

# On Polyhedral and Second-Order-Cone Decompositions of Semidefinite Optimization Problems

Dimitris Bertsimas<sup>a,\*</sup>, Ryan Cory-Wright<sup>b</sup>

<sup>a</sup>*Sloan School of Management and Operations Research Center, Massachusetts Institute of Technology Cambridge, MA, USA*

<sup>b</sup>*Operations Research Center, Massachusetts Institute of Technology Cambridge, MA, USA*

---

## Abstract

We study a cutting-plane method for semidefinite optimization problems (SDOs), and supply a proof of the method's convergence, under a boundedness assumption. By relating the method's rate of convergence to an initial outer approximation's diameter, we argue that the method performs well when initialized with a second-order-cone approximation, instead of a linear approximation. We invoke the method to provide bound gaps of 0.5-6.5% for sparse PCA problems with 1000s of covariates, and solve nuclear norm problems over  $500 \times 500$  matrices.

**Keywords:** Semidefinite optimization, Cutting planes, Nuclear norm, Sparse PCA.

---

## 1. Introduction

We study decomposition schemes for semidefinite optimization problems (SDOs) of the form:

$$\begin{aligned} \min_{X \in \mathbb{R}^{n \times n}} \quad & \langle C, X \rangle \\ \text{s.t.} \quad & \langle A_i, X \rangle = b_i, \quad \forall i \in [m], \quad X \succeq 0. \end{aligned} \quad (1)$$

Problem (1) is tractable from a traditional complexity theory perspective, because it is usually<sup>1</sup> solvable in polynomial time via interior point methods (IPMs). However, it is notoriously difficult to solve in practice, because IPMs memory requirements scale at a demanding rate. Indeed, state-of-the-art SDO solvers such as MOSEK cannot solve constrained instances of Problem (1) with  $n > 250$  variables on a standard laptop, and it is optimization folklore that there is a gap between SDOs theoretical and practical tractability.

Motivated by the demanding memory requirements of IPMs, a stream of literature studies *inexact* methods for SDOs, which replace the semidefinite constraint with weaker yet less computationally demanding constraints. This approach was first investigated by Kim and Kojima [13], who observed that relaxing a positive semidefinite constraint to the weaker constraint that all  $2 \times 2$  minors of a matrix are positive semidefinite yields a second order cone (SOC)-representable outer approximation of the positive semidefinite (PSD) cone.

In a related line of work, Krishnan and Mitchell [15] propose applying Kelley [12]'s cutting plane method to generate

an improving sequence of outer approximations of the positive semidefinite cone. Their approach converges whenever the feasible region is bounded, albeit exponentially slowly in the worst-case<sup>2</sup>. Unfortunately, in [15]'s implementation, Kelley's method performs poorly in practice and often provides weaker bounds than those obtained by enforcing SOC-representable  $2 \times 2$  minor constraints, as found empirically in [21, 1].

From a traditional complexity theory perspective, the slow rate of convergence of Kelley [12]'s method is unsurprising. Indeed, SDOs cannot be approximated arbitrarily well using polynomially-sized linear optimization problems (LOs) [6, Proposition 3], and therefore any LO-based cutting-plane method for SDOs converges exponentially slowly in the worst-case. However, there are numerous examples of cutting-plane methods with slow worst-case rates of convergence which nonetheless successfully solve large-scale problems in practice [see 7, 10, 5, and references therein]. Moreover, optimizers widely believe that there is a gap between the observed and worst-case performance of cutting-plane methods [see 16, Section 5.2]. Therefore, a slow worst-case rate of convergence does not preclude Kelley's method from performing well in practice.

### 1.1. Our Contributions

In this paper, we provide a different perspective on Kelley's method than the traditional complexity theory perspective. By bounding the number of iterations required to obtain an  $\epsilon$ -feasible solution to Problem (1), we establish that the method's worst-case performance depends explicitly on the diameter of an initial outer approximation of the problem, and argue that this dependence also appears in practice.

By examining this worst-case bound, we diagnose why existing implementations of Kelley [12]'s method do not scale well in practice. Namely, they are initialized with LO-representable

---

\*Corresponding author

Email addresses: dbertsim@mit.edu (Dimitris Bertsimas), ryancw@mit.edu (Ryan Cory-Wright)

<sup>1</sup>In general, Problem (1) is not solvable in polynomial time, because it may admit solutions of exponential size [see 2, Section 3.3]. Nonetheless, by bounding the allowable size of solutions to Problem (1) apriori, it can be solved in polynomial time [2]. As most practically relevant SDOs can be bounded in this fashion, we refer to SDOs as "polynomial time solvable"; see Ramana [18] for a complete characterization of the complexity of SDOs.

<sup>2</sup>To our knowledge, no proof of convergence has been supplied for SDOs. Therefore, we supply a proof in Theorem 1.

outer approximations of the PSD cone, which do not capture its inherent nonlinearity. Indeed, as we establish in Section 2, the diameter of an LO-representable approximation grows linearly with  $n \cdot \text{tr}(X)$ , and therefore the rate of convergence of Kelley’s method (under this initial approximation) depends upon  $n \cdot \text{tr}(X)$ . After observing that Kelley’s method performs poorly in both theory and practice with an LO-representable initial approximation, we propose initializing Kelley’s method with a SOC-representable approximation. We establish that (a) the diameter of this approximation grows at the slower rate of  $\text{tr}(X)$ , and (b) initializing a cutting-plane method with a SOC approximation performs better in both theory and practice.

Finally, we apply Kelley’s method to two problems from the machine learning literature: sparse principal component analysis (PCA) and nuclear norm minimization. For sparse PCA, we demonstrate that initializing Kelley’s method with an SOC-representable feasible region provides bounds which are comparable with the SDO-representable bound of d’Aspremont et al. [8], and successfully supplies near-exact bounds for problems with 1000s of covariates. For nuclear norm minimization, we reformulate the nuclear norm in terms of its dual norm, and invoke this reformulation to derive a new class of cutting-planes. Moreover, we demonstrate that combining Kelley’s method with this class of cuts scales better than IPMs for nuclear norm minimization.

### 1.2. Structure

The rest of the paper is laid out as follows:

- In Section 2, we study two outer-approximations of Problem (1): a linear approximation and a second-order-cone approximation. We also supply bounds on the distance between these approximations and the original problem.
- In Section 3, we present an outer-approximation method for solving Problem (1) to certifiable optimality, and supply a proof of its convergence. Furthermore, we establish that the worst-case rate of convergence depends explicitly on the diameter of the initial feasible region.
- In Section 4, we present numerical results demonstrating that the outer-approximation method is competitive with state-of-the-art methods for both sparse principal component analysis and nuclear norm minimization problems.

