Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Convexification of Permutation-Invariant Sets and an Application to Sparse Principal Component Analysis

Jinhak Kim, Mohit Tawarmalani, Jean-Philippe P. Richard

To cite this article:
Jinhak Kim, Mohit Tawarmalani, Jean-Philippe P. Richard (2022) Convexification of Permutation-Invariant Sets and an Application to Sparse Principal Component Analysis. Mathematics of Operations Research 47(4):2547-2584. https://doi.org/10.1287/moor.2021.1219

https://pubsonline.informs.org/journal/moor

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 47, No. 4, November 2022, pp. 2547–2584
ISSN 0364-765X (print), ISSN 1526-5471 (online)

Convexification of Permutation-Invariant Sets and an Application to Sparse Principal Component Analysis

Jinhak Kim,[a] Mohit Tawarmalani,[b] Jean-Philippe P. Richard[c]

[a] College of Business, Northern Illinois University, DeKalb, Illinois 60115; [b] Krannert School of Management, Purdue University, West Lafayette, Indiana 47907; [c] Industrial and Systems Engineering, University of Minnesota, Minneapolis, Minnesota 55455

Contact: jkim23@niu.edu, https://orcid.org/0000-0002-1957-4798 (JK); mtawarma@purdue.edu, https://orcid.org/0000-0003-3085-0084 (MT); jrichar@umn.edu, https://orcid.org/0000-0001-5641-0939 (J-PPR)

Received: October 6, 2019
Revised: November 12, 2020; August 6, 2021
Accepted: October 5, 2021
Published Online in Articles in Advance: December 29, 2021

MSC2020 Classification: Primary: 47N10; secondary: 90C26, 52A27

https://doi.org/10.1287/moor.2021.1219

Abstract. We develop techniques to convexify a set that is invariant under permutation and/or change of sign of variables and discuss applications of these results. First, we convexify the intersection of the unit ball of a permutation and sign-invariant norm with a cardinality constraint. This gives a nonlinear formulation for the feasible set of sparse principal component analysis (PCA) and an alternative proof of the K-support norm. Second, we characterize the convex hull of sets of matrices defined by constraining their singular values. As a consequence, we generalize an earlier result that characterizes the convex hull of rank-constrained matrices whose spectral norm is below a given threshold. Third, we derive convex and concave envelopes of various permutation-invariant nonlinear functions and their level sets over hypercubes, with congruent bounds on all variables. Finally, we develop new relaxations for the exterior product of sparse vectors. Using these relaxations for sparse PCA, we show that our relaxation closes 98% of the gap left by a classical semidefinite programming relaxation for instances where the covariance matrices are of dimension up to 50×50.

Funding: The second and third authors wish to acknowledge the support from the National Science Foundation Division of Civil, Mechanical and Manufacturing Innovation [Grants 1727989 and 1917323].

Keywords: convexification • permutation-invariant sets • majorization • sparse PCA

1. Introduction

This paper is concerned with convexification of permutation-invariant sets. A set $S \subseteq \mathbb{R}^n$ is *permutation invariant* if $x \in S$ implies that $Px \in S$ for all n-dimensional permutation matrices P. Permutation-invariant sets appear in a variety of optimization problems. To support this claim and highlight the relevance of our construction, we next provide a variety of example applications where exploiting permutation invariance proves fruitful. Consider first sparse principal component analysis (PCA), a problem first introduced in Zou et al. [34], which consists in finding sparse vectors that explain the most variance in a data set. Specifically, the problem of finding the first sparse principal component can be formulated as $\max\{x^\top \Sigma x \,|\, x \in S\}$, where $S = \{x \in \mathbb{R}^n | \text{card}(x) \le K, \|x\| \le 1\}$, $\text{card}(x)$ is the number of nonzero components of x, and Σ is the covariance matrix of the given data; see, for example d'Aspremont et al. [7]. The feasible set of this model is permutation invariant because $\text{card}(Px) = \text{card}(x)$ and $\|Px\| = \|x\|$ for any vector $x \in \mathbb{R}^n$ and permutation matrix P. The convex hull of S is the unit ball associated with the K-support norm (Argyriou et al. [2]), a result that yields a tighter approximation of S compared with the elastic net $\{x \in \mathbb{R}^n \,|\, \|x\|_1 \le \sqrt{K}, \|x\| \le 1\}$ where $\|\cdot\|_1$ is the l_1-norm. Recognizing the set as permutation invariant allows for a streamlined derivation of these results. Second, observe that permutation invariance also arises when studying sets of matrices as the rank of a matrix can be thought of as a permutation-invariant cardinality constraint on the singular values. This equivalent formulation suggests that the applicability of this concept extends to sets of matrices whose singular values belong to a permutation-invariant set. For example, Hiriart-Urruty and Le [13] considers the set $\{M \in \mathcal{M}^{m,n}(\mathbb{R}) \,|\, \text{rank}(M) \le K, \|M\|_{sp} \le r\}$, where $\mathcal{M}^{m,n}(\mathbb{R})$ is the set of m-by-n real matrices, and $\|M\|_{sp}$ is the spectral norm of M. Because the spectral norm is the largest singular value, it follows easily that the singular values of these matrices belong to a permutation-invariant set. Using a permutation-invariant perspective to study such sets helps to generalize the results of Hiriart-Urruty and Le [13] in various ways. Finally, observe that a variety of sets with specific structures have been studied in nonconvex optimization because of their uses in creating convex relaxations of more general problems. For example, specialized techniques have been used to identify the convex hull of the multilinear monomial $\prod_{i=1}^n x_i$ over $[0,1]^n$ and $[-1,1]^n$ (Luedtke et al. [23],

Rikun [30]). We will show that these results can be obtained as special cases of our general approach to convexifying permutation-invariant sets. Besides, we will show that we can obtain hitherto unknown polynomial-sized convex hull descriptions for various commonly occurring sets in global optimization.

We now describe why permutation invariance is a useful property to consider while constructing convex hulls of sets. To see this, let $\Delta_\pi := \{x \mid x_{\pi(1)} \geq \cdots \geq x_{\pi(n)}\}$, where π is an n-dimensional permutation, and denote by $\text{conv}(S)$ the convex hull of a given set S. Because S can be expressed as a union of $n!$ sets of the form $S \cap \Delta_\pi$, if there is a (possibly lifted) polynomial-sized representation of $\text{conv}(S \cap \Delta_\pi)$ for each π, disjunctive programming (Balas [3]) can be used to represent $\text{conv}(S)$ in a higher-dimensional space. Unfortunately, this representation is exponentially sized in n, because the construction creates copies, one for each π, for each of the variables $(x_1, \ldots, x_n)$. What is remarkable about the permutation invariance is that by exploiting this property, we will show that a much more compact formulation can be derived. To appreciate the significance of this construction, we first argue that if S is not permutation invariant, such a compact formulation for $\text{conv}(S)$ may not even exist. To illustrate this point, we consider the set $\mathbb{B}$, which is a face of the Boolean-quadric polytope and is defined as $\mathbb{B} := \{xx^\top \mid x \in \{0,1\}^n, x_1 = 1\}$. If we order all the $\frac{n(n+1)}{2}$ products, these imply an order for all x_j variables because $x_1 x_j = x_j$ for $j = 2, \ldots, n$. Moreover, if $x_j \leq x_i$, it follows that $x_j x_i = x_j$. Therefore, the ordering of the $\frac{n(n+1)}{2}$ bilinear terms reduces to that of all x_j variables, possibly with some equalities. In other words, it suffices to consider $\mathbb{B} \cap \Delta'_\pi$, where $\Delta'_\pi = \{xx^\top \mid x_{\pi(r)} x_{\pi(i)} \geq x_{\pi(k)} x_{\pi(j)}$ whenever $\max\{r,i\} \leq \max\{k,j\}\}$ is a lifting of Δ_π and $\pi(1) = 1$. Because $x \in \{0,1\}^n \cap \Delta_\pi$ implies that $x_{\pi(r)} x_{\pi(i)} = x_{\pi(\max\{r,i\})}$, we can write $\text{conv}(\mathbb{B} \cap \Delta'_\pi) = \{X \in [0,1]^{n \times n} \mid X_{ij} = X_{1,\max\{i,j\}}, X_{1\cdot} \in \Delta_\pi, X_{11} = 1\}$. However, it is known that $\text{conv}(\mathbb{B})$ does not have a polynomial-sized formulation (Fiorini et al. [8]). In contrast, treating permutation-invariant sets S in this way provides a significant advantage because the sets $S \cap \Delta_\pi$ are congruent to one another. Exploiting this fact, we show that it is possible to construct a polynomial-sized extended formulation for S whenever a polynomial-sized formulation exists for $\text{conv}(S \cap \Delta_\pi)$. Our construction makes use of well-known extended formulations of a permutahedron alongside the convex hull of $S \cap \Delta_\pi$. The outline of the construction is as follows: first, we consider a permutation-invariant set S and assume that its convex hull over

$$\Delta^n := \{x \in \mathbb{R}^n \mid x_1 \geq \cdots \geq x_n\} \tag{1}$$

has a polynomial description. Then, the convex hull is simply the union of permutahedra where each permutahedron is generated by a point in $\text{conv}(S \cap \Delta^n)$. Each permutahedron can then be modeled using a polynomial number of linear equalities and inequalities to provide an extended formulation for $\text{conv}(S)$. The techniques involved apply to other settings. For example, they can be used to obtain convex hulls of sign-invariant sets using convex hull representations of $S \cap \{x \mid x \geq 0\}$.

The remainder of this paper is organized as follows. We present basic convexification results for permutation- and/or sign-invariant sets in Section 2. We then explore various applications of the results in the ensuing sections. In Section 3, we derive the convex hull of the intersection of a unit ball associated with a permutation-invariant norm and a cardinality constraint. The resulting convex hull defines another norm for which we give an explicit formula. As a result, we show that it is simple to determine whether an arbitrary point belongs to the convex hull, and to construct a separating hyperplane when it does not. We study the connection between permutation-invariant sets and sets of matrices characterized by their singular values. Furthermore, we investigate the semidefinite-representability of rank-constrained sets of matrices.

In Section 4, we develop convex and concave envelope characterizations of various permutation-invariant functions and sets described using such functions. For example, we derive the convex hull of the lower level set of a Schur-concave function, which is convex when all but one variable are fixed, a lifted representation of the convex hull of the graph of $\prod_{i=1}^n x_i$ over $[a,b]^n$, where $a, b \in \mathbb{R}$, and the convex hull of $\prod_{i=1}^m y_i^\alpha \geq \prod_{i=1}^n x_j^\beta$ over $[c,d]^m \times [a,b]^n$ with $a, c \geq 0$ and $\alpha, \beta > 0$. We show numerically that, for general a and b, the convex hull of $\prod_{i=1}^n x_i$ over $[a,b]^n$ is much tighter than the widely used recursive McCormick relaxation, in contrast to the well-known fact that the recursive McCormick procedure yields the convex hull of $\prod_{i=1}^n x_i$ when $a = -1$ and $b = 1$ (Luedtke et al. [23]).

In Section 5, we study the set of rank-one matrices whose generating vectors lie in a permutation-invariant set. We construct semidefinite programming (SDP) relaxations of the convex hull by proposing various valid inequalities derived from the rank-one condition of the matrix and the fact that every extreme point of a permutahedron generated by a vector is a permutation of the generating vector. Finally, we perform computational experiments with our relaxation for sparse PCA on several instances taken from the literature and other randomly generated instances. We compare our results to the relaxations proposed by d'Aspremont et al. [7] and Bertsimas et al. [6].

To increase the readability of this paper, we include a table of some frequently used notations in the appendix.

2. Convex Hull of Permutation-Invariant and Sign-Invariant Sets

In this section, we show that the convex hulls of permutation-invariant and sign-invariant sets can be readily constructed if their convex hulls over a fundamental subdomain are known. Given a set S, we use the notation $\text{int}(S)$ to represent its interior, $\text{vert}(S)$ to denote the set of its extreme points, and $\text{conv}(S)$ to represent its convex hull.

For a positive integer k, we denote the set of k-by-k permutation matrices by $\mathcal{P}_k$. Given a positive integer n and a nonnegative integer p, a set $S \subseteq \{(x,z) \in \mathbb{R}^n \times \mathbb{R}^p\}$ is called *permutation invariant with respect to* x if $(x,z) \in S$ implies that $(Px,z) \in S$ for all permutation matrices $P \in \mathcal{P}_n$. When $S \subseteq \{x \in \mathbb{R}^n\}$ is permutation invariant with respect to x, we simply say that S is permutation invariant. A real-valued function $(x,z) \longmapsto f(x,z)$ with $(x,z) \in \mathbb{R}^n \times \mathbb{R}^p$ is called *permutation invariant with respect to* x if $f(x,z) = f(Px,z)$ for all permutation matrices $P \in \mathcal{P}_n$. When $f : x \longmapsto f(x)$ is permutation invariant with respect to x, we say that f is permutation invariant. Moreover, any permutation-invariant set S can be written as the lower-level set $S = \{(x,z) \mid f(x,z) \le 0\}$ of a permutation-invariant function f by choosing this function to be the indicator function of the S, which takes value zero on the set and ∞ otherwise.

A set $S \subseteq \{(x,z) \in \mathbb{R}^n \times \mathbb{R}^p\}$ where n is a positive integer and p is a nonnegative integer is called *sign invariant* with respect to x if $(x,z) \in S$ implies that $(\bar{x},z) \in S$ for all $\bar{x}$ that satisfy $|\bar{x}| = |x|$.

Lemma 1 describes an important property of the convex hull of sets that are closed under certain linear transformations of the coordinates of their elements.

Lemma 1. *Let $T \in \mathcal{M}^{n,n}(\mathbb{R})$, and let $S \subseteq \mathbb{R}^n$ be such that for each $x \in S$, $Tx \in S$ as well. Then, if $x \in \text{conv}(S)$, $Tx \in \text{conv}(S)$.*

Proof. The result follows because $TS \subseteq S$ implies that $T\text{conv}(S) = \text{conv}(TS) \subseteq \text{conv}(S)$. □

It follows from Lemma 1 that if S is permutation invariant (respectively, sign invariant), then $\text{conv}(S)$ is also permutation invariant (respectively, sign invariant).

For each $x \in \mathbb{R}^n$, we denote the ith largest component of x by $x_{[i]}$ for $i = 1, \dots, n$.

Definition 1. Given two vectors $x, y \in \mathbb{R}^n$, we say that x *majorizes* y, a property we denote by $x \succeq_m y$ if $\sum_{i=1}^{j} x_{[i]} \ge \sum_{i=1}^{j} y_{[i]}$ for $j = 1, \dots, n-1$, and $\sum_{i=1}^{n} x_{[i]} = \sum_{i=1}^{n} y_{[i]}$. We say that y *is weakly submajorized by* x, and denote this relation as $x \succeq_{wm} y$ if $\sum_{i=1}^{j} x_{[i]} \ge \sum_{i=1}^{j} y_{[i]}$, $\forall j = 1, \dots, n$. For simplicity of notation, we shall refer to this relation as *weak majorization* of y by x.

The result of Lemma 2 relates majorization and permutation. Its proof follows from combining Hardy, Littlewood, and Pólya's theorem (Hardy et al. [12]) with Birkhoff's theorem; see Marshall et al. [24, theorems 2.B.2 and 2.A.2].

Lemma 2 (Marshall et al. [24, corollary 2.B.3], Rado [29]). *For $x, y \in \mathbb{R}^n$, $x \succeq_m y$ if and only if y is a convex combination of x and its permutations.*

An extension of majorization, which is known as G-majorization, introduced by Rado [29] in the context of a group of transformations, is defined using the property in Lemma 2 as the set of all doubly stochastic matrices from a semigroup; see Marshall et al. [24, section 14.C] for more detail and references about G-majorization.

Lemma 3. *Let Z be a convex subset of $\mathbb{R}^n \times \mathbb{R}^p$. Then*

$$Y := \left\{ (x,u,z) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^p \;\middle|\; \begin{array}{l} (u,z) \in Z, \\ u \succeq_m x, \\ u_1 \ge \cdots \ge u_n \end{array} \right\}$$

is convex.

Proof. First, observe that $\sum_{i=1}^{j} u_{[i]} = \sum_{i=1}^{j} u_i$ because $u_1 \ge \cdots \ge u_n$. Furthermore, $\sum_{i=1}^{j} x_{[i]}$ is a convex function being the maximum of all possible sums of j elements chosen from x. Next, $\sum_{i=1}^{n} x_{[i]} = \sum_{i=1}^{n} x_i$ and is, therefore, linear. Therefore, Y has the following convex representation:

$$Y = \left\{ (x,u,z) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^p \;\middle|\; \begin{array}{l} (u,z) \in Z, \\ \sum_{i=1}^{j} u_i \ge \sum_{i=1}^{j} x_{[i]}, \text{for } j = 1, \dots, n-1, \\ \sum_{i=1}^{n} u_i = \sum_{i=1}^{n} x_i, \\ u_1 \ge \cdots \ge u_n \end{array} \right\}. \quad \square$$

We next present Theorem 1, which gives an explicit description of the convex hull of a permutation-invariant set when an explicit description of the convex hull of its intersection with the cone $x_1 \geq \cdots \geq x_n$ is available. This description only requires the introduction of a copy u of the variables x together with majorization constraints.

Theorem 1. *Suppose $S \subseteq \{(x,z) | \mathbb{R}^n \times \mathbb{R}^p\}$ is permutation invariant with respect to $x \in \mathbb{R}^n$. Then*

$$\operatorname{conv}(S) = X := \left\{ (x,z) \,\middle|\, \begin{array}{l} (u,z) \in \operatorname{conv}(S_0), \\ u \geq_m x \end{array} \right\}, \tag{2}$$

where $S_0 = S \cap \{(u,z) \,|\, u_1 \geq \cdots \geq u_n\}$.

Proof. To prove that X is convex using Lemma 3, it suffices to show that $(u,z) \in \operatorname{conv}(S_0)$ implies $u \in \Delta^n$. This is clear because $\operatorname{conv}(S_0) \subseteq \operatorname{conv}(S) \cap \{(u,z) \mid u_1 \geq \cdots \geq u_n\}$.

We now show that $S \subseteq X$. As X is convex, this will also show that $\operatorname{conv}(S) \subseteq X$. Consider an arbitrary $(x,z) \in S$ and define u as $u_i = x_{[i]}$ for $i = 1, \ldots, n$. Then, $(u,z) \in S_0$ because S is permutation invariant and u is in descending order. Because $u \geq_m x$, $(x,z) \in X$.

