Foundations and Trends® in Optimization
Vol. 2, No. 1-2 (2015) 1–156
© 2016 A. Lemon, A. M.-C. So, Y. Ye
DOI: 10.1561/2400000009

Low-Rank Semidefinite Programming: Theory and Applications

Alex Lemon
Stanford University
adlemon@stanford.edu

Anthony Man-Cho So
The Chinese University of Hong Kong
manchoso@se.cuhk.edu.hk

Yinyu Ye
Stanford University
yyye@stanford.edu

Contents

Abstract

Finding low-rank solutions of semidefinite programs is important in many applications. For example, semidefinite programs that arise as relaxations of polynomial optimization problems are exact relaxations when the semidefinite program has a rank-1 solution. Unfortunately, computing a minimum-rank solution of a semidefinite program is an NP-hard problem. In this paper we review the theory of low-rank semidefinite programming, presenting theorems that guarantee the existence of a low-rank solution, heuristics for computing low-rank solutions, and algorithms for finding low-rank approximate solutions. Then we present applications of the theory to trust-region problems and signal processing.

A. Lemon, A. M.-C. So, Y. Ye. *Low-Rank Semidefinite Programming:
Theory and Applications*. Foundations and Trends® in Optimization, vol. 2,
no. 1-2, pp. 1–156, 2015.
DOI: 10.1561/2400000009.

Introduction

1

Introduction

1.1 Low-rank semidefinite programming

A semidefinite program (SDP) is an optimization problem of the form

$$\begin{array}{ll} \text{minimize} & C \bullet X \\ \text{subject to} & A_i \bullet X = b_i, \quad i = 1, \dots, m \\ & X \succeq 0. \end{array} \tag{SDP}$$

The optimization variable is $X \in \mathbf{S}^n$, where $\mathbf{S}^n$ is the set of all $n \times n$ symmetric matrices, and the problem data are $A_1, \dots, A_m, C \in \mathbf{S}^n$ and $b \in \mathbf{R}^m$. The trace inner product of $A, B \in \mathbf{R}^{m \times n}$ is

$$A \bullet B = \mathbf{tr}(A^\mathsf{T} B) = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} B_{ij}.$$

The constraint $X \succeq 0$ denotes a generalized inequality with respect to the cone of positive-semidefinite matrices, and means that X is positive semidefinite: that is, $z^\mathsf{T} X z \geq 0$ for all $z \in \mathbf{R}^n$. We can write (SDP) more compactly by defining the operator $\mathcal{A} : \mathbf{S}^n \to \mathbf{R}^m$ such that

$$\mathcal{A}(X) = \begin{bmatrix} A_1 \bullet X \\ \vdots \\ A_m \bullet X \end{bmatrix}.$$

Using this notation we can express (SDP) as

$$\begin{array}{ll} \text{minimize} & C \bullet X \\ \text{subject to} & \mathcal{A}(X) = b \\ & X \succeq 0. \end{array}$$

The dual problem of (SDP) is

$$\begin{array}{ll} \text{maximize} & b^{\mathsf{T}} y \\ \text{subject to} & \sum_{i=1}^{m} y_i A_i + S = C \\ & S \succeq 0, \end{array} \tag{SDD}$$

where the optimization variables are $y \in \mathbf{R}^m$ and $S \in \mathbf{S}^n$. We can write (SDD) more succinctly as

$$\begin{array}{ll} \text{maximize} & b^{\mathsf{T}} y \\ \text{subject to} & \mathcal{A}^*(y) + S = C \\ & S \succeq 0, \end{array}$$

where the adjoint operator $\mathcal{A}^* : \mathbf{R}^m \to \mathbf{S}^n$ is given by

$$\mathcal{A}^*(y) = \sum_{i=1}^{m} y_i A_i.$$

We do not attempt to give a general exposition of the theory of semidefinite programming in this paper – an excellent survey is provided by Vandenberghe and Boyd [96]. The preceding remarks are only meant to establish our particular conventions for talking about SDPs. Additional results about SDPs are given in Appendix A, which presents those aspects of the theory that are most relevant for our purposes.

Semidefinite programs can be solved efficiently using interior-point algorithms. However, such algorithms typically converge to a maximum-rank solution [45], and in many cases we are interested in finding a low-rank solution. For example, it is well known that every polynomial optimization problem has a natural SDP relaxation, and this relaxation is exact when it has a rank-1 solution. (We include the derivation of this important result in Appendix A for completeness.) Unfortunately, finding a minimum-rank solution of an SDP is NP-hard: a special case of this problem is finding a minimum-cardinality solution of a system of linear equations, which is known to be NP-hard [36]. In this paper we review approaches to finding low-rank solutions and approximate solutions of SDPs, and present some applications in which low-rank solutions are important.

1.2 Outline

Chapter 2 discusses reduced-rank exact solutions of SDPs and theorems about rank. We give an efficient algorithm for reducing the rank of a solution. Although the algorithm may not find a minimum-rank solution, it often works well in practice, and we can prove a bound on the rank of the solution returned by the algorithm. Then we give a theorem relating the uniqueness of the rank of a solution to the uniqueness of the solution itself, and show how to use this theorem for sensor-network localization. The chapter concludes with a theorem that allows us to deduce the existence of a low-rank solution from the sparsity structure of the coefficients.

Because finding a minimum-rank solution of an SDP is NP-hard, we do not expect to arrive at an algorithm that accomplishes this task in general. However, there are many heuristics for finding low-rank solutions that often perform well in practice; we discuss these methods in Chapter 3. We also present rounding methods, in which we find a low-rank approximate solution that is close to a given exact solution in some sense. One of the rounding methods that we discuss is the famous Goemans-Williamson algorithm [39]; if the unique-games conjecture is true, then this algorithm achieves the best possible approximation ratio for the maximum-cut problem [57, 58].

The paper concludes with two chapters covering applications of the theoretical results to trust-region problems and signal processing. There are three appendices: the first gives background information, and establishes our notation; the second reviews some classical results about linear programming that we generalize to semidefinite programming in Chapter 2; and the last contains technical probability lemmas that are used in our analysis of rounding methods.

Part I

Theory

2

Exact Solutions and Theorems about Rank

2.1 Introduction

In this chapter we discuss exact reduced-rank solutions and theorems about rank. We begin by extending classical results about the cardinality of solutions of linear programs (reviewed in Appendix B) to results about the rank of solutions of semidefinite programs. Then we present a theorem that allows us to use the sparsity structure of the coefficients of an SDP to guarantee the existence of a low-rank solution.

2.2 Rank reduction for semidefinite programs

We now generalize the well-known analysis of sparsification for linear programs to rank reduction for semidefinite programs. (The corresponding sparsification algorithm for LPs is described in Appendix B; our presentation of the corresponding rank-reduction algorithm for SDPs is deliberately similar.) The main result is Theorem 2.1, which guarantees the existence of a solution of (SDP) whose rank r satisfies $r(r+1)/2 \leq m$, where m is the number of linear equality constraints. This bound was independently discovered by Barvinok [3] and Pataki [76].

Suppose we are given a solution X of (SDP), and we want to find another solution X^+ with $\mathbf{rank}(X^+) < \mathbf{rank}(X)$. If we had an efficient method for this problem that worked on every solution that does not have minimum rank, then we could find a minimum-rank solution by applying this rank-reduction method at most n times. However, we know that the problem of finding a minimum-rank solution of (SDP) is NP-hard. Thus, we do not expect to find an efficient rank-reduction algorithm that always works. Nonetheless, we still hope to find a method that often performs well in practice. We begin by making the following assumption:

$$\mathbf{null}(X^+) \supset \mathbf{null}(X), \tag{2.1}$$

where $\mathbf{null}(X) = \{z \in \mathbf{R}^n \mid Xz = 0\}$ is the nullspace of X. Observe the strong similarity between this assumption, and the assumption (B.1) that forms the basis for the standard LP-sparsification algorithm: in the LP case, we assume that if a component of x is zero, then the corresponding component of x^+ is also zero; in the SDP case, we assume that if X has zero gain in some direction, then X^+ also has zero gain in that direction. (Here we interpret a nonzero vector being in the nullspace of a matrix as meaning that the matrix has zero gain in the direction of the vector.) The following example shows that this assumption can yield suboptimal results in some cases.

Example 2.1. Consider the SDP feasibility problem

$$\begin{aligned} X_{ii} + X_{nn} &= 1, \quad i = 1, \ldots, n-1 \\ X_{ij} &= 0, \quad 1 \le i < j \le n \\ X &\succeq 0, \end{aligned}$$

and suppose we are given the solution $X = \mathbf{diag}(1, \ldots, 1, 0)$. If we assume that $\mathbf{null}(X^+) \supset \mathbf{null}(X)$, then we have that $e_n \in \mathbf{null}(X^+)$ because $e_n \in \mathbf{null}(X)$. (Here e_n is the nth standard basis vector in $\mathbf{R}^n$: that is, the vector of length n whose nth component is equal to 1, and whose other components are all equal to 0.) This implies that $X^+_{nn} = e_n^\mathsf{T} X^+ e_n = 0$. Then the equality constraint $X^+_{ii} + X^+_{nn} = 1$ implies that $X^+_{ii} = 1$ for $i = 1, \ldots, n-1$. Therefore, we have that $X^+ = X$, and we are unable to reduce the rank of X. However, X is not a minimum-rank solution of the feasibility problem: $\mathbf{rank}(X) = n-1$, but $\tilde{X} = \mathbf{diag}(0, \dots, 0, 1)$ is a solution with $\mathbf{rank}(\tilde{X}) = 1$.

Example 2.1 shows that the assumption in (2.1) may not only lead to suboptimal results, but may even lead to arbitrarily poor results: for every positive integer n, there is an instance of (SDP) and a corresponding initial solution such that our algorithm returns a solution whose rank is $n-1$ times the rank of a minimum-rank solution. However, because we do not expect to find an algorithm that works on every instance of (SDP), we need to make a suboptimal assumption at some point. Moreover, we will see that the assumption in (2.1) allows us to derive an algorithm that often works well, and has some performance guarantees.

We have stated our assumption as $\mathbf{null}(X^+) \supset \mathbf{null}(X)$. Although this statement is clear and intuitive, a different formulation will prove useful in the development of our algorithm. Let $r = \mathbf{rank}(X)$. Since X is positive semidefinite, there exists a matrix $V \in \mathbf{R}^{n\times r}$ such that $X = VV^{\mathsf{T}}$. Then assumption (2.1) is equivalent to assuming that X^+ has the form

$$X^+ = V(I + \alpha\Delta)V^{\mathsf{T}},$$

where we think of $\Delta \in \mathbf{S}^n$ as an update direction, and $\alpha \in \mathbf{R}$ as a step size. We will also sometimes find it convenient to write X^+ as

$$X^+ = X + \alpha V \Delta V^{\mathsf{T}}.$$

The fact that the proposed reformulation is equivalent to (2.1) is a consequence of Proposition A.1.

We want to choose α and Δ such that X^+ is a solution of (SDP), and $\mathbf{rank}(X^+) < \mathbf{rank}(X)$. Since the rank of X^+ is strictly less than that of X, we must have that $X^+ \neq X$, and hence that $\alpha \neq 0$.

- In order to maintain optimality, we require that

  $$C \bullet X^+ = C \bullet X.$$

  Substituting in $X^+ = X + \alpha V \Delta V^{\mathsf{T}}$ and simplifying, we obtain the condition

  $$(V^{\mathsf{T}} C V) \bullet \Delta = 0.$$

- We also need X^+ to satisfy the equality constraints
$$A_i \bullet X^+ = b_i, \quad i = 1, \ldots, m.$$
Substituting in our expression for X^+ and simplifying gives the conditions
$$(V^\mathsf{T} A_i V) \bullet \Delta = 0, \quad i = 1, \ldots, m.$$
For convenience we define the mapping $\mathcal{A}_V : \mathbf{S}^r \to \mathbf{R}^m$ such that
$$\mathcal{A}_V(\Delta) = \mathcal{A}(V\Delta V^\mathsf{T}) = \begin{bmatrix} (V^\mathsf{T} A_1 V) \bullet \Delta \\ \vdots \\ (V^\mathsf{T} A_m V) \bullet \Delta \end{bmatrix} = \begin{bmatrix} A_1 \bullet (V\Delta V^\mathsf{T}) \\ \vdots \\ A_m \bullet (V\Delta V^\mathsf{T}) \end{bmatrix}.$$
Then we can express our condition as $\mathcal{A}_V(\Delta) = 0$.
- The updated solution must satisfy $X^+ = V(I+\alpha\Delta)V^\mathsf{T} \succeq 0$. Since V is skinny and full rank, this is equivalent to the condition
$$I + \alpha\Delta \succeq 0.$$
- Finally, we have that $\mathbf{rank}(X^+) < \mathbf{rank}(X)$ if and only if $I+\alpha\Delta$ is singular.

In summary we want to choose α and Δ in order to satisfy the following conditions:
$$\begin{aligned} (V^\mathsf{T} C V) \bullet \Delta &= 0 \\ \mathcal{A}_V(\Delta) &= 0 \\ I + \alpha\Delta &\succeq 0 \\ I + \alpha\Delta &\text{ is singular.} \end{aligned}$$
It turns out that the first constraint is implied by the second constraint. The main idea is that because the nullspace of X^+ contains the nullspace of X, the updated solution X^+ automatically satisfies the complementary-slackness condition. Therefore, X^+ is optimal whenever it is feasible. We make this argument more precise in the proof of the following proposition.

Proposition 2.1. Suppose $X = VV^\mathsf{T}$ is a solution of (SDP). If $\mathcal{A}_V(\Delta) = 0$, then $(V^\mathsf{T} C V) \bullet \Delta = 0$.

Proof. Consider the semidefinite program

$$\begin{array}{ll} \text{minimize} & (V^{\mathsf T}CV) \bullet \tilde{X} \\ \text{subject to} & (V^{\mathsf T}A_iV) \bullet \tilde{X} = b_i, \quad i = 1, \ldots, m \\ & \tilde{X} \succeq 0 \end{array} \tag{2.2}$$

with variable $\tilde{X} \in \mathbf{S}^r$. Note that $\tilde{X} = I$ is strictly feasible for this problem because

$$(V^{\mathsf T}A_iV) \bullet I = A_i \bullet (VV^{\mathsf T}) = A_i \bullet X = b_i,$$

and $I \succ 0$. Similarly, for every feasible point $\tilde{X}$ of (2.2), we have that $V\tilde{X}V^{\mathsf T}$ is feasible for (SDP), and achieves an objective value of

$$C \bullet (V\tilde{X}V^{\mathsf T}) = (V^{\mathsf T}CV) \bullet \tilde{X}.$$

Since X is optimal for (SDP), we have that

$$(V^{\mathsf T}CV) \bullet I = C \bullet (VV^{\mathsf T}) = C \bullet X \le C \bullet (V\tilde{X}V^{\mathsf T}) = (V^{\mathsf T}CV) \bullet \tilde{X}$$

for every feasible point $\tilde{X}$ of (2.2). Thus, we see that I is optimal for (2.2). Moreover, since Slater's condition is satisfied (we remarked above that I is strictly feasible), we can find optimal dual variables $\tilde{S} \in \mathbf{S}^r$ and $\tilde{y} \in \mathbf{R}^m$ satisfying the KKT conditions

$$\begin{aligned} \sum_{i=1}^m \tilde{y}_i(V^{\mathsf T}A_iV) + \tilde{S} &= V^{\mathsf T}CV \\ (V^{\mathsf T}A_iV) \bullet \tilde{X} &= b_i, \qquad i = 1, \ldots, m \\ \tilde{S}, \tilde{X} &\succeq 0 \\ \tilde{S}\tilde{X} &= 0. \end{aligned}$$

Because $\tilde{X} = I$ is a solution of (2.2), the last condition implies that $\tilde{S} = 0$, and hence that every feasible point of (2.2) is optimal because it automatically satisfies complementary slackness. Since $\tilde{S} = 0$, the first KKT condition simplifies to

$$V^{\mathsf T}CV = \sum_{i=1}^m \tilde{y}_i(V^{\mathsf T}A_iV).$$

Thus, if $\mathcal{A}_V(\Delta) = 0$, or equivalently if

$$(V^{\mathsf T}A_iV) \bullet \Delta = 0, \quad i = 1, \ldots, m,$$

then we have that

$$(V^{\mathsf{T}}CV) \bullet \Delta = \left(\sum_{i=1}^{m} \tilde{y}_i (V^{\mathsf{T}} A_i V) \right) \bullet \Delta = \sum_{i=1}^{m} \tilde{y}_i ((V^{\mathsf{T}} A_i V) \bullet \Delta) = 0.$$

□

Note that the argument in the proof of Proposition 2.1 only works if X is a solution of (SDP). In particular, if we had an arbitrary feasible point X, and we wanted to find another feasible point X^+ with the same objective value, we could not ignore the condition $(V^{\mathsf{T}}CV) \bullet \Delta = 0$.

An algorithm for SDP rank reduction. A method for rank reduction is given in Algorithm 2.1. Using the observations above, we will prove that this algorithm returns a solution of (SDP), and derive a bound on the rank of this solution.

Algorithm 2.1: rank reduction for semidefinite programs

```
Input: a solution $X$ of (SDP)
repeat
    compute the factorization $X = VV^{\mathsf{T}}$
    find a nonzero $\Delta \in \mathbf{null}(\mathcal{A}_V)$ (if possible)
    find a maximum-magnitude eigenvalue $\lambda_1$ of $\Delta$
    $\alpha := -1/\lambda_1$
    $X := V\,(I + \alpha\Delta)\,V^{\mathsf{T}}$
until $\mathbf{null}(\mathcal{A}_V) = \{0\}$
```

Proposition 2.2. Given a solution $X = VV^{\mathsf{T}}$ of (SDP), Algorithm 2.1 returns another solution X^+ such that $\mathbf{rank}(X^+) \leq \mathbf{rank}(X)$. Moreover, this inequality is strict if $\mathbf{null}(\mathcal{A}_V) \neq \{0\}$.

Proof. In our preliminary analysis of the rank-reduction problem, we showed that X^+ is a solution of (SDP) with $\mathbf{rank}(X^+) < \mathbf{rank}(X)$ if α and Δ satisfy the following properties:

$$\mathcal{A}_V(\Delta) = 0 \tag{2.3}$$

$$I + \alpha\Delta \succeq 0 \tag{2.4}$$

$$I + \alpha\Delta \text{ is singular.} \tag{2.5}$$

In Algorithm 2.1 we choose Δ in order to satisfy (2.3). We then choose $\alpha = -1/\lambda_1$, where λ_1 is a maximum-magnitude eigenvalue of Δ. Note that λ_1 is nonzero because Δ is nonzero, so our choice of α is defined. Let $\Delta = Q\Lambda Q^{\mathsf{T}}$ be the eigenvalue decomposition of Δ, where $Q \in \mathbf{R}^{r\times r}$ is orthogonal, and $\Lambda \in \mathbf{R}^{r\times r}$ is diagonal. Then we have that

$$\begin{aligned} I + \alpha\Delta &= QQ^{\mathsf{T}} + \alpha Q\Lambda Q^{\mathsf{T}} \\ &= Q(I + \alpha\Lambda)Q^{\mathsf{T}} \\ &= Q\tilde{\Lambda}Q^{\mathsf{T}}, \end{aligned}$$

where we define the matrix

$$\begin{aligned} \tilde{\Lambda} &= I + \alpha\Lambda \\ &= \mathbf{diag}(1 + \alpha\lambda_1, \ldots, 1 + \alpha\lambda_n) \\ &= \mathbf{diag}\left(1 - \frac{\lambda_1}{\lambda_1}, \ldots, 1 - \frac{\lambda_n}{\lambda_1}\right). \end{aligned}$$

Our choice of α implies that $\tilde{\Lambda}$ is singular because it is a diagonal matrix whose first diagonal entry is zero. Additionally, $\tilde{\Lambda}$ is positive semidefinite: since we ordered the eigenvalues of Δ in descending order of magnitude, we have that $|\lambda_1| \geq |\lambda_i|$, and hence that

$$1 - \frac{\lambda_i}{\lambda_1} \geq 1 - \left|\frac{\lambda_i}{\lambda_1}\right| \geq 0$$

for $i = 1, \ldots, r$. Thus, conditions (2.4) and (2.5) are also satisfied.

This analysis shows that, after each iteration of the algorithm, we have that X^+ is still a solution of (SDP), and its rank is no larger than the rank of X. (Moreover, if $\mathbf{null}(\mathcal{A}_V) \neq \{0\}$, then the rank of X^+ is strictly less than that of X.) □

Theorem 2.1 (Barvinok [3] and Pataki [76]). If (SDP) is solvable, then it has a solution X with $\mathbf{rank}(X) = r$ such that $r(r+1)/2 \leq m$. Moreover, Algorithm 2.1 efficiently finds such a solution.

Proof. The termination condition for Algorithm 2.1 is

$$\mathbf{null}(\mathcal{A}_V) = \{0\},$$

where $\mathcal{A}_V : \mathbf{S}^r \to \mathbf{R}^m$ is a linear mapping, and r is the rank of the solution returned by the algorithm. Every linear mapping whose input

m	1	2	3	4	5	6	7	8	9	10
bound	1	1	2	2	2	3	3	3	3	4

Table 2.1: upper bounds on the minimum rank of a solution of (SDP)

space has strictly larger dimension than its output space has a nontrivial nullspace. Therefore, when the algorithm terminates, we must have that the dimension of the input space of $\mathcal{A}_V$ is less than or equal to the dimension of the output space of $\mathcal{A}_V$: that is,

$$\dim(\mathbf{S}^r) = \frac{r(r+1)}{2} \leq \dim(\mathbf{R}^m) = m.$$

Thus, Algorithm 2.1 returns a solution X with $\mathbf{rank}(X) = r$ satisfying $r(r+1)/2 \leq m$. □

Another way of stating the result of Theorem 2.1 is that there is a solution X of (SDP) such that

$$\mathbf{rank}(X) \leq \left\lfloor \frac{\sqrt{8m+1}-1}{2} \right\rfloor.$$

The values of this bound for small values of m are given in Table 2.1. As we will see later, it is particularly important for applications to note that if $m \leq 2$, then (SDP) has a solution with rank at most 1.

The following example shows that the bound in Theorem 2.1 cannot be improved without additional hypotheses.

Example 2.2. Suppose $r \leq n$, and consider the SDP feasibility problem

$$\begin{aligned} X_{ii} &= 1, \quad i = 1, \dots, r \\ X_{ij} &= 0, \quad 1 \leq i < j \leq r \\ X &\succeq 0 \end{aligned}$$

with variable $X \in \mathbf{S}^n$. The minimum-rank solution of this problem is $X = e_1 e_1^\mathsf{T} + \cdots + e_r e_r^\mathsf{T}$, which has rank r. There are r equality constraints of the form $X_{ii} = 1$, and $r(r-1)/2$ equality constraints of the form $X_{ij} = 0$. The total number of equality constraints is therefore

$$m = r + \frac{r(r-1)}{2} = \frac{r(r+1)}{2}.$$

Thus, the minimum-rank solution X for this problem has $\mathbf{rank}(X) = r$ satisfying $r(r+1)/2 = m$, where m is the number of linear equality constraints.

Example 2.3. Consider a norm-constrained quadratic optimization problem of the form

$$\begin{array}{ll} \text{minimize} & x^\mathsf{T} P x + 2q^\mathsf{T} x + r \\ \text{subject to} & \|x\| = 1, \end{array}$$

where $x \in \mathbf{R}^n$ is the optimization variable, and $P \in \mathbf{S}^n$, $q \in \mathbf{R}^n$, and $r \in \mathbf{R}$ are problem data. In particular, note that we do not assume that P is positive semidefinite, so the objective function may not be convex. We will have much more to say about such problems (which are called simple trust-region problems) in Chapter 4. The natural SDP relaxation of this problem is

$$\begin{array}{ll} \text{minimize} & \begin{bmatrix} P & q \\ q^\mathsf{T} & r \end{bmatrix} \bullet X \\ \text{subject to} & \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \bullet X = 1 \\ & \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \bullet X = 1 \\ & X \succeq 0, \end{array}$$

where the optimization variable is $X \in \mathbf{S}^{n+1}$. (See Chapter A for a review of the natural SDP relaxation of a polynomial optimization problem, and how to construct a solution of the polynomial optimization problem from a rank-1 solution of the SDP relaxation.) Theorem 2.1 allows us to conclude that the SDP relaxation has a rank-1 solution. Thus, we can solve the norm-constrained quadratic optimization problem by finding a rank-1 solution $X = vv^\mathsf{T} \in \mathbf{S}^{n+1}$ of the associated SDP, where $v \in \mathbf{R}^{n+1}$, and taking $x = v_{n+1}(v_1, \ldots, v_n)$.

Remark 2.1. Consider what happens when we apply Algorithm 2.1 to an instance of (SDP) with homogeneous equality constraints (that is, with $b = 0$). Then we can always choose $\Delta = I \in \mathbf{S}^r$. This choice of Δ works because

$$\mathcal{A}_V(I) = \mathcal{A}(VV^\mathsf{T}) = \mathcal{A}(X) = 0.$$

For this value of Δ, we have that $\alpha = -1$, and hence that

$$X^+ = V(I + \alpha\Delta)V^{\mathsf{T}} = V(I - I)V^{\mathsf{T}} = 0.$$

Thus, Algorithm 2.1 tells us that $X = 0$ is a solution of every solvable instance of SDP with homogeneous equality constraints. Note in particular the (easy-to-overlook) hypothesis in Theorem 2.1 that (SDP) is solvable. For example, consider the semidefinite program

$$\begin{array}{ll} \text{minimize} & -X_{11} \\ \text{subject to} & X_{22} = 0 \\ & X \succeq 0. \end{array}$$

The linear constraint for this problem is homogeneous, but $X = 0$ is not a solution: the problem is unbounded below, and not solvable.

Pataki [76] also showed how to use Theorem 2.1 to obtain an upper bound on the minimum rank of an optimal dual slack variable.

Corollary 2.2 (Pataki [76])**.** Consider an instance of (SDD) such that $A_1, \ldots, A_m$ are linearly independent. If (SDD) is solvable, then it has a solution (y, S) with $\mathbf{rank}(S) = s$ such that $s(s+1)/2 \leq n(n+1)/2 - m$.

Proof. We will prove the bound by converting (SDD) into a standard-form primal SDP in the dual slack variable S, and then applying Theorem 2.1. Under the assumption that $A_1, \ldots, A_m$ are linearly independent, the Gram matrix

$$G = \begin{bmatrix} A_1 \bullet A_1 & \cdots & A_1 \bullet A_m \\ \vdots & \ddots & \vdots \\ A_m \bullet A_1 & \cdots & A_m \bullet A_m \end{bmatrix}$$

is invertible. Therefore, we can use the constraint $\sum_{i=1}^m y_i A_i + S = C$ of (SDD) to solve for y in terms of S:

$$y = G^{-1} \begin{bmatrix} A_1 \bullet (C - S) \\ \vdots \\ A_m \bullet (C - S) \end{bmatrix}.$$

Defining $\beta = G^{-1}b$, we can write the objective function of (SDD) as

$$\begin{aligned} b^\mathsf{T} y &= b^\mathsf{T} G^{-1} \begin{bmatrix} A_1 \bullet (C - S) \\ \vdots \\ A_m \bullet (C - S) \end{bmatrix} \\ &= \beta^\mathsf{T} \begin{bmatrix} A_1 \bullet (C - S) \\ \vdots \\ A_m \bullet (C - S) \end{bmatrix} \\ &= \sum_{i=1}^m \beta_i (A_i \bullet (C - S)) \\ &= \left(\sum_{i=1}^m \beta_i A_i \right) \bullet C - \left(\sum_{i=1}^m \beta_i A_i \right) \bullet S. \end{aligned}$$

Another consequence of the assumption that $A_1, \ldots, A_m$ are linearly independent is that $\dim(\mathbf{span}(A_1, \ldots, A_m)) = m$. Therefore, we can find an orthonormal basis $Q_1, \ldots, Q_m \in \mathbf{S}^n$ for $\mathbf{span}(A_1, \ldots, A_m)$. Then we can extend $Q_1, \ldots, Q_m$ to an orthonormal basis $Q_1, \ldots, Q_{\dim(\mathbf{S}^n)}$ for all of $\mathbf{S}^n$. For a fixed value of S, there exists $y \in \mathbf{R}^m$ such that $\sum_{i=1}^m y_i A_i + S = C$ if and only if $C - S \in \mathbf{span}(A_1, \ldots, A_m)$. An equivalent condition in terms of the orthonormal basis defined above is

$$Q_i \bullet (C - S) = 0, \quad i = m+1, \ldots, \dim(\mathbf{S}^n).$$

Combining these observations, we can eliminate y from (SDD), giving the following problem:

$$\begin{array}{ll} \text{maximize} & (\sum_{i=1}^m \beta_i A_i) \bullet C - (\sum_{i=1}^m \beta_i A_i) \bullet S \\ \text{subject to} & Q_i \bullet (C - S) = 0, \quad i = m+1, \ldots, \dim(\mathbf{S}^n) \\ & S \succeq 0. \end{array}$$

We can convert this problem into a standard-form primal SDP by negating the objective to obtain a minimization problem, ignoring the additive constant in the objective, and rearranging the equality constraints:

$$\begin{array}{ll} \text{minimize} & (\sum_{i=1}^m \beta_i A_i) \bullet S \\ \text{subject to} & Q_i \bullet S = Q_i \bullet C, \quad i = m+1, \ldots, \dim(\mathbf{S}^n) \\ & S \succeq 0. \end{array}$$

	n				
m	1	2	3	4	5
1	0	1	2	3	4
2		1	2	3	4
3		0	2	3	4
4			1	3	4
5			1	2	4

Table 2.2: upper bounds on the minimum rank of an optimal dual slack variable

This problem is an instance of (SDP) with $\dim(\mathbf{S}^n) - m$ equality constraints; moreover, this problem is solvable because we assume that (SDD) is solvable. Therefore, we can apply Theorem 2.1 to conclude that there is a solution S with $s = \mathbf{rank}(S)$ such that

$$\frac{s(s+1)}{2} \leq \dim(\mathbf{S}^n) - m = \frac{n(n+1)}{2} - m.$$

□

Another way of stating the result of Corollary 2.2 is that there is a solution S of (SDD) such that

$$\mathbf{rank}(S) \leq \left\lfloor \frac{\sqrt{4n(n+1) - 8m + 1} - 1}{2} \right\rfloor.$$

(Note that the assumption that $A_1, \ldots, A_m$ are linearly independent implies that $m \leq \dim(\mathbf{S}^n) = n(n+1)/2$, so the quantity in the square root is always strictly greater than 1.) The values of this bound for small values of m and n are given in Table 2.2. (Values of m and n for which $m > \dim(\mathbf{S}^n) = n(n+1)/2$ are left blank because the bound does not apply.) There are a couple of particularly interesting features of this table.

- If $m = \dim(\mathbf{S}^n) = n(n+1)/2$, then $S = 0$ is a solution. Note that if $S = 0$ is feasible for a primal-dual pair with a feasible primal, then $S = 0$ is optimal because it satisfies the complementary slackness condition $XS = 0$ for every matrix X. If

$m = \dim(\mathbf{S}^n)$, then our linear-independence assumption implies that $A_1, \dots, A_m$ span $\mathbf{S}^n$; thus, for every matrix $C \in \mathbf{S}^n$, there exist scalars $y_1, \dots, y_m$ such that

$$\sum_{i=1}^{m} y_i A_i = C.$$

This implies that $S = 0$ is feasible.

- If $m = n(n+1)/2 - 1$ or $m = n(n+1)/2 - 2$, then (SDD) has a solution (y, S) with $\mathbf{rank}(S) = 1$.

Further rank reduction for feasibility problems. It is also worth noting that, with some additional mild hypotheses, there is a bound for SDP feasibility problems that is sometimes slightly stronger than the bound in Theorem 2.1. This bound (which we state without proof) was first given by Barvinok [4], who provided a nonconstructive proof; an algorithm for finding a solution satisfying the bound was supplied by Ai, Huang, and Zhang [1].

Theorem 2.3. Consider the set

$$\mathcal{F} = \{X \in \mathbf{S}^n \mid A_i \bullet X = b_i,\ i = 1, \dots, m\},$$

where $A_1, \dots, A_m \in \mathbf{S}^n$ and $b \in \mathbf{R}^m$ are given. If $\mathcal{F}$ is nonempty and bounded, and $m = (r+1)(r+2)/2$ for some positive integer $r \leq n-2$, then there exists $X \in \mathcal{F}$ such that $\mathbf{rank}(X) \leq r$.

For example, Theorem 2.1 tells us that every solvable SDP with $m = 3$ equality constraints has a solution with rank at most 2, while Theorem 2.3 tells us that every bounded and feasible SDP feasibility problem with $m = 3$ equality constraints and variable size $n \geq 5$ has a solution with rank at most 1.

2.3 Rank and uniqueness

In this section we present a theorem relating rank and uniqueness for semidefinite programs. This result was given by Zhu [105], and generalizes the classical result in Theorem B.2, which relates cardinality and uniqueness for linear programs. Our discussion of the theorem for semidefinite programs is intentionally similar to the development of the corresponding result for linear programs.

Theorem 2.4. Let $X = VV^\mathsf{T}$ be a solution of (SDP), where $V \in \mathbf{R}^{n\times r}$ and $r = \mathbf{rank}(X)$. This solution is unique if and only if

(i) X has the maximum rank among all solutions, and

(ii) $\mathbf{null}(\mathcal{A}_V) = \{0\}$,

where we define $\mathcal{A}_V : \mathbf{S}^r \to \mathbf{R}^m$ such that $\mathcal{A}_V(Z) = \mathcal{A}(VZV^\mathsf{T})$.

Proof. First, suppose X is the unique solution of (SDP). It is trivially true that X has the maximum rank among all solutions because it is the only solution. In order to show that $\mathbf{null}(\mathcal{A}_V) = \{0\}$, we will argue by contradiction. Suppose there exists a nonzero $\Delta \in \mathbf{S}^r$ such that $\mathcal{A}_V(\Delta) = 0$. Then Algorithm 2.1 finds a solution $\tilde{X}$ of (SDP) whose rank is strictly less than that of X. This contradicts the assumption that X is the unique solution of (SDP), and thereby proves that $\mathbf{null}(\mathcal{A}_V) = \{0\}$.

Conversely, suppose that X and $\tilde{X}$ are distinct solutions of (SDP). We can assume without loss of generality that X has the maximum rank among all solutions of (SDP). First, observe that $(1/2)(X + \tilde{X})$ is a solution of (SDP) with

$$\begin{aligned}\mathbf{range}((1/2)(X+\tilde{X})) &= \mathbf{range}((1/2)X) + \mathbf{range}((1/2)\tilde{X})\\ &= \mathbf{range}(X) + \mathbf{range}(\tilde{X}),\end{aligned}$$

where we have used Lemma A.7 and the fact that nonzero scaling does not change the range of a matrix. Since X is assumed to be a maximum-rank solution, we must have that $\mathbf{range}(\tilde{X}) \subset \mathbf{range}(X)$ because otherwise $\mathbf{range}(X)$ would be a strict subset of $\mathbf{range}((1/2)(X+\tilde{X}))$, and the rank of $(1/2)(X+\tilde{X})$ would be strictly greater than that of X, contradicting our assumption that X is a maximum-rank solution. Taking orthogonal complements, and noting that X and $\tilde{X}$ are symmetric, we find that

$$\mathbf{range}(X)^\perp = \mathbf{null}(X) \subset \mathbf{range}(\tilde{X})^\perp = \mathbf{null}(\tilde{X}).$$

Let $X = VV^\mathsf{T}$, where $V \in \mathbf{R}^{n\times r}$ and $r = \mathbf{rank}(X)$. Having shown that $\mathbf{null}(X) \subset \mathbf{null}(\tilde{X})$, we can use Proposition A.1 to conclude that there exists a matrix $Q \in \mathbf{S}^r$ such that $\tilde{X} = VQV^\mathsf{T}$. Then we have that

$$\mathcal{A}_V(I - Q) = \mathcal{A}(V(I - Q)V^\mathsf{T}) = \mathcal{A}(X) - \mathcal{A}(\tilde{X}) = b - b = 0.$$

Additionally, $I - Q$ is nonzero because $X = VV^\mathsf{T}$ and $\tilde{X} = VQV^\mathsf{T}$ are distinct. Thus, $I - Q$ is a nonzero matrix in $\mathbf{null}(\mathcal{A}_V)$. □

Corollary 2.5. If (SDP) is solvable, and every solution has the same rank, then (SDP) has a unique solution.

Proof. Let $X = VV^\mathsf{T}$ be a solution of (SDP), where $V \in \mathbf{R}^{n\times r}$ and $r = \mathbf{rank}(X)$. Since every solution of (SDP) has the same rank, X must have the maximum rank among all solutions. Another consequence of the fact that every solution has the same rank is that Algorithm 2.1 must terminate on the first iteration: that is, $\mathbf{null}(\mathcal{A}_V) = \{0\}$. Thus, Theorem 2.4 tells us that X is the unique solution of (SDP). □

2.3.1 An example of sensor-network localization

Suppose $x_{\text{true}} \in \mathbf{R}^d$ is the unknown location of a sensor, and $a_1, \ldots, a_m \in \mathbf{R}^d$ are the known locations of points called anchors. We are given the distances between the sensor and the anchors:

$$d_i = \|x_{\text{true}} - a_i\|, \quad i = 1, \ldots, m.$$

We can attempt to determine x_{true} from the distance measurements by solving the optimization problem

$$\begin{array}{ll} \text{minimize} & 0 \\ \text{subject to} & \|x - a_i\|^2 = d_i^2, \quad i = 1, \ldots, m \end{array}$$

with variable $x \in \mathbf{R}^d$. The primal SDP relaxation of this problem is

$$\begin{array}{ll} \text{minimize} & 0 \bullet X \\ \text{subject to} & \begin{bmatrix} I & -a_i \\ -a_i^\mathsf{T} & 0 \end{bmatrix} \bullet X = d_i^2 - \|a_i\|^2, \quad i = 1, \ldots, m \\ & \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \bullet X = 1 \\ & X \succeq 0 \end{array}$$

with variable $X \in \mathbf{S}^{d+1}$, and the corresponding dual SDP relaxation is

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^m (d_i^2 - \|a_i\|^2) y_i + z \\ \text{subject to} & \sum_{i=1}^m y_i \begin{bmatrix} I & -a_i \\ -a_i^\mathsf{T} & 0 \end{bmatrix} + z \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} + S = 0 \\ & S \succeq 0 \end{array} \tag{2.6}$$

with variables $y \in \mathbf{R}^m$, $z \in \mathbf{R}$, and $S \in \mathbf{S}^{d+1}$. (See Section A.2.4 for a development of the natural SDP relaxation of a polynomial optimization problem.) We can check that

$$X_0 = \begin{bmatrix} x_{\text{true}} \\ 1 \end{bmatrix} \begin{bmatrix} x_{\text{true}} \\ 1 \end{bmatrix}^\mathsf{T}$$

is a solution of the primal SDP relaxation. However, if the solution of the primal SDP relaxation is not unique, then we cannot necessarily solve the sensor-network-localization problem using the SDP relaxation. The following theorem gives conditions under which the primal SDP relaxation has a unique solution.

Proposition 2.3. If $a_1, \ldots, a_m$ are affinely independent (that is, not contained in a hyperplane), then the primal SDP relaxation of the sensor-network-localization problem has a unique solution.

Proof. We have that $a_1, \ldots, a_m$ are affinely dependent if and only if there exist a scalar β and a nonzero vector $\eta \in \mathbf{R}^d$ such that

$$\eta^\mathsf{T} a_i + \beta = 0, \quad i = 1, \ldots, m.$$

(The vector η is a normal vector of the hyperplane containing $a_1, \ldots, a_m$, and β determines the offset of the hyperplane from the origin.) Collecting these conditions into a matrix equation gives

$$\begin{bmatrix} a_1^\mathsf{T} & 1 \\ \vdots & \vdots \\ a_m^\mathsf{T} & 1 \end{bmatrix} \begin{bmatrix} \eta \\ \beta \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Thus, we have that $a_1, \ldots, a_m$ are affinely independent if and only if

$$\begin{bmatrix} a_1^\mathsf{T} & 1 \\ \vdots & \vdots \\ a_m^\mathsf{T} & 1 \end{bmatrix}$$

is skinny and full rank, or, equivalently,

$$\begin{bmatrix} a_1 & \cdots & a_m \\ 1 & \cdots & 1 \end{bmatrix}$$

is fat and full rank. Therefore, if $a_1, \ldots, a_m$ are affinely independent, then we can find a vector $y \in \mathbf{R}^m$ such that

$$\begin{bmatrix} a_1 & \cdots & a_m \\ 1 & \cdots & 1 \end{bmatrix} y = \begin{bmatrix} \sum_{i=1}^m y_i a_i \\ \sum_{i=1}^m y_i \end{bmatrix} = - \begin{bmatrix} x_{\text{true}} \\ 1 \end{bmatrix}.$$

Define $z = -\|x_{\text{true}}\|^2$, and

$$\begin{aligned} S &= \sum_{i=1}^m y_i \begin{bmatrix} -I & a_i \\ a_i^{\mathsf{T}} & 0 \end{bmatrix} - z \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} -\left(\sum_{i=1}^m y_i\right) I & \sum_{i=1}^m y_i a_i \\ \left(\sum_{i=1}^m y_i a_i\right)^{\mathsf{T}} & -z \end{bmatrix} \\ &= \begin{bmatrix} I & -x_{\text{true}} \\ -x_{\text{true}}^{\mathsf{T}} & \|x_{\text{true}}\|^2 \end{bmatrix} \\ &= \begin{bmatrix} I & -x_{\text{true}} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} I & -x_{\text{true}} \end{bmatrix}. \end{aligned}$$

We have that (y, z, S) is a solution of (2.6) because it is feasible by construction, and satisfies the complementarity condition

$$\begin{aligned} SX_0 &= \left(\begin{bmatrix} I & -x_{\text{true}} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} I & -x_{\text{true}} \end{bmatrix} \right) \left(\begin{bmatrix} x_{\text{true}} \\ 1 \end{bmatrix} \begin{bmatrix} x_{\text{true}} \\ 1 \end{bmatrix}^{\mathsf{T}} \right) \\ &= \begin{bmatrix} I & -x_{\text{true}} \end{bmatrix}^{\mathsf{T}} \left(\begin{bmatrix} I & -x_{\text{true}} \end{bmatrix} \begin{bmatrix} x_{\text{true}} \\ 1 \end{bmatrix} \right) \begin{bmatrix} x_{\text{true}} \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} I & -x_{\text{true}} \end{bmatrix}^{\mathsf{T}} \left(x_{\text{true}} - x_{\text{true}} \right) \begin{bmatrix} x_{\text{true}} \\ 1 \end{bmatrix} \\ &= 0. \end{aligned}$$

The last expression given for S implies that $\mathbf{rank}(S) = d$. Let X be a solution of the primal SDP relaxation. Then we can use the complementarity condition $S \bullet X = 0$ and Lemma A.5 to conclude that $\mathbf{rank}(X) \leq 1$. Moreover, the constraint

$$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \bullet X = 1$$

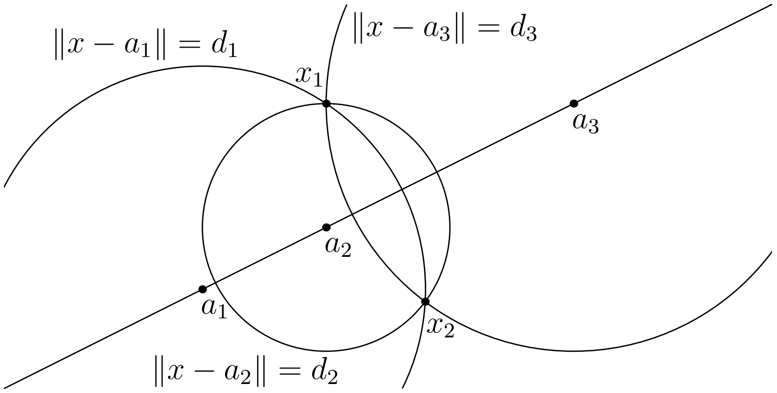


Figure 2.1: sensor-network localization fails with affinely dependent anchors

guarantees that X is nonzero, and hence that $\mathbf{rank}(X) = 1$. Since the rank of a solution of the primal SDP relaxation must be unique, we can use Corollary 2.5 to conclude that the primal SDP relaxation has a unique solution. □

The assumption that $a_1, \dots, a_m$ are affinely independent is reasonable because if this condition is violated, then we cannot uniquely identify x based on the distance measurements $\|x - a_i\|$ for $i = 1, \dots, m$. The geometry of a simple example with $m = 3$ is shown in Figure 2.1. Note that the points x_1 and x_2 cannot be distinguished on the basis of the measurements $\|x - a_i\|$ for $i = 1, \dots, m$.