### 1.3. Notation

We let nonbold face characters such as  $b$  denote scalars, lowercase bold faced characters such as  $\mathbf{x}$  denote vectors, uppercase bold faced characters such as  $\mathbf{X}$  denote matrices, and calligraphic uppercase characters such as  $\mathcal{Z}$  denote sets. We assume that matrices  $\mathbf{X}$  are symmetric, unless stated otherwise. We let  $[n]$  denote the set of running indices  $\{1, \dots, n\}$ . We let  $\mathbf{e}$  denote a vector of all 1’s,  $\mathbf{0}$  denote a vector of all 0’s, and  $\mathbb{I}$  denote the identity matrix, with dimension implied by the context. We let  $S_+^n$  denote the  $n \times n$  positive semidefinite cone.

We also use an assortment of matrix operators. We let  $\|\cdot\|_*$  denote the spectral norm of a matrix,  $\langle \cdot, \cdot \rangle$  denote the Euclidean inner product between two matrices,  $\|\cdot\|_F$  denote the Frobenius norm of a matrix, and  $\|\cdot\|_\sigma$  denote the singular value norm of

a matrix; see Horn and Johnson [11] for a general theory of matrix operators.

Finally, we let  $D(\mathcal{X})$  denote the diameter of a set  $\mathcal{X}$ , i.e.,

$$D(\mathcal{X}) = \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} \|\mathbf{x} - \mathbf{y}\|,$$

where  $\|\cdot\|$  is the  $l_2$  (Frobenius) norm for vectors (matrices).

## 2. Two Outer Approximations of the PSD cone

In this section, we introduce two outer approximations of the positive semidefinite cone, bound the diameters of the approximations, and propose two classes of separating hyperplanes which iteratively refine these approximations.

### 2.1. A Linear Outer Approximation

The following problem is a valid relaxation of Problem (1), which is LO-representable [see 17, Table 1]:

$$\begin{aligned} \min_{\mathbf{X} \in S^n} \quad & \langle \mathbf{C}, \mathbf{X} \rangle \\ \text{s.t.} \quad & \langle \mathbf{A}_i, \mathbf{X} \rangle = b_i, & \forall i \in [m], \\ & X_{i,i} \geq 0, & \forall i \in [n], \\ & X_{i,i} + X_{j,j} \pm 2X_{i,j} \geq 0, & \forall i \in [n], \forall j \in [n]. \end{aligned} \quad (2)$$

In Section 3, we will iteratively improve the quality of outer approximations of SDOs, such as Problem (2), by adding cutting planes, and prove that any limit point of this procedure solves Problem (1). As Section 3’s proof of convergence relies on a compactness argument, which in turn rests on the boundedness of the initial feasible region, we now prove that imposing the constraint  $\text{tr}(\mathbf{X}) \leq T$  in Problem (2) ensures that all feasible solutions are contained within a convex, compact set.

**Proposition 1.** *Let there exist a  $T \geq 0$  such that all solutions to Problem (1) are bounded in trace norm by  $T$ , i.e.,  $\text{tr}(\mathbf{X}) \leq T$ . Then, the solutions to Problem (2) form a convex, compact set.*

*Proof.* The boundedness of the set follows because:

$$\sum_{i=1}^n \sum_{j=1}^n |X_{i,j}| \leq \sum_{i=1}^n \sum_{j=1}^n \frac{1}{2}(X_{i,i} + X_{j,j}) \leq n \cdot \text{tr}(\mathbf{X}) \leq nT.$$

Therefore, Problem (2)’s feasible region is a polytope (i.e., a bounded polyhedron) and the result holds.  $\square$

A corollary of Proposition 1 is that  $\|\mathbf{X}\|_F \leq nT$ , because  $\|\cdot\|_2 \leq \|\cdot\|_1$ . Consequently, the set of feasible solutions to Problem (2) has a diameter of  $D(\mathcal{X}) \leq 2nT$ .

Unfortunately, while Problem (2) is a very tractable outer approximation, various authors [21, 1] have observed that it poorly approximates Problem (1) in practice. Motivated by these observations, we now argue that Problem (2) should *not* be used as an initial outer-approximation for a cutting-plane method, because it is dominated by an equally tractable yet much tighter nonlinear approximation. To this end, we now bound the degree to which solutions to Problem (2) violate the constraint  $\mathbf{X} \geq \mathbf{0}$ , in terms of  $\text{tr}(\mathbf{X})$ , and demonstrate that (2) offers poor constraint violation guarantees, since the worst-case value of  $\lambda_{\min}(\mathbf{X})$  is more negative in Problem (2) than (3):

**Proposition 2.** Let  $X \in \mathbb{R}^{n \times n}$  be a feasible solution to Problem (2) such that  $\text{tr}(X) = T$  is held constant. Then,

$$\lambda_{\min}(X) \geq \frac{T}{n} - \frac{T \sqrt{(n^3 - 1)(n - 1)}}{n} \geq \frac{T}{n} - nT.$$

*Proof.* By Proposition 1, we have that  $\text{tr}(X^2) \leq n^2 T^2$ . Therefore, the result follows from the eigenvalue-trace bound [see 22, Theorem 2.1], which requires that for any matrix  $X \in S^n$

$$\lambda_{\min}(X) \geq \frac{\text{tr}(X)}{n} - \sqrt{(n - 1) \left( \frac{\text{tr}(X^2)}{n} - \frac{\text{tr}(X)^2}{n^2} \right)}.$$

Note that this inequality is tight in the absence of further problem structure [see 22, Theorem 2.1].  $\square$

Proposition 2 demonstrates that *even if* the optimal choice of  $X$  is bounded in trace norm by a constant, the quality of Problem (2)'s approximation degrades with  $n$ . This result explains why various authors [21, 14, 1] have found that Problem (2) poorly approximates Problem (1) in a high-dimensional setting.

We now present a SOC-representable outer approximation of Problem (1), and establish that it enjoys more attractive worst-case guarantees than a linear outer approximation.

