23. *Let P be such that for all positive definite completable $\mathcal{T}(P,D)$, the determinant maximizer $T^\star$ is Toeplitz. Then P is of the form P_1, P_2, or P_3.*

Proof. Let $P=\{0,k_1,\dots,k_s\}$ for some $s\in\{1,2,\dots,n-1\}$ and let $D=\{t_0,t_1,\dots,t_s\}$ be arbitrary data so that $\mathcal{T}(P,D)$ is positive definite completable. We assume here that t_j corresponds to diagonal k_j. Since we have assumed that $T^\star$ is Toeplitz, Lemma 7.1.13 gives us that $(T^\star)^{-1}$ has non-zero entries only in the diagonals specified by P (and their symmetric counterparts). Let $a=\begin{pmatrix}a_0 & \cdots & a_{n-1}\end{pmatrix}^T$ denote the first column of $(T^\star)^{-1}$ and let us define $\hat{a}:=\begin{pmatrix}a^T & 0\end{pmatrix}^T$. Then by Fact 7.1.9 we have,

$$(T^\star)^{-1}=\frac{1}{a_0}\left(A(\hat{a})A(\hat{a})^T-B(\hat{a})B(\hat{a})^T\right).$$

where by (7.1.4),

$$A(\hat{a})=\sum_{j=0}^{s}a_jJ^{k_j},\quad B(\hat{a})=\sum_{j=1}^{s}a_jJ^{n-k_j}.$$

In the above we define $k_0:=0$. Substituting in the expressions for $A(\hat{a})$ and $B(\hat{a})$ and expanding, we see that $(T^\star)^{-1}$ is a linear combination of three types of terms, along with their symmetric counterparts:

$$\begin{cases} J^{k_j}\left(J^{k_j}\right)^T, \\ J^{k_0}\left(J^{k_j}\right)^T, \\ J^{k_j}\left(J^{k_\ell}\right)^T-J^{n-k_j}\left(J^{n-k_\ell}\right)^T,\ j\neq\ell. \end{cases}$$

By Lemma 7.1.21, the terms $J^{k_j}(J^{k_j})^T$ have non-zero entries only on the main diagonal, and the terms $J^{k_0}(J^{k_j})^T$ have non-zero entries only on the diagonals belonging to P. The third type of term, $J^{k_j}(J^{k_\ell})^T - J^{n-k_j}(J^{n-k_\ell})^T$, where $j \neq \ell$ has non-zero entries only on the diagonals $\pm|k_j - k_\ell|$. As we have already observed in the proof of Proposition 7.1.22, the set of data for which $\mathcal{T}$ is positive definite completable is an open set. We may therefore perturb the data of $\mathcal{T}$ so that the entries of a do not all lie on the same proper linear manifold. Then terms of the form $J^{k_j}(J^{k_\ell})^T - J^{n-k_j}(J^{n-k_\ell})^T$ with $j \neq \ell$ do not cancel each other out. Therefore, for each pair j and ℓ where $j \neq \ell$, it holds that either $|k_j - k_\ell| \in P$ or $J^{k_j}(J^{k_\ell})^T - J^{n-k_j}(J^{n-k_\ell})^T = 0$. By Lemma 7.1.21 the second alternative is equivalent to $k_\ell = n - k_j$. Therefore for $j \neq \ell$ one of the following alternatives holds:

$$|k_j - k_\ell| \in P \text{ or } k_\ell = n - k_j. \tag{7.1.7}$$

Using this observation we now proceed to show that P has one of the specified forms.

Let $1 \leq r \leq s$ be the largest integer such that $\{0, k_1, \ldots, k_r\}$ is of the form P_1. That is, $k_2 = 2k_1$, $k_3 = 3k_1$, and so on. If $r = s$, then we are done. Therefore we may assume that $s \geq r+1$. Now we show that in fact $s = r+1$. By (7.1.7) it holds that $k_{r+1} - k_1 \in P$ or $k_{r+1} = n - k_1$. We show that the first case does not hold. Indeed, if $k_{r+1} - k_1 \in P$, then it follows that $k_{r+1} - k_1 \in \{k_1, \ldots, k_r\}$. This implies that,

$$k_{r+1} \in \{2k_1, \ldots, rk_1, (r+1)k_1\} = \{k_2, \ldots, k_r, (r+1)k_1\}.$$

Clearly $k_{r+1} \notin \{k_2, \ldots, k_r\}$, and if $k_{r+1} = (r+1)k_1$, then r is not maximal, a contradiction. Therefore $k_{r+1} = n - k_1$. To show that $s = r+1$, suppose to the contrary that $s \geq r+2$. Again by (7.1.7), it holds that $k_{r+2} - k_1 \in P$ or $k_{r+2} = n - k_1$. The latter does not hold since then $k_{r+2} = k_{r+1}$. Thus we have $k_{r+2} - k_1 \in \{k_1, \ldots, k_r, k_{r+1}\}$, which implies that

$$k_{r+2} \in \{2k_1, \ldots, rk_1, (r+1)k_1, k_{r+1} + k_1\} = \{k_2, \ldots, k_r, k_r + k_1, n\}.$$

Clearly $k_{r+2} \notin \{k_2, \ldots, k_r, n\}$ giving us that $k_{r+2} = k_r + k_1$. Since $k_r < k_{r+1} < k_{r+2}$, it follows that $0 < k_{r+2} - k_{r+1} < k_1$. Therefore $k_{r+2} - k_{r+1} \notin P$. Then by (7.1.7) it holds that $k_{r+2} = n - k_{r+1} = k_1$, a contradiction.

We have shown that $P = \{0, k_1, 2k_1, \ldots, rk_1, k_s\}$ with $k_s = n - k_1$. If $r = 1$, then P is of the form P_3. On the other hand if $r \geq 2$ it holds by (7.1.7) that,

$$\{k_s - k_r, \ldots, k_s - k_2\} \subseteq P \setminus \{0, k_1\}.$$

Equivalently,

$$\{k_s - k_r, \ldots, k_s - k_2\} \subseteq \{k_2, \ldots, k_r\}.$$

Since these two sets of distinct increasing elements have identical cardinality, we conclude that $k_s - k_2 = k_r$. Rearranging, we obtain that $k_s = (r+2)k_1$ and P is of the form P_2, as desired. □

7.1.6 Maximum Rank Positive Semidefinite Toeplitz Completions

Theorem 7.1.14 can be extended to partial Toeplitz matrices that are only positive semidefinite completable using the central path of (6.1.1).

Corollary 7.1.24. *Let $\mathcal{T} = \mathcal{T}(P, D)$ be a partial Toeplitz matrix where P is of the form P_1, P_2, or P_3. If $\mathcal{T}$ is positive semidefinite completable, then $\mathcal{T}$ admits a Toeplitz positive semidefinite completion with maximum rank over all positive semidefinite completions of $\mathcal{T}$.*

Proof. Let $\mathcal{T}(P, D)$ be as in the hypothesis with $P = \{0, k_1, \ldots, k_s\}$ and $D = \{t_0, t_1, \ldots, t_s\}$ for a positive integer $s \leq n-1$. Let $\mathcal{F} = \mathcal{F}(\mathcal{A}, b)$ be the spectrahedron consisting of the positive semidefinite completions of $\mathcal{T}$. As in (7.0.1) we may construct $\mathcal{A}$ so that it consists of matrices E_{ij} such that $j - i \in P$.

Let us address two simple cases. First, if $\mathcal{F}$ contains a positive definite matrix, then $\mathcal{T}$ is positive definite completable. By Theorem 7.1.14 the maximum determinant completion is Toeplitz, satisfying the claim. Second, if $\mathcal{F}$ is the zero matrix then the maximum rank over $\mathcal{F}$ is 0 and is attained by the zero matrix which is Toeplitz, as desired.