We next prove that $X \subseteq \operatorname{conv}(S)$. Let $(x,z) \in X$. We show that this point can be expressed as a convex combination of points in S. Because $(x,z) \in X$, there exists u such that $(u,z) \in \operatorname{conv}(S_0) \subseteq \operatorname{conv}(S)$ and $u \geq_m x$. It follows from the permutation invariance of S with respect to x and Lemma 1 that $\operatorname{conv}(S)$ is permutation invariant with respect to x. By Lemma 2, x can be written as $x = \sum_i \lambda_i (P_i u)$, where P_i is a permutation matrix, $\lambda_i \geq 0$, and $\sum_i \lambda_i = 1$. Therefore, $(x,z) = \left(\sum_i \lambda_i (P_i u), z\right) = \sum_i \lambda_i (P_i u, z)$, concluding the proof. □

We next present classical results that allow for a linear formulation of the majorization constraints; see Nemirovski [27, section 3.3.4] for a more thorough discussion. To model $\sum_{i=1}^{j} x_{[i]}$, we express it as the value function of the following optimization problem, where $x_1, \ldots, x_n \in \mathbb{R}$ and $j \in \{1, \ldots, n-1\}$:

$$\begin{array}{lll} \max & \displaystyle\sum_{i=1}^{n} x_i s_i & \\ \text{s.t.} & \displaystyle\sum_{i=1}^{n} s_i = j, & \\ & 0 \leq s_i \leq 1, & i = 1, \ldots, n. \end{array} \tag{3}$$

To model majorization constraints, we need to enforce that for all feasible s, $\sum_{i=1}^{n} x_i s_i$ does not exceed $\sum_{i=1}^{j} u_j$. By taking the dual of (3), we exchange the quantifier "for all" to "there exists" and obtain the following formulation, which is amenable to direct inclusion in the model:

$$\begin{array}{lll} \text{LS}(j): \min & jr + \displaystyle\sum_{i=1}^{n} t_i & \\ \text{s.t.} & x_i \leq t_i + r, & i = 1, \ldots, n, \\ & t_i \geq 0, & i = 1, \ldots, n, \end{array} \tag{4}$$

where the dual variables r and t correspond to the primal constraints $\sum_{i=1}^{n} s_i = j$ and $s \leq 1$, respectively. Because (3) is feasible, (4) exhibits no duality gap. The constraint $\sum_{i=1}^{j} u_i \geq \sum_{i=1}^{j} x_{[i]}$ is then modeled through the existence of an (r, t) that is feasible to (4) such that $\sum_{i=1}^{j} u_i \geq jr + \sum_{i=1}^{n} t_i$.

Theorem 2. *Suppose $S \subseteq \{(x,z) \in \mathbb{R}^n \times \mathbb{R}^p\}$ is permutation invariant with respect to x. Then,*

$$\operatorname{conv}(S) = \left\{ (x,z) \,\middle|\, \begin{array}{ll} (u,z) \in \operatorname{conv}(S_0) & \\ u_1 \geq \cdots \geq u_n & \\ \sum_{i=1}^{n} u_i = \sum_{i=1}^{n} x_i & \\ \sum_{i=1}^{j} u_i \geq j r^j + \sum_{i=1}^{n} t_i^j, & j = 1, \ldots, n-1, \\ x_i \leq t_i^j + r^j, & j = 1, \ldots, n-1,\ i = 1, \ldots, n, \\ t_i^j \geq 0, & j = 1, \ldots, n-1,\ i = 1, \ldots, n \end{array} \right\}, \tag{5}$$

where $S_0 = S \cap \{(u,z) \,|\, u_1 \geq \cdots \geq u_n\}$.

Remark 1. In fact, Theorems 1 and 2 remain valid for any choice of S_0 that satisfies

$$\operatorname{conv}(S) \cap \{(u,z) \mid u_1 \geq \cdots \geq u_n\} \supseteq S_0 \supseteq S \cap \{(u,z) \mid u_1 \geq \cdots \geq u_n\}.$$

The formulation given in (5) expresses $\operatorname{conv}(S)$ as the projection of a convex set with n^2+n+p variables. This formulation is much smaller than that which would have been obtained using a classical application of disjunctive programming (see Balas [3, chapter 2]). An even smaller representation is possible using a more compact formulation of the permutahedron. Goemans [10] proposed such an extended formulation for the permutahedron, the convex hull of all possible permutations of a fixed vector $u \in \mathbb{R}^n$ using a sorting network where the numbers of variables and inequalities of the extended formulation depend on the number of comparators of the associated sorting network. When the Ajtai–Komlós–Szemerédi sorting network (Ajtai et al. [1]) is used, the extended formulation for the permutahedron has $\Theta(n \log n)$ variables and inequalities. As imposing the majorization constraint $u \geq_m x$ is equivalent to requiring that x belongs to the permutahedron generated by u, alternative extended formulations of $\operatorname{conv}(S)$ can be obtained by replacing $u \geq_m x$ in (2) with such extended formulations. This results in a formulation of $\operatorname{conv}(S)$ that is more compact than (5). For relaxations of sparse principal component analysis, we show in Section 5.2 that this smaller formulation provides some computational benefit on larger instances in our test set.

The ideas underlying the proof of Theorem 2 can also be applied when sets are invariants with respect to collections of linear transformations that are not permutation matrices. In particular, we describe next a related convexification result for sign-invariant sets.

Theorem 3. *Suppose $S \subseteq \{(x,z) \in \mathbb{R}^n \times \mathbb{R}^p\}$ is sign invariant with respect to x. Then,*

$$\operatorname{conv}(S) = X := \{(x,z) \mid (u,z) \in \operatorname{conv}(S_0), u \geq |x|\}, \tag{6}$$

where $S_0 = S \cap (\mathbb{R}^n_+ \times \mathbb{R}^p)$.

Proof. Set X is convex because it is the projection of an intersection of two convex sets. We now show that $S \subseteq X$. For an arbitrary $(x,z) \in S$, define $u = |x|$. By sign invariance of S, $(u,z) \in S$, and hence $(u,z) \in S_0 \subseteq \operatorname{conv}(S_0)$. By definition, u satisfies $u \geq |x|$. This shows that $\operatorname{conv}(S) \subseteq X$.

We next show that $X \subseteq \operatorname{conv}(S)$. Let $(x,z) \in X$. There exists $u \in \mathbb{R}^n$ such that $(u,z) \in \operatorname{conv}(S_0) \subseteq \operatorname{conv}(S)$ and $u \geq |x|$. Because $\operatorname{conv}(S)$ is sign invariant by Lemma 1, it follows that $\{(\bar{x},z) \mid \bar{x}_i \in \{u_i, -u_i\}\} \subseteq \operatorname{conv}(S)$. Therefore, $(x,z) \in \{(\bar{x},z) \mid |\bar{x}_i| \leq u_i\} \subseteq \operatorname{conv}(S)$, where the containment follows from the convexity of $\operatorname{conv}(S)$. □

Next, we consider the case where the target set S is both sign invariant and permutation invariant. Although two separate representations (2) and (6) for the convex hull already exist, we can derive a unified representation. Inequality $u \geq_m |x|$ would not be a useful description because $\sum_{i=1}^n u_i = \sum_{i=1}^n |x_i|$ is a nonconvex constraint. In fact, such a set would not even be convex, as can be seen from $S = \{-1, 1\}$. Instead, we prove that the convex hull can be represented using weak majorization. More precisely, suppose that $S \subseteq \{(x,z) \mid \mathbb{R}^n \times \mathbb{R}^p\}$ is sign and permutation invariant with respect to $x \in \mathbb{R}^n$. Then, $\operatorname{conv}(S) = X := \{(x,z) \mid (u,z) \in \operatorname{conv}(S_0), u \geq_{wm} |x|\}$, where $S_0 = S \cap \{(u,z) \mid u_1 \geq \cdots \geq u_n \geq 0\}$. We first prove that $S \subseteq X$. Consider an arbitrary $(x,z) \in S$, and define u as $u_i = |x|_{[i]}$ for $i = 1, \ldots, n$. Then, $(u,z) \in S_0 \subseteq \operatorname{conv}(S_0)$ and $u \geq_{wm} |x|$, showing the inclusion. We next prove $X \subseteq \operatorname{conv}(S)$. Consider an arbitrary $(x,z) \in X$. Then, there exists u such that $(u,z) \in \operatorname{conv}(S_0) \subseteq \operatorname{conv}(S)$ and $u \geq_{wm} |x|$. It follows that there exists u' such that $u \geq_m u'$ and $u' \geq |x|$; see Marshall et al. [24, result 5.A.9, p. 177], for example. Then, it can be shown that $(u',z) \in \operatorname{conv}(S)$ using the same arguments that were used in the proof of Theorem 1. Because $\operatorname{conv}(S)$ is sign invariant, this implies that all sign variants of $(u',z) \in \operatorname{conv}(S)$. Then, by the last part of the proof of Theorem 3, this implies that $(x,z) \in \operatorname{conv}(S)$. (This convex hull description can also be derived by first constructing $\operatorname{conv}(S \cap (\mathbb{R}^n_+ \times \mathbb{R}^p))$ using Theorem 1. Then, $\operatorname{conv}(S)$ is obtained by Theorem 3. The variables u' introduced by Theorem 3 can finally be projected using Marshall et al. [24, 5.A.9].)

The above convexification results can be easily extended to the sets that are permutation invariant or/and sign invariant with respect to multiple subsets of independent variables.

Theorem 4. *Let $S \subseteq \{(x^1, \ldots, x^m, z) \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_m} \times \mathbb{R}^p\}$.*

1. *Suppose S is a permutation-invariant set with respect to x^k for $k = 1, \ldots, m$. Then,*

$$\operatorname{conv}(S) = \left\{ (x^1, \ldots, x^m, z) \;\middle|\; \begin{array}{l} (u^1, \ldots, u^m, z) \in \operatorname{conv}(S_0), \\ u^k \geq_m x^k, k = 1, \ldots, m \end{array} \right\}, \tag{7}$$

where $S_0 = S \cap (\Delta^{n_1} \times \cdots \times \Delta^{n_m} \times \mathbb{R}^p)$.

2. *Suppose S is sign invariant with respect to x^k for $k = 1, \dots, m$. Then,*

$$\operatorname{conv}(S) = \left\{ (x^1, \dots, x^m, z) \,\middle|\, \begin{array}{l} (u^1, \dots, u^m, z) \in \operatorname{conv}(S_0), \\ u^k \geq |x^k|, k = 1, \dots, m \end{array} \right\}, \tag{8}$$

where $S_0 = S \cap (\mathbb{R}^{n_1}_+ \times \cdots \times \mathbb{R}^{n_m}_+ \times \mathbb{R}^p)$.

3. *Suppose S is permutation invariant and sign invariant with respect to x^k for $k = 1, \dots, m$. Then,*

$$\operatorname{conv}(S) = \left\{ (x^1, \dots, x^m, z) \,\middle|\, \begin{array}{l} (u^1, \dots, u^m, z) \in \operatorname{conv}(S_0), \\ u^k \geq_{wm} |x^k|, k = 1, \dots, m \end{array} \right\}, \tag{9}$$

where

$$S_0 = S \cap \left\{ (u^1, \dots, u^m, z) \,|\, u^k_1 \geq \cdots \geq u^k_{n_k} \geq 0,\ k = 1, \dots, m \right\}.$$

We remark that our convexification result on sign-invariant sets can also be used to convexify reflections on a hyperplane (Kaibel and Pashkovich [16]). In particular, consider a set S and assume that the set is closed under reflection on $\langle a, \cdot \rangle = 0$ for some a with $\|a\|_2 = 1$. Assume further that $S_0 = S \cap \{x \,|\, \langle a, x \rangle \geq 0\}$ is available. Consider any orthogonal matrix U that aligns a along the first principal direction, so that $Ua = e_1$. Observe that US is sign invariant with respect to variable x_1. Then, $\operatorname{conv}(US) = \{z \,|\, (t, z_2, \dots, z_n) \in U \operatorname{conv}(S_0), t \geq |z_1|\}$ and so, $\operatorname{conv}(S) = \{x \,|\, x + (t - z_1)a \in \operatorname{conv}(S_0), t \geq |z_1|\}$, where we have used the fact that $U^\top e_1 = a$.

3. Sparsity Theorem

In this section, we first study the convex hull of the set

$$N^K_{\|\cdot\|_s} = \{x \in \mathbb{R}^n \,|\, \|x\|_s \leq 1, \operatorname{card}(x) \leq K\}, \tag{10}$$

where $\|\cdot\|_s$ is a sign- and permutation-invariant norm (also known as a *symmetric gauge function*), and introduce the sparsity theorem (Theorem 6), which shows that the optimal value of

$$\begin{array}{ll} \min & \|u\|_s \\ \text{s.t.} & u_1 \geq \cdots \geq u_K \geq 0, \\ & u_{K+1} = \cdots = u_n = 0, \\ & u \geq_{wm} |x| \end{array} \tag{11}$$

is no more than one if and only if $x \in \operatorname{conv}(N^K_{\|\cdot\|_s})$ and that an optimal solution to (11) can be obtained in closed form. In other words, given $x \in \mathbb{R}^n$, Theorem 6 gives a closed-form expression for a sparse vector u, in terms of x, in the representative disjunct $\Delta^n \cap \mathbb{R}^n_+$ that weakly majorizes $|x|$ and minimizes $\|u\|_s$. In fact, this u majorizes $|x|$. This is not surprising. If the optimal u to (11) did not majorize $|x|$, there would exist u' such that $u \geq u' \geq_m |x|$; see Marshall et al. [24, 5.A.9]. Then, because u' is in the convex hull of u and its sign variants, it follows by sign invariance of $\|\cdot\|_s$ that $\|u'\|_s \leq \|u\|_s$, and so u' is also optimal and majorizes $|x|$.

We denote the set $\{x \in \mathbb{R}^n \,|\, \|x\|_s \leq r\}$ by $B_s(r)$. When $K = 1$, the convex hull is an ℓ_1-norm ball. The set is trivial when $K = n$. Therefore, we assume $1 < K < n$. When the associated norm $\|\cdot\|_s$ is the ℓ_2-norm, $N^K_{\|\cdot\|}$ is the feasible set of the sparse principal component analysis problem; see d'Aspremont et al. [7].

We define

$$\Delta^n_+ := \Delta^n \cap \mathbb{R}^n_+ \tag{12}$$

and, for any vector $x \in \mathbb{R}^n$, define

$$(x_{\Delta^n})_i = x_{[i]} \text{ and } (x_{\Delta^n_+})_i = |x|_{[i]}, \quad \text{for } i = 1, \dots, n.$$

When the dimension n of the set and the associated vector is clear in the context, we use the simpler notations $\Delta, \Delta_+, x_\Delta$, and x_{Δ_+}.

By sign and permutation invariance of the norm $\|\cdot\|_s$ and that of the cardinality requirement, $N^K_{\|\cdot\|_s}$ is sign and permutation invariant, and hence we can apply Theorem 4 to obtain its convex hull as a projection of a higher-dimensional set

$$\operatorname{conv}\left(N^K_{\|\cdot\|_s}\right) = \left\{ x \in \mathbb{R}^n \,\middle|\, \begin{array}{l} u \in N^K_{\|\cdot\|_s} \cap \Delta_+, \\ u \geq_{wm} |x| \end{array} \right\} = \left\{ x \in \mathbb{R}^n \,\middle|\, \begin{array}{l} \|u\|_s \leq 1, \\ u_1 \geq \cdots \geq u_K \geq 0, \\ u_{K+1} = \cdots = u_n = 0, \\ u \geq_{wm} |x| \end{array} \right\}. \tag{13}$$

The extended formulation (13) can be written in closed form with $O(nK)$ additional variables and constraints based on the modeling technique described in Section 2. Other extended formulations are proposed in Lim [20] and Lim and Wright [21] for the case where $\|\cdot\|_s$ is an ℓ_p-norm. In these papers, the formulations are obtained either through dynamic programming concepts or Goemans' [9] extended formulation of the permutahedron using a sorting network.

In this section, we describe the convex hull as a norm ball in the original variable space. The induced norm is easily calculable if the associated norm $\|\cdot\|_s$ is calculable. Moreover, given an arbitrary point in $\mathbb{R}^n$ not in the convex hull, we devise an algorithm to construct a separating hyperplane.

We first present the following lemma introduced in Li and Mehta [19].

Lemma 4. *Suppose $x \geq_m y$. Then, for any permutation-invariant norm $f(\cdot)$, $f(x) \geq f(y)$.*

A set in $\mathbb{R}^n$ is called a *convex body* if it is a compact convex set with nonempty interior. In the next proposition, we show that $\operatorname{conv}(N^K_{\|\cdot\|_s})$ is a convex body.

Proposition 1. *The set $\operatorname{conv}(N^K_{\|\cdot\|_s})$ is a convex body.*

Proof. Because $N^K_{\|\cdot\|_s}$ is a compact set, it follows that $\operatorname{conv}(N^K_{\|\cdot\|_s})$ is a compact convex set (Barvinok [4, corollary I.2.4]). To see that $\operatorname{conv}(N^K_{\|\cdot\|_s})$ has a nonempty interior, observe that there exists $\epsilon > 0$ such that $B_1(\epsilon) \subseteq B_s(1)$, where $B_1(\epsilon)$ represents the ℓ_1-norm ball with radius ϵ. This follows from the equivalence of norms in a finite vector space. Notice that $\operatorname{vert}(B_1(\epsilon)) = \{\pm\epsilon e_i \mid i = 1, \ldots, n\}$, where e_i is ith canonical vector. Because for any $x \in \operatorname{vert}(B_1(\epsilon))$, $\operatorname{card}(x) = 1$, it follows that $\operatorname{vert}(B_1(\epsilon)) \subseteq N^K_{\|\cdot\|_s}$, and, so, $B_1(\epsilon) \subseteq \operatorname{conv}(N^K_{\|\cdot\|_s})$. The result follows because $0 \in \operatorname{int}(B_1(\epsilon))$. □

It is well known that there exists a one-to-one correspondence between norms in $\mathbb{R}^n$ and convex bodies symmetric about zero and containing zero in their interior; see Matoušek [25, section 14.4] for instance. Given an arbitrary norm $\|\cdot\|$, we can construct its unit ball $\{x \mid \|x\| \leq 1\}$, which is a convex body of the desired type. Conversely, given any compact convex body C that is symmetric about zero and contains zero in its interior, we can define the function

$$f_C(x) := \min\left\{t > 0 \,\middle|\, \frac{x}{t} \in C\right\} \tag{14}$$

for $x \in \mathbb{R}^n$. It is known that the function f_C satisfies the properties of norms; see Matoušek [25, section 14.4], for example. Furthermore, the convex body C is a lower-level set of this norm, that is, $C = \{x \mid f_C(x) \leq 1\}$.

Because $\operatorname{conv}(N^K_{\|\cdot\|_s})$ is a compact convex body that is symmetric about zero and contains zero in its interior, a norm associated with $\operatorname{conv}(N^K_{\|\cdot\|_s})$ can be defined as in (14). We denote the corresponding norm by $\|\cdot\|_c$. Because $\|\cdot\|_c$ is sign and permutation invariant, the following result holds.

Proposition 2. *The set $\operatorname{conv}(N^K_{\|\cdot\|_s})$ is the unit ball associated with a sign- and permutation-invariant norm, that is, $\operatorname{conv}(N^K_{\|\cdot\|_s}) = B_c(1)$.*

We next show that the values of the norms $\|\cdot\|_c$ and $\|\cdot\|_s$ are the same for vectors that satisfy the cardinality constraint.

Proposition 3. *If $\operatorname{card}(x) \leq K$, $\|x\|_c = \|x\|_s$.*

Proof. We first show that $\|x\|_c \geq \|x\|_s$. Because $N^K_{\|\cdot\|_s} \subseteq B_s(1)$, it follows that $B_c(1) = \operatorname{conv}(N^K_{\|\cdot\|_s}) \subseteq B_s(1)$. This implies that $B_c(r) \subseteq B_s(r)$ for any r. In particular, when $r = \|x\|_c$, we have $x \in B_c(r) \subseteq B_s(r)$, which implies that $\|x\|_c \geq \|x\|_s$. We now show $\|x\|_c \leq \|x\|_s$ when $\operatorname{card}(x) \leq K$. Let $r = \|x\|_s$, and observe that $\frac{x}{r} \in N^K_{\|\cdot\|_s} \subseteq B_c(1)$. Therefore, $x \in B_c(r)$ or $\|x\|_c \leq \|x\|_s$. □

We present an explicit formula to evaluate $\|\cdot\|_c$. For an arbitrary $x \in \mathbb{R}^n$, define $s(x) \in \mathbb{R}^{K+2}$ as

$$s(x)_i = \frac{\sum_{j=i}^{n} |x|_{[j]}}{K - i + 1}, i = 1, \ldots, K; \; s(x)_0 = s(x)_{K+1} = \infty.$$

Let i_x be the minimum among those indices that minimize $s(x)_i$, and let $\delta(x) = s(x)_{i_x}$. Now, define $u(x) \in \mathbb{R}^n$ as

$$u(x)_i = \begin{cases} |x|_{[i]}, & \text{if} \quad i \in \{1, \ldots, i_x - 1\} \\ \delta(x), & \text{if} \quad i \in \{i_x, \ldots, K\} \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

In the following proposition, we show that, for arbitrary $x \in \mathbb{R}^n$, we can construct a vector $u(x) \in \Delta_+$ that satisfies the cardinality constraint and majorizes $|x|$.