## 2.2. A Second-Order Cone Outer Approximation

The following problem is a valid relaxation of Problem (1), which is SOC representable [see 13]:

$$\begin{aligned} \min_{X \in S^n} \quad & \langle C, X \rangle \\ \text{s.t.} \quad & \langle A_i, X \rangle = b_i, \quad \forall i \in [m], \\ & \left\| \begin{pmatrix} 2X_{i,j} \\ X_{i,i} - X_{j,j} \end{pmatrix} \right\|_2 \leq X_{i,i} + X_{j,j}, \quad \forall i \in [n], \forall j \in [n]. \end{aligned} \quad (3)$$

We now relate the diameter of Problem (3)'s feasible region to a trace constraint:

**Proposition 3.** Let there exists a  $T \geq 0$  such that all feasible solutions to Problem (1) are bounded in trace norm by  $T$ , i.e.,  $\text{tr}(X) \leq T, \forall X \in \mathcal{X}$ . Then, Problem (3)'s feasible region is bounded in Frobenius norm by  $T$ , i.e.,  $\|X\|_F \leq T, \forall X \in \mathcal{X}$ .

*Proof.* The  $2 \times 2$  minor constraints in Problem (3) are equivalent to requiring that  $X_{i,j}^2 \leq X_{i,i}X_{j,j}$ . Therefore, this result follows directly from the observation that:

$$\|X\|_F^2 = \sum_{i=1}^n \sum_{j=1}^n X_{i,j}^2 \leq \sum_{i=1}^n \sum_{j=1}^n X_{i,i}X_{j,j} = \text{tr}(X)^2 \leq T^2. \quad \square$$

We now bound the degree to which solutions to Problem (3) violate the constraint  $X \geq \mathbf{0}$ , and establish that Problem (3) enjoys better constraint violation guarantees than Problem (2):

**Proposition 4.** Let  $X \in S^n$  be a feasible solution to Problem (3) such that  $\text{tr}(X) = T$ . Then

$$\lambda_{\min}(X) \geq \frac{2T}{n} - T.$$

*Proof.* The result follows from observing that  $\text{tr}(X^2) \leq T^2$ , and invoking the eigenvalue-trace bound. The bound is tight in the absence of additional problem structure [22, Theorem 2.1].  $\square$

Observe that Proposition 4's bound holds upon relaxing  $\text{tr}(X) = T$  to  $\text{tr}(X) \leq T$ , as the later constraint is equivalent to writing  $\text{tr}(X) = t$  for some  $t \in [0, T]$ , and Proposition 4's bound is minimized over  $t \in [0, T]$  by setting  $t = T$ .

Two important instances of Proposition 4's bound arise in correlation matrices where  $\text{Diag}(X) = \mathbf{e}$  and spectraplex matrices where  $\text{tr}(X) = 1$ . In these cases:

$$\lambda_{\min}(X) \geq 2 - n, \text{ and } \lambda_{\min}(X) \geq \frac{2}{n} - 1,$$

respectively. Indeed, imposing the constraint  $\text{tr}(X) = 1$  in Problem (3) provides a dimension-independent violation bound, while imposing the same constraint in Problem (2) yields a bound of the order  $\frac{1}{n} - n$ .

The worst-case value of  $\lambda_{\min}(\cdot)$  is  $n$  times smaller in Problem (3) than Problem (2). Therefore, we have established that Problem (3) is a tighter approximation than Problem (2). As optimizers widely believe that LOs and SOCPs are equally tractable in the age of modern IPMs, we will use Problem (3) as an initial approximation when developing a cutting-plane method in Section 3, and not consider Problem (2) for the rest of the paper.

## 2.3. Two Classes of Separating Hyperplanes

We now present two classes of separating hyperplanes which allow us to improve the outer approximations presented in Problems (2)-(3). Our first class of separating hyperplanes rests on an almost tautological result, namely:

**Proposition 5.** The following two constraints are equivalent:

$$\begin{aligned} X &\geq \mathbf{0}, \\ \langle X, Y \rangle &\geq 0, \quad \forall Y \in \{Y \in S_+^n : \|Y\|_* \leq 1\}. \end{aligned}$$

*Proof.* The result follows from the self-duality of  $S_+^n$ .  $\square$

For a given  $\hat{X}$ , Proposition 5 yields an extreme ray  $Y(\hat{X}) \in S_+^n$  in  $O(n^2)$  time via the power method, see [11]; namely

$$Y(\hat{X}) = yy^\top, \quad \text{where } y = \arg \min_{x: \|x\|_2=1} x^\top \hat{X} x.$$

Moreover,  $Y$  either verifies membership of  $\hat{X} \in S_+^n$  if  $\langle \hat{X}, Y \rangle \geq 0$ , or provides a separating hyperplane such that  $\langle \hat{X}, Y \rangle < 0$  and  $\langle X, Y \rangle \geq 0$  for any feasible  $X \in S_+^n$ . We refer to this class of cuts as "trailing eigenvalue cuts".

The next class of separating hyperplanes is derived from a reformulation of the PSD cone which is, to our knowledge, new:

**Proposition 6.** The following two constraints are equivalent:

$$\begin{aligned} X &\geq \mathbf{0}, \\ \langle X, Y \rangle &\leq \text{tr}(X), \quad \forall Y \in \{Y \in S_+^n : \|Y\|_\sigma \leq 1\}. \end{aligned}$$

*Proof.* The first constraint is equivalent to requiring that

$$\|X\|_* \leq \text{tr}(X),$$

because  $X \geq \mathbf{0}$  if and only if  $\lambda_{\min}(X) \geq 0$ . Moreover, all eigenvalues  $\lambda_i(X)$  are non-negative if and only if the singular values  $\sigma_i(X)$  are such that  $\sum_i \sigma_i(X) \leq \sum_i \lambda_i(X)$ , since  $\sigma_i(X) = |\lambda_i(X)|$  for any symmetric matrix  $X$ ; see [11, 2.6.P15]. Thus, the result follows from the duality of the spectral and nuclear norms.  $\square$

The following proposition demonstrates that an extreme ray of Proposition 6's reformulation of the PSD cone is given via a singular value decomposition (SVD):

**Proposition 7.** *An optimal solution to the optimization problem*

$$\max_{Y \in S^n: \|Y\|_{\sigma} \leq 1} \langle X, Y \rangle$$

is given by setting  $Y^* := UU^\top$ , where  $X := U\Lambda U^\top$  is a singular value decomposition of  $X$ .

