On the Partial Convexification of the Low-Rank Spectral Optimization: Rank Bounds and Algorithms

Yongchun Li
Stewart School of Industrial & Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, ycli@gatech.edu

Weijun Xie
Stewart School of Industrial & Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, wxie@gatech.edu

A Low-rank Spectral Optimization Problem (LSOP) minimizes a linear objective function subject to multiple two-sided linear inequalities intersected with a low-rank and spectral constrained domain set. Although solving LSOP is, in general, NP-hard, its partial convexification (i.e., replacing the domain set by its convex hull) termed "LSOP-R", is often tractable and yields a high-quality solution. This motivates us to study the strength of LSOP-R. Specifically, we derive rank bounds for any extreme point of the feasible set of LSOP-R with different matrix spaces and prove their tightness. The proposed rank bounds recover two well-known results in the literature from a fresh angle and allow us to derive sufficient conditions under which the relaxation LSOP-R is equivalent to the original LSOP. To effectively solve LSOP-R, we develop a column generation algorithm with a vector-based convex pricing oracle, coupled with a rank-reduction algorithm, which ensures that the output solution always satisfies the theoretical rank bound. Finally, we numerically verify the strength of the LSOP-R and the efficacy of the proposed algorithms.

Key words: Low Rank, Partial Convexification, Rank Bounds, Column Generation, Rank Reduction.

1. Introduction

This paper studies the Low-rank Spectral Optimization Problem (LSOP) of the form:

$$\mathbf{V}_{\text{opt}} := \min_{\boldsymbol{X}\in\mathcal{D}} \left\{ \langle \boldsymbol{A}_0, \boldsymbol{X}\rangle : b_i^l \le \langle \boldsymbol{A}_i, \boldsymbol{X}\rangle \le b_i^u, \forall i \in [m] \right\}, \tag{LSOP}$$

where the decision $\boldsymbol{X}$ is a matrix variable with domain $\mathcal{D}$, $\langle\cdot,\cdot\rangle$ denotes the Frobenius inner product of two equal-sized matrices, we have that $-\infty \le b_i^l \le b_i^u \le +\infty$ for each $i \in [m]$, and matrices $\boldsymbol{A}_0$ and $\{\boldsymbol{A}_i\}_{i\in[m]}$ are symmetric if the matrix variable $\boldsymbol{X}$ is symmetric. Throughout, we let $\tilde{m}$ denote the number of linearly independent matrices in the set $\{\boldsymbol{A}_i\}_{i\in[m]}$.

Specifically, the domain set $\mathcal{D}$ in LSOP is defined as below

$$\mathcal{D} := \{\boldsymbol{X} \in \mathcal{Q} : \text{rank}(\boldsymbol{X}) \le k, F(\boldsymbol{X}) := f(\boldsymbol{\lambda}(\boldsymbol{X})) \le 0\}, \tag{1}$$

which consists of a low-rank and a closed convex spectral constraint. Note that (i) the matrix space $\mathcal{Q}$ can denote positive semidefinite matrix space $\mathcal{S}_+^n$, non-symmetric matrix space $\mathbb{R}^{n\times p}$, symmetric indefinite matrix space $\mathcal{S}^n$, or diagonal matrix space with $k \le n \le p$ being positive integers; (ii) function $F(\boldsymbol{X}): \mathcal{Q} \to \mathbb{R}$ is continuous, convex, and spectral which only depends on the eigenvalue or singular value vector $\boldsymbol{\lambda}(\boldsymbol{X}) \in \mathbb{R}^n$ of matrix $\boldsymbol{X}$. Thus, we can rewrite it as $F(\boldsymbol{X}) := f(\boldsymbol{\lambda}(\boldsymbol{X})) : \mathbb{R}^n \to \mathbb{R}$; and (iii) when there are multiple convex spectral constraints, i.e., $F_j(\boldsymbol{X}) \le 0, \forall j \in J$, we can integrate them into a single convex spectral constraint by defining a function $F(\boldsymbol{X}) := \max_{j\in J} F_j(\boldsymbol{X})$. Hence, the domain set $\mathcal{D}$ readily covers multiple spectral constraints. Such a set $\mathcal{D}$ naturally appears in many machine learning and optimization problems with a low-rank constraint (see Subsection 1.1).

The low-rank constraint dramatically complicates LSOP, which often turns out to be an intractable nonconvex bilinear program. Thus, we leverage the convex hull of domain set $\mathcal{D}$, denoted by conv($\mathcal{D}$), to obtain a partial convexification over $\mathcal{D}$ for LSOP, termed LSOP-R throughout:

$$\mathbf{V}_{\text{rel}} := \min_{\boldsymbol{X}\in \text{conv}(\mathcal{D})} \left\{ \langle \boldsymbol{A}_0, \boldsymbol{X} \rangle : b_i^l \le \langle \boldsymbol{A}_i, \boldsymbol{X} \rangle \le b_i^u, \forall i \in [m] \right\} \le \mathbf{V}_{\text{opt}}, \tag{LSOP-R}$$

where the inequality is because that LSOP-R serves as a convex relaxation of LSOP.

Since the LSOP requires the solution to be at most rank-k and the more tractable relaxation LSOP-R may not enforce the rank constraint, a guaranteed low-rank solution of the LSOP-R can consistently favor various application requirements. That is, we are interested in finding the smallest integer $\widehat{k} \ge k$ such that there is an optimal solution $\boldsymbol{X}^*$ of LSOP-R satisfying rank$(\boldsymbol{X}^*) \le \widehat{k}$. More importantly, the rank bounds are useful to develop approximation algorithms for LSOP and solution algorithms for LSOP-R, as well as to understand when the relaxation LSOP-R meets LSOP (see, e.g., Burer and Monteiro 2003, 2005, Burer and Ye 2020, Lau et al. 2011, Tantipongpipat et al. 2019). The analysis of rank bounds further inspires us a rank-reduction algorithm for LSOP-R. Hence, this paper aims to provide: (i) theoretical rank bounds of LSOP-R solutions; and (ii) effective algorithms for solving LSOP-R while satisfying rank bounds.

1.1. Scope and Flexibility of Our LSOP Framework

In this subsection, we discuss several interesting low-rank constrained problems in different matrix space $\mathcal{Q}$ where the proposed LSOP framework can be applied (as a substructure). For those application examples, we specify their corresponding LSOP-Rs along with the rank bounds later.

Quadratically Constrained Quadratic Program (QCQP) with $\mathcal{Q} := \mathcal{S}^n_+$. The QCQP has been widely studied in many application areas, including optimal power flow, sensor network problems, signal processing (Josz et al. 2016, Gharanjik et al. 2016, Khobahi et al. 2019), among others. The QCQP of matrix form can be viewed as a special case of the proposed LSOP:

$$\min_{\boldsymbol{X}\in \mathcal{D}} \left\{ \langle \boldsymbol{A}_0, \boldsymbol{X} \rangle : b_i^l \le \langle \boldsymbol{A}_i, \boldsymbol{X} \rangle \le b_i^u, \forall i \in [m] \right\}, \quad \mathcal{D} := \{ \boldsymbol{X} \in \mathcal{S}^n_+ : \text{rank}(\boldsymbol{X}) \le 1 \}, \tag{QCQP}$$

where $F(\boldsymbol{X})=0$. In fact, the domain set $\mathcal{D}$ can be extended to incorporate any closed convex spectral function $F(\boldsymbol{X})$. For example, if there is a ball constraint (i.e., $\boldsymbol{x}^\top \boldsymbol{x}\le 1$) in the QCQP like trust region subproblem, we can add a spectral constraint $\operatorname{tr}(\boldsymbol{X})=1+\boldsymbol{x}^\top\boldsymbol{x}\le 2$ into set $\mathcal{D}$.

Low-Rank Kernel Learning with $\mathcal{Q}:=\mathcal{S}_+^n$. Given an input kernel matrix $\boldsymbol{Y}\in\mathcal{S}_+^n$ of rank up to k, the low-rank kernel learning aims to find a rank $\le k$ matrix $\boldsymbol{X}$ that closely approximates $\boldsymbol{Y}$, subject to additional linear constraints. The Log-Determinant divergence is a popular measure of the kernel learning quality (see, e.g., Kulis et al. 2009), defined as $\langle \boldsymbol{Y}^\dagger,\boldsymbol{X}\rangle-\log\det[(\boldsymbol{X}+\alpha\boldsymbol{I}_n)(\boldsymbol{Y}+\alpha\boldsymbol{I}_n)^{-1}]$, where $\alpha>0$ is small and $\boldsymbol{I}_n$ is the identify matrix for addressing the rank deficiency issue. Thus, the kernel learning under Log-Determinant divergence, as formulated by Kulis et al. (2009), becomes

$$\begin{aligned}&\min_{z\in\mathbb{R},\boldsymbol{X}\in\mathcal{D}}\left\{\langle \boldsymbol{Y}^\dagger,\boldsymbol{X}\rangle-z+\log\det(\boldsymbol{Y}+\alpha\boldsymbol{I}_n):b_i^l\le\langle\boldsymbol{A}_i,\boldsymbol{X}\rangle\le b_i^u,\forall i\in[m]\right\},\\ &\mathcal{D}:=\{\boldsymbol{X}\in\mathcal{S}_+^n:\operatorname{rank}(\boldsymbol{X})\le k,\log\det(\boldsymbol{X}+\alpha\boldsymbol{I}_n)\ge z\}.\end{aligned}\qquad\text{(Kernel Learning)}$$

Here, a closed convex spectral function is defined as $F(\boldsymbol{X}):=z-\log\det(\boldsymbol{X}+\alpha\boldsymbol{I}_n)$ for any z. The proposed LSOP naturally performs a substructure over variable $\boldsymbol{X}$ in Kernel Learning.

Fair PCA with $\mathcal{Q}:=\mathcal{S}_+^n$. It is recognized that the conventional Principal Component Analysis (PCA) may generate biased learning results against sensitive attributes, such as gender, race, or education level (Samadi et al. 2018). Fairness has recently been introduced to the PCA. Formally, the seminal work by Tantipongpipat et al. (2019) formulates the fair PCA as follows in which the substructure over $\boldsymbol{X}$ is a special case of the proposed LSOP

$$\max_{z\in\mathbb{R}_+,\boldsymbol{X}\in\mathcal{D}}\{z:z\le\langle\boldsymbol{A}_i,\boldsymbol{X}\rangle,\forall i\in[m]\},\quad \mathcal{D}:=\{\boldsymbol{X}\in\mathcal{S}_+^n:\operatorname{rank}(\boldsymbol{X})\le k,||\boldsymbol{X}||_2\le 1\},\qquad\text{(Fair PCA)}$$

where $F(\boldsymbol{X}):=||\boldsymbol{X}||_2-1$ denotes the spectral norm (i.e., the largest singular value) of a matrix and matrices $\boldsymbol{A}_1,\cdots\boldsymbol{A}_m\in\mathcal{S}_+^n$ denote the sample covariance matrices from m different groups.

Fair SVD with $\mathcal{Q}:=\mathbb{R}^{n\times p}$. Similar to Fair PCA, fair Singular Value Decomposition (SVD) seeks a fair representation of m different data matrices that are non-symmetric, as formulated below

$$\max_{z\in\mathbb{R},\boldsymbol{X}\in\mathcal{D}}\{z:z\le\langle\boldsymbol{A}_i,\boldsymbol{X}\rangle,\forall i\in[m]\},\quad \mathcal{D}:=\{\boldsymbol{X}\in\mathbb{R}^{n\times p}:\operatorname{rank}(\boldsymbol{X})\le k,||\boldsymbol{X}||_2\le 1\}.\qquad\text{(Fair SVD)}$$

Here, matrices $\boldsymbol{A}_1,\cdots\boldsymbol{A}_m\in\mathbb{R}^{n\times p}$ are non-symmetric.

Matrix Completion with $\mathcal{Q}:=\mathbb{R}^{n\times p}$. In the matrix completion problem, the objective is to recover a low-rank matrix $\boldsymbol{X}\in\mathbb{R}^{n\times p}$ from a subset of observed entries $\{A_{ij}\}_{(i,j)\in\Omega\subseteq[n]\times[p]}$, which arises from a variety of applications including recommendation systems, computer vision, and signal processing

(Chakraborty et al. 2013, Miao and Kou 2021). The proposed LSOP appears as a substructure of variable $\boldsymbol{X}$ in the following matrix completion formulation

$$\min_{\boldsymbol{X}\in\mathcal{D}, z\in\mathbb{R}_+} \{z : X_{ij} = A_{ij}, \forall (i,j)\in\Omega\}, \quad \mathcal{D} := \{\boldsymbol{X}\in\mathbb{R}^{n\times p} : \operatorname{rank}(\boldsymbol{X}) \le k, \|\boldsymbol{X}\|_* \le z\}, \tag{Matrix Completion}$$

where $F(\boldsymbol{X}) := \|\boldsymbol{X}\|_* - z$ builds on the nuclear norm. The LSOP-R counterpart of matrix completion reduces to a widely studied convex relaxation in the literature (see Section 3).

Sparse Ridge Regression with $\mathcal{Q} := \mathcal{S}^n$. As machine learning may encounter with datasets with many features, the sparsity has been enforced to select a handful of important ones to improve the interpretability of the learning outcomes. The zero norm that denotes the sparsity of a vector is equal to the rank of the corresponding diagonal matrix constructed by this vector. Hence, when $\mathcal{Q} := \mathcal{S}^n$ in (1), the special diagonal matrix space allows us to formulate LSOP with the known sparse ridge regression problem (see, e.g., Xie and Deng 2020) as

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \left\{ \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2/n + \alpha\|\boldsymbol{x}\|_2^2 : \|\boldsymbol{x}\|_0 \le k \right\},$$

where $\boldsymbol{A}\in\mathbb{R}^{m\times n}$ represents the data matrix, $\alpha > 0$ is a regularizer, and $\|\boldsymbol{x}\|_0$ denotes the zero norm. By introducing the auxiliary variables $\boldsymbol{y} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}$, $z \ge \|\boldsymbol{x}\|_2^2$ and letting $\boldsymbol{X} = \operatorname{Diag}(\boldsymbol{x})$, the sparse ridge regression reduces to

$$\begin{aligned} &\min_{\boldsymbol{X}\in\mathcal{D}, \boldsymbol{y}\in\mathbb{R}^m, z\in\mathbb{R}_+} \left\{ \|\boldsymbol{y}\|_2^2/n + \alpha z : \boldsymbol{y} = \boldsymbol{b} - \boldsymbol{A}\operatorname{diag}(\boldsymbol{X}) \right\}, \\ &\mathcal{D} := \left\{ \boldsymbol{X}\in\mathcal{S}^n : \boldsymbol{X} = \operatorname{Diag}(\operatorname{diag}(\boldsymbol{X})), \operatorname{rank}(\boldsymbol{X}) \le k, \|\boldsymbol{X}\|_F^2 \le z \right\}, \end{aligned} \tag{Sparse Ridge Regression}$$

where the constraint $\boldsymbol{X} = \operatorname{Diag}(\operatorname{diag}(\boldsymbol{X}))$ enforces matrix $\boldsymbol{X}$ to be diagonal and thus $\|\boldsymbol{X}\|_F^2 = \|\operatorname{diag}(\boldsymbol{X})\|_2^2 = \|\boldsymbol{x}\|_2^2 \le z$ must hold. The subproblem over $\boldsymbol{X}$ above follows our LSOP framework.

1.2. Relevant Literature

In this subsection, we survey the relevant literature on two aspects.

Convexification of Low-Rank Spectral Domain Set $\mathcal{D}$. There are few works on the convexification of a low-rank spectral domain set $\mathcal{D}$ in (1). The work by Bertsimas et al. (2021) has successfully extended the perspective technique to convexify a special low-rank set $\mathcal{D}$ with $\mathcal{Q} := \mathcal{S}^n$ in which all the eigenvalues in the function $F(\boldsymbol{X}) := f(\boldsymbol{\lambda}(\boldsymbol{X}))$ are separable. Such approach, however, may fail to cover the general set $\mathcal{D}$, e.g., the spectral norm function $F(\boldsymbol{X}) := \|\boldsymbol{X}\|_2$ in the Fair PCA which is not separable. Another seminal work (Kim et al. 2022) leverages the majorization technique on the convexification of any permutation-invariant set. We observe that our domain set $\mathcal{D}$ in (1) is permutation-invariant with eigenvalues or singular values; thus, the majorization technique can be applied. It should be noted that the majorization technique may not always provide an explicit description for our convex hull $\text{conv}(\mathcal{D})$, except making further assumptions on the spectral constraint, as shown in Section 4. To resolve this limitation, this paper studies the inner approximation of the convex hull $\text{conv}(\mathcal{D})$ to obtain a tractable relaxation of LSOP with the matrix space $\mathcal{Q} := \mathcal{S}^n$. This paper focuses on bounding the ranks of LSOP-R solutions, and the convexification result of Kim et al. (2022) paves the path for the derivation of our bounds.

Rank Bounds for LSOP-R. Given a domain set $\mathcal{D} = \{\boldsymbol{X} \in \mathcal{S}_+^n : \text{rank}(\boldsymbol{X}) \leq 1\}$ particularly adopted in QCQP, its convex hull is the positive semidefinite matrix space and the corresponding LSOP-R feasible set is a spectrahedron. Barvinok (1995), Deza et al. (1997), Pataki (1998) independently showed that the rank of any extreme point in the spectrahedron is in the order of $\mathcal{O}(\sqrt{2\tilde{m}})$. Recently, in the celebrated paper on Fair PCA (Tantipongpipat et al. 2019), the authors also proved a rank bound for all feasible extreme points of the corresponding LSOP-R, which is $\mathcal{O}(\sqrt{2m})$, and the proof technique extended that of Pataki (1998). Another relevant work pays particular attention to the sparsity bound for the sparse optimization problem (Askari et al. 2022), which relies on a different continuous relaxation from the LSOP-R. Since the LSOP can encompass the sparse optimization when $\mathcal{Q}$ in (1) denotes the diagonal matrix space, our rank bounds are in fact sparsity bounds under this setting. To the best of our knowledge, this is the first work to study the rank bounds for the generic partial convexification– LSOP-R. Our rank bounds recover all the ones reviewed here for QCQP and Fair PCA from a different perspective, and successfully reduce the sparsity bound of Askari et al. (2022) when applying to Sparse Ridge Regression.

1.3. Contributions and Outline

We theoretically guarantee the solution quality of the relaxation LSOP-R by bounding the ranks of all feasible extreme points with three different matrix spaces (i.e., $\mathcal{Q} = \mathcal{S}_+^n$, $\mathcal{Q} = \mathbb{R}^{n \times p}$, and $\mathcal{Q} = \mathcal{S}^n$), respectively, and each matrix space has its own advantages and applications, as exemplified in Subsection 1.1. Notably, our rank bounds hold for any domain set $\mathcal{D}$ in the form of (1) and for any $\tilde{m}$ linearly independent inequalities, and they are attainable by the worst-case instances. Below summarizes the major contributions of this paper.

(i) Section 2 studies LSOP-R with the positive semidefinite matrix space, i.e., $\mathcal{Q} := \mathcal{S}_+^n$. We show that the rank of any extreme point in the feasible set of LSOP-R deviates at most $\lfloor \sqrt{2\tilde{m} + 9/4} - 3/2 \rfloor$ from the original rank-k requirement in LSOP. We establish this rank bound from a novel perspective, specifically that of analyzing the rank of various faces in the convex hull of the domain set (i.e., $\text{conv}(\mathcal{D})$). Besides, the rank bound gives a sufficient condition under which the LSOP-R exactly solves the original LSOP.

We conclude Section 2 by revisiting the three examples with matrix space $\mathcal{Q} := \mathcal{S}_+^n$ in Subsection 1.1 and deriving the rank bounds for their corresponding LSOP-Rs.

(ii) Section 3 explores the non-symmetric matrix space, i.e., $\mathcal{Q} = \mathbb{R}^{n\times p}$. Advancing the analysis in Section 2, we show that the LSOP-R admits the same rank bound as the one with $\mathcal{Q} := \mathcal{S}_+^n$.

In Subsection 3.1, we present the non-symmetric applications: Fair SVD and Matrix Completion, and derive the first-known rank bounds for their LSOP-Rs.

(iii) Section 4 discusses the symmetric indefinite matrix space, i.e., $\mathcal{Q} := \mathcal{S}^n$ under two cases depending on whether the function $f(\cdot)$ in (1) is sign-invariant with eigenvalues or not.

In Subsection 4.1, the sign-invariant property allows for an identical rank bound of LSOP-R to the positive semidefinite matrix space.

Beyond sign-invariance in Subsection 4.2, the convexification of domain set $\mathcal{D}$ involves the convex hull of a union of multiple convex sets, which inspires us a tighter relaxation than LSOP-R by replacing $\text{conv}(\mathcal{D})$ with the union set, further termed LSOP-R-I. Then, we derive a rank bound of the order of $\mathcal{O}(\sqrt{4\tilde{m}})$ for the LSOP-R-I.

Finally, we extend the analysis to derive sparsity bounds in the diagonal matrix space and apply the result to Sparse Ridge Regression.

(iv) In Section 5, we develop an efficient column generation algorithm with a vector-based convex pricing oracle to solve LSOP-R, where the rank of the output solution can be reduced to be no larger than the theoretical bound using a rank-reduction algorithm. In Section 6, we numerically test the proposed algorithms.

The detailed rank bounds for LSOP-R with different matrix spaces and those for application examples can be found in Table 1. Note that we further tighten the rank bounds using $\tilde{k}$ in Definition 1. We also show that the rank bounds marked with asterisk in Table 1 are tight.

Notation and Definition. We use bold lower-case letters (e.g., $\boldsymbol{x}$) and bold upper-case letters (e.g., $\boldsymbol{X}$) to denote vectors and matrices, respectively, and use corresponding non-bold letters (e.g., x_i) to denote their components. We let $[n] := \{1, 2, \cdots, n\}$. We let $\boldsymbol{I}_n$ denote the $n \times n$ identity matrix. For any $\lambda \in \mathbb{R}$, we let $\lfloor \lambda \rfloor$ denote the greatest integer less than or equal to λ, let $(\lambda)_+ := \max\{0, \lambda\}$, and let $\text{sign}(\lambda)$ be 1 if $\lambda \geq 0$, otherwise, -1. For a set S, we let $|S|$ denote its cardinality. For a vector $\boldsymbol{x} \in \mathbb{R}^n$, we let $|\boldsymbol{x}| := (|x_1|, \dots, |x_n|)^\top$ contain the absolute entries of $\boldsymbol{x}$, let $\|\boldsymbol{x}\|_0$ denote its zero norm, let $\|\boldsymbol{x}\|_1$ denote its one norm, let $\|\boldsymbol{x}\|_2$ denote its two norm, and let $\text{Diag}(\boldsymbol{x})$ denote a diagonal matrix with diagonal entries from $\boldsymbol{x}$. For a matrix $\boldsymbol{X} \in \mathcal{Q}$, we let $\|\boldsymbol{X}\|_2$ be its largest singular value, let $\|\boldsymbol{X}\|_*$ be its nuclear norm, let $\|\boldsymbol{X}\|_F$ be its Frobenius norm, and if matrix $\boldsymbol{X}$ is square, let $\boldsymbol{X}^\dagger$ be its Moore–Penrose inverse, let $\text{tr}(\boldsymbol{X})$ be the trace, let $\text{diag}(\boldsymbol{X})$ be a vector including its diagonal elements, and let $\lambda_{\min}(\boldsymbol{X}), \lambda_{\max}(\boldsymbol{X})$ denote the smallest and largest eigenvalue of matrix $\boldsymbol{X}$, respectively. For a matrix $\boldsymbol{X} \in \mathcal{Q}$ and an integer $i \in [n]$, we let $\|\boldsymbol{X}\|_{(i)}$ denote the sum of first i largest singular values of matrix $\boldsymbol{X}$. Note that $\|\boldsymbol{X}\|_{(n)} = \text{tr}(\boldsymbol{X})$ if matrix $\boldsymbol{X}$ is positive semidefinite. For a set T over $(\boldsymbol{X}, \boldsymbol{x})$, we let $\text{proj}_{\boldsymbol{X}}(T) := \{\boldsymbol{X} : \exists \boldsymbol{x}, (\boldsymbol{x}, \boldsymbol{X}) \in T\}$ denote the projection of set T

Table 1 Summary of Our Rank Bounds

LSOP Relaxation	**Matrix Space $\mathcal{Q}$**	**Spectral Constraint**	**Rank Bound**
LSOP-R	$\mathcal{S}_+^n$	Any	$\tilde{k}+\lfloor\sqrt{2\tilde{m}+9/4}-3/2\rfloor$ (Theorem 2)*
LSOP-R	$\mathbb{R}^{n\times p}$	Any	$\tilde{k}+\lfloor\sqrt{2\tilde{m}+9/4}-3/2\rfloor$ (Theorem 4)*
LSOP-R	$\mathcal{S}^n$	Sign-invariant	$\tilde{k}+\lfloor\sqrt{2\tilde{m}+9/4}-3/2\rfloor$ (Theorem 5)*
LSOP-R-I	$\mathcal{S}^n$	Any	$\tilde{k}+\lfloor\sqrt{4\tilde{m}+9}\rfloor-3$ (Theorem 7)
LSOP-R-I	Diagonal $\mathcal{S}^n$	Any	$\tilde{k}+\tilde{m}$ (Theorem 9)*
LSOP-R Examples	**Matrix Space $\mathcal{Q}$**	**Spectral Constraint**	**Rank Bound**
Kernel Learning-R	$\mathcal{S}_+^n$	$z-\log\det(\boldsymbol{X}+\alpha\boldsymbol{I}_n)\le 0$	$k+\lfloor\sqrt{2\tilde{m}+9/4}-3/2\rfloor$ (Corollary 1)
Fair PCA-R	$\mathcal{S}_+^n$	$\lVert\boldsymbol{X}\rVert_2-1\le 0$	$k+\lfloor\sqrt{2m+1/4}-3/2\rfloor$ (Corollary 2)
QCQP-R	$\mathcal{S}_+^n$	N/A	$1+\lfloor\sqrt{2\tilde{m}+1/4}-3/2\rfloor$ (Corollary 3)
Fair SVD-R	$\mathbb{R}^{n\times p}$	$\lVert\boldsymbol{X}\rVert_2-1\le 0$	$k+\lfloor\sqrt{2m+1/4}-3/2\rfloor$ (Corollary 4)
Matrix Completion-R	$\mathbb{R}^{n\times p}$	$\lVert\boldsymbol{X}\rVert_*-z\le 0$	$\lfloor\sqrt{2\lvert\Omega\rvert+9/4}-1/2\rfloor$ (Corollary 5)
Sparse Ridge Regression-R	Diagonal $\mathcal{S}^n$	$\lVert\boldsymbol{X}\rVert_F^2-z\le 0$	$k+\mathrm{rank}(\boldsymbol{A})$ (Corollary 6)

into element $\boldsymbol{X}$. A function $f(\boldsymbol{\lambda}):\mathbb{R}^n\to\mathbb{R}$ is *symmetric* if it is invariant with any permutation of $\boldsymbol{\lambda}$ and is *sign-invariant* if $f(\boldsymbol{\lambda})=f(|\boldsymbol{\lambda}|)$. Additional notation will be introduced later as needed.

2. The Rank Bound in Positive Semidefinite Matrix Space

To guarantee the solution quality of the partial convexification LSOP-R, this section derives a rank bound for all extreme points and for an optimal solution to LSOP-R, provided that $\mathcal{Q}:=\mathcal{S}_+^n$ denotes the positive semidefinite matrix space. Notably, our results recover the existing rank bounds for two special LSOPs: QCQP and Fair PCA, and are also applicable to Kernel Learning.

2.1. Convexifying Domain Set

This subsection provides an explicit characterization of the convex hull of the domain set $\mathcal{D}$, i.e., $\mathrm{conv}(\mathcal{D})$, which is a key component in the feasible set of LSOP-R,

Before deriving $\mathrm{conv}(\mathcal{D})$, let us define an integer $\tilde{k}\le k$ as the strengthened rank. When necessary in the theoretical analysis of rank bounds, we replace the rank requirement k with the smaller $\tilde{k}$.

Definition 1 (Strengthened Rank $\tilde{k}$) *For a domain set $\mathcal{D}$ in (1), we let $1 \leq \tilde{k} \leq k$ be the smallest integer such that*

$$\operatorname{conv}(\mathcal{D}) := \operatorname{conv}\left(\left\{\boldsymbol{X} \in \mathcal{S}_+^n : \operatorname{rank}(\boldsymbol{X}) \leq \tilde{k}, F(\boldsymbol{X}) := f(\boldsymbol{\lambda}(\boldsymbol{X})) \leq 0\right\}\right).$$

The following example confirms that $\tilde{k}$ can be indeed strictly less than k.

Example 1 Suppose a domain set $\mathcal{D} := \{\boldsymbol{X} \in \mathcal{Q} : \operatorname{rank}(\boldsymbol{X}) \leq k, f(\|\boldsymbol{X}\|_*) \leq 0\}$ with $k \geq 2$ and $f(\cdot)$ is a closed convex bounded function. Then sets $\mathcal{D}$ and $\operatorname{conv}(\mathcal{D})$ are bounded. In this example, we show below that $\tilde{k} = 1 < k$.

For any matrix $\widehat{\boldsymbol{X}}$ in domain set $\mathcal{D}$ with rank $r \geq 2$, let its singular value decomposition be $\widehat{\boldsymbol{X}} = \boldsymbol{Q} \operatorname{Diag}(\widehat{\boldsymbol{\lambda}}) \boldsymbol{P}^\top$. Next, let us construct r vectors $\{\boldsymbol{\lambda}^i\}_{i \in [r]} \subseteq \mathbb{R}_+^n$ as below, where for each $i \in [r]$,

$$\lambda_\ell^i = \begin{cases} \|\widehat{\boldsymbol{\lambda}}\|_1, & \text{if } \ell = i; \\ 0, & \text{if } \ell \in [n] \setminus \{i\}. \end{cases}.$$

Then, for each $i \in [r]$, we have that $\|\boldsymbol{\lambda}^i\|_0 = 1 \leq k$ and $f(\|\boldsymbol{\lambda}^i\|_1) = f(\|\widehat{\boldsymbol{\lambda}}\|_1) \leq 0$, which means that the inclusion $\boldsymbol{Q} \operatorname{Diag}(\boldsymbol{\lambda}^i) \boldsymbol{P}^\top \in \mathcal{D}$ holds for all $i \in [r]$. Also, we have $\widehat{\boldsymbol{X}} = \sum_{i \in [r]} \frac{\widehat{\lambda}_i}{\|\widehat{\boldsymbol{\lambda}}\|_1} \boldsymbol{Q} \operatorname{Diag}(\boldsymbol{\lambda}^i) \boldsymbol{P}^\top$ by the construction of vector $\{\boldsymbol{\lambda}^i\}_{i \in [r]} \subseteq \mathbb{R}_+^n$, implying that $\widehat{\boldsymbol{X}}$ cannot be an extreme point of the set $\operatorname{conv}(\mathcal{D})$. That is, any extreme point in the bounded set $\operatorname{conv}(\mathcal{D})$ has a rank at most one. Hence, we must have $\tilde{k} = 1$ by Definition 1. ◇

We now turn to the explicit characterization of $\operatorname{conv}(\mathcal{D})$, which builds on the majorization below.