Proposition 4. *Let $s(x), i_x, \delta(x)$, and $u(x)$ be defined as above. Then,*

1. $s(x)_{i+1} - s(x)_i = \frac{1}{K-i+1}(s(x)_{i+1} - |x|_{[i]}) = \frac{1}{K-i}(s(x)_i - |x|_{[i]})$ *for* $i = 1, \dots, K-1$,
2. $s(x)_1 \geq \cdots \geq s(x)_{i_x}$ *and* $s(x)_{i_x} \leq \cdots \leq s(x)_K$,
3. $u(x) \geq_m |x|$,
4. $u(x)_1 = \max\{|x|_{[1]}, s(x)_1\}$.

Proof. By definition of $s(x)$,

$$(K-i)s(x)_{i+1} - (K-i+1)s(x)_i = (K-i)\frac{\sum_{j=i+1}^n |x|_{[j]}}{K-i} - (K-i+1)\frac{\sum_{j=i}^n |x|_{[j]}}{K-i+1}$$
$$= \sum_{j=i+1}^n |x|_{[j]} - \sum_{j=i}^n |x|_{[j]} = -|x|_{[i]}.$$

By adding $s(x)_{i+1}$ on both sides, we have $(K-i+1)(s(x)_{i+1} - s(x)_i) = s(x)_{i+1} - |x|_{[i]}$, implying the first equality of part 1. The remainder of the part can be shown similarly by adding $s(x)_i$ on both sides.

To see part 2, if there is no index i in $\{1, \dots, K-1\}$ so that $s(x)_i < s(x)_{i+1}$, the result holds trivially. Assume now that i' is the smallest such index. Then, $s(x)_{i'+1} > s(x)_{i'} > |x|_{[i']} \geq |x|_{[i'+1]}$, where the second inequality follows because $s(x)_{i'+1} - s(x)_{i'} > 0$ implies $s(x)_{i'} > |x|_{[i']} > 0$ from part 1. This in turn shows, using part 1, that $s(x)_{i'+2} > s(x)_{i'+1}$ as long as $i' < K-1$. By induction, $s(x)_{i'} < \cdots < s(x)_K$, and by the definition of i', $s(x)_1 \geq \cdots \geq s(x)_{i'}$. Then, part 2 follows by defining i_x as the first index such that $s(x)_{i_x} = s(x)_{i'}$.

We next prove part 3. We first show that $u(x)$ is nonincreasing. If $i_x = 1$, then all components of $u(x)$ equal $\delta(x)$, and hence it is nonincreasing. Now, assume that $i_x \geq 2$. By definition of $u(x)$, it suffices to show that $|x|_{[i_x-1]} \geq \delta(x)$. Then, $s(x)_{i_x} - |x|_{[i_x-1]} = (K-i_x)(s(x)_{i_x} - s(x)_{i_x-1}) \leq 0$ by plugging in $i = i_x - 1$ in the first equality of part 1. Therefore, $|x|_{[i_x-1]} \geq s(x)_{i_x} = \delta(x)$, completing the proof that $u(x)$ is nonincreasing. Thus, $u(x)_{[i]} = u(x)_i$ for all $i = 1, \dots, n$. Next, observe that $\sum_{i=i_x}^n u(x)_i = \sum_{i=i_x}^n |x|_{[i]}$ by definition of $\delta(x)$. This implies in turn that $\sum_{i=1}^n u(x)_i = \sum_{i=1}^n |x|_{[i]}$. We next show that $\sum_{i=1}^j u(x)_i \geq \sum_{i=1}^j |x|_{[i]}$ for all $j = 1, \dots, n-1$. When $j = 1, \dots, i_x - 1$, the inequality holds with equality by definition of $u(x)$. We next consider the case $j \geq i_x$. If $i_x = K$, the inequality holds because $\sum_{i=1}^j u(x)_i = \sum_{i=1}^K u(x)_i = \sum_{i=1}^n u(x)_i = \sum_{i=1}^n |x|_{[i]} \geq \sum_{i=1}^j |x|_{[i]}$, where the first and the second equalities follow because $j \geq K$ and $u(x)_{K+1} = \cdots = u(x)_n = 0$. Now assume that $i_x < K$. Because $s(x)_{i_x+1} \geq s(x)_{i_x}$ and $s(x)_{i_x+1} - s(x)_{i_x} = \frac{1}{K-i_x}(s(x)_{i_x} - |x|_{[i_x]})$ by part 1, $s(x)_{i_x} \geq |x|_{[i_x]}$, and hence $\delta(x) = s(x)_{i_x} \geq |x|_{[i]}$ for all $i \geq i_x$. Therefore, $\sum_{i=1}^j u(x)_i - \sum_{i=1}^j |x|_{[i]} = \sum_{i=i_x}^j u(x)_i - \sum_{i=i_x}^j |x|_{[i]} = \sum_{i=i_x}^j (\delta(x) - |x|_{[i]}) \geq 0$.

For part 4, first assume $i_x = 1$. Then, $u(x)_1 = s(x)_1$. By part 1, $s(x)_1 \geq |x|_{[1]}$, and hence $u(x)_1 = \max\{|x|_{[1]}, s(x)_1\}$. Next, assume that $i_x \geq 2$. Then, $u(x)_1 = |x|_{[1]}$. By part 2, $s(x)_2 \leq s(x)_1$, and hence, by part 1, $s(x)_1 \leq |x|_{[1]}$. Therefore, $u(x)_1 = \max\{|x|_{[1]}, s(x)_1\}$. □

We next show in the following theorem that $u(x)$ can be used to compute $\|x\|_c$ if $\|\cdot\|_s$ is calculable. To this end, we introduce the following notations. Let $\hat{\beta}$ be an optimal solution to

$$\max\{\beta^\top u(x) \mid \|\beta\|_{s*} \leq 1\}, \tag{16}$$

where $\|\cdot\|_{s*}$ is the dual norm of $\|\cdot\|_s$. We now define θ and $\chi \in \mathbb{R}^n$ as follows:

$$\theta_i = \begin{cases} \hat{\beta}_i, & i = 1, \dots, i_x - 1, \\ \dfrac{\sum_{j=i_x}^K \hat{\beta}_j}{K - i_x + 1}, & i = i_x, \dots, K, \\ 0 & \text{otherwise,} \end{cases} \qquad \chi_j = \begin{cases} \hat{\beta}_i & \text{for } i = 1, \dots, i_x - 1, \\ \dfrac{\sum_{i=i_x}^K \hat{\beta}_i}{K - i_x + 1} & \text{otherwise.} \end{cases} \tag{17}$$

Theorem 5. *For an arbitrary $x \in \mathbb{R}^n$, let $u(x)$ be defined as in (15). Then, $\|x\|_c = \|u(x)\|_s$.*

Proof. Recall the definitions of $s(x)$, i_x, and $\delta(x)$. We have $\|x\|_c \leq \|u(x)\|_c = \|u(x)\|_s$, where the first inequality is because of part 3 of Proposition 4 and Lemma 4, whereas the equality is due to Proposition 3. We next show

$\|x\|_c \geq \|u(x)\|_s$. Let $r = \|u(x)\|_s$ so that $u(x) \in B_s(r)$. By the definition of $\|\cdot\|_s$ and $\hat{\beta}$, we have $\hat{\beta}^\top u(x) = r$. Moreover, $\hat{\beta}^\top u \leq \|u\|_s \leq r$ for all $u \in B_s(r)$, where the first inequality is because $\hat{\beta} \in B_{s^*}(1)$, and the second inequality is because $u \in B_s(r)$. Because $u(x) \in \Delta_+$, it follows from rearrangement inequality that, without loss of generality, we may assume that $\hat{\beta} \in \Delta_+$.

We next show that $\theta^\top u \leq r$ for all $u \in B_s(r)$, where θ is as defined in (17). Define $\bar{\beta} := (\hat{\beta}_1, \ldots, \hat{\beta}_K, 0, \ldots, 0)$ and $\check{\beta} := (\hat{\beta}_1, \ldots, \hat{\beta}_K, -\hat{\beta}_{K+1}, \ldots, -\hat{\beta}_n)$, and observe that $\bar{\beta} = \frac{\hat{\beta}+\check{\beta}}{2}$. By sign invariance of $\|\cdot\|_s$ and thus of $\|\cdot\|_{s^*}, \|\cdot\|_{s^*}$, $\check{\beta} \in B_{s^*}(1)$, and so $\bar{\beta} \in B_{s^*}(1)$. However, because $\bar{\beta} \geq_m \theta$, Lemma 4 shows that $\theta \in B_{s^*}(1)$. This in turn shows that $\theta^\top u \leq r$ is valid for $B_s(r)$.

We next claim that $\chi^\top u \leq 1$ is valid for $N^K_{\|\cdot\|_s}$, where χ is as defined in (17). Assume that, on the contrary, there exists $\hat{u} \in N^K_{\|\cdot\|_s}$ such that $\chi^\top \hat{u} > 1$. Because of the rearrangement inequality and the fact that $\chi \in \Delta_+$, we may assume that $\hat{u} \in \Delta_+$. This yields the contradiction $1 \geq \theta^\top \hat{u} = \chi^\top \hat{u} > 1$, where the first inequality is by the validity of $\theta^\top u \leq 1$ for $B_s(1)$, which outer-approximates $N^K_{\|\cdot\|_s}$, and the equality can be arranged by choosing $\hat{u}$ with support at most K. It follows that $\chi^\top u \leq r$ is valid for $B_c(r)$ or, in other words, that $\chi \in B_{c^*}(1)$, where $\|\cdot\|_{c^*}$ is the dual norm of $\|\cdot\|_c$. Therefore,

$$\|x\|_c = \|x_{\Delta_+}\|_c = \max\{\beta^\top x_{\Delta_+} \mid \|\beta\|_{c^*} \leq 1\} \geq \chi^\top x_{\Delta_+}, \tag{18}$$

where the inequality is because $\|\chi\|_{c^*} \leq 1$. However, the following calculation shows that:

$$\chi^\top x_{\Delta_+} = \sum_{i=1}^{i_x-1} \hat{\beta}_i |x|_{[i]} + \frac{\sum_{i=i_x}^{K} \hat{\beta}_i}{K - i_x + 1} \sum_{i=i_x}^{n} |x|_{[i]} = \sum_{i=1}^{i_x-1} \hat{\beta}_i |x|_{[i]} + \sum_{i=i_x}^{K} \hat{\beta}_i \frac{\sum_{i=i_x}^{n} |x|_{[i]}}{K - i_x + 1} = \hat{\beta}^\top u(x) = r. \tag{19}$$

Combining (18) and (19), we conclude that $\|x\|_c \geq \|u(x)\|_s$, where we have used that r was defined to be $\|u(x)\|_s$. □

Corollary 1. *For a fixed $x \in \mathbb{R}^n$, let χ be as defined in (17). Then, $\chi^\top u \leq 1$ is valid for $N^K_{\|\cdot\|_s}$, $\|\chi\|_{c^*} = 1$, and $\chi^\top x_{\Delta_+} = \|x\|_c$. In particular, $\chi^\top u \leq 1$ separates x_{Δ_+} if $x \notin \operatorname{conv}(N^K_{\|\cdot\|_s})$.*

Proof. It was already shown in the proof of Theorem 5 that $\chi^\top u \leq 1$ is valid for $N^K_{\|\cdot\|_s}$. Let $u(x)$ be as defined in (15). Then, $\chi^\top x_{\Delta_+} = \|u(x)\|_s = \|x\|_c$, where the first equality is from (19) and the second equality is from Theorem 5. Therefore, $\|x_{\Delta_+}\|_c = \chi^\top x_{\Delta_+} \leq \|\chi\|_{c^*} \|x_{\Delta_+}\|_c \leq \|x_{\Delta_+}\|_c$, where the first inequality is due to Cauchy–Schwarz inequality, and the second inequality is because the proof of Theorem 5 shows that $\|\chi\|_{c^*} \leq 1$. Therefore, equality holds throughout, and, in particular, $\|\chi\|_{c^*} = 1$. If $x \notin \operatorname{conv}(N^K_{\|\cdot\|_s}) = B_c(1)$, then $\chi^\top x_{\Delta_+} = \|x\|_c > 1$, where the inequality is because $x \notin B_c(1)$. □

Remark 2. In the proof of Corollary 1, let T be the transformation (a composition of sign conversions and permutations) that maps x to x_{Δ_+}. Then, the hyperplane that separates x and $N^K_{\|\cdot\|_s}$ is $T^{-1}(\chi)^\top u \leq 1$.

Theorem 6 (Sparsity Theorem). *For an arbitrary $x \in \mathbb{R}^n$, $u(x)$ as defined in (15) is an optimal solution to*

$$\begin{aligned} \min \quad & \|u\|_s \\ \text{s.t.} \quad & u_1 \geq \cdots \geq u_K \geq 0, \\ & u_{K+1} = \cdots = u_n = 0, \\ & u \geq_{wm} |x|. \end{aligned}$$

Proof. First, $u(x)$ is feasible because of its definition and part 3 of Proposition 4. Now, it suffices to show that $\|u(x)\|_s \leq \|u\|_s$ for any feasible solution u. Because $u \geq_{wm} |x|$, there exists u' such that $u \geq_m u'$ and $u' \geq |x|$; see Marshall et al. [24, 5.A.9], for example. Then, $\|u\|_s = \|u\|_c \geq \|u'\|_c \geq \|x\|_c = \|u(x)\|_s$, where the first equality follows from Proposition 3, the first inequality follows from Lemma 4, and the last equality follows from Theorem 5. Finally, to prove the second inequality, observe that $|x|$ can be written as a convex combination of u' and its sign invariants because $u' \geq |x|$. By the triangle inequality, positive homogeneity, and the sign invariance of $\|\cdot\|_c, \|x\|_c \leq \|u'\|_c$. □

Example 1. Consider the set $N^3_{\|\cdot\|_2}$, where $n = 6$. Let $x := \left(\frac{27}{28}, \frac{5}{28}, \frac{4}{28}, \frac{3}{28}, \frac{2}{28}, \frac{1}{28}\right)$. Note that $\|x\|_2 = 1$ and $x \in \Delta_+$. For illustration, we establish that $x \notin \operatorname{conv}(N^3_{\|\cdot\|_2})$ by constructing an explicit separating hyperplane described in (16)

and (17). First, we construct the vector $s(x) \in \mathbb{R}^3$ as follows:

$$s(x)_1 = \frac{\sum_{j=1}^{6} x_j}{3-1+1} = \frac{1}{3}\left(\frac{27}{28} + \frac{5}{28} + \frac{4}{28} + \frac{3}{28} + \frac{2}{28} + \frac{1}{28}\right) = \frac{28}{56},$$
$$s(x)_2 = \frac{\sum_{j=2}^{6} x_j}{3-2+1} = \frac{1}{2}\left(\frac{5}{28} + \frac{4}{28} + \frac{3}{28} + \frac{2}{28} + \frac{1}{28}\right) = \frac{15}{56},$$
$$s(x)_3 = \frac{\sum_{j=3}^{6} x_j}{3-3+1} = \frac{1}{1}\left(\frac{4}{28} + \frac{3}{28} + \frac{2}{28} + \frac{1}{28}\right) = \frac{20}{56}.$$

Observe that $s(x)_2 = \min\{s(x)_1, s(x)_2, s(x)_3\}$. Next, we compute that

$$u(x)_1 = x_1 = \frac{27}{28}, \quad u(x)_2 = u(x)_3 = s(x)_2 = \frac{15}{56}, \quad u(x)_4 = u(x)_5 = u(x)_6 = 0.$$

Because $\|u(x)\|_2 = 1.036\cdots > 1$, we conclude from Theorem 5 that $x \notin \operatorname{conv}(N^3_{\|\cdot\|_2})$. We now derive the separating hyperplane. We first separate $u(x)$ from $B_2(1)$. Because $\|\cdot\|_2$ is self-dual, $\hat{\beta} = u(x)/\|u(x)\|_2$ is an optimal solution to (16). Then, the inequality $\hat{\beta}^\top u \le 1$ (or, equivalently, $u(x)^\top u \le \|u(x)\|_2$) is valid for $B_2(1)$. Furthermore, it separates $u(x)$ because $u(x)^\top u(x) = \|u(x)\|_2^2 > \|u(x)\|_2$. We next construct a hyperplane that separates x from $N^3_{\|\cdot\|_2}$. Define θ and χ as follows:

$$\theta_1 = \beta_1 = \frac{1}{\|u(x)\|_2} u(x)_1 = \frac{1}{\|u(x)\|_2}\frac{27}{28}, \quad \theta_2 = \theta_3 = \frac{1}{3-2+1}(\beta_2 + \beta_3) = \frac{1}{\|u(x)\|_2}\frac{15}{56}, \quad \theta_4 = \cdots = \theta_6 = 0,$$
$$\chi_1 = \beta_1 = \frac{1}{\|u(x)\|_2} u(x)_1 = \frac{1}{\|u(x)\|_2}\frac{27}{28}, \quad \chi_2 = \cdots = \chi_6 = \frac{1}{3-2+1}(\beta_2 + \beta_3) = \frac{1}{\|u(x)\|_2}\frac{15}{56}.$$

Observe that $\|\beta\|_2 = \|\theta\|_2 = 1$, but $\|\chi\|_2 > 1$. Now consider the inequality $\chi^\top y \le 1$. It is valid for $N^3_{\|\cdot\|_2}$ because for any $u \in N^3_{\|\cdot\|_2}$, $\chi^\top u \le \chi^\top u_{\Delta_+} = \theta^\top u_{\Delta_+} \le \|\theta\|_2 \|u_{\Delta_+}\|_2 \le 1$. Moreover, it separates x because $\chi^\top x = \frac{1}{\|u(x)\|_2}\left(\frac{27}{28}, \frac{15}{56}, \frac{15}{56}, \frac{15}{56}, \frac{15}{56}, \frac{15}{56}\right)^\top \left(\frac{27}{28}, \frac{5}{28}, \frac{4}{28}, \frac{3}{28}, \frac{2}{28}, \frac{1}{28}\right) = 1.036\cdots > 1$. □

Next, we consider some special cases of the set $N^K_{\|\cdot\|_s}$ defined in (10) and provide explicit convex hull descriptions.