*Proof.* Let  $\lambda_i$  denote the eigenvalues of  $X$ , and  $\sigma_i = |\lambda_i|$  denote the singular values of  $X$ . Then, the result follows from observing that (1), by Hölder's inequality:

$$\langle X, Y \rangle \leq \|X\|_* \|Y\|_{\sigma}, \quad \forall X, Y \in S^n,$$

because the spectral norm is dual to the nuclear norm, and (2) Hölder's inequality is tight when  $Y = Y^*$ , because  $u_i \perp u_j$ ,  $\forall i \neq j$  and therefore  $\langle X, Y^* \rangle = \sum_{i=1}^n |\lambda_i| = \sum_{i=1}^n \sigma_i = \|X\|_*$ .  $\square$

Observe that  $Y^*$  either verifies membership of  $\hat{X} \in S_+^n$  if  $\langle \hat{X}, Y^* \rangle \leq \text{tr}(\hat{X})$ , or provides a separating hyperplane such that  $\langle X, Y \rangle > \text{tr}(\hat{X})$  and  $\langle X, Y \rangle \leq \text{tr}(X)$ ,  $\forall X \in S_+^n$ . We refer to this class of separating hyperplanes as “nuclear norm cuts”.

#### Pros and Cons of The Two Separation Oracles

In this section, we restated a well-known separation oracle for the PSD cone, and derived a new separation oracle. Conceptually, both oracles are similar, as they both rest upon semi-infinite reformulations of the positive semidefinite cone. However, they admit practical differences. For a matrix  $X$  with  $k$  negative eigenvalues, the trailing eigenvalue oracle provides  $k$  orthogonal separating hyperplanes, while the nuclear norm oracle only provides one hyperplane. Moreover, as observed by Permenter and Parrilo [17],  $k$  trailing-eigenvalue cuts can be aggregated into a single nonlinear cut, which ensures that  $X$  is positive semidefinite on the subspace spanned by all  $k$  trailing eigenvectors, by imposing a  $k \times k$  semidefinite constraint [17].

In spite of the aggregate power of trailing-eigenvalue cuts, nuclear norm cuts may sometimes be preferable. Indeed, while imposing trailing-eigenvalue cuts sometimes causes the multiplicity of  $X$ 's smallest eigenvalue to increase [21], nuclear norm cuts penalize the absolute sum of all  $n$  eigenvalues, rather than the worst-case eigenvalue, and therefore do not have this defect. Moreover, as we show in Section 4.2, nuclear norm cuts minimize the nuclear norm of a matrix in the problem's original space, while trailing eigenvalue cuts can only be applied to nuclear norm minimization after reformulating the problem in a higher-dimensional space, which is inefficient. We summarize the advantages and disadvantages of both cuts in Table 1.

Table 1: Summary of advantages (+)/ disadvantages (−) of the cuts.

Cut	Characteristics
Trailing EV	(+) Each cut generated in $\Omega(n^2)$ time
	(+) Multiple cuts per iterate
	(+) Cuts can be aggregated
	(−) Aggregating cuts is expensive
	(−) Models nuclear norm in extended space
Nuclear Norm	(+) Penalizes trailing eigenvalue clustering
	(+) Models nuclear norm in original space
	(−) Each cut generated in $\Omega(n^3)$ time

### 3. An Efficient Algorithmic Approach

In this section, we present an efficient numerical approach to solve Problem (1). Our approach is an outer-approximation strategy, which takes a tractable SOC-representable approximation of the original problem, and iteratively improves the approximation by adding cutting planes.

#### 3.1. A Cutting-Plane Method

The previous section presents tractable outer-approximations of Problem (1) and demonstrates that these approximations can be iteratively improved by adding cutting planes. Consequently, a valid numerical strategy for solving Problem (1) is the cutting-plane method, which was originally proposed by Kelley [12] under an assumption that the problem's feasible region is polyhedral. The cutting-plane method replaces the constraint  $X \geq \mathbf{0}$  with infinitely many linear constraints of the form  $f(X, Y) \leq 0$ , which must hold for each  $Y$  in some convex set  $\mathcal{Y}$ . Specifically,

$$f(X, Y) = -\langle X, Y \rangle \leq 0, \forall Y \in \mathcal{Y} = \{Y \in S_+^n : \|Y\|_* \leq 1\} \quad (4)$$

for trailing eigenvalue cuts and

$$f(X, Y) = \langle X, Y - \mathbb{I} \rangle \leq 0, \forall Y \in \mathcal{Y} = \{Y \in S_+^n : \|Y\|_{\sigma} \leq 1\} \quad (5)$$

for nuclear-norm cuts. After performing this reformulation, the cutting-plane method iteratively minimizes over outer-approximations of Problem (1), which are given by imposing finitely many constraints of the form  $f(X, Y^t) \leq 0, \forall t \in [T]$ .

These approximations are iteratively improved by (a) minimizing over them to obtain optimal but infeasible iterates  $X^t$ , and (b) refining them by adding a constraint corresponding to a most violated  $Y$  at  $X^t, Y^t$ , where

$$Y^t = \arg \max_{Y \in \mathcal{Y}} f(X^t, Y).$$

This numerical strategy yields a sequence of optimal but infeasible iterates, until some iterate  $X^t$  is  $\epsilon$ -feasible, i.e.,

$$f(X^t, Y^t) \leq \epsilon$$

at which point the method terminates with an optimal and  $\epsilon$ -feasible solution. We present this strategy in Algorithm 1.

We now prove that Algorithm 1 converges. Proofs of convergence for cutting-plane methods rely on compactness argu-

---

**Algorithm 1** Outer-approximation scheme

---

**Require:** Feasibility tolerance  $\epsilon$

$t \leftarrow 1$

**repeat**

  Compute  $X^{t+1}$  solution of

    min  $\langle C, X \rangle$

    s.t.  $\langle A_i, X \rangle = b_i, \forall i \in [m], f(X, Y^i) \leq 0, \forall i \in [t-1],$

$$\left\| \begin{pmatrix} 2X_{i,j} \\ X_{i,i} - X_{j,j} \end{pmatrix} \right\|_2 \leq X_{i,i} + X_{j,j}, \forall i \in [n], \forall j \in [n].$$

  Compute  $Y^{t+1}$  solution of  $\max_{Y \in \mathcal{Y}} f(X^t, Y)$ .

  Impose constraint  $f(X, Y^{t+1}) \leq 0$ .

$t \leftarrow t + 1$

**until**  $f(X^t, Y^t) \leq \epsilon$

**return**  $X^t$ 

---

ments, which can be established by bounding Problem (3)'s feasible region. To supply a proof of convergence, we require:

**Assumption 1.** All feasible solutions to Problem (1) are contained within a convex set  $\mathcal{X}$ , such that  $\text{tr}(X) \leq T, \forall X \in \mathcal{X}$ .

Assumption 1 is often satisfied naturally. For instance, the trace of  $X$  is constrained in SDO relaxations of binary quadratic and sparse principal component analysis problems. Alternatively, Assumption 1 can be satisfied by imposing the big-M constraint  $\text{tr}(X) \leq M$  for a large positive number  $M$ .