Definition 2 (Majorization) *Given two vectors $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n$, we let $\boldsymbol{x}_1 \succ \boldsymbol{x}_2$ denote that $\boldsymbol{x}_1$ weakly majorizes $\boldsymbol{x}_2$ (i.e., $\max_{\boldsymbol{z} \in \{0,1\}^n}\{\boldsymbol{z}^\top \boldsymbol{x}_1 : \mathbf{e}^\top \boldsymbol{z} = \ell\} \geq \max_{\boldsymbol{z} \in \{0,1\}^n}\{\boldsymbol{z}^\top \boldsymbol{x}_2 : \mathbf{e}^\top \boldsymbol{z} = \ell\}$ for all $\ell \in [n]$), and let $\boldsymbol{x}_1 \succeq \boldsymbol{x}_2$ denote that $\boldsymbol{x}_1$ majorizes $\boldsymbol{x}_2$, i.e., $\boldsymbol{x}_1$ weakly majorizes $\boldsymbol{x}_2$ and $\boldsymbol{e}^\top \boldsymbol{x}_1 = \boldsymbol{e}^\top \boldsymbol{x}_2$, where $\mathbf{e} \in \mathbb{R}^n$ denotes all-ones vector.*

The spectral function $F(\boldsymbol{X}) := f(\boldsymbol{\lambda}(\boldsymbol{X}))$ in (1) is invariant with any permutation of eigenvalues (see, e.g., Drusvyatskiy and Kempton 2015), implying a permutation-invariant domain set $\mathcal{D}$. This motivates us to apply the convexification result for a permutation-invariant set, as shown in the seminal work by Kim et al. (2022).

Proposition 1 *For a domain set $\mathcal{D}$ in positive semidefinite matrix space, i.e., $\mathcal{Q} := \mathcal{S}_+^n$ in (1), its convex hull $\operatorname{conv}(\mathcal{D})$ is equal to*

$$\left\{\boldsymbol{X} \in \mathcal{Q} : \exists \boldsymbol{x} \in \mathbb{R}_+^n, f(\boldsymbol{x}) \leq 0, x_1 \geq \cdots \geq x_n, x_{k+1} = 0, \|\boldsymbol{X}\|_{(\ell)} \leq \sum_{i \in [\ell]} x_i, \forall \ell \in [k], \operatorname{tr}(\boldsymbol{X}) = \sum_{i \in [k]} x_i\right\},$$

and is a closed set.

Proof. See Appendix A.1. □

The function $\|\boldsymbol{X}\|_{(i)}$ in Proposition 1 is known to admit a tractable semidefinite representation for each $i \in [n]$ (see, e.g., Ben-Tal and Nemirovski 2001) .

2.2. Rank Bound

In this subsection, we derive an upper bound of the ranks of all extreme points in the feasible set of LSOP-R, which also sheds light on the rank gap between the relaxation LSOP-R and the original LSOP at optimality. Let us first introduce a key lemma to facilitate the analysis of rank bound.

Lemma 1 *Given two vectors* $\boldsymbol{\lambda}, \boldsymbol{x} \in \mathbb{R}^n$ *with their elements sorted in descending order and* $\boldsymbol{x} \succeq \boldsymbol{\lambda}$, *suppose that there exists an index* $j_1 \in [n-1]$ *such that* $\sum_{i\in[j_1]} \lambda_i < \sum_{i\in[j_1]} x_i$. *Then we have*

$$\sum_{i\in[j]} \lambda_i < \sum_{i\in[j]} x_i, \forall j \in [j_0, j_2 - 1],$$

where the indices j_0, j_2 *satisfy* $\lambda_{j_0} = \cdots = \lambda_{j_1} \geq \lambda_{j_1+1} = \cdots = \lambda_{j_2}$ *with* $1 \leq j_0 \leq j_1 \leq j_2 - 1 \leq n-1$.

Proof. See Appendix A.2. □

Next, we are ready to prove one of our main contributions that relies on the inequalities over $[j_0, j_1]$ in Lemma 1. Specifically, we show that the rank bound of a face in $\text{conv}(\mathcal{D})$ depends on the dimension of this face. Below is the formal definition of faces and dimension of a closed convex set.

Definition 3 (Face & Dimension) *A convex subset* F *of a closed convex set* $\mathcal{T}$ *is called a face of* $\mathcal{T}$ *if for any line segment* $[a, b] \subseteq \mathcal{T}$ *such that* $F \cap (a, b) \neq \emptyset$, *we have* $[a, b] \subseteq F$. *The dimension of a face is equal to the dimension of its affine hull.*

Theorem 1 *Given* $\mathcal{Q} := \mathcal{S}_+^n$ *in domain set* $\mathcal{D}$ *in* (1), *suppose that* $F^d \subseteq \text{conv}(\mathcal{D})$ *is a* d*-dimensional face of the convex hull of the domain set. Then any point in face* F^d *has a rank at most* $\tilde{k} + \lfloor\sqrt{2d + 9/4} - 3/2\rfloor$, *where* $\tilde{k} \leq k$ *follows Definition 1.*

Proof. See Appendix A.3. □

Theorem 1 suggests that the dimension of a face in set $\text{conv}(\mathcal{D})$ determines the rank bound. When intersecting set $\text{conv}(\mathcal{D})$ with m linear inequalities in LSOP-R, the recent work by Li and Xie (2022) established the one-to-one correspondence between the extreme points of the intersection set and no larger than m-dimensional faces of $\text{conv}(\mathcal{D})$ (see lemma 1 therein).

Lemma 2 (Li and Xie 2022) *For any closed convex set* $\mathcal{T}$, *if* $\boldsymbol{X}$ *is an extreme point in the intersection of set* $\mathcal{T}$ *and* $\tilde{m}$ *linearly independent inequalities, then* $\boldsymbol{X}$ *must be contained in a no larger than* $\tilde{m}$*-dimensional face of set* $\mathcal{T}$.

Taken these results together, we next derive a rank bound for LSOP-R.

Theorem 2 *Given* $\mathcal{Q} := \mathcal{S}_+^n$ *in* (1) *and integer* $\tilde{k} \leq k$ *following Definition 1, we have* *(i) Each feasible extreme point in the LSOP-R has a rank at most $\tilde{k}+\lfloor\sqrt{2\tilde{m}+9/4}-3/2\rfloor$; and*

(ii) There is an optimal solution to the LSOP-R of rank at most $\tilde{k}+\lfloor\sqrt{2\tilde{m}+9/4}-3/2\rfloor$ if the LSOP-R yields a finite optimal value, i.e., $\mathbf{V}_{\rm rel}>-\infty$.

Proof. The proof can be divided into two parts.

Part (i). Since the LSOP-R feasible set intersects the closed convex set $\text{conv}(\mathcal{D})$ with $\tilde{m}$ linearly independent inequalities, according to Lemma 2, each extreme point $\widehat{\boldsymbol{X}}$ in the intersection belongs to a no larger than $\tilde{m}$-dimensional face in set $\text{conv}(\mathcal{D})$. Using Theorem 1, the face containing $\widehat{\boldsymbol{X}}$ must satisfy the rank-($\tilde{k}+\lfloor\sqrt{2\tilde{m}+9/4}-3/2\rfloor$) constraint and so does the extreme point $\widehat{\boldsymbol{X}}$.

Part (ii). Since the LSOP-R admits a finite optimal value, i.e., $\mathbf{V}_{\rm rel}>-\infty$ and a line-free feasible set, according to Rockafellar (1972), the LSOP-R can achieve extreme point optimum. Thus, according to Part (i), there exists an optimal extreme point satisfying the desired rank bound. ☐

We make the following remarks about Theorem 2.

(i) Theorem 2 shows that any extreme point in the feasible set of LSOP-R, as a convex relaxation of LSOP, violates the rank-k constraint by at most $(\tilde{k}-k+\lfloor\sqrt{2\tilde{m}+9/4}-3/2\rfloor)_+$.

(ii) Given $\mathcal{Q}:=\mathcal{S}_+^n$ in (1), the most striking aspect is that the rank bound of Theorem 2 is independent of any domain set $\mathcal{D}$, any linear objective function, and any $\tilde{m}$ linearly independent inequalities in LSOP-R. We present three examples in Subsection 2.3 to demonstrate the versatility of our rank bound;

(iii) From a new angle, the proposed rank bound for LSOP-R arises from bounding the rank of various faces in set $\text{conv}(\mathcal{D})$, as shown in Theorem 1 through perturbing majorization constraints. As a result, the derivation of our rank bounds in Theorem 2 differs from those for two specific LSOP-Rs of QCQP (Barvinok 1995, Deza et al. 1997, Pataki 1998) and Fair PCA (Tantipongpipat et al. 2019). In particular, Pataki (1998) derived the rank bound from analyzing faces of the whole feasible set of LSOP-R of QCQP, whereas our result, inspired by Theorem 1, only focuses on the faces of a set $\text{conv}(\mathcal{D})$;

(iv) The worse-case example in Appendix B can attain the proposed rank bound in Theorem 2, which confirms the tightness; and

(v) An extra benefit of Theorem 2 is to provide a sufficient condition about when LSOP-R is equivalent to the original LSOP, as shown below.

Proposition 2 *Given $\mathcal{Q}:=\mathcal{S}_+^n$ in (1) and integer $\tilde{k}\leq k$ following Definition 1, if $\tilde{m}\leq(k-\tilde{k}+2)(k-\tilde{k}+3)/2-2$ holds, we have that*

(i) Each feasible extreme point in the LSOP-R has a rank at most k; and

(ii) The LSOP-R achieves the same optimal value as the original LSOP, i.e., $\mathbf{V}_{\rm opt}=\mathbf{V}_{\rm rel}$ if the LSOP-R yields a finite optimal value, i.e., $\mathbf{V}_{\rm rel}>-\infty$.

Proof. When $\tilde{m} \le (k-\tilde{k}+2)(k-\tilde{k}+3)/2-2$, the inequality $\tilde{k}+\lfloor\sqrt{2\tilde{m}+9/4}-3/2\rfloor \le k$ holds. Then, using the result in Theorem 2, the rank of each extreme point in the LSOP-R feasible set is no larger than k, and the LSOP-R has an optimal extreme point $\boldsymbol{X}^*$ of rank at most k. Given $-\infty < \mathbf{V}_{\text{rel}} \le \mathbf{V}_{\text{opt}}$, the feasible $\boldsymbol{X}^*$ must be also optimal to LSOP. Thus completes the proof. □

Proposition 2 contributes to the literature of the LSOP-R objective exactness that refers to the equation $\mathbf{V}_{\text{opt}} = \mathbf{V}_{\text{rel}}$, as it indicates that for any LSOP, the corresponding LSOP-R always achieves the same optimal value whenever $\tilde{m} \le 1$ and $\mathbf{V}_{\text{rel}} > -\infty$.

2.3. Applying to Kernel Learning, Fair PCA, and QCQP

This subsection revisits three LSOP examples in positive semidefinite matrix space. More specifically, our proposed rank bound in Theorem 2 recovers the existing results for Fair PCA and QCQP from a different perspective, and can be applied to Kernel Learning, providing the first-known rank bound of its LSOP-R counterpart.

Kernel Learning. First, recall that in the Kernel Learning, its domain set $\mathcal{D}$ is defined for any given value of variable z, i.e., $\mathcal{D} := \{\boldsymbol{X} \in \mathcal{S}_+^n : \text{rank}(\boldsymbol{X}) \le k, \log\det(\boldsymbol{X}+\alpha\boldsymbol{I}) \ge z\}$, which, according to Proposition 1, has an explicit representation of the convex hull $\text{conv}(\mathcal{D})$. Thus, the LSOP-R corresponding to Kernel Learning can be written as

$$\mathbf{V}_{\text{rel}} := \min_{\boldsymbol{X}\in\text{conv}(\mathcal{D}), z\in\mathbb{R}} \left\{\langle \boldsymbol{Y}^\dagger, \boldsymbol{X}\rangle - z + \log\det(\boldsymbol{Y}+\alpha\boldsymbol{I}) : b_i^l \le \langle \boldsymbol{A}_i, \boldsymbol{X}\rangle \le b_i^u, \forall i \in [m]\right\}, \tag{Kernel Learning-R}$$

where $\text{conv}(\mathcal{D}) := \{\boldsymbol{X} \in \mathcal{S}_+^n : \exists \boldsymbol{x} \in \mathbb{R}_+^n, \sum_{i\in[n]} \log(\alpha + x_i) \ge z, x_1 \ge \cdots \ge x_n, x_{k+1} = 0, \|\boldsymbol{X}\|_{(\ell)} \le \sum_{i\in[\ell]} x_i, \forall \ell \in [k], \text{tr}(\boldsymbol{X}) = \sum_{i\in[k]} x_i\}$.

Note that the existing works on rank bounds fail to cover Kernel Learning-R (see, e.g., Barvinok 1995, Pataki 1998, Tantipongpipat et al. 2019). Our results in Theorem 2 and Proposition 2 can fill this gap since by exploring the formulation structure, the Kernel Learning-R can be viewed as a special case of the LSOP-R.

Corollary 1 (Kernel Learning) *There exists an optimal solution of the Kernel Learning-R with rank at most $k+\left\lfloor\sqrt{2\tilde{m}+9/4}-3/2\right\rfloor$. Besides, if there is $\tilde{m} \le 1$ linearly independent inequality, the Kernel Learning-R achieves the same optimal value as Kernel Learning.*

Proof. Suppose that $(\widehat{\boldsymbol{X}}, \widehat{z})$ is an optimal solution to the Kernel Learning-R. Then for a given solution $\widehat{z}$, all the points in the following set are also optimal to Kernel Learning-R

$$\left\{\boldsymbol{X} \in \text{conv}(\mathcal{D}) : \langle \boldsymbol{A}_i, \boldsymbol{X}\rangle = \langle \boldsymbol{A}_i, \widehat{\boldsymbol{X}}\rangle, \forall i \in [m]\right\},$$

where $\mathcal{D} := \{\boldsymbol{X} \in \mathcal{S}_+^n : \text{rank}(\boldsymbol{X}) \le k, \log\det(\boldsymbol{X}+\alpha\boldsymbol{I}) \ge \widehat{z}\}$ is based on $\widehat{z}$. As $\tilde{k} \le k$ and the domain set $\mathcal{D}$ is bounded, according to Theorem 2, all the extreme points in the optimal set above have a rank at most $k+\lfloor\sqrt{2\tilde{m}+9/4}-3/2\rfloor$. Thus, there is an optimal solution $(\boldsymbol{X}^*,\widehat{z})$ to the Kernel Learning-R with $\boldsymbol{X}^*$ satisfying the desired rank bound. Finally, the rank bound does not exceed k when $\tilde{m}\leq 1$, implying that $(\boldsymbol{X}^*,\widehat{z})$ is also optimal to the original Kernel Learning. □

Fair PCA. Another application of Theorem 2 is to provide a rank bound for partial convexification LSOP-R of the Fair PCA defined as follows

$$\begin{aligned}\mathbf{V}_{\text{rel}} := &\max_{z\in\mathbb{R},\boldsymbol{X}\in\text{conv}(\mathcal{D})}\{z: z\leq\langle\boldsymbol{A}_i,\boldsymbol{X}\rangle,\forall i\in[m]\},\\ &\text{conv}(\mathcal{D})=\{\boldsymbol{X}\in\mathcal{S}_+^n:\text{tr}(\boldsymbol{X})\leq k,||\boldsymbol{X}||_2\leq 1\}.\end{aligned} \tag{Fair PCA-R}$$

Note that the Fair PCA is equipped with a domain set $\mathcal{D}:=\{\boldsymbol{X}\in\mathcal{S}_+^n:\text{rank}(\boldsymbol{X})\leq k,\|\boldsymbol{X}\|_2\leq 1\}$, whose convex hull can be readily derived without the majorization technique, as shown in Fair PCA-R. The rank bound below recovers the one in Tantipongpipat et al. (2019).

Corollary 2 (Fair PCA) *There exists an optimal solution of the Fair PCA-R with rank at most* $k+\left\lfloor\sqrt{2m+1/4}-3/2\right\rfloor$. *Besides, if there are* $m\leq 2$ *sample covariance matrices, the Fair PCA-R achieves the same optimal value as Fair PCA.*

Proof. See Appendix A.4. □

We close this subsection by discussing a special domain set $\mathcal{D}:=\{\boldsymbol{X}\in\mathcal{S}_+^n:\text{rank}(\boldsymbol{X})\leq 1\}$, which can be naturally adopted in QCQP. The LSOP-R under this setting reduces to a semidefinite program relaxation of QCQP:

$$\mathbf{V}_{\text{rel}} := \min_{\boldsymbol{X}\in\text{conv}(\mathcal{D})}\left\{\langle\boldsymbol{A}_0,\boldsymbol{X}\rangle: b_i^l\leq\langle\boldsymbol{A}_i,\boldsymbol{X}\rangle\leq b_i^u,\forall i\in[m]\right\},\quad \text{conv}(\mathcal{D}):=\mathcal{S}_+^n. \tag{QCQP-R}$$

We show that Theorem 2 can recover a well-known rank bound of QCQP-R (see, e.g., Pataki 1998).

Corollary 3 (QCQP) *There exists an optimal solution of QCQP-R with rank at most* $1+\lfloor 2\tilde{m}+1/4-3/2\rfloor$ *if* $\mathbf{V}_{\text{rel}}>-\infty$. *Besides, if there are* $\tilde{m}\leq 2$ *linearly independent inequalities, the QCQP-R achieves the same optimal value as the original QCQP if* $\mathbf{V}_{\text{rel}}>-\infty$.

Proof. See Appendix A.5. □

Corollary 3 implies that the QCQP-R coincides with the original QCQP when there are $\tilde{m}\leq 2$ linearly independent constraints, which is consistent with those of previous studies on the objective exactness of QCQP-R (see Burer and Ye 2020, Kılınç-Karzan and Wang 2021, Li and Xie 2022).

3. The Rank Bound in Non-symmetric Matrix Space

This section derives a rank bound of LSOP-R in non-symmetric matrix space, i.e., $\mathcal{Q}=\mathbb{R}^{n\times p}$ in (1). Due to the non-symmetry, the analysis relies on the singular value decomposition in this section. More importantly, the nonnegative singular values enable us to rewrite function $f(\cdot)$ in (1) as

$f(\boldsymbol{\lambda}(\boldsymbol{X})) := f(|\boldsymbol{\lambda}(\boldsymbol{X})|)$, which admits the sign-invariant property. Therefore, when $\mathcal{Q} = \mathbb{R}^{n\times p}$, the domain set $\mathcal{D}$ in (1) is permutation- and sign-invariant with singular values, and its convex hull $\operatorname{conv}(\mathcal{D})$ can be readily described according to Kim et al. (2022) (see $\operatorname{conv}(\mathcal{D})$ in Appendix A.6).

We derive a rank bound below for any fixed-dimension face in the set $\operatorname{conv}(\mathcal{D})$ with $\mathcal{Q} := \mathbb{R}^{n\times p}$.

Theorem 3 *Given $\mathcal{Q} := \mathbb{R}^{n\times p}$ in domain set $\mathcal{D}$ in (1), suppose that $F^d \subseteq \operatorname{conv}(\mathcal{D})$ is a d-dimensional face in the convex hull of the domain set $\mathcal{D}$. Then any point in face F^d has a rank at most $\tilde{k} + \lfloor\sqrt{2d+9/4} - 3/2\rfloor$, where $\tilde{k} \le k$ follows Definition 1.*

Proof. See Appendix A.6. ☐

In Theorem 3, we guarantee the rank bounds for all faces of set $\operatorname{conv}(\mathcal{D})$ with $\mathcal{Q} := \mathbb{R}^{n\times p}$. Analogous to the link between Theorem 1 and Theorem 2, the statement in Lemma 2 still holds for $\mathcal{Q} := \mathbb{R}^{n\times p}$. Hence, Theorem 3 directly implies a rank bound for LSOP-R.

Theorem 4 *Given $\mathcal{Q} := \mathbb{R}^{n\times p}$ in (1) and integer $\tilde{k} \le k$ following Definition 1, suppose that the LSOP-R admits a line-free feasible set, then we have that*

(i) Each feasible extreme point in the LSOP-R has a rank at most $\tilde{k} + \lfloor\sqrt{2\tilde{m}+9/4} - 3/2\rfloor$; and

(ii) There is an optimal solution to LSOP-R of rank at most $\tilde{k} + \lfloor\sqrt{2\tilde{m}+9/4} - 3/2\rfloor$ if the LSOP-R yields a finite optimal value, i.e., $\mathbf{V}_{\text{rel}} > -\infty$.

We make the several remarks about the rank bound in Theorem 4 with $\mathcal{Q} = \mathbb{R}^{n\times p}$.

(a) The premise of a line-free feasible set ensures the existence of extreme points in the feasible set of LSOP-R (Rockafellar 1972);

(b) When $\mathcal{Q} := \mathcal{S}_+^n$ or $\mathcal{Q} := \mathbb{R}^{n\times p}$, both Theorems 1 and 3 are based on analyzing the perturbation of majorization constraints in a permutation-invariant set $\operatorname{conv}(\mathcal{D})$, which leads to the same rank bounds in Theorems 2 and 4. This motivates us to explore another type of permutation-invariant set beyond $\operatorname{conv}(\mathcal{D})$ to derive the rank bounds in symmetric yet indefinite matrix space in Section 4;

(c) The rank bound in Theorem 4 is tight (see Appendix B); and

(d) A notable side product of Theorem 4 is that by letting $\tilde{k} + \lfloor\sqrt{2m+9/4} - 3/2\rfloor \le k$, Theorem 4 reduces to a sufficient condition of the exactness of LSOP-R.

Proposition 3 *Given $\mathcal{Q} := \mathbb{R}^{n\times p}$ in (1) and integer $\tilde{k} \le k$ following Definition 1, if $\tilde{m} \le (k-\tilde{k}+2)(k-\tilde{k}+3)/2-2$ holds, then we have that*

(i) Each feasible extreme point in the LSOP-R has a rank at most k; and

(ii) The LSOP-R achieves the same optimal value as the original LSOP, i.e., $\mathbf{V}_{\text{opt}} = \mathbf{V}_{\text{rel}}$ if the LSOP-R yields a finite optimal value, i.e., $\mathbf{V}_{\text{rel}} > -\infty$ and admits a line-free feasible set.

Proof. Using Theorem 4, the proof is identical to that of Proposition 2 and thus is omitted. ☐

3.1. Applying to Fair SVD and Matrix Completion

This subsection investigates two LSOP application examples with $\mathcal{Q} := \mathbb{R}^{n\times p}$ in Subsection 1.1 and provides the first-known rank bounds for their corresponding LSOP-Rs by leveraging Theorem 4.

Fair SVD. The domain set in Fair SVD is $\mathcal{D} := \{\boldsymbol{X} \in \mathbb{R}^{n\times p} : \mathrm{rank}(\boldsymbol{X}) \le 1, \|\boldsymbol{X}\|_2 \le 1\}$. Therefore, for Fair SVD, its LSOP-R can be formulated below

$$\begin{aligned} \mathbf{V}_{\mathrm{rel}} := \max_{z\in\mathbb{R}, \boldsymbol{X}\in\mathrm{conv}(\mathcal{D})} \left\{ z : z \le \langle \boldsymbol{A}_i, \boldsymbol{X}\rangle, \forall i \in [m] \right\}, \\ \mathrm{conv}(\mathcal{D}) = \{\boldsymbol{X} \in \mathbb{R}^{n\times p} : \|\boldsymbol{X}\|_* \le k, ||\boldsymbol{X}||_2 \le 1\}. \end{aligned} \tag{Fair SVD-R}$$

Note that in Fair SVD-R, we have that $\tilde{k} = k$ and the domain set $\mathcal{D}$ is compact, which enables us to directly apply Theorem 4.

Corollary 4 (Fair SVD) *There exists an optimal solution of Fair SVD-R with rank at most* $k + \lfloor \sqrt{2m + 1/4} - 3/2 \rfloor$. *Besides, if there are* $m \le 2$ *groups of data matrices, the Fair SVD-R yields the same optimal value as the original Fair SVD.*

Proof. See Appendix A.7. □

Matrix Completion. Matrix Completion that performs low-rank recovery from observed samples is a popular technique in machine learning, which is known to be notoriously NP-hard. Alternatively, we study the following convex relaxation of Matrix Completion:

$$\mathbf{V}_{\mathrm{rel}} := \min_{\boldsymbol{X}\in\mathrm{conv}(\mathcal{D}), z\in\mathbb{R}_+} \left\{ z : X_{ij} = A_{ij}, \forall (i,j) \in \Omega \right\}, \quad \mathrm{conv}(\mathcal{D}) := \{\boldsymbol{X} \in \mathbb{R}^{n\times p} : \|\boldsymbol{X}\|_* \le z\}, \tag{Matrix Completion-R}$$

which is equivalent to $\min_{\boldsymbol{X}\in\mathbb{R}^{n\times p}} \left\{ \|\boldsymbol{X}\|_* : X_{ij} = A_{ij}, \forall (i,j) \in \Omega \right\}$. Despite the popularity of the convex relaxation– Matrix Completion-R in the literature (see, e.g., Cai et al. 2010, Miao and Kou 2021), there is no guaranteed rank bound of its solution so far. We fill this gap by deriving the following rank bound for the Matrix Completion-R.

Corollary 5 (Matrix Completion) *Suppose that there are* $m := |\Omega|$ *observed entries in Matrix Completion. Then there exists an optimal solution to Matrix Completion-R with rank at most* $\lfloor \sqrt{2|\Omega| + 9/4} - 1/2 \rfloor$. *Besides, if there are* $|\Omega| \le (k+1)(k+2)/2 - 2$ *observed entries, the Matrix Completion-R can achieve the same optimal value as the original Matrix Completion.*

Proof. See Appendix A.8. □

4. Rank Bounds in Symmetric Indefinite Matrix Space

This section focuses on the symmetric indefinite matrix space, i.e., $\mathcal{Q} := \mathcal{S}^n$ and its special case– diagonal matrix space which can be applied to Sparse Ridge Regression. So far, scant attention has been paid to the LSOP of $\mathcal{Q} := \mathcal{S}^n$ in literature. In contrast to a matrix in $\mathcal{Q} := \mathcal{S}_+^n$ or $\mathcal{Q} := \mathbb{R}^{n\times p}$ that formulates the spectral and rank constraints by the nonnegative eigenvalues or singular values, respectively, a symmetric indefinite matrix may have both negative and positive eigenvalues. This requires a different analysis of rank bounds. Specifically, we discuss two cases of $\mathcal{Q} := \mathcal{S}^n$ depending on whether the function $f(\cdot)$ in (1) is sign-invariant or not.

4.1. The Rank Bound with Sign-Invariance

In this subsection, we study the case where the function $f(\cdot)$ in domain set $\mathcal{D}$ in (1) is sign-invariant, i.e., $f(\boldsymbol{\lambda}) = f(|\boldsymbol{\lambda}|)$ for any $\boldsymbol{\lambda} \in \mathbb{R}^n$. In this case, it suffices to focus on the absolute eigenvalues of a symmetric matrix, i.e., singular values. Therefore, the rank bound in Theorem 4 for the non-symmetric matrix space can be extended, as shown below.

Theorem 5 *Given $\mathcal{Q} := \mathcal{S}^n$ in (1) and integer $\tilde{k} \leq k$ following Definition 1, suppose that the function $f(\cdot)$ in (1) is sign-invariant and LSOP-R admits a line-free feasible set, then we have that*

(i) Each feasible extreme point in the LSOP-R has a rank at most $\tilde{k} + \lfloor \sqrt{2\tilde{m} + 9/4} - 3/2 \rfloor$; and

(ii) There is an optimal solution to LSOP-R of rank at most $\tilde{k} + \lfloor \sqrt{2\tilde{m} + 9/4} - 3/2 \rfloor$ if the LSOP-R yields a finite optimal value, i.e., $\mathbf{V}_{\text{rel}} > -\infty$.

Proof. See Appendix A.9. □

For the case of $\mathcal{Q} := \mathcal{S}^n$ with sign-invariant property, we obtain the same rank bound as those in Theorems 2 and 4. Besides, there is an example of LSOP-R that attains this rank bound (see Appendix B). Following the spirit of Proposition 2, Theorem 5 also gives rise to a sufficient condition about when the LSOP-R matches the original LSOP as shown in Appendix A.9.

4.2. Rank Bounds Beyond Sign-Invariance

When the function $f(\cdot)$ in the domain set $\mathcal{D}$ with $\mathcal{Q} := S^n$ in (1) is not sign-invariant, we analyze a tighter tractable relaxation for LSOP than the LSOP-R, termed LSOP-R-I and then derive the rank bound of LSOP-R-I, which can be applied to sparse optimization in the next subsection.

As indicated previously, the positive and negative eigenvalues may simultaneously exist in a symmetric matrix; therefore, even if we have eigenvalues of a symmetric matrix sorted in descending order, the rank constraint cannot be readily dropped without precisely locating zero eigenvalues. This indicates a need of disjunctive description for $\operatorname{conv}(\mathcal{D})$ as detailed below.

Proposition 4 *For a domain set $\mathcal{D}$ with $\mathcal{Q} := \mathcal{S}^n$ and an integer $\tilde{k} \leq k$ in Definition 1, we have*

(i) $\operatorname{conv}(\mathcal{D}) = \operatorname{conv}\left(\cup_{s\in[\tilde{k}+1]}\mathcal{Y}^s\right)$*; and*

(ii) $\mathcal{D} \subseteq \cup_{s\in[\tilde{k}+1]}\mathcal{Y}^s$,

where for each $s \in [\tilde{k}+1]$*, set* $\mathcal{Y}^s$ *is closed and is defined as*

$$\mathcal{Y}^s := \left\{ \boldsymbol{X} \in \mathcal{S}^n : \exists \boldsymbol{x} \in \mathbb{R}^n, f(\boldsymbol{x}) \le 0, x_1 \ge \cdots \ge x_n, x_s = x_{s+n-\tilde{k}-1} = 0, \boldsymbol{x} \succeq \boldsymbol{\lambda}(\boldsymbol{X}) \right\}. \tag{2}$$

Note that $\boldsymbol{\lambda}(\boldsymbol{X}) \in \mathbb{R}^n$ *denotes the eigenvalue vector of the symmetric matrix* $\boldsymbol{X} \in \mathcal{S}^n$*.*

Proof. See Appendix A.10. □

Note that (i) the majorization $\boldsymbol{x} \succeq \boldsymbol{\lambda}(\boldsymbol{X})$ admits a tractable semidefinite representation according to Ben-Tal and Nemirovski (2001) and (ii) the inclusions $\mathcal{D} \subseteq \cup_{s\in[\tilde{k}+1]} \mathcal{Y}^s \subseteq \mathrm{conv}(\mathcal{D})$ in Proposition 4 inspire us a stronger partial relaxation for LSOP with $\mathcal{Q} := \mathcal{S}^n$, i.e.,

$$\mathrm{V}_{\mathrm{rel}} \le \mathrm{V}_{\mathrm{rel\text{-}I}} := \min_{s\in[\tilde{k}+1]} \min_{\boldsymbol{X}\in\mathcal{Y}^s} \left\{ \langle \boldsymbol{A}_0, \boldsymbol{X} \rangle : b_i^l \le \langle \boldsymbol{A}_i, \boldsymbol{X} \rangle \le b_i^u, \forall i \in [m] \right\} \le \mathrm{V}_{\mathrm{opt}}, \tag{LSOP-R-I}$$

where integer $\tilde{k}$ follows Definition 1 and for each $s \in [\tilde{k}+1]$, set $\mathcal{Y}^s$ is defined in (2).

Then, to derive the rank bound for the tighter LSOP-R-I, it is sufficient to study all inner minimization problems over matrix variable $\boldsymbol{X}$. For each $s \in [\tilde{k}+1]$, we see that the set $\mathcal{Y}^s$ in (2) that builds on the majorization constraints is permutation-invariant with eigenvalues of $\boldsymbol{X}$, which motivates us to derive the following rank bounds for faces of set $\mathcal{Y}^s$.