Now we may assume that $\mathcal{F} \subset \mathbb{S}^n_+ \setminus \mathbb{S}^n_{++}$ and that $\mathcal{F} \neq \{0\}$. We have assumed that $\mathcal{T}$ is positive semidefinite completable, hence $\mathcal{F} \neq \emptyset$. Moreover, since $0 \in P$ it follows that $\mathcal{F}$ is bounded. Thus $\mathcal{F}$ satisfies Assumption 6.1.1 and the central path of (6.1.1) is well-defined for $\mathcal{F}$. Let us construct this path with $B = I$. Note that b consists of the data D (up to scaling) and $\mathcal{A}(I)$ is a vector that contains 1s in positions corresponding to the diagonal and 0s elsewhere. Thus, for $\alpha > 0$, the vector $b(\alpha) := b + \alpha\mathcal{A}(I)$ differs from b by exactly α in elements corresponding to the diagonal and is the same as b in all other elements. Therefore $\mathcal{F}(\alpha)$ consists of positive definite completions of $\mathcal{T}(P, D(\alpha))$ where,

$$D(\alpha) := \{t_0 + \alpha, t_1, \ldots, t_s\}.$$

Since $\mathcal{F}(\alpha) \cap \mathbb{S}^n_{++} \neq \emptyset$ it follows that $\mathcal{T}(P, D(\alpha))$ is positive definite completable. Thus Theorem 7.1.14 implies that $X(\alpha)$ is Toeplitz for every $\alpha > 0$.

Now let $\bar{X}$ denote the limit point of $X(\alpha)$ as $\alpha \searrow 0$. Then $\bar{X}$ is a positive semidefinite completion of $\mathcal{T}$ with maximum rank over all positive semidefinite completions according to Theorem 6.1.9. Moreover, since the Toeplitz matrices form a linear subspace of $\mathbb{S}^n$ and $X(\alpha)$ is Toeplitz for every $\alpha > 0$ then $\bar{X}$, the limit point of $X(\alpha)$, is also Toeplitz, as desired. □

7.2 Singularity Degree of Toeplitz Cycles

Let us turn our attention back to the Toeplitz completion problem, Problem 7.0.1, presented in the introduction to this chapter. From a graph theoretic perspective, the pattern

of the partial matrix in Problem 7.0.1 corresponds to one large cycle. For this reason, the problem has been referred to as "cycle completion". The data ensuring that Problem 7.0.1 is solvable is characterized in [4].

Fact 7.2.1 (Corollary 6, [4]). *Let $\mathcal{T} = \mathcal{T}(P, D)$ be a partial Toeplitz matrix of order $n \geq 4$ with,*

$$P = \{0, 1, n-1\} \text{ and } D = \{1, \cos(\theta), \cos(\phi)\}, \quad \theta, \phi \in [0, \pi].$$

Then $\mathcal{T}$ is positive semidefinite completable if, and only if,

$$\frac{\phi}{n-1} \leq \theta \leq \frac{(n-2)\pi + \phi}{n-1} \text{ when } n \text{ is even,}$$

and,

$$\frac{\phi}{n-1} \leq \theta \leq \frac{(n-1)\pi - \phi}{n-1} \text{ when } n \text{ is odd.}$$

By choosing θ and ϕ so that the inequalities of Fact 7.2.1 are tight, we obtain a solution set that satisfies Assumption 6.1.1. Then we may construct the central path of (6.1.1) as in the proof of Corollary 7.1.24. Moreover, since the pattern of Fact 7.2.1 is of the form P_3, the matrix $X(\alpha)$ in the central path is Toeplitz for every $\alpha > 0$. Then we may use the fact that $Z(\alpha)$ in the central path is a scalar multiple of $X(\alpha)^{-1}$ and the Gohberg-Semencul formula of Fact 7.1.9, to analyze the singularity degree of the set of positive semidefinite completions of $\mathcal{T}$. To this end we record the following special case of Fact 7.1.9.

Lemma 7.2.2. *Let $T \in \mathbb{S}^n$ be Toeplitz and non-singular. If,*

$$e_1^T T^{-1} = \begin{pmatrix} a & c & 0 \cdots & 0 & d \end{pmatrix}$$

then,

$$T^{-1} = \begin{bmatrix} a & c & 0 & & d \\ c & b & c & \ddots & \\ 0 & c & \ddots & \ddots & 0 \\ & \ddots & \ddots & b & c \\ d & & 0 & c & a \end{bmatrix}, \tag{7.2.1}$$

where $b := \frac{1}{a}(a^2 + c^2 - d^2)$.

Proof. By Fact 7.1.9 it holds that,

$$\begin{aligned} T^{-1} &= \frac{1}{a}\left(aJ^0 + cJ^1 + dJ^{n-1}\right)\left(aJ^0 + cJ^1 + dJ^{n-1}\right)^T \\ &\qquad - \frac{1}{a}\left(dJ^1 + cJ^{n-1}\right)\left(dJ^1 + cJ^{n-1}\right)^T \\ &= aI + \frac{c^2 - d^2}{a} J^1 \left(J^1\right)^T + \frac{d^2 - c^2}{a} J^{n-1}\left(J^{n-1}\right)^T && (7.2.2) \\ &\qquad + c\left(J^0\left(J^1\right)^T + J^1\left(J^0\right)^T\right) && (7.2.3) \\ &\qquad + d\left(J^0\left(J^{n-1}\right)^T + J^{n-1}\left(J^0\right)^T\right). && (7.2.4) \end{aligned}$$

By Lemma 7.1.21 the non-zero elements of (7.2.2) are restricted to the diagonal of T^{-1}. Therefore, (7.2.3) implies that every entry of diagonals 1 and -1 of T^{-1} takes on the value c. Similarly, (7.2.4) implies that $T^{-1}_{n,1} = T^{-1}_{1,n} = d$. We have shown that T^{-1} has the desired form of (7.2.1) everywhere except the diagonal. Now it is not difficult to verify that for $k \in \{1, \ldots, n-1\}$,

$$J^k \left(J^k\right)^T = \begin{bmatrix} 0 & 0 \\ 0 & I_{n-k} \end{bmatrix},$$

where I_{n-k} is the identity matrix of order $n-k$. Then from (7.2.2) we have,

$$T^{-1}_{kk} = \begin{cases} a & \text{if } k = 1, \\ \frac{1}{a}\left(a^2 + c^2 - d^2\right) & \text{if } k \in \{2, \ldots, n-1\}, \\ \frac{1}{a}\left(a^2 + c^2 - d^2 + d^2 - c^2\right) = a & \text{if } k = n, \end{cases}$$

as desired. □

As it turns out the partial Toeplitz matrix of Fact 7.2.1 has a unique positive semidefinite completion.