2.4 Rank and sparsity

Consider an SDP of the form

$$\begin{array}{ll} \text{minimize} & A_0 \bullet X \\ \text{subject to} & A_k \bullet X \leq b_k, \quad k = 1, \dots, m \\ & X \succeq 0, \end{array} \tag{2.7}$$

where $X \in \mathbf{S}^n$ is the optimization variable, and $A_0, \dots, A_m \in \mathbf{S}^n$ and $b \in \mathbf{R}^m$ are problem data. In many applications the coefficients $A_0, \dots, A_m$ are sparse, not only in the sense that each A_k has only a few nonzero entries, but also in the much stronger sense that

$(A_0)_{ij} = \cdots = (A_m)_{ij} = 0$ for most i and j. We can encode the sparsity pattern of $A_0, \ldots, A_m$ using a graph $G = (V, E)$ with vertex set $V = \{1, \ldots, n\}$ and edge set

$$E = \{(i,j) \mid (A_k)_{ij} \neq 0 \text{ for some } k\}.$$

If the coefficients are sparse in the strong sense described above, then the corresponding graph will be sparse. We will show that if the graph possesses certain properties, then the associated instance of (2.7) is guaranteed to have a low-rank solution, and we can efficiently construct such a solution using the graph. Many results of this type were developed in the context of power networks [61, 62, 63, 88, 86, 103]. However, related theorems have also been stated in terms of general quadratic optimization [9, 59], and applied to distributed control problems [54].

We will present a simple example of a theorem connecting rank and sparsity that was given by Sojoudi and Lavaei [87]. Before stating this result, we need to introduce some additional terminology. Define the sign of an edge $(i,j) \in E$ to be

$$\sigma(i,j) = \begin{cases} 1 & (A_0)_{ij}, \ldots, (A_m)_{ij} \geq 0, \\ -1 & (A_0)_{ij}, \ldots, (A_m)_{ij} \leq 0, \\ 0 & \text{otherwise.} \end{cases}$$

We say that the edge (i,j) is sign definite if $\sigma(i,j) \neq 0$, positive if $\sigma(i,j) = 1$, and negative if $\sigma(i,j) = -1$.

The construction used in the following theorem is illustrated for a simple example in Example 2.4. Since concrete examples are often easier to understand than abstract proofs, we encourage the reader to work through the proof and the example in parallel.

Theorem 2.6 (Sojoudi and Lavaei [87]). Consider a solvable instance of (2.7) with associated graph $G = (V, E)$. If

(1) every edge of G is sign definite, and

(2) every cycle of G has an even number of positive edges,

then (2.7) has a rank-1 solution.

Proof. The first step in the proof is to assign a label $\alpha(k) \in \{\pm 1\}$ to each vertex k in such a way that

$$\sigma(i,j) = -\alpha(i)\alpha(j)$$

for every edge (i,j) in the graph. Let T be a rooted spanning tree for G, r be the root of the tree, and $p(i)$ denote the parent of i in T. Choose the labels

$$\alpha(r) = 1 \quad \text{and} \quad \alpha(i) = -\frac{\sigma(i,p(i))}{\alpha(p(i))}, \quad i \neq r.$$

By construction we have that $\sigma(i,j) = -\alpha(i)\alpha(j)$ for every edge (i,j) in T. We will show that our assumptions about the structure of G imply that this property actually holds for all edges in G, not just the ones contained in the spanning tree T. Consider an edge (i,j) that is in G but not in T. Recall that adding an edge to a tree creates a cycle. Thus, adding (i,j) to T yields a cycle:

$$v_1 = j, v_2, \ldots, v_{\ell-1}, v_\ell = i, v_{\ell+1} = j,$$

where (v_k, v_{k+1}) is in T for $k = 1, \ldots, \ell - 1$. Because every cycle in G has an even number of positive edges, we have that

$$\prod_{k=1}^{\ell} \sigma(v_k, v_{k+1}) = (-1)^{\ell}.$$

(If ℓ is even, then the cycle contains an even number of negative edges, so the product of the signs of all of the edges in the cycle is equal to 1; if ℓ is odd, then the cycle contains an odd number of negative edges, so the product of the signs of all of the edges in the cycle is equal to -1.) Since (v_k, v_{k+1}) is in T for $k = 1, \ldots, \ell - 1$, we have that $\sigma(v_k, v_{k+1}) = -\alpha(v_k)\alpha(v_{k+1})$ for $k = 1, \ldots, \ell - 1$, and hence that

$$\begin{aligned}
\prod_{k=1}^{\ell} \sigma(v_k, v_{k+1}) &= \left(\prod_{k=1}^{\ell-1} \sigma(v_k, v_{k+1}) \right) \sigma(v_\ell, v_{\ell+1}) \\
&= \left(\prod_{k=1}^{\ell-1} (-\alpha(v_k)\alpha(v_{k+1})) \right) \sigma(v_\ell, v_{\ell+1}) \\
&= (-1)^{\ell-1} \alpha(v_1) \left(\prod_{k=2}^{\ell-1} \alpha(v_k)^2 \right) \alpha(v_\ell)\sigma(v_\ell, v_{\ell+1}).
\end{aligned}$$

We can simplify this expression by observing that $\alpha(v_k)^2 = 1$ since $\alpha(v_k) \in \{\pm 1\}$ for $k = 2, \ldots, \ell - 1$. Also, we note that $v_1 = v_{\ell+1} = j$ and $v_\ell = i$. Combining these results, we find that

$$\prod_{k=1}^{\ell} \sigma(v_k, v_{k+1}) = (-1)^{\ell-1} \alpha(i)\alpha(j)\sigma(i,j) = (-1)^{\ell}.$$

Solving for $\sigma(i,j)$ gives

$$\sigma(i,j) = -\frac{1}{\alpha(i)\alpha(j)} = -\alpha(i)\alpha(j),$$

where $1/\alpha(k) = \alpha(k)$ because $\alpha(k) \in \{\pm 1\}$.

Now we use the labels $\alpha(1), \ldots, \alpha(n)$ to construct a rank-1 solution of the SDP. Let X be a solution of (2.7), and define

$$\tilde{X} = \begin{bmatrix} \alpha(1)\sqrt{X_{11}} \\ \vdots \\ \alpha(n)\sqrt{X_{nn}} \end{bmatrix} \begin{bmatrix} \alpha(1)\sqrt{X_{11}} \\ \vdots \\ \alpha(n)\sqrt{X_{nn}} \end{bmatrix}^{\mathsf{T}}.$$

Note that $X \succeq 0$ because X is a solution of (2.7); therefore, $X_{ii} \geq 0$ for $i = 1, \ldots, n$, so the square roots in our definition of $\tilde{X}$ are real numbers. Using our definition of $\tilde{X}$, and the fact that $\alpha(i)\alpha(j) = -\sigma(i,j)$, we have that

$$\begin{aligned} (A_k)_{ij}\tilde{X}_{ij} &= (A_k)_{ij}\alpha(i)\alpha(j)\sqrt{X_{ii}X_{jj}} \\ &= -\sigma(i,j)(A_k)_{ij}\sqrt{X_{ii}X_{jj}} \end{aligned}$$

for all $(i,j) \in E$. The definition of the sign of an edge implies that $\sigma(i,j)(A_k)_{ij} = |(A_k)_{ij}|$, and hence that

$$(A_k)_{ij}\tilde{X}_{ij} = -|(A_k)_{ij}|\sqrt{X_{ii}X_{jj}}.$$

Another consequence of X being positive semidefinite is that

$$\det\left(\begin{bmatrix} X_{ii} & X_{ij} \\ X_{ij} & X_{jj} \end{bmatrix}\right) = X_{ii}X_{jj} - X_{ij}^2 \geq 0.$$

Rearranging this inequality, we find that

$$-\sqrt{X_{ii}X_{jj}} \leq -|X_{ij}|.$$

This inequality allows us to bound $(A_k)_{ij}\tilde{X}_{ij}$:

$$\begin{aligned}(A_k)_{ij}\tilde{X}_{ij} &= -|(A_k)_{ij}|\sqrt{X_{ii}X_{jj}}\\ &\leq -|(A_k)_{ij}||X_{ij}|\\ &= -|(A_k)_{ij}X_{ij}|\\ &\leq (A_k)_{ij}X_{ij}.\end{aligned}$$

Summing both sides of this inequality over i and j gives

$$A_k \bullet \tilde{X} = \sum_{i=1}^{n}\sum_{j=1}^{n}(A_k)_{ij}\tilde{X}_{ij} \leq \sum_{i=1}^{n}\sum_{j=1}^{n}(A_k)_{ij}X_{ij} = A_k \bullet X.$$

(Although our earlier analysis only held for $(i,j) \in E$, we can still sum over i and j because $(A_k)_{ij} = 0$ if $(i,j) \notin E$.) With $k = 1,\dots,m$, this inequality tells us that $\tilde{X}$ satisfies the inequality constraints of (2.7):

$$A_k \bullet \tilde{X} \leq A_k \bullet X \leq b_k, \quad k = 1,\dots,m.$$

With $k = 0$, this inequality tells us that $\tilde{X}$ is optimal for (2.7) because X is optimal, and $A_0 \bullet \tilde{X} \leq A_0 \bullet X$. Thus, we can conclude that $\tilde{X}$ is a rank-1 solution of (2.7). □

Example 2.4. Consider the following instance of (2.7):

$$\begin{array}{ll}\text{minimize} & X_{11}\\ \text{subject to} & X_{ii} \leq 1, \quad i = 1,2,3,4\\ & -X_{ii} \leq -1, \quad i = 1,2,3,4\\ & X_{12} + X_{13} - X_{23} + X_{34} \leq 2\\ & X \succeq 0.\end{array}$$

A rank-4 solution of this problem is given by $X = I$. The graph G representing the sparsity pattern of the coefficients is given in Figure 2.2. All of the edges are sign definite, and are labeled with the corresponding signs. Note that G satisfies the conditions of Theorem 2.6: we have already remarked that all of the edges are sign definite; there is only one cycle, and it has two positive edges. Thus, this problem must have a rank-1 solution. We can construct such a solution by applying the procedure given in the proof of Theorem 2.6. We will use the rooted

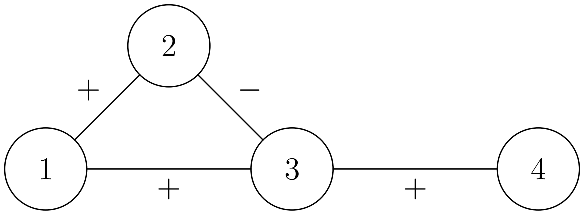


Figure 2.2: the graph G representing the sparsity pattern of Example 2.4

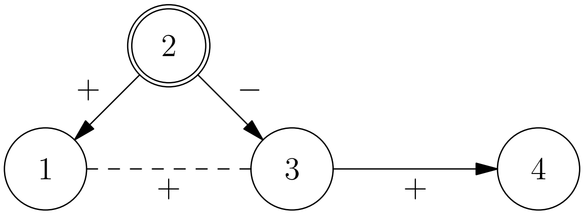


Figure 2.3: a rooted spanning tree for the graph G in Figure 2.2

spanning tree T shown in Figure 2.3. First, we assign the label $\alpha(2) = 1$ to the root node. Then we assign labels to the children of the root:

$$\alpha(1) = -\frac{\sigma(1,2)}{\alpha(2)} = -1 \quad \text{and} \quad \alpha(3) = -\frac{\sigma(2,3)}{\alpha(2)} = 1.$$

Next, we assign a label to the node at depth 2:

$$\alpha(4) = -\frac{\sigma(3,4)}{\alpha(3)} = -1.$$

Note that the edge $(1,3)$ is in G, but not in T; however, since G satisfies the hypotheses of Theorem 2.6, we still have that

$$\alpha(1)\alpha(3) = -\sigma(1,3) = -1.$$

Using our vertex labels and the initial solution $X = I$, we construct the rank-1 solution

$$\tilde{X} = \begin{bmatrix} \alpha(1)\sqrt{X_{11}} \\ \alpha(2)\sqrt{X_{22}} \\ \alpha(3)\sqrt{X_{33}} \\ \alpha(4)\sqrt{X_{44}} \end{bmatrix} \begin{bmatrix} \alpha(1)\sqrt{X_{11}} \\ \alpha(2)\sqrt{X_{22}} \\ \alpha(3)\sqrt{X_{33}} \\ \alpha(4)\sqrt{X_{44}} \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} 1 & -1 & -1 & 1 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}.$$

Finally, we list two important cases where the hypotheses of Theorem 2.6 are easy to check.

Corollary 2.7. Consider a solvable instance of (2.7) with associated graph G.

(1) If G is acyclic, and every edge of G is sign definite, then (2.7) has a rank-1 solution.

(2) If every edge of G is negative, then (2.7) has a rank-1 solution.

Bose et al. [9] gave a result similar to the first part of Corollary 2.7; the second part of the corollary was shown by Kim and Kojima [59].

3

Heuristics and Approximate Solutions

3.1 Introduction

This chapter is about heuristics and approximate methods for low-rank semidefinite programming. First, we describe the nonlinear-programming method of Burer and Monteiro [15, 16, 17, 18]. This is a popular heuristic that often works well in practice, particularly when we can guarantee the existence of a low-rank solution using, for example, one of the theorems from Chapter 2. Next, we discuss the nuclear-norm heuristic. One of the foundations of compressed sensing is the fact that the ℓ_1-heuristic often finds a minimum-cardinality solution of a system of linear equations [2, 20, 21, 22, 23, 30]. Similarly, the nuclear-norm heuristic often recovers a minimum-rank solution of an SDP feasibility problem [80]. We do not prove guarantees about the nuclear-norm heuristic, focusing instead on showing how to minimize the nuclear norm by solving a semidefinite program. The chapter concludes with a presentation of methods for rounding exact solutions to low-rank approximate solutions. These techniques are widely used in approximation algorithms, including the famous Goemans-Williamson algorithm for the maximum-cut problem.

3.2 Nonlinear-programming algorithms

Interior-point methods are often impractical for large-scale semidefinite programs. This has prompted the development of first-order algorithms such as the spectral-bundle algorithm of Helmberg and Rendl [48]; a survey of such first- and second-order algorithms is given by Monteiro [71]. Large-scale SDPs frequently arise as relaxations of combinatorial optimization problems, such as the maximum-cut problem (see Section 3.4.3). Extending an algorithm for solving relaxations of maximum-cut problems [14], Burer and Monteiro [15, 16, 17, 18] proposed a nonlinear-programming algorithm for low-rank semidefinite programming, and demonstrated that it is often effective in practice. This algorithm is based on the fact that there is a one-to-one correspondence between the set of $n \times n$ positive-semidefinite matrices with rank at most r, and the set of matrices that can be written in the form RR^T for some $R \in \mathbf{R}^{n\times r}$. Having made this observation, we consider the optimization problem

$$\begin{array}{ll} \text{minimize} & C \bullet (RR^\mathsf{T}) \\ \text{subject to} & A_i \bullet (RR^\mathsf{T}) = b_i, \quad i = 1, \ldots, m, \end{array} \tag{SDP-r}$$

with variable $R \in \mathbf{R}^{n\times r}$, where the positive integer r is a problem parameter. If r is chosen to be at least as large as the rank of a minimum-rank solution of (SDP), then (SDP-r) is equivalent to (SDP). (For example, we can use the bounds in Chapter 2 to choose r.) If r is less than the rank of a minimum-rank solution of (SDP), then we can think of (SDP-r) as giving us a low-rank approximate solution of (SDP).

The constraint $X \succeq 0$ is the source of the difficulty in solving (SDP) since the objective and equality constraints are both linear. Thus, (SDP-r) has the advantage of eliminating the difficult matrix inequality; however, this comes at the expense of turning the linear objective and constraint functions into (possibly nonconvex) quadratic functions. Nonetheless, making this trade-off gives rise to an algorithm that often works well in practice, at least when r is small, and there exists a solution with rank less than or equal to r.

We can attempt to solve (SDP-r) using standard nonlinear-programming algorithms. Burer and Monteiro suggest using an augmented-Lagrangian method with a limited-memory Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm to solve the unconstrained subproblems. Augmented-Lagrangian methods were introduced by Hestenes [49] and Powell [79]; the BFGS algorithm was independently proposed by Broyden [11, 12], Fletcher [34], Goldfarb [40], and Shanno [82]. We will not review augmented-Lagrangian methods or the BFGS algorithm because both are covered very well in several texts on nonlinear programming and numerical optimization [7, 35, 65].

It is also worth mentioning that other nonlinear-programming algorithms have been suggested for solving (SDP-r) (for example, see [53]). However, the approach suggested by Burer and Monteiro is widely used in the literature on applications of low-rank semidefinite programming.

3.3 The nuclear-norm heuristic

Suppose we want to find a minimum-rank solution of the SDP feasibility problem

$$\begin{aligned} A_i \bullet X &= b_i, \quad i = 1, \dots, m \\ X &\succeq 0, \end{aligned}$$

where $X \in \mathbf{S}^n$ is the variable, and $A_1, \dots, A_m \in \mathbf{S}^n$ and $b \in \mathbf{R}^m$ are problem data. Although we focus on feasibility problems for simplicity, we can readily extend our results to optimization problems. First, we compute a solution $X^\star$ of (SDP) using an interior-point method. Then we find a minimum-rank solution of the feasibility problem

$$\begin{aligned} C \bullet X &= C \bullet X^\star \\ A_i \bullet X &= b_i, \qquad i = 1, \dots, m \\ X &\succeq 0. \end{aligned}$$

A common technique for finding low-rank solutions of SDP feasibility problems involves solving an optimization problem with a specially chosen objective. For example, Barvinok's proof of the rank bound in Theorem 2.1 was based on considering an instance of (SDP) with a generic value of C. Minimizing the "trace objective" $\mathbf{tr}(X)$ (that is, the special case when $C = I$) is common in the control community [69, 75].

The following example shows the effect of different objective functions on a specific problem.

Example 3.1. Suppose we want to find the smallest possible dimension d and corresponding points $x_1, x_2, x_3 \in \mathbf{R}^d$ satisfying the distance constraints

$$\|x_1\| = 1, \quad \|x_1 - x_2\| = 1, \quad \text{and} \quad \|x_2 - x_3\| = 1. \tag{3.1}$$

Consider the SDP feasibility problem

$$\begin{aligned} A_i \bullet X &= 1, \quad i = 1, 2, 3 \\ X &\succeq 0, \end{aligned} \tag{3.2}$$

where we define

$$A_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \; A_2 = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \text{ and } A_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix}.$$

Given points $x_1, x_2, x_3 \in \mathbf{R}^d$ satisfying (3.1), the matrix

$$X = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^\mathsf{T} \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}$$

is a rank-d solution of (3.2). Conversely, given a rank-d solution X of (3.2), we can use the factorization above to find $x_1, x_2, x_3 \in \mathbf{R}^d$ satisfying (3.1). Thus, our original problem is equivalent to finding a minimum-rank solution of (3.2).

Because (3.2) has $m = 3$ equality constraints, Theorem 2.1 guarantees that there exists a solution whose rank is at most 2. This is convenient because it allows us to draw pictures in the plane. We can think of our problem as choosing the configuration of a mechanical linkage, as shown in Figure 3.1. The linkage has a fixed pivot at the origin, and floating pivots at locations x_1 and x_2; the end of the linkage is x_3. Since the origin is fixed, and the length of the segment between the origin and x_1 is fixed at $\|x_1\| = 1$, x_1 must lie on the unit circle centered at the origin. Similarly, x_2 must lie on the unit circle centered at x_1, and x_3 must lie on the unit circle centered at x_2. In Figure 3.1, these circles are shown as dashed lines, while the thick solid lines represent the segments of the linkage.

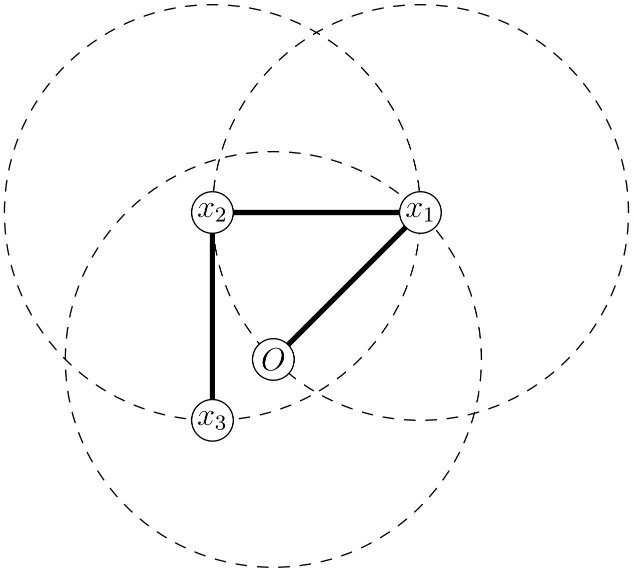


Figure 3.1: representation of the feasibility problem as a mechanical linkage

Suppose we apply the trace objective to this problem. This corresponds to configuring the linkage in order to minimize

$$\mathbf{tr}(X) = \|x_1\|^2 + \|x_2\|^2 + \|x_3\|^2 = 1 + \|x_2\|^2 + \|x_3\|^2.$$

Thus, we want x_2 and x_3 to be as close to the origin as possible. In terms of the intuition provided by the mechanical linkage, we can imagine attaching elastic bands that stretch from the origin to the floating pivot at x_2 and the end of the linkage at x_3. Because x_2 and x_3 are equally weighted in the objective function, the strengths of the corresponding elastic bands are equal as well. The solution of the SDP using the trace objective is shown in Figure 3.2. Observe that the origin is the midpoint of the line segment joining x_2 and x_3. In terms of our mechanical-linkage analogy, this reflects the fact that the forces due to the elastic bands joining the origin to x_2 and x_3 must be balanced.

The standard trace objective gives us a rank-2 solution of the feasibility problem, which can be used to construct a set of points in $\mathbf{R}^2$ satisfying the given conditions. Returning to the mechanical-linkage analogy, suppose we pull x_3 as far from the origin as possible. Then

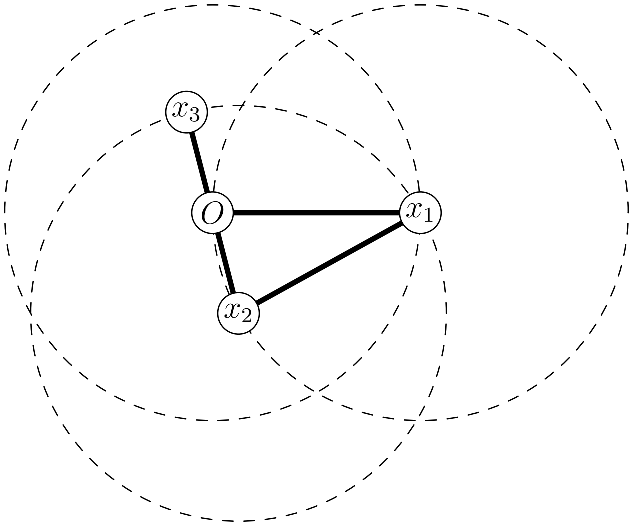


Figure 3.2: a solution of the SDP with $C = I$

the linkage will be configured in a straight line as shown in Figure 3.3, allowing us to find a set of points in $\mathbf{R}$ satisfying the given conditions. Based on this intuition, we compute the solution of the feasibility problem that minimizes the objective

$$(-e_3 e_3^\mathsf{T}) \bullet X = -\|x_3\|^2.$$

(This is equivalent to finding the solution that maximizes $\|x_3\|^2$, which corresponds to our intuition of pulling x_3 as far from the origin as possible.)

Another objective function commonly used in the literature is the log-det heuristic: $\log(\det(X+\delta I))$, where $\delta > 0$ is a small regularization term. This approach was proposed by Fazel, Hindi, and Boyd [33].

One limitation of the trace and log-det heuristics is that they can only be applied to square matrices. Thus, we cannot apply these heuristics to a system of linear matrix equations:

$$A_i \bullet X = b_i, \quad i = 1, \ldots, p,$$

where $X \in \mathbf{R}^{m\times n}$ is the variable, and $A_1, \ldots, A_p \in \mathbf{R}^{m\times n}$ and $b \in \mathbf{R}^m$ are problem data. For a problem of this type, Fazel [31, 32] suggested

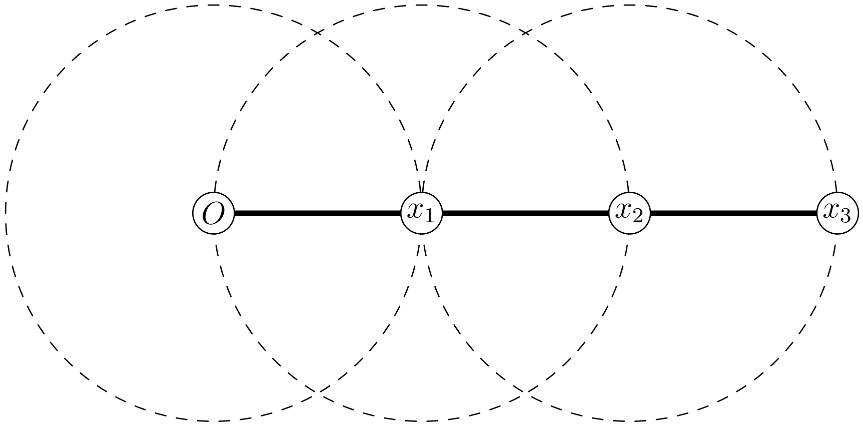


Figure 3.3: a solution of the SDP with $C = -e_3 e_3^\mathsf{T}$

minimizing the nuclear norm subject to the constraints $A_i \bullet X = b_i$ for $i = 1, \ldots, p$. Recall that the nuclear norm of a matrix $X \in \mathbf{R}^{m \times n}$ is defined to be

$$\|X\|_* = \sum_{i=1}^{\min(m,n)} \sigma_i,$$

where σ_i denotes the ith singular value of X. In the special case when X is symmetric and positive semidefinite, we have that $\|X\|_* = \mathbf{tr}(X)$, so we can think of the nuclear-norm heuristic as a generalization of the trace heuristic.

There are other intuitively appealing justifications for using the nuclear-norm heuristic. Recall that the convex envelope of a function $f : \mathbf{R}^n \to \mathbf{R}$ is defined to be the convex function $g : \mathbf{R}^n \to \mathbf{R}$ such that $h(x) \leq g(x) \leq f(x)$ for all $x \in \mathbf{dom}(f)$, and all convex functions $h : \mathbf{R}^n \to \mathbf{R}$. Thus, we can think of g as the best convex approximation of f. It is possible to show that the nuclear norm is the convex envelope of the rank function (see [31]). Therefore, the nuclear-norm heuristic provides the best convex approximation of the problem of minimizing the rank subject to affine constraints.

We can also think of the nuclear-norm heuristic as the matrix analog of the ℓ_1-heuristic because

$$\|\mathbf{diag}(x)\|_* = \sum_{i=1}^{n} |x_i| = \|x\|_1$$

for every vector $x \in \mathbf{R}^n$. It has been shown that the ℓ_1-heuristic yields a minimum-cardinality solution in many cases [2, 20, 21, 22, 23, 30]. Moreover, similar guarantees can often be made for the nuclear-norm heuristic [80]. We will not attempt to prove these guarantees here; our focus will instead be on showing how to minimize the nuclear norm by solving a semidefinite program.

Proposition 3.1. The nuclear norm of $A \in \mathbf{R}^{m\times n}$ is the common optimal value of the semidefinite program

$$\begin{array}{ll} \text{maximize} & A \bullet Y \\ \text{subject to} & \begin{bmatrix} I_m & Y \\ Y^\mathsf{T} & I_n \end{bmatrix} \succeq 0 \end{array}$$

with variable $Y \in \mathbf{R}^{m\times n}$, and its dual

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}(W_1) + \mathbf{tr}(W_2) \\ \text{subject to} & \begin{bmatrix} W_1 & -(1/2)A \\ -(1/2)A^\mathsf{T} & W_2 \end{bmatrix} \succeq 0, \end{array}$$

with variables $W_1 \in \mathbf{S}^m$ and $W_2 \in \mathbf{S}^n$.

Proof. Let $r = \mathbf{rank}(A)$, and $A = U\Sigma V^\mathsf{T}$ be the (reduced) singular-value decomposition of A, where $U \in \mathbf{R}^{m\times r}$ and $V \in \mathbf{R}^{n\times r}$ have orthonormal columns, and $\Sigma \in \mathbf{R}^{r\times r}$ is diagonal and nonsingular. Consider the matrix $Y = UV^\mathsf{T}$. Corollary A.12 tells us that

$$\begin{bmatrix} I_m & Y \\ Y^\mathsf{T} & I_n \end{bmatrix} \succeq 0$$

if and only if $I_n - Y^\mathsf{T}Y \succeq 0$. Let $\tilde{V} \in \mathbf{R}^{n\times(n-r)}$ be the matrix whose columns are the right singular vectors of A corresponding to the zero singular values. Then we have that

$$I_n - Y^\mathsf{T}Y = I_n - (UV^\mathsf{T})^\mathsf{T}(UV^\mathsf{T}) = I_n - VV^\mathsf{T} = \tilde{V}\tilde{V}^\mathsf{T} \succeq 0.$$

This proves that Y is feasible for the primal problem. Moreover, Y achieves an objective value of

$$A \bullet Y = \mathbf{tr}(A^\mathsf{T} Y) = \mathbf{tr}((U\Sigma V^\mathsf{T})^\mathsf{T}(UV^\mathsf{T})) = \mathbf{tr}(\Sigma) = \|A\|_*.$$

Similarly, the matrices $W_1 = (1/2)U\Sigma U^\mathsf{T}$ and $W_2 = (1/2)V\Sigma V^\mathsf{T}$ are feasible for the dual problem because

$$\begin{aligned}\begin{bmatrix} W_1 & -(1/2)A \\ -(1/2)A^\mathsf{T} & W_2 \end{bmatrix} &= \begin{bmatrix} (1/2)U\Sigma U^\mathsf{T} & -(1/2)U\Sigma V^\mathsf{T} \\ -(1/2)V\Sigma U^\mathsf{T} & (1/2)V\Sigma V^\mathsf{T} \end{bmatrix} \\ &= \frac{1}{2}\begin{bmatrix} U\Sigma^{\frac{1}{2}} \\ -V\Sigma^{\frac{1}{2}} \end{bmatrix}\begin{bmatrix} U\Sigma^{\frac{1}{2}} \\ -V\Sigma^{\frac{1}{2}} \end{bmatrix}^\mathsf{T} \\ &\succeq 0.\end{aligned}$$

These matrices achieve an objective value of

$$\begin{aligned}\mathbf{tr}(W_1) + \mathbf{tr}(W_2) &= \mathbf{tr}((1/2)U\Sigma U^\mathsf{T}) + \mathbf{tr}((1/2)V\Sigma V^\mathsf{T}) \\ &= \mathbf{tr}(\Sigma) \\ &= \|A\|_*.\end{aligned}$$

For all feasible matrices Y, W_1, and W_2, we have that

$$\mathbf{tr}(W_1) + \mathbf{tr}(W_2) - A \bullet Y = \begin{bmatrix} W_1 & -(1/2)A \\ -(1/2)A^\mathsf{T} & W_2 \end{bmatrix} \bullet \begin{bmatrix} I_m & Y \\ Y^\mathsf{T} & I_n \end{bmatrix} \geq 0$$

since the trace inner product of two positive semidefinite matrices is nonnegative. Thus, we have that $\mathbf{tr}(W_1)+\mathbf{tr}(W_2) \geq A \bullet Y$ for all feasible matrices Y, W_1, and W_2. For the values of Y, W_1, and W_2 given above, we found that

$$A \bullet Y = \mathbf{tr}(W_1) + \mathbf{tr}(W_2) = \|A\|_*.$$

Therefore, we can conclude that the values of Y, W_1, and W_2 given above are the solutions of the corresponding optimization problems, and that the common optimal value of the two problems is $\|A\|_*$. □

Corollary 3.1. Consider the problem of minimizing the nuclear norm subject to affine constraints:

$$\begin{array}{ll} \text{minimize} & \|X\|_* \\ \text{subject to} & A_i \bullet X = b_i, \quad i = 1, \ldots, p, \end{array}$$

where $X \in \mathbf{R}^{m\times n}$ is the optimization variable, and $A_1, \ldots, A_p \in \mathbf{R}^{m\times n}$ and $b \in \mathbf{R}^p$ are problem data. We can solve this problem by solving the semidefinite program

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}(W_1) + \mathbf{tr}(W_2) \\ \text{subject to} & A_i \bullet X = b_i, \quad i = 1, \ldots, p \\ & \begin{bmatrix} W_1 & -(1/2)X \\ -(1/2)X^\mathsf{T} & W_2 \end{bmatrix} \succeq 0 \end{array}$$

with variables $X \in \mathbf{R}^{m\times n}$, $W_1 \in \mathbf{S}^m$, and $W_2 \in \mathbf{S}^n$.

3.4 Rounding methods

Many popular approximation algorithms for NP-hard problems are based on SDP relaxations of integer-programming problems. In order to recover an approximate solution from the SDP relaxation, we typically need to round the solution of the SDP. The most famous example of such an algorithm is the Goemans-Williamson algorithm for the maximum-cut problem [39]. This section surveys some of the most popular methods for rounding the solutions of SDP problems to low-rank approximate solutions, and proves some guarantees on the quality of the resulting approximations.

3.4.1 Low-rank projection

Consider a matrix $X \in \mathbf{S}^n_+$ with eigenvalue decomposition

$$X = \sum_{i=1}^{n} \lambda_i v_i v_i^\mathsf{T},$$

where $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$ are the eigenvalues of X, and $v_1, \ldots, v_n \in \mathbf{R}^n$ form an orthonormal set of corresponding eigenvectors. Suppose X is the solution of a semidefinite program, and we desire a solution with rank at most r. Perhaps the most natural rounding method is to find the rank-r matrix that is closest to X in some norm. If we use either the operator norm or the Frobenius norm, then this matrix is

$$\tilde{X} = \sum_{i=1}^{r} \lambda_i v_i v_i^\mathsf{T}.$$

Although this method can be effective in some cases, it can also perform very poorly in other cases, as shown in the following example. For simplicity we present the example as a linear program for which we seek a low-cardinality solution.

Example 3.2. Consider the linear program

$$\begin{array}{ll} \text{minimize} & (n-2)x_1 - 2x_2 - \cdots - 2x_{n-1} \\ \text{subject to} & x_1 + 2x_n = 2 \\ & x_i + x_n = 1, \quad i = 2, \ldots, n-1 \\ & x \geq 0 \end{array}$$

with variable $x \in \mathbf{R}^n$. The set of solutions of this problem is

$$\mathcal{X}^\star = \{(2\theta, \theta, \ldots, \theta, 1-\theta) \in \mathbf{R}^n \mid 0 \leq \theta \leq 1\}.$$

Suppose we are given the solution corresponding to $\theta = 1/2$:

$$x^\star = (1, 1/2, \ldots, 1/2) \in \mathbf{R}^n.$$

Rounding this solution to the nearest vector with one nonzero component gives $\tilde{x} = e_1$, which violates all of the equality constraints, and achieves an objective value of $n-2$. However, the best solution with one nonzero component is e_n, which is the solution of the linear program corresponding to $\theta = 0$, and achieves an objective value of 0.

3.4.2 Binary quadratic maximization

Consider the binary quadratic maximization problem

$$\begin{array}{ll} \text{maximize} & x^\mathsf{T} Q x \\ \text{subject to} & x_i \in \{\pm 1\}, \quad i = 1, \ldots, n, \end{array} \tag{3.3}$$

where $x \in \mathbf{R}^n$ is the variable, and $Q \in \mathbf{S}^n$ is problem data. Note that we do not assume that Q is positive semidefinite. Let $z^\star$ be the optimal value of this problem. We can formulate the constraint $x_i \in \{\pm 1\}$ as the quadratic constraint $x_i^2 = 1$. This gives us a polynomial optimization problem whose natural SDP relaxation of (3.3) is

$$\begin{array}{ll} \text{maximize} & Q \bullet X \\ \text{subject to} & E_{ii} \bullet X = 1, \quad i = 1, \ldots, n \\ & X \succeq 0, \end{array} \tag{3.4}$$

where $X \in \mathbf{S}^n$ is the variable, and E_{ii} is the matrix whose (i,i)-entry is equal to 1, and whose other entries are all equal to 0. Let $\tilde{z}^\star$ be the optimal value of the SDP relaxation. Because (3.4) is a relaxation of (3.3), we have that $\tilde{z}^\star \geq z^\star$. The proof of the following theorem shows that, in the special case when Q is positive semidefinite, we can randomly round a solution of the SDP relaxation to an approximate solution of (3.3) whose expected objective value is within a factor of $2/\pi \approx 0.6366$ of optimal. For such a randomized rounding algorithm, we typically compute several rounded solutions, and then report the solution with the highest objective value.

Theorem 3.2 (Nesterov [74]). If Q is positive semidefinite, then

$$z^\star \geq (2/\pi)\tilde{z}^\star.$$

Proof. Let $X \in \mathbf{S}^n$ be a solution of (3.4), and $x \in \mathbf{R}^n$ be a normal random vector with mean vector 0 and covariance matrix X. Define the vector $\hat{x} \in \mathbf{R}^n$ such that

$$\hat{x}_i = \begin{cases} 1 & x_i \geq 0, \\ -1 & \text{otherwise.} \end{cases}$$

The constraint $E_{ii} \bullet X = 1$ in (3.4) implies that x_i has unit variance. Because x_i and x_j have unit variance, X_{ij} is the correlation of x_i and x_j. Thus, Corollary C.4 tells us that

$$\mathbf{E}\Big((\hat{x}\hat{x}^\mathsf{T})_{ij}\Big) = \mathbf{E}(\hat{x}_i\hat{x}_j) = \frac{2}{\pi}\arcsin(X_{ij}).$$

Applying this result, we find that

$$z^\star \geq \mathbf{E}\Big(\hat{x}^\mathsf{T}Q\hat{x}\Big) = Q \bullet \mathbf{E}\Big(\hat{x}\hat{x}^\mathsf{T}\Big) = \frac{2}{\pi}\left(Q \bullet \arcsin(X)\right).$$

Since $X \succeq 0$, Corollary A.9 tells us that $\arcsin(X) \succeq X$. Combined with the assumption that $Q \succeq 0$, this implies that

$$z^\star = \frac{2}{\pi}\left(Q \bullet \arcsin(X)\right) \geq \frac{2}{\pi}\left(Q \bullet X\right) = (2/\pi)\tilde{z}^\star.$$

□

3.4.3 The maximum-cut problem

Given an undirected graph $G = (V, E)$ with nonnegative edge weights, the maximum-cut problem is to partition the set of vertices into two sets in order to maximize the total weight of the edges between the two sets. More concretely, suppose the vertex set is $V = \{1, \ldots, n\}$, and let

$$w_{ij} = \begin{cases} \text{the weight of edge } (i,j) & (i,j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

The maximum-cut problem is to find a subset S of V that maximizes

$$\sum_{i \in S} \sum_{j \notin S} w_{ij},$$

It is well known that this problem is NP-hard – the corresponding decision problem was in Karp's original list of NP-complete problems [56].

The Goemans-Williamson algorithm [39] is an approximation algorithm for the maximum-cut problem that achieves an approximation ratio of

$$\alpha = \frac{2}{\pi} \min_{0 \le \theta \le \pi} \left(\frac{\theta}{1 - \cos(\theta)} \right) \approx 0.8786.$$

This is currently the best known approximation ratio for the maximum-cut problem among all polynomial-time algorithms. Moreover, under the unique-games conjecture [57], it is NP-hard to obtain an approximation ratio that is better than that of the Goemans-Williamson algorithm [58]. Without relying on any unproven conjectures, it is possible to show that it is NP-hard to obtain an approximation ratio better than $16/17 \approx 0.9412$ [47, 94].

We can describe the set S using the vector $x \in \mathbf{R}^n$ such that

$$x_i = \begin{cases} 1 & i \in S, \\ -1 & i \notin S. \end{cases}$$

Then we have that

$$\frac{1 + x_i}{2} = \begin{cases} 1 & i \in S, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad \frac{1 - x_j}{2} = \begin{cases} 0 & j \in S, \\ 1 & \text{otherwise.} \end{cases}$$

Thus, we can write the objective of the maximum-cut problem as

$$\begin{aligned}\sum_{i\in S}\sum_{j\notin S} w_{ij} &= \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(\frac{1+x_i}{2}\right)\left(\frac{1-x_j}{2}\right)\\ &= \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(1+x_i-x_j-x_ix_j)\\ &= \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(1-x_ix_j),\end{aligned}$$

where the last step follows from the fact that

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}x_i = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}x_j$$

because $w_{ij} = w_{ji}$, so that $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - x_j) = 0$. Since $x_i \in \{\pm 1\}$, we have that $x_i^2 = 1$ for $i = 1, \ldots, n$. Therefore,

$$\begin{aligned}\sum_{i\in S}\sum_{j\notin S} w_{ij} &= \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i^2 - x_ix_j)\\ &= \sum_{i=1}^{n}\left(\frac{1}{4}\sum_{j=1}^{n} w_{ij}\right)x_i^2 - \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{1}{4}w_{ij}\right)x_ix_j\\ &= x^{\mathsf{T}}Qx,\end{aligned}$$

where we define the matrix $Q \in \mathbf{S}^n$ such that

$$Q_{ij} = \begin{cases}\left(\sum_{k=1}^{n} w_{ik} - w_{ij}\right)/4 & i = j,\\ -w_{ij}/4 & \text{otherwise.}\end{cases}$$

Thus, the maximum-cut problem can be represented as a binary-quadratic optimization problem:

$$\begin{array}{ll}\text{maximize} & x^{\mathsf{T}}Qx\\ \text{subject to} & x_i \in \{\pm 1\}, \quad i = 1, \ldots, n.\end{array} \tag{3.5}$$

Theorem 3.2 tells us that the optimal value of (3.5) is within a factor of $2/\pi \approx 0.6366$ of the optimal value of its SDP relaxation. However, we can use the special structure of Q in (3.5) to show that the approximation ratio is actually much better.