Theorem 6 *Given* $\mathcal{Q} := \mathcal{S}^n$ *in* (1)*, for each* $s \in [\tilde{k}+1]$*, suppose that* $F^d \subseteq \mathcal{Y}^s$ *is a d-dimensional face of the closed convex set* $\mathcal{Y}^s$ *defined in* (2)*, then any point in face* F^d *has rank at most*

$$\begin{cases} \tilde{k} + \lfloor \sqrt{2d + 9/4} \rfloor - 3/2, & \text{if } s \in \{1, \tilde{k}+1\}; \\ \tilde{k} + \lfloor \sqrt{4d+9} \rfloor - 3, & \text{otherwise.} \end{cases},$$

where $\tilde{k} \le k$ *follows Definition 1.*

Proof. See Appendix A.11. □

The results in Theorem 6 lead to the following rank bounds for the relaxation LSOP-R-I.

Theorem 7 *Given* $\mathcal{Q} := \mathcal{S}^n$ *in* (1) *and integer* $\tilde{k} \le k$ *following Definition 1, suppose that the inner minimization of LSOP-R-I always admits a line-free feasible set, then we have that*

(i) For each $s \in [\tilde{k}+1]$*, any feasible extreme point in the sth inner minimization of LSOP-R-I has a rank at most* $\tilde{k} + \lfloor \sqrt{2\tilde{m} + 9/4} - 3/2 \rfloor$ *if* $s \in \{1, \tilde{k}+1\}$*, and* $\tilde{k} + \lfloor \sqrt{4\tilde{m}+9} - 3 \rfloor$*, otherwise;*

(ii) There exists an optimal solution to LSOP-R-I of rank at most $\tilde{k} + \lfloor \sqrt{4\tilde{m}+9} - 3 \rfloor$ *if the LSOP-R-I yields a finite optimal value, i.e.,* $\mathrm{V}_{\mathrm{rel\text{-}I}} > -\infty$*.*

For the rank bound results in Theorem 7, we remark that

(a) When $s \in \{1, \tilde{k}+1\}$, based on the definition of set $\mathcal{Y}^s$ in (2), all the eigenvalues of a feasible extreme point of the sth inner minimization of LSOP-R-I are either nonpositive or nonnegative. Hence, the rank bound recovers the one in Theorem 2;

(b) When $s \in [2, \tilde{k}]$, each inner minimization of LSOP-R-I has the same rank bound. In this setting, we split a symmetric indefinite matrix into two parts with respective to positives and negative eigenvalues; then analyze the rank bounds of one positive definite matrix and one negative definite matrix, respectively; and integrate the rank bounds. As a result, the rank bound gets slightly worse and is in the order of $\mathcal{O}(\sqrt{4\tilde{m}})$ rather than $\mathcal{O}(\sqrt{2\tilde{m}})$; and

(c) The rank bounds in Theorem 7 shed light on LSOP-R-I objective exactness as shown below.

Proposition 5 *Given $\mathcal{Q} := \mathcal{S}^n$ in (1) and integer $\tilde{k} \leq k$ following Definition 1, when $\tilde{m} \leq \frac{(k-\tilde{k}+3)(k-\tilde{k}+5)}{4} - 2$, the LSOP-R-I achieves the same optimal value as the original LSOP if the LSOP-R-I yields a finite optimal value, i.e., $\mathbf{V}_{\text{rel -I}} > -\infty$ and admits a line-free feasible set.*

4.3. Rank Bound in Diagonal Matrix Space and Applying to Sparse Ridge Regression

This subsection studies the partial convexification LSOP-R with respect to the sparse optimization in vector space, which restricts the number of nonzero elements of any feasible solution (see, e.g., Li and Xie 2020). Since a vector can be equivalently converted into a diagonal matrix, we introduce the following special domain set $\mathcal{D}$ adopted in the sparse optimization problem:

$$\mathcal{D} := \{\boldsymbol{X} \in \mathcal{S}^n : \boldsymbol{X} = \operatorname{Diag}(\operatorname{diag}(\boldsymbol{X})), \operatorname{rank}(\boldsymbol{X}) \leq k, F(\boldsymbol{X}) := f(\operatorname{diag}(\boldsymbol{X})) \leq 0\}, \tag{3}$$

where the variable $\boldsymbol{X}$ is a diagonal matrix and function $f(\cdot)$ is continuous, convex, and symmetric. Thus, the rank bound of LSOP-R over the convex hull of the domain set $\mathcal{D}$ in (3) is in fact the sparsity bound. To be consistent, we stick to the rank bound in the below.

The sparse domain set $\mathcal{D}$ in (3) admits a diagonal matrix space, a special case of $\mathcal{Q} := \mathcal{S}^n$, which enables us to extend the rank bounds in Theorem 5 and Theorem 7 depending on whether the function $f(\cdot)$ in (3) is sign-invariant or not, as summarized below. Similarly, given a domain set $\mathcal{D}$ in (3), the analysis of rank bounds of LSOP-R or LSOP-R-I requires an explicit description of the corresponding set $\operatorname{conv}(\mathcal{D})$ or the union of sets $\{\mathcal{Y}^s\}_{s \in [\tilde{k}+1]}$ (see Appendix A.12).

Theorem 8 *Given a sparse domain set $\mathcal{D}$ in (3) and integer $\tilde{k} \leq k$ in Definition 1, suppose that the function $f(\cdot)$ in (3) is sign-invariant and LSOP-R admits a line-free feasible set. Then we have*

(i) Each feasible extreme point in the LSOP-R has a rank at most $\tilde{k} + \tilde{m}$; and

(ii) There is an optimal solution to LSOP-R of rank at most $\tilde{k} + \tilde{m}$ if the LSOP-R yields a finite optimal value, i.e., $\mathbf{V}_{\text{rel}} > -\infty$.

Theorem 9 *Given a sparse domain set $\mathcal{D}$ in (3) and integer $\tilde{k} \leq k$ following Definition 1, suppose that the inner minimization of LSOP-R-I always admits a line-free feasible set. Then we have*

(i) For any $s \in [\tilde{k}+1]$, each feasible extreme point in the sth inner minimization of LSOP-R-I has a rank at most $\tilde{k} + \tilde{m}$; and

(ii) There is an optimal solution to LSOP-R-I of rank at most $\tilde{k}+\tilde{m}$ if the LSOP-R-I yields a finite optimal value, i.e., $\mathbf{V}_{\text{rel -I}} > -\infty$.

The proof of the rank bound in Theorems 8 and 9 can be found in Appendix A.12. We note that

(i) For the LSOP with a sparse domain set $\mathcal{D}$ in (3), its relaxation corresponds to LSOP-R or LSOP-R-I which depends on the property of function $f(\cdot)$. Interestingly, either relaxation violates the original low-rankness by at most $\tilde{m}$;

(ii) Applying the diagonal matrix space to the LSOP-R and LSOP-R-I dramatically changes their rank bounds to be linear in $\tilde{m}$, in contrast to those of $\sqrt{\tilde{m}}$ derived in Theorems 5 and 7. This is because an $n\times n$ diagonal matrix maps onto a vector of only length n, whereas a general symmetric matrix equals a size-$\mathcal{O}(n^2)$ vector; and

(iii) We give an example in Appendix B that attains the rank bound in Theorems 8 and 9.

As indicated previously, the Sparse Ridge Regression admits a sparse domain set in (3), i.e.,

$$\mathcal{D} := \{\boldsymbol{X}\in\mathcal{S}^n : \boldsymbol{X} = \operatorname{Diag}(\operatorname{diag}(\boldsymbol{X})), \operatorname{rank}(\boldsymbol{X})\le k, \|\boldsymbol{X}\|_F^2 := \|\operatorname{diag}(\boldsymbol{X})\|_2^2 \le z\},$$

where the function $f(\operatorname{diag}(\boldsymbol{X}))$ in (3) is specified to be $\|\operatorname{diag}(\boldsymbol{X})\|_2^2 - z$ and thus is sign-invariant. Specifically, the LSOP-R corresponding to Sparse Ridge Regression can be formulated by

$$\mathbf{V}_{\text{rel}} := \min_{\boldsymbol{X}\in\operatorname{conv}(\mathcal{D}),\boldsymbol{y}\in\mathbb{R}^m, z\in\mathbb{R}_+} \left\{\frac{1}{n}\|\boldsymbol{y}\|_2^2 + \alpha z : \boldsymbol{A}\operatorname{diag}(\boldsymbol{X}) = \boldsymbol{b}-\boldsymbol{y}\right\}, \qquad \text{(Sparse Ridge Regression-R)}$$

where given the variable $z\in\mathbb{R}_+$, $\operatorname{conv}(\mathcal{D}) := \Big\{\boldsymbol{X}\in\mathcal{S}^n : \boldsymbol{X} = \operatorname{Diag}(\operatorname{diag}(\boldsymbol{X})), \exists \boldsymbol{x}\in\mathbb{R}_+^n, \|\boldsymbol{x}\|_2^2 \le z, x_1\ge\cdots\ge x_n, x_{\tilde{k}+1}=0, \|\boldsymbol{X}\|_{(\ell)} \le \sum_{i\in[\ell]} x_i, \forall \ell\in[\tilde{k}-1], \|\boldsymbol{X}\|_* \le \sum_{i\in[\tilde{k}]} x_i\Big\}$.

Corollary 6 (Sparse Ridge Regression) *Given the data $\boldsymbol{A}\in\mathbb{R}^{m\times n}$ in Sparse Ridge Regression, there is an optimal solution $\boldsymbol{X}^*$ of Sparse Ridge Regression-R of rank at most $k+\operatorname{rank}(\boldsymbol{A})$.*

Proof. See Appendix A.13. □

It is worth mentioning that the diagonal elements of an optimal diagonal solution $\boldsymbol{X}^*$ to Sparse Ridge Regression-R can serve as a relaxed solution for the original Sparse Ridge Regression in vector space. According to the rank bound in Corollary 6, the zero-norm of such relaxed solution (i.e., $\operatorname{diag}(\boldsymbol{X}^*)$) is no larger than $k+\operatorname{rank}(\boldsymbol{A})$, smaller than the recent sparsity bound (i.e., $k+\operatorname{rank}(\boldsymbol{A})+1$) shown in Askari et al. (2022).

5. Solving LSOP-R by Column Generation and Rank-Reduction Algorithms

This section studies an efficient column generation algorithm for solving LSOP-R, coupled with a rank-reduction algorithm. The algorithms together return an optimal solution to LSOP-R whose rank must not exceed the theoretical guarantees. Throughout this section, we suppose that LSOP-R yields a finite optimal value, i.e, $\mathbf{V}_{\text{rel}} > -\infty$ and set $\operatorname{conv}(\mathcal{D})$ contains no line.

5.1. Column Generation Algorithm

The column generation does not require an explicit description of the convex hull of domain set $\mathcal{D}$, and instead use the representation theorem 18.5 in Rockafellar (1972) to describe the line-free convex set $\mathrm{conv}(\mathcal{D})$ via the linear programming formulation (with possibly infinite number of variables), implying an equivalent linear reformulation of LSOP-R as shown below

$$\text{(LSOP-R)} \quad \mathbf{V}_{\mathrm{rel}} := \min_{\boldsymbol{\alpha}\in\mathbb{R}_+^{|\mathcal{H}|},\boldsymbol{\gamma}\in\mathbb{R}_+^{|\mathcal{J}|}} \left\{ \sum_{h\in\mathcal{H}} \alpha_h \langle \boldsymbol{A}_0, \boldsymbol{X}_h\rangle + \sum_{j\in\mathcal{J}} \gamma_j \langle \boldsymbol{A}_0, \boldsymbol{d}_j\rangle : \sum_{h\in\mathcal{H}} \alpha_h = 1, \right.$$
$$\left. b_i^l \le \sum_{h\in\mathcal{H}} \alpha_h \langle \boldsymbol{A}_i, \boldsymbol{X}_h\rangle + \sum_{j\in\mathcal{J}} \gamma_j \langle \boldsymbol{A}_i, \boldsymbol{d}_j\rangle \le b_i^u, \forall i\in[m] \right\}, \tag{4}$$

where $\{\boldsymbol{X}_h\}_{h\in\mathcal{H}}$ and $\{\boldsymbol{d}_j\}_{j\in\mathcal{J}}$ consist of all the extreme points and extreme directions in the set $\mathrm{conv}(\mathcal{D})$, respectively.

It is often difficult to enumerate all the possible extreme points $\{\boldsymbol{X}_h\}_{h\in H}$ and extreme directions $\{\boldsymbol{d}_j\}_{j\in J}$ in the set $\mathrm{conv}(\mathcal{D})$ and directly solve the (semi-infinite) linear programming problem (4) to optimality. Alternatively, the column generation algorithm starts with the *restricted master problem* (RMP) which only contains a small portion of extreme points and extreme directions and then solves the *pricing problem* (PP) to iteratively generate an improving point. The detailed implementation is presented in Algorithm 1. The column generation Algorithm 1 can be effective in practice since a small number of points in $\mathrm{conv}(\mathcal{D})$ are often needed to return a (near-)optimal solution to LSOP-R (4) according to Carathéodory theorem.

As shown in Algorithm 1, at each iteration, we first solve a linear program–RMP based on a subset of extreme points $\{\boldsymbol{X}_h\}_{h\in\mathcal{H}}$ and a subset of extreme directions $\{\boldsymbol{d}_j\}_{j\in\mathcal{J}}$ collected from the set $\mathrm{conv}(\mathcal{D})$. Using the Lagrangian multipliers $(\nu, \boldsymbol{\mu}^l, \boldsymbol{\mu}^u)$, the dual problem of the RMP is equal to

$$\max_{\boldsymbol{\mu}^l\in\mathbb{R}_+^m,\boldsymbol{\mu}^u\in\mathbb{R}_+^m} \left\{ -\max_{h\in\mathcal{H}} \left\langle -\boldsymbol{A}_0 + \sum_{i\in[m]} (\mu_i^l - \mu_i^u)\boldsymbol{A}_i, \boldsymbol{X}_h \right\rangle + (\boldsymbol{b}^l)^\top \boldsymbol{\mu}^l - (\boldsymbol{b}^u)^\top \boldsymbol{\mu}^u : \right.$$
$$\left. 0 \ge \left\langle -\boldsymbol{A}_0 + \sum_{i\in[m]} (\mu_i^l - \mu_i^u)\boldsymbol{A}_i, \boldsymbol{d}_j \right\rangle, \forall j\in\mathcal{J} \right\}. \tag{5}$$

Then, given an optimal dual solution $(\nu^*, (\boldsymbol{\mu}^l)^*, (\boldsymbol{\mu}^u)^*)$ at Step 4, the Pricing Problem at Step 6 of Algorithm 1 finds an improving point or direction for LSOP-R (4) in set $\mathrm{conv}(\mathcal{D})$.

When the Pricing Problem is unbounded, there exists a direction $\widehat{\boldsymbol{d}}$ in the set $\mathrm{conv}(\mathcal{D})$ such that $\langle -\boldsymbol{A}_0 + \sum_{i\in[m]}((\mu_i^l)^* - (\mu_i^u)^*)\boldsymbol{A}_i, \widehat{\boldsymbol{d}}\rangle > 0$, which violates the constraint in the dual problem (5). Hence, adding this direction to the RMP leads to a smaller objective value. Besides, when the Pricing Problem yields a finite optimal value no larger than $\nu^* + \epsilon$, we can show that Algorithm 1 finds an $\epsilon-$optimal solution to the LSOP-R (4), as detailed below.

Proposition 6 *The output solution of column generation Algorithm 1 is ϵ-optimal to the LSOP-R (4) if the LSOP-R (4) yields a finite optimal value, i.e., $\mathbf{V}_{\rm rel} > -\infty$. That is, suppose that Algorithm 1 returns $\boldsymbol{X}^*$, then the inequalities $\mathbf{V}_{\rm rel} \leq \langle \boldsymbol{A}_0, \boldsymbol{X}^* \rangle \leq \mathbf{V}_{\rm rel} + \epsilon$ hold.*

Proof. See Appendix A.14. □

The convergence of the column generation Algorithm 1 is guaranteed by the theory of linear programming, since the Pricing Problem in fact seeks the positive reduced cost. The finite termination or the rate of convergence of this method has been established under the assumption of finiteness of extreme points and extreme directions or revising the problem setting (e.g., adding penalty term or enforcing the strict improvement of the RMP at each iteration). For more details, we refer interested readers to the works in Amor et al. (2004, 2009), Desrosiers and Lübbecke (2005). The time complexity of Algorithm 1 lies in the number of iterations as well as in solving the Pricing Problem. Fortunately, we show that the Pricing Problem can be easily solved and even admits closed-form solutions given some special domain set $\mathcal{D}$ in the next subsection.

5.2. Efficient Pricing Oracle

In this subsection, we show that despite not describing the convex hull of domain set $\mathcal{D}$, the Pricing Problem at Step 6 of Algorithm 1 is efficiently solvable and even admits the closed-form solution for some cases. Given the linear objective function, the Pricing Problem can be simplified by replacing $\operatorname{conv}(\mathcal{D})$ with the domain set $\mathcal{D}$, i.e.,

$$\mathbf{V}_{\rm P} := \max_{\boldsymbol{X} \in \mathcal{Q}} \left\{ \left\langle \tilde{\boldsymbol{A}}, \boldsymbol{X} \right\rangle : \operatorname{rank}(\boldsymbol{X}) \leq k, f(\boldsymbol{\lambda}(\boldsymbol{X})) \leq 0 \right\}, \qquad \text{(Pricing Problem-S)}$$

where we define $\tilde{\boldsymbol{A}} := -\boldsymbol{A}_0 + \sum_{i \in [m]} ((\mu_i^l)^* - (\mu_i^u)^*) \boldsymbol{A}_i$ throughout this subsection.

Next, let us explore the solution structure of Pricing Problem-S. Note that matrices $\boldsymbol{A}_0$ and $\{\boldsymbol{A}_i\}_{i \in [m]}$ are symmetric if matrix variable $\boldsymbol{X}$ is symmetric, and so does matrix $\tilde{\boldsymbol{A}}$.

Theorem 10 *The Pricing Problem-S has an optimal solution $\boldsymbol{X}^*$, as specified follows*

(i) Given $\mathcal{Q} := \mathcal{S}_+^n$ in (1), suppose that $\tilde{\boldsymbol{A}} := \boldsymbol{U} \operatorname{Diag}(\boldsymbol{\beta}) \boldsymbol{U}^\top$ denotes the eigen-decomposition of matrix $\tilde{\boldsymbol{A}}$ with eigenvalues $\beta_1 \geq \cdots \geq \beta_n$. Then $\boldsymbol{X}^* = \boldsymbol{U} \operatorname{Diag}(\boldsymbol{\lambda}^*) \boldsymbol{U}^\top$ and $\boldsymbol{\lambda}^* \in \mathbb{R}_+^n$ equals

$$\boldsymbol{\lambda}^* \in \arg\max_{\boldsymbol{\lambda} \in \mathbb{R}_+^n} \left\{ \boldsymbol{\lambda}^\top \boldsymbol{\beta} : \lambda_i = 0, \forall i \in [k+1, n], f(\boldsymbol{\lambda}) \leq 0 \right\}. \tag{6}$$

(ii) Given $\mathcal{Q} := \mathbb{R}^{n \times p}$ in (1), suppose that $\tilde{\boldsymbol{A}} := \boldsymbol{U} \operatorname{Diag}(\boldsymbol{\beta}) \boldsymbol{V}^\top$ is the singular value decomposition of matrix $\tilde{\boldsymbol{A}}$ with singular values $\beta_1 \geq \cdots \geq \beta_n$. Then $\boldsymbol{X}^* = \boldsymbol{U} \operatorname{Diag}(\boldsymbol{\lambda}^*) \boldsymbol{V}^\top$ and $\boldsymbol{\lambda}^* \in \mathbb{R}_+^n$ equals

$$\boldsymbol{\lambda}^* \in \arg\max_{\boldsymbol{\lambda} \in \mathbb{R}_+^n} \left\{ \boldsymbol{\lambda}^\top \boldsymbol{\beta} : \lambda_i = 0, \forall i \in [k+1, n], f(\boldsymbol{\lambda}) \leq 0 \right\}. \tag{7}$$

Algorithm 1 The Column Generation Algorithm for Solving LSOP-R (4) with $\mathbf{V}_{\text{rel}} > -\infty$

Input: Data $\boldsymbol{A}_0$, $\{\boldsymbol{A}_i, b_i^l, b_i^m\}_{i\in[m]}$, domain set $\mathcal{D}$ with a line-free convex hull, and optimality gap $\epsilon > 0$

Initialize $\mathcal{H} = \{1\}$, $\mathcal{J} = \emptyset$, and a feasible solution $\boldsymbol{X}_1$ to LSOP-R (4)

for $t = 1, 2, \cdots$ **do**

  Solve the RMP below and denote by $(\boldsymbol{\alpha}^*, \boldsymbol{\gamma}^*)$ the optimal solution

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}_+^{|\mathcal{H}|}, \boldsymbol{\gamma}\in\mathbb{R}_+^{|\mathcal{J}|}} \Bigg\{ \sum_{h\in\mathcal{H}} \alpha_h \langle \boldsymbol{A}_0, \boldsymbol{X}_h\rangle + \sum_{j\in\mathcal{J}} \gamma_j \langle \boldsymbol{A}_0, \boldsymbol{d}_j\rangle : \sum_{h\in\mathcal{H}} \alpha_h = 1, \quad b_i^l \le \sum_{h\in\mathcal{H}} \alpha_h \langle \boldsymbol{A}_i, \boldsymbol{X}_h\rangle + \sum_{j\in\mathcal{J}} \gamma_j \langle \boldsymbol{A}_i, \boldsymbol{d}_j\rangle \le b_i^u, \forall i \in [m] \Bigg\}, \tag{RMP}$$

  and let $(\nu^*, (\boldsymbol{\mu}^l)^*, (\boldsymbol{\mu}^u)^*)$ denote the corresponding optimal dual solutions

  Update solution $\boldsymbol{X}^* := \sum_{h\in\mathcal{H}} \alpha_h^* \boldsymbol{X}_i + \sum_{j\in\mathcal{J}} \gamma_j^* \boldsymbol{d}_j$

  Solve the Pricing Problem and let $\widehat{\boldsymbol{X}}$ denote its optimal solution

$$\mathbf{V}_{\text{P}} := \max_{\boldsymbol{X}\in\text{conv}(\mathcal{D})} \Big\langle -\boldsymbol{A}_0 + \sum_{i\in[m]} ((\mu_i^l)^* - (\mu_i^u)^*)\boldsymbol{A}_i, \boldsymbol{X} \Big\rangle. \tag{Pricing Problem}$$

  if $\mathbf{V}_{\text{P}} = \infty$ **then**

    $\widehat{\boldsymbol{X}}$ can be expressed by $\tilde{\boldsymbol{X}} + \theta\widehat{\boldsymbol{d}}$ with $\theta > 0$

    Add direction $\widehat{\boldsymbol{d}}$ to set $\{\boldsymbol{d}_j\}_{j\in\mathcal{J}}$ and update $\mathcal{J} := \mathcal{J} \cup \{|\mathcal{J}| + 1\}$

  else if $\mathbf{V}_{\text{P}} > \nu^* + \epsilon$ **then**

    Add point $\widehat{\boldsymbol{X}}$ to set $\{\boldsymbol{X}_h\}_{h\in\mathcal{H}}$ and update $\mathcal{H} := \mathcal{H} \cup \{|\mathcal{H}| + 1\}$

  else

    Terminate the iteration

  end if

end for

Output: Solution $\boldsymbol{X}^*$

(iii) Given $\mathcal{Q} := \mathcal{S}^n$ in (1), suppose that $\tilde{\boldsymbol{A}} := \boldsymbol{U}\,\text{Diag}(\boldsymbol{\beta})\boldsymbol{U}^\top$ denotes the eigen-decomposition of matrix $\tilde{\boldsymbol{A}}$ with eigenvalues $\beta_1 \ge \cdots \ge \beta_n$. Then $\boldsymbol{X}^* = \boldsymbol{U}\,\text{Diag}(\boldsymbol{\lambda}^*)\boldsymbol{U}^\top$ and $\boldsymbol{\lambda}^* \in \mathbb{R}^n$ equals

$$\boldsymbol{\lambda}^* \in \arg\max_{s\in[k+1]} \max_{\boldsymbol{\lambda}\in\mathbb{R}^n} \left\{ \boldsymbol{\lambda}^\top \boldsymbol{\beta} : \lambda_1 \ge \cdots \ge \lambda_n, \lambda_s = \lambda_{s+n-k-1} = 0, f(\boldsymbol{\lambda}) \le 0 \right\}. \tag{8}$$

(iv) Given $\mathcal{Q} := \mathcal{S}^n$ and a sign-invariant function $f(\cdot)$ in (1), suppose that $\tilde{\boldsymbol{A}} := \boldsymbol{U}\,\text{Diag}(\boldsymbol{\beta})\boldsymbol{U}^\top$ denotes the eigen-decomposition of matrix $\tilde{\boldsymbol{A}}$ with eigenvalues satisfying $|\beta_1| \ge \cdots \ge |\beta_n|$. Then $\boldsymbol{X}^* = \boldsymbol{U}\,\text{Diag}(\boldsymbol{\lambda}^*)\boldsymbol{U}^\top$ and $\boldsymbol{\lambda}^* \in \mathbb{R}^n$ equals

$$\boldsymbol{\lambda}^* \in \arg\max_{\boldsymbol{\lambda}\in\mathbb{R}^n} \left\{ |\boldsymbol{\lambda}|^\top |\boldsymbol{\beta}| : |\lambda_i| = 0, \forall i \in [k+1, n], f(|\boldsymbol{\lambda}|) \le 0 \right\}. \tag{9}$$

Proof. See Appendix A.15. □

We make the following remarks about Theorem 10:

(a) For any domain set $\mathcal{D}$ in (1), the Pricing Problem-S involves a rank constraint and is thus nonconvex. Theorem 10 provides a striking result that the Pricing Problem-S can be equivalent to solving a convex program over a vector variable of length n (see problems (6)-(9)). Therefore, the column generation Algorithm 1 can admit an efficient implementation for solving LSOP-R;

(b) By leveraging Theorem 10, we can derive the closed-form solutions to the Pricing Problem-S for a special family of domain set $\mathcal{D}$ as shown in Corollary 7;

(c) Although the convex hull of the domain set $\mathcal{D}$ may admit an explicit description based on the majorization technique (see, e.g., Proposition 1), the description involves many auxiliary variables and constraints and is semidefinite representable. Therefore, the original Pricing Problem over $\text{conv}(\mathcal{D})$ can still be challenging to optimize and scale. Quite differently, Theorem 10 allows us to target a simple convex program for solving the Pricing Problem, which significantly enhances the Algorithm 1 performance in our numerical study; and

(d) The column generation Algorithm 1 can be extended to solving LSOP-R-I with $\mathcal{Q} := \mathcal{S}^n$ whose pricing oracle can be also simplified, as shown in Corollary 8.

According to Theorem 10, the Pricing Problem-S can be equivalently converted to solving the maximization problem over variable $\boldsymbol{\lambda}$. Therefore, it suffices to derive closed-form solutions to problems (6)-(9). For any $\ell \in [1, \infty]$, we let $\|\cdot\|_\ell$ denote the ℓ-norm of a vector.

Corollary 7 *For a domain set $\mathcal{D} := \{\boldsymbol{X} \in \mathcal{Q} : \text{rank}(\boldsymbol{X}) \le k, \|\boldsymbol{\lambda}(\boldsymbol{X})\|_\ell \le c\}$, where $c \ge 0$, $\ell \in [1, \infty]$, and $1/\ell + 1/q = 1$, we have that*

(i) Given $\mathcal{Q} := \mathcal{S}_+^n$, the solution $\boldsymbol{\lambda}^*$ is optimal to problem (6) where $\lambda_i^* = c\sqrt[\ell]{\frac{(\beta_i)_+^q}{\sum_{j\in[k]}(\beta_j)_+^q}}$ for all $i \in [k]$ and $\lambda_i^* = 0$ for all $i \in [k+1, n]$;

(ii) Given $\mathcal{Q} := \mathbb{R}^{n\times p}$, the solution $\boldsymbol{\lambda}^*$ is optimal to problem (7) where $\lambda_i^* = c\sqrt[\ell]{\frac{\beta_i^q}{\sum_{j\in[k]}\beta_j^q}}$ for all $i \in [k]$ and $\lambda_i^* = 0$ for all $i \in [k+1, n]$;

(iii) Given $\mathcal{Q} := \mathcal{S}^n$, the solution $\boldsymbol{\lambda}^*$ is optimal to problem (9) where $\lambda_i^* = \text{sign}(\beta_i)c\sqrt[\ell]{\frac{|\beta_i|^q}{\sum_{j\in[k]}|\beta_j|^q}}$ for all $i \in [k]$ and $\lambda_i^* = 0$ for all $i \in [k+1, n]$.

Proof. See Appendix A.16. □

Since the LSOP-R-I is equivalent to solving $(\tilde{k}+1)$ subproblems over sets $\{\mathcal{Y}^s\}_{s\in[\tilde{k}+1]}$ defined in (2), we apply column generation Algorithm 1 to solving each subproblem and then take the best one among $(\tilde{k}+1)$ output solutions. To be specific, for each $s \in [\tilde{k}+1]$, the Algorithm 1 generates a new column by solving

$$\mathrm{V}_{\mathrm{P}}^s := \max_{\boldsymbol{X}\in\mathcal{Y}^s} \left\langle -\boldsymbol{A}_0 + \sum_{i\in[m]}((\mu_i^l)^* - (\mu_i^u)^*)\boldsymbol{A}_i, \boldsymbol{X} \right\rangle = \max_{\boldsymbol{X}\in\mathcal{Y}^s} \left\langle \tilde{\boldsymbol{A}}, \boldsymbol{X} \right\rangle, \tag{10}$$

which replaces the feasible set $\text{conv}(\mathcal{D})$ in Pricing Problem with set $\mathcal{Y}^s$ in (2). We close this subsection by simplifying the pricing problem (10) as follows.

Corollary 8 *Given $\mathcal{Q} := \mathcal{S}^n$ in (1), suppose that $\tilde{\boldsymbol{A}} := \boldsymbol{U}\,\text{Diag}(\boldsymbol{\beta})\boldsymbol{U}^\top$ denotes the eigen-decomposition of matrix $\tilde{\boldsymbol{A}}$ with eigenvalues $\beta_1 \geq \cdots \geq \beta_n$. Then for each $s \in [\tilde{k}+1]$, the pricing problem (10) has an optimal solution $\boldsymbol{X}^* = \boldsymbol{U}\,\text{Diag}(\boldsymbol{\lambda}^*)\boldsymbol{U}^\top$, where $\boldsymbol{\lambda}^* \in \mathbb{R}^n$ equals*

$$\boldsymbol{\lambda}^* := \max_{\boldsymbol{\lambda}\in\mathbb{R}^n} \left\{\boldsymbol{\lambda}^\top \boldsymbol{\beta} : \lambda_1 \geq \cdots \geq \lambda_n, \lambda_s = \lambda_{s+n-k-1} = 0, f(\boldsymbol{\lambda}) \leq 0\right\}. \tag{11}$$

Proof. The proof is identical to that of Part (iii) in Theorem 10 and thus is omitted. □

5.3. Rank-Reduction Algorithm

The column generation Algorithm 1 may not always be able to output a solution that satisfies the theoretical rank bounds. To resolve this, this subsection designs a rank-reduction Algorithm 2 to find an alternative solution of the same or smaller objective value while satisfying the desired rank bounds.

Given a (near-)optimal solution $\boldsymbol{X}^*$ of LSOP-R returned by Algorithm 1 that violates the proposed rank bound, our rank-reduction Algorithm 2 runs as follows: (i) since $\boldsymbol{X}^*$ is not an extreme point of LSOP-R, then we can find a direction $\boldsymbol{Y}$ in its feasible set, along which the objective value will decrease or stay the same; (ii) we move $\boldsymbol{X}^*$ along the direction $\boldsymbol{Y}$ until a point on the boundary of the feasible set of LSOP-R; (iii) we update $\boldsymbol{X}^*$ to be the new boundary point found; and (iv) finally, we terminate the iteration when no further movement is available, i.e., $\boldsymbol{X}^*$ is a feasible extreme point in LSOP-R that must satisfy the rank bound. Hence, the rank-reduction procedure in Algorithm 2 can be viewed as searching for an extreme point.