Our proof of convergence also rests on the requirement that the functions  $f(X, Y^t)$  are Lipschitz continuous, i.e., there exists a Lipschitz constant  $L$  such that for any feasible  $X_1, X_2$  in Problem (3) and any separating hyperplane  $Y$ ,

$$|f(X_1, Y) - f(X_2, Y)| \leq L \|X_1 - X_2\|_F.$$

This holds for both cuts outlined previously, by combining the Cauchy-Schwarz inequality with norm equivalence. Indeed,

$$|f(X_1, Y) - f(X_2, Y)| \leq \|X_1 - X_2\|_F \|Y\|_F = \|X_1 - X_2\|_F$$

for trailing eigenvalue cuts and

$$|f(X_1, Y) - f(X_2, Y)| \leq \|X_1 - X_2\|_F \|Y - \mathbb{I}\|_F \leq 2\sqrt{n} \|X_1 - X_2\|_F$$

for nuclear norm cuts. We now state our main result:

**Theorem 1.** Suppose that Assumption 1 holds and the separation oracle in Algorithm 1 is exact. Let  $X^t$  be a feasible solution returned by the  $t$ -th iterate of Algorithm 1, where

$$t \geq \left( \frac{LT}{\epsilon} + 1 \right)^n$$

$L$  is the Lipschitz constant and  $T$  is the aforementioned trace bound. Then,  $X^t$  is an optimal and  $\epsilon$ -feasible solution to Problem (1). Moreover, any limit point of  $\{X_t\}_{t=1}^\infty$  solves (1).

**Remark 1.** Note that the sequence of iterates  $\{X^t\}_{t=1}^\infty$  need not converge to a single solution  $X^*$ . For instance, there could be

multiple optimal solutions and multiple subsequences of  $\{X^t\}_{t=1}^\infty$  which each converge towards different minimizers. However, if Problem (1) admits a unique optimal solution and Assumption 1 holds then  $\{X^t\}_{t=1}^\infty$  converges to the unique optimal  $X^*$ .

**Remark 2.** An alternative to Assumption 1 is to regularize the objective with a strongly convex term  $\frac{1}{\gamma} \|X\|_F^2$  for some  $\gamma > 0$ . This term ensures the boundedness of a cutting-plane method's iterates without Assumption 1, by making the objective function coercive. Moreover, it guarantees that Algorithm 1 converges to the unique optimal  $X^*$ , because the regularized problem admits a unique optimal solution [see 4, Proposition 1.1.2].

Proofs of convergence of general cutting-plane methods have been supplied by a number of authors since the work of Kelley [12]; see for instance [4, Section 7.5.3], [16, Section 5.2]. To keep this paper self-contained, and make explicit Theorem 1's dependence upon Assumption 1, we now supply a proof of convergence; essentially due to [16]:

*Proof.* Suppose that at some iteration  $k > 1$ , Algorithm 1 has not converged. Then, we have that  $f(X^k, Y^k) > \epsilon$ , but  $f(X^k, Y^i) \leq 0, \forall i < k$ , which implies that

$$\epsilon < f(X^k, Y^k) - f(X^i, Y^k) \leq L \|X^i - X^k\|_F,$$

where the second inequality follows by Lipschitz continuity. Rearranging the above inequality implies  $\|X^i - X^k\|_F > \frac{\epsilon}{L}$ , i.e., Algorithm 1 never visits any ball of radius  $\frac{\epsilon}{L}$  twice. Moreover, by iteration  $k$ , Algorithm 1 has visited non-overlapping balls with combined volume

$$k \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2} + 1)} \left( \frac{\epsilon}{L} \right)^n,$$

and these balls must be centered at feasible points, i.e., contained within a ball of radius  $T + \frac{\epsilon}{L}$  which has volume

$$\frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2} + 1)} \left( T + \frac{\epsilon}{L} \right)^n.$$

That is, if Algorithm 1 has not converged at iteration  $k$ , we have:

$$k < \left( \frac{LT}{\epsilon} + 1 \right)^n,$$

which implies that we converge to an  $\epsilon$ -optimal solution within  $k \leq \left( \frac{LT}{\epsilon} + 1 \right)^n$  iterations. The result follows by re-arranging to establish that at each iteration  $k$  we obtain an  $\epsilon$ -feasible solution where  $\epsilon \leq \frac{nLT}{\log(k)-n}$ , and letting  $k \mapsto \infty$ .  $\square$

We note that Theorem 1's bound depends explicitly on Assumption 1's trace bound. Therefore, Algorithm 1 is particularly well-suited to problems with a constraint on  $\text{tr}(X)$ . Moreover, the above bound illustrates why a SOC-representable initial feasible region will outperform an LO-representable initial feasible region in practice: a worst-case rate of convergence which depends on  $\text{tr}(X)$ , rather than  $n \cdot \text{tr}(X)$ , often translates to numerically superior performance in practice.

We close this section by remarking that while IPMs require that strong duality holds, Algorithm 1 does not. Consequently,

Algorithm 1 may be a good method for SDOs which IPMs cannot solve because of constraint qualification failures.

#### 4. Numerical Results

In this section, we apply Algorithm 1 to large-scale sparse principal component analysis and nuclear norm minimization problems, using trailing eigenvalue (resp. nuclear norm) cuts for sparse PCA (resp. nuclear norm minimization). All experiments were implemented in Julia 1.0 using MOSEK 9.0 and JuMP.jl 0.18.5 [9], and performed on one Intel Xeon E5-2690 v4 2.6GHz CPU, using 32 GB RAM and six CPU cores with two threads each (12 threads total).

##### 4.1. Sparse Principal Component Analysis

Given high-dimensional data  $\mathbf{A} \in \mathbb{R}^{n \times p}$  and its normalized centered covariance matrix  $\mathbf{\Sigma} := \frac{1}{p-1} \mathbf{A}^\top \mathbf{A} \in \mathbb{S}^n$ , a common task is to compress  $\mathbf{A}$ , by projecting  $\mathbf{A}$  onto a small number of principal components. This task is known as principal component analysis (PCA), and is achieved by performing singular value decomposition to obtain  $\mathbf{\Sigma} = \mathbf{S} \mathbf{\Lambda} \mathbf{S}^\top$ , and projecting  $\mathbf{A}$  onto the leading eigenvectors via  $\mathbf{A}_{\text{new}} = \mathbf{S}_{[1:k]} \mathbf{A}$ .