Theorem 3.3 (Goemans and Williamson [39])**.** Let $z^\star$ and $\tilde{z}^\star$ denote the optimal values of (3.5) and its SDP relaxation, respectively. Then, $z^\star \geq \alpha \tilde{z}^\star$, where

$$\alpha = \frac{2}{\pi} \min_{0 \leq \theta \leq \pi} \left(\frac{\theta}{1 - \cos(\theta)} \right) \approx 0.8786.$$

Proof. The SDP relaxation of (3.5) is

$$\begin{array}{ll} \text{minimize} & Q \bullet X \\ \text{subject to} & E_{ii} \bullet X = 1, \quad i = 1, \ldots, n \\ & X \succeq 0. \end{array} \tag{3.6}$$

Let $X \in \mathbf{S}^n$ be a solution of (3.6), and $x \in \mathbf{R}^n$ be a normal random vector with mean vector 0 and covariance matrix X. Define the vector $\hat{x} \in \mathbf{R}^n$ such that

$$\hat{x}_i = \begin{cases} 1 & x_i \geq 0, \\ -1 & \text{otherwise}. \end{cases}$$

The constraint $E_{ii} \bullet X = 1$ in (3.6) implies that x_i has unit variance. Because x_i and x_j have unit variance, X_{ij} is the correlation of x_i and x_j. Thus, Corollary C.4 tells us that

$$\mathbf{E}\Big((\hat{x}\hat{x}^\mathsf{T})_{ij} \Big) = \mathbf{E}(\hat{x}_i \hat{x}_j) = \frac{2}{\pi} \arcsin(X_{ij}).$$

Our definition of the matrix Q implies that

$$\hat{x}^\mathsf{T} Q \hat{x} = \frac{1}{4} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(1 - \hat{x}_i \hat{x}_j)$$

for all $\hat{x} \in \mathbf{R}^n$ such that $\hat{x}_i^2 = 1$, and

$$Q \bullet X = \frac{1}{4} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(1 - X_{ij})$$

for all $X \in \mathbf{S}^n$ such that $X_{ii} = 1$. Since the sum of the acute angles in a right triangle is $\pi/2$, we have that $\arccos(X_{ij}) + \arcsin(X_{ij}) = \pi/2$, and hence that

$$1 - \frac{2}{\pi} \arcsin(X_{ij}) = \frac{2}{\pi} \arccos(X_{ij}).$$

Combining these results, we find that

$$\begin{aligned}\mathbf{E}\Big(\hat{x}^{\mathsf{T}}Q\hat{x}\Big) &= \mathbf{E}\left(\frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(1-\hat{x}_i\hat{x}_j)\right)\\ &= \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(1-\mathbf{E}(\hat{x}_i\hat{x}_j))\\ &= \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(1-\frac{2}{\pi}\arcsin(X_{ij})\right)\\ &= \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(\frac{2}{\pi}\arccos(X_{ij})\right).\end{aligned}$$

Consider any constant α such that

$$\frac{2}{\pi}\arccos(t) \geq \alpha(1-t), \quad -1 \leq t \leq 1.$$

(We only need the inequality to be satisfied for $-1 \leq t \leq 1$ because X_{ij} is a correlation, so $-1 \leq X_{ij} \leq 1$.) For such an α, we have that

$$\mathbf{E}\Big(\hat{x}^{\mathsf{T}}Q\hat{x}\Big) \geq \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\alpha(1-X_{ij}) = \alpha(Q\bullet X) = \alpha\tilde{z}^{\star}.$$

In order to obtain the tightest bound possible, we choose α to be the largest value satisfying the constraint above: that is,

$$\alpha = \frac{2}{\pi}\min_{-1\leq t\leq 1}\left(\frac{\arccos(t)}{1-t}\right) = \frac{2}{\pi}\min_{0\leq\theta\leq\pi}\left(\frac{\theta}{1-\cos(\theta)}\right) \approx 0.8786.$$

(The largest lower bound for a function over an interval is the minimum value of the function on the interval.) Having shown how to construct an approximate solution $\hat{x}$ of (3.5) such that $\mathbf{E}\Big(\hat{x}^{\mathsf{T}}Q\hat{x}\Big) \geq \alpha\tilde{z}^{\star}$, we can conclude that $z^{\star} \geq \alpha\tilde{z}^{\star}$. □

3.4.4 Problems with positive-semidefinite coefficients

It is also possible to establish bounds on the quality of a rounded solution in the case when all of the coefficient matrices are positive-semidefinite. This result is based on the analysis of So, Ye, and

Zhang [85], although we present somewhat different bounds here. For simplicity we only consider feasibility problems

$$\begin{aligned} A_i \bullet X &= b_i, \quad i = 1, \ldots, m, \\ X &\succeq 0 \end{aligned}$$

with variable $X \in \mathbf{S}^n$, and problem data $A_1, \ldots, A_m \in \mathbf{S}^n_+$ and $b \in \mathbf{R}^m_+$. Note that $A_i \bullet X \geq 0$ because A_i and X are both positive semidefinite; thus, the assumption that the b_i are nonnegative only serves to exclude problem instances that are trivially infeasible. We can extend our analysis to optimization problems by finding the optimal value $z^\star$ of (SDP), and then considering the SDP feasibility problem with equality constraints $C \bullet X = z^\star$ and $A_i \bullet X = b_i$ for $i = 1, \ldots, m$.

Theorem 3.4. Suppose $A_1, \ldots, A_m, X \in \mathbf{S}^n_+$ and $b \in \mathbf{R}^m_+$ satisfy

$$A_i \bullet X = b_i, \quad i = 1, \ldots, m.$$

Because X is positive semidefinite, there exists a matrix $V \in \mathbf{R}^{n \times r}$ such that $X = VV^\mathsf{T}$, where $r = \mathbf{rank}(X)$. Let $z_1, \ldots, z_d \in \mathbf{R}^r$ be independent standard normal random vectors, and define

$$Z = \frac{1}{d} \sum_{k=1}^{d} z_k z_k^\mathsf{T} \quad \text{and} \quad \tilde{X} = VZV^\mathsf{T}.$$

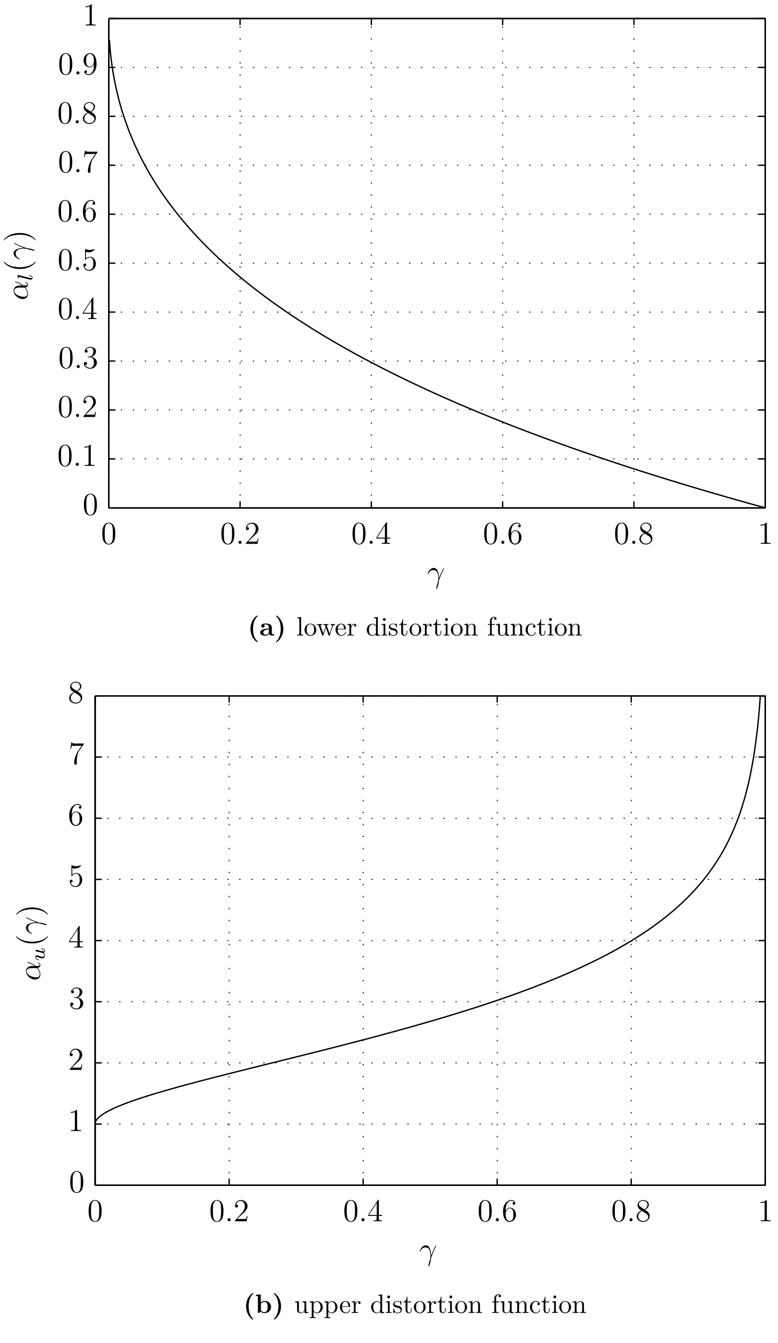
Then, for all $\gamma \in (0, 1)$, we have that

$$\mathbf{prob}\Big(\alpha_l(\gamma) b_i \leq A_i \bullet \tilde{X} \leq \alpha_u(\gamma) b_i, i = 1, \ldots, m\Big) \geq 1 - 2m(1-\gamma)^{\frac{d}{2}},$$

where we define the distortion functions

$$\alpha_l(\gamma) = -W_0\left(\frac{\gamma - 1}{e}\right) \quad \text{and} \quad \alpha_u(\gamma) = -W_{-1}\left(\frac{\gamma - 1}{e}\right),$$

and W_k is the kth branch of the Lambert W function (see Remark 3.1). The distortion functions $\alpha_l(\gamma)$ and $\alpha_u(\gamma)$ are shown in Figure 3.4. In particular, if $\gamma > 1 - (2m)^{-\frac{2}{d}}$, then there is positive probability that $\tilde{X}$ satisfies the distortion bounds $\alpha_l(\gamma) b_i \leq A_i \bullet \tilde{X} \leq \alpha_u(\gamma) b_i$ for all $i = 1, \ldots, m$.



(a) lower distortion function

(b) upper distortion function

Figure 3.4: distortion functions in Theorem 3.4

Proof. Suppose $b_i = 0$. Then we have that

$$A_i \bullet (VV^\mathsf{T}) = A_i \bullet X = b_i = 0,$$

so we can use Lemma A.3 to conclude that $A_i V = 0$. This implies that

$$A_i \bullet \tilde{X} = A_i \bullet (VZV^\mathsf{T}) = (A_i V) \bullet (ZV^\mathsf{T}) = (0) \bullet (ZV^\mathsf{T}) = 0.$$

Thus, $\tilde{X}$ satisfies all homogeneous equality constraints exactly. In the rest of the proof, we assume that the homogeneous equality constraints have been discarded, so $b_i > 0$ for $i = 1, \ldots, m$.

We are interested in bounding the probability that the distortion bounds are satisfied:

$$\begin{aligned}
&\mathbf{prob}\Big(\alpha_l(\gamma) b_i \le A_i \bullet \tilde{X} \le \alpha_u(\gamma) b_i, i = 1, \ldots, m\Big) \\
&\quad = 1 - \mathbf{prob}\left(\bigcup_{i=1}^m \Big(A_i \bullet \tilde{X} < \alpha_l(\gamma) b_i \text{ or } A_i \bullet \tilde{X} > \alpha_u(\gamma) b_i\Big)\right).
\end{aligned}$$

Applying the union bound yields

$$\begin{aligned}
&\mathbf{prob}\Big(\alpha_l(\gamma) b_i \le A_i \bullet \tilde{X} \le \alpha_u(\gamma) b_i, i = 1, \ldots, m\Big) \\
&\quad \ge 1 - \sum_{i=1}^m \Big(\mathbf{prob}\Big(A_i \bullet \tilde{X} < \alpha_l(\gamma) b_i\Big) + \mathbf{prob}\Big(A_i \bullet \tilde{X} > \alpha_u(\gamma) b_i\Big)\Big) \\
&\quad = 1 - \sum_{i=1}^m \mathbf{prob}\left(\frac{(A_i \bullet \tilde{X})d}{b_i} < \alpha_l(\gamma) d\right) \\
&\qquad - \sum_{i=1}^m \mathbf{prob}\left(\frac{(A_i \bullet \tilde{X})d}{b_i} > \alpha_u(\gamma) d\right).
\end{aligned}$$

(We can divide by b_i without changing the direction of the inequality due to our assumption that the b_i are all strictly positive.) Observe that we can write $A_i \bullet \tilde{X}$ and b_i as

$$\begin{aligned}
A_i \bullet \tilde{X} &= A_i \bullet (VZV^\mathsf{T}) = (V^\mathsf{T} A_i V) \bullet Z, \\
b_i &= A_i \bullet X = A_i \bullet (VV^\mathsf{T}) = \mathbf{tr}(V^\mathsf{T} A_i V).
\end{aligned}$$

Using these observations, we can express our bound as

$$\begin{aligned}
&\mathbf{prob}\Big(\alpha_l(\gamma)b_i \le A_i \bullet \tilde{X} \le \alpha_u(\gamma)b_i, i = 1, \dots, m\Big) \\
&\quad \ge 1 - \sum_{i=1}^{m} \mathbf{prob}\left(\frac{((V^\mathsf{T} A_i V) \bullet Z)d}{\mathbf{tr}(V^\mathsf{T} A_i V)} < \alpha_l(\gamma)d\right) \\
&\quad\quad - \sum_{i=1}^{m} \mathbf{prob}\left(\frac{((V^\mathsf{T} A_i V) \bullet Z)d}{\mathbf{tr}(V^\mathsf{T} A_i V)} > \alpha_u(\gamma)d\right).
\end{aligned}$$

Since $V^\mathsf{T} A_i V \in \mathbf{S}^r$ is symmetric, it has an eigenvalue expansion

$$V^\mathsf{T} A_i V = \sum_{j=1}^{r} \lambda_{ij} q_{ij} q_{ij}^\mathsf{T},$$

where $\lambda_{i1}, \dots, \lambda_{ir} \in \mathbf{R}$ are the eigenvalues of A_i, and $q_{i1}, \dots, q_{ir} \in \mathbf{R}^r$ form an orthonormal set of corresponding eigenvectors. Note that $V^\mathsf{T} A_i V$ is positive semidefinite because A_i is positive semidefinite; therefore, the eigenvalues λ_{ij} are nonnegative. Using this eigenvalue expansion of $V^\mathsf{T} A_i V$, the definition of Z, and the fact that the trace of a matrix is the sum of its eigenvalues, we have that

$$\begin{aligned}
\frac{((V^\mathsf{T} A_i V) \bullet Z)d}{\mathbf{tr}(V^\mathsf{T} A_i V)} &= \frac{1}{\sum_{\tilde{j}=1}^{r} \lambda_{i\tilde{j}}} \left(\left(\sum_{j=1}^{r} \lambda_{ij} q_{ij} q_{ij}^\mathsf{T}\right) \bullet \left(\frac{1}{d}\sum_{k=1}^{d} z_k z_k^\mathsf{T}\right)\right) d \\
&= \sum_{j=1}^{r} \frac{\lambda_{ij}}{\sum_{\tilde{j}=1}^{r} \lambda_{i\tilde{j}}} \sum_{k=1}^{d} (q_{ij} z_k)^2 \\
&= \sum_{j=1}^{r} \theta_{ij} \sum_{k=1}^{d} (q_{ij}^\mathsf{T} z_k)^2,
\end{aligned}$$

where we define

$$\theta_{ij} = \frac{\lambda_{ij}}{\sum_{\tilde{j}=1}^{r} \lambda_{i\tilde{j}}}.$$

Observe that the θ_{ij} are nonnegative, and satisfy

$$\sum_{j=1}^{r} \theta_{ij} = \sum_{j=1}^{r} \frac{\lambda_{ij}}{\sum_{\tilde{j}=1}^{r} \lambda_{i\tilde{j}}} = 1.$$

Because each z_k is a standard normal random vector, and each q_{ij} is a unit vector, we have that each $q_{ij}^\mathsf{T} z_k$ is a standard normal random variable. Additionally, because the z_k are independent standard normal random vectors, and $q_{i1}, \dots, q_{ir}$ form an orthonormal set, we have that

$$\begin{aligned}\mathbf{E}\Big((q_{ij_1}^\mathsf{T} z_{k_1})(q_{ij_2}^\mathsf{T} z_{k_2})^\mathsf{T}\Big) &= q_{ij_1}^\mathsf{T}\, \mathbf{E}\Big(z_{k_1} z_{k_2}^\mathsf{T}\Big) q_{ij_2} \\ &= \delta_{k_1 k_2} q_{ij_1}^\mathsf{T} q_{ij_2} \\ &= \delta_{k_1 k_2} \delta_{j_1 j_2}.\end{aligned}$$

Thus, $q_{ij_1}^\mathsf{T} z_{k_1}$ and $q_{ij_2}^\mathsf{T} z_{k_2}$ are uncorrelated unless $j_1 = j_2$ and $k_1 = k_2$. Since uncorrelated jointly normal random variables are independent, the $q_{ij}^\mathsf{T} z_k$ are independent standard normal random variables. This implies that

$$y_{ij} = \sum_{k=1}^{d} (q_{ij}^\mathsf{T} z_k)^2$$

is a chi-squared random variable with d degrees of freedom because it is the sum of the squares of d independent standard normal random variables. Moreover, the y_{ij} are independent because the $q_{ij}^\mathsf{T} z_k$ are independent for $j = 1, \dots, r$ and $k = 1, \dots, d$. Thus, we have that

$$\frac{((V^\mathsf{T} A_i V) \bullet Z) d}{\mathbf{tr}(V^\mathsf{T} A_i V)} = \sum_{j=1}^{d} \theta_{ij} y_{ij},$$

where the y_{ij} are independent chi-squared random variables with d degrees of freedom, and $\theta_{i1}, \dots, \theta_{ir}$ are nonnegative scalars summing to 1 for all $i = 1, \dots, m$. Therefore, we can apply Lemma C.2:

$$\begin{aligned}&\mathbf{prob}\left(\frac{((V^\mathsf{T} A_i V) \bullet Z) d}{\mathbf{tr}(V^\mathsf{T} A_i V)} < \alpha_l(\gamma) d\right) \\ &\quad = \mathbf{prob}\left(\sum_{j=1}^{r} \theta_{ij} y_{ij} < \alpha_l(\gamma) d\right) \\ &\quad \le (e\alpha_l(\gamma)\exp(-\alpha_l(\gamma)))^{\frac{d}{2}} \\ &\quad = \left(-eW_0\left(\frac{\gamma-1}{e}\right)\exp\left(W_0\left(\frac{\gamma-1}{e}\right)\right)\right)^{\frac{d}{2}} \\ &\quad = \left(-e\left(\frac{\gamma-1}{e}\right)\right)^{\frac{d}{2}} \\ &\quad = (1-\gamma)^{\frac{d}{2}},\end{aligned}$$

where we have made use of the fact that $W_0(y)\exp(W_0(y)) = y$ for all $y \in (-1/e, 0)$ (see Remark 3.1). Note that we can apply the lemma even though the inequality inside the probability is strict since chi-squared random variables are continuous, so the probability is the same whether the inequality is weak or strict. Similarly, Lemma C.1 tells us that

$$\begin{aligned}
&\mathbf{prob}\left(\frac{((V^{\mathsf{T}}A_iV)\bullet Z)d}{\mathbf{tr}(V^{\mathsf{T}}A_iV)} > \alpha_u(\gamma)d\right)\\
&= \mathbf{prob}\left(\sum_{j=1}^{r}\theta_{ij}y_{ij} > \alpha_u(\gamma)d\right)\\
&\leq (e\alpha_u(\gamma)\exp(-\alpha_u(\gamma)))^{\frac{d}{2}}\\
&= \left(-eW_{-1}\left(\frac{\gamma-1}{e}\right)\exp\left(W_{-1}\left(\frac{\gamma-1}{e}\right)\right)\right)^{\frac{d}{2}}\\
&= \left(-e\left(\frac{\gamma-1}{e}\right)\right)^{\frac{d}{2}}\\
&= (1-\gamma)^{\frac{d}{2}}.
\end{aligned}$$

Combining these results gives

$$\mathbf{prob}\Big(\alpha_l(\gamma)b_i \leq A_i \bullet \tilde{X} \leq \alpha_u(\gamma)b_i, i = 1, \ldots, m\Big) \geq 1 - 2m(1-\gamma)^{\frac{d}{2}}.$$

□

Remark 3.1. A general discussion of the Lambert W function is given by Corless, et al. [26]. For our purposes it suffices to know that the W function satisfies $W(y)\exp(W(y)) = y$ for all $y \in (-1/e, 0)$. A sketch of the function that maps z to $z\exp(z)$ is shown in Figure 3.5. For every $y \in (-1/e, 0)$, there are exactly two values of z such that $z\exp(z) = y$. The value of z such that $z\exp(z) = y$ and $z \in (-1, 0)$ is given by $W_0(y)$; the value of z such that $z\exp(z) = y$ and $z \in (-\infty, -1)$ is given by $W_{-1}(y)$.

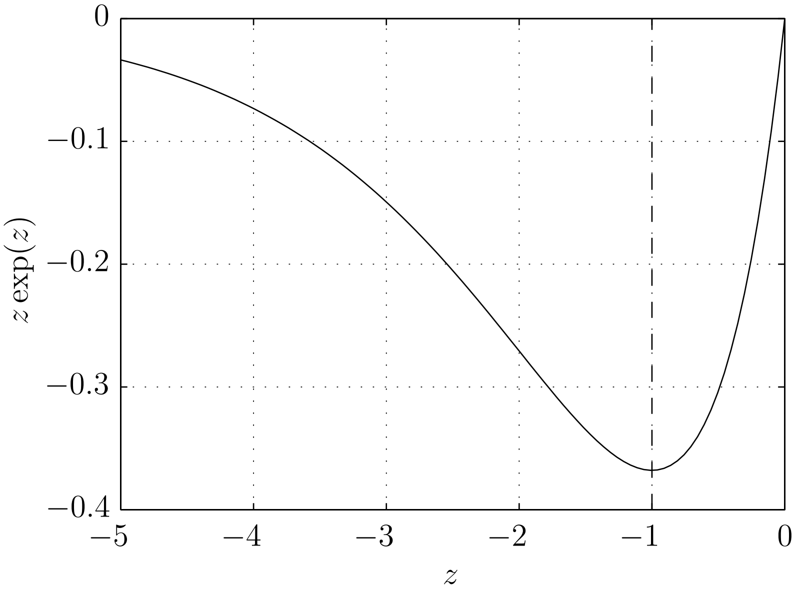


Figure 3.5: the function that maps z to $z\exp(z)$

Part II

Applications

4

Trust-Region Problems

A trust-region problem (4.1) is an optimization problem of the form

$$\begin{array}{ll} \text{minimize} & x^\mathsf{T} P_0 x + 2q_0^\mathsf{T} x + r_0 \\ \text{subject to} & a_i^\mathsf{T} x \leq b_i, \quad i = 1, \ldots, m \\ & x^\mathsf{T} P_i x + 2q_i^\mathsf{T} x + r_i \leq 0, \quad i = 1, \ldots, p \\ & \|x\| = 1. \end{array} \tag{4.1}$$

The optimization variable is $x \in \mathbf{R}^n$, and the problem data are $a_1, \ldots, a_m \in \mathbf{R}^n$, $b_1, \ldots, b_m \in \mathbf{R}$, $P_0, \ldots, P_p \in \mathbf{S}^n$, $q_0, \ldots, q_p \in \mathbf{R}^n$, and $r_0, \ldots, r_p \in \mathbf{R}$. We assume that $P_1, \ldots, P_p$ are positive semidefinite, so each quadratic inequality constraint represents an ellipsoid (or degenerate ellipsoid). However, note that the objective may not be convex because we do not assume that P_0 is positive semidefinite.

An important special case of (4.1) is the simple trust-region problem, where $m = p = 0$. For example, the Levenberg-Marquardt algorithm for nonlinear programming solves an instance of the simple trust-region problem in each step of the algorithm. The simple trust-region problem is known to be much easier than the general trust-region problem, and has been studied extensively [29, 37, 41, 44, 72, 73, 81, 90]. Additionally, some algorithms for nonconvex quadratic programming, which is NP-hard in general, use the simple trust-region problem as a subproblem [55, 101]. It is possible to show that there is no duality gap for the simple trust-region problem [91]; however, a duality gap may exist for the general trust-region problem (4.1).

Special cases of (4.1) with $m = 0$ and $0 < s \leq 2$ have also received considerable attention. For example, Peng and Yuan [77] considered the problem of minimizing a quadratic function subject to two quadratic constraints. Using properties of local minimizers [66], Martinez and Santos [67] presented an algorithm for minimizing a quadratic function subject to two strictly convex quadratic constraints. Zhang [104] proposed an algorithm for the general quadratic case, and the two-dimensional trust-region problem was solved by Williamson [99]. Unfortunately, a duality gap may exist for all of these problems.

In this chapter we describe approaches to solving trust-region problems using semidefinite programming. The key step in the analysis is typically to demonstrate that the SDP relaxation has a rank-1 solution for certain special cases of (4.1).

4.1 SDP relaxation of a trust-region problem

Note that (4.1) is a quadratic optimization problem. The derivations of the primal and dual SDP relaxations of such a problem are given in Section A.2.4. In particular, the primal SDP relaxation of (4.1) is

$$\begin{array}{ll} \text{minimize} & Q_0 \bullet X \\ \text{subject to} & L_i \bullet X \leq 0, \quad i = 1, \dots, m \\ & Q_i \bullet X \leq 0, \quad i = 1, \dots, p \\ & F_i \bullet X = 1, \quad i = 1, 2 \\ & X \succeq 0, \end{array} \tag{4.2}$$

where the optimization variable is $X \in \mathbf{S}^{n+1}$, and we define

$$Q_i = \begin{bmatrix} P_i & q_i \\ q_i^{\mathsf{T}} & r_i \end{bmatrix}, \quad i = 0, \dots, p,$$

$$L_i = \begin{bmatrix} 0 & (1/2)a_i \\ (1/2)a_i^{\mathsf{T}} & -b_i \end{bmatrix}, \quad i = 1, \dots, m,$$

$$F_1 = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \quad \text{and} \quad F_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

The corresponding dual relaxation is

$$\begin{array}{ll} \text{maximize} & \nu_1 + \nu_2 \\ \text{subject to} & \sum_{i=1}^{2} \nu_i F_i - \sum_{i=1}^{m} \lambda_i L_i - \sum_{i=1}^{p} \mu_i Q_i + S = Q_0 \\ & \lambda \geq 0 \\ & \mu \geq 0 \\ & S \succeq 0, \end{array} \tag{4.3}$$

with variables $\lambda \in \mathbf{R}^m$, $\mu \in \mathbf{R}^p$, $\nu \in \mathbf{R}^2$, and $S \in \mathbf{S}^{n+1}$. The complementarity conditions for (4.2) and (4.3) are

$$\begin{aligned} \lambda_i(L_i \bullet X) &= 0, \quad i = 1, \ldots, m \\ \mu_i(Q_i \bullet X) &= 0, \quad i = 1, \ldots, p \\ S \bullet X &= 0. \end{aligned}$$

Since X and S are positive semidefinite, the last complementarity condition is equivalent to $SX = 0$.

We argue in Section A.2.4 that (4.2) is an exact relaxation if it has a rank-1 solution. The following lemma gives sufficient conditions for the existence of a rank-1 solution of (4.2).

Lemma 4.1. Suppose (4.2) and (4.3) are both solvable, and there is no duality gap. Let X and (λ, μ, ν, S) be solutions of (4.2) and (4.3), respectively. The boundary of the second-order cone is

$$\mathbf{bd}(\mathbf{SOC}) = \{(x, t) \in \mathbf{R}^{n+1} \mid \|x\|_2 = t\}.$$

If there exists a nonzero vector $z \in \mathbf{range}(X) \cap \mathbf{bd}(\mathbf{SOC})$ such that

(i) $z^\mathsf{T} L_i z \leq 0$ for $i = 1, \ldots, m$,

(ii) $z^\mathsf{T} Q_i z \leq 0$ for $i = 1, \ldots, p$,

(iii) $\lambda_i(z^\mathsf{T} L_i z) = 0$ for $i = 1, \ldots, m$, and

(iv) $\mu_i(z^\mathsf{T} Q_i z) = 0$ for $i = 1, \ldots, p$,

then (4.2) is an exact relaxation of (4.1).

Proof. Since we assume $z \in \mathbf{bd}(\mathbf{SOC})$, we have that $\|z_{1:n}\| = z_{n+1}$, where $z_{1:n} = (z_1, \ldots, z_n)$ is the vector consisting of the first n components of z. Because we also assume that z is nonzero, it must be the case that z_{n+1} is nonzero, and hence that we can define the matrix

$$\tilde{X} = \frac{1}{z_{n+1}^2} z z^\mathsf{T}.$$

We have that $\mathbf{rank}(\tilde{X}) = 1$ since $\tilde{X}$ is a positive multiple of a nonzero dyad. We will show that $\tilde{X}$ is a solution of (4.2) by showing that it is feasible, and satisfies the complementarity conditions. Since we assume that $z^\mathsf{T} L_i z \leq 0$ for $i = 1, \ldots, m$, we have that

$$L_i \bullet \tilde{X} = \frac{z^\mathsf{T} L_i z}{z_{n+1}^2} \leq 0, \quad i = 1, \ldots, m.$$

Similarly, the assumption that $z^\mathsf{T} Q_i z \leq 0$ for $i = 1, \ldots, p$ implies that

$$Q_i \bullet \tilde{X} = \frac{z^\mathsf{T} Q_i z}{z_{n+1}^2} \leq 0, \quad i = 1, \ldots, p.$$

Using the definitions of F_1 and F_2, we find that

$$F_1 \bullet (z z^\mathsf{T}) = z^\mathsf{T} F_1 z = \|z_{1:n}\|^2 \quad \text{and} \quad F_2 \bullet (z z^\mathsf{T}) = z^\mathsf{T} F_2 z = z_{n+1}^2.$$

Therefore, we have that

$$F_1 \bullet \tilde{X} = \frac{\|z_{1:n}\|^2}{z_{n+1}^2} = 1 \quad \text{and} \quad F_2 \bullet \tilde{X} = \frac{z_{n+1}^2}{z_{n+1}^2} = 1.$$

Additionally, we have that $\tilde{X}$ is positive semidefinite because it is a positive multiple of a dyad. Taken together, these results show that $\tilde{X}$ is feasible for (4.2).

Our assumptions that $\lambda_i(z^\mathsf{T} L_i z) = 0$ and $\mu_i(z^\mathsf{T} Q_i z) = 0$ imply that

$$\lambda_i(L_i \bullet \tilde{X}) = \frac{\lambda_i(z^\mathsf{T} L_i z)}{z_{n+1}^2} = 0 \quad \text{and} \quad \mu_i(Q_i \bullet \tilde{X}) = \frac{\mu_i(z^\mathsf{T} Q_i z)}{z_{n+1}^2} = 0.$$

Since $z \in \mathbf{range}(X)$, there exists a vector $\tilde{z}$ such that $z = X\tilde{z}$. We have that $SX = 0$ because S and X are solutions of their respective problems, and therefore satisfy complementary slackness. Combining these results, we find that

$$S\tilde{X} = \frac{1}{z_{n+1}^2} S z z^{\mathsf{T}} = \frac{1}{z_{n+1}^2} S X \tilde{z} z^{\mathsf{T}} = 0.$$

Thus, we have shown that $\tilde{X}$ is feasible, and satisfies the complementarity conditions. This implies that $\tilde{X}$ is a solution of (4.2). □

4.2 The simple trust-region problem

The simple trust-region problem (4.4) is a special case of (4.1) with $m = p = 0$:

$$\begin{array}{ll} \text{minimize} & x^{\mathsf{T}} P_0 x + 2 q_0^{\mathsf{T}} x + r_0 \\ \text{subject to} & \|x\| = 1. \end{array} \tag{4.4}$$

SDP relaxation of the simple trust-region problem

The SDP relaxation of (4.4) is

$$\begin{array}{ll} \text{minimize} & Q_0 \bullet X \\ \text{subject to} & F_i \bullet X = 1, \quad i = 1, 2 \\ & X \succeq 0. \end{array} \tag{4.5}$$

This problem has $m = 2$ equality constraints, so Theorem 2.1 guarantees the existence of a rank-1 solution. Thus, this SDP relaxation is exact. In particular, if $X = vv^{\mathsf{T}} \in \mathbf{S}^{n+1}$ is a rank-1 solution of (4.5), where $v \in \mathbf{R}^{n+1}$, then $x = v_{n+1}(v_1, \ldots, v_n)$ is a solution of (4.4).

4.3 Linear equality constraints

It turns out that adding linear equality constraints to (4.4) does not make the problem more difficult – we can reduce such a problem to an instance of (4.4). Consider the problem

$$\begin{array}{ll} \text{minimize} & x^{\mathsf{T}} P_0 x + 2 q_0^{\mathsf{T}} x + r_0 \\ \text{subject to} & Ax = b \\ & \|x\| = 1, \end{array}$$

where we assume that $A \in \mathbf{R}^{m\times n}$ is fat and full rank. (Note that if A is not fat and full rank, then some of the equality constraints are either redundant or inconsistent. We can use Gaussian elimination to detect inconsistent constraints, or to identify and remove redundant constraints.) Let $x_{\text{mn}} = A^\dagger b$ denote the minimum-norm solution of $Ax = b$, where $A^\dagger$ is the pseudoinverse of A. The problem is infeasible if $\|x_{\text{mn}}\| > 1$; if $\|x_{\text{mn}}\| = 1$, then x_{mn} is the solution of (4.4) because it is the only feasible point. (Note that the minimum-norm solution of $Ax = b$ is unique because we assume that A is fat and full rank.) Now consider the case when $\|x_{\text{mn}}\| < 1$. We can write every solution of $Ax = b$ in the form $x = x_{\text{mn}} + z$, where $z \in \mathbf{null}(A)$. We can then reformulate (4.4) in terms of the variable z:

$$\begin{array}{ll}\text{minimize} & z^\mathsf{T} P_0 z + 2(q_0 + P_0 x_{\text{mn}})^\mathsf{T} z + (x_{\text{mn}}^\mathsf{T} P_0 x_{\text{mn}} + 2q_0^\mathsf{T} x_{\text{mn}} + r_0)\\ \text{subject to} & Az = 0\\ & \|z\| = (1 - \|x_{\text{mn}}\|^2)^{\frac{1}{2}}.\end{array}$$

In this reformulation we have used the fact that the minimum-norm solution x_{mn} is in the orthogonal complement of $\mathbf{null}(A)$, so that

$$\|x_{\text{mn}} + z\|^2 = \|x_{\text{mn}}\|^2 + 2x_{\text{mn}}^\mathsf{T} z + \|z\|^2 = \|x_{\text{mn}}\|^2 + \|z\|^2.$$

Therefore, we can write the condition $\|x\| = 1$ as

$$\|z\| = (1 - \|x_{\text{mn}}\|^2)^{\frac{1}{2}}.$$

We have that $\dim(\mathbf{null}(A)) = m - n$ since we assume that A is fat and full rank. Let $N \in \mathbf{R}^{n\times(m-n)}$ be a matrix whose columns form an orthonormal basis for $\mathbf{null}(A)$. Then every $z \in \mathbf{null}(A)$ can be written as $z = (1 - \|x_{\text{mn}}\|^2)^{\frac{1}{2}} N w$, where $w \in \mathbf{R}^{m-n}$. Expressing our optimization problem in terms of w gives a simple trust-region problem:

$$\begin{array}{ll}\text{minimize} & w^\mathsf{T} \tilde{P} w + 2\tilde{q}^\mathsf{T} w + \tilde{r}\\ \text{subject to} & \|w\| = 1,\end{array}$$

where we define

$$\begin{aligned}\tilde{P} &= (1 - \|x_{\text{mn}}\|^2) N^\mathsf{T} P_0 N\\ \tilde{q} &= (1 - \|x_{\text{mn}}\|^2)^{\frac{1}{2}} N^\mathsf{T}(P_0 x_{\text{mn}} + q_0)\\ \tilde{r} &= x_{\text{mn}}^\mathsf{T} P_0 x_{\text{mn}} + 2q_0^\mathsf{T} x_{\text{mn}} + r_0.\end{aligned}$$

We obtain the constraint in our reformulated problem by noting that

$$\|w\| = \|Nw\| = \|(1 - \|x_{\mathrm{mn}}\|^2)^{-\frac{1}{2}} z\| = 1$$

because N has orthonormal columns, and $\|z\| = (1 - \|x_{\mathrm{mn}}\|^2)^{\frac{1}{2}}$.

4.4 Linear inequality constraints

We have shown that a trust-region problem with linear equality constraints can be reduced to an instance of (4.4). Now consider a trust-region problem with linear inequality constraints

$$\begin{array}{ll} \text{minimize} & x^{\mathsf{T}} P_0 x + 2q_0^{\mathsf{T}} x + r_0 \\ \text{subject to} & a_i^{\mathsf{T}} x \le b_i, \quad i = 1, \dots, m \\ & \|x\| = 1. \end{array} \tag{4.6}$$

4.4.1 Feasibility problems

First, we consider the feasibility problem associated with (4.6): that is, given vectors $a_1, \dots, a_m \in \mathbf{R}^n$ and $b \in \mathbf{R}^m$, we want to determine if there is a vector $x \in \mathbf{R}^n$ satisfying the constraints

$$\begin{aligned} & a_i^{\mathsf{T}} x \le b_i, \quad i = 1, \dots, m, \\ & \|x\| = 1. \end{aligned} \tag{4.7}$$

We assume that $a_1, \dots, a_m$ are linearly independent. (If $a_1, \dots, a_m$ are linearly dependent, then either there are variables that can be eliminated, or there are constraints that are either redundant or inconsistent; we can address these issues using Gaussian elimination.) Since we assume that $a_1, \dots, a_m$ are linearly independent, it must be the case that the system of equations

$$a_i^{\mathsf{T}} x = b_i, \quad i = 1, \dots, m$$

has a solution. Moreover, a system of equations and inequalities obtained by relaxing some of these equations to inequalities must also have a solution because every relaxation of a feasible problem is also feasible. First, we show that (4.7) is NP-hard in the general case.

Theorem 4.2. It is NP-hard to decide if (4.7) has a solution.

Proof. We will give a polynomial-time reduction to (4.7) from the partition problem, which is known to be NP-hard [36]. In the partition problem, we are given a set of integers $\{a_1, \ldots, a_N\}$, and we want to determine if there exists a subset S of $\{1, \ldots, N\}$ such that

$$\sum_{i \in S} a_i = \sum_{i \notin S} a_i.$$

Define the vector $a = (a_1, \ldots, a_N) \in \mathbf{R}^N$, and consider the following instance of (4.7):

$$\begin{aligned} \pm\sqrt{N}(a, -a)^{\mathsf{T}} x &\le 0 \\ \pm(e_j, e_j)^{\mathsf{T}} x &\le \pm 1/\sqrt{N}, && j = 1, \ldots, N \\ (-e_j, 0)^{\mathsf{T}} x &\le 0, && j = 1, \ldots, N \\ (0, -e_j)^{\mathsf{T}} x &\le 0, && j = 1, \ldots, N \\ \|x\| &= 1, \end{aligned}$$

where $x \in \mathbf{R}^{2N}$. Note that this problem has $m = 4N + 2$ linear inequality constraints. Let $x = (u, v)$, where $u, v \in \mathbf{R}^N$. We can express our instance of (4.7) in terms of u and v as

$$\begin{aligned} (\sqrt{N} a)^{\mathsf{T}} u &= (\sqrt{N} a)^{\mathsf{T}} v \\ u_j + v_j &= 1/\sqrt{N}, && j = 1, \ldots, N \\ u_j, v_j &\ge 0, && j = 1, \ldots, N \\ \|u\|^2 + \|v\|^2 &= 1. \end{aligned}$$

If u and v satisfy the second and fourth of these conditions, then

$$\begin{aligned} \sum_{j=1}^N u_j v_j &= \frac{1}{2} \sum_{j=1}^N ((u_j + v_j)^2 - (u_j^2 + v_j^2)) \\ &= \frac{1}{2} \left(\sum_{j=1}^N \left(\frac{1}{\sqrt{N}} \right)^2 - \left(\sum_{j=1}^N u_j^2 + \sum_{j=1}^N v_j^2 \right) \right) \\ &= \frac{1}{2} \left(1 - (\|u\|^2 + \|v\|^2) \right) \\ &= 0. \end{aligned}$$

Since $u_j, v_j \geq 0$, this implies that $u_j v_j = 0$ for $j = 1, \ldots, N$. Thus, there is a one-to-one correspondence between subsets S of $\{1, \ldots, N\}$, and vectors $x = (u, v)$ satisfying the constraints

$$u_j + v_j = \frac{1}{\sqrt{N}}, \quad u_j, v_j \geq 0, \quad \text{and} \quad \|u\|^2 + \|v\|^2 = 1.$$

In particular, given a subset S of $\{1, \ldots, N\}$, we set

$$u_j = \begin{cases} 1/\sqrt{N} & j \in S, \\ 0 & \text{otherwise}, \end{cases} \quad \text{and} \quad v_j = \begin{cases} 1/\sqrt{N} & j \notin S, \\ 0 & \text{otherwise}. \end{cases}$$

Similarly, given vectors u and v satisfying the constraints, we take

$$S = \{j \in \{1, \ldots, N\} \mid u_j = 1/\sqrt{N}\}.$$

Intuitively, we can think of $\sqrt{N}u$ and $\sqrt{N}v$ as the indicator vectors for the set S and its complement, respectively. This implies that

$$\sum_{i \in S} a_i = (\sqrt{N}a)^{\mathsf{T}} u \quad \text{and} \quad \sum_{i \notin S} a_i = (\sqrt{N}a)^{\mathsf{T}} v.$$

Thus, we have $\sum_{i \in S} a_i = \sum_{i \notin S} a_i$ if and only if $(\sqrt{N}a)^{\mathsf{T}} u = (\sqrt{N}a)^{\mathsf{T}} v$. This proves that the partition problem is equivalent to the given instance of (4.7), and hence that (4.7) is NP-hard. □

Although Theorem 4.2 states that (4.7) is NP-hard in general, the following theorem shows that we can solve (4.7) in polynomial time under an additional technical assumption.

Theorem 4.3. Suppose there exists an integer $\alpha < m$ such that $|\mathcal{A}| \leq \alpha$ for every index set $\mathcal{A} \subseteq \{1, \ldots, m\}$ with the property that

$$\{x \in \mathbf{R}^n \mid \|x\| \leq 1, \ a_i^{\mathsf{T}} x = b_i \text{ for } i \in \mathcal{A}\} \neq \emptyset.$$

Then we can compute a solution of (4.7) in polynomial time.

Proof. Let x_0 be a solution of

$$\begin{array}{ll} \text{minimize} & \|x\|^2 \\ \text{subject to} & a_i^{\mathsf{T}} x \leq b_i, \quad i = 1, \ldots, m. \end{array}$$

(We can solve this problem efficiently using standard algorithms.) If $\|x_0\| > 1$, then (4.7) is infeasible; if $\|x_0\| = 1$, then x_0 is a solution of (4.7). Thus, we can focus on the case when $\|x_0\| < 1$. Choose an index set $\mathcal{A} \subseteq \{1, \dots, m\}$ with $|\mathcal{A}| > \alpha$, and let x_1 be a solution of the convex quadratic program:

$$\begin{array}{ll} \text{minimize} & \|x\|^2 \\ \text{subject to} & a_i^\mathsf{T} x = b_i, \quad i \in \mathcal{A} \\ & a_i^\mathsf{T} x \le b_i, \quad i \notin \mathcal{A}. \end{array}$$

Note that this problem is feasible because we assume that $a_1, \dots, a_m$ are linearly independent. Additionally, because $|\mathcal{A}| > \alpha$, we have that $\|x_1\| > 1$. Let

$$x(\theta) = \theta x_0 + (1-\theta) x_1.$$

For every $\theta \in [0,1]$, the vector $x(\theta)$ satisfies

$$a_i^\mathsf{T} x(\theta) = \theta(a_i^\mathsf{T} x_0) + (1-\theta)(a_i^\mathsf{T} x_1) \le \theta b_i + (1-\theta) b_i = b_i.$$

Note that the function $f(\theta) = \|x(\theta)\|$ is continuous,

$$f(0) = \|x_0\| < 1, \quad \text{and} \quad f(1) = \|x_1\| > 1.$$

Therefore, the intermediate-value theorem guarantees the existence of a $\hat{\theta} \in [0,1]$ such that $f(\hat{\theta}) = 1$. We can find such a $\hat{\theta}$ efficiently using, for example, bisection on θ. Then $x(\hat{\theta})$ is a solution of (4.7). □

The proof of Theorem 4.3 easily generalizes to the case when there are linear equality constraints.