Suppose that $\boldsymbol{\lambda}^* \in \mathbb{R}^n$ denotes eigen-(singular-)value vector of matrix $\boldsymbol{X}^*$ in the descending order. Next, we show how to construct matrix $\boldsymbol{Y}$ and vector $\boldsymbol{x}^*$ with different matrix spaces in the rank-reduction Algorithm 2 (i.e., Step 5 and Step 4), which represents moving direction and determines the moving distance, respectively.

When $\mathcal{Q} := \mathcal{S}_+^n$, according to the convex hull description in Proposition 1, there exists an optimal solution $\boldsymbol{x}^*$ at Step 4 such that $f(\boldsymbol{x}^*) \leq 0$. When $\mathcal{Q} := \mathbb{R}^{n\times p}$, the convex hull $\text{conv}(\mathcal{D})$ described in Proposition 7 also ensures us a solution $\boldsymbol{x}^*$. When $\mathcal{Q} := \mathcal{S}^n$, we note that the solution $\boldsymbol{X}^*$ is feasible to LSOP-R-I and thus belongs to some set $\mathcal{Y}^s$ with $s \in [\tilde{k}+1]$, which, according to Proposition 4, allows us to compute a pair of solutions $(\boldsymbol{x}^*, s^*)$.

Next, we find the matrix $\boldsymbol{Y}$ at Step 5 based on the proofs of rank bounds. Suppose that the near-optimal solution $\boldsymbol{X}^*$ has rank r. When $\mathcal{Q} := \mathcal{S}_+^n$, we define

$$\boldsymbol{Y} := \boldsymbol{Q}_2 \boldsymbol{\Delta} \boldsymbol{Q}_2^\top, \quad \exists \boldsymbol{\Delta} \in \mathcal{S}^{r-\tilde{k}+1}, \langle \boldsymbol{Q}_2^\top \boldsymbol{A}_i \boldsymbol{Q}_2, \boldsymbol{\Delta}\rangle = 0, \forall i \in [m], \text{tr}(\boldsymbol{\Delta}) = 0, \tag{12}$$

where matrix $\boldsymbol{Q}_2 \in \mathbb{R}^{n\times r-\tilde{k}+1}$ consists of eigenvectors corresponding to the $(r-\tilde{k}+1)$ least nonzero eigenvalues (i.e., $\lambda^*_{\tilde{k}}\cdots,\lambda^*_r$) of matrix $\boldsymbol{X}^*$. When $\mathcal{Q}:=\mathbb{R}^{n\times p}$, we define

$$\boldsymbol{Y} := \boldsymbol{Q}_2\boldsymbol{\Delta}\boldsymbol{P}_2^\top, \quad \exists \boldsymbol{\Delta}\in\mathcal{S}^{r-\tilde{k}+1}, \langle \boldsymbol{P}_2^\top \boldsymbol{A}_i\boldsymbol{Q}_2, \boldsymbol{\Delta}\rangle = 0, \forall i\in[m], \operatorname{tr}(\boldsymbol{\Delta})=0, \tag{13}$$

where matrices $\boldsymbol{Q}_2 \in \mathbb{R}^{n\times r-\tilde{k}+1}$, $\boldsymbol{P}_2 \in \mathbb{R}^{p\times r-\tilde{k}+1}$ consist of left and right singular vectors corresponding to the $(r-\tilde{k}+1)$ least nonzero singular values of matrix $\boldsymbol{X}^*$. When $\mathcal{Q}:=\mathcal{S}^n$, we considers the positive and negative eigenvalues in matrix $\boldsymbol{X}^*$, respectively, and suppose $\lambda^*_1\geq\cdots\geq\lambda^*_{d_1-1}>0=\lambda^*_{d_1}=\cdots=\lambda^*_{d_2-1}=0>\lambda^*_{d_2}\geq\cdots\geq\lambda^*_n$. Then, inspired by the proof of Theorem 6, let us define a block matrix $\boldsymbol{\Delta}$ below corresponding to positive and negative eigenvalues, respectively,

$$\begin{aligned}&\boldsymbol{Y} := \left(\boldsymbol{Q}_1^2\ \boldsymbol{Q}_3^1\right)\boldsymbol{\Delta}\left(\boldsymbol{Q}_1^2\ \boldsymbol{Q}_3^1\right)^\top, \quad \exists\boldsymbol{\Delta}\in\mathcal{S}^{r-\tilde{k}+2} := \begin{pmatrix}\boldsymbol{\Delta}_1\in\mathcal{S}^{d_1-s^*+1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Delta}_3\in\mathcal{S}^{s^*+n-\tilde{k}-d_2+1}\end{pmatrix},\\ &\langle\left(\boldsymbol{Q}_1^2\ \boldsymbol{Q}_3^1\right)^\top\boldsymbol{A}_i\left(\boldsymbol{Q}_1^2\ \boldsymbol{Q}_3^1\right),\boldsymbol{\Delta}\rangle = 0, \forall i\in[m], \operatorname{tr}(\boldsymbol{\Delta}_1)=0, \operatorname{tr}(\boldsymbol{\Delta}_3)=0,\end{aligned} \tag{14}$$

where given solution s^* at Step 4 of Algorithm 2, eigenvector matrices $\boldsymbol{Q}_1^2, \boldsymbol{Q}_3^1$ correspond to the positive eigenvalues $(\lambda^*_{s^*-1},\cdots,\lambda^*_{d_1-1})$ and negative eigenvalues $(\lambda^*_{d_2},\cdots,\lambda^*_{s^*+n-\tilde{k}})$, respectively.

We let $\boldsymbol{Y} := -\boldsymbol{Y}$ if $\langle\boldsymbol{A}_0,\boldsymbol{Y}\rangle>0$. Given a direction $\boldsymbol{Y}$, we can find the largest moving distance δ^* at Step 9 of Algorithm 2. To be specific, following the proofs in Theorems 1, 3 and 6, we let $\boldsymbol{\Lambda}_2 := \operatorname{Diag}(\lambda^*_{\tilde{k}},\cdots,\lambda^*_r)$ when $\mathcal{Q}:=\mathcal{S}^n_+$ or $\mathcal{Q}:=\mathbb{R}^{n\times p}$ and let $\boldsymbol{\Lambda}_1^2\in\mathcal{S}^{d_1-s^*+1}_{++} := \operatorname{Diag}(\lambda^*_{s^*-1},\cdots,\lambda^*_{d_1-1})$, $\boldsymbol{\Lambda}_3^1\in -\mathcal{S}^{s^*+n-\tilde{k}-d_2+1}_{++} := \operatorname{Diag}(\lambda^*_{d_2},\cdots,\lambda^*_{s^*+n-\tilde{k}})$ when $\mathcal{Q}:=\mathcal{S}^n$.

Following the searching procedure above, we show that the Algorithm 2 can output an alternative solution satisfying the rank bounds as summarized below. Besides, we show that Algorithm 2 always terminates and can strictly reduce the rank at each iteration for a special case in Corollary 9. More specifically, at each iteration of Algorithm 2, the convex program at Step 4 can be efficiently solved by commercial solvers or first-order methods; we can also readily find the direction $\boldsymbol{Y}$ in (12)–(14) using the matrix $\boldsymbol{\Delta}$ in the null space of linear equations; and the moving distance δ^* at Step 9 requires solving a simple univariate convex optimization problem. In our numerical study, the Algorithm 2 efficiently reduces the rank and converges fast.

Theorem 11 *For the rank-reduction Algorithm 2, the following statements must hold:*

(i) Algorithm 2 always terminates; and

(ii) Let $\boldsymbol{X}^*$ denote the output solution of Algorithm 2. Then $\boldsymbol{X}^*$ is ϵ-optimal to either LSOP-R or LSOP-R-I, i.e.,

$$\begin{cases}\mathbf{V}_{\text{rel}}\leq\langle\boldsymbol{A}_0,\boldsymbol{X}^*\rangle\leq\mathbf{V}_{\text{rel}}+\epsilon & \text{if } \mathcal{Q}:=\mathcal{S}^n_+ \text{ or } \mathcal{Q}:=\mathbb{R}^{n\times p};\\ \mathbf{V}_{\text{rel -I}}\leq\langle\boldsymbol{A}_0,\boldsymbol{X}^*\rangle\leq\mathbf{V}_{\text{rel -I}}+\epsilon & \text{if } \mathcal{Q}:=\mathcal{S}^n.\end{cases},$$

and the rank of solution $\boldsymbol{X}^*$ satisfies

$$\operatorname{rank}(\boldsymbol{X}^*)\leq\begin{cases}\tilde{k}+\lfloor\sqrt{2\tilde{m}+9/4}-3/2\rfloor, & \text{if } \mathcal{Q}:=\mathcal{S}^n_+ \text{ or } \mathcal{Q}:=\mathbb{R}^{n\times p};\\ \tilde{k}+\lfloor\sqrt{4\tilde{m}+9}\rfloor-3, & \text{if } \mathcal{Q}:=\mathcal{S}^n.\end{cases},$$

where integer $\tilde{k}$ is being defined in Definition 1.

Algorithm 2 Rank Reduction for LSOP-R or LSOP-R-I with $\mathbf{V}_{\text{rel}} > -\infty$, $\mathbf{V}_{\text{rel-I}} > -\infty$

Input: Data $\boldsymbol{A}_0$, $\{\boldsymbol{A}_i, b_i^l, b_i^m\}_{i\in[m]}$, domain set $\mathcal{D}$ with a line-free convex hull, matrix space $\mathcal{Q} \in \{\mathcal{S}_+^n, \mathbb{R}^{n\times p}, \mathcal{S}^n\}$, integer $\tilde{k}$ being defined in Definition 1, and scalar $\delta^* := 1$

Initialize an ϵ-optimal solution $\boldsymbol{X}^*$ of LSOP-R or LSOP-R-I returned by Algorithm 1

while $\delta^* > 0$ **do**

  Let $\boldsymbol{\lambda}^* \in \mathbb{R}^n$ denote eigenvalues of matrix $\boldsymbol{X}^*$ in descending order if $\boldsymbol{X}^* \in \mathcal{Q}$ is symmetric, or denote singular values otherwise. Then, compute

$$\begin{cases} \boldsymbol{x}^* \in \arg\min_{\boldsymbol{x}\in\mathbb{R}_+^n} \{f(\boldsymbol{x}) : x_1 \geq \cdots \geq x_n, x_{k+1} = 0, \boldsymbol{x} \succeq \boldsymbol{\lambda}^*\}, & \text{if } \mathcal{Q} := \mathcal{S}_+^n; \\ \boldsymbol{x}^* \in \arg\min_{\boldsymbol{x}\in\mathbb{R}_+^n} \{f(\boldsymbol{x}) : x_1 \geq \cdots \geq x_n, x_{k+1} = 0, \boldsymbol{x} \succ \boldsymbol{\lambda}^*\}, & \text{if } \mathcal{Q} := \mathbb{R}^{n\times p}; \\ (\boldsymbol{x}^*, s^*) \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^n, s\in[\tilde{k}+1]} \{f(\boldsymbol{x}) : x_1 \geq \cdots \geq x_n, x_s = x_{s+n-\tilde{k}-1} = 0, \boldsymbol{x} \succeq \boldsymbol{\lambda}^*\}, & \text{if } \mathcal{Q} := \mathcal{S}^n. \end{cases}$$

  Find a nonzero matrix $\boldsymbol{Y}$

$$\boldsymbol{Y} := \begin{cases} \boldsymbol{Q}_2 \boldsymbol{\Delta} \boldsymbol{Q}_2^\top \text{ as defined in (12)}, & \text{if } \mathcal{Q} := \mathcal{S}_+^n; \\ \boldsymbol{Q}_2 \boldsymbol{\Delta} \boldsymbol{P}_2^\top \text{ as defined in (13)}, & \text{if } \mathcal{Q} := \mathbb{R}^{n\times p}; \\ \left(\boldsymbol{Q}_1^2 \ \boldsymbol{Q}_3^1\right) \boldsymbol{\Delta} \left(\boldsymbol{Q}_1^2 \ \boldsymbol{Q}_3^1\right)^\top \text{ as defined in (14)}, & \text{if } \mathcal{Q} := \mathcal{S}^n. \end{cases}$$

  if $\boldsymbol{Y}$ does not exist **then**

    Set $\delta^* = 0$ and break the loop

  else

    Let $\boldsymbol{Y} := -\operatorname{sign}(\langle \boldsymbol{A}_0, \boldsymbol{Y}\rangle)\boldsymbol{Y}$ and $\boldsymbol{X}(\delta) := \boldsymbol{X}^* + \delta\boldsymbol{Y}$, and compute δ^* by

$$\begin{cases} \arg\max_{\delta\geq 0} \{\delta : \boldsymbol{x}^* \succeq \boldsymbol{\lambda}(\boldsymbol{X}(\delta)), \lambda_{\min}(\boldsymbol{\Lambda}_2 + \delta\boldsymbol{\Delta}) \geq 0\}, & \text{if } \mathcal{Q} := \mathcal{S}_+^n; \\ \arg\max_{\delta\geq 0} \{\delta : \boldsymbol{x}^* \succ \boldsymbol{\lambda}(\boldsymbol{X}(\delta)), \lambda_{\min}(\boldsymbol{\Lambda}_2 + \delta\boldsymbol{\Delta}) \geq 0\}, & \text{if } \mathcal{Q} := \mathbb{R}^{n\times p}; \\ \arg\max_{\delta\geq 0} \{\delta : \boldsymbol{x}^* \succeq \boldsymbol{\lambda}(\boldsymbol{X}(\delta)), \lambda_{\min}(\boldsymbol{\Lambda}_1^2 + \delta\boldsymbol{\Delta}_1) \geq 0, \lambda_{\max}(\boldsymbol{\Lambda}_3^1 + \delta\boldsymbol{\Delta}_3) \leq 0\}, & \text{if } \mathcal{Q} := \mathcal{S}^n. \end{cases}$$

    Update $\boldsymbol{X}^* := \boldsymbol{X}(\delta^*)$

  end if

end while

Output: Solution $\boldsymbol{X}^*$

Proof. See Appendix A.17. □

Corollary 9 *For a domain set $\mathcal{D}$ with $\tilde{k} = 1$ being defined in Definition 1, the rank-reduction Algorithm 2 reduces the solution's rank by at least one at each iteration.*

Proof. See Appendix A.18. □

6. Numerical Study

In this section, we numerically test the proposed rank bounds, column generation Algorithm 1, and rank-reduction Algorithm 2 on three LSOP examples: multiple-input and multiple-output (MIMO) radio communication network, optimal power flow, and Matrix Completion. All the experiments are conducted in Python 3.6 with calls to Gurobi 9.5.2 and MOSEK 10.0.29 on a PC with 10-core CPU, 16-core GPU, and 16GB of memory. All the times reported are wall-clock times.

6.1. LSOP with Trace and Log-Det Spectral Constraints: MIMO Radio Network

In recent years, there has been an increasing interest in a special class of LSOP with trace and log-det spectral constraints in the communication and signal processing field. Specifically, the objective is to find a low-rank transmit covariance matrix $\boldsymbol{X} \in \mathcal{S}_+^n$, subject to the trace and log-det constraints, where the former limits the transmit power and the latter is often used for entropy and mutual information requirements (see, e.g., Yu and Lau 2010 and references therein). Formally, we consider the LSOP built on a special domain set $\mathcal{D}$ below:

$$\mathcal{D} := \{\boldsymbol{X} \in \mathcal{S}_+^n : \operatorname{rank}(\boldsymbol{X}) \le k, \operatorname{tr}(\boldsymbol{X}) \le U, \log\det(\boldsymbol{I}_n + \boldsymbol{X}) \ge L\},$$

where $U, L \ge 0$ are pre-specified constants and here, the spectral function in (1) is defined as $F(\boldsymbol{X}) := \max\{\operatorname{tr}(\boldsymbol{X}) - U, L - \log\det(\boldsymbol{I}_n + \boldsymbol{X})\}$.

Since the domain set $\mathcal{D}$ above is defined in the positive semidefinite matrix space $\mathcal{Q} := \mathcal{S}_+^n$, using the convex hull description in Proposition 1, we can explicitly describe the $\operatorname{conv}(\mathcal{D})$. Then, for this specific LSOP, its LSOP-R can be formulated as

$$\mathbf{V}_{\text{rel}} := \min_{\boldsymbol{X} \in \operatorname{conv}(\mathcal{D})} \left\{ \langle \boldsymbol{A}_0, \boldsymbol{X} \rangle : b_i^l \le \langle \boldsymbol{A}_i, \boldsymbol{X} \rangle \le b_i^u, \forall i \in [m] \right\}, \tag{15}$$

where $\operatorname{conv}(\mathcal{D}) := \{\boldsymbol{X} \in \mathcal{Q} : \exists \boldsymbol{x} \in \mathbb{R}_+^n, \sum_{i \in [n]} x_i \le U, \sum_{i \in [n]} \log(1 + x_i) \ge L, x_1 \ge \cdots \ge x_n, x_{k+1} = 0, \|\boldsymbol{X}\|_{(\ell)} \le \sum_{i \in [\ell]} x_i, \forall \ell \in [k], \operatorname{tr}(\boldsymbol{X}) = \sum_{i \in [k]} x_i\}$.

The LSOP-R (15) can be recast as a semidefinite program and be directly solved by off-the-shelf solvers like MOSEK since the set $\operatorname{conv}(\mathcal{D})$ above is semidefinite representable. However, the semidefinite program is known to be hard to scale. An efficient approach to solving the LSOP-R (15) is through our column generation Algorithm 1, which at each iteration, solves a linear program RMP and a vector-based Pricing Problem 6 based on Theorem 10, as defined below

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}_+^n} \left\{ \boldsymbol{\lambda}^\top \boldsymbol{\beta} : \lambda_i = 0, \forall i \in [k+1, n], \sum_{i \in [k]} \lambda_i \le U, \sum_{i \in [k]} \log(1 + \lambda_i) \ge L \right\}. \tag{16}$$

Alternatively, without using Theorem 10, the original Pricing Problem can be solved as a semidefinite program by plugging the set $\operatorname{conv}(\mathcal{D})$ above when generating a new column. This column generation procedure is termed as the "naive column generation algorithm." In the following, we numerically test the three methods available for solving the LSOP-R (15) on synthetic data in Table 2, which demonstrates the efficiency of our Algorithm 1.

For the LSOP-R (15), we consider a single testing case by fixing the parameters n, m, k and generating random data L, U, $\boldsymbol{A}_0 \in \mathcal{S}^n$, and $\{\boldsymbol{A}_i, b_i^l, b_i^u\}_{i\in[m]} \in \mathcal{S}^n \times \mathbb{R} \times \mathbb{R}$. Particularly, we first generate a rank-one matrix $\boldsymbol{X}_0 := \frac{1}{\|\boldsymbol{x}\|_2}\boldsymbol{x}\boldsymbol{x}^\top$, where the vector $\boldsymbol{x} \in \mathbb{R}^n$ follows the standard normal distribution. Then, we define the constants U, L to be $U := \operatorname{tr}(\boldsymbol{X}_0) + \sigma_1$ and $L := \log\det(\boldsymbol{I}_n + \boldsymbol{X}_0) - \sigma_2$ with random numbers σ_1, σ_2 uniform in $[0,1]$, which ensures the inclusion $\boldsymbol{X}_0 \in \mathcal{D}$. Next, we generate matrices $\{\boldsymbol{A}_i\}_{i\in\{0\}\cup[m]} \in \mathbb{R}^{n\times n}$ from standard normal distribution and symmetrize them by $\boldsymbol{A}_i := (\boldsymbol{A}_i + \boldsymbol{A}_i^\top)/2$ for all $i \in \{0\}\cup[m]$. Finally, for each $i \in [m]$, we let $b_i^l := \langle \boldsymbol{A}_i, \boldsymbol{X}_0\rangle - \theta_i^1$ and $b_i^u := \langle \boldsymbol{A}_i, \boldsymbol{X}_0\rangle + \theta_i^2$ with the random constants θ_i^1, θ_i^2 uniform in $[0,1]$.

In Table 2, we compare the computational time (in seconds) and the rank of output solution of the LSOP-R (15) returned by three methods: MOSEK, column generation with and without the acceleration of Pricing Problem (16). It should be mentioned in Table 2 and all the following tables that we let "CG" denote the column generation and mark "–" in a unsolved case reaching the one-hour limit; for any output solution, we count the number of its eigenvalues or singular values greater than 10^{-10} as its rank; and we set the optimality gap of column generation algorithm to be $\epsilon := 10^{-4}$ such that the output value is no larger ϵ than $\mathbf{V}_{\text{rel}}$. It is seen that our proposed column generation Algorithm 1 performs very well in computational efficiency and solution quality, dominating the performance of commercial solver MOSEK and naive column generation in both respects. In particular, the naive column generation solves two testing cases within one hour limit. With the efficient Pricing Problem (16) used in the proposed Algorithm 1, all the cases are solved within two minutes. It is interesting to see that the solution returned by Algorithm 1 always satisfies the rank-k constraint. That is, for these instances, CG Algorithm 1 finds an optimal solution to the original LSOP. Plugging the output solution into the rank-reduction Algorithm 2, the rank can be further reduced by one or two within two seconds, implying that the integer $\tilde{k}$, albeit unknown, should be much smaller k for these instances. By contrast, MOSEK, based on the interior point method, tends to yield a high-rank solution. Finally, we observe that the theoretical rank bound based on the integer k, as presented in the last column, is consistently higher than that of the solution found by the Algorithm 1 or Algorithm 2. This suggests that the theoretical bound may not be tight for these instances.

6.2. LSOP with Trace Constraints: Optimal Power Flow

This subsection numerically tests a special QCQP widely adopted in the optimal power flow (OPF) problem (Bedoya et al. 2019, Eltved and Burer 2022)

$$\min_{\boldsymbol{X}\in\mathcal{D}} \left\{\langle \boldsymbol{A}_0, \boldsymbol{X}\rangle : \langle \boldsymbol{A}_i, \boldsymbol{X}\rangle \le b_i^u, \forall i \in [m]\right\}, \quad \mathcal{D} := \{\boldsymbol{X} \in \mathcal{S}_+^n : \operatorname{rank}(\boldsymbol{X}) \le 1, L \le \operatorname{tr}(\boldsymbol{X}) \le U\}, \tag{OPF}$$

Table 2 Numerical performance of three approaches to solving LSOP-R (15)

n	m	k	MOSEK		Naive CG		Proposed Algorithms				Rank bound
							CG Algorithm 1		Algorithm 2		
			rank	time(s)	rank	time(s)	rank	time(s)	reduced rank	time(s)	(Theorem 2)
50	5	5	48	17	3	223	3	1	2	1	7
50	10	5	29	19	5	1261	5	1	3	1	8
50	10	10	32	183	–	–	5	1	3	1	13
100	10	10	–	–	–	–	2	2	1	1	13
100	15	10	–	–	–	–	5	2	3	1	14
100	15	15	–	–	–	–	5	3	3	1	19
200	15	15	–	–	–	–	4	3	2	1	19
200	20	15	–	–	–	–	7	9	5	1	20
200	25	25	–	–	–	–	11	21	9	1	30
500	25	25	–	–	–	–	10	24	8	2	30
500	50	25	–	–	–	–	21	99	20	2	33
500	50	50	–	–	–	–	22	104	20	2	58

where the spectral function in (1) is $F(\boldsymbol{X}) := \max\{\operatorname{tr}(\boldsymbol{X}) - U, L - \operatorname{tr}(\boldsymbol{X})\}$. The LSOP-R corresponding to OPF can be formulated as

$$\begin{aligned} \mathbf{V}_{\text{rel}} &:= \min_{\boldsymbol{X} \in \operatorname{conv}(\mathcal{D})} \left\{ \langle \boldsymbol{A}_0, \boldsymbol{X} \rangle : \langle \boldsymbol{A}_i, \boldsymbol{X} \rangle \le b_i^u, \forall i \in [m] \right\}, \\ \operatorname{conv}(\mathcal{D}) &:= \{ \boldsymbol{X} \in \mathcal{S}_+^n : L \le \operatorname{tr}(\boldsymbol{X}) \le U \}, \end{aligned} \tag{OPF-R}$$

and consequently, the Pricing Problem in Algorithm 1 is equivalent to

$$\max_{\boldsymbol{X} \in \operatorname{conv}(\mathcal{D})} \left\langle -\boldsymbol{A}_0 + \sum_{i \in [m]} ((\mu_i^l)^* - (\mu_i^u)^*) \boldsymbol{A}_i, \boldsymbol{X} \right\rangle, \quad \operatorname{conv}(\mathcal{D}) := \{ \boldsymbol{X} \in \mathcal{S}_+^n : L \le \operatorname{tr}(\boldsymbol{X}) \le U \}.$$

The Pricing Problem above, according to Part (i) of Theorem 10, is equivalent to solving problem (6) which now admits a closed-form solution; that is, if $-\boldsymbol{A}_0 + \sum_{i \in [m]} ((\mu_i^l)^* - (\mu_i^u)^*) \boldsymbol{A}_i = \boldsymbol{U} \operatorname{Diag}(\boldsymbol{\beta}) \boldsymbol{U}^\top$ with eigenvalues $\beta_1 \ge \cdots \ge \beta_n$, then an optimal solution of the pricing problem is $\boldsymbol{X}^* = \boldsymbol{U} \operatorname{Diag}(\boldsymbol{\lambda}^*) \boldsymbol{U}^\top$, where $\lambda_1^* = U$ if $\beta_1 \ge 0$, and L, otherwise and $\lambda_i^* = 0$ for all $i \in [2, n]$.

Following the synthetic data generation detailed in the previous subsection, we compare the performance of our column generation Algorithm 1 with MOSEK and naive column generation when solving OPF-R. The computational results are displayed in Table 3. We can see that despite with a simple description of set $\operatorname{conv}(\mathcal{D})$, the MOSEK and naive column generation are difficult to scale under the curse of semidefinite program. Our column generation Algorithm 1 can efficiently return a low-rank solution for OPF-R that satisfies the theoretical rank bound in Theorem 2; however, the output rank is larger than one. We observe that the output solution of Algorithm 1 may not attain all m linear equations in practice, and the rank bound can be strengthened using only the binding constraints. Thus, we tailor the rank-reduction Algorithm 2 for OPF-R with $\tilde{k} = k = 1$, where we let $\boldsymbol{x}^* := (\sum_{i \in [n]} \lambda_i^*, 0, \cdots, 0)$ at Step 4 in Algorithm 2 and compute matrix $\boldsymbol{Y}$ at Step 5 based on the binding constraints, which successfully reduces the output rank of Algorithm 1, as presented in Table 3. In fact, the rank-reduction procedure can slightly improve the output solution from Algorithm 1 due to the objective function decreasing along the direction $\boldsymbol{Y}$.

Table 3 Numerical performance of three approaches to solving OPF-R

n	m	k	MOSEK		Naive CG		Proposed Algorithms: CG Algorithm 1		Proposed Algorithms: Algorithm 2		Rank bound
			rank	time(s)	rank	time(s)	rank	time(s)	reduced rank	time(s)	(Theorem 2)
1000	50	1	28	160	–	–	3	42	2	3	9
1000	60	1	32	195	–	–	5	80	3	10	10
1500	60	1	27	642	–	–	3	113	2	11	10
1500	75	1	186	724	–	–	6	344	4	35	11
2000	75	1	40	1850	–	–	5	594	3	67	11
2000	90	1	12	2236	–	–	4	483	2	27	13
2500	90	1	–	–	–	–	5	1323	3	122	13
2500	100	1	–	–	–	–	4	1326	2	114	13

6.3. Matrix Completion

This subsection solves the Matrix Completion-R by the following bilevel optimization form that allows for the noisy observations:

$$\begin{aligned} \mathbf{V}_{\text{rel}} := \min_{z\in\mathbb{R}_+} \quad & z \\ \text{s.t.} \quad & \underline{\delta} \geq \min_{\boldsymbol{X},\delta} \{\delta : |X_{ij} - A_{ij}| \leq \delta, \forall (i,j)\in\Omega, \boldsymbol{X}\in \operatorname{conv}(\mathcal{D}(z))\}, \\ & \operatorname{conv}(\mathcal{D}(z)) := \{\boldsymbol{X}\in\mathbb{R}^{n\times p} : \|\boldsymbol{X}\|_* \leq z\} \end{aligned} \quad \text{(Matrix Completion-R1)}$$

where $\underline{\delta}\geq 0$ is small and depends on the noisy level of observed entries $\{A_{ij}\}_{(i,j)\in\Omega}$ with Ω denoting the indices of the observed entries. If $\underline{\delta}=0$, then Matrix Completion-R1 is equivalent to the noiseless Matrix Completion-R. Following the work of Zhang et al. (2012), we generate a synthetic rank-k matrix $\boldsymbol{A}\in\mathbb{R}^{n\times p}$ from the model $\boldsymbol{A}=\boldsymbol{U}\boldsymbol{V}^\top+\epsilon\boldsymbol{Z}$, where $\boldsymbol{U}\in\mathbb{R}^{n\times k}, \boldsymbol{V}\in\mathbb{R}^{p\times n}$ compose i.i.d. entries from the standard normal distribution and $\boldsymbol{Z}\in\mathbb{R}^{n\times p}$ is a Gaussian white noise. In addition, we set a relatively small noise level $\epsilon=10^{-4}$ and accordingly, we let $\underline{\delta}=10^{-4}$ in Matrix Completion-R1. The set Ω is sampled uniformly from $[n]\times[p]$ at random.

For any solution z of the Matrix Completion-R1, our column generation Algorithm 1 is able to efficiently solve the embedded optimization problem over $(\boldsymbol{X},\delta)$ as this problem falls into our LSOP-R framework. This motivates us to use the binary search algorithm for solving Matrix Completion-R1. To do so, we begin with a pair of upper and lower bounds of z, where we let $z_U=\|\boldsymbol{A}_\Omega\|_*$ and $z_L=0$ with matrix $\boldsymbol{A}_\Omega$ containing the observed entries $\{A_{ij}\}_{(i,j)\in\Omega}$ and all other zeros; at each step t, we let $z_t=(z_U+z_L)/2$ and compute $(\boldsymbol{X}_t,\delta_t)$ by the proposed Algorithm 1; and if $\delta_t\geq\underline{\delta}$, we let $z_L=z_t$, otherwise, we let $z_U=z_t$; and we terminate this searching procedure when $z_U-z_L\leq 10^{-4}$ holds.

Table 4 presents the numerical results, which show that even if being nested within the iterative binary search algorithm, our Algorithm 1 is more scalable than the other two methods on the testing cases. However, unlike solving LSOP-R (15) and OPF-R, the Algorithm 1 fails to return a low-rank solution, provided that Matrix Completion-R1 admits a rank bound of $\lfloor\sqrt{2|\Omega|+9/4}-1/2\rfloor$ based on Corollary 5. In addition, we have that $\underline{k}=1$. Then, according to Corollary 9, the rank-reduction Algorithm 2 converges fast to reach the rank bound, as shown in Table 4. It is seen that the reduced rank is always less than the original rank k, which demonstrates the strength of Matrix Completion-R1 in exactly solving the original Matrix Completion. We note that Algorithm 2 may not yield a rank strictly less than the theoretical bound, which is because given $\underline{\delta}=10^{-4}$, all the linear inequalities in Matrix Completion-R1 are nearly binding.