Lemma 7.2.3. *Let $\mathcal{T} = \mathcal{T}(P, D)$ be a partial Toeplitz matrix of order $n \geq 4$ with,*

$$P = \{0, 1, n-1\} \text{ and } D = \{1, \cos(\theta), \cos((n-1)\theta)\}, \quad \theta \in \left[0, \frac{\pi}{n-1}\right].$$

Then the unique positive semidefinite completion of $\mathcal{T}$ is the Toeplitz matrix T satisfying,

$$T_{ij} = \cos(|i-j|\theta), \ \forall i, j \in \{1, \ldots, n\}, \tag{7.2.5}$$

with,

$$\operatorname{rank}(T) = \begin{cases} 1 & \text{if } \theta = 0, \\ 2 & \text{otherwise.} \end{cases} \tag{7.2.6}$$

Proof. Let X be a positive semidefinite completion of $\mathcal{T}$. The first row of X is,

$$e_1^T X = \begin{pmatrix} \cos\theta_0 & \cos\theta_1 & \cos\theta_2 & \cdots & \cos\theta_{n-1} \end{pmatrix},$$

where $\theta_0 = 0, \theta_1 = \theta$, $\theta_{n-1} = (n-1)\theta$ and $\theta_2, \ldots, \theta_{n-2} \in [0, \pi]$. The principal submatrix of X consisting of rows and columns 1, $n-1$ and n, is,

$$X_{\{1,n-1,n\}} := \begin{bmatrix} 1 & \cos(\theta_{n-2}) & \cos((n-1)\theta) \\ \cos(\theta_{n-2}) & 1 & \cos(\theta) \\ \cos((n-1)\theta) & \cos(\theta) & 1 \end{bmatrix}.$$

Since $X_{\{1,n-1,n\}} \succeq 0$, Proposition 2 of [4] gives us that $(n-1)\theta \leq \theta_{n-2} + \theta$. Thus

$$\theta_{n-2} \geq (n-2)\theta. \tag{7.2.7}$$

Next, let us denote by $\widehat{X}$ the $(n-1)\times(n-1)$ upper left block of X. Note that $\widehat{X}$ is a positive semidefinite completion of the partial Toeplitz matrix $\widehat{\mathcal{T}} = \mathcal{T}(\widehat{P}, \widehat{D})$ of order $n-1$ with,

$$P = \{0, 1, n-2\},\ D = \{1, \cos(\theta), \cos(\theta_{n-2})\}.$$

By Corollary 2 of [4] it holds that,

$$2\max\{\theta_{n-2}, \theta\} \leq (n-2)\theta + \theta_{n-2}. \tag{7.2.8}$$

Since (7.2.7) implies that $\theta_{n-2} = \max\{\theta_{n-2}, \theta\}$, (7.2.8) simplifies to,

$$\theta_{n-2} \leq (n-2)\theta. \tag{7.2.9}$$

Combining (7.2.9) with (7.2.7) gives us that $\theta_{n-2} = (n-2)\theta$.

Using the same arguments, but with the principal submatrix consisting of rows and columns 1,2, and n and the $(n-1)\times(n-1)$ lower right block of X, we get that the $(2,n)$ entry of X also takes the value $\cos((n-2)\theta)$. We have thus shown that along diagonal $(n-2)$, the value of X is $\cos((n-2)\theta)$. By inductively considering other principal submatrices of X we obtain (7.2.5). The claim regarding rank follows from the observation that $T = B^T B$ where,

$$B := \begin{bmatrix} 1 & \cos\theta & \cos(2\theta) & \cdots & \cos((n-1)\theta) \\ 0 & \sin\theta & \sin(2\theta) & \cdots & \sin((n-1)\theta) \end{bmatrix}.$$

□

The main result of this section is a statement about the singularity degree of a family of partial Toeplitz matrices.

Theorem 7.2.4. *Let $\mathcal{T} = \mathcal{T}(P, D)$ be a partial Toeplitz matrix of order $n \geq 4$ with,*

$$P = \{0, 1, n-1\} \text{ and } D = \{1, \cos(\theta), \cos((n-1)\theta)\}, \quad \theta \in \left[0, \frac{\pi}{n-1}\right].$$

Let $\mathcal{F}$ denote the spectrahedron that consists of the positive semidefinite completions of $\mathcal{T}$. Then,

$$\mathrm{sd}(\mathcal{F}) = \begin{cases} 1, & \text{if } \theta \in [0, \frac{\pi}{n-1}), \\ \geq 2, & \text{if } \theta = \frac{\pi}{n-1}. \end{cases}$$

Proof. By Fact 7.2.1, $\mathcal{F}$ is non-empty. Moreover, it is bounded and satisfies Assumption 6.1.1. Adopting the notation of Chapter 6, we let $(X(\alpha), y(\alpha), Z(\alpha))$ denote the central path of (6.1.1) for $\mathcal{F}$, constructed with $B = I$, as in Corollary 7.1.24. In the proof of Corollary 7.1.24, we also showed that $X(\alpha)$ is Toeplitz. Therefore, $Z(\alpha) = \alpha X(\alpha)^{-1}$

has the form of Lemma 7.2.2. If we let $\bar{Z}$ denote the limit of $Z(\alpha)$ as $\alpha \searrow 0$, then there exist a, b, c, and d such that,

$$\bar{Z} = \begin{bmatrix} a & c & 0 & & d \\ c & b & c & \ddots & \\ 0 & c & \ddots & \ddots & 0 \\ & \ddots & \ddots & b & c \\ d & & 0 & c & a \end{bmatrix}. \tag{7.2.10}$$

First we show $a > 0$. For the sake of contradiction, suppose that $a = 0$. Since $\bar{Z} \succeq 0$ we have $c = d = 0$. Moreover, $\bar{X}$ and $\bar{Z}$ are orthogonal hence,

$$0 = \langle \bar{X}, \bar{Z} \rangle = (n-2)b,$$

implying that $b = 0$. Then $\bar{Z} = 0$, contradicting Theorem 6.1.9. One implication of $a > 0$ is the relation,

$$b = \frac{1}{a}(a^2 + c^2 - d^2). \tag{7.2.11}$$

Next, let us make some observations regarding the rank of $\bar{Z}$. If $b > 0$ then columns 2 through $(n-2)$ of $\bar{Z}$ are linearly independent, implying that $\text{rank}(\bar{Z}) \geq n-2$. On the other hand, when $b = 0$ it follows, by $\bar{Z} \succeq 0$, that $c = 0$. Moreover, (7.2.11) implies that d has the same magnitude as a and therefore, $\text{rank}(\bar{Z}) = 1$. Summarizing, we have,

$$\text{rank}(\bar{Z}) = \begin{cases} 1 & \text{if } b = 0, \\ \geq n-2 & \text{if } b > 0. \end{cases} \tag{7.2.12}$$

Since $\bar{X}$ and $\bar{Z}$ are orthogonal, Fact 2.1.1 implies that $\bar{X}\bar{Z} = 0$, giving us the equations,

$$0 = a + c\cos(\theta) + d\cos((n-1)\theta), \tag{7.2.13}$$

$$0 = b + 2c\cos(\theta). \tag{7.2.14}$$

Now we show that,

$$\theta \in \left(0, \frac{\pi}{n-1}\right) \implies \text{rank}(\bar{Z}) \geq n-2. \tag{7.2.15}$$

To see this, suppose that $b = 0$. We have already shown that this assumption implies that $c = 0$ and that $d = \pm a$. If $d = a$, then (7.2.13) implies that $\theta = \pi/(n-1)$. On the other hand if $d = -a$, (7.2.13) gives us that $\theta = 0$. These two observations give us that,

$$b = 0 \implies \theta \in \left(0, \frac{\pi}{n-1}\right), \tag{7.2.16}$$

implying (7.2.15). Moreover, (7.2.6) implies that $\text{rank}(\bar{X}) = 2$ for these values of θ. Therefore complementary slackness (see Section 4.5) and the fact that $\bar{Z}$ is a maximum rank exposing vector for the first iteration of the facial reduction algorithm give us that,

$$\theta \in \left(0, \frac{\pi}{n-1}\right) \implies \text{sd}(\mathcal{F}) = 1. \tag{7.2.17}$$

Next we we consider $\theta = \pi/(n-1)$. In this case, (7.2.13) and (7.2.14) give us that,

$$c = -\frac{b}{\left(2\cos\left(\frac{\pi}{n-1}\right)\right)} \text{ and } d = a - \frac{b}{2}.$$

Substituting into (7.2.11) and rearranging we get,

$$\frac{b^2}{a}\left(\frac{1}{4} - \frac{1}{4\cos^2(\frac{\pi}{n-1})}\right) = 0.$$

Consequently $b = 0$, and $\text{rank}(\bar{Z}) = 1$ implying that $\text{sd}(\mathcal{F}) \geq 2$.