Theorem 4.4. Consider the feasibility problem

$$\begin{aligned} a_i^\mathsf{T} x \le b_i&, \quad i = 1, \dots, m \\ Cx = d& \\ \|x\| = 1& \end{aligned}$$

with variable $x \in \mathbf{R}^n$, and problem data $a_1, \dots, a_m \in \mathbf{R}^n$, $b \in \mathbf{R}^m$, $C \in \mathbf{R}^{p \times n}$, and $d \in \mathbf{R}^p$. We can assume without loss of generality that the rows of the matrix

$$\begin{bmatrix} a_1^\mathsf{T} \\ \vdots \\ a_m^\mathsf{T} \\ C \end{bmatrix}$$

are linearly independent. Suppose there exists an integer $\alpha < m$ such that $|\mathcal{A}| \leq \alpha$ for every index set $\mathcal{A} \subseteq \{1, \ldots, m\}$ with the property that

$$\{x \in \mathbf{R}^n \mid \|x\| \leq 1,\ Cx = d,\ a_i^\mathsf{T} x = b_i \text{ for } i \in \mathcal{A}\} \neq \emptyset.$$

Then we can compute a solution of the feasibility problem above in polynomial time.

4.4.2 SDP-SOCP relaxation

Adding redundant (also called valid) constraints to an optimization problem does not change the optimal set or optimal value of the problem. Intuitively, the Lagrangian relaxation constructs a lower bound on the optimal value of an optimization problem using linear combinations of the constraints. Therefore, adding redundant constraints may result in tighter relaxations (that is, smaller duality gaps) because there are more constraints that can be used to construct the lower bound. Perhaps the most famous examples of valid inequalities are Gomory cuts, which were introduced by Gomory [42, 43], and used in practical algorithms by Cornuéjols [27, 28]. Adding valid inequalities to (4.6) gives the problem

$$\begin{array}{ll}
\text{minimize} & x^\mathsf{T} P_0 x + 2q_0^\mathsf{T} x + r_0 \\
\text{subject to} & \|(b_i - a_i^\mathsf{T} x)x\| \leq b_i - a_i^\mathsf{T} x, \quad i = 1, \ldots, m \\
& (b_i - a_i^\mathsf{T} x)(b_j - a_j^\mathsf{T} x) \geq 0, \quad 1 \leq i < j \leq m \\
& \|x\| = 1.
\end{array}$$

Due to the constraint $\|x\| = 1$, we have that

$$\|(b_i - a_i^\mathsf{T} x)x\| = |b_i - a_i^\mathsf{T} x|\|x\| = |b_i - a_i^\mathsf{T} x|.$$

Thus, the constraint $\|(b_i - a_i^\mathsf{T} x)x\| = |b_i - a_i^\mathsf{T} x| \leq b_i - a_i^\mathsf{T} x$ is equivalent to the inequality $b_i - a_i^\mathsf{T} x \geq 0$, which can also be written as $a_i^\mathsf{T} x \leq b_i$. However, the more complicated formulation of these constraints gives a tighter relaxation. Similarly, because $a_i^\mathsf{T} x \leq b_i$, the constraints $(b_i - a_i^\mathsf{T} x)(b_j - a_j^\mathsf{T} x) \geq 0$ are redundant; nonetheless, we include these constraints because they result in a tighter relaxation.

We can express the problem with the added valid constraints as

$$\begin{array}{ll}\text{minimize} & \begin{bmatrix} P_0 & q_0 \\ q_0^\mathsf{T} & r_0 \end{bmatrix} \bullet \left(\begin{bmatrix} x \\ 1 \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix}^\mathsf{T} \right) \\ \text{subject to} & \left(\begin{bmatrix} x \\ 1 \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix}^\mathsf{T} \right) \begin{bmatrix} -a_i \\ b_i \end{bmatrix} \in \mathbf{SOC}, \quad i = 1, \dots, m \\ & \begin{bmatrix} -a_i \\ b_i \end{bmatrix}^\mathsf{T} \left(\begin{bmatrix} x \\ 1 \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix}^\mathsf{T} \right) \begin{bmatrix} -a_j \\ b_j \end{bmatrix} \geq 0, \quad 1 \leq i < j \leq m \\ & \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \bullet \left(\begin{bmatrix} x \\ 1 \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix}^\mathsf{T} \right) = 1. \end{array}$$

Since the set of dyads is equal to the set of rank-1, positive-semidefinite matrices, we can rewrite this problem as

$$\begin{array}{ll}\text{minimize} & C \bullet X \\ \text{subject to} & F_i \bullet X = 1, \quad i = 1, 2 \\ & c_i^\mathsf{T} X c_j \geq 0, \quad 1 \leq i < j \leq m \\ & X c_i \in \mathbf{SOC}, \quad i = 1, \dots, m \\ & X \succeq 0 \\ & \mathbf{rank}(X) = 1, \end{array}$$

where we define the matrices $F_1, F_2 \in \mathbf{S}^{n+1}$, and the vectors $c_1, \dots, c_m \in \mathbf{R}^{n+1}$ such that

$$F_1 = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \quad F_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad \text{and} \quad c_i = \begin{bmatrix} -a_i \\ b_i \end{bmatrix}.$$

Ignoring the rank constraint gives an SDP-SOCP relaxation of (4.6):

$$\begin{array}{ll}\text{minimize} & C \bullet X \\ \text{subject to} & F_i \bullet X = 1, \quad i = 1, 2 \\ & A_{ij} \bullet X \geq 0, \quad 1 \leq i < j \leq m \\ & X c_i \in \mathbf{SOC}, \quad i = 1, \dots, m \\ & X \succeq 0, \end{array} \tag{4.8}$$

where $A_{ij} \in \mathbf{S}^{n+1}$ is the symmetric part of $c_i c_j^\mathsf{T}$:

$$A_{ij} = \frac{1}{2}(c_i c_j^\mathsf{T} + c_j c_i^\mathsf{T}).$$

If we can show that (4.8) has a rank-1 solution, then (4.8) is an exact relaxation of (4.6): that is, (4.8) and (4.6) have the same optimal value, and $X = vv^\mathsf{T}$ is a rank-1 solution of (4.8) if and only if $x = v_{n+1}v_{1:n}$ is a solution of (4.6). Our analysis will use the dual of (4.8):

$$\begin{array}{ll} \text{maximize} & \lambda_1 + \lambda_2 \\ \text{subject to} & \sum_{i=1}^2 \lambda_i F_i + \sum_{i=1}^{m-1} \sum_{j=i+1}^m \mu_{ij} A_{ij} \\ & \quad +(1/2) \sum_{i=1}^m (\nu_i c_i^\mathsf{T} + c_i \nu_i^\mathsf{T}) + S = C \\ & \mu_{ij} \geq 0, \quad 1 \leq i < j \leq m \\ & \nu_i \in \mathbf{SOC}, \quad i = 1, \ldots, m \\ & S \succeq 0, \end{array} \tag{4.9}$$

where the variables are $\lambda_i, \mu_{ij} \in \mathbf{R}$, $\nu_i \in \mathbf{R}^{n+1}$, and $S \in \mathbf{S}^n$. The complementarity conditions for (4.8) and (4.9) are

$$\begin{aligned} \mu_{ij}(A_{ij} \bullet X) = 0, &\quad 1 \leq i < j \leq m \\ \nu_i^\mathsf{T} X c_i = 0, &\quad i = 1, \ldots, m \\ S \bullet X = 0. & \end{aligned}$$

Since X and S are positive semidefinite, the last complementarity condition is equivalent to $SX = 0$. The following lemma gives sufficient conditions for (4.8) to be an exact relaxation of (4.6).

Lemma 4.5. Suppose (4.8) and (4.9) are both solvable, and there is no duality gap. Let X and (λ, μ, ν, S) be solutions of (4.8) and (4.9), respectively. If there exists a nonzero vector $z \in \mathbf{range}(X) \cap \mathbf{bd}(\mathbf{SOC})$ such that

(i) $c_i^\mathsf{T} z \geq 0$ for $i = 1, \ldots, m$,

(ii) $\mu_{ij}(c_i^\mathsf{T} z)(c_j^\mathsf{T} z) = 0$ for $1 \leq i < j \leq m$, and

(iii) $(c_i^\mathsf{T} z)(\nu_i^\mathsf{T} z) = 0$ for $i = 1, \ldots, m$,

then (4.8) is an exact relaxation of (4.6).

Proof. Since z is on the boundary of $\mathbf{SOC}$, we have that $\|z_{1:n}\| = z_{n+1}$. Because we also assume that z is nonzero, z_{n+1} must be nonzero, so we can define the matrix

$$\tilde{X} = \frac{1}{z_{n+1}^2} zz^\mathsf{T}.$$

We have that $\mathbf{rank}(\tilde{X}) = 1$ since $\tilde{X}$ is a positive multiple of a dyad. We will show that $\tilde{X}$ is a solution of (4.8) by showing that it is feasible, and satisfies the complementarity conditions. The definitions of F_1 and F_2 imply that

$$\begin{aligned} F_1 \bullet \tilde{X} &= \frac{z^{\mathsf{T}} F_1 z}{z_{n+1}^2} = \frac{\|z_{1:n}\|^2}{z_{n+1}^2} = 1 \\ F_2 \bullet \tilde{X} &= \frac{z^{\mathsf{T}} F_2 z}{z_{n+1}^2} = \frac{z_{n+1}^2}{z_{n+1}^2} = 1. \end{aligned}$$

We have that $\tilde{X}$ is positive semidefinite because it is a positive multiple of a dyad. The assumption that $c_i^{\mathsf{T}} z \geq 0$ for $i = 1, \ldots, m$ guarantees

$$A_{ij} \bullet \tilde{X} = \frac{(c_i^{\mathsf{T}} z)(c_j^{\mathsf{T}} z)}{z_{n+1}^2} \geq 0.$$

Similarly, because $c_i^{\mathsf{T}} z \geq 0$, $z \in \mathbf{SOC}$, and $\mathbf{SOC}$ is closed under nonnegative scaling, we have that

$$\tilde{X} c_i = \frac{c_i^{\mathsf{T}} z}{z_{n+1}^2} z \in \mathbf{SOC}.$$

We have now shown that $\tilde{X}$ is feasible for (4.8).

Our assumption that $\mu_{ij}(c_i^{\mathsf{T}} z)(c_j^{\mathsf{T}} z) = 0$ implies that

$$\mu_{ij}(A_{ij} \bullet \tilde{X}) = \frac{\mu_{ij}(c_i^{\mathsf{T}} z)(c_j^{\mathsf{T}} z)}{z_{n+1}^2} = 0$$

for $1 \leq i < j \leq m$. Similarly, the assumption that $(c_i^{\mathsf{T}} z)(\nu_i^{\mathsf{T}} z) = 0$ guarantees that

$$\nu_i^{\mathsf{T}} \tilde{X} c_i = \frac{(c_i^{\mathsf{T}} z)(\nu_i^{\mathsf{T}} z)}{z_{n+1}^2} = 0$$

for $i = 1, \ldots, m$. Since $z \in \mathbf{range}(X)$, there exists a vector $\tilde{z}$ such that $z = X\tilde{z}$. We have that $SX = 0$ because S and X are solutions of their respective problems, and therefore satisfy complementary slackness. Combining these results, we find that

$$S\tilde{X} = \frac{1}{z_{n+1}^2} S z z^{\mathsf{T}} = \frac{1}{z_{n+1}^2} S X \tilde{z} z^{\mathsf{T}} = 0.$$

Thus, we have shown that $\tilde{X}$ is feasible, and satisfies the complementarity conditions. This implies that $\tilde{X}$ is a solution of (4.8). □

At most one inactive inequality constraint

Consider the optimization problem

$$\begin{array}{ll} \text{minimize} & x^\mathsf{T} P_0 x + 2q_0^\mathsf{T} x + r_0 \\ \text{subject to} & a_i^\mathsf{T} x \le b_i, \quad i = 1, \ldots, m \\ & (b_i - a_i^\mathsf{T} x)(b_j - a_j^\mathsf{T} x) = 0, \quad 1 \le i < j \le m \\ & \|x\| = 1. \end{array} \tag{4.10}$$

Recall that an inequality constraint is said to be active or binding if it holds with equality. Thus, (4.10) requires that at most one inequality constraint is inactive because the constraint $(b_i - a_i^\mathsf{T} x)(b_j - a_j^\mathsf{T} x) = 0$ is satisfied if and only if $a_i^\mathsf{T} x = b_i$ or $a_j^\mathsf{T} x = b_j$. The SDP-SOCP relaxation of (4.10) is

$$\begin{array}{ll} \text{minimize} & C \bullet X \\ \text{subject to} & F_i \bullet X = 1, \quad i = 1, 2 \\ & A_{ij} \bullet X = 0, \quad 1 \le i < j \le m \\ & Xc_i \in \mathbf{SOC}, \quad i = 1, \ldots, m \\ & X \succeq 0. \end{array} \tag{4.11}$$

The corresponding dual relaxation is

$$\begin{array}{ll} \text{maximize} & \lambda_1 + \lambda_2 \\ \text{subject to} & \sum_{i=1}^{2} \lambda_i F_i + \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \mu_{ij} A_{ij} \\ & \quad +(1/2) \sum_{i=1}^{m} (\nu_i c_i^\mathsf{T} + c_i \nu_i^\mathsf{T}) + S = C \\ & \nu_i \in \mathbf{SOC}, \quad i = 1, \ldots, m \\ & S \succeq 0. \end{array} \tag{4.12}$$

The complementarity conditions for these relaxations are

$$\begin{aligned} \nu_i^\mathsf{T} X c_i &= 0, \quad i = 1, \ldots, m \\ S \bullet X &= 0. \end{aligned}$$

Since S and X are positive semidefinite, the last complementarity condition is equivalent to $SX = 0$. We will show that if there is no duality gap between (4.11) and (4.12), then (4.11) is an exact relaxation of (4.10). Our analysis will use the following lemma, which gives sufficient conditions for the SDP-SOCP relaxation to be exact.

Lemma 4.6. Suppose there is no duality gap between (4.11) and (4.12). If there exists a nonzero vector $z \in \mathbf{range}(X) \cap \mathbf{bd}(\mathbf{SOC})$ such that

(i) $c_i^\mathsf{T} z \geq 0$ for $i = 1, \dots, m$,

(ii) $(c_i^\mathsf{T} z)(c_j^\mathsf{T} z) = 0$ for $1 \leq i < j \leq m$, and

(iii) $(c_i^\mathsf{T} z)(\nu_i^\mathsf{T} z) = 0$ for $i = 1, \dots, m$,

then (4.11) is an exact relaxation of (4.10).

Proof. Let X be a solution of (4.11), and (λ, μ, ν, S) be a solution of (4.12). In addition to satisfying the constraints of their respective problems, these solutions must satisfy the complementarity conditions

$$SX = 0 \quad \text{and} \quad \nu_i^\mathsf{T} X c_i = 0, \quad i = 1, \dots, m.$$

Suppose z satisfies the hypotheses above. Because $z \in \mathbf{bd}(\mathbf{SOC})$, we have that $\|z_{1:n}\| = z_{n+1}$. Since, in addition, z is assumed to be nonzero, this implies that z_{n+1} is nonzero. Therefore, we can define the matrix

$$\tilde{X} = \frac{1}{z_{n+1}^2} z z^\mathsf{T}.$$

The definitions of F_1 and F_2, imply that

$$\begin{aligned} F_1 \bullet \tilde{X} &= \frac{z^\mathsf{T} F_1 z}{z_{n+1}^2} = \frac{\|z_{1:n}\|^2}{z_{n+1}^2} = 1, \\ F_2 \bullet \tilde{X} &= \frac{z^\mathsf{T} F_2 z}{z_{n+1}^2} = \frac{z_{n+1}^2}{z_{n+1}^2} = 1. \end{aligned}$$

Our assumption that $(c_i^\mathsf{T} z)(c_j^\mathsf{T} z) = 0$ for $1 \leq i < j \leq m$ implies that

$$A_{ij} \bullet \tilde{X} = \frac{(c_i^\mathsf{T} z)(c_j^\mathsf{T} z)}{z_{n+1}^2} = 0.$$

Since $z \in \mathbf{SOC}$, $c_i^\mathsf{T} z \geq 0$, and $\mathbf{SOC}$ is closed under nonnegative scaling, we have that

$$\tilde{X} c_i = \frac{c_i^\mathsf{T} z}{z_{n+1}^2} z \in \mathbf{SOC}.$$

Note that $\tilde{X}$ is positive semidefinite because it is a positive multiple of a dyad. We have now shown that $\tilde{X}$ satisfies the constraints of (4.11);

it remains to check that $\tilde{X}$ satisfies the complementarity conditions. Our assumption that $(c_i^\mathsf{T} z)(\nu_i^\mathsf{T} z) = 0$ implies that

$$\nu_i^\mathsf{T} \tilde{X} c_i = \frac{(c_i^\mathsf{T} z)(\nu_i^\mathsf{T} z)}{z_{n+1}^2} = 0.$$

Finally, because $z \in \mathbf{range}(X)$, we have that $z = X\tilde{z}$ for some vector $\tilde{z}$, and hence that

$$S\tilde{X} = \frac{1}{z_{n+1}^2} S z z^\mathsf{T} = \frac{1}{z_{n+1}^2} S X \tilde{z} z^\mathsf{T} = 0,$$

where we use the fact that $SX = 0$. Having shown that $\tilde{X}$ is feasible, and satisfies the complementarity conditions, we can conclude that $\tilde{X}$ is optimal. Moreover, it is clear that $\mathbf{rank}(\tilde{X}) = 1$ because $\tilde{X}$ is a scalar multiple of a dyad. □

We are now prepared to show that (4.11) is an exact relaxation of (4.10). We will rely heavily on Lemma 4.6.

Theorem 4.7. If there is no duality gap between (4.11) and (4.12), then (4.11) has a rank-1 solution, which implies that (4.11) is an exact relaxation of (4.10).

Proof. Let X be a solution of (4.11), and (λ, μ, ν, S) be a solution of (4.12). We will show how to construct a rank-1 solution of (4.11). There are three cases to consider.

(1) First, suppose $Xc_1 = \cdots = Xc_m = 0$. Let $r = \mathbf{rank}(X)$, and use Lemma A.10 to find vectors $z_1, \dots, z_r$ such that

$$X = \sum_{j=1}^{r} z_i z_i^\mathsf{T} \quad \text{and} \quad z_j^\mathsf{T}(F_1 - F_2) z_j = \frac{(F_1 - F_2) \bullet X}{r} = 0,$$

where $(F_1 - F_2) \bullet X = (F_1 \bullet X) - (F_2 \bullet X) = 1 - 1 = 0$ because X is feasible for (4.11). Then we have that $z_j \in \mathbf{bd}(\mathbf{SOC})$ because

$$z^\mathsf{T}(F_1 - F_2) z_j = \|(z_j)_{1:n}\|^2 - (z_j)_{n+1}^2 = 0.$$

Note that $z_1, \dots, z_r$ must be linearly independent since they form a dyadic decomposition of the rank-r matrix X. This implies that the z_j are all nonzero. Moreover, the fact that

$$Xc_i = \sum_{j=1}^{r} (c_i^\mathsf{T} z_j) z_j = 0$$

implies that $c_i^\mathsf{T} z_j = 0$ for all i and j. We have that $z_j \in \mathbf{range}(X)$ because the z_j form a dyadic decomposition for X. To summarize, we have shown that the z_j are nonzero vectors contained in $\mathbf{range}(X) \cap \mathbf{bd}(\mathbf{SOC})$ such that $c_i^\mathsf{T} z_j = 0$ for all i and j. This implies that each of the z_j satisfies the hypotheses of Lemma 4.6, and hence that (4.11) is an exact relaxation of (4.10).

(2) Next, consider the case when at least one of the Xc_i is a nonzero vector on the boundary of $\mathbf{SOC}$. We can assume without loss of generality that Xc_1 is such a vector. Let $z = Xc_1$. We have that $c_1^\mathsf{T} z = c_1^\mathsf{T} X c_1 \geq 0$ because $X \succeq 0$; we also have that

$$c_j^\mathsf{T} z = c_1^\mathsf{T} X c_j = A_{1j} \bullet X = 0$$

for $j = 2, \ldots, m$ because X is feasible for (4.11). This implies that $(c_i^\mathsf{T} z)(c_j^\mathsf{T} z) = 0$ for $1 \leq i < j \leq m$. Complementarity requires that $\nu_1^\mathsf{T} X c_1 = 0$, and hence that $(c_1^\mathsf{T} z)(\nu_1^\mathsf{T} z) = (c_1^\mathsf{T} z)(\nu_1^\mathsf{T} X c_1) = 0$. Having already shown that $c_j^\mathsf{T} z = 0$, we can conclude that $(c_j^\mathsf{T} z)(\nu_j^\mathsf{T} z) = 0$ for $j = 2, \ldots, m$. Since X and S are solutions of their respective problems, they satisfy the complementary-slackness condition $SX = 0$; this implies that $Sz = SXc_1 = 0$. We have now shown that z satisfies the hypotheses of Lemma 4.6, and hence that (4.11) is an exact relaxation of (4.10).

(3) Finally, consider the case when at least one of the Xc_i is contained in the interior of $\mathbf{SOC}$. Without loss of generality, we can assume that Xc_1 is in the interior of $\mathbf{SOC}$. Define the matrix $P \in \mathbf{S}^{n+1}$ such that

$$P = \sum_{i=1}^{m} (X^{\frac{1}{2}} c_i)(X^{\frac{1}{2}} c_i)^\dagger.$$

In our expression for P, we take the pseudoinverse of column vectors; recall that the pseudoinverse of a column vector q is given by

$$q^{\dagger} = \begin{cases} (1/\|q\|^2)q^{\mathsf{T}} & q \neq 0, \\ 0 & q = 0. \end{cases}$$

Note that Xc_1 is nonzero because it is in the interior of $\mathbf{SOC}$. Thus, we have that

$$\begin{aligned} P = {} & \frac{1}{c_1^{\mathsf{T}} X c_1}(X^{\frac{1}{2}}c_1)(X^{\frac{1}{2}}c_1)^{\mathsf{T}} \\ & + \sum_{i=2}^{m} \begin{cases} (1/(c_i^{\mathsf{T}} X c_i))(X^{\frac{1}{2}}c_i)(X^{\frac{1}{2}}c_i)^{\mathsf{T}} & X^{\frac{1}{2}}c_i \neq 0, \\ 0 & X^{\frac{1}{2}}c_i = 0. \end{cases} \end{aligned}$$

Define the matrix $Z \in \mathbf{S}^{n+1}$ such that

$$\begin{aligned} Z &= X^{\frac{1}{2}}(I - P)X^{\frac{1}{2}} \\ &= X - \frac{1}{c_1^{\mathsf{T}} X c_1}(Xc_1)(Xc_1)^{\mathsf{T}} \\ &\qquad - \sum_{i=2}^{m} \begin{cases} (1/(c_i^{\mathsf{T}} X c_i))(Xc_i)(Xc_i)^{\mathsf{T}} & X^{\frac{1}{2}}c_i \neq 0, \\ 0 & X^{\frac{1}{2}}c_i = 0. \end{cases} \end{aligned}$$

Because $Xc_i \in \mathbf{SOC}$, we have that

$$\begin{aligned} (F_1 - F_2) \bullet ((Xc_i)(Xc_i)^{\mathsf{T}}) &= (Xc_i)^{\mathsf{T}} F_1 (Xc_i) - (Xc_i)^{\mathsf{T}} F_2 (Xc_i) \\ &= \|(Xc_i)_{1:n}\|^2 - (Xc_i)_{n+1}^2 \\ &\leq 0. \end{aligned}$$

Moreover, this inequality is strict for $i = 1$ because Xc_1 is in the interior of $\mathbf{SOC}$. Combined with the fact that $(F_1 - F_2) \bullet X = 0$ since X is feasible, this allows us to conclude that

$$\begin{aligned} & (F_1 - F_2) \bullet Z \\ & \quad = (F_1 - F_2) \bullet X - \frac{(F_1 - F_2) \bullet ((Xc_1)(Xc_1)^{\mathsf{T}})}{c_1^{\mathsf{T}} X c_1} \\ & \qquad - \begin{cases} ((F_1 - F_2) \bullet ((Xc_i)(Xc_i)^{\mathsf{T}}))/(c_i^{\mathsf{T}} X c_i) & X^{\frac{1}{2}}c_i \neq 0, \\ 0 & X^{\frac{1}{2}}c_i = 0. \end{cases} \\ & \quad > 0. \end{aligned}$$

This implies that Z is nonzero, and hence that $s = \mathbf{rank}(Z) > 0$. Use Lemma A.10 to find vectors $u_1, \dots, u_s$ such that

$$Z = \sum_{j=1}^{s} u_j u_j^\top \quad \text{and} \quad u_j^\top (F_1 - F_2) u_j = \frac{(F_1 - F_2) \bullet Z}{s} > 0.$$

Then our choice of $u_1, \dots, u_s$ implies that

$$u_j^\top (F_1 - F_2) u_j = \|(u_j)_{1:n}\|^2 - (u_j)_{n+1}^2 > 0,$$

so that $u_j \notin \mathbf{SOC}$. We will use u_1 to construct a vector z satisfying the hypotheses of Lemma 4.6; however, this choice is somewhat arbitrary: any of the u_j would serve as well. Because Xc_1 is in the interior of $\mathbf{SOC}$, and u_1 is not contained in $\mathbf{SOC}$, there exists $\theta \in (0, 1)$ such that

$$z = \theta X c_1 + (1 - \theta) u_1$$

is on the boundary of $\mathbf{SOC}$. Suppose z is equal to zero. Then we have that $u_1 = -(\theta/(1-\theta)) X c_1$, and hence that

$$\begin{aligned} u_1^\top (F_1 - F_2) u_1 &= \left(\frac{\theta}{1-\theta}\right)^2 (Xc_1)^\top (F_1 - F_2)(Xc_1) \\ &= \left(\frac{\theta}{1-\theta}\right)^2 (\|(Xc_1)_{1:n}\|^2 - (Xc_1)_{n+1}^2) \\ &< 0, \end{aligned}$$

where we use the fact that $\|(Xc_1)_{1:n}\|^2 < (Xc_1)_{n+1}^2$ because Xc_1 is in the interior of $\mathbf{SOC}$. This contradicts the fact that u_1 was chosen such that $u_1^\top (F_1 - F_2) u_1 > 0$, and thereby proves that z must be nonzero. We have that $u_1, \dots, u_s$ form a basis for $\mathbf{range}(Z) \subset \mathbf{range}(X^{\frac{1}{2}}) = \mathbf{range}(X)$. Thus, we have that $u_1 \in \mathbf{range}(X)$. Since we also have that $Xc_1 \in \mathbf{range}(X)$, we can conclude that

$$z = \theta X c_1 + (1 - \theta) u_1 \in \mathbf{range}(X).$$

For a column vector q, we think of $qq^\dagger$ as the projection onto q. We have that $X^{\frac{1}{2}} c_i$ and $X^{\frac{1}{2}} c_j$ are orthogonal because

$$(X^{\frac{1}{2}} c_i)^\top (X^{\frac{1}{2}} c_j) = c_i^\top X c_j = A_{ij} \bullet X = 0.$$

This implies that P is the orthogonal projection onto the subspace $\mathbf{span}(X^{\frac{1}{2}}c_1, \ldots, X^{\frac{1}{2}}c_m)$. Therefore, we have that $PX^{\frac{1}{2}}c_i = X^{\frac{1}{2}}c_i$, and hence that

$$\begin{aligned} Zc_i &= X^{\frac{1}{2}}(I-P)X^{\frac{1}{2}}c_i \\ &= X^{\frac{1}{2}}(X^{\frac{1}{2}}c_1 - PX^{\frac{1}{2}}c_i) \\ &= X^{\frac{1}{2}}(X^{\frac{1}{2}}c_i - X^{\frac{1}{2}}c_i) \\ &= 0. \end{aligned}$$

Note that $u_1, \ldots, u_s$ are linearly independent because they form a dyadic expansion of the rank-s matrix Z. Then the fact that

$$Zc_i = \sum_{j=1}^{s} (c_i^\mathsf{T} u_j) u_j = 0$$

implies that $c_i^\mathsf{T} u_j = 0$ for all i and j. Since $X \succeq 0$, we have that $c_1^\mathsf{T} X c_1 \geq 0$. Combining these results, we find that

$$c_1^\mathsf{T} z = \theta c_1^\mathsf{T} X c_1 + (1-\theta) c_1^\mathsf{T} u_1 = \theta c_1^\mathsf{T} X c_1 \geq 0.$$

Similarly, we have that

$$c_j^\mathsf{T} z = \theta c_1^\mathsf{T} X c_j + (1-\theta) c_j^\mathsf{T} u_1 = \theta(A_{1j} \bullet X) = 0$$

for $j = 2, \ldots, m$. Taken together, these results imply that $c_i^\mathsf{T} z \geq 0$ for $i = 1, \ldots, m$, $(c_i^\mathsf{T} z)(c_j^\mathsf{T} z) = 0$ for $1 \leq i < j \leq m$, and $(c_i^\mathsf{T} z)(\nu_i^\mathsf{T} z) = 0$ for $i = 2, \ldots, m$. Since Xc_1 is in the interior of **SOC**, complementarity implies that $\nu_1 = 0$, and hence that $(c_1^\mathsf{T} z)(\nu_1^\mathsf{T} z) = 0$. We have now shown that z satisfies the hypotheses of Lemma 4.6, which proves that (4.11) is an exact relaxation of (4.10).

The cases above are collectively exhaustive, and therefore suffice to show that (4.11) is an exact relaxation of (4.10). □

Non-intersecting pairs of linear constraints

Now we consider an instance of (4.6) with two non-intersecting linear constraints. When we say the linear constraints of (4.6) do not intersect, we mean that there is no vector x such that $\|x\| \leq 1$ and $a_i^\mathsf{T} x = b_i$ for $i = 1, 2$. The results in this section are due to Burer and Anstreicher [13], and Burer and Yang [19].

Theorem 4.8. Suppose $m = 2$, (4.8) and (4.9) are both solvable, and there is no duality gap. If

$$\{x \in \mathbf{R}^n \mid a_1^\mathsf{T} x = b_1,\ a_2^\mathsf{T} x = b_2,\ \|x\| \leq 1\} = \emptyset,$$

then (4.8) is an exact relaxation of (4.6).

Proof. Let X be a solution of (4.8), and (λ, μ, ν, S) be a solution of (4.9). If $A_{12} \bullet X = 0$, then we can find a rank-1 solution of (4.8) using the construction in the proof of Theorem 4.7. Thus, we only need to consider the case when $A_{12} \bullet X > 0$. In this case complementarity implies that $\mu_{12} = 0$, and hence that we can use Lemma 4.5 to deduce that there exists a rank-1 solution of (4.8) if we can find a nonzero $z \in \mathbf{range}(X) \cap \mathbf{bd}(\mathbf{SOC})$ such that

$$c_1^\mathsf{T} z, c_2^\mathsf{T} z \geq 0 \quad \text{and} \quad (c_1^\mathsf{T} z)(\nu_1^\mathsf{T} z) = (c_2^\mathsf{T} z)(\nu_2^\mathsf{T} z) = 0.$$

Because we are considering the case when $c_1^\mathsf{T} X c_2 = A_{12} \bullet X > 0$, we have that Xc_1 and Xc_2 are nonzero. We divide the analysis into cases.

(1) First, consider the case when Xc_1 and Xc_2 are linearly dependent. Since Xc_1 and Xc_2 are nonzero, this implies that there exists a nonzero scalar α such that $Xc_1 = \alpha X c_2$. Moreover, α must be positive because $Xc_1, Xc_2 \in \mathbf{SOC}$.

(1)(a) Suppose $Xc_1 \in \mathbf{bd}(\mathbf{SOC})$. Then we claim that the vector $z = Xc_1 = \alpha X c_2$ has the desired properties. Because X is positive semidefinite, we have that $c_1^\mathsf{T} z = c_1^\mathsf{T} X c_1 \geq 0$. Similarly, we have that $c_2^\mathsf{T} z = \alpha(c_2^\mathsf{T} X c_2) \geq 0$ since $X \succeq 0$ and $\alpha > 0$. Because X and ν_1 are optimal, they satisfy the complementarity condition $\nu_1^\mathsf{T} X c_1 = 0$. Thus, z satisfies

$$\nu_1^\mathsf{T} z = \nu_1^\mathsf{T} X c_1 = 0 \quad \text{and} \quad \nu_2^\mathsf{T} z = \alpha(\nu_2^\mathsf{T} X c_2) = 0.$$

This completes the proof that z satisfies the hypotheses of Lemma 4.5, and hence that (4.8) is an exact relaxation of (4.6).

(1)(b) Next, suppose $Xc_1 \in \mathbf{int}(\mathbf{SOC})$. Then we also have that $Xc_2 = (1/\alpha)Xc_1 \in \mathbf{int}(\mathbf{SOC})$, and hence that $\nu_1 = \nu_2 = 0$ due to complementarity. The orthogonal projection onto the subspace $\mathbf{span}(X^{\frac{1}{2}}c_1) = \mathbf{span}(X^{\frac{1}{2}}c_2)$ is represented by the matrix

$$\begin{aligned} P &= (X^{\frac{1}{2}}c_1)((X^{\frac{1}{2}}c_1)^{\mathsf{T}}(X^{\frac{1}{2}}c_1))^{-1}(X^{\frac{1}{2}}c_1)^{\mathsf{T}} \\ &= \frac{1}{c_1^{\mathsf{T}}Xc_1}(X^{\frac{1}{2}}c_1)(X^{\frac{1}{2}}c_1)^{\mathsf{T}}. \end{aligned}$$

Define the matrix

$$Z = X^{\frac{1}{2}}(I-P)X^{\frac{1}{2}} = X - \frac{1}{c_1^{\mathsf{T}}Xc_1}(Xc_1)(Xc_1)^{\mathsf{T}}.$$

Since $Xc_1 \in \mathbf{int}(\mathbf{SOC})$, we have that

$$\begin{aligned} (F_1 - F_2) \bullet ((Xc_1)(Xc_1)^{\mathsf{T}}) &= (Xc_1)^{\mathsf{T}}(F_1 - F_2)(Xc_1) \\ &= \|(Xc_1)_{1:n}\|^2 - (Xc_1)_{n+1}^2 \\ &< 0. \end{aligned}$$

Combined with the fact that $(F_1 - F_2) \bullet X = 0$ because X is feasible for (4.8), this allows us to conclude that

$$\begin{aligned} (F_1 - F_2) \bullet Z &= (F_1 - F_2) \bullet X \\ &\quad - \frac{1}{c_1^{\mathsf{T}}Xc_1}(F_1 - F_2) \bullet ((Xc_1)(Xc_1)^{\mathsf{T}}) \\ &> 0. \end{aligned}$$

This implies that Z is nonzero. Let $s = \mathbf{rank}(Z) > 0$, and use Lemma A.10 to find $z_1, \ldots, z_s$ such that

$$Z = \sum_{i=1}^{s} z_i z_i^{\mathsf{T}} \quad \text{and} \quad z_i^{\mathsf{T}}(F_1 - F_2)z_i = \frac{(F_1 - F_2) \bullet Z}{s} > 0.$$

We have that $z_i \notin \mathbf{SOC}$ because

$$z_i^{\mathsf{T}}(F_1 - F_2)z_i = \|(z_i)_{1:n}\|^2 - (z_i)_{n+1}^2 > 0.$$

Since $Xc_1 \in \mathbf{int}(\mathbf{SOC})$ and $z_i \notin \mathbf{SOC}$, there exists $\theta \in (0,1)$ such that

$$z = \theta Xc_1 + (1-\theta)z_1 \in \mathbf{bd}(\mathbf{SOC}).$$

Suppose $z = 0$. Then we have that

$$z_1 = -\frac{\theta}{1-\theta} X c_1,$$

and hence that

$$z_1^\mathsf{T}(F_1 - F_2) z_1 = \left(\frac{\theta}{1-\theta}\right)^2 (Xc_1)^\mathsf{T}(F_1 - F_2)(Xc_1) < 0.$$

This contradicts our choice of z_1, and thereby proves that z is nonzero. We have that $PX^{\frac{1}{2}}c_i = X^{\frac{1}{2}}c_i$ because P is the projection onto $\mathbf{span}(X^{\frac{1}{2}}c_1) = \mathbf{span}(X^{\frac{1}{2}}c_2)$. This implies that

$$\begin{aligned} Zc_i &= X^{\frac{1}{2}}(I - P)X^{\frac{1}{2}}c_i \\ &= X^{\frac{1}{2}}(X^{\frac{1}{2}}c_i - PX^{\frac{1}{2}}c_i) \\ &= X^{\frac{1}{2}}(X^{\frac{1}{2}}c_i - X^{\frac{1}{2}}c_i) \\ &= 0. \end{aligned}$$

Therefore, we have that

$$Zc_i = \sum_{j=1}^{s} (c_i^\mathsf{T} z_j) z_j = 0,$$

and hence that $c_i^\mathsf{T} z_j = 0$ because $z_1, \ldots, z_s$ are linearly independent because they form a dyadic expansion of the rank-s matrix Z. Since X is positive semidefinite, we have that $c_i^\mathsf{T} X c_i \geq 0$ for $i = 1, 2$. Combined with the facts that $c_i^\mathsf{T} z_j = 0$ and $\alpha > 0$, this allows us to conclude that

$$\begin{aligned} c_1^\mathsf{T} z &= \theta(c_1^\mathsf{T} X c_1) + (1-\theta)(c_1^\mathsf{T} z_1) \geq 0, \\ c_2^\mathsf{T} z &= \alpha\theta(c_2^\mathsf{T} X c_2) + (1-\theta)(c_2^\mathsf{T} z_1) \geq 0. \end{aligned}$$

This completes the proof that z satisfies the hypotheses of Lemma 4.5, and hence that (4.8) is an exact relaxation of (4.6).

(2) Now consider the case when Xc_1 and Xc_2 are linearly independent.

(2)(a) Suppose one of Xc_1 and Xc_2 is on the boundary of **SOC**, and the other is in the interior of **SOC**. Without loss of generality, assume that Xc_1 is on the boundary of **SOC**, and Xc_2

is in the interior of **SOC**. We claim that $z = Xc_1$ satisfies the hypotheses of Lemma 4.5. We have already shown that Xc_1 is nonzero. It is clear that $Xc_1 \in \mathbf{range}(X)$, and we assume that $Xc_1 \in \mathbf{bd}(\mathbf{SOC})$. Since $X \succeq 0$, we have that $c_1^\mathsf{T} z = c_1^\mathsf{T} X c_1 \geq 0$. Similarly, because X is feasible, we find that

$$c_2^\mathsf{T} z = c_1^\mathsf{T} X c_2 = A_{12} \bullet X \geq 0.$$

Since X and ν_1 satisfy complementarity, we have that

$$\nu_1^\mathsf{T} z = \nu_1^\mathsf{T} X c_1 = 0.$$

Because $Xc_2 \in \mathbf{int}(\mathbf{SOC})$, we have that $\nu_2 = 0$, and hence that $\nu_2^\mathsf{T} z = 0$. This completes the proof that z satisfies the hypotheses of Lemma 4.5.

(2)(b) Now consider the case when $Xc_1, Xc_2 \in \mathbf{int}(\mathbf{SOC})$. Let P be the projection onto $\mathbf{span}(X^{\frac{1}{2}}c_1, X^{\frac{1}{2}}c_2)$, and define the matrix $Z = X^{\frac{1}{2}}(I - P)X^{\frac{1}{2}}$.

(2)(b)(i) Suppose $(F_1 - F_2) \bullet Z \geq 0$. Use Lemma A.10 to find $z_1, \ldots, z_s$ such that

$$Z = \sum_{i=1}^{s} z_i z_i^\mathsf{T} \quad \text{and} \quad z_i^\mathsf{T}(F_1 - F_2) z_i = \frac{(F_1 - F_2) \bullet Z}{s} \geq 0.$$

Because P is the projection onto $\mathbf{span}(X^{\frac{1}{2}}c_1, X^{\frac{1}{2}}c_2)$, we have that $PX^{\frac{1}{2}}c_i = X^{\frac{1}{2}}c_i$, and hence that

$$Zc_i = \sum_{j=1}^{s} (c_i^\mathsf{T} z_j) z_j = X^{\frac{1}{2}}(I - P)X^{\frac{1}{2}}c_i = 0$$

for $i = 1, 2$. Since $z_1, \ldots, z_s$ are linearly independent, this implies that $c_i^\mathsf{T} z_j = 0$ for $i = 1, 2$ and $j = 1, \ldots, s$. Because z_1 is nonzero and contained in **SOC**, we have that

$$x = \frac{1}{(z_1)_{n+1}} (z_1)_{1:n}$$

satisfies $\|x\| \leq 1$. Recall that we defined $c_i = (-a_i, b_i)$. Thus, we have that

$$c_i^\mathsf{T} z_1 = b_i (z_1)_{n+1} - a_i^\mathsf{T} (z_1)_{1:n} = 0,$$

and hence that

$$a_i^\mathsf{T} x = \frac{a_i^\mathsf{T}(z_1)_{1:n}}{(z_1)_{n+1}} = b_i.$$

Combining these observations, we see that we have found a vector x such that $a_1^\mathsf{T} x = b_1$, $a_2^\mathsf{T} x = b_2$, and $\|x\| \leq 1$. This violates the non-intersection assumption, so this case cannot happen.

(2)(b)(ii) Now consider the case when $(F_1 - F_2) \bullet Z < 0$. Use Lemma A.10 to find $z_1, \ldots, z_s$ such that $Z = \sum_{i=1}^s z_i z_i^\mathsf{T}$, and $z_i^\mathsf{T}(F_1 - F_2)z_i = ((F_1 - F_2) \bullet Z)/s$, where $s = \mathbf{rank}(Z)$. Then we have that $Xc_2 \in \mathbf{int}(\mathbf{SOC})$, and $z_1 \notin \mathbf{SOC}$, and we can use the construction in (1)(b) to find a rank-1 solution of the SDP-SOCP relaxation.