Table 4 Numerical performance of three approaches to solving Matrix Completion-R1

n	p	$\lvert\Omega\rvert$	k	MOSEK		Naive CG		Proposed Algorithms				Rank bound
								CG Algorithm 1		Algorithm 2		
				rank	time(s)	rank	time(s)	rank	time(s)	reduced rank	time(s)	(Theorem 4)
100	50	30	10	22	57	–	–	23	19	7	1	7
100	100	40	10	29	303	–	–	31	52	8	1	8
100	200	50	10	–	–	–	–	36	120	9	3	9
300	200	60	15	–	–	–	–	49	342	10	11	10
300	300	70	15	–	–	–	–	60	502	11	16	11
300	400	80	15	–	–	–	–	64	1434	12	29	12

7. Conclusion

This paper studied the low-rank spectral constrained optimization problem by deriving the rank bounds for its partial convexification. Our rank bounds remain consistent across different domain sets in the form of (1) with a fixed matrix space. This paper specifically investigated positive semidefinite, non-symmetric, symmetric, and diagonal matrix spaces. Due to the flexible domain set, we have applied the proposed rank bounds to various low-rank application examples, including kernel learning, QCQP, fair PCA, fair SVD, matrix completion, and sparse ridge regression. To harvest the promising theoretical results, we develop an efficient column generation algorithm for solving the partial convexification coupled with a rank-reduction algorithm. One possible future direction is to study the general nonlinear objective functions.

Acknowledgments

This research has been supported in part by the Georgia Tech ARC-ACO fellowship. The authors would like to thank Prof. Fatma Kılınç-Karzan from Carnegie Mellon University for her valuable suggestions on the earlier version of this paper.

References

Amor HB, Desrosiers J, Frangioni A (2004) *Stabilization in column generation* (Groupe d'études et de recherche en analyse des décisions).

Amor HMB, Desrosiers J, Frangioni A (2009) On the choice of explicit stabilizing terms in column generation. *Discrete Applied Mathematics* 157(6):1167–1184.

Askari A, d'Aspremont A, Ghaoui LE (2022) Approximation bounds for sparse programs. *SIAM Journal on Mathematics of Data Science* 4(2):514–530.

Barvinok AI (1995) Problems of distance geometry and convex properties of quadratic maps. *Discrete & Computational Geometry* 13(2):189–202.

Bedoya JC, Abdelhadi A, Liu CC, Dubey A (2019) A qcqp and sdp formulation of the optimal power flow including renewable energy resources. *2019 International Symposium on Systems Engineering (ISSE)*, 1–8 (IEEE).

Ben-Tal A, Nemirovski A (2001) *Lectures on modern convex optimization: analysis, algorithms, and engineering applications* (SIAM).

Bertsimas D, Cory-Wright R, Pauphilet J (2021) A new perspective on low-rank optimization. *arXiv preprint arXiv:2105.05947* .

Burer S, Monteiro RD (2003) A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming* 95(2):329–357.

Burer S, Monteiro RD (2005) Local minima and convergence in low-rank semidefinite programming. *Mathematical programming* 103(3):427–444.

Burer S, Ye Y (2020) Exact semidefinite formulations for a class of (random and non-random) nonconvex quadratic programs. *Mathematical Programming* 181(1):1–17.

Cai JF, Candès EJ, Shen Z (2010) A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization* 20(4):1956–1982.

Chakraborty S, Zhou J, Balasubramanian V, Panchanathan S, Davidson I, Ye J (2013) Active matrix completion. *2013 IEEE 13th international conference on data mining*, 81–90 (IEEE).

Desrosiers J, Lübbecke ME (2005) A primer in column generation. *Column generation*, 1–32 (Springer).

Deza MM, Laurent M, Weismantel R (1997) *Geometry of cuts and metrics*, volume 2 (Springer).

Drusvyatskiy D, Kempton C (2015) Variational analysis of spectral functions simplified. *arXiv preprint arXiv:1506.05170* .

Eltved A, Burer S (2022) Strengthened sdp relaxation for an extended trust region subproblem with an application to optimal power flow. *Mathematical Programming* 1–26.

Gharanjik A, Shankar B, Soltanalian M, Oftersten B (2016) An iterative approach to nonconvex qcqp with applications in signal processing. *2016 IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 1–5 (IEEE).

Hardy GH, Littlewood JE, Pólya G, Pólya G, et al. (1952) *Inequalities* (Cambridge university press).

Josz C, Fliscounakis S, Maeght J, Panciatici P (2016) Ac power flow data in matpower and qcqp format: itesla, rte snapshots, and pegase. *arXiv preprint arXiv:1603.01533* .

Khobahi S, Soltanalian M, Jiang F, Swindlehurst AL (2019) Optimized transmission for parameter estimation in wireless sensor networks. *IEEE Transactions on Signal and Information Processing over Networks* 6:35–47.

Kılınç-Karzan F, Wang AL (2021) Exactness in sdp relaxations of qcqps: Theory and applications. *arXiv preprint arXiv:2107.06885* .

Kim J, Tawarmalani M, Richard JPP (2022) Convexification of permutation-invariant sets and an application to sparse principal component analysis. *Mathematics of Operations Research* 47(4):2547–2584.

Kulis B, Sustik MA, Dhillon IS (2009) Low-rank kernel learning with bregman matrix divergences. *Journal of Machine Learning Research* 10(2).

Lau LC, Ravi R, Singh M (2011) *Iterative methods in combinatorial optimization*, volume 46 (Cambridge University Press).

Li Y, Xie W (2020) Exact and approximation algorithms for sparse pca. *arXiv preprint arXiv:2008.12438* .

Li Y, Xie W (2021) Beyond symmetry: Best submatrix selection for the sparse truncated svd. *arXiv preprint arXiv:2105.03179* .

Li Y, Xie W (2022) On the exactness of dantzig-wolfe relaxation for rank constrained optimization problems. *arXiv preprint arXiv:2210.16191* .

Miao J, Kou KI (2021) Color image recovery using low-rank quaternion matrix completion algorithm. *IEEE Transactions on Image Processing* 31:190–201.

Pataki G (1998) On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of operations research* 23(2):339–358.

Rockafellar RT (1972) *Convex analysis* (Princeton university press).

Samadi S, Tantipongpipat U, Morgenstern JH, Singh M, Vempala S (2018) The price of fair pca: One extra dimension. *Advances in neural information processing systems* 31.

Tantipongpipat U, Samadi S, Singh M, Morgenstern JH, Vempala S (2019) Multi-criteria dimensionality reduction with applications to fairness. *Advances in neural information processing systems* 32.

Xie W, Deng X (2020) Scalable algorithms for the sparse ridge regression. *SIAM Journal on Optimization* 30(4):3359–3386.

Yu H, Lau VK (2010) Rank-constrained schur-convex optimization with multiple trace/log-det constraints. *IEEE transactions on signal processing* 59(1):304–314.

Zhang D, Hu Y, Ye J, Li X, He X (2012) Matrix completion by truncated nuclear norm regularization. *2012 IEEE Conference on computer vision and pattern recognition*, 2192–2199 (IEEE).

Appendix A. Proofs

A.1 Proof of Proposition 1

Proposition 1 *For a domain set $\mathcal{D}$ in positive semidefinite matrix space, i.e., $\mathcal{Q} := \mathcal{S}_+^n$ in (1), its convex hull* $\operatorname{conv}(\mathcal{D})$ *is equal to*

$$\left\{\boldsymbol{X} \in \mathcal{Q} : \exists \boldsymbol{x} \in \mathbb{R}_+^n, f(\boldsymbol{x}) \leq 0, x_1 \geq \cdots \geq x_n, x_{k+1} = 0, \|\boldsymbol{X}\|_{(\ell)} \leq \sum_{i \in [\ell]} x_i, \forall \ell \in [k], \operatorname{tr}(\boldsymbol{X}) = \sum_{i \in [k]} x_i \right\},$$

and is a closed set.

Proof. The derivation of set $\operatorname{conv}(\mathcal{D})$ can be found in Kim et al. (2022)[theorems 4 and 7]. We thus focus on proving the closeness of set $\operatorname{conv}(\mathcal{D})$. First, we define set $\mathcal{T}$ below

$$\mathcal{T} = \left\{(\boldsymbol{X}, \boldsymbol{x}) \in \mathcal{Q} \times \mathbb{R}_+^n : f(\boldsymbol{x}) \leq 0, x_1 \geq \cdots \geq x_n, x_{k+1} = 0, \|\boldsymbol{X}\|_{(\ell)} \leq \sum_{i \in [\ell]} x_i, \forall \ell \in [k], \operatorname{tr}(\boldsymbol{X}) = \sum_{i \in [k]} x_i \right\},$$

and the equation $\operatorname{conv}(\mathcal{D}) = \operatorname{proj}_{\boldsymbol{X}}(\mathcal{T})$ holds.

For any convergent sequence $\{\boldsymbol{X}_t\}_{t \in [T]}$ in set $\operatorname{conv}(\mathcal{D})$, suppose that the sequence converges to $\boldsymbol{X}^*$, i.e., $\lim_{t \to \infty} \boldsymbol{X}_t = \boldsymbol{X}^*$. Without loss of generality, we can assume that the convergent sequence $\{\boldsymbol{X}_t\}_{t \in [T]}$ is bounded. Then, for each $\boldsymbol{X}_t \in \operatorname{conv}(\mathcal{D})$ with bounded eigenvalues, there exists a bounded vector $\boldsymbol{x}_t \in \mathbb{R}_+^n$ such that $(\boldsymbol{x}_t, \boldsymbol{X}_t) \in \mathcal{T}$. We now obtain a bounded sequence $\{(\boldsymbol{x}_t, \boldsymbol{X}_t)\}_{t \in [T]}$ in set $\mathcal{T}$. According to the Bolzano-Weierstrass theorem, a bounded sequence has a convergent subsequence. Suppose that a convergent subsequence of $\{(\boldsymbol{x}_t, \boldsymbol{X}_t)\}_{t \in [T]}$ converges to $(\boldsymbol{x}^*, \widehat{\boldsymbol{X}})$. On the other hand, every subsequence of a convergent sequence converges to the same limit, implying that $\widehat{\boldsymbol{X}} = \boldsymbol{X}^*$. Given that function $f(\cdot)$ is closed in (1), set $\mathcal{T}$ is clearly closed and thus we have $(\boldsymbol{x}^*, \boldsymbol{X}^*) \in \mathcal{T}$. Since $\operatorname{conv}(\mathcal{D})$ is the projection of set $\mathcal{T}$ onto matrix space, the limit $\boldsymbol{X}^*$ must belong to $\operatorname{conv}(\mathcal{D})$. This proves the closeness. □

A.2 Proof of Lemma 1

Lemma 1 *Given two vectors $\boldsymbol{\lambda}, \boldsymbol{x} \in \mathbb{R}^n$ with their elements sorted in descending order and $\boldsymbol{x} \succeq \boldsymbol{\lambda}$, suppose that there exists an index $j_1 \in [n-1]$ such that $\sum_{i \in [j_1]} \lambda_i < \sum_{i \in [j_1]} x_i$. Then we have*

$$\sum_{i \in [j]} \lambda_i < \sum_{i \in [j]} x_i, \forall j \in [j_0, j_2 - 1],$$

where the indices j_0, j_2 satisfy $\lambda_{j_0} = \cdots = \lambda_{j_1} \geq \lambda_{j_1+1} = \cdots = \lambda_{j_2}$ with $1 \leq j_0 \leq j_1 \leq j_2 - 1 \leq n-1$.

Proof. The proof includes two parts by dividing the inequalities over $[j_0, j_2 - 1]$ into two subintervals: $[j_0, j_1]$ and $[j_1 + 1, j_2 - 1]$.

(i) If the equality can be attained for some $j^* \in [j_0, j_1 - 1]$, i.e., $\sum_{i\in[j^*]} \lambda_i = \sum_{i\in[j^*]} x_i$, given $\sum_{i\in[j_1]} \lambda_i < \sum_{i\in[j_1]} x_i$, then we have

$$\sum_{i\in[j^*+1,j_1]} \lambda_i < \sum_{i\in[j^*+1,j_1]} x_i \Longrightarrow \lambda_{j^*} = \cdots = \lambda_{j_1} < \frac{\sum_{i\in[j^*+1,j_1]} x_i}{j_1 - j^*} \le x_{j^*},$$

and thus $\sum_{i\in[j^*-1]} \lambda_i = \sum_{i\in[j^*-1]} x_i + x_{j^*} - \lambda_{j^*} > \sum_{i\in[j^*-1]} x_i$, which contradicts with $\boldsymbol{x} \succeq \boldsymbol{\lambda}$.

(ii) If the equality can be attained for some $j^* \in [j_1 + 1, j_2 - 1]$, i.e., $\sum_{i\in[j^*]} \lambda_i = \sum_{i\in[j^*]} x_i$, given $\sum_{i\in[j_1]} \lambda_i < \sum_{i\in[j_1]} x_i$, then we have

$$\sum_{i\in[j_1+1,j^*]} \lambda_i > \sum_{i\in[j_1+1,j^*]} x_i \Longrightarrow \lambda_{j_1+1} = \cdots = \lambda_{j^*+1} > \frac{\sum_{i\in[j_1+1,j^*]} x_i}{j^* - j_1} \ge x_{j^*+1},$$

and thus $\sum_{i\in[j^*+1]} \lambda_i = \sum_{i\in[j^*]} x_i + \lambda_{j^*+1} > \sum_{i\in[j^*+1]} x_i$, which contradicts with $\boldsymbol{x} \succeq \boldsymbol{\lambda}$. □

A.3 Proof of Theorem 1

Theorem 1 *Given $\mathcal{Q} := \mathcal{S}^n_+$ in domain set $\mathcal{D}$ in (1), suppose that $F^d \subseteq \operatorname{conv}(\mathcal{D})$ is a d-dimensional face of the convex hull of the domain set. Then any point in face F^d has a rank at most $\tilde{k} + \lfloor\sqrt{2d + 9/4} - 3/2\rfloor$, where $\tilde{k} \le k$ follows Definition 1.*

Proof. We use contradiction to prove this result and replace k by $\tilde{k}$. Given $\mathcal{Q} = \mathcal{S}^n_+$, suppose that point $\boldsymbol{X}^* \in \mathcal{S}^n_+$ belongs to face F^d with a rank $r > \tilde{k} + \lfloor\sqrt{2d + 9/4} - 3/2\rfloor$, i.e.,

$$d + 1 < (r - \tilde{k} + 1)(r - \tilde{k} + 2)/2. \tag{17}$$

Then we let $\lambda_1 \ge \cdots \ge \lambda_r > 0 = \lambda_{r+1} = \cdots = \lambda_n$ denote the eigenvalues of $\boldsymbol{X}^*$ and form the eigenvalue vector $\boldsymbol{\lambda} \in \mathbb{R}^n_+$. Since $\lambda_{r+1} = \cdots = \lambda_n = 0$, we can rewrite $\boldsymbol{X}^*$ as $\boldsymbol{X}^* = \boldsymbol{Q}_1 \boldsymbol{\Lambda}_1 \boldsymbol{Q}_1^\top + \boldsymbol{Q}_2 \boldsymbol{\Lambda}_2 \boldsymbol{Q}_2^\top$, where

$$\boldsymbol{\Lambda}_1 = \operatorname{Diag}(\lambda_1, \cdots, \lambda_{\tilde{k}-1}) \in \mathcal{S}^{\tilde{k}-1}_{++}, \quad \boldsymbol{\Lambda}_2 = \operatorname{Diag}(\lambda_{\tilde{k}}, \cdots, \lambda_r) \in \mathcal{S}^{r-\tilde{k}+1}_{++},$$

and $\boldsymbol{Q}_1 \in \mathbb{R}^{n\times\tilde{k}-1}$, $\boldsymbol{Q}_2 \in \mathbb{R}^{n\times r-\tilde{k}+1}$ are corresponding eigenvector matrices.

For any d-dimensional face F^d with $d \ge 1$, there are $(d + 1)$ points such that $F^d \subseteq \operatorname{aff}(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_{d+1})$, where $\operatorname{aff}(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_{d+1}) = \{\sum_{i\in[d+1]} \alpha_i \boldsymbol{X}_i : \boldsymbol{\alpha} \in \mathbb{R}^{d+1}, \sum_{i\in[d+1]} \alpha_i = 1\}$ denotes the affine hull of these points. Note that any size–$n \times n$ symmetric matrix can be recast into a vector of length $n(n + 1)/2$. By inequality (17), there is a nonzero symmetric matrix $\boldsymbol{\Delta} \in \mathcal{S}^{r-\tilde{k}+1}$ satisfying $(d + 1)$ equations below

$$\langle \boldsymbol{Q}_2^\top (\boldsymbol{X}_1 - \boldsymbol{X}_i) \boldsymbol{Q}_2, \boldsymbol{\Delta} \rangle = 0, \forall i \in [d], \quad \langle \boldsymbol{Q}_2^\top \boldsymbol{Q}_2, \boldsymbol{\Delta} \rangle = \operatorname{tr}(\boldsymbol{\Delta}) = 0, \tag{18}$$

where the first d equations indicate that matrix $\boldsymbol{Q}_2 \boldsymbol{\Delta} \boldsymbol{Q}_2^\top$ is orthogonal to the face F^d, i.e., $\boldsymbol{Q}_2 \boldsymbol{\Delta} \boldsymbol{Q}_2^\top \perp F^d$.

Then let us construct the two matrices $\boldsymbol{X}^+(\delta)\in\mathcal{S}_+^n$ and $\boldsymbol{X}^-(\delta)\in\mathcal{S}_+^n$ as below

$$\boldsymbol{X}^+(\delta)=\boldsymbol{X}^*+\delta\boldsymbol{Q}_2\boldsymbol{\Delta}\boldsymbol{Q}_2^\top=\boldsymbol{Q}_1\boldsymbol{\Lambda}_1\boldsymbol{Q}_1^\top+\boldsymbol{Q}_2(\boldsymbol{\Lambda}_2+\delta\boldsymbol{\Delta})\boldsymbol{Q}_2^\top,$$
$$\boldsymbol{X}^-(\delta)=\boldsymbol{X}^*-\delta\boldsymbol{Q}_2\boldsymbol{\Delta}\boldsymbol{Q}_2^\top=\boldsymbol{Q}_1\boldsymbol{\Lambda}_1\boldsymbol{Q}_1^\top+\boldsymbol{Q}_2(\boldsymbol{\Lambda}_2-\delta\boldsymbol{\Delta})\boldsymbol{Q}_2^\top,$$

for any $\delta>0$. Thus, $\langle\boldsymbol{A}_i,\boldsymbol{X}^+(\delta)\rangle=\langle\boldsymbol{A}_i,\boldsymbol{X}^-(\delta)\rangle=\langle\boldsymbol{A}_i,\boldsymbol{X}^*\rangle$ for all $i\in[m]$ and $\operatorname{tr}(\boldsymbol{X}^+(\delta))=\operatorname{tr}(\boldsymbol{X}^-(\delta))=\operatorname{tr}(\boldsymbol{X}^*)$ based on equations (18). We further show that when $\delta>0$ is small enough, $\boldsymbol{X}^+(\delta),\boldsymbol{X}^-(\delta)$ also belong to set $\operatorname{conv}(\mathcal{D})$ as detailed by the claim below. If so, the equation $\boldsymbol{X}^*=\frac{1}{2}\boldsymbol{X}^+(\delta)+\frac{1}{2}\boldsymbol{X}^-(\delta)$ implies that $\boldsymbol{X}^+(\delta),\boldsymbol{X}^-(\delta)\in F^d$ according to the definition of face F^d. A contradiction with fact that $\boldsymbol{Q}_2\boldsymbol{\Delta}\boldsymbol{Q}_2^\top\perp F^d$. Therefore, the point $\boldsymbol{X}^*$ must have a rank no larger than $\tilde{k}+\lfloor\sqrt{2d+9/4}-3/2\rfloor$. Indeed, we can show that

Claim 1 *There is a scalar $\underline{\delta}>0$ such that for any $\delta\in(0,\underline{\delta})$, we have $\boldsymbol{X}^+(\delta),\boldsymbol{X}^-(\delta)\in\operatorname{conv}(\mathcal{D})$.*

Proof. First, according to the characterization of $\operatorname{conv}(\mathcal{D})$ in Proposition 1 and $\boldsymbol{X}^*\in\operatorname{conv}(\mathcal{D})$, there is a vector $\boldsymbol{x}^*\in\mathbb{R}_+^n$ such that $\boldsymbol{\lambda}\preceq\boldsymbol{x}^*$. Let $\boldsymbol{\lambda}^+(\delta)\in\mathbb{R}_+^n$ and $\boldsymbol{\lambda}^-(\delta)\in\mathbb{R}_+^n$ denote the eigenvalue vectors of $\boldsymbol{X}^+(\delta)$ and $\boldsymbol{X}^-(\delta)$, respectively, provided that $\delta>0$ is small enough to ensure $\boldsymbol{X}^+(\delta),\boldsymbol{X}^-(\delta)\in\mathcal{S}_+^n$. According to Proposition 1, $\boldsymbol{X}^+(\delta),\boldsymbol{X}^-(\delta)\in\operatorname{conv}(\mathcal{D})$ holds if $\boldsymbol{\lambda}^+(\delta),\boldsymbol{\lambda}^-(\delta)\preceq\boldsymbol{x}^*$. Next, we will show that if the scalar $\delta>0$ only imposes slight perturbation on eigenvalue vector $\boldsymbol{\lambda}$ of matrix $\boldsymbol{X}^*$, then $\boldsymbol{\lambda}^+(\delta),\boldsymbol{\lambda}^-(\delta)\preceq\boldsymbol{x}^*$ still holds given $\boldsymbol{\lambda}\preceq\boldsymbol{x}^*$. Let us first analyze the perturbation from $\boldsymbol{\lambda}$ to $\boldsymbol{\lambda}^+(\delta)$ caused by scalar δ.

Specifically, according to Proposition 1, a pair $(\boldsymbol{X}^*,\boldsymbol{x}^*)$ satisfies $x_1^*\geq\cdots\geq x_{\tilde{k}}^*\geq 0=x_{\tilde{k}+1}^*=\cdots=x_n^*$, $\|\boldsymbol{X}^*\|_{(\ell)}\leq\sum_{i\in[\ell]}x_i^*$ for each $\ell\in[\tilde{k}-1]$ and $\operatorname{tr}(\boldsymbol{X}^*)=\sum_{i\in[\tilde{k}]}x_i^*$. Thus, we have

$$\sum_{i\in[\ell]}\lambda_i=\|\boldsymbol{X}^*\|_{(\ell)}<\|\boldsymbol{X}^*\|_{(r)}=\operatorname{tr}(\boldsymbol{X}^*)=\sum_{i\in[\tilde{k}]}x_i^*=\sum_{i\in[\ell]}x_i^*,\quad\forall\ell\in[\tilde{k},r-1],$$

where the first inequality and the second equality are from the fact that $\operatorname{rank}(\boldsymbol{X}^*)=r$ and $\boldsymbol{X}^*\in\mathcal{S}_+^n$, and the last equality is due to $x_{\tilde{k}+1}^*=\cdots=x_n^*=0$. Suppose that $\lambda_0=\infty$ and $\lambda_{j_0-1}>\lambda_{j_0}=\cdots=\lambda_{\tilde{k}}$ for some $j_0\in[\tilde{k}]$. The according to Lemma 1 with $j_1=\tilde{k}$, we have

$$\lambda_{j_0-1}>\lambda_{j_0},\quad\sum_{i\in[\ell]}\lambda_i<\sum_{i\in[\ell]}x_i^*,\forall\ell\in[j_0,\tilde{k}].\tag{19}$$

Then we define a constant $c^*>0$ by

$$c^*:=\min\left\{\lambda_{j_0-1}-\lambda_{j_0},\min_{\ell\in[j_0,\tilde{k}]}\left(\sum_{i\in[\ell]}x_i^*-\sum_{i\in[\ell]}\lambda_i\right)\right\}.$$

Suppose that matrix $\boldsymbol{\Lambda}_2+\delta\boldsymbol{\Delta}\in\mathcal{S}_+^{r-\tilde{k}+1}$ admits eigenvector matrix $\boldsymbol{P}\in\mathbb{R}^{(r-\tilde{k}+1)\times(r-\tilde{k}+1)}$. Then it is easy to check that $\begin{pmatrix}\boldsymbol{Q}_1^\top\\ \boldsymbol{P}^\top\boldsymbol{Q}_2^\top\end{pmatrix}(\boldsymbol{Q}_1\ \boldsymbol{Q}_2\boldsymbol{P})=\boldsymbol{I}_r$, since matrix $(\boldsymbol{Q}_1\ \boldsymbol{Q}_2)\in\mathbb{R}^{n\times r}$ is orthonormal.

It follows that the nonzero eigenvalues of matrix $\boldsymbol{X}^+(\delta)$ contain those of $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2+\delta\boldsymbol{\Delta}$. That is, $\{\lambda_i\}_{i\in[\tilde{k}-1]}$ are also eigenvalues of matrix $\boldsymbol{X}^+(\delta)$. Therefore, the perturbation from $\boldsymbol{\lambda}$ to $\boldsymbol{\lambda}^+(\delta)$ is captured by the impact of matrix $\delta\boldsymbol{\Delta}$ on $\boldsymbol{\Lambda}_2$. We let $\lambda_{\max}$ denote the largest singular value of matrix $\boldsymbol{\Delta}$, according to Weyl's inequality, the eigenvalue vector $\boldsymbol{\lambda}^+(\delta)$ of matrix $\boldsymbol{X}^+(\delta)$ satisfies

$$\boldsymbol{\lambda}^+(\delta)\leq\left(\lambda_1,\cdots,\lambda_{\tilde{k}-1},\lambda_{\tilde{k}}+\delta\lambda_{\max},\cdots,\lambda_r+\delta\lambda_{\max},0,\cdots,0\right), \tag{20}$$

where note that the eigenvalue vector $\boldsymbol{\lambda}^+(\delta)$ above may not follow a descending order and the inequality is the component-wise convention for vectors. Without loss of generality, we let $\lambda_1^+(\delta)\geq\cdots\geq\lambda_n^+(\delta)\geq 0$ denote the eigenvalues of $\boldsymbol{X}^+(\delta)$, i.e., a permutation of elements in vector $\boldsymbol{\lambda}^+(\delta)$ by a descending order.

By letting the scalar $\underline{\delta}\leq\frac{c^*}{(r-\tilde{k}+1)\lambda_{\max}}$, after perturbation of eigenvalue vector $\boldsymbol{\lambda}^+(\delta)$ by any $\delta\in(0,\underline{\delta})$, we still have that $\boldsymbol{x}^*\succeq\boldsymbol{\lambda}^+(\delta)$ as shown follows.

(a) We show that $\lambda_i^+(\delta)=\lambda_i$ for any $i\in[j_0-1]$. If not, there must exist $\lambda_\ell^+(\delta)>\lambda_\ell$ for some $\ell\in[j_0-1]$, which means that $\lambda_\ell^+\notin\{\lambda_i\}_{i\in[\tilde{k}-1]}$ as $\{\lambda_i\}_{i\in[\tilde{k}-1]}$ are eigenvalues of matrix $\boldsymbol{X}^+(\delta)$. According to the inequality (20), $\lambda_\ell^+(\delta)$ must satisfy $\lambda_\ell^+(\delta)\leq\lambda_{\tilde{k}}+\delta\lambda_{\max}\leq\lambda_{\tilde{k}}+\lambda_{j_0-1}-\lambda_{j_0}\leq\lambda_{j_0-1}\leq\lambda_\ell$, where the second inequality is due to $\delta\lambda_{\max}\leq c^*/(r-\tilde{k}+1)$ and the third one is from $j_0\leq\tilde{k}$. Therefore, $\lambda_i^+(\delta)=\lambda_i$ must hold for any $i\in[j_0-1]$ and we have $\sum_{i\in[\ell]}\lambda_i^+(\delta)=\sum_{i\in[\ell]}\lambda_i\leq\sum_{i\in[\ell]}x_i^*$ for all $\ell\in[j_0-1]$.

(b) According to the inequality (20), there are at most $(r-\tilde{k}+1)$ entries in $\boldsymbol{\lambda}^+(\delta)$ that go beyond those of $\boldsymbol{\lambda}$ up to $\delta\lambda_{\max}$. Hence, we have

$$\sum_{i\in[\ell]}\lambda_i^+(\delta)\leq\sum_{i\in[\ell]}\lambda_i+(r-\tilde{k}+1)\delta\lambda_{\max}\leq\sum_{i\in[\ell]}\lambda_i+c^*\leq\sum_{i\in[\ell]}x_i^*,\forall\ell\in[j_0,\tilde{k}],$$

where the second inequality is from the definition of constant c^*.

(c) Since $\operatorname{tr}(\boldsymbol{X}^+(\delta))=\operatorname{tr}(\boldsymbol{X}^*)=\sum_{i\in[\tilde{k}]}\boldsymbol{x}^*$, we have $\sum_{i\in[\ell]}\lambda_i^+(\delta)\leq\operatorname{tr}(\boldsymbol{X}^+(\delta))=\sum_{i\in[\tilde{k}]}\boldsymbol{x}^*=\sum_{i\in[\ell]}\boldsymbol{x}^*$ for each $\ell\in[\tilde{k}+1,n]$.

Similarly, there exists a scalar $\underline{\delta}^*>0$ such that $\boldsymbol{x}^*\succeq\boldsymbol{\lambda}^-(\delta)$ for any $\delta\in(0,\underline{\delta}^*)$. Letting $\underline{\delta}:=\min\{\underline{\delta}^*,\underline{\delta}\}$, we complete the proof of the rank bound. ◇

□

A.4 Proof of Corollary 2

Corollary 2 (Fair PCA) *There exists an optimal solution of the Fair PCA-R with rank at most* $k+\left\lfloor\sqrt{2m+1/4}-3/2\right\rfloor$. *Besides, if there are $m\leq 2$ sample covariance matrices, the Fair PCA-R achieves the same optimal value as Fair PCA.*

Proof. It is easy to check that the domain set $\mathcal{D}$ in Fair PCA satisfies $\tilde{k}=k$. According to Li and Xie (2022), at most $(m-1)$ linear inequalities are binding at any extreme points in the feasible set of Fair PCA-R. Besides, the optimal value of Fair PCA-R must be finite due to the boundedness of set $\operatorname{conv}(\mathcal{D})$. Hence, in this example, the rank bound in Theorem 2 reduces to $k+\left\lfloor\sqrt{2m+1/4}-3/2\right\rfloor$. When $m\le 2$, the rank bound reaches k, and thus an optimal extreme point of Fair PCA-R coincides with that of the original Fair PCA. □

A.5 Proof of Corollary 3

Corollary 3 (QCQP) *There exists an optimal solution of QCQP-R with rank at most* $1+\lfloor 2\tilde{m}+1/4-3/2\rfloor$ *if* $\mathbf{V}_{\mathrm{rel}}>-\infty$. *Besides, if there are* $\tilde{m}\le 2$ *linearly independent inequalities, the QCQP-R achieves the same optimal value as the original QCQP if* $\mathbf{V}_{\mathrm{rel}}>-\infty$.

Proof. Given a domain set $\mathcal{D}:=\{\boldsymbol{X}\in\mathcal{S}_+^n:\operatorname{rank}(\boldsymbol{X})\le 1\}$, we have that $F(\boldsymbol{X})=0$ in (1) and $\tilde{k}=1$ by Definition 1. According to the characterization of $\operatorname{conv}(\mathcal{D})$ in Proposition 1, we have

$$\operatorname{conv}(\mathcal{D})=\left\{\boldsymbol{X}\in\mathcal{S}_+^n:\exists\boldsymbol{x}\in\mathbb{R}_+^n, x_1\ge\cdots\ge x_n, x_{k+1}=0, \|\boldsymbol{X}\|_{(\ell)}\le\sum_{i\in[\ell]}x_i, \forall\ell\in[k], \operatorname{tr}(\boldsymbol{X})=\sum_{i\in[k]}x_i\right\}.$$

Note that the vector $\boldsymbol{x}$ can be unbounded. Hence, in the proof of Theorem 1, it suffices to construct a symmetric matrix $\boldsymbol{\Delta}\in\mathcal{S}^r$ to satisfy the first d equations in (18) since we can arbitrarily adjust the vector $\boldsymbol{x}^*$ therein. When $\mathbf{V}_{\mathrm{rel}}>-\infty$, this observation leads to a better rank bound $r(r+1)/2\le m$, where r denotes the rank of an extreme point to QCQP-R. When $\tilde{m}\le 2$, the equivalence between QCQP and QCQP-R is directly from Proposition 2. □

A.6 Proof of Theorem 3

Before proving Theorem 3, let us give an explicit characterization of the convex hull of domain set $\mathcal{D}$ in the non-symmetric matrix space.