PCA is sometimes criticized for producing uninterpretable features, because each new feature is a linear combination of all  $n$  original features. This is often unacceptable because:

- In medical applications, downstream decisions taken via PCA must be interpretable.
- In financial applications such as investing across index funds, each non-zero entry in each PC incurs a cost.

One approach to obtain interpretable principal components is to stipulate that they are sparse, i.e., have at most  $k$  non-zero entries. This approach leads to the following problem [see 8]:

$$\lambda_{\max}^k(\mathbf{\Sigma}) := \max_{\mathbf{x} \in \mathbb{R}^n} \mathbf{x}^\top \mathbf{\Sigma} \mathbf{x} \text{ s.t. } \mathbf{x}^\top \mathbf{x} \leq 1, \|\mathbf{x}\|_0 \leq k. \quad (6)$$

Unfortunately, this problem cannot be solved to certifiable optimality when  $n > 100$  [see 3, for a certifiably optimal approach].

Motivated by the observation that SDO relaxations provide near-exact upper bounds on  $\lambda_{\max}^k(\mathbf{\Sigma})$  in practice, d’Aspremont et al. [8] proposed the relaxation:

$$\max_{\mathbf{X} \in \mathbb{S}_+^n} \langle \mathbf{\Sigma}, \mathbf{X} \rangle \text{ s.t. } \text{tr}(\mathbf{X}) = 1, \|\mathbf{X}\|_1 \leq k. \quad (7)$$

Unfortunately, modern IPMs cannot solve Problem (7) for  $n > 250$ , because all known reformulations of the constraint  $\|\mathbf{X}\|_1 \leq k$  require  $\Omega(n^2)$  linear inequality constraints. Therefore, we apply Algorithm 1 (with trailing eigenvalue cuts) to this problem, and compare the bounds obtained via Algorithm 1 to the best feasible solution obtained by the method of Berk and Bertsimas [3] (with a time limit of 60 seconds). We also compare our bounds with the exact SDO bound given in Problem (7), wherever it can be computed by MOSEK within our 32GB peak memory budget.

For completeness, we now describe the datasets which we benchmark our approach on. Note that the first 3 datasets are distributed via the UCI Machine Learning Repository<sup>3</sup>:

- The Pitprops dataset: a  $13 \times 13$  covariance matrix derived from 180 observations of 13 features.
- The normalized communities dataset: a  $101 \times 101$  correlation matrix derived from 1994 observations of 128 features, after eliminating 27 categorical variables.
- The normalized Arrhythmia dataset: a  $274 \times 274$  correlation matrix derived from 452 observations of 279 features, after eliminating 5 categorical variables.
- Wilshire 5000: a  $2130 \times 2130$  correlation matrix derived from 2768 observations of changes in daily stock prices from 3 January 2007 to 31 December 2017, obtained via Yahoo! Finance using the R package *quantmod* [see 20], after eliminating stocks which were not traded in all 2768 days considered. As optimizing over  $n^2$  SOC constraints is too expensive when  $n > 2000$ , we adopt a more tractable initial approximation for the Wilshire 5000 dataset. Namely, we aggregate the  $n^2$  SOC constraints into  $n$  constraints, by (a) observing that

$$X_{i,i} X_{j,j} \geq X_{i,j}^2 \implies X_{i,i} \geq \sum_{j=1}^n X_{i,j}^2,$$

since  $\text{tr}(\mathbf{X}) = 1$ , and (b) introducing the auxiliary variables  $z$  to model a relaxation of the sparsity constraint  $\|\mathbf{x}\|_0 \leq k$ . This yields the following SOC relaxation, which is weaker than d’Aspremont et al. [8]’s relaxation, but has  $O(n)$ , rather than  $O(n^2)$ , SOC/linear inequality constraints:

$$\min_{\mathbf{X} \in \mathbb{S}_+^n, \mathbf{z} \in \mathbb{R}_+^n} \langle \mathbf{\Sigma}, \mathbf{X} \rangle \text{ s.t. } \text{tr}(\mathbf{X}) = 1, \mathbf{z} \leq \mathbf{e}, \sum_{j=1}^n X_{i,j}^2 \leq z_i X_{i,i}, \forall i \in [n]$$

We now present our experimental results. Table 2 depicts the time required for Algorithm 1 to generate a bound with 0, 5 or 20 cuts, and the time required for MOSEK to solve the SDO relaxation when  $k = 10$ ; Table 3 depicts the magnitude of the bound gaps for the different bounds when  $k = 10$ ; Figure 1 depicts the quality of the SOC and SDO bounds, for varying  $k$ , against the optimal solution obtained by the algorithm of [3] (for  $n \leq 101$ ) or the best solution obtained after 60 seconds (for  $n > 101$ ). We do not generate cuts for the Wilshire 5000 dataset, as the SOC bound is sufficiently tight here.

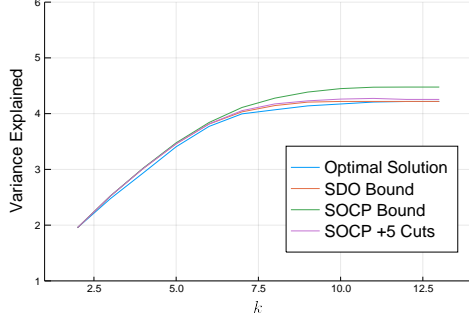
Table 2: Runtime (s) for  $k = 10$  (resp.  $k = 50$ ) for UCI (Wilshire) datasets.

Problem	Approach runtime (s)			
	SOC	SOC 5 cuts	SOC 20 cuts	SDO
pitprops	0.1	0.2	0.7	0.1
norm communities	2.8	9.7	26.3	233.6
norm arrhythmia	37.5	188.6	564.1	n/a
norm wilshire	870.6	n/a	n/a	n/a

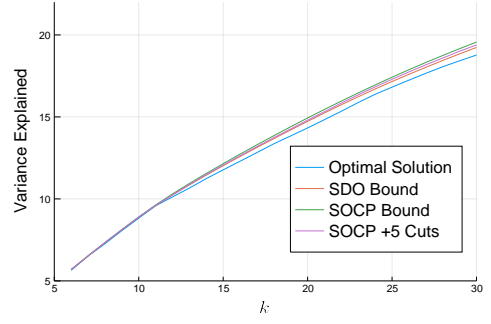
Our main findings from this set of experiments are as follows:

- Algorithm 1 provides bounds for sparse PCA problems which are as tight as those given by the SDO relaxation of [8], yet scale to problems an order of magnitude larger.
- For sparse PCA problems where  $n \leq 1000$ , Problem (3)’s outer approximation provides an accurate upper bound in a tractable fashion. Moreover, for larger problems where

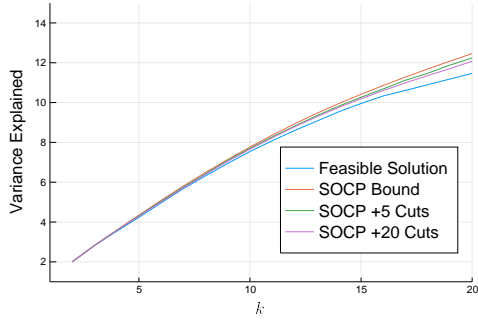
<sup>3</sup><https://archive.ics.uci.edu/ml/datasets.html>.