(2)(c) Finally, consider the case when $Xc_1, Xc_2 \in \mathbf{bd}(\mathbf{SOC})$. Let P be the projection onto $\mathbf{span}(X^{\frac{1}{2}}c_1, X^{\frac{1}{2}}c_2)$, and define the matrix $Z = X^{\frac{1}{2}}(I - P)X^{\frac{1}{2}}$. Since X is feasible, it satisfies the constraint

$$c_1^\mathsf{T} X c_2 = (X^{\frac{1}{2}}c_1)^\mathsf{T}(X^{\frac{1}{2}}c_2) = 0.$$

Thus, we have that $X^{\frac{1}{2}}c_1$ and $X^{\frac{1}{2}}c_2$ are orthogonal, which implies that the projection matrix P is given by

$$\begin{aligned} P &= \sum_{i=1}^2 (X^{\frac{1}{2}}c_i)(X^{\frac{1}{2}}c_i)^\dagger \\ &= \sum_{i=1}^2 \frac{1}{c_i^\mathsf{T} X c_i}(X^{\frac{1}{2}}c_i)(X^{\frac{1}{2}}c_i)^\mathsf{T}. \end{aligned}$$

Using this expression for P, we find that

$$\begin{aligned} Z &= X^{\frac{1}{2}}(I - P)X^{\frac{1}{2}} \\ &= X - X^{\frac{1}{2}} P X^{\frac{1}{2}} \\ &= X - X^{\frac{1}{2}} \left(\sum_{i=1}^2 \frac{1}{c_i^\mathsf{T} X c_i}(X^{\frac{1}{2}}c_i)(X^{\frac{1}{2}}c_i)^\mathsf{T} \right) X^{\frac{1}{2}} \\ &= X - \sum_{i=1}^2 \frac{1}{c_i^\mathsf{T} X c_i}(Xc_i)(Xc_i)^\mathsf{T}. \end{aligned}$$

Because Xc_1 and Xc_2 are on the boundary of the second-order cone, we have that

$$(Xc_i)^\mathsf{T}(F_1 - F_2)(Xc_i) = \|(Xc_i)_{1:n}\|^2 - (Xc_i)_{n+1}^2 = 0$$

for $i = 1, 2$. Additionally, we have that $(F_1 - F_2) \bullet X = 0$ since X is feasible. Combining these results, we find that

$$\begin{aligned}(F_1 - F_2) \bullet Z &= (F_1 - F_2) \bullet \left(X - \sum_{i=1}^{2} \frac{1}{c_i^\mathsf{T} X c_i} (Xc_i)(Xc_i)^\mathsf{T} \right) \\ &= (F_1 - F_2) \bullet X - \sum_{i=1}^{2} \frac{(Xc_i)^\mathsf{T}(F_1 - F_2)(Xc_i)}{c_i^\mathsf{T} X c_i} \\ &= 0.\end{aligned}$$

Thus, we have that $(F_1 - F_2) \bullet Z = 0$, and we can use the construction in (2)(b)(i) to show that the non-intersection assumption is violated. Therefore, this case cannot happen.

□

Burer and Anstreicher [13] give an example showing that the SDP relaxation of an instance of (4.6) with $m = 2$ may not be exact if the linear constraints intersect inside the unit ball.

Non-intersecting sets of linear constraints

Burer and Yang [19] extended Theorem 4.8 to the case of an arbitrary number of linear constraints such that no two constraints intersect inside the unit ball.

Theorem 4.9. Suppose (4.8) and (4.9) are both solvable, and there is no duality gap. If

$$\{x \in \mathbf{R}^n \mid a_i^\mathsf{T} x = b_i,\ a_j^\mathsf{T} x = b_j,\ \|x\| \le 1\} = \emptyset$$

for all distinct i and j in $\{1, \dots, m\}$, then (4.8) is an exact relaxation of (4.6).

Bienstock and Michalka [8] showed that (4.6) can be solved efficiently if the linear constraints satisfy the non-intersection condition given in the following theorem.

Theorem 4.10. The problem (4.6) can be solved in polynomial time if there exists a positive integer $k > 1$ with the property that

$$\{x \in \mathbf{R}^n \mid \|x\| \leq 1,\ a_i^\mathsf{T} x \leq b_i \text{ for all } i \in \mathcal{I}\} = \emptyset$$

for all index sets $\mathcal{I} \subset \{1, \ldots, m\}$ such that $|\mathcal{I}| = k$.

4.5 Ellipsoidal quadratic inequality constraints

Consider an instance of (4.1) with $m = 0$:

$$\begin{array}{ll} \text{minimize} & x^\mathsf{T} P_0 x + 2q_0^\mathsf{T} x + r_0 \\ \text{subject to} & x^\mathsf{T} P_i x + 2q_i^\mathsf{T} x + r_i \leq 0, \quad i = 1, \ldots, p \\ & \|x\| = 1. \end{array} \tag{4.13}$$

When there are no linear inequality constraints, (4.2) simplifies to

$$\begin{array}{ll} \text{minimize} & Q_0 \bullet X \\ \text{subject to} & Q_i \bullet X \leq 0, \quad i = 1, \ldots, p \\ & F_i \bullet X = 1, \quad i = 1, 2 \\ & X \succeq 0. \end{array} \tag{4.14}$$

Similarly, (4.3) simplifies to

$$\begin{array}{ll} \text{minimize} & \nu_1 + \nu_2 \\ \text{subject to} & \sum_{i=1}^2 \nu_i F_i - \sum_{i=1}^p \mu_i Q_i + S = Q_0 \\ & \mu \geq 0 \\ & S \succeq 0. \end{array} \tag{4.15}$$

4.5.1 Inactive quadratic constraints

The following result was given by Ye and Zhang [102].

Theorem 4.11. If $p = 1$, and there exists a solution X of (4.14) such that the quadratic inequality constraint is inactive at X (that is, $Q_1 \bullet X < 0$)), then (4.14) is an exact relaxation of (4.13).

Proof. Suppose $p = 1$, X is a solution of (4.14) such that $Q_1 \bullet X < 0$, and (μ, ν, S) is a solution of (4.15). The fact that $Q_1 \bullet X < 0$ implies that X is nonzero, and hence that $r = \mathbf{rank}(X) > 0$. Use Lemma A.10 to find $z_1, \ldots, z_r \in \mathbf{R}^n$ such that

$$X = \sum_{i=1}^{r} z_i z_i^{\mathsf{T}} \quad \text{and} \quad z_i^{\mathsf{T}}(F_1 - F_2) z_i = \frac{(F_1 - F_2) \bullet X}{r} = 0,$$

where $(F_1 - F_2) \bullet X = 0$ because X is feasible for (4.14). We have that

$$Q_1 \bullet X = Q_1 \bullet \left(\sum_{i=1}^{r} z_i z_i^{\mathsf{T}} \right) = \sum_{i=1}^{r} z_i^{\mathsf{T}} Q_1 z_i < 0,$$

which implies that $z_i^{\mathsf{T}} Q_1 z_i < 0$ for some i. Without loss of generality, assume that $z_1^{\mathsf{T}} Q_1 z_1 < 0$. We claim that

$$z = \mathbf{sign}((z_1)_{n+1}) z_1$$

satisfies the hypotheses of Lemma 4.1. We have that $z \in \mathbf{range}(X)$ because $z_1, \ldots, z_r$ form a basis for $\mathbf{range}(X)$. We chose z_1 such that

$$z_1^{\mathsf{T}}(F_1 - F_2) z_1 = \|(z_1)_{1:n}\|^2 - (z_1)_{n+1}^2 = 0,$$

which implies that z is on the boundary of $\mathbf{SOC}$. Since we have that $z^{\mathsf{T}} Q_1 z = z_1^{\mathsf{T}} Q_1 z_1 < 0$, the vector z is nonzero, and $z^{\mathsf{T}} Q_1 z \leq 0$. Complementarity requires that $\mu_1 = 0$ because $Q_1 \bullet X < 0$, and hence that $\mu_1(z^{\mathsf{T}} Q_1 z) = 0$. We have now checked that z satisfies all of the hypotheses of Lemma 4.1. Thus, we can conclude that (4.14) is an exact relaxation of (4.13). □

Burer and Anstreicher [13] showed that (4.14) may not be an exact relaxation if there does not exist a solution for which the quadratic inequality constraint is strictly satisfied.

Convex quadratic constraints

We can also show that (4.14) is an exact relaxation of (4.13) when $Q_1, \ldots, Q_p$ are positive semidefinite. The characterization of positive-semidefinite block matrices given in Corollary A.12 tells us that the block matrix Q_i is positive semidefinite if and only if P_i is positive semidefinite, $q_i \in \mathbf{range}(P_i)$, and $r_i - q_i^{\mathsf{T}} P_i^{\dagger} q_i \geq 0$. Thus, requiring that $Q_1, \ldots, Q_p$ be positive semidefinite is a stronger condition than requiring that (4.13) be a convex optimization problem.

Theorem 4.12. Suppose $Q_1, \dots, Q_p \succeq 0$. If (4.14) and (4.15) are both solvable, and there is no duality gap, then (4.14) is an exact relaxation of (4.13).

Proof. Suppose $Q_1, \dots, Q_p$ are positive semidefinite, X is a solution of (4.14), and (μ, ν, S) is a solution of (4.15). Since X and Q_i are positive semidefinite, Lemma A.1 tells us that $Q_i \bullet X \geq 0$. We also have that $Q_i \bullet X \leq 0$ for $i = 1, \dots, p$ because X is feasible for (4.14). Therefore, it must be the case that $Q_i \bullet X = 0$ for $i = 1, \dots, p$. Since $X_{n+1,n+1} = 1$, we have that $r = \mathbf{rank}(X) > 0$. Then we can use Lemma A.10 to find $z_1, \dots, z_r \in \mathbf{R}^n$ such that

$$X = \sum_{i=1}^{r} z_i z_i^{\mathsf{T}} \quad \text{and} \quad z_i^{\mathsf{T}}(F_1 - F_2) z_i = \frac{(F_1 - F_2) \bullet X}{r} = 0,$$

where $(F_1 - F_2) \bullet X = 0$ because X is feasible for (4.14). We claim that

$$z = \mathbf{sign}((z_1)_{n+1}) z_1$$

satisfies the hypotheses of Lemma 4.1. We have that z is nonzero and contained in $\mathbf{range}(X)$ because $z_1, \dots, z_r$ form a basis for $\mathbf{range}(X)$. Additionally, since

$$z_1^{\mathsf{T}}(F_1 - F_2) z_1 = z^{\mathsf{T}}(F_1 - F_2) z = \|z_{1:n}\|^2 - z_{n+1}^2 = 0,$$

and $z_{n+1} \geq 0$, we have that z is on the boundary of $\mathbf{SOC}$. Because $Q_i \bullet X = 0$ for $i = 1, \dots, p$, we have that

$$Q_i \bullet X = Q_i \bullet \left(\sum_{j=1}^{r} z_j z_j^{\mathsf{T}} \right) = \sum_{j=1}^{r} z_j^{\mathsf{T}} Q_i z_j = 0.$$

Since $Q_i \succeq 0$, every term in the summation is nonnegative, and it must be the case that $z_j^{\mathsf{T}} Q_i z_j = 0$ for all i and j. In particular, we have that $z^{\mathsf{T}} Q_i z = z_1^{\mathsf{T}} Q_i z_1 = 0$ for $i = 1, \dots, p$. This proves that z satisfies the hypotheses of Lemma 4.1, and hence that (4.14) is an exact relaxation of (4.13) when $Q_1, \dots, Q_p \succeq 0$. □

5

QCQPs with Complex Variables

5.1 Introduction

Quadratically constrained quadratic programs (QCQPs) with complex variables appear frequently in applications. We define a complex QCQP in standard form to be an optimization problem of the form

$$\begin{array}{ll} \text{minimize} & z^{\mathsf{H}} Q z \\ \text{subject to} & z^{\mathsf{H}} A_j z \geq b_j, \quad j = 1, \ldots, m, \end{array} \tag{5.1}$$

where the optimization variable is $z \in \mathbf{C}^n$, and the problem data are $Q, A_1, \ldots, A_m \in \mathcal{H}^n$ and $b \in \mathbf{R}^m$. We use $\mathcal{H}^n$ to denote the set of $n \times n$ Hermitian matrices, and z^{H} to denote the conjugate transpose of z (also called the Hermitian transpose). Throughout this chapter we will reserve i for the imaginary unit. Note that the objective function and the left sides of the constraints are real even though z, Q, and $A_1, \ldots, A_m$ are complex since the value of all quadratic forms with Hermitian matrices is real.

Because the objective function or constraints may be nonconvex, (5.1) is intractable in general. Therefore, it is common to consider the natural SDP relaxation

$$\begin{array}{ll} \text{minimize} & Q \bullet Z \\ \text{subject to} & A_j \bullet Z \geq b_j, \quad j = 1, \ldots, m \\ & Z \succeq 0 \end{array} \tag{5.2}$$

with optimization variable $Z \in \mathcal{H}^n$. This problem is a (complex) SDP, and can be solved efficiently. Moreover, the SDP relaxation is tight if it has a rank-1 solution.

Throughout the chapter we will use the complex extensions of results that we only prove for real vectors and matrices. Most of the proofs are easily adapted to the complex case, and we will not usually explicitly state that we are using the complex versions.

5.2 Rank of SDP solutions

5.2.1 Bounds via constraint counting

Consider the standard-form complex SDP

$$\begin{array}{ll} \text{minimize} & C \bullet Z \\ \text{subject to} & A_j \bullet Z = b_j, \quad j = 1, \ldots, m \\ & Z \succeq 0 \end{array} \tag{5.3}$$

with variable $Z \in \mathcal{H}^n$ and optimal value $v^\star$, where $C, A_1, \ldots, A_m \in \mathcal{H}^n$ and $b \in \mathbf{R}^m$ are problem data. The following theorem is the complex analog of Theorem 2.1.

Theorem 5.1. If (5.3) is solvable, then it has a solution Z such that $\mathbf{rank}(Z) \leq \lfloor \sqrt{m} \rfloor$. Moreover, we can find such a solution efficiently.

Proof. Let Z be a solution of (5.3) with $\mathbf{rank}(Z) = r$. Then there exists $V \in \mathbf{C}^{n \times r}$ such that $VV^{\mathsf{H}} = Z$. Define $\tilde{C} = V^{\mathsf{H}} C V \in \mathcal{H}^r$ and $\tilde{A}_j = V^{\mathsf{H}} A_j V \in \mathcal{H}^r$ for $j = 1, \ldots, m$, and consider the problem

$$\begin{array}{ll} \text{minimize} & \tilde{C} \bullet \tilde{Z} \\ \text{subject to} & \tilde{A}_j \bullet \tilde{Z} = b_j, \quad j = 1, \ldots, m \\ & \tilde{Z} \succeq 0 \end{array} \tag{5.4}$$

with variable $\tilde{Z} \in \mathcal{H}^r$ and optimal value $\tilde{v}^\star$. We claim that $\tilde{Z} = I$ is a solution of (5.4), and $\tilde{v}^\star = v^\star$. First, observe that $\tilde{Z} = I$ is feasible for (5.4) since $I \succeq 0$, and

$$\tilde{A}_j \bullet I = (VA_jV^{\mathsf{H}}) \bullet I = A_j \bullet (VV^{\mathsf{H}}) = A_j \bullet Z = b_j.$$

Similarly, $W = I$ achieves an objective value of

$$\tilde{C} \bullet I = (V^{\mathsf{H}}CV) \bullet I = C \bullet (VV^{\mathsf{H}}) = C \bullet Z = v^\star.$$

This implies that $\tilde{v}^\star \leq v^\star$. Conversely, it is straightforward to check that if $\tilde{Z}$ is feasible for (5.4), then $V\tilde{Z}V^{\mathsf{H}}$ is feasible for (5.3), and achieves an objective value of $\tilde{C} \bullet \tilde{Z}$, which implies that $v^\star \leq \tilde{v}^\star$. Taken together these results allow us to conclude that $\tilde{Z} = I$ is optimal for (5.4), and $\tilde{v}^\star = v^\star$.

Next, we show that every $\tilde{Z}$ that is feasible for (5.4) is also optimal. Towards that end, consider the dual of (5.4):

$$\begin{array}{ll} \text{maximize} & b^{\mathsf{T}}\tilde{y} \\ \text{subject to} & \sum_{j=1}^m \tilde{y}_j \tilde{A}_j + \tilde{S} = \tilde{C} \\ & \tilde{S} \succeq 0 \end{array} \qquad (5.5)$$

with variables $\tilde{y} \in \mathbf{R}^m$ and $\tilde{S} \in \mathcal{H}^r$. Since (5.4) is bounded below and strictly feasible, strong duality holds, and (5.5) has a solution $(\tilde{y}, \tilde{S})$. Because $\tilde{Z} = I$ is optimal for (5.4), complementarity requires that $\tilde{S} \bullet I = 0$, and hence that $\tilde{S} = 0$. It follows that every $\tilde{Z}$ that is feasible for (5.4) satisfies the complementarity condition $\tilde{S} \bullet \tilde{Z} = 0$. Therefore, every feasible point of (5.4) is optimal.

To complete the proof, consider the system of linear equations

$$\tilde{A}_j \bullet \Delta = 0, \quad j = 1, \ldots, m \qquad (5.6)$$

with variable $\Delta \in \mathcal{H}^r$. Since Δ is conjugate symmetric, it is completely determined by the entries on and above the diagonal. Note that Δ has r diagonal entries (which must be real, and are therefore be specified by one real number), and $r(r-1)/2$ entries above the diagonal (which may be complex, and are therefore be specified by two real numbers). It follows that (5.6) is a system of m equations in $r + 2(r(r-1)/2) = r^2$

real variables. If $r^2 > m$, then (5.6) has a nonzero solution $\Delta \in \mathcal{H}^n$. Let λ_1 be a maximum-magnitude eigenvalue of Δ, and consider the matrix

$$\tilde{Z}^+ = I - (1/\lambda_1)\Delta \in \mathcal{H}^r.$$

Using an argument similar to the one in the proof of Proposition 2.2, we can verify that $\tilde{Z}^+ \succeq 0$ and $\mathbf{rank}(\tilde{Z}^+) < r$. Moreover, because $\tilde{A}_j \bullet \Delta = 0$ for $j = 1, \ldots, m$, we have that

$$\tilde{A}_j \bullet \tilde{Z}^+ = \tilde{A}_j \bullet \left(I - \frac{1}{\lambda_1}\Delta \right) = \tilde{A}_j \bullet I = b_j, \quad j = 1, \ldots, m.$$

It follows that $\tilde{Z}^+$ is feasible and hence optimal for (5.4). This allows us to conclude that $Z^+ = V\tilde{Z}^+V^{\mathsf{H}}$ is optimal for (5.3), and satisfies $\mathbf{rank}(Z^+) \leq \mathbf{rank}(\tilde{Z}^+) < r$.

We can now repeat the above procedure with Z^+ as our initial solution of (5.3). Iteratively applying this method until $\Delta = 0$ is the only solution of (5.6), we obtain a solution Z of (5.3) such that $\mathbf{rank}(Z)^2 \leq m$. Because $\mathbf{rank}(Z)$ is an integer, this inequality is equivalent to the bound $\mathbf{rank}(Z) \leq \lfloor \sqrt{m} \rfloor$. Additionally, the procedure used in this proof allows us to find such a solution efficiently. □

Corollary 5.2. Consider the SDP

$$\begin{array}{ll} \text{minimize} & C \bullet Z \\ \text{subject to} & A_j \bullet Z \geq b_j, \quad j = 1, \ldots, m \\ & A_j \bullet Z = b_j, \quad j = m+1, \ldots, m+p \\ & Z \succeq 0, \end{array}$$

with variable $Z \in \mathcal{H}^n$, and problem data $C, A_1, \ldots, A_{m+p} \in \mathcal{H}^n$ and $b \in \mathbf{R}^{m+p}$. If this problem is solvable, then it has a solution Z satisfying $\mathbf{rank}(Z) \leq \lfloor \sqrt{m+p} \rfloor$. Moreover, we can find such a solution efficiently.

Proof. Let Z_0 be a solution of the SDP. Then Z_0 is also a solution of the optimization problem

$$\begin{array}{ll} \text{minimize} & C \bullet Z \\ \text{subject to} & A_j \bullet Z = A_j \bullet Z_0, \quad j = 1, \ldots, m+p \\ & Z \succeq 0 \end{array} \qquad (5.7)$$

with variable $Z \in \mathcal{H}^n$. Using Theorem 5.1 we can efficiently find a solution Z of (5.7) satisfying $\mathbf{rank}(Z) \leq \lfloor \sqrt{m+p} \rfloor$. To complete the proof, we note that Z is also optimal for the original SDP. □

Using Corollary 5.2, we obtain our first tightness result concerning the relaxation (5.2).

Corollary 5.3. The semidefinite relaxation (5.2) is tight for (5.1) when $m \leq 3$.

Remark 5.1. Theorem 5.1 extends the corresponding result for real SDPs given in Theorem 2.1 to complex SDPs. A slightly different formulation of Theorem 5.1 was given by Huang and Zhang [51].

We can extend Theorem 5.1 to SDPs with block structure. Specifically, consider the problem

$$\begin{array}{ll} \text{minimize} & \sum_{k=1}^K C_k \bullet Z_k \\ \text{subject to} & \sum_{k=1}^K A_{jk} \bullet Z_k = b_j, \quad j = 1, \ldots, m \\ & Z_k \succeq 0, \quad k = 1, \ldots, K, \end{array} \tag{5.8}$$

where the optimization variables are $Z_k \in \mathcal{H}^{n_k}$ for $k = 1, \ldots, K$, and the problem data are $C_k, A_{1k}, \ldots, A_{mk} \in \mathcal{H}^{n_k}$ for $k = 1, \ldots, K$, and $b \in \mathbf{R}^m$. Let $v^\star$ be the optimal value of this problem. The following theorem gives a bound on the rank of a minimum-rank solution of a problem of this form.

Theorem 5.4. If (5.8) is solvable, then it has a solution $(Z_1, \ldots, Z_K)$ such that $\sum_{k=1}^K \mathbf{rank}(Z_k)^2 \leq m$. Moreover, we can find such a solution efficiently.

Proof. The proof is similar to that of Theorem 5.1. Let $(Z_1, \ldots, Z_k)$ be a solution of (5.8) with $\mathbf{rank}(Z_k) = r_k$ for $k = 1, \ldots, K$. Then there exists $V_k \in \mathbf{C}^{n_k \times r_k}$ such that $V_k V_k^{\mathsf{H}} = Z_k$ for $k = 1, \ldots, K$. Consider the auxiliary SDP

$$\begin{array}{ll} \text{minimize} & \sum_{k=1}^K \tilde{C}_k \bullet \tilde{Z}_k \\ \text{subject to} & \sum_{k=1}^K \tilde{A}_{jk} \bullet \tilde{Z}_k = b_j, \quad j = 1, \ldots, m \\ & \tilde{Z}_k \succeq 0, \quad k = 1, \ldots, K, \end{array} \tag{5.9}$$

with optimization variables $\tilde{Z}_k \in \mathcal{H}^{r_k}$ for $k = 1, \ldots, K$, where we define $\tilde{C}_k = V_k^{\mathsf{H}} C_k V_k \in \mathcal{H}^{r_k}$ for $k = 1, \ldots, K$, and $\tilde{A}_{jk} = V_k^{\mathsf{H}} A_{jk} V_k \in \mathcal{H}^{r_k}$ for $k = 1, \ldots, K$ and $j = 1, \ldots, m$. Let $\tilde{v}^\star$ be the optimal value of this auxiliary SDP. As in the proof of Theorem 5.1, we can show that every feasible point $(\tilde{Z}_1, \ldots, \tilde{Z}_k)$ of (5.9) is optimal, and corresponds to a solution $(V_1 \tilde{Z}_1 V_1^{\mathsf{H}}, \ldots, V_K \tilde{Z}_K V_K^{\mathsf{H}})$ of (5.8). Consider the system of equations

$$\sum_{k=1}^{K} \tilde{A}_{jk} \bullet \Delta_k = 0, \quad j = 1, \ldots, m \tag{5.10}$$

with variables $\Delta_k \in \mathcal{H}^{r_k}$ for $k = 1, \ldots, K$. Note that the number of real variables in (5.10) is $r^2 = \sum_{k=1}^{K} r_k^2$. Thus, if $r^2 > m$, then there exist matrices $\Delta_k \in \mathcal{H}^{r_k}$ for $k = 1, \ldots, K$ satisfying (5.10) such that at least one of the Δ_k is nonzero. Let $\tilde{\Lambda}$ be the set of all eigenvalues of the Δ_k:

$$\tilde{\Lambda} = \bigcup_{k=1}^{K} \{\lambda \in \mathbf{R} \mid \lambda \text{ is an eigenvalue of } \Delta_k\}$$

and let λ_1 be a maximum-magnitude element of $\tilde{\Lambda}$. Define

$$\tilde{Z}_k^+ = I_{r_k} - (1/\lambda_1)\Delta_k, \quad k = 1, \ldots, K$$

where I_{r_k} is the $r_k \times r_k$ identity matrix. We can then check that $\tilde{Z}_k^+ \succeq 0$ and $\mathbf{rank}(\tilde{Z}_k^+) \leq r_k$ for all $k = 1, \ldots, K$, and $\mathbf{rank}(\tilde{Z}_k^+) < r_k$ for some $k \in \{1, \ldots, K\}$. Additionally, because $\Delta_1, \ldots, \Delta_K$ satisfy (5.10), we have that

$$\begin{aligned} \sum_{k=1}^{K} \tilde{A}_{jk} \bullet \tilde{Z}_k^+ &= \sum_{k=1}^{K} \tilde{A}_{jk} \bullet (I_{r_k} - (1/\lambda)\Delta_k) \\ &= \sum_{k=1}^{K} \tilde{A}_{jk} \bullet I_{r_k} \\ &= \sum_{k=1}^{K} A_{jk} \bullet Z_k \\ &= b_j \end{aligned}$$

for $j = 1, \ldots, m$. It follows that $(\tilde{Z}_1^+, \ldots, \tilde{Z}_K^+)$ is feasible and hence optimal for (5.9). This implies that $Z_k^+ = V_k \tilde{Z}_k^+ V_k^{\mathsf{H}}$ for $k = 1, \ldots, K$ is a solution of (5.8) satisfying $\sum_{k=1}^{K} \mathbf{rank}(Z_k^+)^2 < r^2$.

To complete the proof, we repeat the procedure given above with Z_k^+ for $k = 1, \ldots, K$ as our initial solution. Iteratively applying this method until $(\Delta_1, \ldots, \Delta_K) = (0, \ldots, 0)$ is the only solution of (5.10) yields a solution of $(Z_1, \ldots, Z_K)$ of (5.8) with $\sum_{k=1}^K \mathbf{rank}(Z_k)^2 \leq m$. □

An analog of Corollary 5.2 for complex SDPs with block structure follows directly from Theorem 5.4.

Corollary 5.5. Consider the SDP

$$\begin{array}{ll} \text{minimize} & \sum_{k=1}^K C_k \bullet Z_k \\ \text{subject to} & \sum_{k=1}^K A_{jk} \bullet Z_k = b_j, \quad j = 1, \ldots, m \\ & \sum_{k=1}^K A_{jk} \bullet Z_k \geq b_j, \quad j = m+1, \ldots, m+p \\ & Z_1, \ldots, Z_k \succeq 0, \quad k = 1, \ldots, K \end{array}$$

with optimization variables $Z_k \in \mathcal{H}^{n_k}$ for $k = 1, \ldots, K$, and problem data $C_k, A_{1k}, \ldots, A_{mk} \in \mathcal{H}^{n_k}$ for $k = 1, \ldots, K$ and $b \in \mathbf{R}^m$. If this problem is solvable, then it has a solution $(Z_1, \ldots, Z_K)$ such that $\sum_{k=1}^K \mathbf{rank}(Z_k)^2 \leq m + p$. Moreover, we can compute such a solution efficiently.

Remark 5.2. Theorem 5.4 and Corollary 5.5 are due to Huang and Palomar [50].

5.2.2 Bound via complementarity

We have seen that an upper bound on the rank of a minimum-rank solution of an SDP can be obtained by counting the number of constraints. Now we describe an alternative approach, which exploits the complementarity property of primal and dual solutions. The dual of (5.3) is

$$\begin{array}{ll} \text{maximize} & b^\mathsf{T} y \\ \text{subject to} & \sum_{j=1}^m y_j A_j + S = C \\ & S \succeq 0, \end{array} \tag{5.11}$$

where the optimization variables are $y \in \mathbf{R}^m$ and $S \in \mathcal{H}^n$. Suppose Z and (y, S) are solutions of (5.3) and (5.11), respectively. These solutions must satisfy the complementarity condition $S \bullet Z = 0$. Then Lemma A.5 tells us that

$$\mathbf{rank}(Z) \leq n - \mathbf{rank}(S).$$

Thus, if we can argue that there exists a high-rank optimal dual slack variable, then we can conclude that every primal solution has low rank.

5.3 Connection to the $\mathcal{S}$-procedure

We can use the rank bounds in the previous section to develop the $\mathcal{S}$-procedure, which can be viewed as a theorem of alternatives for quadratic systems. The $\mathcal{S}$-procedure plays a fundamental role in the development of the duality theory for nonconvex quadratic optimization [5, 52, 95], and has applications in many areas of science and engineering [6]. For a historical perspective on the $\mathcal{S}$-procedure, we refer the reader to [46, 78]. We will use Theorem 5.1 to prove the following version of the $\mathcal{S}$-procedure.

Theorem 5.6. Suppose $A_1, A_2, Q \in \mathcal{H}^n$, and there exists $z_0 \in \mathbf{C}^n$ such that $z_0^{\mathsf{H}} A_j z_0 > 0$ for $j = 1, 2$. Then the following are equivalent:

(i) $z^{\mathsf{H}} Q z \geq 0$ whenever $z^{\mathsf{H}} A_1 z, z^{\mathsf{H}} A_2 z \geq 0$;

(ii) there exist $\lambda_1, \lambda_2 \geq 0$ such that $Q \succeq \lambda_1 A_1 + \lambda_2 A_2$.

Proof. Suppose there exist $\lambda_1, \lambda_2 \geq 0$ such that $Q \succeq \lambda_1 A_1 + \lambda_2 A_2$, and the vector $z \in \mathbf{C}^n$ satisfies $z^{\mathsf{H}} A_j z \geq 0$ for $j = 1, 2$. The assumption that $Q \succeq \lambda_1 A_1 + \lambda_2 A_2$ implies that

$$z^{\mathsf{H}} Q z \geq z^{\mathsf{H}} (\lambda_1 A_1 + \lambda_2 A_2) z = \lambda_1 (z^{\mathsf{H}} A_1 z) + \lambda_2 (z^{\mathsf{H}} A_2 z).$$

Then, using the assumptions that $\lambda_j \geq 0$ and $z^{\mathsf{H}} A_j z \geq 0$ for $j = 1, 2$, we can conclude that

$$z^{\mathsf{H}} Q z \geq \lambda_1 (z^{\mathsf{H}} A_1 z) + \lambda_2 (z^{\mathsf{H}} A_2 z) \geq 0.$$

Conversely, suppose $z^{\mathsf{H}} Q z \geq 0$ whenever $z^{\mathsf{H}} A_j z \geq 0$ for $j = 1, 2$. Consider the optimization problem

$$\begin{array}{ll} \text{minimize} & z^{\mathsf{H}} Q z \\ \text{subject to} & z^{\mathsf{H}} A_j z \geq 0, \quad j = 1, 2 \\ & \|z\| = 1 \end{array}$$

with variable $z \in \mathbf{C}^n$, and its SDP relaxation

$$\begin{array}{ll} \text{minimize} & Q \bullet Z \\ \text{subject to} & A_j \bullet Z \geq 0, \quad j = 1, 2 \\ & I \bullet Z = 1 \\ & Z \succeq 0 \end{array} \tag{5.12}$$

with variable $Z \in \mathcal{H}^n$. Let $v^\star$ be the optimal value of the SDP relaxation. Observe that (5.12) is solvable because its feasible region is compact. Then Theorem 5.1 tells us that the SDP relaxation has a rank-1 solution $Z = zz^{\mathsf{H}}$, where $z \in \mathbf{C}^n$ satisfies

$$A_j \bullet Z = z^{\mathsf{H}} A_j z \geq 0, \quad j = 1, 2.$$

Since $z^{\mathsf{H}} Q z \geq 0$ whenever $z^{\mathsf{H}} A_j z \geq 0$ for $j = 1, 2$, we have that

$$v^\star = Q \bullet Z = z^{\mathsf{H}} Q z \geq 0.$$

The dual of (5.12) is

$$\begin{array}{ll} \text{maximize} & \mu \\ \text{subject to} & Q \succeq \lambda_1 A_1 + \lambda_2 A_2 + \mu I \\ & \lambda_1, \lambda_2 \geq 0, \end{array} \tag{5.13}$$

where the variables are $\lambda_1, \lambda_2, \mu \in \mathbf{R}$. We assume that there exists a vector $z_0 \in \mathbf{C}^n$ such that $z_0^{\mathsf{T}} A_j z_0 > 0$ for $j = 1, 2$. This implies that (5.12) is strictly feasible. Therefore, strong duality holds, and (5.13) has a solution $(\mu, \lambda_1, \lambda_2)$, and $\mu = v^\star \geq 0$. Because $(\mu, \lambda_1, \lambda_2)$ is feasible for (5.13), we have that $\lambda_j \geq 0$ for $j = 1, 2$, and

$$Q \succeq \lambda_1 A_1 + \lambda_2 A_2 + \mu I \succeq \lambda_1 A_1 + \lambda_2 A_2.$$

□

Having found that the $\mathcal{S}$-procedure is a consequence of our results on the rank of solutions of complex SDPs, we can derive various extensions. For example, consider the following inhomogeneous version of Theorem 5.6.

Corollary 5.7. Given $A_1, A_2, P \in \mathcal{H}^n$, $b_1, b_2, q \in \mathbf{C}^n$, and $c_1, c_2, r \in \mathbf{R}$, define the functions $f_1, f_2, g : \mathbf{C}^n \to \mathbf{R}$ such that

$$\begin{aligned} f_1(z) &= z^{\mathsf{H}} A_1 z + 2\operatorname{Re}(b_1^{\mathsf{H}} z) + c_1, \\ f_2(z) &= z^{\mathsf{H}} A_2 z + 2\operatorname{Re}(b_2^{\mathsf{H}} z) + c_2, \\ g(z) &= z^{\mathsf{H}} P z + 2\operatorname{Re}(q^{\mathsf{H}} z) + r. \end{aligned}$$

Suppose there exists a vector $z_0 \in \mathbf{C}^n$ such that $f_1(z_0), f_2(z_0) > 0$. Then the following are equivalent:

(i) $g(z) \geq 0$ whenever $f_1(z), f_2(z) \geq 0$;

(ii) there exist $\lambda_1, \lambda_2 \geq 0$ such that

$$\begin{bmatrix} P & q \\ q^{\mathsf{H}} & r \end{bmatrix} \succeq \lambda_1 \begin{bmatrix} A_1 & b_1 \\ b_1^{\mathsf{H}} & c_1 \end{bmatrix} + \lambda_2 \begin{bmatrix} A_2 & b_2 \\ b_2^{\mathsf{H}} & c_2 \end{bmatrix}.$$

Proof. Define the functions $\tilde{f}_1, \tilde{f}_2, \tilde{g} : \mathbf{C}^n \times \mathbf{C} \to \mathbf{R}$ such that

$$\begin{aligned} \tilde{f}_j(z,t) &= \begin{bmatrix} z \\ t \end{bmatrix}^{\mathsf{H}} \begin{bmatrix} A_j & b_j \\ b_j^{\mathsf{H}} & c_j \end{bmatrix} \begin{bmatrix} z \\ t \end{bmatrix} = (z,t)^{\mathsf{H}} \tilde{A}_i (z,t), \quad j = 1, 2, \\ \tilde{g}(z,t) &= \begin{bmatrix} z \\ t \end{bmatrix}^{\mathsf{H}} \begin{bmatrix} P & q \\ q^{\mathsf{H}} & r \end{bmatrix} \begin{bmatrix} z \\ t \end{bmatrix} = (z,t)^{\mathsf{H}} \tilde{Q} (z,t), \end{aligned}$$

where we define

$$\tilde{A}_j = \begin{bmatrix} A_j & b_j \\ b_j^{\mathsf{H}} & c_j \end{bmatrix}, \quad j = 1, 2 \quad \text{and} \quad \tilde{Q} = \begin{bmatrix} P & q \\ q^{\mathsf{H}} & r \end{bmatrix}.$$

Note that $f_j(z) = \tilde{f}_j(z,1)$ for $j = 1, 2$, and $g(z) = \tilde{g}(z,1)$. We claim that (i) is equivalent to

$$\tilde{g}(z,t) \geq 0 \quad \text{whenever} \quad \tilde{f}_j(z,t) \geq 0 \quad \text{for } j = 1, 2. \tag{5.14}$$

If we are able to prove this equivalence, then the desired result follows from Theorem 5.6. Setting $t = 1$, we see that (5.14) implies (i). Conversely, suppose (i) holds. Fix values of $z \in \mathbf{C}^n$ and $t \in \mathbf{C}$ such that $\tilde{f}_j(z,t) \geq 0$ for $j = 1, 2$. There are two cases to consider.

(1) First, consider the case when $t \neq 0$. Then we have that

$$\begin{aligned} f_j(z/t) &= (z/t)^{\mathsf{H}} A_j (z/t) + 2\,\mathrm{Re}(b_j^{\mathsf{H}}(z/t)) + c \\ &= \frac{1}{|t|^2}\left(z^{\mathsf{H}} A_j z + z^{\mathsf{H}} b_j t + t^{\mathsf{H}} b_j^{\mathsf{H}} z + t^{\mathsf{H}} c t\right) \\ &= \frac{\tilde{f}_j(z,t)}{|t|^2} \\ &\geq 0. \end{aligned}$$

Because we assume that (i) holds, this implies that $g(z/t) \geq 0$, and hence that

$$\tilde{g}(z,t) = |t|^2 g(z/t) \geq 0.$$

(2) Now consider the case when $t = 0$. Write the complex number $(A_j z_0 + b_j)^{\mathsf{H}} z$ in polar form as

$$(A_j z_0 + b_j)^{\mathsf{H}} z = r_j \exp(i\theta_j), \quad j = 1, 2,$$

where $r_j \geq 0$ is the magnitude, and $\theta_j \in (-\pi, \pi]$ is the argument. For $\alpha, \phi \in \mathbf{R}$ and $\epsilon \in \{\pm 1\}$, we have that

$$f(\alpha\epsilon\exp(i\phi)z + z_0) = \tilde{f}_j(z,t)\alpha^2 + 2r_j\cos(\theta_j + \phi)\epsilon\alpha + f(z_0),$$

where $\tilde{f}_j(z,t) = z^{\mathsf{H}} A_j z$ since we are considering the case when $t = 0$. If we choose

$$\phi = -\frac{\theta_1 + \theta_2}{2} \quad \text{and} \quad \epsilon = \begin{cases} 1 & \cos((\theta_1 - \theta_2)/2) \geq 0, \\ -1 & \cos((\theta_1 - \theta_2)/2) < 0 \end{cases}$$

then we have that

$$\begin{aligned} &f(\alpha\epsilon\exp(i\phi)z + z_0) \\ &\quad = \tilde{f}_j(z,t)\alpha^2 + 2r_j\left|\cos\left(\frac{\theta_1 - \theta_2}{2}\right)\right|\alpha + f(z_0). \end{aligned}$$

We have assumed that $\tilde{f}_j(z,t) \geq 0$ and $r_j \geq 0$ for $j = 1, 2$, and $f(z_0) > 0$. Thus, we have that $f(\alpha\epsilon\exp(i\phi)z + z_0) > 0$ for all $\alpha \geq 0$. Since we are considering the case when (5.14) holds, this implies that

$$\begin{aligned} &g(\alpha\epsilon\exp(i\phi)z + z_0) \\ &\quad = \tilde{g}(z,t)\alpha^2 + 2\,\mathrm{Re}((Pz_0 + q)^{\mathsf{H}} z\exp(i\phi))\epsilon\alpha + g(z_0) \end{aligned}$$

is nonnegative for all $\alpha \geq 0$. Because concave quadratic functions are unbounded below, it must be the case that $g(\alpha\epsilon \exp(i\phi)z+z_0)$ is either a convex quadratic function of α or a linear function of α. Equivalently, we must have that $\tilde{g}(z,t) \geq 0$.

□

As another illustration of the power of the rank bound in Theorem 5.1, we give a variation of the $\mathcal{S}$-procedure with both equality and inequality constraints.

Corollary 5.8. Given $A, P \in \mathcal{H}^n$, $b, q \in \mathbf{C}^n$, and $c, r \in \mathbf{R}$, define the functions $f, g : \mathbf{C}^n \to \mathbf{R}$ such that

$$\begin{aligned} f(z) &= z^{\mathsf{H}}Az + 2\operatorname{Re}(b^{\mathsf{H}}z) + c, \\ g(z) &= z^{\mathsf{H}}Pz + 2\operatorname{Re}(q^{\mathsf{H}}z) + r. \end{aligned}$$

Additionally, suppose there exist $z_1, z_2 \in \mathbf{C}^n$ such that $\|z_1\|, \|z_2\| \leq 1$, and $f(z_1) < 0 < f(z_2)$. Then the following are equivalent:

(i) $g(z) \geq 0$ whenever $\|z\| \leq 1$ and $f(z) = 0$;

(ii) there exist $\lambda_1 \geq 0$ and $\lambda_2 \in \mathbf{R}$ such that

$$\begin{bmatrix} P & q \\ q^{\mathsf{H}} & r \end{bmatrix} \succeq \lambda_1 \begin{bmatrix} -I & 0 \\ 0 & 1 \end{bmatrix} + \lambda_2 \begin{bmatrix} A & b \\ b^{\mathsf{H}} & c \end{bmatrix}.$$

Proof. Suppose there exist $\lambda_1 \geq 0$ and $\lambda_2 \in \mathbf{R}$ such that

$$\begin{bmatrix} P & q \\ q^{\mathsf{H}} & r \end{bmatrix} \succeq \lambda_1 \begin{bmatrix} -I & 0 \\ 0 & 1 \end{bmatrix} + \lambda_2 \begin{bmatrix} A & b \\ b^{\mathsf{H}} & c \end{bmatrix},$$

and $z \in \mathbf{C}^n$ satisfies $\|z\| \leq 1$ and $f(z) = 0$. Then we have that

$$\begin{aligned} g(z) &= \begin{bmatrix} z \\ 1 \end{bmatrix}^{\mathsf{H}} \begin{bmatrix} P & q \\ q^{\mathsf{H}} & r \end{bmatrix} \begin{bmatrix} z \\ 1 \end{bmatrix} \\ &\geq \begin{bmatrix} z \\ 1 \end{bmatrix}^{\mathsf{H}} \left(\lambda_1 \begin{bmatrix} -I & 0 \\ 0 & 1 \end{bmatrix} + \lambda_2 \begin{bmatrix} A & b \\ b^{\mathsf{H}} & c \end{bmatrix} \right) \begin{bmatrix} z \\ 1 \end{bmatrix} \\ &= \lambda_1(1 - \|z\|^2) + \lambda_2 f(z). \end{aligned}$$

Since $\lambda_1 \geq 0$, $\|z\| \leq 1$, and $f(z) = 0$, this implies that $g(z) \geq 0$.

Conversely, suppose $g(z) \geq 0$ whenever $\|z\| \leq 1$ and $f(z) = 0$. Consider the optimization problem

$$\begin{array}{ll} \text{minimize} & g(z) \\ \text{subject to} & \|z\|^2 \leq 1 \\ & f(z) = 0, \end{array}$$

with variable $z \in \mathbf{C}^n$, and its SDP relaxation

$$\begin{array}{ll} \text{minimize} & \begin{bmatrix} P & q \\ q^{\mathsf{H}} & r \end{bmatrix} \bullet Z \\ \text{subject to} & \begin{bmatrix} -I & 0 \\ 0 & 1 \end{bmatrix} \bullet Z \geq 0 \\ & \begin{bmatrix} A & b \\ b^{\mathsf{H}} & c \end{bmatrix} \bullet Z = 0 \\ & \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \bullet Z = 1 \\ & Z \succeq 0 \end{array} \tag{5.15}$$

with variable $Z \in \mathcal{H}^{n+1}$. Let $v^\star$ be the optimal value of the SDP relaxation. Note that (5.15) is solvable because its feasible region is compact. Then Corollary 5.2 tells us that (5.15) has a rank-1 solution. Let $Z = vv^{\mathsf{H}}$ be such a rank-1 solution, and define $z = v_{n+1}v_{1:n}$, where $v_{1:n} = (v_1, \ldots, v_n) \in \mathbf{C}^n$. The constraints of (5.15) imply that $\|z\| \leq 1$ and $f(z) = 0$, and hence that $v^\star = g(z) \geq 0$. Define the matrix

$$Z_0 = \begin{bmatrix} (n+1)^{-1}I & 0 \\ 0 & 1 \end{bmatrix}.$$

Note that Z_0 satisfies

$$Z_0 \succeq 0 \quad \text{and} \quad \begin{bmatrix} I & 0 \\ 0 & -1 \end{bmatrix} \bullet Z_0 = -\frac{1}{n+1} < 0.$$

(Our subsequent analysis holds for every matrix Z_0 with these two properties; we only give a specific choice of Z_0 for concreteness.) Then define

$$\theta = \begin{bmatrix} A & b \\ b^{\mathsf{H}} & c \end{bmatrix} \bullet Z_0.$$

We claim that (5.15) is always strictly feasible. There are three cases to consider.