Proposition 7 *For a domain set* $\mathcal{D}$ *in non-symmetric matrix space, i.e.,* $\mathcal{Q}:=\mathbb{R}^{n\times p}$ *in* (1), *its convex hull* $\operatorname{conv}(\mathcal{D})$ *is equal to*

$$\left\{\boldsymbol{X}\in\mathbb{R}^{n\times p}:\exists\boldsymbol{x}\in\mathbb{R}_+^n, f(\boldsymbol{x})\le 0, x_1\ge\cdots\ge x_n, x_{k+1}=0, \|\boldsymbol{X}\|_{(\ell)}\le\sum_{i\in[\ell]}x_i, \forall\ell\in[k-1], \|\boldsymbol{X}\|_*\le\sum_{i\in[k]}x_i\right\}$$

and is a closed set.

Proof. The derivation of $\operatorname{conv}(\mathcal{D})$ has been shown in Kim et al. (2022)[theorem 7]. The closeness proof of set $\operatorname{conv}(\mathcal{D})$ follows that of Proposition 1. □

Theorem 3 *Given $\mathcal{Q} := \mathbb{R}^{n\times p}$ in domain set $\mathcal{D}$ in (1), suppose that $F^d \subseteq \operatorname{conv}(\mathcal{D})$ is a d-dimensional face in the convex hull of the domain set $\mathcal{D}$. Then any point in face F^d has a rank at most $\tilde{k} + \lfloor\sqrt{2d+9/4}-3/2\rfloor$, where $\tilde{k} \le k$ follows Definition 1.*

Proof. Suppose that there is a non-symmetric matrix $\boldsymbol{X}^* \in F^d$ with rank $r > \tilde{k}+\lfloor\sqrt{2d+9/4}-3/2\rfloor$. Then, we denote the singular value decomposition of matrix $\boldsymbol{X}^*$ by $\boldsymbol{X}^* = \boldsymbol{Q}_1\boldsymbol{\Lambda}_1\boldsymbol{P}_1^\top + \boldsymbol{Q}_2\boldsymbol{\Lambda}_2\boldsymbol{P}_2^\top$ and denote its singular values by $\lambda_1 \ge \cdots \ge \lambda_r > 0 = \lambda_{r+1} = \cdots = \lambda_n$, where

$$\boldsymbol{\Lambda}_1 := \operatorname{Diag}(\lambda_1, \cdots, \lambda_{\tilde{k}-1}) \in \mathcal{S}_{++}^{\tilde{k}-1}, \quad \boldsymbol{\Lambda}_2 := \operatorname{Diag}(\lambda_{\tilde{k}}, \cdots, \lambda_r) \in \mathcal{S}_{++}^{r-\tilde{k}+1},$$

where $\boldsymbol{Q}_1, \boldsymbol{Q}_2$ and $\boldsymbol{P}_1, \boldsymbol{P}_2$ are corresponding left and right singular vectors.

Following the proof of Theorem 1 and using matrices $\{\boldsymbol{X}_i\}_{i\in[d+1]}$ therein, we can construct a nonzero symmetric matrix $\boldsymbol{\Delta} \in \mathcal{S}^{r-\tilde{k}+1}$ satisfying the equations below

$$\langle \boldsymbol{P}_2^\top(\boldsymbol{X}_i - \boldsymbol{X}_{d+1})\boldsymbol{Q}_2, \boldsymbol{\Delta}\rangle = 0, \forall i \in [d], \quad \operatorname{tr}(\boldsymbol{\Delta}) = 0.$$

It should be noted that according to Proposition 7, the description of convex hull set $\operatorname{conv}(\mathcal{D})$ with $\mathcal{Q} = \mathbb{R}^{n\times p}$ relies on the nuclear norm of a non-symmetric matrix.

Next, we define two matrices $\boldsymbol{X}^{\pm}(\delta) = \boldsymbol{X}^* \pm \delta\boldsymbol{Q}_2\boldsymbol{\Delta}\boldsymbol{P}_2^\top$, where $\delta > 0$ is a small scalar such that matrices $\begin{pmatrix}\boldsymbol{\Lambda}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Lambda}_2 \pm \delta\boldsymbol{\Delta}\end{pmatrix} \in \mathcal{S}_+^r$ are positive semidefinite. Hence, we can show that matrix $\boldsymbol{X}^+(\delta)$ has the same nonzero singular value vector as that of positive semidefinite matrix $\begin{pmatrix}\boldsymbol{\Lambda}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Lambda}_2 + \delta\boldsymbol{\Delta}\end{pmatrix}$, which means that

$$\|\boldsymbol{X}^+(\delta)\|_* = \left\|\begin{pmatrix}\boldsymbol{\Lambda}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Lambda}_2 + \delta\boldsymbol{\Delta}\end{pmatrix}\right\|_* = \operatorname{tr}(\boldsymbol{\Lambda}_1) + \operatorname{tr}(\boldsymbol{\Lambda}_2 + \delta\boldsymbol{\Delta}) = \operatorname{tr}(\boldsymbol{\Lambda}_1) + \operatorname{tr}(\boldsymbol{\Lambda}_2) = \|\boldsymbol{X}^*\|_*,$$

where the second equation is because for any positive semidefinite matrix, the eigenvalues meet singular values. Analogously, the equation $\|\boldsymbol{X}^-(\delta)\|_* = \|\boldsymbol{X}^*\|_*$ holds for matrix $\boldsymbol{X}^-(\delta)$. Then, the existence of such a symmetric matrix $\boldsymbol{\Delta}$ can form a contradiction as shown in Theorem 1, leading to the desired rank bound $d+1 \ge (r-\tilde{k}+1)(r-\tilde{k}+2)/2$. ☐

A.7 Proof of Corollary 4

Corollary 4 (Fair SVD) *There exists an optimal solution of Fair SVD-R with rank at most $k+\lfloor\sqrt{2m+1/4}-3/2\rfloor$. Besides, if there are $m \le 2$ groups of data matrices, the Fair SVD-R yields the same optimal value as the original Fair SVD.*

Proof. Analogous to Fair PCA-R, there are at most $(m-1)$ linear equations for all extreme points in the feasible set of Fair SVD-R. According to Theorem 4, any optimal extreme point of its Fair SVD-R has a rank at most $k + \left\lfloor\sqrt{2m+1/4}-3/2\right\rfloor$. When $m \le 2$, the rank bound becomes k and thus the Fair SVD-R can yield a rank-k solution that is also optimal to the original Fair SVD. ☐

A.8 Proof of Corollary 5

Corollary 5 (Matrix Completion) *Suppose that there are $m := |\Omega|$ observed entries in Matrix Completion. Then there exists an optimal solution to Matrix Completion-R with rank at most $\lfloor\sqrt{2|\Omega|+9/4}-1/2\rfloor$. Besides, if there are $|\Omega| \le (k+1)(k+2)/2-2$ observed entries, the Matrix Completion-R can achieve the same optimal value as the original Matrix Completion.*

Proof. Let $(\widehat{z}, \widehat{\boldsymbol{X}})$ denote an optimal solution to the Matrix Completion-R. Then given an optimal solution $\widehat{z}$, the compact set below can be viewed as an optimal set of variable $\boldsymbol{X}$ to Matrix Completion-R

$$\left\{\boldsymbol{X} \in \operatorname{conv}(\mathcal{D}) : X_{ij} = \widehat{X}_{ij}, \forall (i,j) \in \Omega\right\},$$

where now we have a domain set $\mathcal{D} := \{\boldsymbol{X} \in \mathbb{R}^{n\times p} : \operatorname{rank}(\boldsymbol{X}) \le k, \|\boldsymbol{X}\|_* \le \widehat{z}\}$.

It is seen that the domain set $\mathcal{D}$ above can be viewed as a special case of that in Example 1; hence, we now have $\tilde{k} = 1 \le k$ by Definition 1. Since there are $m = |\Omega|$ linear inequalities, according to Theorem 4, any optimal extreme point $\boldsymbol{X}^*$ has a rank at most $1 + \lfloor\sqrt{|\Omega|+9/4}-3/2\rfloor$. When $\lfloor\sqrt{|\Omega|+9/4}-1/2\rfloor \le k$, the Matrix Completion-R can achieve the desired rank-k solution as the original Matrix Completion based on Proposition 3. □

A.9 Proof of Theorem 5 and Its Implication of LSOP-R Exactness

Theorem 5 *Given $\mathcal{Q} := \mathcal{S}^n$ in (1) and integer $\tilde{k} \le k$ following Definition 1, suppose that the function $f(\cdot)$ in (1) is sign-invariant and LSOP-R admits a line-free feasible set, then we have that*

(i) Each feasible extreme point in the LSOP-R has a rank at most $\tilde{k} + \lfloor\sqrt{2\tilde{m}+9/4}-3/2\rfloor$; and

(ii) There is an optimal solution to LSOP-R of rank at most $\tilde{k} + \lfloor\sqrt{2\tilde{m}+9/4}-3/2\rfloor$ if the LSOP-R yields a finite optimal value, i.e., $\mathbf{V}_{\text{rel}} > -\infty$.

Proof. Since the domain set $\mathcal{D}$ is sign-invariant, according to Proposition 7, its convex hull $\operatorname{conv}(\mathcal{D})$ is equal to

$$\left\{\boldsymbol{X} \in \mathcal{Q} : \exists \boldsymbol{x} \in \mathbb{R}^n_+, f(\boldsymbol{x}) \le 0, x_1 \ge \cdots \ge x_n, x_{k+1} = 0, \|\boldsymbol{X}\|_{(\ell)} \le \sum_{i\in[\ell]} x_i, \forall \ell \in [k-1], \|\boldsymbol{X}\|_* \le \sum_{i\in[k]} x_i\right\}.$$

Then, the result in Theorem 1 can be readily extended to set $\operatorname{conv}(\mathcal{D})$ with the symmetric indefinite matrix space, except considering the singular value decomposition rather than eigen-decomposition. That is, any d-dimensional face in set $\operatorname{conv}(\mathcal{D})$ satisfies the rank-$\tilde{k} + \lfloor\sqrt{2d+9/4}-3/2\rfloor$ constraint. Using Lemma 2, the remaining proof is identical to that of Theorem 2 and thus is omitted. □

Notable, Theorem 5 implies that by letting $\tilde{k} + \lfloor\sqrt{2d+9/4}-3/2\rfloor \le k$, on can obtain a sufficient condition under which the LSOP-R coincides with the LSOP.

Proposition 8 *Given $\mathcal{Q} := \mathcal{S}^n$ in (1) and integer $\tilde{k} \leq k$ following Definition 1, suppose that the function $f(\cdot)$ in the domain set $\mathcal{D}$ is sign-invariant and the LSOP-R admits a line-free feasible set. Then, if $\tilde{m} \leq (k-\tilde{k}+2)(k-\tilde{k}+3)/2-2$ holds, we have that*

(i) Each feasible extreme point in LSOP-R has a rank at most k; and

(ii) The LSOP-R achieves the same optimal value as the original LSOP, i.e., $\mathbf{V}_{\text{opt}} = \mathbf{V}_{\text{rel}}$ if the LSOP-R yields a finite optimal value, i.e., $\mathbf{V}_{\text{rel}} > -\infty$.

Proof. Using Theorem 5, the proof is identical to that of Proposition 2 and thus is omitted. □

A.10 Proof of Proposition 4

Proposition 4 *For a domain set $\mathcal{D}$ with $\mathcal{Q} := \mathcal{S}^n$ and an integer $\tilde{k} \leq k$ in Definition 1, we have*

(i) $\operatorname{conv}(\mathcal{D}) = \operatorname{conv}\left(\cup_{s\in[\tilde{k}+1]}\mathcal{Y}^s\right)$; and

(ii) $\mathcal{D} \subseteq \cup_{s\in[\tilde{k}+1]}\mathcal{Y}^s$,

where for each $s \in [\tilde{k}+1]$, set $\mathcal{Y}^s$ is closed and is defined as

$$\mathcal{Y}^s := \left\{ \boldsymbol{X} \in \mathcal{S}^n : \exists \boldsymbol{x} \in \mathbb{R}^n, f(\boldsymbol{x}) \leq 0, x_1 \geq \cdots \geq x_n, x_s = x_{s+n-\tilde{k}-1} = 0, \boldsymbol{x} \succeq \boldsymbol{\lambda}(\boldsymbol{X}) \right\}. \tag{2}$$

Note that $\boldsymbol{\lambda}(\boldsymbol{X}) \in \mathbb{R}^n$ denotes the eigenvalue vector of the symmetric matrix $\boldsymbol{X} \in \mathcal{S}^n$.

Proof. By Definition 1, we will replace k by $\tilde{k}$ in the below. According to Kim et al. (2022)[theorem 4], we have that $\operatorname{conv}(\mathcal{D}) = \operatorname{proj}_{\boldsymbol{X}}(\mathcal{T})$, where $\mathcal{T} := \{(\boldsymbol{X}, \boldsymbol{x}) \in \mathcal{S}^n \times \mathbb{R}^n : \boldsymbol{x} \in \operatorname{conv}(\tilde{\mathcal{D}} \cap \{x_1 \geq \cdots \geq x_n\}), \boldsymbol{x} \succeq \boldsymbol{\lambda}(\boldsymbol{X})\}$ and $\tilde{\mathcal{D}} := \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_0 \leq \tilde{k}, f(\boldsymbol{x}) \leq 0\}$.

Next, let us consider the key inner convex hull over $\boldsymbol{x}$ above, i.e., $\operatorname{conv}(\tilde{\mathcal{D}} \cap \{x_1 \geq \cdots \geq x_n\})$, where there may exist both positive and negative elements in $\boldsymbol{x}$. Given $x_1 \geq x_2 \geq \ldots \geq x_n$, $x_{i_1} = \cdots = x_{i_2} = 0$ must hold whenever $x_{i_1} = x_{i_2} = 0$ with $1 \leq i_1 \leq i_2 \leq n$ and $i_2 - i_1 \geq n - \tilde{k} - 1$. Besides, the constraint $\|\boldsymbol{x}\|_0 \leq \tilde{k}$ implies that there are at least $(n-\tilde{k})$ zero entries in the vector $\boldsymbol{x}$. Therefore, to remove the constraint $\|\boldsymbol{x}\|_0 \leq \tilde{k}$, we can split set $\tilde{\mathcal{D}} \cap \{x_1 \geq \cdots \geq x_n\}$ into $(\tilde{k}+1)$ subsets, i.e., $\cup_{s\in[\tilde{k}+1]}\tilde{\mathcal{D}}^s$ depending on where zero entries of $\boldsymbol{x}$ are located, where for each $s \in [\tilde{k}+1]$, set $\tilde{\mathcal{D}}^s$ is defined by

$$\tilde{\mathcal{D}}^s := \left\{ \boldsymbol{x} \in \mathbb{R}^n : f(\boldsymbol{x}) \leq 0, x_1 \geq \cdots \geq x_n, x_s = x_{s+n-\tilde{k}-1} = 0 \right\}.$$

Finally, plugging $\cup_{s\in[\tilde{k}+1]}\tilde{\mathcal{D}}^s$ into set $\mathcal{T}$ and moving the convex hull operation outside, we have $\operatorname{conv}(\mathcal{D}) = \operatorname{proj}_{\boldsymbol{X}}(\mathcal{T}) \supseteq \operatorname{conv}(\cup_{s\in[\tilde{k}+1]}\mathcal{Y}^s)$. Note that the convex hull and projection operators can interchange. On the other hand, for a matrix $\boldsymbol{X} \in \mathcal{D}$, suppose its eigenvalue vector $\boldsymbol{\lambda}(\boldsymbol{X})$ is sorted in descending order. Then we see that $\boldsymbol{X} \in \cup_{s\in[\tilde{k}+1]}\mathcal{Y}^s$. Thus, we conclude that $\operatorname{conv}(\mathcal{D}) = \operatorname{conv}(\cup_{s\in[\tilde{k}+1]}\mathcal{Y}^s)$. Besides, we have $\mathcal{D} \subseteq \cup_{s\in[\tilde{k}+1]}\mathcal{Y}^s$. Following the same proof of Proposition 1, each set $\mathcal{Y}^s$ must be closed for any $s \in [\tilde{k}+1]$. □

A.11 Proof of Theorem 6

Theorem 6 *Given $\mathcal{Q} := \mathcal{S}^n$ in (1), for each $s \in [\tilde{k}+1]$, suppose that $F^d \subseteq \mathcal{Y}^s$ is a d-dimensional face of the closed convex set $\mathcal{Y}^s$ defined in (2), then any point in face F^d has rank at most*

$$\begin{cases} \tilde{k} + \lfloor \sqrt{2d+9/4} \rfloor - 3/2, & \text{if } s \in \{1, \tilde{k}+1\}; \\ \tilde{k} + \lfloor \sqrt{4d+9} \rfloor - 3, & \text{otherwise.} \end{cases},$$

where $\tilde{k} \leq k$ follows Definition 1.

Proof. We prove the rank bounds by contradiction. There are two cases to be discussed depending on the value of $s \in [\tilde{k}+1]$.

Case I. Suppose $s = 1$ or $s = \tilde{k}+1$. Then for any matrix $\boldsymbol{X}^* \in \mathcal{Y}^s \subseteq \mathcal{S}^n$, its eigenvalue vector is either all nonpositive or all nonnegative as established in Proposition 4. Following the proof of Theorem 1, we arrive at the same rank bound.

Case II. Suppose $2 \leq s \leq \tilde{k}$. Then for any point $\boldsymbol{X}^* \in F^d$ with rank $r > \tilde{k}$. Without loss of generality, we let $\boldsymbol{X}^* = \boldsymbol{Q} \operatorname{Diag}(\boldsymbol{\lambda}) \boldsymbol{Q}^\top$ be its eigen-decomposition and denote the eigenvalues by $\lambda_1 \geq \cdots \geq \lambda_{d_1-1} > 0 = \lambda_{d_1} = \cdots = \lambda_{d_2-1} = 0 > \lambda_{d_2} \geq \cdots \geq \lambda_n$ with $1 \leq d_1 \leq d_2 \leq n$. When $d_1 = 1$, all eigenvalues of matrix $\boldsymbol{X}^*$ are nonpositive and the proof of Theorem 1 simply follows. Next, let us consider $2 \leq d_1 \leq d_2 \leq n$. Since there are only r nonzero eigenvalues in matrix $\boldsymbol{X}^* \in \mathcal{S}^n$, we have

$$d_2 - d_1 = n - r.$$

In addition, the eigenvalue vector of matrix $\boldsymbol{X}^*$ corresponds to a vector $\boldsymbol{x}^* \in \mathbb{R}^n$ such that $(\boldsymbol{X}^*, \boldsymbol{x}^*) \in \mathcal{Y}^s$, and according to the definition of set $\mathcal{Y}^s$ (2) in Proposition 4, we have

$$\begin{aligned} &\boldsymbol{x}^* = [x_1^*, \cdots, x_{s-1}^*, 0, \cdots, 0, x_{s+n-\tilde{k}}^*, \cdots, x_n^*]^\top, \quad x_1^* \geq \cdots \geq x_n^*, \\ &\boldsymbol{\lambda} = [\lambda_1, \cdots, \lambda_{d_1-1}, 0, \cdots, 0, \lambda_{d_2}, \cdots, \lambda_n]^\top, \quad \lambda_1 \geq \cdots \geq \lambda_n, \\ &\sum_{i \in [j]} \lambda_i \leq \sum_{i \in [j]} x_i^*, \forall j \in [n-1], \quad \sum_{i \in [n]} \lambda_i = \sum_{i \in [n]} x_i^*. \end{aligned} \tag{21}$$

Note that there must exist both positive and negative entries in vector $\boldsymbol{x}^*$. Then, we claim that

$$\sum_{i \in [d_1-1]} \lambda_i = \sum_{i \in [d_2-1]} \lambda_i \leq \sum_{i \in [s-1]} x_i^*. \tag{22}$$

Otherwise we have $\sum_{i \in [d_1-1]} x_i^* \geq \sum_{i \in [d_1-1]} \lambda_i > \sum_{i \in [s-1]} x_i^*$, which contradicts with the fact that $\max_{T \subseteq [n]} \sum_{i \in T} x_i^* = \sum_{i \in [s-1]} x_i^*$. Then there are three parts to be discussed depending on the relation among two pairs: (s, d_1) and $(d_2, s+n-\tilde{k})$.

(i) Suppose $2 \leq s < d_1 \leq d_2 < s+n-\tilde{k} \leq n$. Then according to the inequality (22), we have

$$\begin{aligned} &\sum_{i \in [\ell]} \lambda_i < \sum_{i \in [d_1-1]} \lambda_i \leq \sum_{i \in [s-1]} x_i^* = \sum_{i \in [\ell]} x_i^*, \quad \forall \ell \in [s-1, d_1-2], \\ &\sum_{i \in [\ell]} \lambda_i < \sum_{i \in [d_2-1]} \lambda_i \leq \sum_{i \in [s-1]} x_i^* = \sum_{i \in [\ell]} x_i^*, \quad \forall \ell \in [d_2, s+n-\tilde{k}-1], \end{aligned} \tag{23}$$

where the strict inequalities are due to $\lambda_{d_1-1} > 0$ and $\lambda_{d_2} < 0$ and the equations result from $x_s^* = \cdots = x_{s+n-\tilde{k}-1}^* = 0$ as detailed in (21).

Next, to facilitate our analysis, we split $\boldsymbol{x}^*$ and $\boldsymbol{\lambda}$ in (21) into three sign-definite subvectors

$$\begin{aligned}&(\boldsymbol{x}^*)^1 \in \mathbb{R}_+^{d_1-1} := \boldsymbol{x}^*_{[d_1-1]}, \quad (\boldsymbol{x}^*)^2 \in \mathbb{R}^{d_2-d_1} := \boldsymbol{x}^*_{[d_1,d_2-1]} = \mathbf{0}, \quad (\boldsymbol{x}^*)^3 \in -\mathbb{R}_+^{n-d_2+1} := \boldsymbol{x}^*_{[d_2,n]},\\ &\boldsymbol{\lambda}^1 \in \mathbb{R}_{++}^{d_1-1} := \boldsymbol{\lambda}_{[d_1-1]}, \quad \boldsymbol{\lambda}^2 \in \mathbb{R}^{d_2-d_1} := \boldsymbol{\lambda}_{[d_1,d_2-1]} = \mathbf{0}, \quad \boldsymbol{\lambda}^3 \in -\mathbb{R}_{++}^{n-d_2+1} := \boldsymbol{\lambda}_{[d_2,n]}.\end{aligned} \tag{24}$$

Suppose $\boldsymbol{\Lambda}_i = \mathrm{Diag}(\boldsymbol{\lambda}^i)$ for $i = 1,2,3$, then $\boldsymbol{\Lambda}_1 \in \mathcal{S}_{++}^{d_1-1}$ and $\boldsymbol{\Lambda}_3 \in -\mathcal{S}_{++}^{n-d_2+1}$. Following the proof in Theorem 1, given the strict inequalities from $s-1$ to d_1-2 on the first line of (23), we can split $\boldsymbol{\Lambda}_1$ into $\boldsymbol{\Lambda}_1^1 \in \mathcal{S}_{++}^{s-2} := \mathrm{Diag}(\lambda_1, \cdots, \lambda_{s-2})$ and $\boldsymbol{\Lambda}_1^2 \in\in \mathcal{S}_{++}^{d_1-s+1} := \mathrm{Diag}(\lambda_{s-1}, \cdots, \lambda_{d_1-1})$. Similarly, using the strict inequalities from d_2 to $s+n-\tilde{k}-1$ on the second line of (23), we split $\boldsymbol{\Lambda}_3$ into $\boldsymbol{\Lambda}_3^1 \in -\mathcal{S}_{++}^{s+n-\tilde{k}-d_2+1} := \mathrm{Diag}(\lambda_{d_2}, \cdots, \lambda_{s+n-\tilde{k}})$ and $\boldsymbol{\Lambda}_3^2 \in -\mathcal{S}_{++}^{\tilde{k}-s} := \mathrm{Diag}(\lambda_{s+n-\tilde{k}+1}, \cdots, \lambda_n)$.

Thus, we rewrite the matrix $\boldsymbol{X}^* \in \mathcal{S}^n$ as the eigen-decomposition below

$$\begin{aligned}\boldsymbol{X}^* &= \boldsymbol{Q}_1\boldsymbol{\Lambda}_1\boldsymbol{Q}_1^\top + \boldsymbol{Q}_2\boldsymbol{\Lambda}_2\boldsymbol{Q}_2^\top + \boldsymbol{Q}_3\boldsymbol{\Lambda}_3\boldsymbol{Q}_3^\top = \boldsymbol{Q}_1\boldsymbol{\Lambda}_1\boldsymbol{Q}_1^\top + \boldsymbol{Q}_3\boldsymbol{\Lambda}_3\boldsymbol{Q}_3^\top\\ &= \left(\boldsymbol{Q}_1^1\ \boldsymbol{Q}_1^2\right)\begin{pmatrix}\boldsymbol{\Lambda}_1^1 & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Lambda}_1^2\end{pmatrix}\left(\boldsymbol{Q}_1^1\ \boldsymbol{Q}_1^2\right)^\top + \left(\boldsymbol{Q}_3^1\ \boldsymbol{Q}_3^2\right)\begin{pmatrix}\boldsymbol{\Lambda}_3^1 & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Lambda}_3^2\end{pmatrix}\left(\boldsymbol{Q}_3^1\ \boldsymbol{Q}_3^2\right)^\top\end{aligned}$$

where the eigenvector matrix $\boldsymbol{Q} = \left(\boldsymbol{Q}_1\ \boldsymbol{Q}_2\ \boldsymbol{Q}_3\right)$ can be decomposed accordingly and $\boldsymbol{Q}_2$ corresponds to zero eigenvalues and is thus omitted.

For any d-dimensional face F^d, there are $(d+1)$ points such that $F^d \subseteq \mathrm{aff}(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_{d+1})$, where $\mathrm{aff}(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_{d+1}) = \{\sum_{i\in[d+1]} \alpha_i \boldsymbol{X}_i : \boldsymbol{\alpha} \in \mathbb{R}^{d+1}, \sum_{i\in[d+1]} \alpha_i = 1\}$ denotes the affine hull of these points. Suppose that the inequality $d+2 < 1/2((d_1-s+1)(d_1-s+2) + (s+n-\tilde{k}-d_2+1)(s+n-\tilde{k}-d_2+2))$ holds, then there exists a block symmetric matrix $\boldsymbol{\Delta} = \begin{pmatrix}\boldsymbol{\Delta}_1 \in \mathcal{S}^{d_1-s+1} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Delta}_3 \in \mathcal{S}^{s+n-\tilde{k}-d_2+1}\end{pmatrix}$ such that

$$\langle \left(\boldsymbol{Q}_1^2\ \boldsymbol{Q}_3^1\right)^\top (\boldsymbol{X}_i - \boldsymbol{X}_{d+1})\left(\boldsymbol{Q}_1^2\ \boldsymbol{Q}_3^1\right), \boldsymbol{\Delta}\rangle = 0, \forall i \in [d], \mathrm{tr}(\boldsymbol{\Delta}_1) = 0, \mathrm{tr}(\boldsymbol{\Delta}_3) = 0, \tag{25}$$

where the first d equations imply

$$\left(\boldsymbol{Q}_1^1\ \boldsymbol{Q}_1^2\right)\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Delta}_1\end{pmatrix}\left(\boldsymbol{Q}_1^1\ \boldsymbol{Q}_1^2\right)^\top + \begin{pmatrix}\boldsymbol{\Delta}_3 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\left(\boldsymbol{Q}_3^1\ \boldsymbol{Q}_3^2\right)^\top \perp F^d. \tag{26}$$

Then let us construct the two matrices $\boldsymbol{X}^+ \in \mathcal{S}^n$ and $\boldsymbol{X}^- \in \mathcal{S}^n$ as below

$$\boldsymbol{X}^\pm(\delta) = \left(\boldsymbol{Q}_1^1\ \boldsymbol{Q}_1^2\right)\begin{pmatrix}\boldsymbol{\Lambda}_1^1 & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Lambda}_1^2 \pm \delta\boldsymbol{\Delta}_1\end{pmatrix}\left(\boldsymbol{Q}_1^1\ \boldsymbol{Q}_1^2\right)^\top + \left(\boldsymbol{Q}_3^1\ \boldsymbol{Q}_3^2\right)\begin{pmatrix}\boldsymbol{\Lambda}_3^1 \pm \delta\boldsymbol{\Delta}_3 & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Lambda}_3^2\end{pmatrix}\left(\boldsymbol{Q}_3^1\ \boldsymbol{Q}_3^2\right)^\top,$$

where $\delta > 0$ is small enough and the eigenvalues of $\boldsymbol{X}^+(\delta)$ and $\boldsymbol{X}^-(\delta)$ can be written as

$$\boldsymbol{\lambda}(\boldsymbol{X}^\pm(\delta)) \in \mathbb{R}^n := \left(\boldsymbol{\lambda}(\boldsymbol{\Lambda}_1^1),\ \boldsymbol{\lambda}(\boldsymbol{\Lambda}_1^2 \pm \delta\boldsymbol{\Delta}_1),\ \mathbf{0},\ \boldsymbol{\lambda}(\boldsymbol{\Lambda}_3^1 \pm \delta\boldsymbol{\Delta}_3),\ \boldsymbol{\lambda}(\boldsymbol{\Lambda}_3^2)\right).$$

First, let us focus on the nonnegative pair $((\boldsymbol{x}^*)^1, \boldsymbol{\lambda}^1)$. Given the first line in (23) and the inequality (22) that implies $\sum_{i\in[d_1-1]} \lambda_i \leq \sum_{i\in[d_1-1]} x_i^* = \sum_{i\in[s]} x_i^*$, according to Claim 1 in Theorem 1, there exists a positive scalar $\underline{\delta} > 0$ such that

$$(\boldsymbol{x}^*)^1 \succeq \left(\boldsymbol{\lambda}(\boldsymbol{\Lambda}_1^1),\, \boldsymbol{\lambda}(\boldsymbol{\Lambda}_1^2 \pm \delta \boldsymbol{\Delta}_1)\right), \quad \boldsymbol{\Lambda}_1^2 \pm \delta \boldsymbol{\Delta}_1 \in \mathcal{S}_{++}^{d_1-s+1}, \quad \forall \delta \in (0, \underline{\delta}).$$

Given $\operatorname{tr}(\boldsymbol{\Delta}_1) = 0$ and $\left(\boldsymbol{\lambda}(\boldsymbol{\Lambda}_1^1),\, \boldsymbol{\lambda}(\boldsymbol{\Lambda}_1^2 \pm \delta \boldsymbol{\Delta}_1)\right) \in \mathbb{R}_+^{d_1-1}$, we can obtain

$$\left((\boldsymbol{x}^*)^1,\, (\boldsymbol{x}^*)^2\right) \succeq \left(\boldsymbol{\lambda}(\boldsymbol{\Lambda}_1^1),\, \boldsymbol{\lambda}(\boldsymbol{\Lambda}_1^2 \pm \delta \boldsymbol{\Delta}_1),\, \boldsymbol{0}\right).$$

In a similar vein, for the nonpositive pair $((\boldsymbol{x}^*)^3, \boldsymbol{\lambda}^3)$, motivated by the second line in (23), we can let $j_0 = j_1 = s + n - \tilde{k} - 1 < j_2$ in Lemma 1 and $\lambda_{s+n-\tilde{k}} = \cdots = \lambda_{j_2} > \lambda_{j_2+1}$. Analogous to the proof of Claim 1 in Theorem 1, along with the majorization result above, we can show that there exists a $\underline{\delta} > 0$ such that for any $\delta \in (0, \underline{\delta})$,

$$\boldsymbol{x}^* := \left((\boldsymbol{x}^*)^1,\, (\boldsymbol{x}^*)^2,\, (\boldsymbol{x}^*)^3\right) \succeq \left(\boldsymbol{\lambda}(\boldsymbol{\Lambda}_1^1),\, \boldsymbol{\lambda}(\boldsymbol{\Lambda}_1^2 \pm \delta \boldsymbol{\Delta}_1),\, \boldsymbol{0},\, \boldsymbol{\lambda}(\boldsymbol{\Lambda}_3^1 \pm \delta \boldsymbol{\Delta}_3),\, \boldsymbol{\lambda}(\boldsymbol{\Lambda}_3^2)\right) =: \boldsymbol{\lambda}(\boldsymbol{X}^{\pm}(\delta)). \tag{27}$$

Combining with the conditions on $\boldsymbol{\Delta}$ in (25) and the fact $\boldsymbol{x}^* \succeq \boldsymbol{\lambda}(\boldsymbol{X}^{\pm}(\delta))$ in (27), we can conclude that $\boldsymbol{X}^+(\delta), \boldsymbol{X}^-(\delta) \in \mathcal{Y}^s$. Thus, the equation $\boldsymbol{X}^* = (\boldsymbol{X}^+(\delta) + \boldsymbol{X}^-(\delta))/2$ contradicts with the orthogonalization in (26). We must have

$$\begin{aligned} d+2 &\geq \frac{(d_1-s+1)(d_1-s+2)}{2} + \frac{(s+n-\tilde{k}-d_2+1)(s+n-\tilde{k}-d_2+2)}{2} \\ &= \frac{(d_1-s+1)(d_1-s+2)}{2} + \frac{(r-\tilde{k}+s+1-d_1)(r-\tilde{k}+s+2-d_1)}{2} \geq \frac{(r-\tilde{k}+2)(r-\tilde{k}+4)}{4}, \end{aligned}$$

where the first equation is from $d_2 - d_1 = n - r$ and the last inequality is obtained by minimizing the convex function over d_1 that $d_1^* = s + (r-\tilde{k})/2$ at optimality. Therefore, the largest rank of all points in d-dimensional face F^d must not exceed $\tilde{k} + \sqrt{9+4d} - 3$.