Proposition 5. *Let* $S = \{x \in \mathbb{R}^q \mid \operatorname{card}(x) \le K, \|x\|_\infty \le r\}$, *where* $\|x\|_\infty = |x|_{[1]}$. *Then,*

$$\operatorname{conv}(S) = \{x \in \mathbb{R}^q \mid \|x\|_1 \le rK, \|x\|_\infty \le r\}. \tag{20}$$

Proof. Observe that $S/r = \{x \in \mathbb{R}^q \mid \operatorname{card}(x) \le K, \|x\|_\infty \le 1\}$. Then,

$$\operatorname{conv}(S) = r \operatorname{conv}(S/r) = \{y \in \mathbb{R}^q \mid \|u(y)\|_\infty \le r\} = \{y \in \mathbb{R}^q \mid max\{|y|_{[1]}, s(y)_1\} \le r\},$$

where the second equality follows from Proposition 2 and Theorem 5, and the third equality from part 4 of Proposition 4. Because $s(y)_1 = \frac{1}{K}\sum_{j=1}^{q} |y|_{[j]}$ by definition of $s(y)$, the result follows. □

When $\|\cdot\|_s$ is the ℓ_2-norm, the norm $\|\cdot\|_c$ associated with $\operatorname{conv}(N^K_{\|\cdot\|_2})$ is known to be the *K-support norm* (or *K-overlap norm*). An explicit formula for this norm is introduced in Argyriou et al. [2]. We next provide an alternate derivation of this formula using our arguments. For consistency with literature, we denote the *K*-support norm by $\|\cdot\|_K^{sp}$.

Lemma 5. *The unique integer* $r \in \{0, \ldots, K-1\}$ *that satisfies*

$$|x|_{[K-r-1]} > s(x)_{K-r} \ge |x|_{[K-r]} \tag{21}$$

is $r = K - i_x$, *where* $|x|_{[0]} = \infty$ *by convention.*

Proof. This result follows from Proposition 4, and we refer to its parts directly in the proof. Let $|x|_{[i]} \le s(x)_i < |x|_{[i-1]}$ for some $i \in \{1, \ldots, K\}$. We show that $i = i_x$. First, we show that $|x|_{[j]} \le s(x)_j$ for all $j \ge i$. Let $j+1$ be the first

index no less than i so that $|x|_{[j+1]} > s(x)_{j+1}$, and observe that, in fact, $j+1 > i$. Then, we obtain the contradiction $|x|_{[j+1]} \le |x|_{[j]} \le s(x)_j \le s(x)_{j+1} < |x|_{[j+1]}$, where the third inequality is by part 1. Therefore, $|x|_{[j]} \le s(x)_j$ for $j \ge i$, which implies by part 1 that $s(x)_i \le \cdots \le s(x)_K$ and by part 2 that $i \ge i_x$. Because $|x|_{[i-1]} > s(x)_i$, it follows by part 1 that either $i = 1$ or $s(x)_{i-1} > s(x)_i$. In either case, it follows that $i \le i_x$. Therefore, $i = i_x = K - r$. □

Proposition 6 (Argyriou et al. [2, proposition 2.1]). *Suppose r the unique integer in $\{0, \dots, K-1\}$ satisfying (21). Then,*

$$\|x\|_K^{sp} = \left(\sum_{i=1}^{K-r-1} x_{[i]}^2 + \frac{1}{r+1} \left(\sum_{i=K-r}^{n} |x|_{[i]} \right)^2 \right)^{\frac{1}{2}}. \tag{22}$$

Proof. By Theorem 5,

$$\|x\|_K^{sp} = \|u(x)\|_2 = \left(\sum_{i=1}^{i_x-1} |x|_{[i]}^2 + (K - i_x + 1)\delta(x)^2 \right)^{\frac{1}{2}} = \left(\sum_{i=1}^{i_x-1} |x|_{[i]}^2 + \frac{1}{K - i_x + 1} \left(\sum_{i=i_x}^{n} |x|_{[i]} \right)^2 \right)^{\frac{1}{2}}.$$

The result then follows because Lemma 5 establishes that $r = K - i_x$. □

3.1. Convexification of Sets of Matrices Characterized by Their Singular Values

Let $\mathcal{M}^{m,n}(\mathbb{R})$ be the set of $m \times n$ real-valued matrices. For $M \in \mathcal{M}^{m,n}(\mathbb{R})$, let $\sigma_1(M) \ge \cdots \ge \sigma_q(M)$ denote the singular values of M, where $q = \min\{m,n\}$, and let $\sigma : \mathcal{M}^{m,n}(\mathbb{R}) \to \mathbb{R}^q$ be defined as $\sigma(M) = (\sigma_1(M), \dots, \sigma_q(M))$. Let $\|M\|_{sp} = \sigma_1(M)$ and $\|M\|_* = \sum_{i=1}^q \sigma_i(M)$ be the *spectral norm* and the *nuclear norm* of M, respectively. In this subsection, we consider sets of matrices that are characterized by their singular values. More specifically, we are interested in sets of the form $\bar{S} = \{M \in \mathcal{M}^{m,n}(\mathbb{R}) \mid f_i(\sigma(M)) \le 1, i = 1, \dots, r\}$ and their convex hulls, where each f_i is a sign- and permutation-invariant function. Define $S = \{x \in \mathbb{R}^q \mid f_i(x) \le 1, i = 1, \dots, r\}$, where $q = \min\{m,n\}$. It is clear that $M \in \bar{S}$ if and only if $\sigma(M) \in S$. As we show next, Theorem 4 implies that convex hulls of sets of the form $\bar{S}$ can be obtained by studying S instead. Similar results, although dealing with closed convex hulls, can be derived using conjugacy results of Lewis [18]. We include a direct proof based on Theorem 4.

Theorem 7. *For $p \in \mathbb{Z}_{++}$ and $q \in \mathbb{Z}_+$ and each $i \in \{1, \dots, r\}$, let $f_i : (x,z) \mapsto \mathbb{R}$, where $(x,z) \in \mathbb{R}^p \times \mathbb{R}^q$, be sign- and permutation-invariant functions with respect to $x \in \mathbb{R}^p$. Let $m, n \in \mathbb{Z}$, such that $\min\{m,n\} = p$, and define $\bar{S} = \{(M,z) \in \mathcal{M}^{m,n}(\mathbb{R}) \times \mathbb{R}^q \mid f_i(\sigma(M), z) \le 1, i = 1, \dots, r\}$. Furthermore, define $S = \{(x,z) \in \mathbb{R}^p \times \mathbb{R}^q \mid f_i(x,z) \le 1, i = 1, \dots, r\}$. Then,*

$$\operatorname{conv}(\bar{S}) = X := \{(M,z) \in \mathcal{M}^{m,n}(\mathbb{R}) \times \mathbb{R}^q \mid (\sigma(M), z) \in \operatorname{conv}(S)\}.$$

Proof. We first show that $\operatorname{conv}(\bar{S}) \subseteq X$. Consider $(M,z) \in \operatorname{conv}(\bar{S})$, so that $(M,z) = \sum_j \gamma_j (M^j, z^j)$, where $(M^j, z^j) \in \bar{S}$ and γ_j are convex multipliers. For $k \in \{1, \dots, p\}$ and any $Y \in \mathcal{M}^{m,n}(\mathbb{R})$, define $s^k(Y) := \sum_{i=1}^k \sigma(Y)_{[i]}$ to be the kth Ky Fan norm. Then, it follows by sublinearity and positive homogeneity of norms that $s^k(M) \le \sum_j \gamma_j s^k(M^j)$. In other words, $\sum_j \gamma_j \sigma(M^j) \ge_{wm} \sigma(M)$. Let $\sigma^j = \sigma(M^j)$. Because $(M^j, z^j) \in \bar{S}$, it follows that $f_i(\sigma^j, z^j) \le 1$. Therefore, $(\sum_j \gamma_j \sigma^j, z) \in \operatorname{conv}(S_0)$, where $S_0 = S \cap \{(\sigma, z) \mid \sigma_1 \ge \cdots \ge \sigma_p \ge 0\}$. Then, it follows by part 3 of Theorem 4 that $(\sigma(M), z) \in \operatorname{conv}(S)$ and $(M,z) \in X$.

We now show that $\operatorname{conv}(\bar{S}) \supseteq X$. Let $(M,z) \in X$, and let $U \operatorname{diag}(\sigma) V^\top$ be the singular value decomposition of M, where $\operatorname{diag}(\sigma) \in \mathcal{M}^{m,n}(\mathbb{R})$ is the diagonal matrix, whose diagonal is the vector σ, and $\sigma_1 \ge \cdots \ge \sigma_p \ge 0$. Because $(\sigma, z) \in \operatorname{conv}(S)$, it follows by part 3 of Theorem 4 that there exist $\sigma' \in \mathbb{R}^p$, $(\sigma^j, z^j) \in S_0$ and convex multipliers γ_j so that $(\sigma', z) = \sum_j \gamma_j (\sigma^j, z^j)$ and $\sigma' \ge_{wm} \sigma$. Now, if θ^j is obtained by permuting σ^j or changing the sign of some of its entries, it follows readily that $(U \operatorname{diag}(\theta^j) V^T, z^j) \in \bar{S}$, because these operations do not alter the singular values of the matrix. Because $\sigma' \ge_{wm} \sigma$, $\sigma = \sum_k \chi_k T_k \sigma'$, where χ_k are convex multipliers, and each T_k permutes the entries of σ' and possibly changes the sign of a few of the entries. Then, it follows that $(\sigma, z) = \sum_k \chi_k (T_k \sigma', z) = \sum_k \sum_j \chi_k \gamma_j (T_k \sigma^j, z^j)$. Because $U \operatorname{diag}(\theta) V^T$ is a linear operator of θ, $\sum_j \sum_k \chi_k \gamma_j = 1$, and we have already shown that $(U \operatorname{diag}(T_k \sigma^j) V^T, z^j) \in \bar{S}$, it follows that $(M,z) \in \operatorname{conv}(\bar{S})$. □

In the following, we denote the set of $p \times p$ real symmetric matrices by $\mathcal{S}^p$, and for any $M \in \mathcal{S}^p$, we denote the eigenvalues by $\lambda(M)$. In this context, a similar result can be shown using eigenvalues instead of singular values.

Theorem 8. *For $p \in \mathbb{Z}_{++}$ and $q \in \mathbb{Z}_+$ and each $i \in \{1,\dots,r\}$, let $f_i : (x,z) \mapsto \mathbb{R}$, where $(x,z) \in \mathbb{R}^p \times \mathbb{R}^q$ be permutation-invariant functions with respect to $x \in \mathbb{R}^p$. Define $\bar{S} = \{(M,z) \in \mathcal{S}^p \times \mathbb{R}^q \mid f_i(\lambda(M),z) \le 1, i = 1,\dots,r\}$. Furthermore, define $S = \{(x,z) \in \mathbb{R}^p \times \mathbb{R}^q \mid f_i(x,z) \le 1, i = 1,\dots,r\}$. Then,*

$$\operatorname{conv}(\bar{S}) = X := \{(M,z) \in \mathcal{S}^p \times \mathbb{R}^q \mid (\lambda(M),z) \in \operatorname{conv}(S)\}.$$

Proof. We provide only a proof sketch because the proof is similar to that of Theorem 7. To show that $\operatorname{conv}(\bar{S}) \subseteq X$, we consider $(M,z) = \sum_j \gamma_j (M^j, z^j) \in \operatorname{conv}(\bar{S})$ and use the fact that $\sum_j \gamma_j \lambda(M^j) \succeq_m \lambda(M)$ (Horn and Johnson [14, theorem 4.3.27]). Then, the result follows from part 1 of Theorem 4. To show that $\operatorname{conv}(\bar{S}) \supseteq X$, we consider $(M,z) \in X$ and express $\sum_j \gamma_j (\lambda^j, z^j) \succeq_m (\lambda(M),z)$ for some $(\lambda^j, z^j) \in S$. Then, we observe that this implies that for any orthogonal matrix U and a permutation π, $(U \operatorname{diag}(\pi(\lambda^j)) U^\top, z) \in \bar{S}$. The result is then derived in a manner similar to that in the proof of Theorem 4 except that instead of using the singular value decomposition $U \operatorname{diag}(\sigma) V^\top$ of M, we use the eigenvalue decomposition $M = U \lambda(M) U^\top$. □

The rank of a matrix can be represented as the cardinality of the vector of singular values. Because cardinality is a sign- and permutation-invariant function, we obtain the following result as a special case of Theorem 7.

Corollary 2. *Let $\bar{S} = \{M \in \mathcal{M}^{m,n}(\mathbb{R}) \mid \operatorname{rank}(M) \le K, \|\sigma(M)\|_s \le r\}$. Then,*

$$\operatorname{conv}(\bar{S}) = \{M \in \mathcal{M}^{m,n}(\mathbb{R}) \mid \|\sigma(M)\|_c \le r\}.$$

Consider $\bar{S}$ in Corollary 2. Recall that determining whether an arbitrary matrix $M \in \mathcal{M}^{m,n}(\mathbb{R})$ is in the convex hull $\operatorname{conv}(\bar{S})$ can be easily done when the norm $\|\cdot\|_s$ is calculable. In particular, when $\|\cdot\|_s$ is the Euclidean norm, a given matrix M is in $\operatorname{conv}(\bar{S})$ if $\|\sigma(M)\|_K^{sp} \le r$; see (22) for an explicit formula for $\|\cdot\|_K^{sp}$. Semidefinite representability of the convex hull will be discussed in Section 3.2.

Next, we consider the special case where $\|\cdot\|_s$ is the l_∞ norm. Proposition 5 and Theorem 7 together give an alternative proof for the following result.

Proposition 7 (Hiriart-Urruty and Le [13, theorem 1]). *Let $\bar{S} = \{M \in \mathcal{M}^{m,n}(\mathbb{R}) \mid \operatorname{rank}(M) \le K, \|M\|_{sp} \le r\}$. Then, $\operatorname{conv}(\bar{S}) = \{M \in \mathcal{M}^{m,n}(\mathbb{R}) \mid \|M\|_* \le rK, \|M\|_{sp} \le r\}$.*

3.2. Semidefinite Representability of Sets of Matrices Characterized by Their Singular Values

We presented in Corollary 2 a convex hull result for a set of matrices $\bar{S}$ that is described using their singular values. The resulting convex hull is written in a norm $\|\cdot\|_c$ induced by the defining norm $\|\cdot\|_s$ of $\bar{S}$. In this subsection, we discuss the representability of this set as the feasible set of a semidefinite programming problem. A set is called *semidefinite representable* if it is a projection of a set expressed by a linear matrix inequality. We remark the following well-known results about semidefinite representability; see section 4.2 of Ben-Tal and Nemirovski [5].

Lemma 6. *The following sets are semidefinite representable:*

1. *the epigraph of the sum of p largest singular values of a rectangular matrix,*
2. *the epigraph of the sum of p largest eigenvalues of a symmetric matrix,*
3. *the graph of the sum of all eigenvalues of a symmetric matrix,*
4. *the set $A \cap B$, where A and B are semidefinite representable.*

In particular, we consider the set $\mathcal{S} = \{M \in \mathcal{M}^{m,n}(\mathbb{R}) \mid \operatorname{rank}(M) \le K, f(\sigma(M)) \le r\}$, where $q = \min\{m,n\}$, and $f : \Delta_+ \to \mathbb{R}$ is a quasiconvex function. A function f is said to be *quasiconvex* on Δ_+ if for $\lambda \in [0,1]$ and $x, y \in \Delta_+$, we have $f(\lambda x + (1-\lambda)y) \le \max\{f(x), f(y)\}$. We assume this function has semidefinite-representable lower-level sets. Then, we show that the convex hull of $\mathcal{S}$, the set of rank-constrained matrices whose singular values belong to a lower-level set of f, are semidefinite representable.

Theorem 9. *Let* $q = \min\{m,n\}$ *and* $\bar{S} = \{M \in \mathcal{M}^{m,n}(\mathbb{R}) \,|\, \text{rank}(M) \le K, f(\sigma(M)) \le r\}$, *where* $f : \Delta_+ \to \mathbb{R}$ *has semidefinite-representable lower-level sets. Then,* $\text{conv}(\bar{S})$ *is semidefinite representable.*

Proof. Define $g : \mathbb{R}^q \longmapsto \mathbb{R}$ so that $g(x) = f(x_{\Delta_+})$. Observe further that g is sign and permutation invariant. Then, let $S = \{x \in \mathbb{R}^q \,|\, \text{card}(x) \le K, g(x) \le r\}$. By sign and permutation invariance of S and Theorem 4,

$$\text{conv}(S) = \left\{ x \in \mathbb{R}^q \;\middle|\; \begin{array}{l} f(u) \le r, \\ u_1 \ge \cdots \ge u_K \ge 0, \\ u_{K+1} = \cdots = u_n = 0, \\ u \ge_{wm} |x| \end{array} \right\}.$$

Therefore, by Theorem 7,

$$\text{conv}(\bar{S}) = \left\{ M \in \mathcal{M}^{m,n}(\mathbb{R}) \;\middle|\; \begin{array}{l} f(u) \le r, \\ u_1 \ge \cdots \ge u_K \ge 0, \\ u_{K+1} = \cdots = u_n = 0, \\ u \ge_{wm} \sigma(M) \end{array} \right\}.$$

By the definition of weak majorization, the convex hull has the following representation:

$$\begin{cases} f(u) \le r, \\ u_1 \ge \cdots \ge u_K \ge 0, \\ u_{K+1} = \cdots = u_n = 0, \\ \sum_{i=1}^{j} u_i \ge \sum_{i=1}^{j} \sigma_j(M), \quad j = 1, \ldots, K. \end{cases} \tag{23}$$

The semidefinite representability of (23) follows from Lemma 6 and the semidefinite representability of the level set $\{u \,|\, f(u) \le r\}$ and the introduced linear inequalities. □

Although Theorem 9 is similar to proposition 4.2.2 in Ben-Tal and Nemirovski [5], we discuss next how these results differ. First, we introduce a rank constraint and thus treat a nonconvex set. Second, we discuss the representation of the convex hull rather than the set itself. Third, we do not require monotonicity of $f(x)$ and require semidefinite representability only over $\Delta_+ (= \Delta_+^q)$ instead of $\mathbb{R}_+^q$. We briefly describe why the added assumptions are required in proposition 4.2.2 in Ben-Tal and Nemirovski [5], but not in our result. This is because, when $f(x)$ is not monotone but is quasiconvex over Δ_+, its extension to $\mathbb{R}_+^q$ defined using $g(x) := f(x_{\Delta_+})$ is not necessarily quasiconvex. As such, the lower-level sets of $g(x)$ are not always semidefinite representable. To see this, consider $f(x) = 1 - x$, where $x \in \mathbb{R}$. Then, $\{x \,|\, g(x) \le 0\}$ is not a convex set because $g(1) = g(-1) = 0$, whereas $g(0) = 1$. On the contrary, consider an $f(x)$ that is monotone, permutation invariant, and quasiconvex over $\mathbb{R}_+^q$. Let $C = \{x \,|\, g(x) \le r\}$. We argue that C is convex and can be expressed as $X = \{x \,|\, f(u) \le r, u \in \Delta_+, u \ge_{wm} |x|\}$. First, observe that Theorem 4 shows that $X = \text{conv}(C) \supseteq C$. Now, we argue that $X \subseteq C$. To see this, assume $x \in X$. Then, $x_{\Delta_+} \in X$ because X, being $\text{conv}(C)$, inherits the sign and permutation invariance of C. Then, there exist u and u' such that $u \ge_m u' \ge x_{\Delta_+}$ and $f(u) \le r$. Therefore, we have $g(x) = f(x_{\Delta_+}) \le f(u') \le f(u) \le r$, where the first inequality is from monotonicity of f. The second inequality is because u and u' are nonnegative, u' is in the convex hull of u and its permutation variants, and f is quasiconvex and permutation invariant. The third inequality follows by definition of u. Therefore, it follows that $x \in C$.