(1) If $\theta = 0$, then Z_0 is strictly feasible for (5.15).

(2) Now suppose $\theta > 0$. We assume that there exists $z_1 \in \mathbf{C}^n$ such that $\|z_1\| \leq 1$ and $f(z_1) < 0$. This assumption implies that the matrix

$$Z_1 = \begin{bmatrix} z_1 \\ 1 \end{bmatrix} \begin{bmatrix} z_1 \\ 1 \end{bmatrix}^{\mathsf{H}}$$

satisfies

$$\begin{bmatrix} I & 0 \\ 0 & -1 \end{bmatrix} \bullet Z_1 = \|z_1\|^2 - 1 \leq 0,$$
$$\begin{bmatrix} A & b \\ b^{\mathsf{H}} & c \end{bmatrix} \bullet Z_1 = f(z_1) < 0.$$

Therefore, we can find $\alpha \in (0,1)$ such that $\alpha Z_0 + (1-\alpha)Z_1$ is strictly feasible for (5.15).

(3) Finally, consider the case when $\theta < 0$. We assume that there exists $z_2 \in \mathbf{C}^n$ such that $\|z_2\| \leq 1$ and $f(z_2) > 0$. This assumption implies that the matrix

$$Z_2 = \begin{bmatrix} z_2 \\ 1 \end{bmatrix} \begin{bmatrix} z_2 \\ 1 \end{bmatrix}^{\mathsf{H}}$$

satisfies

$$\begin{bmatrix} I & 0 \\ 0 & -1 \end{bmatrix} \bullet Z_2 = \|z_2\|^2 - 1 \leq 0,$$
$$\begin{bmatrix} A & b \\ b^{\mathsf{H}} & c \end{bmatrix} \bullet Z_2 = f(z_2) > 0.$$

Thus, we can find $\beta \in (0,1)$ such that $\beta Z_0 + (1-\beta)Z_2$ is strictly feasible for (5.15).

The dual of (5.15):

$$\begin{array}{ll} \text{maximize} & \mu \\ \text{subject to} & \begin{bmatrix} P & q \\ q^{\mathsf{H}} & r \end{bmatrix} \succeq \lambda_1 \begin{bmatrix} -I & 0 \\ 0 & 1 \end{bmatrix} + \lambda_2 \begin{bmatrix} A & b \\ b^{\mathsf{H}} & c \end{bmatrix} + \mu \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \\ & \lambda_1 \geq 0 \end{array}$$

with variables $\lambda \in \mathbf{R}^2$ and $\mu \in \mathbf{R}$. Since we have shown that (5.15) is solvable and strictly feasible, the dual problem has a solution $(\mu, \lambda_1, \lambda_2)$, and there is no duality gap: that is, $\mu = v^\star \geq 0$. Thus, we find that

$$\begin{aligned} \begin{bmatrix} P & q \\ q^{\mathsf{H}} & r \end{bmatrix} &\succeq \lambda_1 \begin{bmatrix} -I & 0 \\ 0 & 1 \end{bmatrix} + \lambda_2 \begin{bmatrix} A & b \\ b^{\mathsf{H}} & c \end{bmatrix} + \mu \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \\ &\succeq \lambda_1 \begin{bmatrix} -I & 0 \\ 0 & 1 \end{bmatrix} + \lambda_2 \begin{bmatrix} A & b \\ b^{\mathsf{H}} & c \end{bmatrix}. \end{aligned}$$

□

Remark 5.3. In general the question of whether an $\mathcal{S}$-procedure exists for systems with quadratic equality constraints is a delicate one. For some recent progress in this direction, see [100].

5.4 Applications to signal processing

5.4.1 Unicast transmit downlink beamforming

Consider a base station with N antennae transmitting individual data streams to M single-antenna users. The signal transmitted by the base station at time t is

$$x(t) = \sum_{j=1}^{M} s_j(t) w_j, \quad t = 1, \ldots, T,$$

where $s_j(t) \in \mathbf{C}$ is the stream of unit-power data symbols, and $w_j \in \mathbf{C}^N$ is the beamforming vector for user j. The signal received by the jth user at time t is given by

$$\begin{aligned} y_j(t) &= h_j^{\mathsf{H}} x(t) + n_j(t) \\ &= \sum_{k=1}^{M} h_j^{\mathsf{H}} w_k s_k(t) + n_j(t) \end{aligned} \tag{5.16}$$

for $t = 1, \ldots, T$, where $h_j \in \mathbf{C}^N$ and $n_j(t) \in \mathbf{C}$ are, respectively, the channel vector and additive noise for user j. We assume that $n_j(t)$ has a complex normal distribution with mean vector 0 and covariance matrix $\sigma_j^2 I$, the channel vectors $h_1, \ldots, h_M$ are randomly fading, and only the second-order statistics $R_j = \mathbf{E}\left(h_j h_j^{\mathsf{H}}\right) \in \mathcal{H}^N$ are known for $j = 1, \ldots, M$. The signal-to-interference-plus-noise ratio (SINR) for user j is defined to be

$$\mathrm{SINR}_j = \frac{w_j^{\mathsf{H}} R_j w_j}{\sigma_j^2 + \sum_{k \neq j} w_k^{\mathsf{H}} R_j w_k}.$$

In the SINR-balancing problem, we want to choose $w_1, \ldots, w_M$ in order to minimize the total transmitter power subject to the constraint that the SINR for user j be greater than or equal to a given constant $\gamma_j > 0$ for $j = 1, \ldots, M$. We can express this problem mathematically as

$$\begin{array}{ll} \text{minimize} & \sum_{j=1}^M \|w_j\|^2 \\ \text{subject to} & \mathrm{SINR}_j \geq \gamma_j, \quad j = 1, \ldots, M, \end{array} \tag{5.17}$$

where the variables are $w_1, \ldots, w_M \in \mathbf{C}^N$. (See Gershman et al. [38] for more details on this problem formulation.) Note that (5.17) can be expressed as a QCQP by clearing the denominators in the inequality constraints. The natural SDP relaxation of this QCQP is the problem

$$\begin{array}{ll} \text{minimize} & \sum_{j=1}^M I \bullet W_j \\ \text{subject to} & \sum_{k=1}^M A_{jk} \bullet W_k \geq \gamma_j \sigma_j^2, \quad j = 1, \ldots, M \\ & W_j \succeq 0, \quad j = 1, \ldots, M \end{array} \tag{5.18}$$

with variables $W_1, \ldots, W_M \in \mathcal{H}^N$, where we define

$$A_{jk} = \begin{cases} R_j & j = k, \\ -\gamma_j R_j & \text{otherwise.} \end{cases}$$

It can be shown that the dual of (5.18) is strictly feasible. Therefore, if (5.18) is feasible, then strong duality holds, and the SDP relaxation of the SINR-balancing problem is solvable. Then we can use Theorem 5.4 to find a solution $(W_1, \ldots, W_M)$ such that $\sum_{j=1}^M \mathbf{rank}(W_j)^2 \leq M$. Note that W_j is nonzero since otherwise $\mathrm{SINR}_j = 0$, and we cannot satisfy the SINR constraint for the jth user. Therefore, we have that $\mathbf{rank}(W_j) = 1$ for $j = 1, \ldots, M$, and hence (5.18) is an exact relaxation of (5.17).

5.4.2 Transmit design for MISO channel secrecy

Consider a base station with N_0 antennae transmitting a data stream to a legitimate single-antenna receiver. Suppose there are M illegitimate multi-antenna receivers eavesdropping on the transmission. Let N_j be the number of antennae of the jth illegitimate receiver for $j = 1, \ldots, M$. In the literature the base station and legitimate receiver are typically called Alice and Bob, respectively; the jth eavesdropper is usually called the jth Eve.

A fundamental problem in such a scenario is to design a transmit scheme for the base station that allows it to reliably communicate with the legitimate receiver while preventing the eavesdroppers from obtaining information from the transmitted signals. Let $x(t) \in \mathbf{C}^{N_0}$ be the signal transmitted by Alice, $h \in \mathbf{C}^{N_0}$ be the multiple-input-single-output (MISO) channel response between Alice and Bob, $G_j \in \mathbf{C}^{N_0 \times N_j}$ be the multiple-input-multiple-output (MIMO) channel response between Alice and the jth Eve, and $n(t) \in \mathbf{C}$ and $v_j(t) \in \mathbf{C}^{N_j}$ be the additive white Gaussian noises at Bob and the jth Eve, respectively. Then the signals received at time t by Bob and the jth Eve are

$$y_0(t) = h^{\mathsf{H}} x(t) + n(t) \quad \text{and} \quad y_j(t) = G_j^{\mathsf{H}} x(t) + v_j(t),$$

respectively. Without loss of generality, we assume that $n(t)$ and $v_j(t)$ have unit variance. Furthermore, let $W = \mathbf{E}\Big(x(t)x(t)^{\mathsf{H}}\Big) \in \mathcal{H}^{N_0}$ be the transmitter covariance.

The following analysis is due to Li and Ma [64]. We are interested in minimizing the average transmitter power subject to the constraint that the achievable secrecy rate exceed a given lower bound $R > 0$. We can express this problem mathematically as

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}(W) \\ \text{subject to} & f_j(W) \geq R, \quad j = 1, \ldots, M \\ & W \succeq 0, \end{array} \tag{5.19}$$

where the variable is $W \in \mathcal{H}^{N_0}$, and

$$f_j(W) = \log_2(1 + h^{\mathsf{H}} W h) - \log_2(\det(I + G_j^{\mathsf{H}} W G_j))$$

is the secrecy-rate function for the jth Eve. Note that (5.19) is not an SDP because the secrecy-rate constraint is nonconvex. In order to obtain an SDP relaxation of (5.19), we need the following fact about positive-semidefinite Hermitian matrices.

Lemma 5.9. Suppose $A \in \mathcal{H}_+^n$, where $\mathcal{H}_+^n$ is the set of $n \times n$ positive-semidefinite Hermitian matrices:

$$\mathcal{H}_+^n = \{A \in \mathcal{H}^n \mid z^{\mathsf{H}} A z \geq 0 \text{ for all } z \in \mathbf{C}^n\}.$$

Then we have that

$$\det(I + A) \geq 1 + \mathbf{tr}(A),$$

with equality if and only if $\mathbf{rank}(A) \leq 1$.

Proof. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A. Because the determinant of a matrix is the product of its eigenvalues, we have that

$$\det(I + A) = \prod_{i=1}^{n} (1 + \lambda_i) = \sum_{S \subset \{1,\ldots,n\}} \prod_{j \in S} \lambda_j.$$

Note that $\lambda_i \geq 0$ for $i = 1, \ldots, n$ because A is positive semidefinite. Therefore, we can obtain a lower bound on $\det(I + A)$ from the expression above by ignoring the terms in the summation corresponding to subsets S with more than one element:

$$\det(I + A) \geq 1 + \sum_{i=1}^{n} \lambda_i = 1 + \mathbf{tr}(A).$$

Moreover, we have equality if and only if at most one of the λ_i is nonzero. Since the rank of A is the number of nonzero λ_i, this implies that the bound $\det(I + A) \geq 1 + \mathbf{tr}(A)$ holds with equality if and only if $\mathbf{rank}(A) \leq 1$. □

Using our lemma we find that

$$2^{f_j(W)} = \frac{1 + h^{\mathsf{H}} W h}{\det(I + G_j^{\mathsf{H}} W G_j)} \leq \frac{1 + h^{\mathsf{H}} W h}{1 + \mathbf{tr}(G_j^{\mathsf{H}} W G_j)}.$$

This bound gives us the following relaxation of (5.19):

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}(W) \\ \text{subject to} & 1 + h^{\mathsf{H}} W h \geq 2^R (1 + \mathbf{tr}(G_j^{\mathsf{H}} W G_j)), \quad j = 1, \ldots, M \\ & W \succeq 0. \end{array}$$

In order to make it clear that this problem is an SDP, we write it as

$$\begin{array}{ll} \text{minimize} & I \bullet W \\ \text{subject to} & (hh^{\mathsf{H}} - 2^R G_j G_j^{\mathsf{H}}) \bullet W \geq 2^R - 1, \quad j = 1, \ldots, M \\ & W \succeq 0. \end{array} \tag{5.20}$$

Note that this relaxation is tight if it has a rank-1 solution because the inequality in Lemma 5.9 holds with equality for rank-1 matrices. The dual of (5.20) is

$$\begin{array}{ll} \text{maximize} & (2^R - 1)\mathbf{1}^{\mathsf{T}} y \\ \text{subject to} & \sum_{j=1}^{M} y_j (hh^{\mathsf{H}} - 2^R G_j G_j^{\mathsf{H}}) + S = I \\ & y \geq 0 \\ & S \succeq 0, \end{array} \tag{5.21}$$

where the optimization variables are $y \in \mathbf{R}^M$ and $S \in \mathcal{H}^{N_0}$. Suppose (5.20) is feasible, and let W_0 be a feasible point. Then, for sufficiently large α, the matrix $\bar{W} = \alpha W_0 + I$ is strictly feasible for (5.20). We also have that

$$y_0 = \beta \mathbf{1} \quad \text{and} \quad S_0 = I - \sum_{j=1}^{M} (y_0)_j (hh^{\mathsf{H}} - 2^R G_j G_j^{\mathsf{H}})$$

are strictly feasible for (5.21), where $\beta = 1/(1 + |\lambda_0|)$, and λ_0 is a maximum-magnitude eigenvalue of

$$\sum_{j=1}^{M} (hh^{\mathsf{H}} - 2^R G_j G_j^{\mathsf{H}}).$$

Since (5.20) and (5.21) are both strictly feasible, strong duality holds, and we can find solutions W and (y, S) of the SDP relaxation and its dual, respectively. Define the matrix

$$B = I + 2^R \sum_{j=1}^{M} y_j G_j G_j^{\mathsf{H}} \succ 0.$$

Then we have that

$$S = I - \sum_{j=1}^{M} y_j (hh^{\mathsf{H}} - 2^R G_j G_j^{\mathsf{H}}) = B - \left(\sum_{j=1}^{M} y_j \right) hh^{\mathsf{H}},$$

and hence that

$$\begin{aligned} \mathbf{rank}(S) &= \mathbf{rank}(B^{-\frac{1}{2}} S B^{-\frac{1}{2}}) \\ &= \mathbf{rank}\left(I - \left(\sum_{i=1}^{M} y_i \right) \left(B^{-\frac{1}{2}} h \right) \left(B^{-\frac{1}{2}} h \right)^{\mathsf{H}} \right) \\ &\geq N_0 - 1, \end{aligned}$$

where the inequality follows from the fact that subtracting a dyad from a matrix can reduce the rank of the matrix by at most 1. The complementarity condition for (5.20) and (5.21) states that $S \bullet W = 0$. Therefore, we can use Lemma A.5 to conclude that

$$\mathbf{rank}(W) \leq N_0 - \mathbf{rank}(S) \leq 1.$$

Therefore, (5.20) is a tight relaxation of (5.19).

5.4.3 Robust unicast downlink precoder design

Let us revisit the SINR-balancing problem introduced in Section 5.4.1. In practice the channel vectors $h_j \in \mathbf{C}^N$ for $j = 1, \ldots, M$ are unknown, and must be estimated by the base station. In order to account for the channel estimation errors in our design process, we need to specify a model for the channel errors. A popular model for the channel errors is the norm-bounded error model [83, 97], in which the channel vector of user j is given by

$$h_j = \bar{h}_j + e_j,$$

where $\bar{h}_j \in \mathbf{C}^N$ is the nominal value of the channel vector for user j, and the channel-error vector $e_j \in \mathbf{C}^N$ satisfies $\|e_j\|_2 \leq \epsilon_j$ for a given threshold $\epsilon_j \geq 0$. With this error model, the robust precoder design problem is

$$\begin{array}{ll} \text{minimize} & \sum_{j=1}^{M} I \bullet W_j \\ \text{subject to} & \Psi_j(W_1, \ldots, W_M) \geq \gamma_j \sigma_j^2, \quad j = 1, \ldots, M \\ & W_j \succeq 0, \quad j = 1, \ldots, M, \end{array} \tag{5.22}$$

with variables $W_1, \ldots, W_n \in \mathcal{H}^n$, where we define

$$\Psi_j(W_1, \ldots, W_M) = \inf_{\|e_j\|_2 \le \epsilon_j} \Big((d_j + e_j)^{\mathsf{H}} \tilde{W}_j (d_j + e_j) \Big)$$

and

$$\tilde{W}_j = W_j - \gamma_j \sum_{k \ne j} W_k.$$

The constraint $\Psi_j(W_1, \ldots, W_M) \ge \gamma_j \sigma_j^2$ is semi-infinite due to the infimum in the definition of Ψ_j. We can obtain a tractable representation of this constraint using the $\mathcal{S}$-procedure. First, we observe that the robustness constraint is satisfied if and only if

$$\begin{aligned} &(\bar{h}_j + e_j)^{\mathsf{H}} \tilde{W}_j (\bar{h}_j + e_j) - \gamma_j \sigma_j^2 \\ &\quad = e_j^{\mathsf{H}} \tilde{W}_j e_j + 2\,\mathrm{Re}\Big((\tilde{W}_j \bar{h}_j)^{\mathsf{H}} e_j \Big) + (\bar{h}_j^{\mathsf{H}} \tilde{W}_j \bar{h}_j - \gamma_j \sigma_j^2). \\ &\quad \ge 0 \end{aligned}$$

whenever $\epsilon_j^2 - \|e_j\|_2^2 \ge 0$. Then we can apply Corollary 5.7 with

$$\begin{gathered} A_1 = A_2 = -I, \quad b_1 = b_2 = 0, \quad c_1 = c_2 = \epsilon_j^2, \\ P = \tilde{W}_j, \quad q = \tilde{W}_j \bar{h}_j, \quad \text{and} \quad r = \bar{h}_j^{\mathsf{H}} \tilde{W}_j \bar{h}_j - \gamma_j \sigma_j^2 \end{gathered}$$

to express the robustness constraint for the jth user as

$$\begin{bmatrix} \tilde{W}_j & \tilde{W}_j \bar{h}_j \\ \bar{h}_j^{\mathsf{H}} \tilde{W}_j & \bar{h}_j^{\mathsf{H}} \tilde{W}_j \bar{h}_j - \gamma_j \sigma_j^2 \end{bmatrix} \succeq \lambda_j \begin{bmatrix} -I & 0 \\ 0 & \epsilon_j^2 \end{bmatrix}$$
$$\lambda_j \ge 0.$$

(Note that we have combined the λ_1 and λ_2 of Corollary 5.7 into a single λ_j since f_1 and f_2 have the same coefficients.) Using this representation of the robustness constraint, we can write the robust precoder design problem as

$$\begin{array}{ll} \text{minimize} & \sum_{j=1}^M I \bullet W_j \\ \text{subject to} & \begin{bmatrix} \tilde{W}_j & \tilde{W}_j \bar{h}_j \\ \bar{h}_j^{\mathsf{H}} \tilde{W}_j & \bar{h}_j^{\mathsf{H}} \tilde{W}_j \bar{h}_j - \gamma_j \sigma_j^2 \end{bmatrix} \succeq \lambda_j \begin{bmatrix} -I & 0 \\ 0 & \epsilon_j^2 \end{bmatrix}, \ j = 1, \ldots, M \\ & \tilde{W}_j = W_j - \gamma_j \sum_{k \ne j} W_k, \quad j = 1, \ldots, M \\ & \lambda_j \ge 0, \quad j = 1, \ldots, M \\ & W_j \succeq 0, \quad j = 1, \ldots, M, \end{array}$$

which is an SDP. It is an open problem to determine whether this SDP always possesses a solution $(W_1, \dots, W_M)$ with $\mathbf{rank}(W_j) \leq 1$ for $j = 1, \dots, M$, although several authors have analyzed special cases of this problem [25, 89, 98].

Medra et al. [68] recently considered a frequency-division duplex (FDD) system with structured vector quantization, and proposed a channel-error model that can more accurately reflect the nature of estimation errors in such a system. In particular, let $\bar{h}_j \in \mathbf{C}^N$ be the nominal value of the channel vector for user j. In order to determine the direction of the channel for user j, the base station uses a Grassmannian codebook $\mathcal{C} = \{v_1, \dots, v_K\}$, where $v_k \in \mathbf{C}^N$ is a known unit vector for $k = 1, \dots, K$. Then the direction of user j's channel is

$$d_j \in \operatorname*{argmin}_{v \in \mathcal{C}} \left\{ 1 - \frac{|\bar{h}_j^{\mathsf{H}} v|^2}{\|\bar{h}_j\|_2^2} \right\}.$$

Under suitable conditions it is possible to show that the channel vector for user j can be expressed as

$$h_j = \|\bar{h}_j\|_2 (d_j + e_j),$$

where $e_j \in \mathbf{C}^N$ is an error vector whose statistics depend on both the channel and the codebook. As is often the case, the statistics of e_j are difficult to characterize. Thus, let us assume for simplicity that e_j lies in a region defined by the conditions

$$\|e_j\|_2 \leq \epsilon \quad \text{and} \quad \|d_j + e_j\|_2 = 1, \tag{5.23}$$

where $\epsilon \geq 0$ is a given parameter. We can then formulate a robust precoder design problem similar to (5.22). Another possibility is to treat $\epsilon \geq 0$ as a decision variable, and use it to determine the largest region in which the error vectors can reside without compromising the quality of service to the users. Specifically, let $\mathcal{E}_j(\epsilon)$ be the set of all vectors e_j satisfying (5.23), and consider the optimization problem

$$\begin{array}{ll} \text{maximize} & \epsilon \\ \text{subject to} & \Phi_j(W_1, \dots, W_M) \geq \gamma_j \sigma_j^2 / \|\bar{h}_j\|_2^2, \quad j = 1, \dots, M \\ & \sum_{j=1}^M I \bullet W_j \leq K \\ & W_j \succeq 0, \quad j = 1, \dots, M, \end{array}$$

where we define

$$\Phi_j(W_1,\ldots,W_M) = \inf_{e_j\in\mathcal{E}_j(\epsilon)}\Big((d_j+e_j)^{\mathsf{H}}\tilde{W}_j(d_j+e_j)\Big),$$

and $\tilde{W}_j$ is defined as in our analysis of the robust precoder design problem. The constraint $\Phi_j(W_1,\ldots,W_M) \geq \gamma_j\sigma_j^2/\|\bar{h}_j\|_2$ is semi-infinite due to the infimum in the definition of Φ_j. We can obtain a tractable representation of this constraint using the $\mathcal{S}$-procedure. First, we observe that the assumption $\|d_j\|_2 = 1$ implies that

$$\|d_j+e_j\|_2 = 1 \quad \text{if and only if} \quad e_j^{\mathsf{H}}e_j + 2\operatorname{Re}(d_j^{\mathsf{H}}e_j) = 0.$$

Thus, we have that

$$\mathcal{E}_j = \{e_j \in \mathbf{C}^N \mid \|e_j\|_2 \leq \epsilon,\ e_j^{\mathsf{H}}e_j + 2\operatorname{Re}(d_j^{\mathsf{H}}e_j) = 0\}.$$

Then the robustness constraint is satisfied if and only if

$$\begin{aligned}&(d_j+e_j)^{\mathsf{H}}\tilde{W}_j(d_j+e_j) - \gamma_j\sigma_j^2/\|\bar{h}_j\|_2^2\\ &\quad = e_j^{\mathsf{H}}\tilde{W}_je_j + 2\operatorname{Re}((\tilde{W}_jd_j)^{\mathsf{H}}e_j) + (d_j^{\mathsf{H}}\tilde{W}_jd_j - \gamma_j\sigma_j^2/\|\bar{h}_j\|_2^2)\\ &\quad \geq 0\end{aligned}$$

whenever $\|e_j\|_2 \leq \epsilon$ and $e_j^{\mathsf{H}}e_j + 2\operatorname{Re}(d_j^{\mathsf{H}}e_j) = 0$. Writing this condition in terms of $\tilde{e}_j = (1/\epsilon)e_j$ yields the equivalent condition that

$$\tilde{e}_j^{\mathsf{H}}(\epsilon^2\tilde{W}_j)\tilde{e}_j + 2\operatorname{Re}((\epsilon\tilde{W}_jd_j)^{\mathsf{H}}\tilde{e}_j) + (d_j^{\mathsf{H}}\tilde{W}_jd_j - \gamma_j\sigma_j^2/\|\bar{h}_j\|_2^2) \geq 0$$

whenever $\|\tilde{e}_j\|_2 \leq 1$ and $\tilde{e}_j^{\mathsf{H}}(\epsilon^2 I)\tilde{e}_j + 2\operatorname{Re}((\epsilon d_j)^{\mathsf{H}}\tilde{e}_j) = 0$. Note that the vectors

$$\tilde{e}_j^{(1)} = -\min(1/\epsilon,1)d_j \quad \text{and} \quad \tilde{e}_j^{(2)} = d_j$$

satisfy $\|\tilde{e}_j^{(1)}\|_2, \|\tilde{e}_j^{(2)}\|_2 \leq 1$, and

$$\begin{aligned}(\tilde{e}_j^{(1)})^{\mathsf{H}}(\epsilon^2 I)\tilde{e}_j^{(1)} + 2\operatorname{Re}((\epsilon d_j)^{\mathsf{H}}\tilde{e}_j^{(1)}) < 0,\\ (\tilde{e}_j^{(2)})^{\mathsf{H}}(\epsilon^2 I)\tilde{e}_j^{(2)} + 2\operatorname{Re}((\epsilon d_j)^{\mathsf{H}}\tilde{e}_j^{(2)}) > 0.\end{aligned}$$

Thus, we can use Corollary 5.8 with

$$A = \epsilon^2 I, \quad b = \epsilon d_j, \quad c = 0,$$

$$P = \epsilon^2\tilde{W}_j, \quad q = \epsilon\tilde{W}_jd_j, \quad \text{and} \quad r = d_j^{\mathsf{H}}\tilde{W}_jd_j - \frac{\gamma_j\sigma_j^2}{\|\bar{h}_j\|_2},$$

to express the robustness constraint for the jth user as

$$\begin{bmatrix} \epsilon^2 \tilde{W}_j & \epsilon \tilde{W}_j d_j \\ \epsilon d_j^{\mathsf{H}} \tilde{W}_j & d_j^{\mathsf{H}} \tilde{W}_j d_j - \gamma_j \sigma_j^2 / \|\bar{h}_j\|_2 \end{bmatrix} \succeq \tilde{\lambda}_{1j} \begin{bmatrix} -I & 0 \\ 0 & 1 \end{bmatrix} + \lambda_{2i} \begin{bmatrix} \epsilon^2 I & \epsilon d_j \\ \epsilon d_j^{\mathsf{H}} & 0 \end{bmatrix}$$
$$\tilde{\lambda}_{1j} \geq 0.$$

Sylvester's Law of Inertia [70] implies that matrix inequalities are preserved under congruence transformations. Multiplying the matrix inequality above on the left and the right by

$$\begin{bmatrix} 1/\epsilon & 0 \\ 0 & 1 \end{bmatrix}$$

and using the change of variables $\lambda_{1j} = \tilde{\lambda}_{1j}/\epsilon^2$ yields the simplified conditions

$$\begin{bmatrix} \tilde{W}_j & \tilde{W}_j d_j \\ d_j^{\mathsf{H}} \tilde{W}_j & d_j^{\mathsf{H}} \tilde{W}_j d_j - \gamma_j \sigma_j^2 / \|\bar{h}_j\|_2 \end{bmatrix} \succeq \lambda_{1j} \begin{bmatrix} -I & 0 \\ 0 & \epsilon^2 \end{bmatrix} + \lambda_{2j} \begin{bmatrix} I & d_j \\ d_j^{\mathsf{H}} & 0 \end{bmatrix}$$
$$\lambda_{1j} \geq 0.$$

Thus, we can reformulate our problem as

$$\begin{array}{ll} \text{maximize} & \epsilon \\ \text{subject to} & \begin{bmatrix} \tilde{W}_j & \tilde{W}_j d_j \\ d_j^{\mathsf{H}} \tilde{W}_j & d_j^{\mathsf{H}} \tilde{W}_j d_j - \gamma_j \sigma_j^2 / \|\bar{h_j}\|^2 \end{bmatrix} \\ & \quad \succeq \lambda_{1j} \begin{bmatrix} -I & 0 \\ 0 & \epsilon^2 \end{bmatrix} + \lambda_{2j} \begin{bmatrix} I & d_j \\ d_j^{\mathsf{H}} & 0 \end{bmatrix}, \quad j = 1, \ldots, M \\ & \tilde{W}_j = W_j - \gamma_j \sum_{k \neq j} W_k, \quad j = 1, \ldots, M \\ & \lambda_{1j} \geq 0, \quad j = 1, \ldots, M \\ & \sum_{j=1}^{M} I \bullet W_j \leq K \\ & W_j \succeq 0, \quad j = 1, \ldots, M. \end{array}$$

Although the reformulated problem is not an SDP due to the terms of the form $\lambda_{1j}\epsilon^2$, it is an SDP for every fixed value of $\epsilon > 0$. Thus, we can efficiently solve the reformulation to an arbitrary level of accuracy using a bisection search on ϵ.

Acknowledgments

We would like to thank Zhisu Zhu for providing his PhD thesis for us to read. Trisha Lemon was invaluable in proofreading several drafts of the paper. Insightful reviews were provided by Jaehyun Park and Enzo Busseti.

Appendices

A

Background

There are many excellent books about linear algebra [70, 92] and optimization [10, 65]. We do not attempt to give a comprehensive summary of these fields in this appendix – we only discuss results that are either uncommon in more general treatments, or extremely important in the analysis of rank in semidefinite programs. This appendix also serves to set our notation.

A.1 Linear algebra

We write $X \in \mathbf{S}^n$ to indicate that X is an $n \times n$ symmetric matrix, and $X \in \mathbf{S}^n_+$ to specify that X is an $n \times n$ positive-semidefinite symmetric matrix. (Recall that a symmetric matrix X is positive semidefinite if $z^\mathsf{T} X z \geq 0$ for all $z \in \mathbf{R}^n$.) The spectral theorem tells us that every symmetric matrix X has an orthogonal eigenvalue decomposition: that is, there exist scalars $\lambda_1, \ldots, \lambda_n \in \mathbf{R}$ (the eigenvalues of X), and an orthonormal set of vectors $q_1, \ldots, q_n \in \mathbf{R}^n$ (corresponding eigenvectors of X) such that

$$X = \sum_{i=1}^{n} \lambda_i q_i q_i^\mathsf{T} = \tilde{Q}\tilde{\Lambda}\tilde{Q}^\mathsf{T},$$

where $\tilde{\Lambda} = \mathbf{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbf{R}^{n \times n}$ is the diagonal matrix whose diagonal entries are $\lambda_1, \ldots, \lambda_n$, and $\tilde{Q} \in \mathbf{R}^{n \times n}$ is the matrix whose columns are $q_1, \ldots, q_n$. If $\mathbf{rank}(X) = r$, then X has exactly nonzero eigenvalues. We can assume without loss of generality that $\lambda_1, \ldots, \lambda_r$ are the nonzero eigenvalues of X. Then the eigenvalue decomposition of X reduces to

$$X = \sum_{i=1}^{r} \lambda_i q_i q_i^{\mathsf{T}} = Q \Lambda Q^{\mathsf{T}},$$

where $\Lambda = \mathbf{diag}(\lambda_1, \ldots, \lambda_r) \in \mathbf{R}^{r \times r}$ is the diagonal matrix whose diagonal entries are $\lambda_1, \ldots, \lambda_r$, and $Q \in \mathbf{R}^{n \times r}$ is the matrix whose columns are $q_1, \ldots, q_r$. If X is symmetric and positive semidefinite, then we have that $\lambda_1, \ldots, \lambda_r > \lambda_{r+1} = \cdots = \lambda_n = 0$. Then we can use the eigenvalue decomposition to construct a dyadic decomposition of X: if we define $v_i = \sqrt{\lambda_i} q_i$ for $i = 1, \ldots, n$, then

$$X = \sum_{i=1}^{n} v_i v_i^{\mathsf{T}} = \tilde{V} \tilde{V}^{\mathsf{T}} = \sum_{i=1}^{r} v_i v_i^{\mathsf{T}} = V V^{\mathsf{T}},$$

where $\tilde{V} \in \mathbf{R}^{n \times n}$ and $V \in \mathbf{R}^{n \times r}$ are the matrices whose columns are $v_1, \ldots, v_n$ and $v_1, \ldots, v_r$, respectively. (Note that these dyadic decomposition are not unique: for example, if $W \in \mathbf{R}^{r \times r}$ is orthogonal, then $X = (VW)(VW)^{\mathsf{T}}$ is another dyadic decomposition of X. The Cholesky factorization is another common decomposition of this form.)

The trace inner product of $A, B \in \mathbf{R}^{m \times n}$ is

$$A \bullet B = \mathbf{tr}(A^{\mathsf{T}} B) = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} B_{ij}.$$

Using these definitions and facts, we can prove the following results, which are mostly standard, but included here because they are extremely important in the analysis of rank in semidefinite programs.

Lemma A.1. Suppose $X, Y \in \mathbf{S}^n$. If $X, Y \succeq 0$, then $X \bullet Y \geq 0$.

Proof. Since $Y \succeq 0$, there exist vectors $v_1, \ldots, v_n \in \mathbf{R}^n$ such that

$$Y = \sum_{k=1}^{n} v_k v_k^{\mathsf{T}}.$$

Then we have that

$$X \bullet Y = X \bullet \left(\sum_{k=1}^{n} v_k v_k^\mathsf{T} \right) = \sum_{k=1}^{n} X \bullet (v_k v_k^\mathsf{T}) = \sum_{k=1}^{n} v_k^\mathsf{T} X v_k.$$

Because $X \succeq 0$, each term in the summation is nonnegative, which implies that $X \bullet Y \geq 0$. □

Lemma A.2. Suppose $A \in \mathbf{S}_+^n$ and $x \in \mathbf{R}^n$. Then $x^\mathsf{T} A x = 0$ if and only if $Ax = 0$.

Proof. Suppose $Ax = 0$. Then we have that $x^\mathsf{T} A x = x^\mathsf{T}(0) = 0$. Conversely suppose $x^\mathsf{T} A x = 0$. Since $A \succeq 0$, there exist vectors $v_1, \ldots, v_n \in \mathbf{R}^n$ such that

$$A = \sum_{k=1}^{n} v_k v_k^\mathsf{T}.$$

Using this expression for A, we find that

$$x^\mathsf{T} A x = x^\mathsf{T} \left(\sum_{k=1}^{n} v_k v_k^\mathsf{T} \right) x = \sum_{k=1}^{n} (v_k^\mathsf{T} x)^2 = 0.$$

Because every term in the summation is nonnegative, it must be the case that $v_k^\mathsf{T} x = 0$ for $k = 1, \ldots, n$, and hence that

$$Ax = \left(\sum_{k=1}^{n} v_k v_k^\mathsf{T} \right) x = \sum_{k=1}^{n} (v_k^\mathsf{T} x) v_k = \sum_{k=1}^{n} (0) v_k = 0.$$

□

Lemma A.3. Suppose $X \in \mathbf{S}_+^n$ and $V \in \mathbf{R}^{n \times k}$. If $X \bullet (VV^\mathsf{T}) = 0$, then $XV = 0$.

Proof. Let $v_1, \ldots, v_k \in \mathbf{R}^n$ be the columns of V:

$$V = \begin{bmatrix} v_1 & \cdots & v_k \end{bmatrix}.$$

Then we have that

$$VV^\mathsf{T} = \begin{bmatrix} v_1 & \cdots & v_k \end{bmatrix} \begin{bmatrix} v_1 & \cdots & v_k \end{bmatrix}^\mathsf{T} = \sum_{j=1}^{k} v_j v_j^\mathsf{T},$$

and hence that

$$X \bullet (VV^{\mathsf{T}}) = X \bullet \left(\sum_{j=1}^{k} v_j v_j^{\mathsf{T}} \right) = \sum_{j=1}^{k} X \bullet (v_j v_j^{\mathsf{T}}) = \sum_{j=1}^{k} v_j^{\mathsf{T}} X v_j.$$

Each term in this summation is nonnegative because $X \succeq 0$. Thus, if $X \bullet (VV^{\mathsf{T}}) = 0$, then we have that $v_j^{\mathsf{T}} X v_j = 0$ for $j = 1, \ldots, k$. We can then use Lemma A.2 to conclude that $Xv_j = 0$ for $j = 1, \ldots, k$, and hence that

$$XV = X \begin{bmatrix} v_1 & \cdots & v_k \end{bmatrix} = \begin{bmatrix} Xv_1 & \cdots & Xv_k \end{bmatrix} = 0.$$

□

Lemma A.4. Suppose $X, Y \in \mathbf{S}_{+}^{n}$. If $X \bullet Y = 0$, then $XY = 0$.

Proof. Since Y is positive semidefinite, there exists $V \in \mathbf{R}^{n \times n}$ such that $Y = VV^{\mathsf{T}}$. Then we have that $X \bullet Y = X \bullet (VV^{\mathsf{T}})$, so we can use Lemma A.3 to conclude that $XV = 0$, and hence that

$$XY = X(VV^{\mathsf{T}}) = (XV)V^{\mathsf{T}} = (0)V^{\mathsf{T}} = 0.$$

□

Lemma A.5. Suppose $A, B \in \mathbf{S}_{+}^{n}$. If $A \bullet B = 0$, then

$$\mathbf{rank}(A) + \mathbf{rank}(B) \leq n.$$

Proof. Let

$$A = Q \Lambda Q^{\mathsf{T}} = \sum_{i=1}^{n} \lambda_i q_i q_i^{\mathsf{T}}$$

be the (full) eigenvalue decomposition of A. Assume that the eigenvalues are ordered such that $\lambda_1 \geq \cdots \geq \lambda_r > \lambda_{r+1} = \cdots = \lambda_n = 0$, where $r = \mathbf{rank}(A)$. (Note that $\lambda_i \geq 0$ for $i = 1, \ldots, n$ because $A \succeq 0$.) Observe that

$$A \bullet B = \left(\sum_{i=1}^{n} \lambda_i q_i q_i^{\mathsf{T}} \right) \bullet B = \sum_{i=1}^{n} \lambda_i (q_i^{\mathsf{T}} B q_i) = 0.$$

Since $B \succeq 0$, we have that $q_i^\mathsf{T} B q_i \geq 0$ for $i = 1, \dots, n$. Taken together, these results imply that $q_i^\mathsf{T} B q_i = 0$ whenever $\lambda_i > 0$. Then Lemma A.2 tells us that $Bq_i = 0$ for $i = 1, \dots, r$. Therefore,

$$\mathbf{span}(q_1, \dots, q_r) \subset \mathbf{null}(B).$$

Using conservation of dimension, we obtain the bound

$$\mathbf{rank}(B) = n - \dim(\mathbf{null}(B)) \leq n - \dim(\mathbf{span}(q_1, \dots, q_r)) = n - r.$$

Because $\mathbf{rank}(A) = r$, this implies that

$$\mathbf{rank}(A) + \mathbf{rank}(B) \leq r + (n - r) = n.$$

□

Proposition A.1. Suppose $A, B \in \mathbf{S}_+^n$. Let $r = \mathbf{rank}(A)$. Since A is positive semidefinite, there exists $V \in \mathbf{R}^{n \times r}$ such that $A = VV^\mathsf{T}$. We have that $\mathbf{null}(A) \subset \mathbf{null}(B)$ if and only if there exists a matrix $Q \in \mathbf{S}^r$ such that $B = VQV^\mathsf{T}$.

Proof. First, suppose there exists $Q \in \mathbf{S}^r$ such that $B = VQV^\mathsf{T}$. Then, for every vector $z \in \mathbf{null}(A)$, we have that

$$z^\mathsf{T} A z = z^\mathsf{T}(VV^\mathsf{T})z = \|V^\mathsf{T} z\|^2 = 0,$$

and hence that $V^\mathsf{T} z = 0$. This allows us to conclude that

$$Bz = (VQV^\mathsf{T})z = (VQ)(V^\mathsf{T} z) = (VQ)(0) = 0,$$

and hence that $\mathbf{null}(A) \subset \mathbf{null}(B)$.

Conversely, suppose $\mathbf{null}(A) \subset \mathbf{null}(B)$. Let the (reduced) eigenvalue decompositions of A and B be

$$A = U\Lambda U^\mathsf{T} \quad \text{and} \quad B = WMW^\mathsf{T},$$

respectively, where $U \in \mathbf{R}^{n \times r}$ and $W \in \mathbf{R}^{n \times s}$ are matrices with orthonormal columns, $\Lambda \in \mathbf{R}^{r \times r}$ and $M \in \mathbf{R}^{s \times s}$ are diagonal and nonsingular, $r = \mathbf{rank}(A)$, and $s = \mathbf{rank}(B)$. With $Q = (V^\dagger W)M(V^\dagger M)^\mathsf{T}$, we have that

$$\begin{aligned} VQV^\mathsf{T} &= V((V^\dagger W)M(V^\dagger W)^\mathsf{T})V^\mathsf{T} \\ &= (VV^\dagger W)M(VV^\dagger W)^\mathsf{T}. \end{aligned}$$

Because A and B are symmetric, the condition $\mathbf{null}(A) \subset \mathbf{null}(B)$ implies that

$$\mathbf{range}(B) = \mathbf{null}(B)^\perp \subset \mathbf{null}(A)^\perp = \mathbf{range}(A).$$

Let $w_1, \ldots, w_s \in \mathbf{R}^n$ denote the columns of W. Then we have that $w_i \in \mathbf{range}(B) \subset \mathbf{range}(A)$, and hence that $VV^\dagger w_i = w_i$ since $VV^\dagger$ is the projection onto $\mathbf{range}(V) = \mathbf{range}(A)$. Using this observation, we find that

$$VV^\dagger W = \begin{bmatrix} VV^\dagger w_1 & \cdots & VV^\dagger w_s \end{bmatrix} = \begin{bmatrix} w_1 & \cdots & w_s \end{bmatrix} = W.$$

Applying this result, we find that

$$VQV^\mathsf{T} = WMW^\mathsf{T} = B.$$

□

Lemma A.6. If S_1 and S_2 are subspaces of an inner-product space V, then $(S_1 + S_2)^\perp = S_1^\perp \cap S_2^\perp$.

Proof. Suppose $x \in (S_1 + S_2)^\perp$ and $y_1 \in S_1$. Since S_2 is a subspace, it contains the zero vector, so $y_1 = y_1 + 0 \in S_1 + S_2$. Therefore, $x^\mathsf{T} y_1 = 0$. This proves that $x \in S_1^\perp$. Similarly, we have that $x \in S_2^\perp$. Thus, we can conclude that $x \in S_1^\perp \cap S_2^\perp$, and hence that $(S_1 + S_2)^\perp \subset S_1^\perp \cap S_2^\perp$.