(ii) Suppose $d_1 \leq s$, then $d_2 - d_1 = n - r$ implies $d_2 \leq s + n - r < s + n - \tilde{k}$. Thus, the second line in (23) still holds. Now we only focus on the nonpositive pair $((\boldsymbol{x}^*)^3, \boldsymbol{\lambda}^3)$ and construct a nonzero symmetric matrix $\boldsymbol{\Delta}_3 \in \mathcal{S}^{s+n-\tilde{k}-d_2+1}$ as (25) with $\boldsymbol{\Delta}_1 = \boldsymbol{0}$. It follows that

$$d+1 \geq \frac{(s+n-\tilde{k}-d_2+1)(s+n-\tilde{k}-d_2+2)}{2} \geq \frac{(r-\tilde{k}+1)(r-\tilde{k}+2)}{2},$$

where the last inequality is due to $d_2 \leq s + n - r$. Hence, we conclude $r \leq \tilde{k} + \lfloor \sqrt{2d + \frac{9}{4}} - \frac{3}{2} \rfloor$.

(iii) If $s + n - \tilde{k} \leq d_2 \leq n$, then $d_2 - d_1 = n - r$ implies $d_1 \geq s + r - \tilde{k} > s$. Thus, the first line in (23) still holds and we can only focus on the nonnegative pair $((\boldsymbol{x}^*)^1, \boldsymbol{\lambda}^1)$. By constructing a nonzero symmetric matrix $\boldsymbol{\Delta}_1 \in \mathcal{S}^{d_1-s+1}$ as (25) with $\boldsymbol{\Delta}_3 = \boldsymbol{0}$, we have

$$d+1 \geq \frac{(d_1-s+1)(d_1-s+2)}{2} \geq \frac{(r-\tilde{k}+1)(r-\tilde{k}+2)}{2},$$

where the last inequality is due to $d_1 \geq s + r - \tilde{k}$. We thus obtain $r \leq \tilde{k} + \lfloor \sqrt{2d + 9/4} - 3/2 \rfloor$.

Taking the largest one among the rank bounds in parts (i), (ii), and (iii), we finally obtain a rank bound of $\tilde{k} + \lfloor \sqrt{4d+9} \rfloor - 3$. □

A.12 Proofs of Theorems 8 and 9

The rank bound of Theorems 8 and 9 arise from describing the convex hull $\operatorname{conv}(\mathcal{D})$ and analyzing the rank of its faces, as shown below.

Proposition 9 *Given a sparse domain set $\mathcal{D}$ in (3) and integer $\tilde{k}$ following Definition 1, we have*

(i) If function $f(\cdot)$ in (3) is sign-invariant, set $\operatorname{conv}(\mathcal{D})$ is closed and is equal to $\operatorname{conv}(\mathcal{D}) := \Big\{\boldsymbol{X} \in \mathcal{S}^n : \boldsymbol{X} = \operatorname{Diag}(\operatorname{diag}(\boldsymbol{X})), \exists \boldsymbol{x} \in \mathbb{R}^n_+, f(\boldsymbol{x}) \leq 0, x_1 \geq \cdots \geq x_n, x_{\tilde{k}+1} = 0, \|\boldsymbol{X}\|_{(\ell)} \leq \sum_{i\in[\ell]} x_i, \forall \ell \in [\tilde{k}-1], \|\boldsymbol{X}\|_* \leq \sum_{i\in[\tilde{k}]} x_i\Big\}$;

(ii) Otherwise, $\operatorname{conv}(\mathcal{D}) = \operatorname{conv}(\cup_{s\in[\tilde{k}+1]} \widehat{\mathcal{Y}}^s)$*, where for each $s \in [\tilde{k}+1]$, set $\widehat{\mathcal{Y}}^s$ is closed and equals* $\widehat{\mathcal{Y}}^s := \big\{\boldsymbol{X} \in \mathcal{S}^n : \boldsymbol{X} = \operatorname{Diag}(\operatorname{diag}(\boldsymbol{X})), \exists \boldsymbol{x} \in \mathbb{R}^n, f(\boldsymbol{x}) \leq 0, x_1 \geq \cdots \geq x_n, x_s = x_{s+n-\tilde{k}-1} = 0, \boldsymbol{x} \succeq \operatorname{diag}(\boldsymbol{X})\big\}$.

Proof. When applied to the diagonal symmetric indefinite matrix space, Part (i) follows Proposition 7, and Part (ii) can be directly obtained by Proposition 4. □

The set $\operatorname{conv}(\mathcal{D})$ admits different expressions depending on whether the function $f(\cdot)$ in (3) is sign-invariant or not, as specified in Proposition 9. Next, we derive the rank bounds of their faces for two expressions, respectively.

Proposition 10 *Given a sparse domain set $\mathcal{D}$ in (3), suppose that F^d is a d-dimensional face in the set $\operatorname{conv}(\mathcal{D})$ or in the set $\widehat{\mathcal{Y}}^s$ for any $s \in [\tilde{k}+1]$. Then all points in face F^d always admit a rank at most $\tilde{k}+d$, where integer $\tilde{k} \leq k$ follows Definition 1.*

Proof. The proof includes two parts.

(i) $F^d \subseteq \operatorname{conv}(\mathcal{D})$. Suppose that $f(\cdot)$ is sign-invariant. Given that a matrix $\boldsymbol{X}^* \in F^d$ is diagonal and has a rank r greater than $\tilde{k}+d$, following the proof of Theorem 1, we can construct a diagonal matrix $\boldsymbol{\Delta} \in \mathbb{R}^{(r-\tilde{k}+1)\times(r-\tilde{k}+1)}$ satisfying the $(d+1)$ constraints (18) whenever $d+1 < r-\tilde{k}+1$ holds, as any diagonal matrix of size $(r-\tilde{k}+1)\times(r-\tilde{k}+1)$ can be mapped into a vector of length $r-\tilde{k}+1$. The existence of such a diagonal matrix $\boldsymbol{\Delta}$ can form a contradiction as shown in Theorem 1, implying that $d+1 \geq r-\tilde{k}+1$ must hold.

(ii) $F^d \subseteq \widehat{\mathcal{Y}}^s$ for any $s \in [\tilde{k}+1]$. Given that a matrix $\boldsymbol{X}^* \in F^d$ is diagonal and is of rank r, following the similar proof of Theorem 6, we can construct a diagonal matrix $\boldsymbol{\Delta}$ satisfying $d+2$ equality constraints in (25); that is,

$$d+2 \leq d_1 - s + 1 + s + n - \tilde{k} - d_2 + 1 = r - \tilde{k} + 2,$$

where the second equality is due to $d_2 - d_1 = n - r$. This completes the proof. □

For a sparse domain set $\mathcal{D}$ in (3), the two sets $\mathrm{conv}(\mathcal{D})$, $\{\widehat{\mathcal{Y}}^s\}_{s\in[\tilde{k}+1]}$ provide us with the relaxations LSOP-R, LSOP-R-I, respectively. Applying Lemma 2 to either the set $\mathrm{conv}(\mathcal{D})$ or the set $\widehat{\mathcal{Y}}^s$ for each $s\in[\tilde{k}+1]$ and using the results in Proposition 10, one can readily derive the rank bounds of either LSOP-R or LSOP-R-I, as shown in the theorems below.

Theorem 8 *Given a sparse domain set $\mathcal{D}$ in (3) and integer $\tilde{k}\le k$ in Definition 1, suppose that the function $f(\cdot)$ in (3) is sign-invariant and LSOP-R admits a line-free feasible set. Then we have*

(i) Each feasible extreme point in the LSOP-R has a rank at most $\tilde{k}+\tilde{m}$; and

(ii) There is an optimal solution to LSOP-R of rank at most $\tilde{k}+\tilde{m}$ if the LSOP-R yields a finite optimal value, i.e., $\mathbf{V}_{\mathrm{rel}} > -\infty$.

Proof. Plugging the rank bound in Proposition 10 of the set $\mathrm{conv}(\mathcal{D})$ described in Proposition 9, the rest of the proof is nearly identical to that in Theorem 2 and thus is omitted.

Theorem 9 *Given a sparse domain set $\mathcal{D}$ in (3) and integer $\tilde{k}\le k$ following Definition 1, suppose that the inner minimization of LSOP-R-I always admits a line-free feasible set. Then we have*

(i) For any $s\in[\tilde{k}+1]$, each feasible extreme point in the sth inner minimization of LSOP-R-I has a rank at most $\tilde{k}+\tilde{m}$; and

(ii) There is an optimal solution to LSOP-R-I of rank at most $\tilde{k}+\tilde{m}$ if the LSOP-R-I yields a finite optimal value, i.e., $\mathbf{V}_{\text{rel -I}} > -\infty$.

Proof. According to Proposition 10, any d-dimensional face of set $\widehat{\mathcal{Y}}^s$ satisfies the rank-$(\tilde{k}+d)$ for all $s\in[\tilde{k}+1]$. Using Lemma 2 and following the analysis of Theorem 2, we can show that for each inner minimization of LSOP-R-I, all the extreme points in the intersection of set $\widehat{\mathcal{Y}}^s$ and $\tilde{m}$ linearly independent inequalities admit a rank-$(\tilde{k}+\tilde{m})$ bound. □

A.13 Proof of Corollary 6

Corollary 6 (Sparse Ridge Regression) *Given the data $\boldsymbol{A}\in\mathbb{R}^{m\times n}$ in Sparse Ridge Regression, there is an optimal solution $\boldsymbol{X}^*$ of Sparse Ridge Regression-R of rank at most $k+\mathrm{rank}(\boldsymbol{A})$.*

Proof. Let $(\widehat{z},\widehat{\boldsymbol{y}},\widehat{\boldsymbol{X}})$ denote an optimal solution to the Sparse Ridge Regression-R, where $\mathbf{V}_{\mathrm{rel}}\ge 0$ must hold. Then given an optimal solution $\widehat{z}$, any matrix $\boldsymbol{X}$ in the following compact set $\mathcal{T}$ is optimal to Sparse Ridge Regression-R

$$\mathcal{T}:=\left\{\boldsymbol{X}\in\mathrm{conv}(\mathcal{D}):\boldsymbol{A}\,\mathrm{diag}(\boldsymbol{X})=\boldsymbol{A}\,\mathrm{diag}(\widehat{\boldsymbol{X}})\right\},$$

where given $\widehat{z}$, we let $\mathcal{D}:=\{\boldsymbol{X}\in\mathcal{S}^n:\boldsymbol{X}=\mathrm{Diag}(\mathrm{diag}(\boldsymbol{X})),\mathrm{rank}(\boldsymbol{X})\le k,\|\boldsymbol{X}\|_F^2\le\widehat{z}\}$.

For each $i \in [m]$, let $\boldsymbol{a}_i \in \mathbb{R}^n$ denote the ith row vector of matrix $\boldsymbol{A}$. In this way, the equation system $\boldsymbol{A}\,\mathrm{diag}(\widehat{\boldsymbol{X}}) = \boldsymbol{A}\,\mathrm{diag}(\boldsymbol{X})$ can be reformulated by the standard form below

$$\left\langle \boldsymbol{A}_i, \widehat{\boldsymbol{X}} \right\rangle = b_i, \forall i \in [m],$$

where for each $i \in [m]$, $\boldsymbol{A}_i = \mathrm{Diag}(\boldsymbol{a}_i)$ and $b_i = \boldsymbol{a}_i^\top \mathrm{diag}(\widehat{\boldsymbol{X}})$. The rank of the matrix $\boldsymbol{A}$ represents the number of linearly independent row vectors in $\{\boldsymbol{a}_i\}_{i\in[m]}$ and so does for $\{\boldsymbol{A}_i\}_{i\in[m]}$.

Since $\tilde{k} \le k$ and there are $\mathrm{rank}(\boldsymbol{A})$ linearly independent equations in the compact set $\mathcal{T}$, according to Theorem 8, one can find an alternative extreme point with rank at most $k + \mathrm{rank}(\boldsymbol{A})$. □

A.14 Proof of Proposition 6

Proposition 6 *The output solution of column generation Algorithm 1 is ϵ-optimal to the LSOP-R (4) if the LSOP-R (4) yields a finite optimal value, i.e., $\mathbf{V}_{\mathrm{rel}} > -\infty$. That is, suppose that Algorithm 1 returns $\boldsymbol{X}^*$, then the inequalities $\mathbf{V}_{\mathrm{rel}} \le \langle \boldsymbol{A}_0, \boldsymbol{X}^* \rangle \le \mathbf{V}_{\mathrm{rel}} + \epsilon$ hold.*

Proof. Notice that $\mathbf{V}_{\mathrm{rel}} \le \langle \boldsymbol{A}_0, \boldsymbol{X}^* \rangle$ holds, since the RMP only optimizes over a subset of $\mathrm{conv}(\mathcal{D})$. To prove the other bound, we see that at termination, the optimal value of the Pricing Problem satisfies $\mathbf{V}_{\mathrm{P}} \le \nu^* + \epsilon$. The Lagrangian dual of original LSOP-R can be rewritten as

$$\mathbf{V}_{\mathrm{rel}} := \min_{\boldsymbol{X} \in \mathrm{conv}(\mathcal{D})} \left\{ \langle \boldsymbol{A}_0, \boldsymbol{X} \rangle + \sup_{\boldsymbol{\mu}^l \in \mathbb{R}^m_+, \boldsymbol{\mu}^u \in \mathbb{R}^m_+} \left\{ \left\langle \sum_{i \in [m]} (\mu_i^u - \mu_i^l) \boldsymbol{A}_i, \boldsymbol{X} \right\rangle + (\boldsymbol{b}^l)^\top \boldsymbol{\mu}^l - (\boldsymbol{b}^u)^\top \boldsymbol{\mu}^u \right\} \right\}.$$

By fixing the inner dual variables to be $((\boldsymbol{\mu}^l)^*, (\boldsymbol{\mu}^u)^*)$ which correspond to the RMP, we have

$$\begin{aligned}\mathbf{V}_{\mathrm{rel}} \ge \min_{\boldsymbol{X} \in \mathrm{conv}(\mathcal{D})} & \left\{ \langle \boldsymbol{A}_0, \boldsymbol{X} \rangle + \left\langle \sum_{i \in [m]} ((\mu_i^u)^* - (\mu_i^l)^*) \boldsymbol{A}_i, \boldsymbol{X} \right\rangle + (\boldsymbol{b}^l)^\top (\boldsymbol{\mu}^l)^* - (\boldsymbol{b}^u)^\top (\mu^u)^* \right\} \\ &= (\boldsymbol{b}^l)^\top (\boldsymbol{\mu}^l)^* - (\boldsymbol{b}^u)^\top (\mu^u)^* - \mathbf{V}_{\mathrm{P}} \ge (\boldsymbol{b}^l)^\top (\boldsymbol{\mu}^l)^* - (\boldsymbol{b}^u)^\top (\mu^u)^* - \nu^* - \epsilon = \langle \boldsymbol{A}_0, \boldsymbol{X}^* \rangle - \epsilon,\end{aligned}$$

where the second inequality is due to $\mathbf{V}_{\mathrm{P}} \le \nu^* + \epsilon$ and the last equality is from the strong duality of RMP. This completes the proof. □

A.15 Proof of Theorem 10

Theorem 10 *The Pricing Problem-S has an optimal solution $\boldsymbol{X}^*$, as specified follows*

(i) Given $\mathcal{Q} := \mathcal{S}^n_+$ in (1), suppose that $\tilde{\boldsymbol{A}} := \boldsymbol{U}\,\mathrm{Diag}(\boldsymbol{\beta})\boldsymbol{U}^\top$ denotes the eigen-decomposition of matrix $\tilde{\boldsymbol{A}}$ with eigenvalues $\beta_1 \ge \cdots \ge \beta_n$. Then $\boldsymbol{X}^* = \boldsymbol{U}\,\mathrm{Diag}(\boldsymbol{\lambda}^*)\boldsymbol{U}^\top$ and $\boldsymbol{\lambda}^* \in \mathbb{R}^n_+$ equals

$$\boldsymbol{\lambda}^* \in \arg\max_{\boldsymbol{\lambda} \in \mathbb{R}^n_+} \left\{ \boldsymbol{\lambda}^\top \boldsymbol{\beta} : \lambda_i = 0, \forall i \in [k+1, n], f(\boldsymbol{\lambda}) \le 0 \right\}. \qquad (6)$$

(ii) Given $\mathcal{Q} := \mathbb{R}^{n\times p}$ *in* (1)*, suppose that* $\tilde{\boldsymbol{A}} := \boldsymbol{U}\,\mathrm{Diag}(\boldsymbol{\beta})\boldsymbol{V}^\top$ *is the singular value decomposition of matrix* $\tilde{\boldsymbol{A}}$ *with singular values* $\beta_1 \geq \cdots \geq \beta_n$*. Then* $\boldsymbol{X}^* = \boldsymbol{U}\,\mathrm{Diag}(\boldsymbol{\lambda}^*)\boldsymbol{V}^\top$ *and* $\boldsymbol{\lambda}^* \in \mathbb{R}^n_+$ *equals*

$$\boldsymbol{\lambda}^* \in \arg\max_{\boldsymbol{\lambda}\in\mathbb{R}^n_+} \left\{\boldsymbol{\lambda}^\top\boldsymbol{\beta} : \lambda_i = 0, \forall i \in [k+1, n], f(\boldsymbol{\lambda}) \leq 0\right\}. \tag{7}$$

(iii) Given $\mathcal{Q} := \mathcal{S}^n$ *in* (1)*, suppose that* $\tilde{\boldsymbol{A}} := \boldsymbol{U}\,\mathrm{Diag}(\boldsymbol{\beta})\boldsymbol{U}^\top$ *denotes the eigen-decomposition of matrix* $\tilde{\boldsymbol{A}}$ *with eigenvalues* $\beta_1 \geq \cdots \geq \beta_n$*. Then* $\boldsymbol{X}^* = \boldsymbol{U}\,\mathrm{Diag}(\boldsymbol{\lambda}^*)\boldsymbol{U}^\top$ *and* $\boldsymbol{\lambda}^* \in \mathbb{R}^n$ *equals*

$$\boldsymbol{\lambda}^* \in \arg\max_{s\in[k+1]} \max_{\boldsymbol{\lambda}\in\mathbb{R}^n} \left\{\boldsymbol{\lambda}^\top\boldsymbol{\beta} : \lambda_1 \geq \cdots \geq \lambda_n, \lambda_s = \lambda_{s+n-k-1} = 0, f(\boldsymbol{\lambda}) \leq 0\right\}. \tag{8}$$

(iv) Given $\mathcal{Q} := \mathcal{S}^n$ *and a sign-invariant function* $f(\cdot)$ *in* (1)*, suppose that* $\tilde{\boldsymbol{A}} := \boldsymbol{U}\,\mathrm{Diag}(\boldsymbol{\beta})\boldsymbol{U}^\top$ *denotes the eigen-decomposition of matrix* $\tilde{\boldsymbol{A}}$ *with eigenvalues satisfying* $|\beta_1| \geq \cdots \geq |\beta_n|$*. Then* $\boldsymbol{X}^* = \boldsymbol{U}\,\mathrm{Diag}(\boldsymbol{\lambda}^*)\boldsymbol{U}^\top$ *and* $\boldsymbol{\lambda}^* \in \mathbb{R}^n$ *equals*

$$\boldsymbol{\lambda}^* \in \arg\max_{\boldsymbol{\lambda}\in\mathbb{R}^n} \left\{|\boldsymbol{\lambda}|^\top|\boldsymbol{\beta}| : |\lambda_i| = 0, \forall i \in [k+1, n], f(|\boldsymbol{\lambda}|) \leq 0\right\}. \tag{9}$$

Proof. The proof can be split into four parts.

Part (i). When $\mathcal{Q} := \mathcal{S}^n_+$, we represent the matrix variable $\boldsymbol{X} \in \mathcal{S}^n_+$ by its eigen-decomposition $\boldsymbol{X} = \sum_{i\in[n]} \lambda_i \boldsymbol{q}_i\boldsymbol{q}_i^\top$ with eigenvalues sorted in descending order. Hence, the Pricing Problem-S reduces to

$$\text{(PP)} \quad \mathbf{V}_{\mathrm{P}} := \max_{\boldsymbol{\lambda}\in\mathbb{R}^n_+} \max_{\boldsymbol{Q}\in\mathbb{R}^{n\times n}} \left\{\sum_{i\in[n]} \lambda_i \boldsymbol{q}_i^\top \tilde{\boldsymbol{A}}\boldsymbol{q}_i : \boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}_n, \lambda_1 \geq \cdots \geq \lambda_n, \|\boldsymbol{\lambda}\|_0 \leq k, f(\boldsymbol{\lambda}) \leq 0\right\}, \tag{28}$$

where $\|\boldsymbol{\lambda}\|_0 \leq k$ is equivalent to the rank constraint $\mathrm{rank}(\boldsymbol{X}) \leq k$. Next we introduce a key claim.

Claim 2 $\max_{\boldsymbol{Q}\in\mathbb{R}^{n\times\ell}, \boldsymbol{Q}^\top\boldsymbol{Q}=\boldsymbol{I}_\ell} \sum_{i\in[\ell]} \boldsymbol{q}_i^\top\tilde{\boldsymbol{A}}\boldsymbol{q}_i = \sum_{i\in[\ell]} \beta_i$ *for each* $\ell \in [n]$.

Proof. Suppose $\boldsymbol{Y} = \boldsymbol{Q}\boldsymbol{Q}^\top$, then we have

$$\begin{aligned} \max_{\boldsymbol{Q}\in\mathbb{R}^{n\times\ell}, \boldsymbol{Q}^\top\boldsymbol{Q}=\boldsymbol{I}_\ell} \sum_{i\in[\ell]} \boldsymbol{q}_i^\top\tilde{\boldsymbol{A}}\boldsymbol{q}_i &\leq \max_{\boldsymbol{Y}\in\mathcal{S}^n_+} \left\{\langle\tilde{\boldsymbol{A}}, \boldsymbol{Y}\rangle : \boldsymbol{Y} \preceq \boldsymbol{I}_n, \mathrm{tr}(\boldsymbol{Y}) = k\right\} \\ &= \max_{\boldsymbol{Y}\in\mathcal{S}^n_+} \left\{\langle\tilde{\boldsymbol{A}} - \beta_n\boldsymbol{I}_n, \boldsymbol{Y}\rangle + k\beta_n : \boldsymbol{Y} \preceq \boldsymbol{I}_n, \mathrm{tr}(\boldsymbol{Y}) = k\right\} = \sum_{i\in[k]} \beta_i, \end{aligned}$$

where given the positive semidefinite matrix $\tilde{\boldsymbol{A}} - \beta_n\boldsymbol{I}_n$, the last equation is due to lemma 2 in Li and Xie (2021) and the inequality is attainable. ◇

Next we introduce the auxiliary variables $\boldsymbol{y} \in \mathbb{R}^n$ where $y_i = \boldsymbol{q}_i^\top\tilde{\boldsymbol{A}}\boldsymbol{q}_i$ for each $i \in [n]$. Given $\lambda_1 \geq \cdots \geq \lambda_n$, to maximize $\boldsymbol{\lambda}^\top\boldsymbol{y}$, we must have $y_1 \geq \cdots \geq y_n$ at optimality due to the rearrangement inequality (Hardy et al. 1952). Since Claim 2 implies $\boldsymbol{\beta} \succeq \boldsymbol{y}$, we propose a relaxation problem of the inner maximization over matrix $\boldsymbol{Q}$ in problem (28) that provides an upper bound

$$\max_{\boldsymbol{y}\in\mathbb{R}^n} \left\{\sum_{i\in[n]} \lambda_i y_i : \sum_{i\in[\ell]} y_i \leq \sum_{i\in[\ell]} \beta_i, \forall \ell \in [n-1], \sum_{i\in[n]} y_i = \sum_{i\in[n]} \beta_i, y_1 \geq \cdots \geq y_n\right\},$$

which admits an optimal solution $y_i^* = \beta_i$ for each $i \in [n]$ with optimal value $\sum_{i\in[n]} \lambda_i \beta_i$ due to $\beta_1 \geq \cdots \geq \beta_n$ and $\lambda_1 \geq \cdots \geq \lambda_n$. Besides, the inner maximization over matrix $\boldsymbol{Q}$ in problem (28) can also achieve the objective value $\sum_{i\in[n]} \lambda_i \beta_i$ when $\boldsymbol{Q}^* = \boldsymbol{U}$; thus, this solution is optimal. Plugging the optimal solution $\boldsymbol{Q}^* = \boldsymbol{U}$ into the problem 28, we obtain

$$\max_{\boldsymbol{\lambda}\in\mathbb{R}^n_+} \left\{\boldsymbol{\lambda}^\top \boldsymbol{\beta} : \lambda_1 \geq \cdots \geq \lambda_n, \|\boldsymbol{\lambda}\|_0 \leq k, f(\boldsymbol{\lambda}) \leq 0\right\}. \tag{29}$$

Since $\beta_1 \geq \cdots \geq \beta_n$ and function $f(\boldsymbol{\lambda})$ is permutation-invariant with $\boldsymbol{\lambda}$, according to the rearrangement inequality (Hardy et al. 1952), it is evident that at optimality of problem (29), we must have $\lambda_1 \geq \cdots \geq \lambda_k \geq \lambda_{k+1} = \cdots = \lambda_n = 0$. Therefore, problem (29) can be reformulated by a convex optimization below

$$\max_{\boldsymbol{\lambda}\in\mathbb{R}^n_+} \left\{\boldsymbol{\lambda}^\top \boldsymbol{\beta} : \lambda_i = 0, \forall i \in [k+1, n], f(\boldsymbol{\lambda}) \leq 0\right\},$$

where any optimal solution to problem above must satisfy $\lambda_1 \geq \cdots \geq \lambda_k \geq \lambda_{k+1} = \cdots = \lambda_n = 0$.

Part (ii). If matrix $\boldsymbol{X} \in \mathcal{Q} := \mathbb{R}^{n\times p}$ is non-symmetric and $n \leq p$, then we introduce the singular value decomposition of $\boldsymbol{X} = \sum_{i\in[n]} \lambda_i \boldsymbol{q}_i \boldsymbol{p}_i^\top$ to reformulate the Pricing Problem-S as

$$\max_{\boldsymbol{\lambda}\in\mathbb{R}^n_+} \max_{\boldsymbol{Q}\in\mathbb{R}^{n\times n}, \boldsymbol{P}\in\mathbb{R}^{p\times n}} \left\{ \sum_{i\in[n]} \lambda_i \boldsymbol{q}_i^\top \tilde{\boldsymbol{A}} \boldsymbol{p}_i : \boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{I}_n, \lambda_1 \geq \cdots \geq \lambda_n, \|\boldsymbol{\lambda}\|_0 \leq k \right\}. \tag{30}$$

Following the proof of Claim 2 and lemma 1 in Li and Xie (2021), we obtain the following result:

Claim 3 $\max_{\boldsymbol{Q}\in\mathbb{R}^{n\times\ell}, \boldsymbol{P}\in\mathbb{R}^{p\times\ell}, \boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}_\ell} \sum_{i\in[\ell]} \boldsymbol{q}_i^\top \tilde{\boldsymbol{A}} \boldsymbol{p}_i = \sum_{i\in[\ell]} \beta_i$ *for each* $\ell \in [n]$.

Thus, according to Claim 3, we can equivalently reformulate the inner maximization by optimizing over the variables $(\boldsymbol{Q}, \boldsymbol{P})$ of problem (30) as

$$\max_{\boldsymbol{y}\in\mathbb{R}^n} \left\{ \sum_{i\in[n]} \lambda_i y_i : \sum_{i\in[\ell]} y_i \leq \sum_{i\in[\ell]} \beta_i, \forall \ell \in [n], y_1 \geq \cdots \geq y_n \right\}.$$

Then the remaining proof simply follows from Part (i).

Part (iii). When $\mathcal{Q} := \mathcal{S}^n$, following the analysis of Part (i), we can arrive at problem (29) but now the eigenvalue variable $\boldsymbol{\lambda} \in \mathbb{R}^n$ can be negative under the symmetric case, i.e., $\boldsymbol{X} \in \mathcal{S}^n$. Analogous to Proposition 4, the resulting feasible set can be decomposed into $k+1$ subsets, i.e.,

$$\{\boldsymbol{\lambda} \in \mathbb{R}^n : \lambda_1 \geq \cdots \geq \lambda_n, \|\boldsymbol{\lambda}\|_0 \leq k, f(\boldsymbol{\lambda}) \leq 0\} = \cup_{s\in[k+1]} \mathcal{T}^s,$$

where for each $s \in [k+1]$, $\mathcal{T}^s := \{\boldsymbol{\lambda} \in \mathbb{R}^n : \lambda_1 \geq \cdots \geq \lambda_n, \lambda_s = \lambda_{s+n-k-1} = 0, f_j(\boldsymbol{\lambda}) \leq 0, \forall j \in [t]\}$. Replacing the feasible set of problem (29) by $\cup_{s\in[k+1]} \mathcal{T}^s$, we complete the proof.