Corollary 3. *Let* $\mathcal{S} = \{M \in \mathcal{M}^{m,n}(\mathbb{R}) \,|\, \text{rank}(M) \le K, \|\sigma(M)\|_s \le r\}$, *where* $\|\cdot\|_s$ *is a permutation-invariant monotone norm. Then,* $\text{conv}(\mathcal{S})$ *is semidefinite representable. In particular, when* $\|\sigma(\cdot)\|_s$ *is a Ky Fan p-norm (the sum of p largest singular values) for some* $p = 1, \ldots, \min\{m,n\}$, *the convex hull is semidefinite representable.*

Similarly, we can prove the following result, where $\mathcal{S}_+^q \in \mathcal{M}^{q \times q}(\mathbb{R})$ is the set of positive semidefinite symmetric matrices.

Theorem 10. *Let* $\bar{S} = \{M \in \mathcal{S}_+^q \,|\, \text{rank}(M) \le K, f(\lambda(M)) \le r\}$, *where* $f : \Delta_+ \to \mathbb{R}$ *has semidefinite-representable lower-level sets. Then,* $\text{conv}(\bar{S})$ *is semidefinite representable.*

Proof. Define $g : \mathbb{R}^q_+ \mapsto \mathbb{R}$ so that $g(x) = f(x_\Delta)$. Then, let $S = \{x \in \mathbb{R}^q_+ \,|\, \text{card}(x) \le K, g(x) \le r\}$. By permutation invariance of S and Theorem 4,

$$\text{conv}(S) = \left\{ x \in \mathbb{R}^q_+ \;\middle|\; \begin{array}{l} f(u) \le r, \\ u_1 \ge \cdots \ge u_K \ge 0, \\ u_{K+1} = \cdots = u_n = 0, \\ u \succeq_m x \end{array} \right\}.$$

Therefore, by Theorem 8,

$$\text{conv}(\bar{S}) = \left\{ M \in \mathcal{S}^q \;\middle|\; \begin{array}{l} f(u) \le r, \\ u_1 \ge \cdots \ge u_K \ge 0, \\ u_{K+1} = \cdots = u_n = 0, \\ u \succeq_m \lambda(M) \end{array} \right\}.$$

By the definition of majorization, the convex hull has the following representation:

$$\begin{cases} f(u) \le r, \\ u_1 \ge \cdots \ge u_K \ge 0, \\ u_{K+1} = \cdots = u_n = 0, \\ \sum_{i=1}^{j} u_i \ge \sum_{i=1}^{j} \lambda_j(M), \quad j = 1, \ldots, K-1, \\ \sum_{i=1}^{K} u_j = \sum_{i=1}^{q} \lambda_i(M). \end{cases} \tag{24}$$

The semidefinite representability of (24) follows from Lemma 6 and the semidefinite representability of the level set $\{u | f(u) \le r\}$ and the introduced linear inequalities. □

The ideas in the above proof can be extended, using disjunctive programming techniques, to symmetric matrices that are not necessarily positive semidefinite. Because the eigenvalues are no longer nonnegative, we cannot impose the restriction that $u \ge 0$ and, thus, assume that $u_{K+1}, \ldots, u_n = 0$. Instead, we express the rank constraint as a union of sets, each of which satisfies $u_{a+1} = \cdots = u_{a+n-K} = 0$ for some $a \in \{0, \ldots, K\}$. Then, we obtain $\text{conv}(S)$ as the convex hull of a union of semidefinite-representable sets. Using the disjunctive programming argument of Ben-Tal and Nemirovski [5, proposition 3.3.5], this yields a lifted representation of a set that outer-approximates $\text{conv}(S)$ and is contained in $\text{clconv}(S)$.

4. Convex Envelopes of Nonlinear Functions

The problem of finding convex envelopes of nonlinear functions is central to the global solution of factorable problems through branch and bound. When the domain over which the envelope is constructed is a polytope P, it is often the case that the envelope is completely determined by the values that the function takes on a subset of the faces of P, or more generally, a subset of its feasible points. If the above property holds, disjunctive programming techniques can often be employed to provide an explicit (although typically large) description of the envelope, through the introduction of new variables for each of the important subsets of P. In this section, using Theorem 4 as a foundation, we show that for certain functions defined over permutation-invariant polytopes, (i) envelopes can be built without recourse to disjunctive programming (Proposition 8), and (ii) polynomially sized disjunctive programming formulations can be constructed even when the number of faces of P important in the construction of the envelope is exponential (Theorem 11). These results yield compact envelopes descriptions for specific families of functions (Propositions 9 and 10). We also provide numerical evidence that the use of these techniques produces relaxations of multilinear functions over permutation-invariant hypercubes that are significantly stronger than those obtained using a recursive application of McCormick's procedure. The techniques can be extended to handle epigraphs of singular values/eigenvalues of matrices using the ideas presented in Sections 3.1 and 3.2.

Definition 2. A function $\phi : C \mapsto \mathbb{R}$ is said to be *Schur concave* on C if for every $x, y \in C$, $x \succeq_m y$ implies that $\phi(x) \le \phi(y)$.

Various functions have been shown to be Schur concave, including the Shannon entropy $\sum_{i=1}^{n} x_i \log(\frac{1}{x_i})$ and elementary symmetric functions $\sum_{J \subseteq \{1,\ldots,n\}: |J| = k} \prod_{i \in J} x_i$. Symmetric concave functions are also Schur concave. More

complex Schur-concave functions can be constructed from known Schur-concave functions using some compositions or operations that preserve Schur concavity; see Marshall et al. [24, chapter 3] for further detail.

In this section, for any function $\phi : C \mapsto \mathbb{R}$, we denote the convex envelope of ϕ over C by $\text{conv}_C(\phi)$. A common tool in the construction of convex envelopes is to restrict the domain of the function to a smaller subset. We say that a function $\phi : C \mapsto \mathbb{R}$ can be *restricted to* X, where $X \subseteq C$, for the purpose of obtaining $\text{conv}_C(\phi)$ if $\text{conv}_C(\phi|_X) = \text{conv}_C(\phi)$, where $\phi|_X(x)$ is defined as $\phi(x)$ for any $x \in X$ and $+\infty$ otherwise.

First, we establish in Lemma 7 that when deriving the envelope of a Schur-concave function over a permutation-invariant polytope, it is sufficient to restrict attention to those points in the domain that are not majorized by other feasible points. When coupled with a simple domain structure, this result permits a description of convexification results without the use of majorization inequalities, in a smaller dimensional space.

Lemma 7. *Let $\phi : P \mapsto \mathbb{R}$ be a Schur-concave function, where $P \subseteq \mathbb{R}^n$ is a permutation-invariant polytope. Let $M := \{x \in P \mid \nexists u \in P \text{ with } u \geq_m x \text{ and } u_\Delta \neq x_\Delta\}$. Let $S := \{(x,\phi) \mid \phi(x) \leq \phi \leq \alpha, x \in P\}$ and $X := \{(x,\phi) \mid \phi(x) \leq \phi \leq \alpha, x \in M\}$. Then,* $\text{conv}(S) = \text{conv}(X)$.

Proof. Because $M \subseteq P$, it follows that $X \subseteq S$ and, therefore, $\text{conv}(X) \subseteq \text{conv}(S)$. Now, consider $(x',\phi') \in S \backslash X$. Therefore, $\phi(x') \leq \phi' \leq \alpha$ and $x' \in P \backslash M$. Let $x''_i = \frac{1}{n}\sum_{i'=1}^n x'_{i'}$ for all $i \in \{1,\dots,n\}$. That is, all components of x'' are identical. Let $u' := \arg\max\{\|u - x''\| \mid u \geq_m x', u \in P\}$. The maximum is achieved because the feasible set is compact and because the objective is upper semicontinuous. Assume by contradiction that there exists $y' \in P$ such that $y' \geq_m u'$ and $y'_\Delta \neq u'_\Delta$. Because u' can be written as a convex combination of at least two permutations of y' and the objective of the problem defining u' is permutation invariant and strictly convex, it follows that $\|y' - x''\| > \|u' - x''\|$, violating the optimality of u'. Therefore, there does not exist $y' \in P$ such that $y' \geq_m u'$ and $y'_\Delta \neq u'_\Delta$. In other words, $u' \in M$. It follows that, for any permutation matrix $Q \in \mathcal{P}_n$, $Qu' \in M$. Because ϕ is Schur concave, then $\phi(Qu') = \phi(u') \leq \phi(x') \leq \phi' \leq \alpha$. Therefore, $(Qu',\phi') \in X$. Finally, because $x' \leq_m u'$, x' can be written as a convex combination of permutations of u'. Therefore, $(x',\phi') \in \text{conv}(X)$. We conclude that $S \subseteq \text{conv}(X)$. □

Lemma 7 requires the identification of the set M, which consists of points in P that are not expressible as a convex combination of another point in P and its permutations. In Lemma 8, we characterize such points as those that have a supporting hyperplane of a specific form. Later, we discuss how these results can be combined to obtain, in closed form, convex envelopes of various Schur-concave functions.

Lemma 8. *Let $x' \in P$, where P is a permutation-invariant polytope. Let π be a permutation of $\{1,\dots,n\}$ such that for each $i \in \{1,\dots,n-1\}$, $x'_{\pi(i)} \geq x'_{\pi(i+1)}$. Then, there exists $u' \in P$ with $u' \geq_m x'$ and $u'_\Delta \neq x'_\Delta$ if and only if there does not exist $a \in \mathbb{R}^n$ such that $a_{\pi(i)} > a_{\pi(i+1)}$ for all $i \in \{1,\dots,n-1\}$ and $\sum_{i=1}^n a_i(x_i - x'_i) \leq 0$ is valid for P.*

Proof. We first show that if such a exists, there does not exist $u' \in P$ such that $u' \geq_m x'$ and $u'_\Delta \neq x'_\Delta$. Assume by contradiction that such a u' exists. Because P is permutation invariant, we may assume that $u'_{\pi(i)} \geq u'_{\pi(i+1)}$ for all $i \in \{1,\dots,n-1\}$ by sorting u' if necessary. Because $u' \geq_m x'$ and $u' \neq x'$, there exists y', $\theta > 0$, $k \in \{1,\dots,n-1\}$, and $r \in \{1,\dots,n-k\}$ such that $u' \geq_m y' \geq_m x'$, $y'_{\pi(k)} = x'_{\pi(k)} + \theta$, $y'_{\pi(k+r)} = x'_{\pi(k+r)} - \theta$ and $y'_i = x'_i$ otherwise; see Marshall et al. [24, lemma 2.B.1]. Because $u' \in P$, P is convex and permutation invariant, and because y' can be written as a convex combination of u' and its permutations, it is clear that $y' \in P$. Therefore, $\sum_{i=1}^n a_i(y'_i - x'_i) \leq 0$ or $a_{\pi(k)} - a_{\pi(k+r)} \leq 0$. This is a contradiction to the assumed ordering of a.

Now, we show that if there does not exist $u' \in P$ such that $u' \geq_m x'$ and $u'_\Delta \neq x'_\Delta$, we can construct such a vector a. Define a polyhedral cone $K := \{\sum_{i=1}^{n-1} \alpha_{\pi(i)}(e_{\pi(i)} - e_{\pi(i+1)}) \mid \alpha \geq 0\}$, where e_j represents the jth standard basis. Observe that $v \in K$ (equivalently, $v \geq_K 0$) implies that $v \geq_m 0$. More specifically, given $v \in \mathbb{R}^n$, let $S^\pi_k[v]$ be the partial sums so that $S^\pi_k[v] = \sum_{i=1}^k v_{\pi(i)}$. Then, $v \in K$ if and only if $S^\pi_k[v]$ is nonnegative for every k and $S^\pi_n[v] = 0$. To verify that the latter set is contained in K, write any v in that set as $v = \sum_{k=1}^{n-1} S^\pi_k[v](e_{\pi(k)} - e_{\pi(k+1)})$ to see that $v \in K$. To show the reverse inclusion, verify that $S^\pi_k[e_{\pi(i)} - e_{\pi(i+1)}] \geq 0$, where the last inequality is satisfied at equality. Because $\sum_{i=1}^k v_{[i]} \geq S^\pi_k[v]$ for all k and $\sum_{i=1}^n v_{[i]} = S^\pi_n[v]$, $v \in K$ implies that $v \geq_m 0$. We next construct the polyhedron $C := P - K - \{x'\}$, where the difference is the Minkowski difference. Because $x' \in P \cap (K + \{x'\})$, it follows that $0 \in C$. Let $\langle a', x\rangle \leq 0$ define the minimal (possibly trivial) face F of C containing zero. We show next that a can be chosen to be a'. First, note that for $x \in P$, $x - x' \in C$. Therefore, as claimed, $\langle a', x - x'\rangle \leq 0$. Because $e_{\pi(i+1)} - e_{\pi(i)} \in C$ for all $i \in \{1,\dots,n-1\}$, we have that $a'_{\pi(i+1)} - a'_{\pi(i)} \leq 0$. We now show that the inequalities are strict; otherwise, there exists a point in $w \in C$ so that $w \geq_K 0$ such that there is a k for which $S^\pi_k[w] > 0$. This suffices because if there exists $w = u - v \in C \cap K$, where $u \in P$ and $v - x' \in K$, then we have $u \geq_K v \geq_K x'$, which in turn proves the existence of $u \in P$ that majorizes x'. Furthermore, by showing $S^\pi_k[w] > 0$ for some k, we have $u_\Delta \neq x'_\Delta$, which contradicts our

assumption that such a u does not exist. Assume that the inequality is not strict for some $k \in \{1,\dots,n-1\}$ so that $\langle a', e_{\pi(k+1)} - e_{\pi(k)}\rangle = 0$. Then, for $\epsilon > 0$, we define $w_\epsilon = \epsilon(e_{\pi(k)} - e_{\pi(k+1)}) \in K$ and show that, for a sufficiently small ϵ, $w_\epsilon \in C$. Because $-w_\epsilon \in F$ and zero is in the relative interior of F, it follows that there exists an ϵ such that $w_\epsilon \in C$. Moreover, $S^\pi_k[w_\epsilon] = \epsilon > 0$. As we argued above, the existence of such a w_ϵ contradicts our assumption and, so, we conclude that $a'_{\pi(i+1)} - a'_{\pi(i)} < 0$ for all $i \in \{1,\dots,n-1\}$. □

Lemmas 7 and 8 can be combined to develop convex envelopes of Schur-concave functions. This is because, taken together, they prove that it suffices to restrict attention to a subset M of P in order to construct the convex envelope. To better understand the structure of M and to illustrate potential applications, we derive in Proposition 8 the closed-form description of the convex envelope of a Schur-concave function over $[a,b]^n$. In this case, M is contained in the one-dimensional faces of $[a,b]^n$, the key insight that allows for the derivation of the closed form. Although we provide a self-contained argument for this fact in the proof of Proposition 8, this inclusion can be seen as a special case of Lemma 8, a visualization that serves to illustrate the use of Lemma 8 in characterizing M. To see this special case, observe that $x' \in M \cap \Delta^n$ only if there is an inequality $\langle \beta, x - x'\rangle \le 0$ valid for $[a,b]^n$ such that $\beta_1 > \cdots > \beta_n$. Because such an inequality is tight at x', it can be derived as a conic combination of facet-defining inequalities tight at x'. The facet-defining inequalities of $[a,b]^n$ are $-x_i \le -a$ and $x_i \le b$ for $i = 1,\dots,n$. Let $F(x')$ be the set of facet-defining inequalities tight at x', and for any facet-defining inequality in this set, say, $\langle \alpha, x - x'\rangle \le 0$, let $L_\alpha = \{(t,t+1) \mid \alpha_t > \alpha_{t+1}\}$. It is easy to see that, for $x' \in [a,b]^n$, $|L_\alpha| \le 1$ for each $\alpha \in F(x')$; that is, each L_α contains at most one pair. Then, for β to be derived as a conic combination of inequalities in $F(x')$, it must be that $\{(i,i+1)\}_{i=1}^{n-1} \subseteq \cup_{\alpha \in F(x')} L_\alpha$. Because $|L_\alpha| \le 1$, it follows that $|F(x')| \ge n-1$. Therefore, M is a subset of one-dimensional faces of the hypercube. More generally, a similar argument shows that if there exists a $k \in \{1,\dots,n-1\}$ such that $|L_\alpha| \le k$, then $|F(x')| \ge \lceil \frac{n-1}{k} \rceil$, and, consequently, M is a subset of $n - \lceil \frac{n-1}{k} \rceil$ faces of P.

Proposition 8. *Consider a function $\phi(x) : \mathbb{R}^n \mapsto \mathbb{R}$ that is Schur concave over $[a,b]^n$ and let $S^\alpha := \{(x,\phi) \mid \phi(x) \le \phi \le \alpha, x \in [a,b]^n\}$. For any $x \in [a,b]^n$, define $S(x) = \sum_{i=1}^n (x_i - a)$. For any $s \in [0, n(b-a)]$, let $i^s = \max\{i \mid i(b-a) < s\}$ and*

$$u^s_i = \begin{cases} b & \text{if } i \le i^s, \\ a + s - (b-a)i^s & \text{if } i = i^s + 1, \\ a & \text{otherwise.} \end{cases} \tag{25}$$

Let $\Theta^\alpha := \{(x,\phi) \mid \phi(u^{S(x)}) \le \phi \le \alpha, x \in [a,b]^n\}$. Then, $\operatorname{conv}(S^\alpha) = \operatorname{conv}(\Theta^\alpha)$. Moreover, if ϕ is component-wise convex, then Θ^α is convex.

Proof. We first show that $u^{S(x')} \ge_m x'$ for each $x' \in [a,b]^n$. This follows because $u^{S(x')}$ simultaneously maximizes the continuous knapsack problems $\max\{\sum_{i=1}^j x_i \mid \sum_{i=1}^n x_i = S(x') + na, x \in [a,b]^n\}$ for all j because the ratio of objective and knapsack coefficient of x_i reduces with increasing i, and x' is a feasible solution to these knapsack problems.

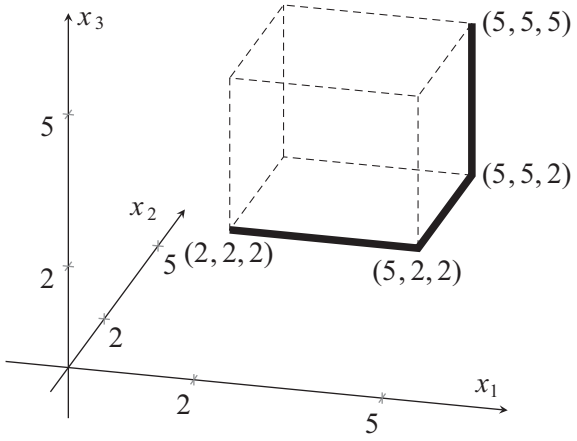
We now show that $S^\alpha \subseteq \Theta^\alpha$. Let $(x',\phi') \in S^\alpha$. Therefore, $\phi(u^{S(x')}) \le \phi(x') \le \phi \le \alpha$, where the first inequality follows from Schur concavity of ϕ and $u^{S(x')} \ge_m x'$, and the remaining inequalities follow because (x',ϕ') is feasible to S^α. Therefore, $(x',\phi') \in \Theta^\alpha$.

Now, we show that $\Theta^\alpha \subseteq \operatorname{conv}(S^\alpha)$. Let $(x',\phi') \in \Theta^\alpha$. Because $u^{S(x')} \in [a,b]^n$ and $\phi(u^{S(x')}) \le \phi' \le \alpha$, we conclude that $(u^{S(x')},\phi') \in S^\alpha$. Then it follows that $(x',\phi') \in \operatorname{conv}(S^\alpha)$ because $u^{S(x')} \ge_m x'$ implies that x' can be written as a convex combination of permutations of $u^{S(x')}$, and S^α is permutation invariant in x.