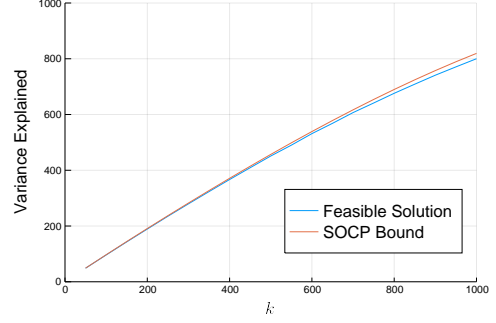
(a) Pitprops data set ( $n = 13$ ).



(b) Normalized communities data set ( $n = 101$ ).



(c) Normalized arrhythmia data set ( $n = 274$ ).



(d) Normalized Wilshire data set ( $n = 2130$ ).

Figure 1: Quality of SOC and SDO bounds on the leading principal component for different UCI datasets.

Table 3: Bound gap (%) for  $k = 10$  (resp.  $k = 50$ ) for UCI (Wilshire) datasets.

Problem	Bound gap (%)			
	SOC	SOC 5 cuts	SOC 20 cuts	SDO
pitprops	6.60	2.10	1.11	1.09
norm communities	0.83	0.75	0.75	0.75
norm arrhythmia	3.01	2.16	1.54	n/a
norm wilshire	0.48	n/a	n/a	n/a

$1000 < n \leq 10000$ , aggregating the SOCP constraints yields a looser yet still tractable bound. This is because Problem (7) contains the constraint  $\text{tr}(\mathbf{X}) = 1$ , which, as established in Proposition 4, provides tight constraint violation guarantees.

#### 4.2. Nuclear Norm Minimization

Given an incomplete set of observations  $A_{(i,j)} \forall (i,j) \in \Omega$  of a low-rank matrix  $\mathbf{A} \in S^n$ , a central problem in machine learning is to recover the entire matrix, by solving:

$$\min \text{Rank}(\mathbf{X}) \text{ s.t. } A_{(i,j)} = X_{(i,j)}, \forall (i,j) \in \Omega. \quad (8)$$

Unfortunately, all known algorithms for solving Problem (8) to certifiable optimality require **doubly** exponential time in both theory and practice [19]. Consequently, various authors including [19] have proposed instead solving Problem (8)'s convex relaxation, which is:

$$\min \|\mathbf{X}\|_* \text{ s.t. } A_{(i,j)} = X_{(i,j)}, \forall (i,j) \in \Omega. \quad (9)$$

By exploiting Schur complements, Recht et al. [19] have established that Problem (9) is SDO-representable. Unfortunately, their reformulation is intractable for even medium-sized problems, as it rests upon lifting the nuclear norm objective to a higher dimensional space. An alternative approach is to apply Algorithm 1 with nuclear norm cuts, i.e., exploit Proposition 6's semi-infinite reformulation of the nuclear norm objective, and iteratively solve a sequence of SOCPs of the form:

$$\begin{aligned} \min \quad & \theta \quad \text{s.t.} \quad \theta \geq \|\mathbf{X}\|_F, \\ & \theta \geq \langle \mathbf{X}, \mathbf{Y} \rangle, \quad \forall \mathbf{Y} \in \bar{\mathcal{Y}}, \\ & A_{(i,j)} = X_{(i,j)}, \quad \forall (i,j) \in \Omega, \end{aligned}$$

where we impose the SOC-representable Frobenius norm constraint, because it is valid by the relation  $\|\mathbf{X}\|_F \leq \|\mathbf{X}\|_*$  [see 11]. Moreover, Proposition 6 proves that at each iteration  $t$  a most-violated  $\mathbf{Y}^t$  in the semi-infinite constraint  $\theta \geq \langle \mathbf{X}, \mathbf{Y} \rangle, \forall \mathbf{Y} \in \bar{\mathcal{Y}}$  is given by taking an SVD of  $\mathbf{X}^t = \mathbf{U}\Sigma\mathbf{U}^\top$  and setting  $\mathbf{Y} = \mathbf{U}\mathbf{U}^\top$ .

When we performed our numerical experiments, rather than solving Problem (9), we applied Algorithm 1 to the following problem (taking  $\gamma = \frac{1}{n}$  in our experiments).

$$\begin{aligned} \min \quad & \|\mathbf{X}\|_* + \frac{1}{\gamma} \|\mathbf{X}\|_F^2 \\ \text{s.t.} \quad & A_{(i,j)} = X_{(i,j)}, \quad \forall (i,j) \in \Omega. \end{aligned} \quad (10)$$

Solving this problem as a surrogate for Problem (9) serves a dual purpose: (a) it ensures that Problem (9) has a unique solution, which prevents degeneracy and discourages stalling be-

haviour, and (b) it ensures that  $\{X_t\}_{t=1}^{\infty}$  converges to Problem (10)’s unique solution; see Remark 2.

We now compare the time required for our approach to obtain a solution within 0.1% of optimality, to the time required for MOSEK to solve a semidefinite reformulation of the problem<sup>4</sup>, where we generate a rank-10 matrix  $R$  from two factors  $U, V$  with i.i.d.  $\mathcal{N}(0, 1)$  entries, and randomly sample half the entries from the matrix, i.e., let  $|\Omega| = \frac{n^2}{2}$ . Figure 2 depicts the runtime requirement for both approaches (averaged over 5 instances), and demonstrates that Algorithm 1 outperforms IPMs when  $n \geq 100$  (i.e., there are  $\geq 5000$  constraints).