Part (iv). For the case of $\mathcal{Q} := \mathcal{S}^n$, suppose that $\boldsymbol{X} = \sum_{i\in[n]} \lambda_i \boldsymbol{q}_i \boldsymbol{q}_i^\top$ denote the eigen-decomposition of matrix variable $\boldsymbol{X}$. Since the rank and spectral functions are sign-invariant, now changing the sign of variable $\boldsymbol{\lambda}$ allows us to reduce problem (28) in Part (i) to

$$\text{(PP)} \quad \mathbf{V}_{\mathrm{P}} := \max_{\boldsymbol{\lambda}\in\mathbb{R}^n} \max_{\boldsymbol{Q}\in\mathbb{R}^{n\times n}} \left\{ \sum_{i\in[n]} |\lambda_i||\boldsymbol{q}_i^\top \tilde{\boldsymbol{A}} \boldsymbol{q}_i| : \boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}_n, |\lambda_1| \geq \cdots \geq |\lambda_n|, \|\boldsymbol{\lambda}\|_0 \leq k, f(|\boldsymbol{\lambda}|) \leq 0 \right\},$$

Using the facts that $\max_{\boldsymbol{Q}\in\mathbb{R}^{n\times\ell}, \boldsymbol{Q}^\top\boldsymbol{Q}=\boldsymbol{I}_\ell} \sum_{i\in[\ell]} |\boldsymbol{q}_i^\top \tilde{\boldsymbol{A}} \boldsymbol{q}_i| = \sum_{i\in[\ell]} |\beta_i|$ for each $\ell \in [n]$ and $|\beta_1| \geq \cdots |\beta_n|$, the analysis of in Part (i) can be readily extended to derive $\boldsymbol{q}_i = \boldsymbol{u}_i$ for each $i \in [n]$ and obtain

$$\boldsymbol{\lambda}^* := \arg\max_{\boldsymbol{\lambda}\in\mathbb{R}^n} \left\{ |\boldsymbol{\lambda}|^\top |\boldsymbol{\beta}| : |\lambda_i| = 0, \forall i \in [k+1, n], f(|\boldsymbol{\lambda}|) \leq 0 \right\}.$$

where the equality is due to the sign-invariant properties of spectral functions $f(\cdot)$. □

A.16 Proof of Corollary 7

Corollary 7 *For a domain set $\mathcal{D} := \{\boldsymbol{X} \in \mathcal{Q} : \operatorname{rank}(\boldsymbol{X}) \leq k, \|\boldsymbol{\lambda}(\boldsymbol{X})\|_\ell \leq c\}$, where $c \geq 0$, $\ell \in [1, \infty]$, and $1/\ell + 1/q = 1$, we have that*

(i) Given $\mathcal{Q} := \mathcal{S}^n_+$, the solution $\boldsymbol{\lambda}^*$ is optimal to problem (6) where $\lambda_i^* = c\sqrt[\ell]{\frac{(\beta_i)_+^q}{\sum_{j\in[k]}(\beta_j)_+^q}}$ for all $i \in [k]$ and $\lambda_i^* = 0$ for all $i \in [k+1, n]$;

(ii) Given $\mathcal{Q} := \mathbb{R}^{n\times p}$, the solution $\boldsymbol{\lambda}^*$ is optimal to problem (7) where $\lambda_i^* = c\sqrt[\ell]{\frac{\beta_i^q}{\sum_{j\in[k]}\beta_j^q}}$ for all $i \in [k]$ and $\lambda_i^* = 0$ for all $i \in [k+1, n]$;

(iii) Given $\mathcal{Q} := \mathcal{S}^n$, the solution $\boldsymbol{\lambda}^*$ is optimal to problem (9) where $\lambda_i^* = \operatorname{sign}(\beta_i) c\sqrt[\ell]{\frac{|\beta_i|^q}{\sum_{j\in[k]}|\beta_j|^q}}$ for all $i \in [k]$ and $\lambda_i^* = 0$ for all $i \in [k+1, n]$.

Proof. **Part (i).** Given $\mathcal{Q} := \mathcal{S}^n_+$ and the ℓ-norm function $f(\cdot) = \|\cdot\|_\ell$, we let r denote the largest integer of matrix $\tilde{\boldsymbol{A}}$ such that $\beta_r \geq 0$. Then, by letting $s := \min\{k, r\}$, the maximization problem (6) over $\boldsymbol{\lambda}$ in Part (i) of Theorem 10 now reduces to

$$\max_{\boldsymbol{\lambda}\in\mathbb{R}^n_+} \left\{ \boldsymbol{\lambda}^\top \boldsymbol{\beta} : \lambda_i = 0, \forall i \in [s+1, n], \|\boldsymbol{\lambda}\|_\ell \leq c \right\} \leq c\sqrt[q]{\sum_{j\in[s]} \beta_j^q} = c\sqrt[q]{\sum_{j\in[k]} (\beta_j)_+^q},$$

where the inequality is due to Holder's inequality and the equation is from the definition $(\beta_j)_+ = \max\{0, \beta_j\}$. The inequality can reach equation when there is $\alpha > 0$ such that $\lambda_i^\ell = \alpha(\beta_i)_+^q$ for all $i \in [n]$, which enables us to construct optimal solution $\boldsymbol{\lambda}^*$ below

$$\lambda_i^* = \begin{cases} c\sqrt[\ell]{\frac{(\beta_i)_+^q}{\sum_{j\in[k]}(\beta_j)_+^q}}, & \forall i \in [k]; \\ 0, & \forall i \in [k+1, n]. \end{cases}.$$

Part (ii). Given $\mathcal{Q} := \mathbb{R}^{n\times p}$, the proof follows that of Part (i) except replacing the eigenvalues by singular values of matrix $\tilde{\boldsymbol{A}}$.

Part (iii). Given $\mathcal{Q} := \mathcal{S}^n$ and sign-invariant ℓ-norm function $f(\cdot) := \|\cdot\|_\ell$, matrix $\tilde{\boldsymbol{A}}$ is symmetric and we let its eigenvalues satisfy $|\beta_1| \geq \cdots \geq |\beta_n|$. Since the sign of variable $\boldsymbol{\lambda}$ can be arbitrary, the maximization problem (9) over $\boldsymbol{\lambda}$ in Part (i) of Theorem 10 now reduces to

$$\max_{\boldsymbol{\lambda}\in\mathbb{R}^n} \{|\boldsymbol{\lambda}|^\top |\boldsymbol{\beta}| : |\lambda_i| = 0, \forall i \in [k+1,n], \|\boldsymbol{\lambda}\|_\ell \leq c\} \leq c \sqrt[q]{\sum_{j\in[k]} |\beta_j|^q},$$

where similar to Part (i), the equality is attainable when $\lambda_i^* = \text{sign}(\beta_i) c \sqrt[\ell]{\frac{|\beta_i|^q}{\sum_{j\in[k]}|\beta_j|^q}}$ for all $i \in [k]$ and $\lambda_i^* = 0$ for all $i \in [k+1,n]$. □

A.17 Proof of Theorem 11

Theorem 11 *For the rank-reduction Algorithm 2, the following statements must hold:*

(i) Algorithm 2 always terminates; and

(ii) Let $\boldsymbol{X}^*$ denote the output solution of Algorithm 2. Then $\boldsymbol{X}^*$ is ϵ-optimal to either LSOP-R or LSOP-R-I, i.e.,

$$\begin{cases} \mathbf{V}_{\text{rel}} \leq \langle \boldsymbol{A}_0, \boldsymbol{X}^* \rangle \leq \mathbf{V}_{\text{rel}} + \epsilon & \text{if } \mathcal{Q} := \mathcal{S}_+^n \text{ or } \mathcal{Q} := \mathbb{R}^{n\times p}; \\ \mathbf{V}_{\text{rel -I}} \leq \langle \boldsymbol{A}_0, \boldsymbol{X}^* \rangle \leq \mathbf{V}_{\text{rel -I}} + \epsilon & \text{if } \mathcal{Q} := \mathcal{S}^n. \end{cases},$$

and the rank of solution $\boldsymbol{X}^*$ satisfies

$$\text{rank}(\boldsymbol{X}^*) \leq \begin{cases} \tilde{k} + \lfloor \sqrt{2\tilde{m} + 9/4} - 3/2 \rfloor, & \text{if } \mathcal{Q} := \mathcal{S}_+^n \text{ or } \mathcal{Q} := \mathbb{R}^{n\times p}; \\ \tilde{k} + \lfloor \sqrt{4\tilde{m} + 9} \rfloor - 3, & \text{if } \mathcal{Q} := \mathcal{S}^n. \end{cases},$$

where integer $\tilde{k}$ is being defined in Definition 1.

Proof. We split the proof into two parts.

Part I. we show that the Algorithm 2 always terminates in different matrix spaces.

(a) Let us begin with the case of $\mathcal{Q} := \mathcal{S}_+^n$. At each iteration of Algorithm 2, for a pair of current and new solutions $(\boldsymbol{X}^*, \boldsymbol{X}(\delta^*))$, they satisfy $\boldsymbol{X}(\delta^*) = \boldsymbol{X}^* + \delta^* \boldsymbol{Y}$, where the direction $\boldsymbol{Y} = \boldsymbol{Q}_2 \boldsymbol{\Delta} \boldsymbol{Q}_2^\top$ is defined in (12). The fact $\text{tr}(\boldsymbol{Y}) = 0$ implies that $\text{tr}(\boldsymbol{X}^*) = \text{tr}(\boldsymbol{X}(\delta^*))$. We let integer $r := \text{rank}(\boldsymbol{X}^*)$. Since $\boldsymbol{Y}$ only perturbs $(r - \tilde{k} + 1)$ smallest positive eigenvalues of matrix $\boldsymbol{X}^*$ (i.e., $\lambda_{\tilde{k}}^*, \cdots, \lambda_r^*$) and the perturbed eigenvalues remain nonnegative, the first $(\tilde{k} - 1)$ largest eigenvalues of $\boldsymbol{X}^*$ are also eigenvalues of new matrix $\boldsymbol{X}(\delta^*)$. Therefore, we must have that $\text{rank}(\boldsymbol{X}(\delta^*)) \leq r$ and $\|\boldsymbol{X}^*\|_{(k)} \leq \|\boldsymbol{X}(\delta^*)\|_{(k)}$, implying that

$$\text{tr}(\boldsymbol{X}^*) - \|\boldsymbol{X}^*\|_{(k)} = \sum_{i\in[\tilde{k},r]} \lambda_i^* \geq \text{tr}(\boldsymbol{X}(\delta^*)) - \|\boldsymbol{X}(\delta^*)\|_{(k)}. \tag{31}$$

which means that the sum of $(r - \tilde{k} + 1)$ smallest positive eigenvalues in the new solution $\boldsymbol{X}(\delta^*)$ must not exceed that of $\boldsymbol{X}^*$.

If the inequality (31) reaches equality and $\operatorname{rank}(\boldsymbol{X}(\delta^*)) = r$, then the $(r-\tilde{k}+1)$ smallest positive eigenvalues of $\boldsymbol{X}^*(\delta^*)$ are exactly those of $\boldsymbol{\Lambda}_2+\delta^*\boldsymbol{\Delta}\in\mathcal{S}_+^{r-\tilde{k}+1}$, where $\operatorname{tr}(\boldsymbol{\Lambda}_2+\delta^*\boldsymbol{\Delta}) = \sum_{i\in[\tilde{k},r]}\lambda_i^*$. Then, following the proof of Theorem 1, we can show that there exists a positive $\widehat{\delta}>0$ such that $\boldsymbol{X}(\delta^*)+\widehat{\delta}\boldsymbol{Y}$ is also feasible to Step 9 of Algorithm 2, which contradicts to the optimality of δ^*. Therefore, for the new solution $\boldsymbol{X}(\delta^*)$, either the sum of its $(r-\tilde{k}+1)$ smallest nonzero eigenvalues strictly decreases (i.e., the inequality (31) is strict) or the rank strictly reduces (i.e., $\operatorname{rank}(\boldsymbol{X}(\delta^*))<r$). Since both values $\sum_{i\in[\tilde{k},r]}\lambda_i^*$ and r are finite, either case ensures that the solution sequence is monotone and thus Algorithm 2 always terminates with $\delta^*=0$ at Step 7.

(b) Similar to Part (a), we can show that Algorithm 2 converges in $\mathcal{Q}=\mathbb{R}^{n\times p}$ as the direction $\boldsymbol{Y}$ in (13) has an impact on only $(r-\tilde{k}+1)$ smallest positive singular values of matrix $\boldsymbol{X}^*$.

(c) In contrast to the case of $\mathcal{Q}:=\mathcal{S}_+^n$ or $\mathcal{Q}:=\mathbb{R}^{n\times p}$, the eigenvectors $\left(\boldsymbol{Q}_1^2\ \boldsymbol{Q}_3^1\right)$ that determine the direction $\boldsymbol{Y}$ in (14) correspond to the (d_1-s^*+1) smallest positive eigenvalues and $(s^*+n-\tilde{k}-d_2+1)$ largest negative eigenvalues, and then the convergence analysis of Part (a) can be readily extended by leveraging those eigenvalues.

Part II. Since Algorithm 2 starts with an ϵ-optimal solution, and we always find a direction $\boldsymbol{Y}$ along which the objective value does not increase, the output solution is at least ϵ-optimal.

We show by contradiction that the output solution $\boldsymbol{X}^*$ of Algorithm 2 must satisfy the rank bounds. We discuss different matrix spaces.

(i) $\mathcal{Q}:=\mathcal{S}_+^n$. At termination of Algorithm 2, we have $\delta^*=0$ and obtain solution $\boldsymbol{X}^*$. Suppose $\operatorname{rank}(\boldsymbol{X}^*)>\tilde{k}+\lfloor\sqrt{2\tilde{m}+9/4}-3/2\rfloor$, i.e., $\tilde{m}+1<(r-\tilde{k}+1)(r-\tilde{k}+2)/2$. Then, there exists a nonzero matrix $\boldsymbol{\Delta}$ to form a nonzero direction $\boldsymbol{Y}$ in (12) and following Claim 1 in the proof of Theorem 1, we can find a nonzero $\underline{\delta}>0$ such that the eigenvalue vector $\boldsymbol{X}^*+\underline{\delta}\boldsymbol{Y}$ is majorized by $\boldsymbol{x}^*$ and $\lambda_{\min}(\boldsymbol{\Lambda}_2+\underline{\delta}\boldsymbol{\Delta})\geq 0$. This contradicts with the maximum value of δ being zero, i.e., $\delta^*=0$. We thus complete the proof.

(ii) $\mathcal{Q}:=\mathbb{R}^{n\times p}$. Similar to Part (i) with $\mathcal{Q}=\mathcal{S}_+^n$, when $\delta^*=0$, we can show that $\operatorname{rank}(\boldsymbol{X}^*)\leq \tilde{k}+\lfloor\sqrt{2\tilde{m}+9/4}-3/2\rfloor$ using the proof of Theorem 3.

(iii) $\mathcal{Q}:=\mathcal{S}^n$. According to the proof of Theorem 6, similar to Part (i), whenever $\operatorname{rank}(\boldsymbol{X}^*)>\tilde{k}+\lfloor\sqrt{4\tilde{m}+9}-3\rfloor$, we can construct a nonzero matrix $\boldsymbol{Y}$ in (14) to move the solution $\boldsymbol{X}^*$ by a nonzero distance δ^* along the direction $\boldsymbol{Y}$, which contradicts the condition $\delta^*=0$. ☐

A.18 Proof of Corollary 9

Corollary 9 *For a domain set $\mathcal{D}$ with $\tilde{k}=1$ being defined in Definition 1, the rank-reduction Algorithm 2 reduces the solution's rank by at least one at each iteration.*

Proof. Given $\mathcal{Q} := \mathcal{S}_+^n$, according to the convergence analysis of Algorithm 2 in the proof of Theorem 11, at each iteration, we have that either the sum of the $(r - \tilde{k} + 1)$ smallest positive eigenvalues strictly decreases (i.e., the inequality (31) is strict) or the rank strictly reduces (i.e., $\text{rank}(\boldsymbol{X}(\delta^*)) < r$). If $\tilde{k} = 1$, then the sum of its r smallest nonzero eigenvalues is exactly the trace of new solution $\boldsymbol{X}(\delta^*)$ that always stays the same as matrix $\boldsymbol{X}^*$. Therefore, the rank must strictly reduce at each iteration, i.e., $\text{rank}(\boldsymbol{X}(\delta^*)) < \text{rank}(\boldsymbol{X}^*)$. The similar analysis follows for the non-symmetric or symmetric indefinite matrix space. □

Appendix B. Tightness of Rank Bounds for LSOP-R

This section presents the worst-case examples of LSOP-R in which the rank bounds in Theorems 2, 4, 5, 8 and 9 are tight, i.e., the rank bounds are attainable by these worst-case instances.

Let us begin with the following key lemma.

Lemma 3 *For any matrix $\boldsymbol{C} \in \mathbb{R}^{n\times n}$, the following results hold for the nuclear norm of $\boldsymbol{C}$*

(i) $\|\boldsymbol{C}\|_* \geq \sum_{i\in[n]} |C_{ii}|$; *and*

(ii) $\|\boldsymbol{C}\|_* \geq \sum_{i\in[n]\setminus S} |C_{ii}| + \|\boldsymbol{C}_{S,S}\|_*$ *for all* $S \subseteq [n]$ *and* $|S| = 2$.

Note that for any set $S \subseteq [n]$, $\boldsymbol{C}_{S,S}$ denotes the submatrix of $\boldsymbol{C}$ with rows and columns in set S.

Proof. According to Li and Xie (2021)[lemma 1], the nuclear norm of matrix $\boldsymbol{C}$ is equal to

$$\|\boldsymbol{C}\|_* := \max_{\boldsymbol{U}\in\mathbb{R}^{n\times n}} \left\{\langle \boldsymbol{C}, \boldsymbol{U}\rangle : \|\boldsymbol{U}\|_2 \leq 1, \|\boldsymbol{U}\|_* \leq n\right\}. \tag{32}$$

Since $\boldsymbol{U} = \text{diag}(\text{sign}(C_{11}), \cdots \text{sign}(C_{nn}))$ is a feasible solution to problem (32) and leads to the objective value $\sum_{i\in[n]} |C_{ii}|$, we arrive at the inequality in Part (i).

In addition, for any subset $S \subseteq [n]$, the problem (32) is lower bounded by

$$\begin{aligned}\|\boldsymbol{C}\|_* \geq \|\boldsymbol{C}_{S,S}\|_* + \|\boldsymbol{C}_{[n]\setminus S,[n]\setminus S}\|_* := &\max_{\boldsymbol{U}\in\mathbb{R}^{2\times 2}} \left\{\langle \boldsymbol{C}_{S,S}, \boldsymbol{U}\rangle : \|\boldsymbol{U}\|_2 \leq 1, \|\boldsymbol{U}\|_* \leq 2\right\} \\ &+ \max_{\boldsymbol{U}\in\mathbb{R}^{(n-2)\times(n-2)}} \left\{\langle \boldsymbol{C}_{[n]\setminus S,[n]\setminus S}, \boldsymbol{U}\rangle : \|\boldsymbol{U}\|_2 \leq 1, \|\boldsymbol{U}\|_* \leq n-2\right\},\end{aligned}$$

where the inequality stems from the fact that the optimal solutions from both maximization problems above always compose a feasible solution to problem (32). By applying Part (i) to the nuclear norm of $\boldsymbol{C}_{[n]\setminus S,[n]\setminus S}$, we thus prove Part (ii). □

Next, we are ready to show that the proposed rank bounds for LSOP-R are tight.

Lemma 4 *Given an integer $\tilde{k} \leq k$ following Definition 1, suppose $r := \tilde{k} + \lfloor\sqrt{2\tilde{m} + 9/4} - 3/2\rfloor$. Then we have that*

(i) Given $\mathcal{Q} := \mathcal{S}_+^n$ in (1), there exists an LSOP-R example that contains a rank-r extreme point;

(ii) Given $\mathcal{Q} := \mathbb{R}^{n\times p}$ in (1), there exists an LSOP-R example that contains a rank-r extreme point;

(iii) Given $\mathcal{Q} := \mathcal{S}^n$ in (1), suppose that function $f(\cdot)$ in the domain set $\mathcal{D}$ is sign-invariant. Then there exists an LSOP-R example that contains a rank-r extreme point.

Proof. First, we have $\tilde{k} \le r \le n$. When $r = \tilde{k}$, the results trivially hold. The following proof that focuses on the case of $r \ge \tilde{k}+1$ splits into three parts.

Part (i). When $\mathcal{Q} := \mathcal{S}_+^n$, let us consider the following example.

Example 2 *Suppose the domain set $\mathcal{D} := \{\boldsymbol{X} \in \mathcal{S}_+^n : \|\boldsymbol{X}\|_2 \le 1, \operatorname{rank}(\boldsymbol{X}) \le k\}$ and $m = (r-k+1)(r-k+2)/2-1$ linear equations in LSOP:*

$$X_{ii} = \frac{1}{r-k+1}, \forall i \in [r-k], \quad X_{ij} = 0, \forall i,j \in [r-k+1]\times[r-k+1], i<j.$$

In Example 2's domain set $\mathcal{D}$, we have that $\tilde{k} = k$, $m = \tilde{m}$, and $r = \tilde{k} + \lfloor\sqrt{2\tilde{m}+9/4}-3/2\rfloor$. It has been proven by Tantipongpipat et al. (2019)[appendix B] that there exists a rank-r extreme point of the feasible set of corresponding LSOP-R.

Part (ii). When $\mathcal{Q} := \mathbb{R}^{n\times p}$, let us consider the following example.

Example 3 *Suppose the domain set $\mathcal{D} := \{\boldsymbol{X} \in \mathbb{R}^{n\times p} : \|\boldsymbol{X}\|_2 \le 1, \operatorname{rank}(\boldsymbol{X}) \le k\}$ with $n = p$ and $m = (r-k+1)(r-k+2)/2-1$ linear equations in LSOP:*

$$X_{ii} = \frac{1}{r-k+1}, \forall i \in [r-k], \quad X_{ij} = 0, \forall i,j \in [r-k+1]\times[r-k+1], i<j.$$

In Example 3's domain set $\mathcal{D}$, we also have $\tilde{k} = k$, $m = \tilde{m}$, and $r = \tilde{k} + \lfloor\sqrt{2\tilde{m}+9/4}-3/2\rfloor$. It is recognized in Li and Xie (2022)[lemma 5] that $\operatorname{conv}(\mathcal{D}) = \{\boldsymbol{X} \in \mathcal{S}_+^n : \|\boldsymbol{X}\|_2 \le 1, \|\boldsymbol{X}\|_* \le k\}$. Thus, in this example, the resulting LSOP-R admits the following feasible set

$$\mathcal{T} = \left\{\boldsymbol{X} \in \mathbb{R}^{n\times n} : \|\boldsymbol{X}\|_2 \le 1, \|\boldsymbol{X}\|_* \le k, X_{ii} = \frac{1}{r-k+1}, \forall i \in [r-k], X_{ij} = 0, \forall 1 \le i < j \le r-k+1\right\}.$$

Then, it suffices to show a rank-r extreme point in the feasible set $\mathcal{T}$ above. Specifically, let us construct a rank-r matrix $\boldsymbol{X}^* \in \mathbb{R}^{n\times n}$ below

$$\boldsymbol{X}^* := \begin{pmatrix} \frac{1}{r-k+1}\boldsymbol{I}_{r-k+1} & \boldsymbol{0}_{r-k+1,k-1} & \boldsymbol{0}_{r-k+1,n-r} \\ \boldsymbol{0}_{k-1,r-k+1} & \boldsymbol{I}_{k-1} & \boldsymbol{0}_{k-1,n-r} \\ \boldsymbol{0}_{n-r,r-k+1} & \boldsymbol{0}_{n-r,k-1} & \boldsymbol{0}_{n-r,n-r} \end{pmatrix}.$$

We prove the extremeness of matrix $\boldsymbol{X}^*$ by contradiction. Suppose that there exist two distinct points $\boldsymbol{X}_1, \boldsymbol{X}_2$ in set $\mathcal{T}$ such that $\boldsymbol{X}^*$ is equal to their convex combination, i.e.,

$$\exists 0 < \alpha < 1, \boldsymbol{X}^* = \alpha\boldsymbol{X}_1 + (1-\alpha)\boldsymbol{X}_2.$$

First, we show that $\mathrm{diag}(\boldsymbol{X}^*) = \mathrm{diag}(\boldsymbol{X}_1) = \mathrm{diag}(\boldsymbol{X}_2)$. The inclusion $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathcal{T}$ indicates $(X_1)_{ii} = (X_2)_{ii} = 1/(r-k+1)$ for all $i \in [r-k]$. Due to the constraint $\|\boldsymbol{X}\|_2 \leq 1$ in set $\mathcal{T}$, we must have $(X_1)_{ii} = (X_2)_{ii} = 1$ for all $i \in [r-k+2, r]$. In addition, according to Part (i) in Lemma 3, points $\boldsymbol{X}_1, \boldsymbol{X}_2$ must satisfy

$$\sum_{i\in[n]} |(X_1)_{ii}| \leq \|\boldsymbol{X}_1\|_* \leq k, \quad \sum_{i\in[n]} |(X_2)_{ii}| \leq \|\boldsymbol{X}_2\|_* \leq k,$$

which further results in equations $(X_1)_{ii} = (X_2)_{ii} = X^*_{ii}$ for all $i \in \{r-k+1\} \cup [r+1, n]$.

Next, if there is a pair $(i^*, j^*) \in [r-k+1] \times [r-k+1]$ and $i^* > j^*$ such that $(X_1)_{i^*j^*} \neq 0$ and $(X_2)_{i^*j^*} \neq 0$, then we let $S := \{i^*, j^*\}$. According to Part (ii) in Lemma 3, we have

$$\|\boldsymbol{X}_1\|_* \geq \sum_{i\in[n]\setminus S} |(X_1)_{ii}| + \|(\boldsymbol{X}_1)_{S,S}\|_* > \sum_{i\in[n]} |(X_1)_{ii}| = k,$$

where the greater inequality is because for any 2×2 submatrix $(\boldsymbol{X}_1)_{S,S}$ with only one nonzero off-diagonal entry, we must have $\|(\boldsymbol{X}_1)_{S,S}\|_* > \sum_{i\in S} |(X_1)_{ii}|$. Similarly, we can show $\|\boldsymbol{X}_2\|_* > k$. This thus forms a contraction with the nuclear norm constraint in set $\mathcal{T}$.

If there is a pair $(i^*, j^*) \in [r-k+2, r] \times [r-k+2, r]$ and $i^* \neq j^*$ such that $(X_1)_{i^*j^*} \neq 0$ and $(X_2)_{i^*j^*} \neq 0$, then we let $S := \{i^*, j^*\}$. Since $(X_1)_{i^*,i^*} = (X_1)_{j^*,j^*} = (X_2)_{i^*,i^*} = (X_2)_{j^*,j^*} = 1$, we have

$$(\boldsymbol{X}_1)_{S,S} := \begin{pmatrix} 1 & (X_1)_{i^*j^*} \\ (X_1)_{j^*i^*} & 1 \end{pmatrix}, \quad (\boldsymbol{X}_2)_{S,S} := \begin{pmatrix} 1 & (X_2)_{i^*j^*} \\ (X_2)_{j^*i^*} & 1 \end{pmatrix}.$$

Then, the simple calculation leads to $\|(\boldsymbol{X}_1)_{S,S}\|_2 > 1$ and $\|(\boldsymbol{X}_2)_{S,S}\|_2 > 1$, which violates the largest singular value constraint in set $\mathcal{T}$.

Finally, if there is a pair $(i^*, j^*) \in [r+1, n] \times [r+1, n]$ and $i^* \neq j^*$ such that $(X_1)_{i^*j^*} \neq 0$ and $(X_2)_{i^*j^*} \neq 0$, then we let $S := \{i^*, j^*\}$. Since $(X_1)_{i^*,i^*} = (X_1)_{j^*,j^*} = (X_2)_{i^*,i^*} = (X_2)_{j^*,j^*} = 0$, we have

$$(\boldsymbol{X}_1)_{S,S} := \begin{pmatrix} 0 & (X_1)_{i^*j^*} \\ (X_1)_{j^*i^*} & 0 \end{pmatrix}, \quad (\boldsymbol{X}_2)_{S,S} := \begin{pmatrix} 0 & (X_2)_{i^*j^*} \\ (X_2)_{j^*i^*} & 0 \end{pmatrix}.$$

Similarly, using Part (ii) of Lemma 3, we can prove $\|(\boldsymbol{X}_1)_{S,S}\|_* > \sum_{i\in S} |(X_1)_{ii}|$, $\|\boldsymbol{X}_1\|_* > k$ and so does matrix $\boldsymbol{X}_2$, which is a contradiction.

Combining the above results together, we must have $(X_1)_{i,j} = 0$ and $(X_2)_{i,j} = 0$ for all $i \neq j$, which means $\boldsymbol{X}^* = \boldsymbol{X}_1 = \boldsymbol{X}_2$. Thus, we complete the proof, provided that $\boldsymbol{X}^*$ is a rank-r extreme point in the feasible set $\mathcal{T}$.

Part (iii). When function $f(\cdot)$ in the domain set $\mathcal{D}$ with $\mathcal{Q} := \mathcal{S}^n$ is sign-invariant, we instead use the absolute eigenvalues, i.e., singular values; thus, the analysis of Part (ii) above can be readily extended, expecting replacing the non-symmetric matrix space with symmetric one. That is, we can construct a worst-case example below where a rank-r extreme point exists.

Example 4 *Suppose the domain set* $\mathcal{D} := \{\boldsymbol{X} \in \mathcal{S}^n : \|\boldsymbol{X}\|_2 \le 1, \operatorname{rank}(\boldsymbol{X}) \le k\}$ *and* $m = (r-k+1)(r-k+2)/2 - 1$ *linear equations in LSOP:*

$$X_{ii} = \frac{1}{r-k+1}, \forall i \in [r-k], \quad X_{ij} = 0, \forall i,j \in [r-k+1] \times [r-k+1], i < j.$$

□

Theorem 12 *Given a sparse domain set* $\mathcal{D}$ *in* (3) *and an integer* $\tilde{k} \le k$ *following Definition 1, we have that*

(i) If function $f(\cdot)$ *in* (3) *is sign-invariant, then there exists an LSOP-R example that contains a rank-*$(\tilde{k}+\tilde{m})$ *extreme point; and*

(ii) There exists an LSOP-R-I example that contains a rank-r *extreme point.*

Proof. Let us introduce an example below where function $f(\cdot)$ is sign-invariant.

Example 5 *Suppose a sparse domain set* $\mathcal{D} := \{\boldsymbol{X} \in \mathcal{S}^n : \boldsymbol{X} = \operatorname{Diag}(\operatorname{diag}(\boldsymbol{X})), \operatorname{rank}(\boldsymbol{X}) \le k, \|\boldsymbol{X}\|_2 \le 1\}$ *and* $m = \tilde{m}$ *linear equations in LSOP:*

$$X_{ii} = \frac{1}{m+1}, \forall i \in [m],$$

which leads to $\mathcal{D} := \{\boldsymbol{X} \in \mathcal{S}^n : \boldsymbol{X} = \operatorname{Diag}(\operatorname{diag}(\boldsymbol{X})), \|\boldsymbol{X}\|_* \le k, \|\boldsymbol{X}\|_2 \le 1\}$.

Then, we can show that $\boldsymbol{X}^* = \operatorname{diag}(\boldsymbol{x}^*)$ is rank-$(k+\tilde{m})$ and is an extreme point in the feasible set of LSOP-R, where $x_i^* = 1/(m+1)$ for all $i \in [m+1]$, $x_i^* = 1$ for all $i \in [m+2, m+k]$, and $x_i^* = 0$ for all $i \in [m+k, n]$. Thus, the rank bound in Theorem 8 is tight.

Theorem 9 provides an identical rank bound for any sparse domain set in (3); hence, Example 5 also serves as a worst-case to show the tightness of Theorem 9. □