To show that Θ^α is convex when ϕ is component-wise convex, we write Θ^α as $\operatorname{proj}_{(x,\phi)} \Xi^\alpha$, where $\Xi^\alpha = \{(x,s,\phi) \mid \varphi(s) \le \phi \le \alpha, x \in [a,b]^n, s = \sum_{i=1}^n (x_i - a)\}$ and $\varphi(s) = \phi(u^s)$. The result follows if $\varphi(s)$ is convex over $[0, n(b-a)]$ because Θ^α is expressed as the projection of the convex set Ξ^α. First, observe that, for $s \in (i(b-a), (i+1)(b-a))$, the convexity of $\varphi(s)$ follows from the assumed convexity of $\phi(u^s)$ because, in this interval, u^s varies only along the ith coordinate. Now, choose $k \in \{0,\dots,n-1\}$, and let $\bar{s} = k(b-a)$. To prove the result, it suffices to verify that the left derivative of $\varphi(s)$ at $\bar{s}$ is no more than the corresponding right derivative. For any $\epsilon > 0$, observe that $u^{\bar{s}} + \epsilon e_k \ge_m u^{\bar{s}} + \epsilon e_{k+1}$. This follows because $(1-\lambda)(u^{\bar{s}}_k + \epsilon, u^{\bar{s}}_{k+1}) + \lambda(u^{\bar{s}}_{k+1}, u^{\bar{s}}_k + \epsilon) = (u^{\bar{s}}_k, u^{\bar{s}}_{k+1} + \epsilon)$, where $0 < \lambda = \frac{\epsilon}{u^{\bar{s}}_k - u^{\bar{s}}_{k+1} + \epsilon} \le 1$, showing that $u^{\bar{s}} + \epsilon e_{k+1}$ can be expressed as a convex combination of $u^{\bar{s}} + \epsilon e_k$ and its permutations. Because $\phi(\cdot)$ is Schur concave, it follows that $\phi(u^{\bar{s}} + \epsilon e_k) \le \phi(u^{\bar{s}} + \epsilon e_{k+1}) = \varphi(\bar{s} + \epsilon)$. Then, the following chain of inequalities follows:

$$\lim_{\epsilon \downarrow 0} \frac{\varphi(\bar{s}) - \varphi(\bar{s} - \epsilon)}{\epsilon} = \lim_{\epsilon \downarrow 0} \frac{\phi(u^{\bar{s}}) - \phi(u^{\bar{s}} - \epsilon e_k)}{\epsilon} \le \lim_{\epsilon \downarrow 0} \frac{\phi(u^{\bar{s}} + \epsilon e_k) - \phi(u^{\bar{s}})}{\epsilon} \le \lim_{\epsilon \downarrow 0} \frac{\varphi(\bar{s} + \epsilon) - \varphi(\bar{s})}{\epsilon},$$

Figure 1. The thicker segments indicate $\{u^{S(x)} \mid x \in [2,5]^3\}$, a set of points in Δ^3 that are not majorized by other points in $\mathbb{R}^3$ (up to permutation).



where the first equality is by the definition of $\varphi(\cdot)$ and $u^{\bar{s}}$, the first inequality is from the assumed convexity of $\phi(\cdot)$ when the argument is perturbated only along the kth coordinate, and the second inequality is because $\phi(u^{\bar{s}} + \epsilon e_k) \leq \varphi(\bar{s} + \epsilon)$ and $\phi(u^{\bar{s}}) = \varphi(\bar{s})$. □

In essence, Proposition 8 shows that we can reduce our attention to the edges of the hypercube belonging to Δ^n in our construction of $\text{conv}(S^\alpha)$; see Figure 1 for an illustration when $a = 2$ and $b = 5$. A similar result can be shown for upper level sets of quasiconcave functions over general polytopes (Tawarmalani and Richard [32]). Symmetric quasiconcave functions are a subclass of Schur-concave functions. In other words, both the results show that for symmetric quasiconcave functions over permutation-invariant polytopes, it suffices to consider the edges of the polytope to construct the convex hull. However, the result in Tawarmalani and Richard [32] applies to general quasiconcave functions over arbitrary polytopes, whereas Proposition 8 applies to Schur-concave functions over a hypercube. Perhaps more importantly, the result in Proposition 8 also applies to level sets of the functions, whereas the result in Tawarmalani and Richard [32] applies only to convex envelope construction.

Permutation invariance also helps with constructing compact extended formulations of certain nonlinear sets. To motivate this statement and introduce the following result, consider the set $Z' = \{(x,z) \in [a,b]^n \times \mathbb{R} \mid z = \prod_{i=1}^n x_i\}$. It is well known that, in order to generate $\text{conv}(Z')$, it suffices to restrict x to the vertices, $\mathcal{F}' = \{a,b\}^n$, of the hypercube $[a,b]^n$. More precisely, $\text{conv}(Z') = \text{conv}(\cup_{x \in \mathcal{F}'}(x, \prod_{i=1}^n x_i))$, where each disjunct in the union is a polytope with compact description (a single point). A higher-dimensional description of the convex hull of this union can be obtained using classical disjunctive programming results. Because the dimension of this formulation depends linearly on the number of disjuncts, $|\mathcal{F}'|$, such an approach has found limited practical use for the given example, as $|\mathcal{F}'|$ is exponential in n. Surprisingly, taking advantage of the permutation invariance of Z' through Theorem 1 allows for a much more economical use of disjunctive programming. Intuitively, this is because the number of elements of $\mathcal{F}'$ required to compute $\text{conv}(S_0)$ in Theorem 1 is polynomial in n. As a result, disjunctive programming provides a polynomial formulation for $\text{conv}(S_0)$, which can then be integrated with the result of Theorem 1 to obtain a polynomially sized higher-dimensional formulation of $\text{conv}(Z')$. The settings where such polynomially sized formulations can be obtained extend far beyond the example presented above, as we describe next in Theorem 11.

Consider now the much more general setup of sets $S(Z,a,b) := \{(x,z) \in [a,b]^n \times \mathbb{R}^m | (x,z) \in Z\}$, where Z is compact and permutation invariant in x. Furthermore, for $\mathcal{F} = \{F_1, \ldots, F_r\}$, where F_i are faces of $[a,b]^n$, define $X(Z,a,b,\mathcal{F}) := \{(x,z) \in [a,b]^n \times \mathbb{R}^m | (x,z) \in Z, x \in \cup_{i=1}^r F_i\}$. Using Theorem 1, we show next that a polynomial-size extended formulation can be constructed for any set $S(Z,a,b)$ for which there exists a collection of faces $\mathcal{F}'$ that completely determines the convex hull, that is, $\text{conv}(S(Z,a,b)) = \text{conv}(X(Z,a,b,\mathcal{F}'))$, and for which a polynomial (possibly extended) formulation of the set on each of these faces $F_i \in \mathcal{F}'$, that is, $\text{conv}(X(Z,a,b,\{F_i\}))$, can be obtained. The strength of this result is that we make no assumption on $|\mathcal{F}'|$ and that this collection may have exponentially many faces.

Theorem 11. *Let $a,b \in \mathbb{R}$, $Z \subseteq \{(x,z) \mid x \in \mathbb{R}^n, z \in \mathbb{R}^m\}$ be a compact permutation-invariant set with respect to x, and let $\mathcal{F} = \{F_1, \ldots, F_r\}$ be a collection of faces of $[a,b]^n$ such that $\operatorname{conv}(S(Z,a,b)) = \operatorname{conv}(X(Z,a,b,\mathcal{F}))$. Moreover, assume that $\operatorname{conv}(X(Z,a,b,\{F_i\}))$ has a polynomial-sized compact extended formulation for each $F_i \in \mathcal{F}$. Then, $\operatorname{conv}(S(Z,a,b))$ has a polynomial-sized extended formulation.*

Proof. For brevity of notation, in this proof, we shall write $S(Z,a,b)$ as S and $X(Z,a,b,\mathcal{F})$ as $X(\mathcal{F})$. We construct $\operatorname{conv}(S)$ using its equivalence to $\operatorname{conv}(X(\mathcal{F}))$. We may assume for computing $\operatorname{conv}(X(\mathcal{F}))$, by taking the union of all permutations of $X(\mathcal{F})$ with respect to x if necessary, that $X(\mathcal{F})$ is permutation invariant in x. This is because a permutation of $X(\mathcal{F})$ with respect to x, say, $X_\pi(\mathcal{F}) := \{(x,z) \mid (\pi(x),z) \in X(\mathcal{F})\}$, is contained in $\operatorname{conv}(X(\mathcal{F}))$, as is seen from $X_\pi(\mathcal{F}) \subseteq S \subseteq \operatorname{conv}(S) = \operatorname{conv}(X(\mathcal{F}))$, where the first inclusion is by permutation invariance of S and the equality is by the assumed hypothesis. Because S is permutation invariant with respect to x, by Lemma 1, $\operatorname{conv}(S)$ is also permutation invariant. We shall use Theorem 1 to construct $\operatorname{conv}(X(\mathcal{F}))$. We first show that we can limit the faces of $[a,b]^n$ that need to be considered in the construction of S_0; see Theorem 1. Consider an arbitrary face F_i of $[a,b]^n$, which is determined by setting a set of variables with indices in $B_i \subseteq \{1,\ldots,n\}$ to their upper bound b and a disjoint set of variables $A_i \subseteq \{1,\ldots,n\}$ to their lower bound a. We will show that the only faces, F_i, $i = 1,\ldots,r$, that need to be considered are such that B_i and A_i are *hole-free*; that is, B_i is of the form $\{1,\ldots,p\}$, and A_i is of the form $\{q,\ldots,n\}$. To see this, let $j(i) = \max\{j \mid j \in B_i\}$ and $X_i = X(\{F_i\}) \cap \{(x,z) \mid x_1 \geq \cdots \geq x_n\}$. It follows from Theorem 1 and permutation invariance of $X(\mathcal{F})$ that it suffices to consider points in $\cup_{i=1}^r X_i$ to construct $\operatorname{conv}(X(\mathcal{F}))$. We will further argue that it suffices to restrict attention to points in $\cup_{i:j(i)=|B_i|} X_i$. Notice that B_i is hole-free if and only if $j(i) = |B_i|$. Assume by contradiction that i' is the index of a face such that $j(i') > |B_{i'}|$ and that $X_{i'}$ contains a point that is not in $\cup_{i:j(i)=|B_i|} X_i$. Among all such faces, we choose i' to minimize $j(i) - |B_i|$. Because $B_{i'}$ is not hole-free, there exists $j \notin B_{i'}$ such that $j < j(i')$. Any point that belongs to $X_{i'}$ must satisfy $b \geq x_j \geq x_{j(i')} = b$. Therefore, $x_j = b$. Because $X_{i'} \neq \emptyset$, $j \notin A_i$. Consider now i'' such that $B_{i''} = B_i \cup \{j\} \setminus \{j(i')\}$ and $A_{i''} = A_{i'}$. Such a face exists in $\mathcal{F}$ because we assumed that for every face $F_i \notin \mathcal{F}$, $\mathcal{F}$ contains all faces obtained by permuting the variables, and $F_{i''}$ is obtained from $F_{i'}$ by exchanging the variables x_j and $x_{j(i')}$. We next show that $X_{i''} \supseteq X_{i'}$. For arbitrary $(x,z) \in X_{i'}$, $x_i = b$ for all $i \in B_{i'}$. Then, it follows that $x_i = b$ for all $i \leq j(i')$ as $x_1 \geq \cdots \geq x_n$. Therefore, $x_i = b$ for all $i \in B_{i''}$. Because $A_{i'} = A_{i''}$ and x_j is not restricted for other indices j, $(x,z) \in X_{i''}$. Therefore, $X_{i''}$ contains a point not in $\cup_{i:j(i)=|B_i|} X_i$, establishing that $j(i'') > |B_{i''}|$. However, because $j(i'') - |B_{i''}| < j(i') - |B_{i'}|$, this contradicts our choice of i'. A similar argument can be used to show that we only need to consider faces F_i such that A_i is hole-free.

There are at most $\binom{n+2}{2}$ such faces, one for each choice of (p,q), where $0 \leq p \leq q-1 \leq n$. Because each $X(\{F_i\})$ is assumed to have a polynomial-sized compact extended formulation, it follows, by disjunctive programming (Ben-Tal and Nemirovski [5, proposition 2.3.5]), that $\operatorname{conv}(S_0)$ has a polynomial-sized compact extended formulation. The result then follows directly from Theorem 1. □

We record and summarize the extended formulation of $\operatorname{conv}(S(Z,a,b))$ for later use in Corollary 4. We also observe that our construction applies even when Z is not compact, as long as the convex hull of $X(Z,a,b,\{F_i\})$ for each F_i of interest is available. For a face F of a polyhedron described by $a^\top x \leq b$, we refer to another face $\bar{F}$ as a *permutation* of F if $\bar{F}$ is described by $\bar{a}^\top \leq b$ where $\bar{a}$ is a permutation of a. We say that a collection of faces $\mathcal{F}$ of a polyhedron is permutation invariant if for a face in $\mathcal{F}$ described by an inequality, all its permutations are included in $\mathcal{F}$. For a face F, we define $l(F) = \{j \in \{1,\ldots,n\} \mid \hat{x}_j = a,\ \forall \hat{x} \in F\}$ and $u(F) = \{j \in \{1,\ldots,n\} \mid \hat{x}_j = b,\ \forall \hat{x} \in F\}$.

Corollary 4. *Let $a,b \in \mathbb{R}$, $Z \subseteq \{(x,z) \mid x \in \mathbb{R}^n, z \in \mathbb{R}^m\}$ be a permutation-invariant set with respect to x, and let $\mathcal{F} = \{F_1,\ldots,F_r\}$ be a permutation-invariant collection of faces of $[a,b]^n$. Let $I = \{i \in \{1,\ldots,r\} \mid \exists p,q, p<q \text{ s.t. } u(F_i) = \{1,\ldots,p\}$ and $l(F_i) = \{q,\ldots,n\}\}$. Then,*

$$\operatorname{conv}(X(Z,a,b,\mathcal{F})) = \left\{(x,z) \mid (u,z) \in \operatorname{conv}\left(\bigcup_{i\in I} X(Z,a,b,\{F_i\})\right), u_1 \geq \cdots \geq u_n, u \geq_m x\right\}.$$

We remark that $\operatorname{conv}(\cup_{i\in I} X(Z,a,b,\{F_i\}))$ can be constructed using disjunctive programming techniques if the recession cone of $X(Z,a,b,\{F_i\})$ does not depend on i (Rockafellar [31, corollary 9.8.1]). Theorem 11 shows that even though the number of faces in $\mathcal{F}$ may be exponentially large, we can exploit the permutation invariance of the set to consider only a polynomial number of faces in the construction. More explicitly, there are $2^{n-d}\binom{n}{d}$ d-dimensional faces of $[a,b]^n$ and $\binom{n+1}{d+1}$ d-dimensional faces of the simplex $b \geq x_1 \geq \cdots \geq x_n \geq a$. But, there are

only $n-d+1$ of the d-dimensional faces of the hypercube, namely, F_l for $l \in \{0,\dots,n-d\}$, that are defined by the hole-free sets $B_l = \{1,\dots,l\}$ and $A_l = \{l+d+1,\dots,n\}$ with the convention that $B_0 = A_{n-d} = \emptyset$.

Applications of Theorem 11 extend beyond Schur-concave functions. For example, consider the convex hull of $\{(x,\alpha) \in [a,b]^n \times \mathbb{R} \,|\, \prod_{i=1}^n x_i \le \alpha\}$, where a is not necessarily positive. The product function is not Schur concave when some of the variables take negative values; for example consider the function $x_1x_2x_3$ and observe that although $(1,-1,-3) \ge_m (0,0,-3)$, the function value is higher at $(1,-1,-3)$ than at $(0,0,-3)$.

There are many functions besides the multilinear monomial that are permutation invariant and whose envelopes are determined by their values on certain low-dimensional faces—the postulates required for the construction in Theorem 11. We discuss some examples next. Observe that all elementary symmetric polynomials satisfy these postulates because, being multilinear, their convex envelope is generated by vertices of $[a,b]^n$. Because Theorem 11 allows z to be multidimensional, this result in fact yields the simultaneous convexification of the elementary symmetric monomials over $[a,b]^n$ and thus the convex envelope of positive linear combinations of elementary symmetric polynomials. Next, consider the function $\prod_{i=1}^n x_i^t$ over $[a,b]^n$, where $a,t>0$. If $t \le 1$, the function is concave when all but one variable is fixed, and therefore the convex envelope is generated by vertices of $[a,b]^n$. If $t \ge 1$, then by restricting attention to any face of the hypercube, where two or more variables are not fixed, say x_1 and x_2, we see that the determinant of the Hessian of $x_1^t x_2^t$ is $-x_1^{2t-2}x_2^{2t-2}t^2(2t-1)$, which is strictly negative. Therefore, there is a direction in which the function is concave and the convex envelope is determined by one-dimensional faces of the hypercube. Because the function is convex on these faces, the convex envelope can be developed using Corollary 4. In fact, when $t \ge 1$, Proposition 8 provides a closed-form expression for the convex envelope. This is because the function $f(x) = \prod_{i=1}^n x_i^t$ is Schur concave because $(x_i - x_j)\left(\frac{\partial f}{\partial x_i} - \frac{\partial f}{\partial x_i}\right) = t(x_i-x_j)^2 \frac{\prod_{i=1}^n x_i^t}{x_i x_j} > 0$.

We next expand on the characterization we gave in Corollary 4 to obtain a more streamlined description of the result for the particular cases of permutation-invariant functions whose convex envelopes are completely determined by the extreme points of their domain, which is assumed to be a hypercube.