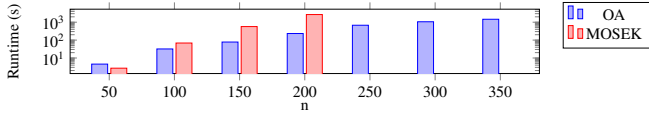


Figure 2: Runtime requirement for Algorithm 1 with nuclear norm cuts (“OA”) vs MOSEK applied to [19]’s reformulation. Note that MOSEK cannot solve instances of [19]’s SDO-representable reformulation with  $n \geq 250$ , as this requires a peak memory budget  $> 32$  GB RAM.

Our approach also generates solutions to Problem (10) for larger problem sizes. Figure 3 depicts the convergence profile of our approach for an instance (generated in the same manner as the previous experiment) where  $n = 500$ ; convergence was attained after 1245 seconds. The convergence profile is typical: we obtain a near-exact lower bound within 10 iterations, and require many more iterations to identify the optimal solution.

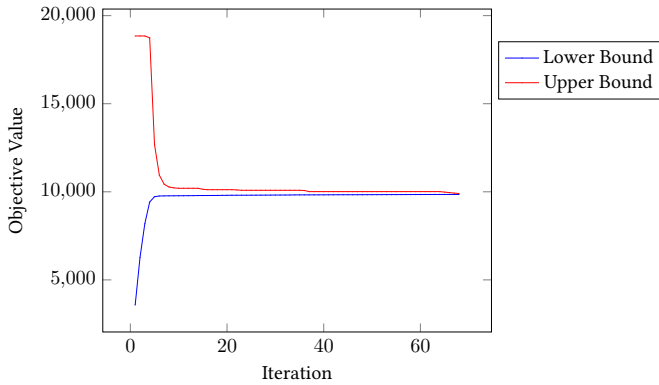


Figure 3: A convergence profile of Algorithm 1 for  $n = 500$ . The upper bound is given by the smallest value of  $\|X^t\|_* + \frac{1}{\gamma}\|X^t\|_F^2$  observed by iteration  $t$ .

Our main set of findings from this experiment are as follows:

- Although IPMs scale better than outer-approximation methods from a traditional complexity theory perspective, in practice Kelley’s cutting-plane method solves nuclear norm minimization problems faster. This is because (a) Kelley’s method solves nuclear norm minimization problems in their original space, while an SDO reformulation lifts the problem

via Schur complements, and (b) while the traditional complexity theory accurately describes the performance of IPMs, it is too conservative with respect to cutting-plane methods.

- The bottleneck in applying Algorithm 1 to larger-scale matrix completion problems is the SVD step, which scales at a rate of  $\Omega(n^3)$ . Therefore, one future direction is to replace the SVD step with an inexact oracle, which performs matrix sketching to yield valid inexact cuts in  $\Omega(n^2)$  time.

## Acknowledgements

We are grateful to Peter Cohen for editorial comments.

## References

- [1] A. A. Ahmadi and A. Majumdar. Dsos and sdsos optimization: more tractable alternatives to sum of squares and semidefinite optimization. *SIAM J. Appl. Alg. Geom.*, 3(2):193–230, 2019.
- [2] F. Alizadeh. Interior point methods in semidefinite programming with applications to combinatorial optimization. *SIAM J. Opt.*, 5(1):13–51, 1995.
- [3] L. Berk and D. Bertsimas. Certifiably optimal sparse principal component analysis. *Math. Prog. Comp.*, to appear, pages 1–40, 2019.
- [4] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.
- [5] D. Bertsimas, R. Cory-Wright, and J. Pauphilet. A unified approach to mixed-integer optimization: Nonlinear formulations and scalable algorithms. *arXiv:1907.02109*, 2019.
- [6] G. Braun, S. Fiorini, S. Pokutta, and D. Steurer. Approximation limits of linear programs (beyond hierarchies). *Math. Oper. Res.*, 40(3):756–772, 2015.
- [7] G. Dantzig, R. Fulkerson, and S. Johnson. Solution of a large-scale traveling-salesman problem. *J. Oper. Res. Soc. Am.*, 2(4):393–410, 1954.
- [8] A. d’Aspremont, L. El Ghaoui, M. Jordan, and G. Lanckriet. A direct formulation for sparse pca using semidefinite programming. In *NIPS*, pages 41–48, 2005.
- [9] I. Dunning, J. Huchette, and M. Lubin. Jump: A modeling language for mathematical optimization. *SIAM Rev.*, 59(2):295–320, 2017.
- [10] M. Fischetti, I. Ljubić, and M. Sinnl. Redesigning benders decomposition for large-scale facility location. *Manag. Sci.*, 63(7):2146–2162, 2016.
- [11] R. A. Horn and C. R. Johnson. *Matrix analysis*. 1990.
- [12] J. E. Kelley, Jr. The cutting-plane method for solving convex programs. *J. Soc. Ind. Appl. Math.*, 8(4):703–712, 1960.
- [13] S. Kim and M. Kojima. Second order cone programming relaxation of nonconvex quadratic optimization problems. *Opt. Meth. Soft.*, 15(3-4): 201–224, 2001.
- [14] B. Kocuk, S. S. Dey, and X. A. Sun. Strong socp relaxations for the optimal power flow problem. *Oper. Res.*, 64(6):1177–1196, 2016.
- [15] K. Krishnan and J. E. Mitchell. A unifying framework for several cutting plane methods for semidefinite programming. *Opt. Meth. Soft.*, 21(1): 57–74, 2006.
- [16] A. Mutapcic and S. Boyd. Cutting-set methods for robust convex optimization with pessimizing oracles. *Opt. Meth. Soft.*, 24(3):381–406, 2009.
- [17] F. Permenter and P. Parrilo. Partial facial reduction: simplified, equivalent sdps via approximations of the psd cone. *Math. Prog.*, 171:1–54, 2018.
- [18] M. V. Ramana. An exact duality theory for semidefinite programming and its complexity implications. *Math. Prog.*, 77(1):129–162, 1997.
- [19] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, 2010.
- [20] J. A. Ryan. quantmod: Quantitative financial modelling framework. *R package*, 2008.
- [21] K. K. Sivaramakrishnan. *A Linear programming approach to semidefinite programming problems*. PhD thesis, RPI, 2002.
- [22] H. Wolkowicz and G. Styann. Bounds for eigenvalues using traces. *Lin. Alg. Appl.*, 29:471–506, 1980.

<sup>4</sup>As MOSEK 9.0 does not allow SOC constraints and quadratic objective terms to be mixed, we model the Frobenius norm term by invoking the equivalence  $t \geq \|x\|_2^2 \iff t + 1 \geq \left\| \begin{pmatrix} 2x \\ t-1 \end{pmatrix} \right\|_2$ .