The only remaining case is that of $\theta = 0$. In this case $\text{rank}(\bar{X}) = 1$. Observe that setting,

$$a = 1, b = 2, c = -1, d = 0,$$

implies that $\bar{Z} \succeq 0$ by diagonal dominance, e.g., Theorem 1.12 of [80]. Moreover it can be verified that $\text{rank}(\bar{Z}) = n-1$ and that $\bar{Z} \in \text{range}(\mathcal{A}^*)$ where $\mathcal{A}$ is the linear map constructed as in the introduction to this chapter. These observations imply that $\text{sd}(\mathcal{F}) = 1$ when $\theta = 0$, completing the proof. □

Chapter 8

Numerical Case Studies

In this chapter we demonstrate how to numerically obtain the bounds derived in Chapter 5. Our analysis is focused on spectrahedra with larger singularity degree, although, we do study one instance with singularity degree 1, in order to demonstrate 'good' convergence. For some of the instances, the exact singularity degree is known, allowing us to test the quality of our bounds. In other instances, the singularity degree is not known and we use our bounds to provide an estimate of it. This chapter is based largely on the preprint [74], coauthored by the author of this thesis.

In order to study the notion that large singularity degree is sufficient, in some sense, for slow convergence, we consider bounded spectrahedra, satisfying Assumption 6.1.1. We follow the primal-dual central path of (6.1.6), where $B = I$, with a path-following algorithm based on the Gauss-Newton search direction, see [16, 42]. In our implementation of the algorithm, we initialize with $\alpha = 1$, and decrease α at each iteration by a factor of 0.6. For this reason it seems natural to consider sequences of the type $\{\sigma^k\}$, as in Chapter 5, where $\sigma = 0.6$.

For each value of k, we approximate the ratios $R_i(\sigma^k)$, $Q_{i,\sigma}(k)$, and $S_{i,\sigma}(k)$ of Chapter 5, by following the primal-dual central path $(X(\alpha), y(\alpha), Z(\alpha))$ of (6.1.6) until $\alpha = \sigma^k$. We find that the plots of $Q_{i,\sigma}(k)$ and $S_{i,\sigma}(k)$ are very similar and for this reason we have chosen to only report the results for $S_{i,\sigma}(k)$, due to the stronger results of Theorem 5.1.9. Once the ratios have been obtained for k sufficiently large, around 60, we generate plots of the ratios against k for each i. The plots are then used to obtain bounds on maximum rank, forward error, and singularity degree. The bounds as well as the true values (when available) are recorded in Table 8.1.

The various bounds associated with a spectrahedron $\mathcal{F}$ and a proposed solution $\tilde{X}$ are denoted as follows:

- $\bar{r}$ - upper bound on $r = \text{rank}(\mathcal{F})$,
- $\underline{d}$ - lower bound on $\text{sd}(\mathcal{F})$,

- $\underline{\epsilon}$ - lower bound on $\epsilon^f(\tilde{X}, \mathcal{F})$,
- N_λ - number of different rates of convergence among eigenvalues of $X(\alpha)$ that vanish.

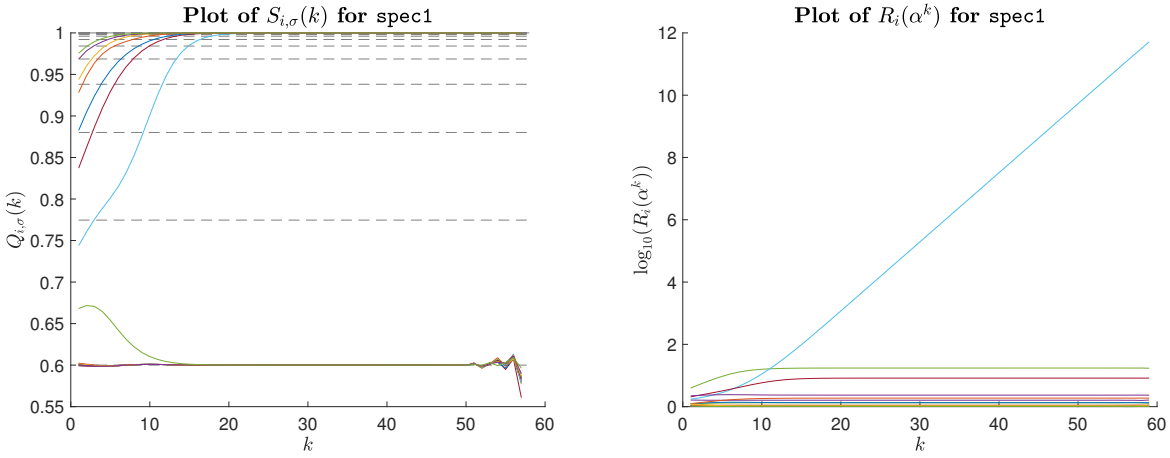

Figure 8.0.1: The dashed lines coincide with the values $\sigma^{\xi(i)}$ for the worst case scenario where $\mathrm{sd}(\mathcal{F}) = n - 1$

.

We denote the first test spectrahedron as `spec1`. Taking the approach of [85], we generate a primal-dual pair of SDPs that satisfy strict complementarity. As in Section 4.5, we take the optimal set of the primal SDP as our spectrahedron. Therefore, the singularity degree of the spectrahedron is 1. We generate `spec1` with $n = 20$ and plots of the ratios $S_{i,\sigma}(k)$ and $R_i(\sigma^k)$ are shown in Figure 8.0.1. In the left image, there is a clear distinction between curves that converge to 1 and curves that do not converge to 1. Moreover, if we disregard the irregularity in the last few values of the curves that do not converge to 1, we may conclude that those curves converge to the smallest dashed line located at 0.6. This observation, together with Theorem 6.2.1, correctly indicates that the spectrahedron has singularity degree 1. Exactly 13 of the curves converge to 0.6, yielding $\overline{r} = 7$, the correct approximation of the maximal rank r. The plot on the right side of the figure shows that exactly one curve blows up and it is the curve corresponding to $i = 7 = r$. This indicates, as expected, two groups of eigenvalues of X: those that converge to positive values and those that vanish.