Proposition 9. *Consider a function $\phi(x) : [a,b]^n \mapsto \mathbb{R}$ that is permutation invariant in x and whose convex envelope remains the same even if its domain is restricted to $\{a,b\}^n$. For $i = 1,\dots,n$ and $j = 0,\dots,n$, let $p_{ij} = a$ if $i > j$ and b otherwise, and let $p_{\cdot j}$ denote the jth column of this matrix. Define $f(x) := \phi(p_{\cdot 0}) + \sum_{i=1}^n \frac{x_i - a}{b-a}(\phi(p_{\cdot i}) - \phi(p_{\cdot i-1}))$. Then, the convex envelope of $\phi(x)$ over $[a,b]^n$ can be expressed as the value function of the following problem:*

$$\operatorname{conv}_{[a,b]^n}(\phi)(x) = \min\{f(u) \,|\, u \ge_m x, b \ge u_1 \ge \cdots \ge u_n \ge a\}. \tag{26}$$

Proof. Observe that the points in $\{a,b\}^n$ that intersect with $x_1 \ge \cdots \ge x_n$ are precisely the columns $p_{\cdot j}$ described in the statement of the result. Consider the column $p_{\cdot j}$, and observe that $f(p_{\cdot j}) = \phi(p_{\cdot j})$. Moreover, f is affine. Let $\Gamma = \{x \in \{a,b\}^n \,|\, x_1 \ge \cdots \ge x_n\}$. Then, we show that $f = \operatorname{conv}_{\operatorname{conv}(\Gamma)}(\phi|_\Gamma)$, where $\phi|_\Gamma$ denotes the restriction of ϕ to Γ. Clearly, $f(x) \le \operatorname{conv}_{\operatorname{conv}(\Gamma)}(\phi|_\Gamma)(x)$ because it matches $\phi|_\Gamma$ at all the points in the domain and is a convex underestimator. Also, $f(x) \ge \operatorname{conv}_{\operatorname{conv}(\Gamma)}(\phi|_\Gamma)(x)$ because of Jensen's inequality applied to $\operatorname{conv}_{\operatorname{conv}(\Gamma)}(\phi)$, observing that $f(x)$ is exact at the extreme points of Γ, and because $f(x)$ is affine. Now, consider Corollary 4. Let $\mathcal{F}$ be the set of extreme points of $[a,b]^n$ and $Z = \{(x,z) \,|\, z \ge \phi(x)\}$. By assumption, $\operatorname{conv}(S(Z,a,b)) = \operatorname{conv}(X(Z,a,b,\mathcal{F}))$. By Corollary 4, it follows that $\operatorname{conv}(X(Z,a,b,\mathcal{F})) = \left\{(x,z) \mid (u,z) \in \operatorname{conv}(\cup_{j=0}^n X(Z,a,b,\{p_{\cdot j}\})), u \ge_m x\right\} = \{(x,z) \mid z \ge f(u), b \ge u_1 \ge \cdots \ge u_n \ge a, u \ge_m x\}$. □

In the next result, we show how the convex hull of $\{(x,y) \in [a,b]^n \times [c,d]^m \,|\, \prod_{i=1}^m y_i^\alpha = \prod_{j=1}^n x_j^\beta\}$ can be constructed. To do so, it suffices to construct the convex hull of $S = \{(x,y) \in [a,b]^n \times [c,d]^m \,|\, \prod_{i=1}^m y_i^\alpha \ge \prod_{j=1}^n x_j^\beta\}$, because the convex hull for the reverse inequality can be developed by switching the x and y variables, and the convex hull for the equality can then be obtained as the intersection of these two convex hulls; see Nguyen et al. [28]. Such a set occurs in polynomial optimization problems where linearized variables are introduced to relax $uv = y$, $u^2 = x_1$, and $v^2 = x_2$ in the form of $y^2 = x_1x_2$. Similar higher-dimensional equality constraints also occur, and their relaxations can be employed in polynomial optimization problems.

Proposition 10. *Consider $S = \{(x,y) \in H \,|\, \prod_{i=1}^m y_i^\alpha \ge \prod_{j=1}^n x_j^\beta\}$, where $H = [a,b]^n \times [c,d]^m$, with $a \ge 0$, $c \ge 0$, $\alpha > 0$, and $\beta > 0$. Let $k = \min\{m, \lfloor \frac{\beta}{\alpha} \rfloor\}$. Define the convex sets*

$$S_{ij} = S \cap \left\{ (x,y) \in H \,\middle|\, \begin{array}{l} (y_r)_{r=1}^i = d, (y_r)_{r=i+k+1}^m = c, \quad y \in \Delta^m \\ (x_s)_{s=1}^j = b, (x_s)_{s=j+2}^n = a \end{array} \right\},$$

$$C_j = S \cap \left\{ (x,y) \in H \,|\, (x_s)_{s=1}^j = b, (x_s)_{s=j+1}^n = a, y \in \Delta^m \right\},$$

where S_{ij} is defined for $i = 0,\dots,m-k$ and $j = 0,\dots,n-1$, and C_j is defined for $j = 0,\dots,n$. Let $T = \cup_{i,j} S_{ij} \cup \cup_j C_j$. Then

$$\operatorname{conv}(S) = X := \{(x,y) \mid v \geq_m y, u \geq_m x, (u,v) \in \operatorname{conv}(T)\}.$$

In particular, if $m\alpha \leq \beta$, then

$$\operatorname{conv}(S) = X' := \left\{(x,y) \in H \;\middle|\; \prod_{i=1}^{m} y_i^{\frac{1}{m}} \geq \prod_{j=1}^{n} u(x)_j^{\frac{\beta}{m\alpha}}\right\}, \tag{27}$$

where $u(x) := u^{S(x)}$, defined as in Proposition 8.

Before providing the proof, we discuss its architecture. This proof will write the convex hull of S as a disjunctive hull of convex subsets of S. Because these convex subsets will be obtained by fixing variables at their bounds, even though they are exponentially many, Corollary 4 will allow the construction of the convex hull. In the first part of the proof, we consider a slice of S at a fixed y and show that it suffices to restrict x to one-dimensional faces of $[a,b]^n$ for constructing the convex hull of the slice. Then, we show that there are two cases. If the remaining x_j variable is also fixed to the bounds, the set is already convex. If not, then we show that any point in such a set cannot be extremal in S unless at least a certain number of y variables, specifically $\max\{0, m - \lceil \frac{\beta}{\alpha} \rceil\}$ of them, are fixed to their bounds. The sets obtained by fixing these variables are once again convex. Then, it will follow by Theorem 11 that we may restrict our attention to the faces in T and, as a result, $\operatorname{conv}(S) = X$.

Proof. Let $\phi(x) := \prod_{j=1}^{n} x_j^{\beta}$, and consider the set $\Upsilon(\gamma) = \{x \in [a,b]^n \mid \phi(x) \leq \gamma\}$. By Marshall et al. [24, theorem 3.A.3], $\phi(x)$ is Schur concave over $[a,b]^n$ because it is permutation invariant and $\frac{\partial \phi}{\partial x_1} \leq \cdots \leq \frac{\partial \phi}{\partial x_n}$ at any point with $x_1 \geq \cdots \geq x_n$. Let $\Upsilon_i(\gamma) = \{x \in \{b\}^{i-1} \times [a,b] \times \{a\}^{n-i} \mid x_i^{\beta} b^{(i-1)\beta} a^{(n-i)\beta} \leq \gamma\}$. Then, it follows by Proposition 8 and Corollary 4 that $\operatorname{conv}(\Upsilon(\gamma)) = \{x \mid u \geq_m x, u_1 \geq \cdots \geq u_n, u \in \operatorname{conv}(\cup_{i=1}^{n} \Upsilon_i(\gamma))\}$.

Because $n-1$ of the x_j variables can be fixed to bounds in the construction of the convex hull of each slice, this restriction can also be imposed in the construction of $\operatorname{conv}(S)$. Now, let $\psi(y) := \prod_{i=1}^{m} y_i$ and consider the slightly more general set Θ that will appear when we fix some of the y variables at their bounds. This set is defined as $\Theta = \{(x,y) \in [a,b] \times [c,d]^m \mid \zeta \psi(y) \geq \delta x^{\frac{\beta}{\alpha}}\}$, where $\delta, \zeta \geq 0$. Consider a point $(x',y') \in \Theta$. Then, by restricting attention to $\bar{y} = \lambda y'$, where $\lambda \in \mathbb{R}$, we obtain an affine transform of a subset Λ of Θ such that $\Lambda = \{(x,\lambda) \in [a,b] \times [c',d'] \mid \lambda\theta \geq \delta' x^{\frac{\beta}{m\alpha}}\}$, where $\theta = \zeta^{\frac{1}{m}} \psi(y')^{\frac{1}{m}}$, $\delta' = \delta^{\frac{1}{m}}$, $c' = \max\{\lambda \mid \lambda y_i' \leq c \text{ for some } i\}$, and $d' = \min\{\lambda \mid \lambda y_i' \geq d \text{ for some } i\}$.

Assume $x' \in (a,b)$ and $y' \in (c,d)^m$. We will first show that if $m > \frac{\beta}{\alpha}$, such a point is not an extreme point of S. By definition, $c' < 1$, $d' > 1$, and $(x',1) \in \Lambda$. Assume $m > \frac{\beta}{\alpha}$ and (x',y') is an extreme point of S. Define $s = \delta' \frac{\beta}{m\alpha} (x')^{\frac{\beta}{m\alpha} - 1}$. If $x' \in (a,b)$, then, for sufficiently small $\epsilon > 0$, we show that $(x',1)$ can be written as a convex combination of $(x' - \epsilon\theta, 1 - s\epsilon)$ and $(x' + \epsilon\theta, 1 + s\epsilon)$. The latter points are feasible in Λ because

$$\delta'(x' \pm \epsilon\theta)^{\frac{\beta}{m\alpha}} \leq \delta' x'^{\frac{\beta}{m\alpha}} + s(x' \pm \epsilon\theta - x') \leq \theta(1 \pm s\epsilon),$$

where the first inequality is by concavity of $x^{\frac{\beta}{m\alpha}}$ for $m \geq \frac{\beta}{\alpha}$, and the second inequality is because $\delta' x'^{\frac{\beta}{m\alpha}} \leq \theta$ as $(x',1)$ belongs to Λ. Because Λ is an affine transform of a subset of Θ, we have expressed (x',y') as a convex combination of $(x' - \epsilon\theta, (1 - s\epsilon)y')$ and $(x' + \epsilon\theta, (1 + s\epsilon)y')$, each of which is feasible to Θ. Because $\epsilon > 0$ and $x' > a \geq 0$ implies $s > 0$, it follows that these points are distinct, thus contradicting the extremality of (x',y').

Because, S is compact, in order to construct $\operatorname{conv}(S)$, we may restrict our attention to the extreme points of S. Therefore, either $x' \in \{a,b\}$ or there exists an i such that $y_i' \in \{c,d\}$. If $x' \in \{a,b\}$, the point belongs to the convex subset of Θ obtained by fixing x' at its current value because the defining inequality can be written as $\zeta^{\frac{1}{m}} \psi(y)^{\frac{1}{m}} \geq \delta x'^{\frac{\beta}{\alpha}}$, a convex inequality. A permutation of such an (x',y') is included in one of the C_js. On the other hand, if $y_i' \in \{c,d\}$, we can reduce the dimension of the set by fixing y_i at y_i' and, thus, effectively reduce m. Therefore, we may assume without loss of generality that $m \leq \frac{\beta}{\alpha}$. Then, we rewrite the defining inequality of Θ as $\zeta^{\frac{1}{m}} \psi(y)^{\frac{1}{m}} \geq \delta^{\frac{1}{m}} x^{\frac{\beta}{m\alpha}}$ and observe that this is a convex inequality because $\psi(y)^{\frac{1}{m}}$ is a concave function and $x^{\frac{\beta}{m\alpha}}$ is a convex function. Therefore, we need to consider faces where either all x_j are fixed at their bounds or where we fix all y_i except for a subset of cardinality $\min\{m, \lfloor \frac{\beta}{\alpha} \rfloor\}$ and fix all x_j except for one; all such sets are, up to permutation, included in one of the C_js or S_{ij}s.

Because $S_{ij} \subseteq S$, $C_j \subseteq S$, and S is permutation invariant, it follows that $\operatorname{conv}(S) \supseteq X$. Because X is convex and S is compact, we only need to show that the extreme points of S are contained in X. However, we have shown that

the extreme points of S belong to T or a set obtained from T by permuting the x and/or y variables. Because X is permutation invariant and contains T, it follows that every extreme point belongs to X. Therefore, $X = \text{conv}(S)$.

Now, consider the case where $m\alpha \le \beta$. Clearly, $k = m$. If we fix y at $\bar{y}$, it follows from the Schur concavity of $\prod_{j=1}^{n} x_i^{\frac{\beta}{m\alpha}}$ over $[a,b]^n$, the convexity of $x_i^{\frac{\beta}{m\alpha}}$ over $[a,b]$, and Proposition 8 that the convex hull of this slice is defined by $\prod_{i=1}^{m} \bar{y}_i^{\frac{1}{m}} \ge \prod_{j=1}^{n} u(x)_j^{\frac{\beta}{m\alpha}}$. This shows that $X' \subseteq \text{conv}(S)$. By Schur concavity of $\prod_{j=1}^{n} x_j^{\frac{\beta}{m\alpha}}$, it follows that $\prod_{j=1}^{n} x_j^{\frac{\beta}{m\alpha}} \ge \prod_{j=1}^{n} u(x)_j^{\frac{\beta}{m\alpha}}$, or $S \subseteq X'$. This implies that $S \subseteq X' \subseteq \text{conv}(S)$. To show that $X' = \text{conv}(S)$, we only need to show that X' is convex. As in the proof of Proposition 8, we let $\varphi(s) = \prod_{j=1}^{n} u(x)_j^{\frac{\beta}{m\alpha}}$, where $s = \sum_{i=1}^{n}(x_i - a)$, and rewrite the above inequality as $\varphi(s) - \prod_{i=1}^{m} y_i^{\frac{1}{m}} \le 0$. Because the left-hand side is jointly convex in (s, y) and s is a linear function of x, this proves that X' is convex in (x, y). □

So far in this section, we have given various results where we describe the convex hull of a set in an extended space by introducing variables u. We now discuss how inequalities in the original space can be obtained by solving a separation problem.

Usually, given a set X and an extended space representation of its convex hull, C, we separate a given point $\bar{x}$ from X by solving the problem $\inf_{(x,u)\in C}\|x - \bar{x}\|$. By duality, the optimal value matches $\max_{\|a\|_* \le 1}\{\langle \bar{x}, a\rangle - h(a)\}$, where $h(\cdot)$ is the support function of C and $\|\cdot\|_*$ is the dual norm. Then, if the optimal value, z^* is strictly larger than zero and the optimal solution to the dual problem is a^*, we have $\langle \bar{x}, a^*\rangle - z^* \ge \langle x, a^*\rangle$ for all $x \in \text{proj}_x C$, and this inequality separates $\bar{x}$ from $\text{proj}_x C$.

However, such an inequality is typically not facet defining for $\text{conv}(X)$ even when the latter set is polyhedral. We now discuss a separation procedure that often generates facet-defining inequalities. The structure of permutation-invariant sets and their extended space representation allow for this alternate approach. Assume we are interested in developing the convex envelope of a permutation-invariant function ϕ, such as $\prod_{i=1}^{n} x_i$, over $[a,b]^n$. As in Theorem 1, the convex envelope of ϕ at x is obtained by expressing x as a convex combination of u and its permutations, where $u \ge_m x$. Moreover, assume that the convex envelope at u is obtained as a convex combination of the extreme points of the simplex $a \le x_1 \le \cdots \le x_n \le b$ with convex multipliers γ. Because $x = Su$ for some doubly stochastic matrix S and $u = V\gamma$, where the columns of V correspond to vertices of the simplex, it follows that $x = SV\gamma$. Therefore, we can find a representation of x as a convex combination of vertices in V and their permutations.

We implement the above procedure for multilinear sets over $[a,b]^n$ to evaluate its impact on the quality of the relaxation. For the purpose of illustration, we consider the special case of $\prod_{i=1}^{n} x_i$ over $[a,b]^n$. In this case, (26) reduces to

$$\begin{aligned} \min \quad & a^n + \sum_{i=1}^{n} b^{i-1} a^{n-i}(u_i - a), \\ \text{s.t.} \quad & u \ge_m x, \\ & b \ge u_1 \ge \cdots \ge u_n \ge a. \end{aligned} \tag{28}$$

Given $x \in \mathbb{R}^n$ in general position inside $[a,b]^n$, assume that the optimal solution to (28) is u. Then, we express $x = Su$, where $S \in \mathcal{M}^{n,n}(\mathbb{R})$ is a doubly stochastic matrix. Given x and u, this can be done through the solution of a linear program, $\min\{0 | x = Su, S\mathbb{1} = \mathbb{1}, \mathbb{1}^\top S = \mathbb{1}^\top, S \ge 0\}$. Although S can also be derived as a product of T-transforms (see Hardy et al. [12, section 2.19, proof of lemma 2]), we use the linear programming approach in our numerical experiments, given its simplicity of implementation. Then, we express S as a convex combination of permutation matrices. Such a representation exists due to the Birkhoff theorem. We obtain it using a straightforward algorithm, which we describe next. Observe first that the desired representation is such that all permutation matrices with nonzero convex multipliers have a support that is contained within the support of S. This implies that the bipartite graph, which we describe next, has a perfect matching. The bipartite graph is constructed with nodes labeled $\{1, \ldots, n\}$ in each partition and edges that connect a node i in the first partition to j in the second partition if and only if $S_{ij} > 0$. Given a perfect matching, we construct a permutation matrix P so that $P_{ij} = 1$ if node i in the first partition is matched to node j in the second partition. Then, we associate P with a convex multiplier π that is chosen to be $\min_{ij}\{S_{ij} | P_{ij} = 1\}$. If $\pi = 1$, we have a representation of S as a convex combination of permutation matrices. Otherwise, observe that $\frac{1}{1-\pi}(S - \pi P)$ is again a doubly stochastic matrix with one less nonzero entry. Therefore, by recursively using the above approach, we obtain S as a convex combination of at most n^2 permutation matrices. Then, we permute u according to these permutation matrices. For each such u,

the convex envelope is given by the optimal function value of (28). Each permuted u can be expressed as a convex combination of the corner points of the permuted simplex $\{b \geq u_1 \geq \cdots \geq u_n \geq a\}$. We claim that an affine underestimator $\breve{\phi}$ of $\text{conv}(\phi)$ that is such that $\breve{\phi}(x) = \text{conv}(\phi)(x)$ must also satisfy $\breve{\phi}(v^i) = \text{conv}(\phi)(v^i)$ for every vertex v^i of $[a,b]^n$ that has a nonzero multiplier in the representation of x computed above. If not, we have convex multipliers λ_i, where $x = \sum_i \lambda_i v^i$, such that

$$\breve{\phi}(x) = \text{conv}(\phi)(x) = \sum_i \lambda_i \,\text{conv}(\phi)(v^i) = \sum_i \lambda_i \phi(v^i) > \sum_i \lambda_i \breve{\phi}(v^i) = \breve{\phi}(x),$$

which yields a contradiction. Here, the first equality is by the definition of $\breve{\phi}$, second equality is because our construction obtains $\text{conv}(\phi)(x)$ as a convex combination of $\text{conv}(\phi)(v^i)$ using multipliers λ^i, the third equality is because v^i are extreme points, the first inequality is because there exists an i such that $\phi(v^i) > \breve{\phi}(v^i)$, and the final equality is because $\breve{\phi}$ is affine. Therefore, it follows that $\breve{\phi}(v^i) = \phi(v^i)$ for all i. Because x was assumed to be in general position, these equality constraints uniquely identify $\breve{\phi}$.

We conclude this section by presenting the results of a numerical experiment that suggests that the bounds obtained when building convex relaxations of $\psi_n(x) = \prod_{i=1}^n x_i$ over $[a,b]^n$ using the procedure described above are significantly stronger than those obtained using factorable relaxations. To this end, we consider functions $\psi_n(x)$ where $n = 10$ over two permutation-invariant hyper-rectangles. The first one, $B_1 = [2,4]^{10}$, is contained in the positive orthant, whereas the second, $B_2 = [-2,3]^{10}$, contains zero in its interior. We generate nine sample points uniformly at random inside of B_1 and B_2. At each point, we compare the value z_r of the relaxation obtained using a recursive application of McCormick's procedure with the value z_e of the convex envelope, obtained using the results described in this section. We then compute the existing gap ("Gap" in Tables 1 and 2) and relative gap ("% gap") using the formulas $z_e - z_r$ and $\frac{z_e - z_r}{z_e}$, respectively. Results are presented in Tables 1 and 2, where it can be observed that the proposed approach leads to substantial improvements in bounds, especially when variables x take both positive and negative values.

5. Set of Rank-One Matrices Associated with Permutation-Invariant Sets

For a positive integer n and a given set $S \in \mathbb{R}^n$, define $M_S := \{(x,X) \in \mathbb{R}^n \times \mathcal{M}^n \mid X = xx^\top, x \in S\}$, where $\mathcal{M}^n = \mathcal{M}^{n,n}(\mathbb{R})$. For each element $(x,X) \in M_S$, it is clear that $\text{rank}(X) = 1$. Studying M_S is particularly important when constructing valid inequalities for semidefinite relaxations of nonconvex optimization problems. In this section, we study the case where the base set S is permutation invariant.