Now suppose $x \in S_1^\perp \cap S_2^\perp$ and $y \in S_1 + S_2$. There exist $y_1 \in S_1$ and $y_2 \in S_2$ such that $y = y_1 + y_2$. Since $x \in S_i^\perp$, we have that $x^\mathsf{T} y_i = 0$ for $i = 1, 2$, and hence that

$$x^\mathsf{T} y = x^\mathsf{T}(y_1 + y_2) = x^\mathsf{T} y_1 + x^\mathsf{T} y_2 = 0 + 0 = 0.$$

This proves that $x \in (S_1 + S_2)^\perp$, and thereby completes the proof that $(S_1 + S_2)^\perp = S_1^\perp \cap S_2^\perp$. □

Lemma A.7. If $A_1, A_2 \in \mathbf{S}_+^n$, then

$$\mathbf{range}(A_1 + A_2) = \mathbf{range}(A_1) + \mathbf{range}(A_2).$$

Proof. Suppose $y \in \mathbf{range}(A_1 + A_2)$. Then there exists $x \in \mathbf{R}^n$ such that $y = (A_1 + A_2)x$, and we have that $y = y_1 + y_2$, where $y_1 = A_1 x \in \mathbf{range}(A_1)$, and $y_2 = A_2 x \in \mathbf{range}(A_2)$. Thus, we see that

$y \in \mathbf{range}(A_1) + \mathbf{range}(A_2)$, which serves to establish the inclusion $\mathbf{range}(A_1 + A_2) \subset \mathbf{range}(A_1) + \mathbf{range}(A_2)$. (Note that this inclusion holds for all matrices $A_1, A_2 \in \mathbf{R}^{m \times n}$ – the additional assumption that A_1 and A_2 are positive semidefinite is only needed to prove the reverse inclusion.)

Taking the orthogonal complement of both sides of the reverse inclusion, we find that

$$\begin{aligned} \mathbf{range}(A_1 + A_2)^\perp &= \mathbf{null}(A_1 + A_2) \\ &\subset (\mathbf{range}(A_1) + \mathbf{range}(A_2))^\perp \\ &= \mathbf{range}(A_1)^\perp + \mathbf{range}(A_2)^\perp \\ &= \mathbf{null}(A_1) + \mathbf{null}(A_2), \end{aligned}$$

where we have used Lemma A.6, and the fact that A_1 and A_2 are symmetric. Thus, we can prove the reverse inclusion by showing the equivalent statement

$$\mathbf{null}(A_1 + A_2) \subset \mathbf{null}(A_1) + \mathbf{null}(A_2).$$

Suppose $x \in \mathbf{null}(A_1 + A_2)$. Then we have that

$$x^\mathsf{T}(A_1 + A_2)x = x^\mathsf{T} A_1 x + x^\mathsf{T} A_2 x = 0.$$

Because A_i is positive semidefinite, we have that $x^\mathsf{T} A_i x \geq 0$ for $i = 1, 2$. If the sum of two nonnegative numbers is equal to zero, then both of those numbers must be equal to zero; thus,

$$x^\mathsf{T} A_1 x = x^\mathsf{T} A_2 x = 0.$$

Lemma A.2 then allows us to conclude that $A_1 x = A_2 x = 0$, and hence that $x \in \mathbf{null}(A_1) \cap \mathbf{null}(A_2)$. □

Lemma A.8 (Schur product theorem)**.** Suppose $X, Y \in \mathbf{S}^n$. If $X, Y \succeq 0$, then $X \circ Y \succeq 0$, where $X \circ Y$ denotes the entrywise (also called the Hadamard or Schur) product of X and Y.

Proof. Since $Y \succeq 0$, there exist vectors $v_1, \ldots, v_n \in \mathbf{R}^n$ such that

$$Y = \sum_{k=1}^{n} v_k v_k^\mathsf{T}.$$

The (i, j)-entry of $X \circ Y$ is then

$$(X \circ Y)_{ij} = X_{ij} Y_{ij} = X_{ij} \left(\sum_{k=1}^{n} v_k v_k^\mathsf{T} \right)_{ij} = \sum_{k=1}^{n} X_{ij} (v_k)_i (v_k)_j$$

and, for all $z \in \mathbf{R}^n$, we have that

$$\begin{aligned} z^\mathsf{T} (X \circ Y) z &= \sum_{i=1}^{n} \sum_{j=1}^{n} (X \circ Y)_{ij} z_i z_j \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} X_{ij} (v_k)_i (v_k)_j z_i z_j \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} X_{ij} (z \circ v_k)_i (z \circ v_k)_j \\ &= \sum_{k=1}^{n} (z \circ v_k)^\mathsf{T} X (z \circ v_k). \end{aligned}$$

Because $X \succeq 0$, we have that $(z \circ v_k)^\mathsf{T} X (z \circ v_k) \geq 0$ for $k = 1, \ldots, n$, and therefore that

$$z^\mathsf{T} (X \circ Y) z = \sum_{k=1}^{n} (z \circ v_k)^\mathsf{T} X (z \circ v_k) \geq 0.$$

Thus, we have that $z^\mathsf{T} (X \circ Y) z \geq 0$ for all $z \in \mathbf{R}^n$, and hence that $X \circ Y \succeq 0$. □

Corollary A.9. Suppose $X \in \mathbf{S}^n$, and $|X_{ij}| \leq 1$ for all i and j. Let $\arcsin(X)$ be the entrywise inverse sine of the matrix X. If $X \succeq 0$, then $\arcsin(X) \succeq X$.

Proof. The Taylor series expansion of $\arcsin(X) - X$ is

$$\arcsin(X) - X = \sum_{n=1}^{\infty} \frac{(1)(3)(5) \cdots (2n-1)}{(2)(4)(6) \cdots (2n)(2n+1)} X^{\langle 2n+1 \rangle},$$

where $X^{\langle k \rangle}$ denotes the entrywise kth power of X. Since $X \succeq 0$, Lemma A.8 (applied inductively) implies that $X^{\langle 2n+1 \rangle} \succeq 0$ for all positive integers n, and hence that $\arcsin(X) - X \succeq 0$. Rearranging this matrix inequality, we find that $\arcsin(X) \succeq X$. □

The following lemma is a generalization of a result given by Sturm and Zhang [93].

Lemma A.10. Suppose $A, X \in \mathbf{S}^n$. If $X \succeq 0$ and $\mathbf{rank}(X) = r$, then there exist vectors $x_1, \dots, x_r \in \mathbf{R}^n$ such that

$$X = \sum_{k=1}^{r} x_k x_k^\mathsf{T} \quad \text{and} \quad x_k^\mathsf{T} A x_k = \frac{A \bullet X}{r}, \quad k = 1, \dots, r.$$

Moreover, such vectors can be found efficiently.

Algorithm A.1: equilibration of a dyadic expansion

```
**Input**: $v_1, \dots, v_r \in \mathbf{R}^n$ such that $X = \sum_{k=1}^r v_k v_k^\mathsf{T}$
**repeat**
    find $i, j \in \{1, \dots, r\}$ such that $v_i^\mathsf{T} A v_i < (A \bullet X)/r < v_j^\mathsf{T} A v_j$
    find $\hat\theta \in (0, \pi/2)$ such that
    $(\cos(\hat\theta) v_i + \sin(\hat\theta) v_j)^\mathsf{T} A (\cos(\hat\theta) v_i + \sin(\hat\theta) v_j) = (A \bullet X)/r$
    $\begin{bmatrix} v_i & v_j \end{bmatrix} := \begin{bmatrix} v_i & v_j \end{bmatrix} \begin{bmatrix} \cos(\hat\theta) & -\sin(\hat\theta) \\ \sin(\hat\theta) & \cos(\hat\theta) \end{bmatrix}$
**until** $v_k^\mathsf{T} A v_k = (A \bullet X)/r$ for $k = 1, \dots, r$
```

Proof. We will argue that Algorithm A.1 can be used to find the vectors $x_1, \dots, x_r$. Since $X \succeq 0$ and $\mathbf{rank}(X) = r$, we can use the eigenvalue decomposition of X to find vectors $v_1, \dots, v_r \in \mathbf{R}^n$ such that

$$X = \sum_{k=1}^{r} v_k v_k^\mathsf{T}.$$

These vectors are a dyadic expansion of X, but may not satisfy the condition $v_k^\mathsf{T} A v_k = (A \bullet X)/r$ for $k = 1, \dots, r$. Observe that

$$\frac{1}{r} \sum_{k=1}^{r} v_k^\mathsf{T} A v_k = \frac{1}{r} \sum_{k=1}^{r} A \bullet (v_k v_k^\mathsf{T}) = \frac{1}{r} \left(A \bullet \left(\sum_{k=1}^{r} v_k v_k^\mathsf{T} \right) \right) = \frac{A \bullet X}{r}.$$

Thus, the average value of $v_1^\mathsf{T} A v_1, \dots, v_r^\mathsf{T} A v_r$ is equal to $(A \bullet X)/r$. This implies that if $v_1^\mathsf{T} A v_1, \dots, v_r^\mathsf{T} A v_r$ are not all equal to $(A \bullet X)/r$, then there must exist $i, j \in \{1, \dots, r\}$ such that

$$v_i^\mathsf{T} A v_i < \frac{A \bullet X}{r} < v_j^\mathsf{T} A v_j.$$

Consider the function $g_{ij} : [0, \pi/2] \to \mathbf{R}$ such that

$$g_{ij}(\theta) = (\cos(\theta) v_i + \sin(\theta) v_j)^\mathsf{T} A (\cos(\theta) v_i + \sin(\theta) v_j).$$

It is clear that g_{ij} is a continuous function of θ. Moreover, we have that

$$g_{ij}(0) = v_i^\mathsf{T} A v_i < \frac{A \bullet X}{r} \quad \text{and} \quad g_{ij}(\pi/2) = v_j^\mathsf{T} A v_j > \frac{A \bullet X}{r}.$$

Therefore, we can use the intermediate-value theorem to conclude that there exists $\hat{\theta} \in (0, \pi/2)$ such that $g_{ij}(\hat{\theta}) = (A \bullet X)/r$. Suppose we replace v_i and v_j by

$$\tilde{v}_i = \cos(\hat{\theta}) v_i + \sin(\hat{\theta}) v_j \quad \text{and} \quad \tilde{v}_j = -\sin(\hat{\theta}) v_i + \cos(\hat{\theta}) v_j.$$

We can write the vector of replacement variables as

$$\begin{bmatrix} \tilde{v}_i & \tilde{v}_j \end{bmatrix} = \begin{bmatrix} v_i & v_j \end{bmatrix} Q,$$

where we define the matrix $Q \in \mathbf{R}^{2 \times 2}$ such that

$$Q = \begin{bmatrix} \cos(\hat{\theta}) & -\sin(\hat{\theta}) \\ \sin(\hat{\theta}) & \cos(\hat{\theta}) \end{bmatrix}.$$

Since Q is orthogonal, we have that

$$\begin{aligned} \tilde{v}_i \tilde{v}_i^\mathsf{T} + \tilde{v}_j \tilde{v}_j^\mathsf{T} &= \begin{bmatrix} \tilde{v}_i & \tilde{v}_j \end{bmatrix} \begin{bmatrix} \tilde{v}_i & \tilde{v}_j \end{bmatrix}^\mathsf{T} \\ &= \left(\begin{bmatrix} v_i & v_j \end{bmatrix} Q \right) \left(\begin{bmatrix} v_i & v_j \end{bmatrix} Q \right)^\mathsf{T} \\ &= \begin{bmatrix} v_i & v_j \end{bmatrix} (Q Q^\mathsf{T}) \begin{bmatrix} v_i & v_j \end{bmatrix}^\mathsf{T} \\ &= \begin{bmatrix} v_i & v_j \end{bmatrix} \begin{bmatrix} v_i & v_j \end{bmatrix}^\mathsf{T} \\ &= v_i v_i^\mathsf{T} + v_j v_j^\mathsf{T}. \end{aligned}$$

Thus, we still have a dyadic expansion for X after replacing v_i and v_j by $\tilde{v}_i$ and $\tilde{v}_j$. Moreover, we chose $\hat{\theta}$ so that

$$\tilde{v}_i^\mathsf{T} A \tilde{v}_i = g_{ij}(\hat{\theta}) = \frac{A \bullet X}{r}.$$

Combining these observations, we see that each iteration of the loop in Algorithm A.1 strictly reduces the number of indices k such that

$v_k^\mathsf{T} A v_k \neq (A \bullet X)/r$. This implies that the algorithm terminates after at most $r-1$ iterations, and that $v_1, \dots, v_r$ satisfy

$$\sum_{k=1}^{r} v_k v_k^\mathsf{T} = X \quad \text{and} \quad v_k^\mathsf{T} A v_k = \frac{A \bullet X}{r}, \quad k = 1, \dots, r$$

when the algorithm terminates. Therefore, we can take $x_1, \dots, x_r$ to be the output of Algorithm A.1. □

The following lemma shows how to minimize a general quadratic function of a vector variable. This result is used in the next section, but is included here because it is a result from linear algebra.

Lemma A.11. Suppose $P \in \mathbf{S}^n$, $q \in \mathbf{R}^n$, and $r \in \mathbf{R}$. Then

$$\min_{x \in \mathbf{R}^n} \left\{ x^\mathsf{T} P x + 2q^\mathsf{T} x + r \right\} = \begin{cases} r - q^\mathsf{T} P^\dagger q & P \succeq 0,\ q \in \mathbf{range}(P), \\ -\infty & \text{otherwise.} \end{cases}$$

Proof. There are three cases to consider.

(1) First, consider the case when $P \not\succeq 0$. Then there exists a vector $x_1 \in \mathbf{R}$ such that $x_1^\mathsf{T} P x_1 < 0$, and the function

$$\begin{aligned} g(t) &= (tx_1)^\mathsf{T} P (tx_1) + 2q^\mathsf{T}(tx_1) + r \\ &= (x_1^\mathsf{T} P x_1) t^2 + 2(q^\mathsf{T} x_1) t + r \end{aligned}$$

is a strictly concave single-variable quadratic function. Since such functions are unbounded below, we can conclude that

$$\min_{x \in \mathbf{R}^n} \left\{ x^\mathsf{T} P x + 2q^\mathsf{T} x + r \right\} = -\infty.$$

(2) Next, consider the case when $q \notin \mathbf{range}(P)$. Let x_2 be the orthogonal projection of q onto $\mathbf{range}(P)^\perp = \mathbf{null}(P)$. Because $q \notin \mathbf{range}(P)$, x_2 must be nonzero. Then the function

$$h(t) = (tx_2)^\mathsf{T} P (tx_2) + 2q^\mathsf{T}(tx_2) + r = 2\|x_2\|^2 t + r$$

is a non-constant linear function. Since such functions are unbounded below, we can conclude that

$$\min_{x \in \mathbf{R}^n} \left\{ x^\mathsf{T} P x + 2q^\mathsf{T} x + r \right\} = -\infty.$$

(3) Finally, consider the case when $P \succeq 0$ and $q \in \mathbf{range}(P)$. The orthogonal projection of q onto $\mathbf{range}(P)$ is $PP^\dagger q$. Our assumption that $q \in \mathbf{range}(P)$ implies that $q = PP^\dagger q$. Then completing the square gives

$$x^\mathsf{T} Px + 2q^\mathsf{T} x + r = (x + P^\dagger q)^\mathsf{T} P(x + P^\dagger q) + (r - q^\mathsf{T} P^\dagger q).$$

Since $P \succeq 0$, this implies that

$$x^\mathsf{T} Px + 2q^\mathsf{T} x + r \geq r - q^\mathsf{T} P^\dagger q$$

for all $x \in \mathbf{R}^n$. Moreover, equality holds if we choose $x = -P^\dagger q$. Therefore, we can conclude that

$$\min_{x \in \mathbf{R}^n} \left\{ x^\mathsf{T} Px + 2q^\mathsf{T} x + r \right\} = r - q^\mathsf{T} P^\dagger q.$$

□

The following result is frequently called the Schur-complement characterization of positive-semidefinite block matrices, and is often used to convert inequalities involving quadratic forms into matrix inequalities. We will see an example of such a conversion in the next section when we show that the Lagrangian relaxation of a quadratic optimization problem is a semidefinite program.

Corollary A.12. Suppose $A \in \mathbf{S}^n$, $B \in \mathbf{R}^{n \times m}$, and $C \in \mathbf{S}^m$. The matrix

$$M = \begin{bmatrix} A & B \\ B^\mathsf{T} & C \end{bmatrix}$$

is positive semidefinite if and only if $A \succeq 0$, $\mathbf{range}(B) \subset \mathbf{range}(A)$, and $C - B^\mathsf{T} A^\dagger B \succeq 0$.

Proof. We have that $M \succeq 0$ if and only if

$$\min_{(x,y) \in \mathbf{R}^n \times \mathbf{R}^m} \left\{ \begin{bmatrix} x \\ y \end{bmatrix}^\mathsf{T} \begin{bmatrix} A & B \\ B^\mathsf{T} & C \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \right\} = 0.$$

Using Lemma A.11, we find that

$$\begin{aligned}
&\min_{(x,y)\in\mathbf{R}^n\times\mathbf{R}^m}\left\{\begin{bmatrix}x\\y\end{bmatrix}^\mathsf{T}\begin{bmatrix}A & B\\B^\mathsf{T} & C\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix}\right\}\\
&\quad=\min_{y\in\mathbf{R}^m}\left\{\min_{x\in\mathbf{R}^n}\left\{x^\mathsf{T}Ax+2x^\mathsf{T}By+y^\mathsf{T}Cy\right\}\right\}\\
&\quad=\min_{y\in\mathbf{R}^m}\left\{\begin{matrix}y^\mathsf{T}(C-B^\mathsf{T}A^\dagger B)y & A\succeq 0,\ By\in\mathbf{range}(A),\\ -\infty & \text{otherwise}\end{matrix}\right\}\\
&\quad=\begin{cases}0 & A\succeq 0,\ \mathbf{range}(B)\subset\mathbf{range}(A),\ C-B^\mathsf{T}A^\dagger B\succeq 0,\\ -\infty & \text{otherwise.}\end{cases}
\end{aligned}$$

Thus, we see that $M \succeq 0$ if and only if $A \succeq 0$, $\mathbf{range}(B) \subset \mathbf{range}(A)$, and $C - B^\mathsf{T}A^\dagger B \succeq 0$. □

Remark A.1. The matrix $S = C - B^\mathsf{T}A^\dagger B$ in Corollary A.12 is called the Schur complement of A in M.

A.2 Optimization

A.2.1 Lagrangian duality

Consider a general optimization problem of the form

$$\begin{array}{ll}\text{minimize} & f_0(x)\\ \text{subject to} & f_i(x)\le 0,\quad i=1,\ldots,m\\ & h_i(x)=0,\quad i=1,\ldots,p.\end{array}$$

The Lagrangian of this problem is defined to be

$$L(x,\lambda,\nu)=f_0(x)+\sum_{i=1}^m\lambda_i f_i(x)+\sum_{i=1}^p\nu_i h_i(x),$$

and the Lagrange dual function is defined to be

$$g(\lambda,\nu)=\inf_{x\in\mathcal{D}}L(x,\lambda,\nu),$$

where $\mathcal{D}=(\cap_{i=0}^m\,\mathbf{dom}(f_i))\cap(\cap_{i=1}^p\,\mathbf{dom}(h_i))$ is the domain of the optimization problem. The vectors λ and ν are called Lagrange multipliers for the inequality and equality constraints, respectively. For every

$\tilde{x} \in \mathbf{R}^n$ that is feasible for the original optimization problem, and all vectors $\lambda \in \mathbf{R}_+^m$ and $\nu \in \mathbf{R}^p$, we have that

$$L(\tilde{x}, \lambda, \nu) = f_0(\tilde{x}) + \sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) + \sum_{i=1}^{p} \nu_i h_i(\tilde{x}) \leq f_0(\tilde{x})$$

because $f_i(\tilde{x}) \leq 0$ and $\lambda_i \geq 0$ for $i = 1, \ldots, m$, and $h_i(\tilde{x}) = 0$ for $i = 1, \ldots, p$. Thus, we have that

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) \leq L(\tilde{x}, \lambda, \nu) \leq f_0(\tilde{x}).$$

Therefore, for every feasible $\tilde{x}$, every nonnegative vector λ, and every vector ν, we have that $g(\lambda, \nu) \leq f_0(\tilde{x})$. This implies that $g(\lambda, \nu)$ is a lower bound on the optimal value of the original optimization problem. We call $g(\lambda, \nu)$ the Lagrange dual function. In order to find the tightest such lower bound, we solve the Lagrange dual problem

$$\begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \geq 0. \end{array}$$

Note that the Lagrange dual problem is convex even when the original optimization problem is not: the dual function is concave because it is the pointwise infimum of affine functions of the dual variables λ and ν. We have argued that the Lagrange dual problem provides a lower bound on the optimal value of the original optimization problem; this result is called weak duality. The difference between the optimal value of the original optimization problem and the optimal value of the Lagrange dual problem is called the duality gap. Because the Lagrange dual problem is convex, and provides a lower bound on the original optimization problem, we sometimes call it the Lagrangian relaxation of the original problem.

Intuitively, the Lagrangian dual problem uses linear combinations of the constraints to construct lower bounds on the optimal value of an optimization problem. Perhaps surprisingly, this simple method of generating lower bounds actually produces tight lower bounds in many cases. We say that strong duality holds when the optimal value of the Lagrangian dual problem is equal to the optimal value of the original optimization problem. Although strong duality is not guaranteed to hold in general, there are a number of sufficient conditions for strong duality that are easy to check and often satisfied in practical problems. Sufficient conditions for strong duality that are stated in terms of properties of the constraint functions are sometimes called constraint qualifications. The most common constraint qualification is Slater's theorem; in order to state this result, we need some additional terminology: we say that an optimization problem is strictly feasible if there exists a vector $x_0 \in \mathbf{R}^n$ such that $f_i(x_0) < 0$ for $i = 1, \ldots, m$, and $h_i(x_0) = 0$ for $i = 1, \ldots, p$; a convex optimization problem is an optimization problem such that $f_0, \ldots, f_m$ are convex, and $h_1, \ldots, h_p$ are affine. We are now prepared to state Slater's theorem. (We omit the proof, which is technical, and given in many texts on convex optimization, such as [10].)

Theorem A.13 (Slater's theorem)**.** There is strong duality for strictly feasible convex optimization problems.

A.2.2 Linear programming

A linear program in primal standard form is an optimization problem of the form

$$\begin{array}{ll} \text{minimize} & c^\mathsf{T} x \\ \text{subject to} & Ax = b \\ & x \geq 0, \end{array} \tag{LP}$$

where $x \in \mathbf{R}^n$ is the optimization variable, and $A \in \mathbf{R}^{m \times n}$, $b \in \mathbf{R}^m$, and $c \in \mathbf{R}^n$ are problem data. The Lagrangian for this problem is

$$L(x, y, s) = c^\mathsf{T} x + y^\mathsf{T}(b - Ax) - s^\mathsf{T} x = (c - (A^\mathsf{T} y + s))^\mathsf{T} x + b^\mathsf{T} y.$$

This is a linear function of x, which is unbounded below unless the coefficient of x is equal to zero. Thus, we have that

$$\begin{aligned} g(y, s) &= \min_{x \in \mathbf{R}^n} L(x, y, s) \\ &= \min_{x \in \mathbf{R}^n} \left\{ (c - (A^\mathsf{T} y + s))^\mathsf{T} x + b^\mathsf{T} y \right\} \\ &= \begin{cases} b^\mathsf{T} y & A^\mathsf{T} y + s = c, \\ -\infty & \text{otherwise.} \end{cases} \end{aligned}$$

Then the Lagrange dual problem of (LP) is

$$\begin{array}{ll} \text{maximize} & b^\mathsf{T} y \\ \text{subject to} & A^\mathsf{T} y + s = c \\ & s \geq 0, \end{array} \tag{LD}$$

where $y \in \mathbf{R}^m$ and $s \in \mathbf{R}^n$ are the optimization variables. Note that (LD) is also a linear program, and that its dual is (LP). We call (LD) a linear program in dual standard form. It is a remarkable fact that strong duality holds for linear programs whenever one of (LP) and (LD) is feasible.

A.2.3 Semidefinite programming

A semidefinite program (SDP) in primal standard form is an optimization problem of the form

$$\begin{array}{ll} \text{minimize} & C \bullet X \\ \text{subject to} & A_i \bullet X = b_i, \quad i = 1, \ldots, m \\ & X \succeq 0, \end{array} \tag{SDP}$$

where $X \in \mathbf{S}^n$ is the optimization variable, and $A_1, \ldots, A_m \in \mathbf{S}^n$, $b \in \mathbf{R}^m$, and $C \in \mathbf{S}^n$ are problem data. The Lagrangian for (SDP) is

$$\begin{aligned} L(X, y, S) &= C \bullet X + \sum_{i=1}^m y_i (b_i - A_i \bullet X) - S \bullet X \\ &= \left(C - \left[\sum_{i=1}^m y_i A_i + S \right] \right) \bullet X + b^\mathsf{T} y, \end{aligned}$$

the Lagrange dual function is

$$\begin{aligned} g(y, S) &= \min_{X \in \mathbf{S}^n} L(X, y, S) \\ &= \min_{X \in \mathbf{S}^n} \left\{ \left(C - \left[\sum_{i=1}^m y_i A_i + S \right] \right) \bullet X + b^\mathsf{T} y \right\} \\ &= \begin{cases} b^\mathsf{T} y & \sum_{i=1}^m y_i A_i + S = C, \\ -\infty & \text{otherwise.} \end{cases} \end{aligned}$$

and the Lagrange dual problem is

$$\begin{array}{ll} \text{maximize} & b^{\mathsf{T}} y \\ \text{subject to} & \sum_{i=1}^{m} y_i A_i + S = C \\ & S \succeq 0, \end{array} \tag{SDD}$$

where $y \in \mathbf{R}^m$ and $S \in \mathbf{S}^n$ are the optimization variables. We call (SDD) a semidefinite program in dual standard form, and it is easy to check that its Lagrangian relaxation is (SDP).

Note the strong similarities between (LP) and (SDP). In order to emphasize these similarities, we define the linear operator $\mathcal{A} : \mathbf{S}^n \to \mathbf{R}^m$ such that

$$\mathcal{A}(X) = \begin{bmatrix} A_1 \bullet X \\ \vdots \\ A_m \bullet X \end{bmatrix}.$$

Then we can write (SDP) as

$$\begin{array}{ll} \text{minimize} & C \bullet X \\ \text{subject to} & \mathcal{A}(X) = b \\ & X \succeq 0. \end{array}$$

Observe that

$$\langle y, \mathcal{A}(X) \rangle = \sum_{i=1}^{m} y_i (A_i \bullet X) = \left(\sum_{i=1}^{m} y_i A_i \right) \bullet X = \left\langle \sum_{i=1}^{m} y_i A_i, X \right\rangle.$$

Since the adjoint operator $\mathcal{A}^* : \mathbf{R}^m \to \mathbf{S}^n$ of $\mathcal{A} : \mathbf{S}^n \to \mathbf{R}^m$ is defined by the relation $\langle y, \mathcal{A}(X) \rangle = \langle \mathcal{A}^*(y), X \rangle$, we have that

$$\mathcal{A}^*(y) = \sum_{i=1}^{m} y_i A_i.$$

Therefore, we can write (SDD) as

$$\begin{array}{ll} \text{maximize} & b^{\mathsf{T}} y \\ \text{subject to} & \mathcal{A}^*(y) + S = C \\ & S \succeq 0 \end{array}$$

in order to emphasize the similarity to (LD). Since SDPs are convex optimization problems, strong duality holds whenever one of (SDP) and (SDD) is strictly feasible.

A.2.4 Polynomial optimization

A polynomial optimization problem is an optimization problem of the form

$$\begin{array}{ll} \text{minimize} & p_0(x) \\ \text{subject to} & p_i(x) \leq 0, \quad i = 1, \ldots, m, \end{array}$$

where $x \in \mathbf{R}^n$ is the optimization variable, and $p_0, \ldots, p_m$ are polynomials. We claim that every polynomial optimization problem can be reformulated as a quadratic optimization problem (that is, a polynomial optimization problem with quadratic polynomials). We will not prove this claim in general, but the following example will hopefully convince the reader that such a reformulation is always possible.

Example A.1. Consider the polynomial optimization problem

$$\begin{array}{ll} \text{minimize} & x_1^3 + x_1^2 x_2 \\ \text{subject to} & x_2^4 - 1 \leq 0. \end{array}$$

Introduce the variables $x_3 = x_1^2$ and $x_4 = x_2^2$. Then we can write our optimization problem as

$$\begin{array}{ll} \text{minimize} & x_1 x_3 + x_2 x_3 \\ \text{subject to} & x_4^2 - 1 \leq 0 \\ & x_1^2 - x_3 \leq 0 \\ & x_3 - x_1^2 \leq 0 \\ & x_2^2 - x_4 \leq 0 \\ & x_4 - x_2^2 \leq 0. \end{array}$$

This is a quadratic optimization problem. We can convert every polynomial optimization problem into a quadratic optimization problem by introducing variables in this way.

Thus, we can restrict our attention to quadratic optimization problems, which can be written in the form

$$\begin{array}{ll} \text{minimize} & x^\mathsf{T} P_0 x + 2q_0^\mathsf{T} x + r_0 \\ \text{subject to} & x^\mathsf{T} P_i x + 2q_i^\mathsf{T} x + r_i \leq 0, \quad i = 1, \ldots, m, \end{array}$$

where $x \in \mathbf{R}^n$ is the optimization variable, and $P_0, \ldots, P_m \in \mathbf{S}^n$, $q_0, \ldots, q_m \in \mathbf{R}^n$, and $r_0, \ldots, r_m \in \mathbf{R}$ are problem data. The Lagrangian for this problem is

$$
\begin{aligned}
L(x,\lambda) &= (x^\mathsf{T} P_0 x + 2q_0^\mathsf{T} x + r_0) + \sum_{i=1}^m \lambda_i (x^\mathsf{T} P_i x + 2q_i^\mathsf{T} x + r_i) \\
&= x^\mathsf{T} \left(P_0 + \sum_{i=1}^m \lambda_i P_i \right) x + 2 \left(q_0 + \sum_{i=1}^m \lambda_i q_i \right)^\mathsf{T} x \\
&\qquad + \left(r_0 + \sum_{i=1}^m \lambda_i r_i \right) \\
&= x^\mathsf{T} P(\lambda) x + 2q(\lambda)^\mathsf{T} x + r(\lambda),
\end{aligned}
$$

where we define

$$
P(\lambda) = P_0 + \sum_{i=1}^m \lambda_i P_i, \quad q(\lambda) = q_0 + \sum_{i=1}^m \lambda_i q_i,
$$
$$
\text{and} \quad r(\lambda) = r_0 + \sum_{i=1}^m \lambda_i r_i.
$$

We can use Lemma A.11 to compute the Lagrange dual function:

$$
\begin{aligned}
g(\lambda) &= \min_{x \in \mathbf{R}^n} L(x,\lambda) \\
&= \begin{cases} r(\lambda) - q(\lambda)^\mathsf{T} P(\lambda)^\dagger q(\lambda) & P(\lambda) \succeq 0, q(\lambda) \in \mathbf{range}(P(\lambda)), \\ -\infty & \text{otherwise.} \end{cases}
\end{aligned}
$$

Then the Lagrange dual problem is

$$
\begin{array}{ll}
\text{maximize} & r(\lambda) - q(\lambda)^\mathsf{T} P(\lambda)^\dagger q(\lambda) \\
\text{subject to} & P(\lambda) \succeq 0 \\
& q(\lambda) \in \mathbf{range}(P(\lambda)) \\
& \lambda \geq 0.
\end{array}
$$

Applying an epigraph transformation, we obtain the equivalent problem

$$
\begin{array}{ll}
\text{maximize} & \mu \\
\text{subject to} & \mu \leq r(\lambda) - q(\lambda)^\mathsf{T} P(\lambda)^\dagger q(\lambda) \\
& P(\lambda) \succeq 0 \\
& q(\lambda) \in \mathbf{range}(P(\lambda)) \\
& \lambda \geq 0.
\end{array}
$$

Using Corollary A.12, we can combine the first three constraints into a matrix inequality:

$$\begin{array}{ll} \text{maximize} & \mu \\ \text{subject to} & \begin{bmatrix} P(\lambda) & q(\lambda) \\ q(\lambda)^{\mathsf{T}} & r(\lambda)-\mu \end{bmatrix} \succeq 0 \\ & \lambda \geq 0. \end{array}$$

Introducing a slack variable for the matrix inequality, and recalling our definitions of $P(\lambda)$, $q(\lambda)$, and $r(\lambda)$, we arrive at the problem

$$\begin{array}{ll} \text{maximize} & \mu \\ \text{subject to} & \mu E_{n+1,n+1} - \sum_{i=1}^{m} \lambda_i C_i + S = C_0 \\ & \lambda \geq 0 \\ & S \succeq 0, \end{array}$$

where the optimization variables are $\lambda \in \mathbf{R}^m$, $\mu \in \mathbf{R}$, and $S \in \mathbf{S}^{n+1}$, and we define

$$E_{n+1,n+1} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad C_i = \begin{bmatrix} P_i & q_i \\ q_i^{\mathsf{T}} & r_i \end{bmatrix}, \quad i = 0, \ldots, m.$$

Note that this is a semidefinite program in dual form. (It is not quite in standard dual form because of the inequality $\lambda \geq 0$, but this is a minor difference). The corresponding primal-form semidefinite program is

$$\begin{array}{ll} \text{minimize} & C_0 \bullet X \\ \text{subject to} & C_i \bullet X \leq 0, \quad i = 1, \ldots, m \\ & E_{n+1,n+1} \bullet X = 1 \\ & X \succeq 0. \end{array}$$

We can also derive the primal form of the relaxation directly from the original quadratic optimization problem. First, we write the quadratic optimization problem as

$$\begin{array}{ll} \text{minimize} & C_0 \bullet \left(\begin{bmatrix} x \\ 1 \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix}^{\mathsf{T}} \right) \\ \text{subject to} & C_i \bullet \left(\begin{bmatrix} x \\ 1 \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix}^{\mathsf{T}} \right) \leq 0, \quad i = 1, \ldots, m. \end{array}$$

Now we observe that

$$\left\{ \begin{bmatrix} x \\ 1 \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix}^{\mathsf{T}} \,\middle|\, x \in \mathbf{R}^n \right\}$$
$$= \{X \in \mathbf{S}^{n+1} \mid E_{n+1,n+1} \bullet X = 1,\ X \succeq 0,\ \ \mathbf{rank}(X) = 1\}.$$

Thus, we can reformulate our quadratic optimization problem in terms of a matrix variable:

$$\begin{array}{ll} \text{minimize} & C_0 \bullet X \\ \text{subject to} & C_i \bullet X \leq 0, \quad i = 1, \ldots, m \\ & E_{n+1,n+1} \bullet X = 1 \\ & X \succeq 0 \\ & \mathbf{rank}(X) = 1. \end{array}$$

Relaxing the rank constraint (which is difficult to handle) yields the primal-form semidefinite relaxation that we found above. The derivation of the primal relaxation directly from the quadratic optimization problem is important because it shows us when the primal relaxation is equivalent to the original problem: we require that there exist a rank-1 solution of the SDP. Moreover, the derivation shows us how to recover a solution of the original problem from a rank-1 solution of the SDP relaxation. In particular, if $X = \tilde{x}\tilde{x}^{\mathsf{T}}$ is a rank-1 solution of the SDP, then $x = \tilde{x}_{n+1}\tilde{x}_{1:n}$ is a solution of the original quadratic optimization problem, where $\tilde{x}_{1:n} = (\tilde{x}_1, \ldots, \tilde{x}_n)$ is the vector consisting of the first n components of $\tilde{x}$.

B

Linear Programs and Cardinality

In this appendix we review techniques for finding reduced-cardinality solutions of linear programs. Techniques for sparsification (also called purification) of solutions to linear programs are often included in proofs of the so-called fundamental theorem of linear programming [65], and are closely related to Carathéodory's theorem [24]. Because computing a minimum-cardinality solution of a system of linear equations is an NP-hard problem [36], we do not hope to develop a sparsification algorithm that always returns a minimum-cardinality solution of the linear program. Nevertheless, we will give an algorithm that is often effective at reducing the cardinality of a given solution of a linear program, and has some performance guarantees. Such an algorithm is important because sparse solutions are often desirable, and interior-point algorithms typically converge to the analytic center of the optimal face, which is a maximum-cardinality solution. We then give a theorem relating cardinality and the uniqueness of solutions for linear programs.

B.1 Sparsification for linear programs

Suppose we are given a solution x of (LP), and we want to find another solution x^+ with $\mathbf{card}(x^+) < \mathbf{card}(x)$. The process of using a given x to find such an x^+ is called sparsification, rounding, or purification. If we had an efficient method for sparsification that worked on every solution that does not have minimum cardinality, then we could find a minimum-cardinality solution by applying our sparsification method at most n times. However, we know that the problem of finding a minimum-cardinality solution of (LP) is NP-hard. Thus, we do not expect to find an efficient sparsification algorithm that always works. Nonetheless, we still hope to find a method that often performs well in practice. We begin by making the assumption

$$x_j^+ = 0 \quad \text{whenever} \quad x_j = 0. \tag{B.1}$$

The following example shows that this assumption can yield suboptimal results in some cases.

Example B.1. Consider the LP feasibility problem

$$\begin{aligned} x_i + x_n &= 1, \quad i = 1, \ldots, n-1 \\ x &\geq 0, \end{aligned}$$

and suppose we are given the solution $x = (1, \ldots, 1, 0)$. If we assume that $x_j^+ = 0$ whenever $x_j = 0$, then we have that $x_n^+ = 0$. Then the constraint $x_i^+ + x_n^+ = 1$ implies that $x_i^+ = 1$ for $i = 1, \ldots, n-1$. Thus, we have that $x^+ = x$, and we are unable to sparsify x. However, x is not a minimum-cardinality solution of the feasibility problem: we have that $\mathbf{card}(x) = n - 1$, but $\tilde{x} = (0, \ldots, 0, 1)$ is a solution with $\mathbf{card}(\tilde{x}) = 1$.

Example B.1 shows that assumption (B.1) may not only lead to suboptimal results, but may even lead to arbitrarily poor results: for every positive integer n, there is an instance of (LP) and a corresponding initial solution such that our algorithm returns a solution whose cardinality is $n - 1$ times the cardinality of a minimum-cardinality solution. However, because we do not expect to find an algorithm that works on every instance of (LP), we need to make a suboptimal assumption at some point.

We have stated our assumption as $x_j^+ = 0$ whenever $x_j = 0$. Although this statement is clear and intuitive, a different formulation will prove useful in the development of our algorithm. An equivalent assumption is that x^+ has the form

$$x^+ = (\mathbf{1} + \alpha\delta) \circ x,$$

where we think of $\delta \in \mathbf{R}^n$ as an update direction, and $\alpha \in \mathbf{R}$ as a step size. We will also sometimes find it convenient to write x^+ as

$$x^+ = x + \alpha X \delta,$$

where $X = \mathbf{diag}(x)$. We want to choose α and δ such that x^+ is a solution of (LP), and $\mathbf{card}(x^+) < \mathbf{card}(x)$. Since the cardinality of x^+ is strictly less than that of x, we must have that $x^+ \neq x$, and hence that $\alpha \neq 0$.

- In order to maintain optimality, we require that

  $$c^\mathsf{T} x^+ = c^\mathsf{T} x.$$

  Substituting in $x^+ = x + \alpha X\delta$ and simplifying, we obtain the condition

  $$(Xc)^\mathsf{T} \delta = 0.$$

- We also need x^+ to satisfy the equality constraint

  $$Ax^+ = b.$$

  As before, we substitute in our expression for x^+, and then simplify; this gives the condition

  $$AX\delta = 0.$$

- The updated solution must satisfy $x^+ = (1 + \alpha\delta) \circ x \geq 0$. Since $x \geq 0$, this is equivalent to the condition

  $$1 + \alpha\delta_j \geq 0 \quad \text{whenever} \quad x_j \neq 0.$$

  Because the value of δ_j does not affect x^+ if $x_j = 0$, we will assume that $\delta_j = 0$ whenever $x_j = 0$. Then our condition for $x^+ \geq 0$ becomes

  $$1 + \alpha\delta_j \geq 0, \quad j = 1, \ldots, n.$$

- Finally, we have that $\mathbf{card}(x^+) < \mathbf{card}(x)$ if and only if there exists $j \in \{1, \ldots, n\}$ such that $1 + \alpha\delta_j = 0$ and $x_j \neq 0$.

To summarize this analysis, we want to choose α and δ satisfying the following conditions:

$$\begin{aligned} (Xc)^{\mathsf{T}}\delta &= 0 \\ (AX)\delta &= 0 \\ \delta_j &= 0 \quad \text{whenever } x_j = 0 \\ 1 + \alpha\delta_j &\geq 0, \quad j = 1, \ldots, n \\ 1 + \alpha\delta_{\hat{\jmath}} &= 0 \quad \text{for some } \hat{\jmath} \in \{1, \ldots, n\}. \end{aligned}$$

It turns out that the first constraint is implied by the second constraint. The main idea is that because $x_j^+ = 0$ whenever $x_j = 0$, the updated solution x^+ automatically satisfies complementary slackness. Therefore, x^+ is optimal whenever it is feasible. We make this argument more precise in the proof of the following proposition.

Proposition B.1. If x is a solution of (LP), then $(AX)\delta = 0$ implies $(Xc)^{\mathsf{T}}\delta = 0$.

Proof. Let x be a solution of (LP), and let (y, s) be a solution of (LD). Then x and (y, s) satisfy the KKT conditions

$$\begin{aligned} A^{\mathsf{T}}y + s &= c \\ Ax &= b \\ x, s &\geq 0 \\ s \circ x &= 0. \end{aligned}$$

Now observe that

$$(Xc)^{\mathsf{T}}\delta = (X(A^{\mathsf{T}}y + s))^{\mathsf{T}}\delta = y^{\mathsf{T}}(AX)\delta + (s \circ x)\delta,$$

where we use $A^{\mathsf{T}}y + s = c$ in the first step, and $Xs = s \circ x$ in the second step. Since we assume that $(AX)\delta = 0$, and complementary slackness implies that $s \circ x = 0$, we see that $(Xc)^{\mathsf{T}}\delta = 0$. □

Note that this argument only works if x is a solution of (LP). In particular, if we had an arbitrary feasible point x, and we wanted to find another feasible point x^+ with the same objective value, we could not ignore the condition $(Xc)^{\mathsf{T}}\delta = 0$.

An algorithm for LP sparsification. An algorithm for LP sparsification is given in Algorithm B.1. Using the observations above, we will prove that the algorithm returns a solution of the LP, and derive a bound on the cardinality of this solution. The statement of our algorithm uses one piece of nonstandard notation: let $I[x] \in \mathbf{R}^{(n-\mathbf{card}(x))\times n}$ denote the matrix whose rows are the rows of the $n \times n$ identity matrix corresponding to the zero components of the vector $x \in \mathbf{R}^n$.

Algorithm B.1: sparsification for linear programs

```
Input: a solution $x$ of (LP)
repeat
    find a nonzero $\delta \in \mathbf{null}\left(\begin{bmatrix} AX \\ I[x] \end{bmatrix}\right)$ (if possible)
    find $\hat{\jmath} \in \underset{j=1,\dots,n}{\operatorname{argmax}}\{|\delta_j|\}$
    $\alpha := -1/\delta_{\hat{\jmath}}$
    $x := (\mathbf{1} + \alpha\delta) \circ x$
until $\mathbf{null}\left(\begin{bmatrix} AX \\ I[x] \end{bmatrix}\right) = \{0\}$
```

Proposition B.2. Algorithm B.1 returns a solution x^+ of (LP) with $\mathbf{card}(x^+) \le \mathbf{card}(x)$.