The second spectrahedron, denoted `spec2`, is the classical 'worst case' problem of [80], as presented in Example 4.2.6, with $n = 5$ and singularity degree $n - 1 = 4$. Plots of the two ratios are in Figure 8.0.2. The left image shows 5 distinct rates of convergence among the eigenvalue sums. All but one of the curves converge to values that are clearly below 1. This indicates, correctly, that the maximum rank is at most 1. The largest of the curves appears to converge to the highest of the dashed lines. Thus we may infer that singularity degree is at least 4. Since $4 = n - 1$, the worst case upper bound, we may conclude that singularity degree is exactly 4. The row corresponding to `spec2` in Table 8.1 shows a very

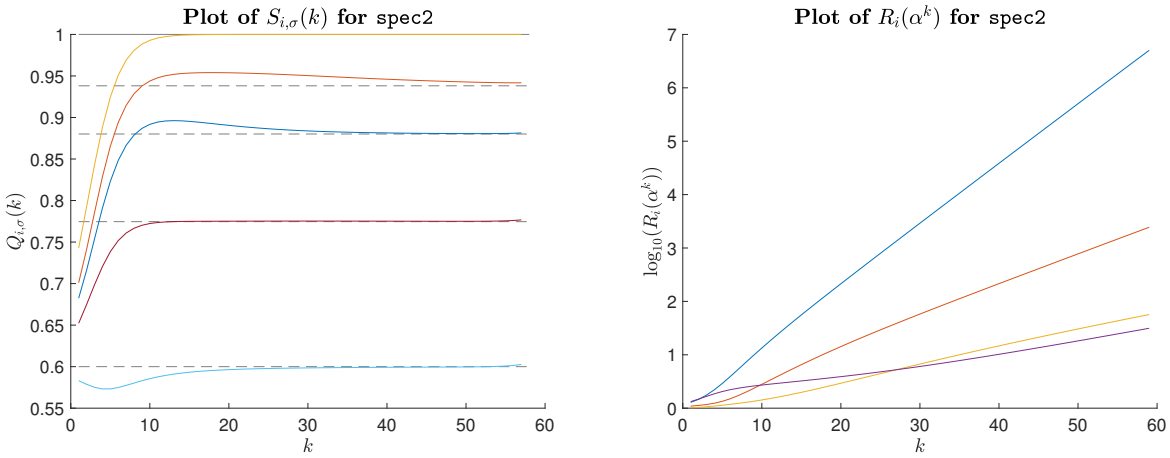

Figure 8.0.2:

large discrepancy between forward error and backward error. Our lower bound is actually quite close to the true forward error. Now let us consider the image on the right. It may be somewhat speculative to assert that the two lower curves blow up. Thus, taking the more cautious approach we assume that only the two larger curves blow up. Checking the indices of these curves yields an upper bound of 3 on the maximum rank. We choose the notably lower estimate of 1 based on the left plot. On the other hand if we are to apply Corollary 6.2.6 then we would want an overestimate of the number of different rates of convergence among the eigenvalues of $X(\alpha)$ that vanish. For this number we include the two lower curves, giving a bound of 4.

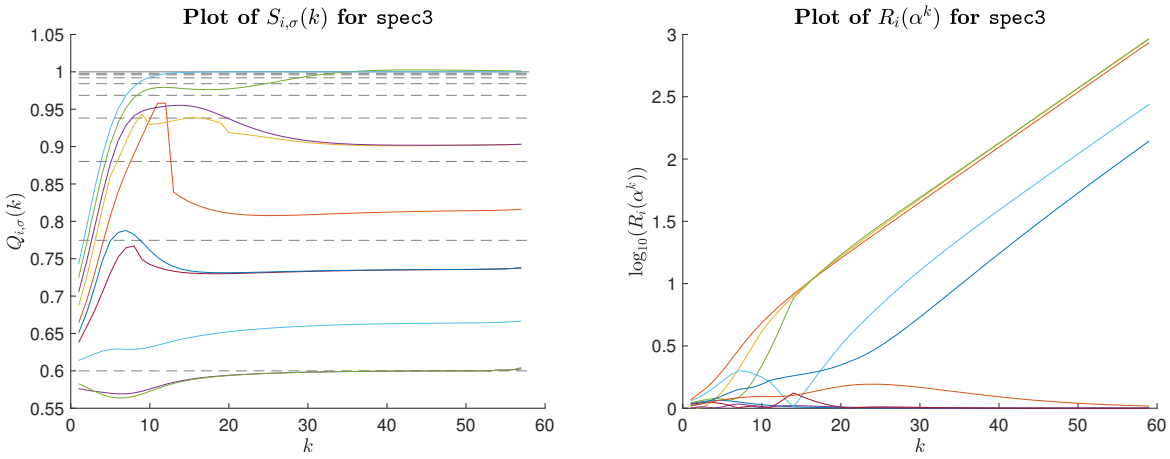

Figure 8.0.3:

For the next spectrahedron, `spec3`, the authors of [84] observed "strange behaviour" when attempting to optimize over it with an interior point method. The dimension is $n = 10$ and the singularity degree is proven to be 5. The left image of Figure 8.0.3

shows six distinct groups of curves. It is clear, for all but two of the curves, that the limit point is different from 1. Thus we have an upper bound of 2 on the maximum rank. The largest of the curves that does not converge to 1 appears to converge to a value that is below the fourth dashed line, indicating a lower bound of 4 on the singularity degree. Unlike the two previous spectrahedra, here the lower bound on singularity degree is a strict one. The image on the right shows exactly five different rates of convergence among the eigenvalues of $X(\alpha)$ that converge to 0. Moreover the upper bound on maximum rank corresponds to the one obtained from the left image.

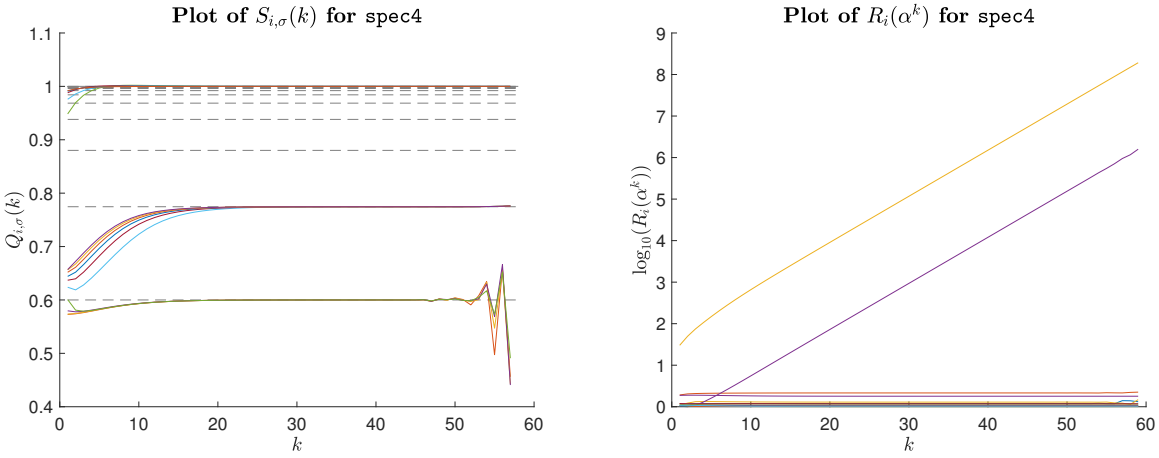

Figure 8.0.4:

The fourth spectrahedron, `spec4`, is generated by the algorithm of [85], just as `spec1` is. However, for this instance we require the existence of a complementarity gap. While we do not know the exact value of the singularity degree, the lack of strict complementarity implies, by Theorem 4.5.3, that singularity degree is at least 2. Plots of the ratios $S_{i,\sigma}(k)$ and $R_i(\sigma^k)$ are shown in Figure 8.0.4. From the left image we can be quite sure that those curves that converge to the second dashed line or below do not converge to 1. A closer inspection reveals that there are 10 such curves, implying an upper bound of 5 on the maximum rank. The corresponding lower bound on forward error, as recorded in Table 8.1, is indeed a lower bound and more informative than the reported backward error. We also obtain a lower bound on the singularity degree that coincides with the theoretical lower bound of 2. The image on right shows exactly two rates of convergence among eigenvalues of $X(\alpha)$ that converge to 0 and, once again, provides the same upper bound on maximum rank as obtained from the image on left.

The final spectrahedron we consider, `spec5`, is a Toeplitz cycle completion problem having the form of Theorem 7.2.4. Setting $n = 10$ and $\theta = \pi/(n-1)$ gives us a spectrahedron with singularity degree at least 2. In Figure 8.0.5, we find images of plots of the ratios $S_{i,\sigma}(k)$ and $R_i(\sigma^k)$. The left image indicates that all but two of the eigenvalues of $X(\alpha)$ converge to 0, yielding an exact approximation of maximum rank. Moreover, the corresponding eight curves appear to have limits below the second dashed line. Hence we

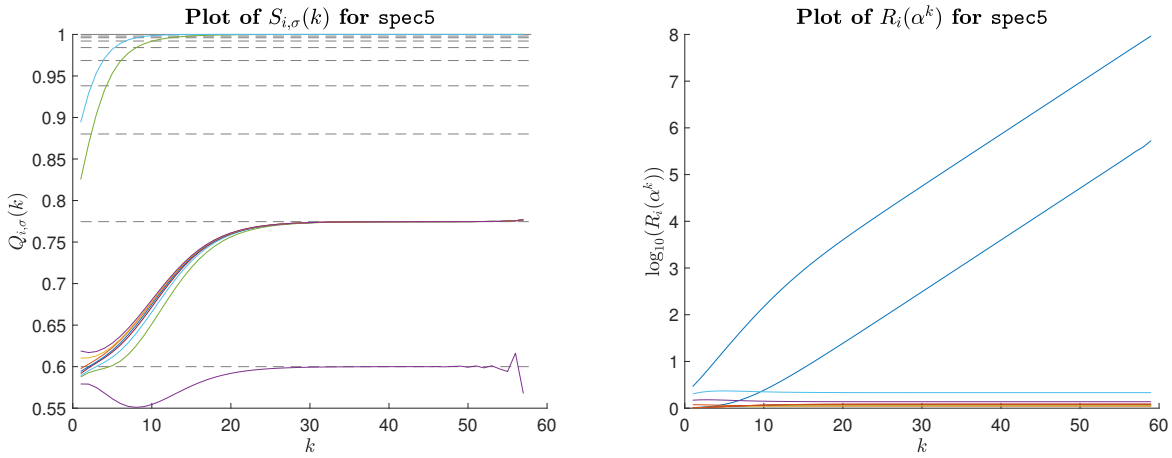

Figure 8.0.5:

have a lower bound of 2 on the singularity degree. This coincides with the theoretical lower bound.

Table 8.1: A record of relevant measures and their bounds for the spectrahedra considered in our analysis.

$\mathcal{F}$	$\epsilon^b(\mathcal{F})$	r	$\overline{r}$	$\epsilon^f(\mathcal{F})$	$\underline{\epsilon}$	$\text{sd}(\mathcal{F})$	$\underline{d}$	N_λ
spec1	6.62×10^{-11}	7	7	4.36×10^{-11}	3.10×10^{-12}	1	1	1
spec2	4.44×10^{-13}	1	1	4.93×10^{-2}	3.19×10^{-2}	4	4	4
spec3	2.47×10^{-13}	2	2	-	9.88×10^{-3}	5	4	5
spec4	1.88×10^{-11}	5	5	1.35×10^{-5}	4.16×10^{-7}	≥ 2	2	2
spec5	2.61×10^{-13}	2	2	1.22×10^{-6}	6.96×10^{-8}	≥ 2	2	2

In these case studies we have demonstrated the ability to upper bound maximum rank quite effectively. The resulting lower bound on forward error is of a much larger magnitude than backward error in all instances with the exception of spec1, where the singularity degree is 1. We see this feature as quite useful, as it alerts practitioners that the proposed solution is of substantially lower accuracy than backward error indicates. For spectrahedra with known singularity degree, we have demonstrated that the lower bound is quite accurate. In the other cases, the lower bound is in agreement with the theoretical lower bound. Lastly, for these test cases (as well as for others we have tested), the value N_λ seems to be an upper bound on singularity degree. Proving this, or demonstrating a counterexample, is an interesting topic for future research.

Chapter 9

Conclusion

We have studied the interplay between error bounds and singularity degree in semidefinite programming. In Chapter 4 we derived several theoretical bounds on singularity degree with respect to transformations of spectrahedra. In Chapter 5 we developed a numerical approach to bound forward error from below and thereby identify sequences that converge poorly to the spectrahedron even when the backward error is small. This approach also leads to a numerical lower bound on singularity degree, a measure that we view as intractable at the present time. In Chapter 6 we showed that larger singularity degree leads to greater irregularity in the convergence of a specific family of central paths. Our results were applied to Toeplitz matrix completions in Chapter 7, where we also proved some results that are independent of the rest of the thesis. The numerical bounds on error and singularity degree were tested on several spectrahedra in Chapter 8.

We hope that the results of this thesis may aid future research on spectrahedra, SDPs, and more generally, linear conic optimization. In conclusion we highlight some points of interest and future research.

1. In the numerical results of Chapter 8, the ratios $S_{i,\sigma}(k)$ always appear to converge. Is it possible, therefore, that additional properties of eigenvalue functions could be used to replace $\liminf$ with $\lim$ in the error bound results of Chapter 5?

2. In showing that singularity degree may be viewed as a measure of hardness for solving spectrahedra, our analysis was restricted to a specific family of central paths. We suspect that the analysis can be extended to other central paths. Is it possible, however, to extend the results to algorithms of a completely different nature? We already mentioned that Drusvyastkiy, Li, and Wolkowicz [17] used the bounds of Sturm to provide upper bounds on forward error for the iterates of the alternating projections algorithm applied to spectrahedra. Can it be shown that convergence of alternating projections is worse for spectrahedra with large singularity degree?

3. The error bounds we have derived rely heavily on properties of symmetric matrices, such as eigenvalues and the orthogonal spectral decomposition. In his extension of the

bounds of Sturm to amenable cones, Lourenço [46] used generalizations of eigenvalues. Given this additional machinery, can our results be extended to amenable cones?

4. In the numerical tests we observed that the number of different rates of convergence among eigenvalues of $X(\alpha)$ that vanish is an upper bound on singularity degree for instances where we have prior knowledge of singularity degree. It remains an open problem to determine whether this is true or not.

5. Structured spectrahedra, such as those arising from Toeplitz completion problems in Chapter 7, allow for theoretical analysis and may lead to interesting test problems. An area of future research is to study singularity degree from a theoretical perspective for other families of structured spectrahedra.

Index

References

[1] Mosek ApS. MOSEK optimization toolbox for MATLAB. *User's Guide and Reference Manual, version*, 4, 2019.

[2] D. Azé and J.-B. Hiriart-Urruty. Optimal Hoffman-type estimates in eigenvalue and semidefinite inequality constraints. *Journal of Global Optimization*, 24(2):133–147, 2002.

[3] M. Bakonyi and H.J. Woerdeman. *Matrix completions, moments, and sums of Hermitian squares*. Princeton University Press, Princeton, NJ, 2011.

[4] W. Barrett, C.R. Johnson, and P. Tarazaga. The real positive definite completion problem for a simple cycle. *Linear Algebra Appl.*, 192:3–31, 1993. Computational linear algebra in algebraic and related problems (Essen, 1992).

[5] C. Berge. *Topological spaces*. Dover Publications, Inc., Mineola, NY, 1997. Including a treatment of multi-valued functions, vector spaces and convexity, Translated from the French original by E. M. Patterson, Reprint of the 1963 translation.

[6] J.M. Borwein and H. Wolkowicz. Characterization of optimality for the abstract convex program with finite-dimensional range. *J. Austral. Math. Soc. Ser. A*, 30(4):390–411, 1980/81.

[7] J.M. Borwein and H. Wolkowicz. Facial reduction for a cone-convex programming problem. *J. Austral. Math. Soc. Ser. A*, 30(3):369–380, 1980/81.