Proof. In our preliminary analysis of the LP-sparsification problem, we showed that x^+ is a solution of (LP) with $\mathbf{card}(x^+) < \mathbf{card}(x)$ if α and δ satisfy the following properties:

$$
\begin{aligned}
(AX)\delta &= 0 && && \text{(B.2)}\\
\delta_j &= 0 && \text{whenever } x_j = 0, && \text{(B.3)}\\
1 + \alpha\delta_j &\ge 0, && j = 1, \dots, n && \text{(B.4)}\\
1 + \alpha\delta_{\hat{\jmath}} &= 0 && \text{for some } \hat{\jmath} \in \{1, \dots, n\}. && \text{(B.5)}
\end{aligned}
$$

In Algorithm B.1 we choose δ such that

$$\begin{bmatrix} AX \\ I[x] \end{bmatrix} \delta = 0.$$

The first block of this equation says that $(AX)\delta = 0$, while the second block says that $I[x]\delta = 0$. Recall that $I[x]$ denotes the matrix whose rows are the rows of the identity matrix corresponding to the zero components of x. Therefore, we have that $\delta_j = 0$ whenever $x_j = 0$. In summary our choice of δ implies that (B.2) and (B.3) are satisfied.

We choose $\alpha = -1/\delta_{\hat{\jmath}}$, where $\delta_{\hat{\jmath}}$ is a maximum-magnitude component of δ. This choice of α implies that $1 + \alpha\delta_{\hat{\jmath}} = 0$, and

$$1 + \alpha\delta_j \geq 1 - |\alpha||\delta_j| = 1 - \left|\frac{\delta_j}{\delta_{\hat{\jmath}}}\right| \geq 0$$

for $j = 1, \ldots, n$. Thus, conditions (B.4) and (B.5) are also satisfied.

This analysis shows that, after each iteration of the algorithm, x is still a solution of (LP), and is at least as sparse as the original solution provided to the algorithm. (If at least one iteration of the loop executes, then x is strictly sparser than the original solution.) □

Theorem B.1. If (LP) is solvable, then it has a solution x with $\mathbf{card}(x) \leq m$. Moreover, Algorithm B.1 finds such a solution.

Proof. The termination condition for Algorithm B.1 is

$$\mathbf{null}\left(\begin{bmatrix} AX \\ I[x] \end{bmatrix}\right) = \{0\},$$

where $\left[\begin{smallmatrix} AX \\ I[x] \end{smallmatrix}\right] \in \mathbf{R}^{(m+n-\mathbf{card}(x))\times n}$. Since every strictly fat matrix has a nontrivial nullspace, we must have that $m + n - \mathbf{card}(x) \geq n$ when the algorithm terminates. Rearranging this inequality, we find that Algorithm B.1 returns a solution x with $\mathbf{card}(x) \leq m$. □

The following example shows that the bound in Theorem B.1 is tight: that is, the bound cannot be improved without additional hypotheses.

Example B.2. Suppose $m \leq n$, and consider the LP feasibility problem

$$\begin{aligned} x_i &= 1, \quad i = 1, \ldots, m \\ x &\geq 0, \end{aligned}$$

with variable $x \in \mathbf{R}^n$. A minimum-cardinality solution of this problem is $x = e_1 + \cdots + e_m$, which satisfies $\mathbf{card}(x) = m$, where m is the number of linear equality constraints.

Remark B.1. Consider what happens when we apply Algorithm B.1 to an instance of (LP) with homogeneous equality constraints (that is, with $b = 0$). Then we can always choose $\delta \in \mathbf{R}^n$ such that

$$\delta_j = \begin{cases} 1 & x_j \neq 0, \\ 0 & x_j = 0. \end{cases}$$

This choice of δ works because

$$(AX)\delta = Ax = 0,$$

and $\mathbf{supp}(\delta) \subset \mathbf{supp}(x)$, where $\mathbf{supp}(z) = \{i \mid z_i \neq 0\}$ is called the support of the vector z. For this value of δ, we have that $\alpha = -1$, and hence that

$$x^+ = (\mathbf{1} + \alpha\delta) \circ x = (\mathbf{1} - \delta) \circ x = 0$$

since either $1 - \delta_j = 0$ or $x_j = 0$ for all $j = 1, \ldots, n$. Thus, Algorithm B.1 tells us that $x = 0$ is a solution of every solvable instance of (LP) with homogeneous equality constraints. Note in particular the (easy-to-overlook) hypothesis in Theorem B.1 that (LP) is solvable. For example, consider the linear program

$$\begin{array}{ll} \text{minimize} & -x_1 \\ \text{subject to} & x_2 = 0 \\ & x \geq 0. \end{array}$$

The linear constraint for this problem is homogeneous, but $x = 0$ is not a solution: the problem is unbounded below, and not solvable.

B.2 Cardinality and uniqueness for linear programs

The following theorems relate the uniqueness of a solution of (LP) to its cardinality.

Theorem B.2. A solution x of (LP) is unique if and only if

(i) x has the maximum cardinality among all solutions, and

(ii) $\mathbf{null}(A_{\mathbf{supp}(x)}) = \{0\}$,

where $A_{\mathbf{supp}(x)}$ is the matrix whose columns are the columns of A corresponding to the nonzero components of x.

Proof. First, suppose x is the unique solution of (LP). It is trivially true that x has the maximum cardinality among all solutions because it is the only solution. In order to show that $\mathbf{null}(A_{\mathbf{supp}(x)}) = \{0\}$, we will argue by contradiction. Suppose there exists a nonzero vector $w \in \mathbf{R}^n$ such that $\mathbf{supp}(w) \subset \mathbf{supp}(x)$ and $Aw = 0$. (This is equivalent to assuming that $\mathbf{null}(A_{\mathbf{supp}(x)}) \neq \{0\}$ because $\mathbf{supp}(w) \subset \mathbf{supp}(x)$ implies that $Aw = A_{\mathbf{supp}(x)} w_{\mathbf{supp}(x)}$.) Define $z \in \mathbf{R}^n$ such that

$$z_j = (X^\dagger w)_j = \begin{cases} w_j/x_j & j \in \mathbf{supp}(x), \\ 0 & \text{otherwise.} \end{cases}$$

Then we have that

$$(AX)z = Aw = 0 \quad \text{and} \quad I[x]z = 0,$$

where $I[x]$ is the matrix consisting of the rows of the identity matrix corresponding to the zero components of x. Thus, z is a nonzero vector in $\mathbf{null}\Big(\Big[\begin{smallmatrix} AX \\ I[x] \end{smallmatrix}\Big]\Big)$, and Algorithm B.1 is able to find a solution $\tilde{x}$ of LP whose cardinality is strictly less than that of x. This contradicts the assumption that x is the unique solution of LP, and thereby proves that $\mathbf{null}(A_{\mathbf{supp}(x)}) = \{0\}$.

Conversely, suppose that x and $\tilde{x}$ are distinct solutions of LP. We can assume without loss of generality that x has the maximum cardinality among all solutions of LP. First, observe that $(x + \tilde{x})/2$ is a solution of LP with

$$\mathbf{supp}((x + \tilde{x})/2) = \mathbf{supp}(x) \cup \mathbf{supp}(\tilde{x})$$

because x and $\tilde{x}$ are nonnegative vectors. This allows us to conclude that $\mathbf{supp}(\tilde{x}) \subset \mathbf{supp}(x)$ since otherwise $(x + \tilde{x})/2$ is a solution whose cardinality is strictly greater than that of x, which violates our assumption that x is a maximum-cardinality solution. Then we have that

$$A_{\mathbf{supp}(x)}(x_{\mathbf{supp}(x)} - \tilde{x}_{\mathbf{supp}(x)}) = Ax - A\tilde{x} = b - b = 0.$$

Additionally, $x_{\mathbf{supp}(x)} - \tilde{x}_{\mathbf{supp}(x)}$ is nonzero because x and $\tilde{x}$ are distinct and $\mathbf{supp}(\tilde{x}) \subset \mathbf{supp}(x)$. Thus, $x_{\mathbf{supp}(x)} - \tilde{x}_{\mathbf{supp}(x)}$ is a nonzero vector in $\mathbf{null}(A_{\mathbf{supp}(x)})$. □

Remark B.2. Note that $\mathbf{null}(A_{\mathbf{supp}(x)}) = \{0\}$ if and only if the columns of A corresponding to the nonzero components of x are linearly independent. We use the slightly more abstruse condition $\mathbf{null}(A_{\mathbf{supp}(x)}) = \{0\}$ because it more easily generalizes to the analogous result for semidefinite programming presented in the main body of the text.

Corollary B.3. If (LP) is solvable, and every solution has the same cardinality, then (LP) has a unique solution.

Proof. Let x be a solution of (LP). Since every solution of (LP) has the same cardinality, x must have the maximum cardinality among all solutions. If we can show that $\mathbf{null}(A_{\mathbf{supp}(x)}) = \{0\}$, then we can use Theorem B.2 to conclude that x is the unique solution of (LP). Suppose $\tilde{w} \in \mathbf{null}(A_{\mathbf{supp}(x)})$. Padding $\tilde{w}$ with zeros in the components corresponding to the zero components of x yields a vector $w \in \mathbf{R}^n$ such that $\mathbf{supp}(w) \subset \mathbf{supp}(x)$ and $w_{\mathbf{supp}(x)} = \tilde{w} \in \mathbf{null}(A_{\mathbf{supp}(x)})$. Another consequence of the fact that every solution has the same cardinality is that Algorithm B.1 must terminate on the first iteration: that is,

$$\mathbf{null}\left(\begin{bmatrix} AX \\ I[x] \end{bmatrix}\right) = \{0\}.$$

Define the vector $z \in \mathbf{R}^n$ such that

$$z_j = (X^\dagger w)_j = \begin{cases} w_j / x_j & j \in \mathbf{supp}(x), \\ 0 & \text{otherwise.} \end{cases}$$

Then we have that

$$(AX)z = Aw = A_{\mathbf{supp}(x)} w_{\mathbf{supp}(x)} = 0.$$

The condition $I[x]z = 0$ is also satisfied since

$$\mathbf{supp}(z) = \mathbf{supp}(w) \subset \mathbf{supp}(x).$$

Thus, we find that $z \in \mathbf{null}\left(\left[\begin{smallmatrix} AX \\ I[X] \end{smallmatrix}\right]\right) = \{0\}$, which allows us to conclude that $z = 0$. Then our definition of z implies that $w = 0$, and hence that $\tilde{w} = 0$, and $\mathbf{null}(A_{\mathbf{supp}(x)}) = \{0\}$. □

C

Technical Probability Lemmas

This appendix contains the proofs of some technical probability lemmas that are used in our analysis of rounding methods in Chapter 3. Although these results are standard, we include the proofs for completeness, and so the results can be referenced in the exact form in which they are needed rather than the more general forms in the literature.

C.1 Convex combinations of chi-squared random variables

More general versions of the lemmas in this section are given by Laurent and Massart [60]. We state and prove specialized lemmas that can be used directly in our analysis of randomized rounding methods.

Lemma C.1. Let $y_1, \ldots, y_r$ be independent, identically distributed chi-squared random variables with d degrees of freedom, and $\theta_1, \ldots, \theta_r$ be nonnegative real numbers such that $\theta_1 + \cdots + \theta_r = 1$. Then, for all $c > 1$, we have that

$$\mathbf{prob}\left(\sum_{i=1}^{r} \theta_i y_i \geq cd\right) \leq (ec\exp(-c))^{\frac{d}{2}}.$$

Proof. For all $t > 0$, we have that

$$\begin{aligned}\mathbf{prob}\left(\sum_{i=1}^{r}\theta_i y_i \geq cd\right) &= \mathbf{prob}\left(t\sum_{i=1}^{r}\theta_i y_i \geq cdt\right)\\ &= \mathbf{prob}\left(\exp\left(t\sum_{i=1}^{r}\theta_i y_i\right) \geq \exp(cdt)\right)\\ &= \mathbf{prob}\left(\prod_{i=1}^{r}\exp(t\theta_i y_i) \geq \exp(cdt)\right).\end{aligned}$$

Applying Markov's inequality gives the bound

$$\mathbf{prob}\left(\sum_{i=1}^{r}\theta_i y_i \geq cd\right) \leq \exp(-cdt)\,\mathbf{E}\left(\prod_{i=1}^{r}\exp(t\theta_i y_i)\right).$$

Because $y_1, \ldots, y_r$ are independent, we can take the product outside the expectation:

$$\mathbf{prob}\left(\sum_{i=1}^{r}\theta_i y_i \geq cd\right) \leq \exp(-cdt)\prod_{i=1}^{r}\mathbf{E}(\exp(t\theta_i y_i)).$$

The expectation on the right side of this inequality is the moment-generating function of y_i evaluated at $t\theta_i$. Since y_i is a chi-squared random variable with d degrees of freedom, this moment-generating function is only defined for $t\theta_i < 1/2$; we will assume that $t < 1/2$ to ensure that this condition is satisfied. Then, using the formula for the moment-generating function of a chi-squared random variable with d degrees of freedom, our bound becomes

$$\mathbf{prob}\left(\sum_{i=1}^{r}\theta_i y_i \geq cd\right) \leq \exp(-cdt)\prod_{i=1}^{r}(1-2t\theta_i)^{-\frac{d}{2}}.$$

The Hessian of the right side of this inequality is

$$\exp(-cdt)\left(\prod_{i=1}^{r}(1-2t\theta_i)^{-\frac{d}{2}}\right)\left(z(\theta)z(\theta)^{\mathsf{T}} + \frac{2}{d}\,\mathbf{diag}(z(\theta))^2\right),$$

where we define the vector

$$z(\theta) = \left(\frac{dt}{1-2t\theta_1}, \ldots, \frac{dt}{1-2t\theta_r}\right).$$

This Hessian is positive semidefinite because it is a nonnegative multiple of the sum of a dyad and a diagonal matrix nonnegative entries. Thus, the right side of our last bound is a convex function of θ. We are interested in making a statement about all $\theta \in \mathbf{conv}(e_1, \dots, e_r)$, where e_j is the jth standard basis vector. Since the maximum of a convex function over a polyhedron is achieved at a vertex, setting $\theta = e_j$ gives a bound that is independent of θ, and satisfied by all values of interest:

$$\begin{aligned}\mathbf{prob}\left(\sum_{i=1}^{r} \theta_i y_i \geq cd\right) &\leq \exp(-cdt)(1-2t)^{-\frac{d}{2}} \\ &= \exp\left(-cdt - \frac{d}{2}\log(1-2t)\right).\end{aligned}$$

(The result is the same for all j.) The first and second derivatives of the right side of this inequality are

$$\frac{2cd}{1-2t}\left(t - \frac{c-1}{2c}\right)\exp\left(-cdt - \frac{d}{2}\log(1-2t)\right)$$

and

$$\left(\left(\frac{d}{1-2t} - cd\right)^2 + \frac{2d}{(1-2t)^2}\right)\exp\left(-cdt - \frac{d}{2}\log(1-2t)\right),$$

respectively. Since the second derivative is positive for all $t \in (0, 1/2)$, the right side of our bound is a strictly convex function of t, which we can minimize in order to obtain the tightest possible bound. By setting the first derivative equal to zero, we find that the minimum occurs at $t = (c-1)/(2c)$. Note that this value of t satisfies our earlier assumption that $0 < t < 1/2$. Substituting this value of t into our bound gives

$$\begin{aligned}\mathbf{prob}\left(\sum_{i=1}^{r} \theta_i y_i \geq cd\right) &\leq \exp\left(-\frac{d}{2}(c - \log(c) - 1)\right). \\ &= \exp(\log(c) - c + 1)^{\frac{d}{2}} \\ &= (ec\exp(-c))^{\frac{d}{2}}.\end{aligned}$$

□

Lemma C.2. Let $y_1, \dots, y_r$ be independent, identically distributed chi-squared random variables with d degrees of freedom, and $\theta_1, \dots, \theta_r$ be nonnegative real numbers such that $\theta_1 + \cdots + \theta_r = 1$. Then, for all $c \in (0, 1)$, we have that

$$\mathbf{prob}\left(\sum_{i=1}^{r} \theta_i y_i \leq cd\right) \leq (ec\exp(-c))^{\frac{d}{2}}.$$

Proof. For all $t > 0$, we have that

$$\begin{aligned}\mathbf{prob}\left(\sum_{i=1}^{r} \theta_i y_i \leq cd\right) &= \mathbf{prob}\left(-t\sum_{i=1}^{r} \theta_i y_i \geq -cdt\right)\\ &= \mathbf{prob}\left(\exp\left(-t\sum_{i=1}^{r} \theta_i y_i\right) \geq \exp(-cdt)\right)\\ &= \mathbf{prob}\left(\prod_{i=1}^{r} \exp(-t\theta_i y_i) \geq \exp(-cdt)\right).\end{aligned}$$

Applying Markov's inequality gives the bound

$$\mathbf{prob}\left(\sum_{i=1}^{r} \theta_i y_i \leq cd\right) \leq \exp(cdt)\,\mathbf{E}\left(\prod_{i=1}^{r} \exp(-t\theta_i y_i)\right).$$

Because $y_1, \ldots, y_r$ are independent, we can take the product outside the expectation:

$$\mathbf{prob}\left(\sum_{i=1}^{r} \theta_i y_i \leq cd\right) \leq \exp(cdt)\prod_{i=1}^{r} \mathbf{E}(\exp(-t\theta_i y_i)).$$

The expectation on the right side of this inequality is the moment-generating function of y_i evaluated at $-t\theta_i$. Since y_i is a chi-squared random variable with d degrees of freedom, this moment-generating function is only defined for $-t\theta_i < 1/2$; this condition is satisfied for all values of t and θ_i satisfying our existing assumptions. Then, using the formula for the moment-generating function of a chi-squared random variable with d degrees of freedom, our bound becomes

$$\mathbf{prob}\left(\sum_{i=1}^{r} \theta_i y_i \leq cd\right) \leq \exp(cdt)\prod_{i=1}^{r}(1 + 2t\theta_i)^{-\frac{d}{2}}.$$

The right side of this inequality is a convex function of $\theta \in \mathbf{R}^r$ (the justification of this fact is similar to that of the corresponding observation in the proof of Lemma C.1). We are interested in making a statement about all $\theta \in \mathbf{conv}(e_1, \dots, e_r)$, where e_j is the jth standard basis vector. Since the maximum of a convex function over a polyhedron is achieved at a vertex, setting $\theta = e_j$ gives a bound that is independent of θ, and satisfied by all values of interest:

$$\begin{aligned}\mathbf{prob}\left(\sum_{i=1}^{r} \theta_i y_i \leq cd\right) &\leq \exp(cdt)(1+2t)^{-\frac{d}{2}} \\ &= \exp\left(cdt - \frac{d}{2}\log(1+2t)\right).\end{aligned}$$

(The result is the same for all j.) The right side of this inequality is a convex function of t, which we can minimize in order to obtain the tightest possible bound. By setting the derivative equal to zero, we find that the minimum occurs at $t = (1-c)/(2c)$. (The calculations needed to prove that the right side of our bound is convex, and find its minimizer are similar to the corresponding calculations in the proof of Lemma C.1.) Note that this value of t satisfies our earlier assumption that $t > 0$. Substituting this value of t into our bound gives

$$\begin{aligned}\mathbf{prob}\left(\sum_{i=1}^{r} \theta_i y_i \leq cd\right) &\leq \exp\left(-\frac{d}{2}(c - \log(c) - 1)\right) \\ &= \exp(\log(c) - c + 1)^{\frac{d}{2}} \\ &= (ec\exp(-c))^{\frac{d}{2}}.\end{aligned}$$

□

C.2 The bivariate normal distribution

Our analysis of randomized rounding methods also uses a classical result due to Sheppard [84]. We provide a modern derivation of this result as well as a geometric interpretation.

Lemma C.3. Suppose $x = (x_1, x_2)$ is a bivariate normal random vector with mean vector and covariance matrix

$$\begin{bmatrix}0\\0\end{bmatrix} \quad \text{and} \quad \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix},$$

respectively. Then the quadrant probabilities of x are

$$\begin{aligned}\textbf{prob}(x_1 \geq 0, x_2 \geq 0) = \textbf{prob}(x_1 \leq 0, x_2 \leq 0) &= \frac{1}{4} + \frac{1}{2\pi}\arcsin(\rho),\\ \textbf{prob}(x_1 \leq 0, x_2 \geq 0) = \textbf{prob}(x_1 \geq 0, x_2 \leq 0) &= \frac{1}{4} - \frac{1}{2\pi}\arcsin(\rho).\end{aligned}$$

Proof. We will derive the first quadrant probability; the other quadrant probabilities follow from symmetry. We can compute the first quadrant probability by integrating the joint probability density function over the first quadrant:

$$\begin{aligned}&\textbf{prob}(x_1 \geq 0, x_2 \geq 0)\\ &\quad = \int_0^\infty \int_0^\infty \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{2(1-\rho^2)}\right) dx_1\, dx_2.\end{aligned}$$

Completing the square in x_1 gives

$$\begin{aligned}&\textbf{prob}(x_1 \geq 0, x_2 \geq 0)\\ &\quad = \int_0^\infty \left(\int_0^\infty \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{(x_1 - \rho x_2)^2}{2(1-\rho^2)}\right) dx_1\right) \exp\left(-\frac{x_2^2}{2}\right) dx_2.\end{aligned}$$

Now we perform a change of variables in the inner integral with

$$\tilde{x}_1 = \frac{x_1 - \rho x_2}{\sqrt{1-\rho^2}}.$$

This change of variables gives

$$\begin{aligned}&\textbf{prob}(x_1 \geq 0, x_2 \geq 0)\\ &\quad = \int_0^\infty \int_{-\rho x_2/\sqrt{1-\rho^2}}^\infty \frac{1}{2\pi} \exp\left(-\frac{\tilde{x}_1^2 + x_2^2}{2}\right) d\tilde{x}_1\, dx_2.\end{aligned}$$

We are now integrating the probability density function of a standard bivariate normal distribution; the region of integration is shown in Figure C.1. Observe that the angle θ defined in the figure is given by $\theta = \arcsin(\rho)$. Because the standard bivariate normal distribution has circular contours, the integral of its probability density function over the shaded region in Figure C.1 is simply the angular width of the region in radians, normalized by 2π: that is,

$$\textbf{prob}(x_1 \geq 0, x_2 \geq 0) = \frac{(\pi/2) + \theta}{2\pi} = \frac{1}{4} + \frac{1}{2\pi}\arcsin(\rho).$$

□

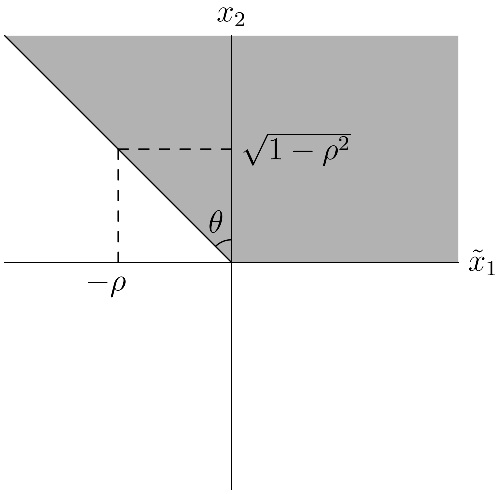


Figure C.1: the region of integration in the $\tilde{x}_1$-x_2 plane

Corollary C.4. Suppose $x = (x_1, x_2)$ is a bivariate normal random vector with mean vector and covariance matrix

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

respectively. Let $\hat{x}_1 = \mathbf{sgn}(x_1)$ and $\hat{x}_2 = \mathbf{sgn}(x_2)$. Then

$$\mathbf{E}(\hat{x}_1\hat{x}_2) = \frac{2}{\pi}\arcsin(\rho).$$

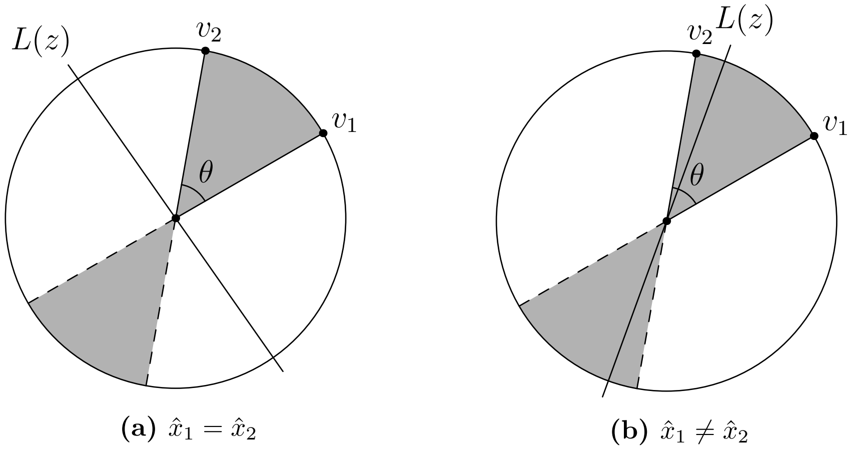
Proof. We have that

$$\begin{aligned} \mathbf{E}(\hat{x}_1\hat{x}_2) &= \mathbf{prob}(\mathbf{sgn}(x_1) = \mathbf{sgn}(x_2)) - \mathbf{prob}(\mathbf{sgn}(x_1) \neq \mathbf{sgn}(x_2)) \\ &= (\mathbf{prob}(x_1 \geq 0, x_2 \geq 0) + \mathbf{prob}(x_1 \leq 0, x_2 \leq 0)) \\ &\quad - (\mathbf{prob}(x_1 \leq 0, x_2 \geq 0) + \mathbf{prob}(x_1 \geq 0, x_2 \leq 0)). \end{aligned}$$

Plugging in the quadrant probabilities from Lemma C.3, we find that

$$\begin{aligned} \mathbf{E}(\hat{x}_1\hat{x}_2) &= 2\left(\frac{1}{4} + \frac{1}{2\pi}\arcsin(\rho)\right) - 2\left(\frac{1}{4} - \frac{1}{2\pi}\arcsin(\rho)\right) \\ &= \frac{2}{\pi}\arcsin(\rho). \end{aligned}$$

□



(a) $\hat{x}_1 = \hat{x}_2$

(b) $\hat{x}_1 \neq \hat{x}_2$

Figure C.2: two possible outcomes for $L(z)$

There is an appealing geometric interpretation of Corollary C.4. Suppose v_1 and v_2 are given unit vectors in $\mathbf{R}^n$. Let $z \in \mathbf{R}^n$ be a standard normal random vector, and define $x_i = v_i^{\mathsf{T}} z$ for $i = 1, 2$. Then

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} v_1^{\mathsf{T}} z \\ v_2^{\mathsf{T}} z \end{bmatrix} = \begin{bmatrix} v_1^{\mathsf{T}} \\ v_2^{\mathsf{T}} \end{bmatrix} z$$

is a bivariate normal random vector with mean vector and covariance matrix

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} v_1^{\mathsf{T}} \\ v_2^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} v_1^{\mathsf{T}} \\ v_2^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} 1 & v_1^{\mathsf{T}} v_2 \\ v_1^{\mathsf{T}} v_2 & 1 \end{bmatrix},$$

respectively. Let $\rho = v_1^{\mathsf{T}} v_2$ be the correlation of x_1 and x_2. Observe that ρ satisfies

$$\rho = v_1^{\mathsf{T}} v_2 = \cos(\theta) = \sin\left(\frac{\pi}{2} - \theta\right).$$

and hence $\arcsin(\rho) = (\pi/2) - \theta$. Let $L(z)$ be the projection of the hyperplane with normal vector z onto the v_1-v_2 plane. We have that $\hat{x}_1 \neq \hat{x}_2$ if and only if v_1 and v_2 are on opposite sides of $L(z)$, as shown in Figure C.2. By symmetry, all lines $L(z)$ are equally likely, so the probability that $\hat{x}_1 \neq \hat{x}_2$ is the area of the wedge between v_1 and v_2 plus the area of the wedge between $-v_1$ and $-v_2$, normalized by 2π

(these wedges are shaded in Figure C.2):

$$\mathbf{prob}(\hat{x}_1 \neq \hat{x}_2) = \frac{\theta + \theta}{2\pi} = \frac{\theta}{\pi}.$$

Then we have that

$$\begin{aligned}
\mathbf{E}(\hat{x}_1\hat{x}_2) &= \mathbf{prob}(\hat{x}_1 = \hat{x}_2) - \mathbf{prob}(\hat{x}_1 \neq \hat{x}_2) \\
&= 1 - 2\,\mathbf{prob}(\hat{x}_1 \neq \hat{x}_2) \\
&= 1 - 2\frac{\theta}{\pi} \\
&= \frac{2}{\pi}\left(\frac{\pi}{2} - \theta\right) \\
&= \frac{2}{\pi}\arcsin(\rho).
\end{aligned}$$

References

[1] W. Ai, Y. Huang, and S. Zhang. On the low rank solutions for linear matrix inequalities. *Mathematics of Operations Research*, 33(4):965 – 975, 2008.

[2] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253 – 263, 2008.

[3] A. Barvinok. Problems of distance geometry and convex properties of quadratic maps. *Discrete & Computational Geometry*, 13(2):189 – 202, 1995.

[4] A. Barvinok. A remark on the rank of positive semidefinite matrices subject to affine constraints. *Discrete & Computational Geometry*, 25(1):23 – 31, 2001.

[5] A. Beck and Y. Eldar. Strong duality in nonconvex quadratic optimization with two quadratic constraints. *SIAM Journal on Optimization*, 17(3):844 – 860, 2006.

[6] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization.* Princeton Series in Applied Mathematics. Princeton University Press, 2009.

[7] D. Bertsekas. *Nonlinear Programming.* Athena Scientific, 2nd edition, 2004.

[8] D. Bienstock and A. Michalka. Polynomial solvability of variants of the trust-region subproblem. *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, 25:380 – 390, 2014.

[9] S. Bose, D. Gayme, K. Chandy, and S. Low. Quadratically constrained quadratic programs on acyclic graphs with application to power flow. *IEEE Transactions on Control of Network Systems*, 2(3):278 – 287, 2015.

[10] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[11] C. Broyden. The convergence of a class of double-rank minimization algorithms 1. General considerations. *IMA Journal of Applied Mathematics*, 6(1):76 – 90, 1970.

[12] C. Broyden. The convergence of a class of double-rank minimization algorithms 2. The new algorithm. *IMA Journal of Applied Mathematics*, 6(3):222 – 231, 1970.

[13] S. Burer and K. Anstreicher. Second-order-cone constraints for extended trust-region subproblems. *SIAM Journal on Optimization*, 23(1):432 – 451, 2013.

[14] S. Burer and R. Monteiro. A projected gradient algorithm for solving the maxcut SDP relaxation. *Optimization Methods and Software*, 15(3 – 4):175 – 200, 2001.

[15] S. Burer and R. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329 – 357, 2003.

[16] S. Burer and R. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427 – 444, 2005.

[17] S. Burer, R. Monteiro, and Y. Zhang. Interior-point algorithms for semidefinite programming based on a nonlinear formulation. *Computational Optimization and Applications*, 22(1):49 – 79, 2002.

[18] S. Burer, R. Monteiro, and Y. Zhang. A computational study of a gradient-based log-barrier algorithm for a class of large-scale SDPs. *Mathematical Programming*, 95(2):359 – 379, 2003.

[19] S. Burer and B. Yang. The trust region subproblem with non-intersecting linear constraints. *Mathematical Programming*, 149(1):253 – 264, 2015.

[20] E. Candès. Compressive sampling. In *Proceedings of the International Congress of Mathematicians*, pages 1433 – 1452. European Mathematical Society, 2006.

[21] E. Candès and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969 – 985, 2007.

[22] E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207 – 1223, 2006.

[23] E. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203 – 4215, 2005.

[24] C. Carathéodory. Über den variabilitätsbereich der koeffizienten von potenzreihen, die gegebene werte nicht annehmen. *Mathematische Annalen*, 64(1):95 – 115, 1097.

[25] T.-H. Chang, W.-K. Ma, and C.-Y. Chi. Worst-case robust multiuser transmit beamforming using semidefinite relaxation: Duality and implications. In *Conference Record of the Asilomar Conference on Signals, Systems and Computers*, volume 45, pages 1579 – 1583. IEEE, 2011.

[26] R. Corless, G. Gonnet, D. Hare, D. Jeffrey, and D. Knuth. On the Lambert W function. *Advances in Computational Mathematics*, 5(1):329 – 359, 1996.

[27] G. Cornuéjols. Revival of the Gomory cuts in the 1990s. *Annals of Operations Research*, 149(1):63 – 66, 2007.

[28] G. Cornuéjols. Valid inequalities for mixed integer linear programs. *Mathematical Programming*, 112(1):3 – 44, 2008.

[29] J. Dennis and R. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, volume 16 of *Classics in Applied Mathematics*. SIAM, 1996.

[30] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289 – 1306, 2006.

[31] M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford University, 2002.

[32] M. Fazel, H. Hindi, and S. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the American Control Conference*, volume 6, pages 4734 – 4739. IEEE, 2001.

[33] M. Fazel, H. Hindi, and S. Boyd. Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices. In *Proceedings of the American Control Conference*, volume 3, pages 2156 – 2162. IEEE, 2003.

[34] R. Fletcher. A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317 – 322, 1970.

[35] R. Fletcher. *Practical Methods of Optimization.* Wiley, 2nd edition, 2000.

[36] M. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W. H. Freeman, 1979.

[37] D. Gay. Computing optimal locally constrained steps. *SIAM Journal on Scientific and Statistical Computing*, 2(2):186 – 197, 1981.

[38] A. Gershman, N. Sidiropoulos, S. Shahbazpanahi, M. Bengtsson, and B. Ottersten. Convex optimization based beamforming: From receive to transmit and network designs. *IEEE Signal Processing Magazine*, 27(3):62 – 75, 2010.

[39] M. Goemans and D. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42(6):1115 – 1145, 1995.

[40] D. Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23 – 26, 1970.

[41] S. Goldfeld, R. Quandt, and H. Trotter. Maximization by quadratic hill-climbing. *Econometrica*, 34(3):541 – 551, 1966.

[42] R. Gomory. An algorithm for the mixed integer problem. Technical report, The RAND Corporation, 1960.

[43] R. Gomory. An algorithm for integer solutions to linear programs. In R. Graves and P. Wolfe, editors, *Recent Advances in Mathematical Programming*, pages 269 – 302. McGraw-Hill, 1963.

[44] N. Gould, S. Lucidi, M. Roma, and P. Toint. Solving the trust-region subproblem using the Lanczos method. *SIAM Journal on Optimization*, 9(2):504 – 525, 1999.

[45] O. Güler and Y. Ye. Convergence behavior of interior-point algorithms. *Mathematical Programming*, 60(1):215 – 228, 1993.

[46] S. Gusev and A. Likhtarnikov. Kalman-Popov-Yakubovich lemma and the S-procedure: A historical essay. *Automation and Remote Control*, 67(11):1768 – 1810, 2006.

[47] J. Håstad. Some optimal inapproximability results. *Journal of the ACM*, 48(4):798 – 859, 2001.

[48] C. Helmberg and F. Rendl. A spectral bundle method for semidefinite programming. *SIAM Journal on Optimization*, 10(3):673 – 696, 2000.

[49] M. Hestenes. Multiplier and gradient methods. *Journal of Optimization Theory and Applications*, 4(5):303 – 320, 1969.

[50] Y. Huang and D. Palomar. Rank-constrained separable semidefinite programming with applications to optimal beamforming. *IEEE Transactions on Signal Processing*, 58(2):664 – 678, 2009.

[51] Y. Huang and S. Zhang. Complex matrix decompositions and quadratic programming. *Mathematics of Operations Research*, 32(3):758 – 768, 2007.

[52] V. Jeyakumar, G. Lee, and G. Li. Alternative theorems for quadratic inequality systems and global quadratic optimization. *SIAM Journal on Optimization*, 20(2):983 – 1001, 2009.

[53] M. Journée, F. Bach, P. Absil, and R. Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, 20(5):2327 – 2351, 2010.

[54] A. Kalbat, R. Madani, G. Fazelnia, and J. Lavaei. Efficient convex relaxation for stochastic optimal distributed control problem. *Allerton Conference on Communication, Control, and Computing*, 52:589 – 596, 2014.

[55] N. Karmarkar, M. Resende, and K. Ramakrishnan. An interior point algorithm to solve computationally difficult set covering problems. *Mathematical Programming*, 52(1):597 – 618, 1991.

[56] R. Karp. *Complexity of Computer Computations*, chapter Reducibility among Combinatorial Problems, pages 85 – 103. Springer, 1972.

[57] S. Khot. On the power of unique 2-prover 1-round games. *Proceedings of the ACM Symposium on Theory of Computing*, 34:767 – 775, 2002.

[58] S. Khot, G. Kindler, E. Mossel, and R. O'Donnell. Optimal inapproximability results for MAX-CUT and other 2-variable CSPs? *SIAM Journal on Computing*, 37(1):319 – 357, 2007.

[59] S. Kim and M. Kojima. Exact solutions of some nonconvex quadratic optimization problems via SDP and SOCP relaxations. *Computational Optimization and Applications*, 26(2):143 – 154, 2003.

[60] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000.

[61] J. Lavaei. Zero duality gap for classical OPF problem convexifies fundamental nonlinear power problems. *Proceedings of the American Control Conference*, pages 4566 – 4573, 2011.

[62] J. Lavaei and S. Low. Zero duality gap in optimal power flow problem. *IEEE Transactions on Power Systems*, 27(1):92 – 107, 2012.

[63] J. Lavaei, D. Tse, and B. Zhang. Geometry of power flows in tree networks. *IEEE Power and Energy General Meeting*, pages 1 – 8, 2012.

[64] Q. Li and W. Ma. Optimal and robust transmit designs for MISO channel secrecy by semidefinite programming. *IEEE Transactions on Signal Processing*, 59(8):3799 – 3812, 2011.

[65] D. Luenberger and Y. Ye. *Linear and Nonlinear Programming.* Springer, 3rd edition, 2008.

[66] J. Martínez. Local minimizers of quadratic functions on Euclidean balls and spheres. *SIAM Journal on Optimization*, 4(1):159 – 176, 1994.

[67] J. Martínez and S. Santos. A trust-region strategy for minimization on arbitrary domains. *Mathematical Programming*, 68(1):267 – 301, 1995.

[68] M. Medra, W. Ma, and T. Davidson. Low-complexity robust MISO downlink precoder optimization for the limited feedback case. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3108 – 3112, 2015.

[69] M. Mesbahi. On the semi-definite programming solution of the least order dynamic output feedback synthesis. *Proceedings of the IEEE Conference on Decision and Control*, 38(2):1851 – 1856, 1999.

[70] C. Meyer. *Matrix Analysis and Applied Linear Algebra.* SIAM, 2001.

[71] R. Monteiro. First- and second-order methods for semidefinite programming. *Mathematical Programming*, 97(1):209 – 244, 2003.

[72] J. Moré. *The Levenberg-Marquardt Algorithm: Implementation and Theory*, pages 105 – 116. Springer, 1978.

[73] J. Moré and D. Sorensen. Computing a trust region step. *SIAM Journal on Scientific and Statistical Computing*, 4(3):553 – 572, 1983.

[74] Y. Nesterov. Quality of semidefinite relaxation for nonconvex quadratic optimization. Technical Report 9719, Center for Operations Research & Econometrics, Universite Catholique de Louvain, 1997.

[75] T. Pare. *Analysis and Control of Nonlinear Systems.* PhD thesis, Stanford University, 2000.

[76] G. Pataki. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of Operations Research*, 23(2):339 – 358, 1998.

[77] J. Peng and Y. Yuan. Optimality conditions for the minimization of a quadratic with two quadratic constraints. *SIAM Journal on Optimization*, 7(3):579 – 594, 1997.

[78] I. Pólik and T. Terlaky. A survey of the S-lemma. *SIAM Review*, 49(3):371 – 418, 2007.

[79] M. Powell. *A Method for Non-Linear Constraints in Minimization Problems.* UKAEA, 1967.

[80] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471 – 501, 2010.

[81] F. Rendl and H. Wolkowicz. A semidefinite framework for trust region subproblems with applications to large scale minimization. *Mathematical Programming*, 77(1):273 – 299, 1997.

[82] D. Shanno. Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24(111):647 – 656, 1970.

[83] M. Shenouda and T. Davidson. Convex conic formulations of robust downlink precoder designs with quality of service constraints. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):714 – 724, 2007.

[84] W. Sheppard. On the calculation of the double integral expressing normal correlation. *Transactions of the Cambridge Philosophical Society*, 19:23 – 66, 1900.

[85] A. So, Y. Ye, and J. Zhang. A unified theorem on SDP rank reduction. *Mathematics of Operations Research*, 33(4):910 – 920, 2008.

[86] S. Sojoudi and J. Lavaei. Physics of power networks makes hard optimization problems easy to solve. *Proceedings of the IEEE Power and Energy Society General Meeting*, pages 1 – 8, 2012.

[87] S. Sojoudi and J. Lavaei. On the exactness of semidefinite relaxation for nonlinear optimization over graphs: Part I. *Proceedings of the IEEE Conference on Decision and Control*, 52:1043 – 1050, 2013.

[88] S. Sojoudi and S. Low. Optimal charging of plug-in hybrid electric vehicles in smart grids. *Proceedings of the IEEE Power and Energy Society General Meeting*, pages 1 – 6, 2011.

[89] E. Song, Q. Shi, M. Sanjabi, R. Sun, and Z. Luo. Robust SINR-constrained MISO downlink beamforming: When is semidefinite programming relaxation tight? *EURASIP Journal on Wireless Communications and Networking*, pages 1 – 11, 2012.

[90] D. Sorensen. Newton's method with a model trust region modification. *SIAM Journal on Numerical Analysis*, 19(2):409 – 426, 1982.

[91] R. Stern and H. Wolkowicz. Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations. *SIAM Journal on Optimization*, 5(2):286 – 313, 1995.

[92] G. Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, 4th edition, 2009.

[93] J. Sturm and S. Zhang. On cones of nonnegative quadratic functions. *Mathematics of Operations Research*, 28(2):246 – 267, 2003.

[94] L. Trevisan, G. Sorkin, M. Sudan, and D. Williamson. Gadgets, approximation, and linear programming. *SIAM Journal on Computing*, 29(6):2074 – 2097, 2000.

[95] H. Tuy and H. Tuan. Generalized S-lemma and strong duality in nonconvex quadratic programming. *Journal of Global Optimization*, 56(3):1045 – 1072, 2013.

[96] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38(1):49 – 95, 1996.

[97] N. Vučić and H. Boche. Robust QoS-constrained optimization of downlink multiuser MISO systems. *IEEE Transactions on Signal Processing*, 57(2):714 – 725, 2009.

[98] Y. Wang and R. Shi. Tightness of semidefinite programming relaxation to robust transmit beamforming with SINR constraints. *Mathematical Problems in Engineering*, 2013:1 – 10, 2013.

[99] K. Williamson. *A Robust Trust Region Algorithm for Nonlinear Programming*. PhD thesis, Rice University, 1990.

[100] Y. Xia, S. Wang, and R. Sheu. S-lemma with equality and its applications. *Mathematical Programming*, 156(1):513 – 547, 2016.

[101] Y. Ye. On affine scaling algorithms for nonconvex quadratic programming. *Mathematical Programming*, 56(1):285 – 300, 1992.

[102] Y. Ye and S. Zhang. New results on quadratic minimization. *SIAM Journal on Optimization*, 14(1):245 – 267, 2003.

[103] B. Zhang and D. Tse. Geometry of injection regions of power networks. *IEEE Transactions on Power Systems*, 28(2):788 – 797, 2013.

[104] Y. Zhang. Computing a Celis-Dennis-Tapia trust-region step for equality constrained optimization. *Mathematical Programming*, 55(1):109 – 124, 1992.

[105] Z. Zhu. *A Semidefinite Programming Method for Graph Realization and the Low Rank Matrix Completion Problem*. PhD thesis, Stanford University, 2011.