[8] J.M. Borwein and H. Wolkowicz. Regularizing the abstract convex program. *J. Math. Anal. Appl.*, 83(2):495–530, 1981.

[9] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[10] Y-L. Cheung, S. Schurr, and H. Wolkowicz. Preprocessing and regularization for degenerate semidefinite programs. In D.H. Bailey, H.H. Bauschke, P. Borwein, F. Garvan, M. Thera, J. Vanderwerff, and H. Wolkowicz, editors, *Computational and Analytical Mathematics, In Honor of Jonathan Borwein's 60th Birthday*, volume 50 of *Springer Proceedings in Mathematics & Statistics*, pages 225–276. Springer, 2013.

[11] C. B. Chua. Analyticity of weighted central paths and error bounds for semidefinite programming. *Mathematical Programming*, 115(2):239–271, 2008.

[12] C.B. Chua and L. Tunçel. Invariance and efficiency of convex representations. *Math. Program.*, 111(1-2, Ser. B):113–140, 2008.

[13] E. de Klerk, C. Roos, and T. Terlaky. Initialization in semidefinite programming via a self-dual skew-symmetric embedding. *Oper. Res. Lett.*, 20(5):213–221, 1997.

[14] E. de Klerk, C. Roos, and T. Terlaky. Infeasible–start semidefinite programming algorithms via self–dual embeddings. In *Topics in Semidefinite and Interior-Point Methods*, volume 18 of *The Fields Institute for Research in Mathematical Sciences, Communications Series*, pages 215–236. American Mathematical Society, 1998.

[15] S. Deng and H. Hu. Computable error bounds for semidefinite programming. *J. Global Optim.*, 14(2):105–115, 1999.

[16] X.V. Doan, S. Kruk, and H. Wolkowicz. A robust algorithm for semidefinite programming. *Optim. Methods Softw.*, 27(4-5):667–693, 2012.

[17] D. Drusvyatskiy, G. Li, and H. Wolkowicz. A note on alternating projections for ill-posed semidefinite feasibility problems. *Math. Program.*, 162(1-2, Ser. A):537–548, 2017.

[18] D. Drusvyatskiy and H. Wolkowicz. The many faces of degeneracy in conic optimization. *Foundations and Trends® in Optimization*, 3(2):77–170, 2017.

[19] Mirjam Dür, Bolor Jargalsaikhan, and Georg Still. Genericity results in linear conic programming—a tour d'horizon. *Math. Oper. Res.*, 42(1):77–94, 2017.

[20] H. Dym and I. Gohberg. Extensions of band matrices with band inverses. *Linear Algebra Appl.*, 36:1–24, 1981.

[21] T. Ehrhardt and K. Rost. Resultant matrices and inversion of Bezoutians. *Linear Algebra and Applications*, 439(3):621–639, 2013.

[22] A.S. El-Bakry, R.A. Tapia, and Y. Zhang. A study of indicators for identifying zero variables in interior-point methods. *SIAM Rev.*, 36(1):45–72, 1994.

[23] R.L. Ellis, I. Gohberg, and D. Lay. Band extensions, maximum entropy and the permanence principle. In *Maximum entropy and Bayesian methods in applied statistics (Calgary, Alta., 1984)*, pages 131–155. Cambridge Univ. Press, Cambridge, 1986.

[24] K. Fan. On a theorem of Weyl concerning eigenvalues of linear transformations i. *Proc. Nat. Acad. Sci. U.S.A.*, 35:652–655, 1949.

[25] G. Finke, R.E. Burkard, and F. Rendl. Quadratic assignment problems. *Ann. Discrete Math.*, 31:61–82, 1987.

[26] I.C. Gohberg and A.A. Semencul. The inversion of finite Toeplitz matrices and their continual analogues. *Mat. Issled.*, 7(2(24)):201–223, 290, 1972.

[27] D. Goldfarb and K. Scheinberg. Interior point trajectories in semidefinite programming. *SIAM J. Optim.*, 8(4):871–886, 1998.

[28] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008. http://stanford.edu/~boyd/graph_dcp.html.

[29] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21. http://cvxr.com/cvx, April 2011.

[30] B. Grone, C.R. Johnson, E. Marques de Sa, and H. Wolkowicz. Positive definite completions of partial Hermitian matrices. *Linear Algebra Appl.*, 58:109–124, 1984.

[31] M. Halická. Analyticity of the central path at the boundary point in semidefinite programming. *European J. Oper. Res.*, 143(2):311–324, 2002. Interior point methods (Budapest, 2000).

[32] M. Halická, E. de Klerk, and C. Roos. On the convergence of the central path in semidefinite optimization. *SIAM J. Optim.*, 12(4):1090–1099 (electronic), 2002.

[33] M. He and M. K. Ng. Toeplitz and positive semidefinite completion problem for cycle graph. *Numer. Math. J. Chinese Univ. (English Ser.)*, 14(1):67–78, 2005.

[34] G. Heinig and K. Rost. Introduction to Bezoutians. In *Numerical methods for structured matrices and applications*, volume 199 of *Oper. Theory Adv. Appl.*, pages 25–118. Birkhäuser Verlag, Basel, 2010.

[35] A.J. Hoffman. On approximate solutions of systems of linear inequalities. *J. of Research of the National Bureau of Standards*, 49:263–265, 1952.

[36] A.J. Hoffman and H.W. Wielandt. The variation of the spectrum of a normal matrix. *Duke Mathematics*, 20:37–39, 1953.

[37] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1990. Corrected reprint of the 1985 original.

[38] C. R. Johnson, M. Lundquist, and G. Nævdal. Positive definite Toeplitz completions. *Journal of the London Mathematical Society*, 59(2):507–520, 1999.

[39] T. Kailath, A. Vieira, and M. Morf. Inverses of Toeplitz operators, innovations, and orthogonal polynomials. *SIAM review*, 20(1):106–119, 1978.

[40] M.G. Kreĭn and M.A. Naĭmark. The method of symmetric and Hermitian forms in the theory of the separation of the roots of algebraic equations. *Linear and Multilinear Algebra*, 10(4):265–308, 1981. Translated from the Russian by O. Boshko and J. L. Howland.

[41] N. Krislock and H. Wolkowicz. Explicit sensor network localization using semidefinite representations and facial reductions. *SIAM J. Optim.*, 20(5):2679–2708, 2010.

[42] S. Kruk, M. Muramatsu, F. Rendl, R.J. Vanderbei, and H. Wolkowicz. The Gauss-Newton direction in semidefinite programming. *Optim. Methods Softw.*, 15(1):1–28, 2001.

[43] F.I. Lander. The Bezoutian and the inversion of Hankel and Toeplitz matrices. *Mat. Issled.*, 9(2):69–87, 1974.

[44] A.S. Lewis and J-S. Pang. Error bounds for convex inequality systems. In *Generalized convexity, generalized monotonicity: recent results (Luminy, 1996)*, volume 27 of *Nonconvex Optim. Appl.*, pages 75–110. Kluwer Acad. Publ., Dordrecht, 1998.

[45] M. Liu and G. Pataki. Exact duality in semidefinite programming based on elementary reformulations. *SIAM Journal on Optimization*, 25(3):1441–1454, 2015.

[46] B. F. Lourenço. Amenable cones: error bounds without constraint qualifications. *arXiv preprint arXiv:1712.06221*, 2017.

[47] B. F. Lourenço, M. Muramatsu, and T. Tsuchiya. Solving SDP completely with an interior point oracle. *arXiv preprint arXiv:1507.08065*, 2015.

[48] B. F. Lourenço, M. Muramatsu, and T. Tsuchiya. A structural geometrical analysis of weakly infeasible SDPs. *Journal of the Operations Research Society of Japan*, 59(3):241–257, 2016.

[49] B. F. Lourenço, M. Muramatsu, and T. Tsuchiya. Facial reduction and partial polyhedrality. *SIAM Journal on Optimization*, 28(3):2304–2326, 2018.

[50] Z-Q. Luo, J. F. Sturm, and S. Zhang. Superlinear convergence of a symmetric primal-dual path-following algorithm for semidefinite programming. *SIAM J. Optim.*, 8:59–81, 1998.

[51] Z-Q. Luo, J.F. Sturm, and S. Zhang. Duality results for conic convex programming. Technical Report Report 9719/A, April, Erasmus University Rotterdam, Econometric Institute EUR, P.O. Box 1738, 3000 DR, The Netherlands, 1997.

[52] Z-Q. Luo, J.F. Sturm, and S. Zhang. Conic convex programming and self-dual embedding. *Optim. Methods Softw.*, 14(3):169–218, 2000.

[53] M. Marden. *Geometry of polynomials*. Second edition. Mathematical Surveys, No. 3. American Mathematical Society, Providence, R.I., 1966.

[54] J. Milnor. *Singular points of complex hypersurfaces*. Annals of Mathematics Studies, No. 61. Princeton University Press, Princeton, N.J.; University of Tokyo Press, Tokyo, 1968.

[55] A. Mohammad-Nezhad and T. Terlaky. On the identification of the optimal partition for semidefinite optimisation. *INFOR: Information Systems and Operational Research*, pages 1–39, 2019.

[56] B.N. Mukherjee and S.S. Maiti. On some properties of positive definite Toeplitz matrices and their possible applications. *Linear Algebra Appl.*, 102:211–240, 1988.

[57] G. Nævdal. On a generalization of the trigonometric moment problem. *Linear Algebra Appl.*, 258:1–18, 1997.

[58] Y.E. Nesterov, M.J. Todd, and Y. Ye. Infeasible-start primal-dual methods and infeasibility detectors for nonlinear programming problems. *Math. Program.*, 84(2, Ser. A):227–267, 1999.

[59] J. Nocedal and S.J. Wright. *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.

[60] J. Pang. Error bounds in mathematical programming. *Math. Programming*, 79(1-3, Ser. B):299–332, 1997. Lectures on mathematical programming (ismp97) (Lausanne, 1997).

[61] G. Pataki. Strong duality in conic linear programming: facial reduction and extended duals. In David Bailey, Heinz H. Bauschke, Frank Garvan, Michel Thera, Jon D. Vanderwerff, and Henry Wolkowicz, editors, *Computational and analytical mathematics*, volume 50 of *Springer Proc. Math. Stat.*, pages 613–634. Springer, New York, 2013.

[62] G. Pataki. Bad semidefinite programs: they all look the same. *SIAM Journal on Optimization*, 27(1):146–172, 2017.

[63] G. Pataki. On positive duality gaps in semidefinite programming. *arXiv preprint arXiv:1812.11796*, 2018.

[64] G. Pataki and L. Tunçel. On the generic properties of convex optimization problems in conic form. *Math. Programming*, to appear.

[65] F. Permenter, H. A. Friberg, and E. D. Andersen. Solving conic optimization problems via self-dual embedding and facial reduction: a unified approach. *SIAM Journal on Optimization*, 27(3):1257–1282, 2017.

[66] F. Permenter and P. Parrilo. Partial facial reduction: simplified, equivalent SDPs via approximations of the PSD cone. Technical Report Preprint arXiv:1408.4685, MIT, Boston, MA, 2014.

[67] I. Pólik and T. Terlaky. Exact duality for optimization over symmetric cones. Technical report, McMaster University, Hamilton, Ontario, Canada, 2007.

[68] F.A. Potra and R. Sheng. A superlinearly convergent primal-dual infeasible-interior-point algorithm for semidefinite programming. Technical Report Reports on Computational Mathematics, 78, University of Iowa, Iowa City, IA, 1995.

[69] M.V. Ramana. An exact duality theory for semidefinite programming and its complexity implications. *Math. Programming*, 77(2):129–162, 1997.

[70] M.V. Ramana, L. Tunçel, and H. Wolkowicz. Strong duality for semidefinite programming. *SIAM J. Optim.*, 7(3):641–662, 1997.

[71] R.T. Rockafellar. *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[72] R.T. Rockafellar and R.J.-B. Wets. *Variational analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1998.

[73] S. Sremac, F. Wang, H. Wolkowicz, and L. Pettersson. Noisy euclidean distance matrix completion with a single missing node. *Journal of Global Optimization*, Aug 2019.

[74] S. Sremac, H. J. Woerdeman, and H. Wolkowicz. Error bounds and singularity degree in semidefinite programming. *arXiv preprint arXiv:1908.04357*, 2019.

[75] S. Sremac, H.J. Woerdeman, and H. Wolkowicz. Maximum determinant positive definite Toeplitz completions. In *Operator Theory, Analysis and the State Space Approach: In Honor of Rien Kaashoek*, volume 271, pages 421–441. Birkhäuser/Springer, Cham, 2018.

[76] J.F. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optim. Methods Softw.*, 11/12(1-4):625–653, 1999. sedumi.ie.lehigh.edu.

[77] J.F. Sturm. Error bounds for linear matrix inequalities. *SIAM J. Optim.*, 10(4):1228–1248 (electronic), 2000.

[78] R.A. Tapia. On the role of slack variables in quasi-Newton methods for constrained optimization. In L.C.W. Dixon and T.P. Szego, editors, *Numerical Optimization of Dynamical Systems*, pages 235–246. North-Holland, Amsterdam, 1980.

[79] V. A. Truong and L. Tunçel. Geometry of homogeneous convex cones, duality mapping, and optimal self-concordant barriers. *Mathematical Programming*, 100(2), 2004.

[80] L. Tunçel. *Polyhedral and Semidefinite Programming Methods in Combinatorial Optimization*, volume 27 of *Fields Institute Monographs*. American Mathematical Society, Providence, RI, 2010.

[81] R. H. Tütüncü, K. C. Toh, and M. J. Todd. Solving semidefinite-quadratic-linear programs using SDPT3. *Math. Program.*, 95(2, Ser. B):189–217, 2003. Computational semidefinite and second order cone programming: the state of the art.

[82] L. Vandenberghe, S. Boyd, and S-P. Wu. Determinant maximization with linear matrix inequality constraints. *SIAM J. Matrix Anal. Appl.*, 19(2):499–533, 1998.

[83] H. Waki and M. Muramatsu. Facial reduction algorithms for conic optimization problems. *J. Optim. Theory Appl.*, 158(1):188–215, 2013.

[84] H. Waki, M. Nakata, and M. Muramatsu. Strange behaviors of interior-point methods for solving semidefinite programming problems in polynomial optimization. *Computational Optimization and Applications*, 53(3):823–844, 2012.

[85] H. Wei and H. Wolkowicz. Generating and measuring instances of hard semidefinite programs. *Math. Program.*, 125(1, Ser. A):31–45, 2010.

[86] H. Wolkowicz, R. Saigal, and L. Vandenberghe, editors. *Handbook of semidefinite programming*. International Series in Operations Research & Management Science, 27. Kluwer Academic Publishers, Boston, MA, 2000. Theory, algorithms, and applications.

[87] Y. Zhu, G. Pataki, and Q. Tran-Dinh. Sieve-SDP: a simple facial reduction algorithm to preprocess semidefinite programs. *Mathematical Programming Computation*, 11(3):503–586, Sep 2019.