For an arbitrary positive integer n, we denote by $\mathbb{1}_n$ the n-dimensional vector of ones. When n is clear in the context, we simply denote it by $\mathbb{1}$. Similarly, we denote by $\mathbb{1}_{n\times n}$ the n-by-n matrix of ones.

As a motivating example, consider sparse PCA, which, for a given covariance matrix Σ, finds a sparse vector x that maximizes the variance $x^\top \Sigma x$. A semidefinite relaxation of sparse PCA therefore aims to approximate

$$M := \{(x,X) \in \mathbb{R}^n \times \mathcal{M}^n \mid X = xx^\top, \|x\| \leq 1, \text{card}(x) \leq K\} \tag{29}$$

for a positive integer $K \in \{1,\ldots,n-1\}$. The set M can be seen as M_S by choosing S to be the permutation-invariant set $S = \{x \in \mathbb{R}^n \mid \|x\| \leq 1, \text{card}(x) \leq K\}$. The separation problem associated with M is known to be NP-hard (Tillmann and Pfetsch [33]). Hence, semidefinite relaxations have been considered that relax the nonconvex constraint $X = xx^\top$ with $X \succeq xx^\top$, which is equivalent to the convex constraint $\begin{bmatrix} X & x \\ x^\top & 1 \end{bmatrix} \succeq 0$. Linear valid inequalities in (x, X) are then developed by exploiting the property that $X = xx^\top$. For example, the authors of d'Aspremont et al. [7] introduce the valid inequality $\mathbb{1}^\top X \mathbb{1} \leq K$ for (29), which is implied by valid inequality $\sum_{i=1}^n x_i \leq \sqrt{K}$ and the condition $X = xx^\top$.

We next show that additional valid inequalities can be constructed in a higher-dimensional space by using the permutation invariance of the base set S. To this end, we prove the following result.

Proposition 11. *Suppose $S \subseteq \mathbb{R}^n$ is a permutation-invariant set. Let*

$$\mathcal{N} = \left\{ (x,u,X,U) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathcal{M}^n \times \mathcal{M}^n \;\middle|\; \begin{array}{l} X = xx^\top, U = uu^\top, \\ x = Pu \text{ for some } P \in \mathcal{P}_n, \\ u \in S \cap \Delta \end{array} \right\}.$$

Then, $M_S = \text{proj}_{(x,X)} \mathcal{N}$.

Proof. To prove $M_S \subseteq \text{proj}_{x,X}\mathcal{N}$, consider $(x, X) \in M_S$. Let $P \in \mathcal{P}_n$ be such that $u := P^{-1}x \in \Delta$, and let $U = uu^\top$. Then, $(x, u, X, U) \in \mathcal{N}$. To prove $M_S \supseteq \text{proj}_{x,X}\mathcal{N}$, consider $(x, u, X, U) \in \mathcal{N}$, and let $P \in \mathcal{P}_n$ be such that $x = Pu$. By permutation invariance of S, $x \in S$, showing $(x, X) \in M_S$. □

Now we develop linear inequalities implied by the conditions in $\mathcal{N}$. For any $(x, u, X, U) \in \mathcal{N}$, observe that $X = xx^\top = (Pu)(Pu)^\top = PUP^\top$ for a permutation matrix P. Therefore, consider linear inequalities implied by the facts that x is a permutation of u, and X is obtained by permuting some columns and rows of U symmetrically. Perhaps, the most straightforward such inequalities are

$$\mathbb{1}^\top u = \mathbb{1}^\top x, \tag{30a}$$

$$\text{trace}(X) = \text{trace}(U), \tag{30b}$$

$$\mathbb{1}^\top X \mathbb{1} = \mathbb{1}^\top U \mathbb{1}. \tag{30c}$$

More generally, consider any function $\phi : \mathbb{R}^n \times \mathcal{M}^n \to \mathbb{R}$ such that $\phi(x, xx^\top)$ is permutation invariant with respect to x. Then, we can impose the equality $\phi(x, X) = \phi(u, U)$ if ϕ is linear in (x, X). In fact, if ϕ is linear in (x, X), then we argue that this identity is implied by (30). To see this, observe that $\phi(x, xx^\top)$ is a quadratic function in x. Let $\psi(x) = \phi(x, xx^\top) = x^\top Cx + d^\top x + e$ for $C \in \mathbb{S}^n, d \in \mathbb{R}^n$, and $e \in \mathbb{R}$. By permutation invariance of ψ, the function $\psi(x) - \psi(Px)$ is the zero function in x for every $P \in \mathcal{P}_n$. Observe that

$$\psi(x) - \psi(Px) = x^\top (C - P^\top CP)x + (d - P^\top d)^\top x \equiv 0.$$

Therefore, $d = P^\top d$ and $C = P^\top CP$. For $i \neq j \in \{1, \ldots, n\}$, consider the permutation matrix P such that $(Px)_i = x_j$, $(Px)_j = x_i$, and $(Px)_k = x_k$ for all $k \neq i, j$. Then, $d_i = d_j, C_{ii} = C_{jj}$, and $C_{ij} = C_{ji}$. Because the choices for i and j are arbitrary, it holds that $d = \rho\mathbb{1}$ for some $\rho \in \mathbb{R}^n$, the diagonal entries of C are identical, and C is symmetric. We next claim that all off-diagonal entries of C are identical. We assume $n \geq 3$ because it is clear otherwise. For any $q \in \{1, \ldots, n\} \setminus \{i, j\}$, $[C - P^\top CP]_{iq} = C_{iq} - C_{jq}$; that is, all entries of qth column of C except for C_{qq} are equal. By symmetry of C, all entries of qth row of C except for C_{qq} are equal. Because q is arbitrary, all off-diagonal entries of C are identical because for any $i < j$ and $p < q$ with $q < j$, $C_{ij} = C_{pj} = C_{pq}$. Therefore,

$$\begin{aligned}\psi(x) &= x^\top (c_1 \text{diag}(\mathbb{1}) + c_2 \mathbb{1}_{n\times n})x + (\rho\mathbb{1})^\top x + e \\ &= c_1 \text{trace}(xx^\top) + c_2 \mathbb{1}^\top (xx^\top)\mathbb{1} + \rho(\mathbb{1}^\top x) + e.\end{aligned}$$

Now, the desired equality $\phi(x, X) = \phi(u, U)$ is

$$c_1 \text{trace}(X) + c_2 \mathbb{1}^\top X \mathbb{1} + \rho(\mathbb{1}^\top x) = c_1 \text{trace}(U) + c_2 \mathbb{1}^\top U \mathbb{1} + \rho(\mathbb{1}^\top u),$$

which is implied by equalities in (30).

Another type of constraints can be obtained when $S \subseteq \mathbb{R}^n_+$; that is, $u \in S$ is chosen to be nonnegative and in descending order. Then, entries in each row of $uu^\top$ are also in descending order, yielding the inequalities

$$U_{i,j} \geq U_{i,j+1}, \quad 1 \leq i \leq n,\ 1 \leq j < n - 1. \tag{31}$$

Similar arguments can be made for column entries. These inequalities, however, are redundant because of the symmetry of U.

Table 1. Gap at a randomly chosen point for $\prod_{i=1}^{10} x_i$ on $[2, 4]^{10}$.

Sample	z_e	z_r	Gap	% gap (%)
1	18,943.5	7,584.8	11,358.6	60.0
2	52,904.9	21,933.9	30,970.9	58.5
3	22,754.2	8,622.1	14,132.1	62.1
4	26,299.0	8,526.7	17,772.3	67.6
5	13,817.1	5,750.7	8,066.3	58.4
6	25,028.6	8,906.2	16,122.4	64.4
7	13,852.4	5,694.1	8,158.3	58.9
8	16,059.4	8,069.1	7,990.2	49.8
9	10,122.1	4,812.2	5,309.9	52.5
Average				59.1

Table 2. Gap at a randomly chosen point for $\prod_{i=1}^{10} x_i$ on $[-2,3]^{10}$.

Sample	z_e	z_r	Gap	% gap (%)
1	−12,314.6	−25,655.4	13,340.9	108.3
2	−16,221.2	−29,559.4	13,338.2	82.2
3	−13,247.0	−29,405.9	16,158.9	122.0
4	−14,069.4	−28,248.4	14,179.0	100.8
5	−10,660.9	−23,134.2	12,463.3	116.9
6	−10,979.5	−21,263.1	10,283.7	93.7
7	−9,367.8	−21,327.4	11,959.6	127.7
8	−10,245.9	−24,782.6	14,536.8	141.9
9	−9,182.8	−21,137.0	11,954.2	130.2
Average				113.7

We next introduce a general framework that constructs tighter linear relaxations by exploring the conceptual relationship that x is a permutation of u. This allows us to model or relax identities of the form $\phi(x, xx^\top) = \phi(u, uu^\top)$, where ϕ is a certain real-valued nonlinear permutation-invariant function. To this end, for fixed integers $p, q \in \{1, \dots, n\}$, $r \in \{1, \dots, \min\{pn, qn\}\}$, and $W \in \mathcal{M}^n$, consider the optimization problem

$$\max \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij} t_{ij}$$

$$\text{s.t.} \sum_{j=1}^{n} t_{ij} \le q, \qquad i \in \{1, \dots, n\}, \tag{32a}$$

$$\sum_{i=1}^{n} t_{ij} \le p, \qquad j \in \{1, \dots, n\}, \tag{32b}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} t_{ij} \le r, \tag{32c}$$

$$0 \le t_{ij} \le 1, i, \qquad j \in \{1, \dots, n\}. \tag{32d}$$

Its dual is

$$\min \quad q\sum_{i=1}^{n} \alpha_i + p\sum_{j=1}^{n} \beta_j + r\gamma + \sum_{i=1}^{n}\sum_{j=1}^{n} \delta_{ij}$$

$$\text{s.t.} \quad \alpha_i + \beta_j + \gamma + \delta_{ij} = W_{ij} + \theta_{ij}\ i,\ j \in \{1, \dots, n\}, \tag{33a}$$

$$\alpha_i \ge 0, \beta_j \ge 0, \gamma \ge 0, \delta_{ij} \ge 0, \theta_{ij} \ge 0\, i,\ j \in \{1, \dots, n\}, \tag{33b}$$

where dual variables α, β, γ, and δ correspond to primal constraints (32a), (32b), and (32c), and the upper-bound constraints of (32d), respectively. We denote these optimization problems by $\max\{f^W(t) \mid t \in \Phi\}$ and $\min\{g(z) \mid z \in \Omega(W)\}$. Strong duality holds because both (32) and (33) are feasible. We denote $h(W) := \max\{f^W(t) \mid t \in \Phi\} = \min\{g(z) \mid z \in \Omega(W)\}$. Observe that $h(ww^\top)$ as a function of w is permutation invariant with respect to w. Therefore, if $(x, u, X, U) \in \mathcal{N}$, it holds that $h(U) = h(X)$. Because the linearity of the identity is not guaranteed, we construct linear inequalities in (x, u, X, U) by taking the identity and the conditions in the set description of $\mathcal{N}$ into account. In the following discussion, we assume that $(x, u, X, U) \in \mathcal{N}$.

We first consider the case where a closed-form description of the optimal value $h(ww^\top)$ is known, $h(U)$ is linear in U, and $h(X)$ is not linear in X. Because $h(X)$ and $f^X(t)$ are both nonlinear in (X, W) and (t, W), respectively, we use the dual objective formulation to reformulate the identity $h(U) = h(X)$ because the dual objective function and the constraints are linear in (z, W). We obtain a reformulation by replacing $h(X)$ with $g(z)$ and adding the conditions in the feasible set $\Omega(X)$ into the formulation.

We next consider the case where either a closed form of $h(ww^\top)$ is unknown or $h(U)$ is nonlinear in U. Then, we construct the relaxation by replacing both $h(U)$ and $h(X)$ with linear functions $g(z^U)$ and $g(z^X)$ with distinct variables $z^U = (\alpha^U, \beta^U, \gamma^U, \delta^U, \theta^U)$ and $z^X = (\alpha^X, \beta^X, \gamma^X, \delta^X, \theta^X)$ and add the conditions in $\Omega(U)$ and $\Omega(X)$. Then, we tighten the relaxation by exploring the permutation relationship between u and x and the rank-one

conditions. Assume that $x = Pu$ and that $z^U = (\alpha^U, \beta^U, \gamma^U, \delta^U, \theta^U)$ is an optimal solution to the dual with $W = uu^\top$. Then, $(P\alpha^U, P\beta^U, \gamma^U, P\delta^U P^\top, P\theta^U P^\top)$ is an optimal solution to the dual with $W = xx^\top$. Therefore, we can tighten the relaxation by considering conditions $\alpha^X = P\alpha^U, \beta^X = P\beta^X, \gamma^U = \gamma^X$, and $\delta^X = P\delta^U P^\top$ for some $P \in \mathcal{P}_n$. For example, we can add the baseline linear conditions $\mathbb{1}^\top\alpha^U = \mathbb{1}^\top\alpha^X, \mathbb{1}^\top\beta^U = \mathbb{1}^\top\beta^X, \gamma^U = \gamma^X$, $\operatorname{trace}(\delta^U) = \operatorname{trace}(\delta^X)$, $\operatorname{trace}(\theta^U) = \operatorname{trace}(\theta^X)$, $\mathbb{1}^\top\delta^U\mathbb{1} = \mathbb{1}^\top\delta^X\mathbb{1}$, and $\mathbb{1}^\top\theta^U\mathbb{1} = \mathbb{1}^\top\theta^X\mathbb{1}$. If $\alpha^U \in \Delta$ (respectively, $\beta^U \in \Delta$) then we can add the linear reformulation for $\alpha^U \geq_m \alpha^X$ (respectively, $\beta^U \geq_m \beta^X$). Furthermore, we can take advantage of a "good" feasible solution of the dual. Let $z_0^U = (\alpha_0^U, \beta_0^U, \gamma_0^U, \delta_0^U)$ be a feasible solution to the dual with $W = uu^\top$. Then, we can replace the left-hand side with $g(z_0^U)$ and the equality with the inequality $\geq$. We may add inequalities that capture the permutation relationships. Obviously, we can add $\gamma_0^U = \gamma^X$. In addition, we can impose the linear reformulations of $\alpha_0^U \geq_m \alpha^X$ and $\beta_0^U \geq_m \beta^X$ because α_0^U and β_0^U are constants. In addition, any linear inequalities implied by the relationship $\delta_0^U = P^\top\delta^X P$ for some $P \in \mathcal{P}_n$, such as $\operatorname{trace}(\delta_0^U) = \operatorname{trace}(\delta^X)$ and $\mathbb{1}^\top\delta_0^U\mathbb{1} = \mathbb{1}^\top\delta^X\mathbb{1}$, can be considered. Similar relations can also be introduced for θ. More generally, arbitrary linear functions that are permutation invariant in α, β, γ, δ, and θ can be considered instead of the specific one in the objective of the dual.

We next present some special but important cases where the closed form of $h(W)$ is known.

Lemma 9. *When $W = ww^\top$ for $w \geq 0$ and $p = q = 1$, the optimal value of* (32) *is $\sum_{i=1}^r w_{[i]}^2$.*

Proof. Without loss of generality, we assume that $w \in \Delta$ and prove that the optimal value is $\sum_{i=1}^r w_i^2$. Define t' as $t'_{ij} = 1$ if $i = j \leq r$ and zero otherwise. Then, t' is feasible for the primal. Its objective value is $\sum_{i=1}^r w_i^2$. We next define $z' := (\alpha', \beta', \gamma', \delta') \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \times \mathcal{M}^n$ as $\alpha'_i = \beta'_i = \max\left\{\frac{w_i^2 - w_r^2}{2}, 0\right\}$ for $i = 1, \ldots, n$, $\gamma' = w_r^2$, and $\delta'_{ij} = 0$ for $i, j \in \{1, \ldots, n\}$ and prove that z' is feasible for the dual. The nonnegativity of z' is clear. We next show that z' satisfies (33a). First, suppose $i \leq r$ and $j \leq r$. Then, $\alpha'_i + \beta'_j + \gamma' + \delta'_{ij} = \frac{w_i^2 + w_j^2}{2} \geq w_i w_j$. Next, consider the case where $i \leq r$ and $j > r$. Then, $\alpha'_i + \beta'_j + \gamma' + \delta'_{ij} = \frac{w_i^2 + w_r^2}{2} \geq w_i w_r \geq w_i w_j$, where the last inequality holds because $w \in \Delta$. The case where $i > r$ and $j \leq r$ is symmetrical, and the case where $i > r$ and $j > r$ is clear. Because t' and z' satisfy complementarity-slackness conditions, they are optimal solutions to the primal and the dual, respectively. Their common objective value is $\sum_{i=1}^r w_i^2$. □

By Lemma 9, when $p = q = 1$, $h(W)$ is the sum of r largest diagonal entries of W. Whereas $h(W)$ is nonlinear, $h(U)$ is linear because it is the sum of the first r diagonal entries. On the other hand, the inequalities $h(U) \geq h(X)$ for $r \in \{1, \ldots, n-1\}$ are equivalent to the inequality parts of the majorization $\operatorname{diag}(U) \geq_m \operatorname{diag}(X)$. The equality part of the majorization is equivalent to the existing constraint (30b). Therefore, $\operatorname{diag}(U) \geq_m \operatorname{diag}(X)$ is a special case of the aforementioned modeling technique.

Lemma 10. *When $W = ww^\top$ for $w \geq 0$ and $r = pq$, the optimal value of* (32) *is $\sum_{i=1}^p \sum_{j=1}^q w_{[i]} w_{[j]}$.*

Proof. Without loss of generality, we assume that $w \in \Delta_+$. First, define t' as $t'_{ij} = 1$ if $i \leq p$ and $j \leq q$ and zero otherwise. It is clear that t' is feasible and that its objective function value is $(\sum_{i=1}^p w_i)(\sum_{j=1}^q w_j)$. Next, we consider its dual (33) and define $z' := (\alpha', \beta', \gamma', \delta') \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \times \mathcal{M}^n$ as follows:

$$\begin{aligned}
&\alpha'_i = \max\{(w_i - w_p)w_q, 0\}, && i = 1, \ldots, n,\\
&\beta'_j = \max\{w_p(w_j - w_q), 0\}, && j = 1, \ldots, n,\\
&\gamma' = w_p w_q,\\
&\delta'_{ij} = \begin{cases} (w_i - w_p)(w_j - w_q) & \text{if } i \leq p \text{ and } j \leq q,\\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

We first show that z' is feasible for the dual. The nonnegativity of the variables is clear from their definition. To prove that they satisfy (33a), we first consider the case where $i \leq p$ and $j \leq q$. Then, we compute that $\alpha'_i + \beta'_j + \gamma' + \delta'_{ij} = w_i w_j$, showing the result. Next, consider the case where $i > p$ or $j > q$. Without loss of generality, we assume that $i > p$. When $j > q$, it holds that $\alpha'_i + \beta'_j + \gamma' + \delta'_{ij} = \gamma' = w_p w_q \geq w_i w_j$, where the inequality holds because $i > p$, $j > q$, and $w \in \Delta_+$. When $j \leq q$, it holds that $\alpha'_i + \beta'_j + \gamma' + \delta'_{ij} = \beta'_j + \gamma' = w_p w_j \geq w_i w_j$, where the inequality holds because $i > p$ and $w \in \Delta_+$. Because t' and z' satisfy complementarity-slackness conditions, they are optimal solutions to the primal and dual, respectively. Their common objective value is $\sum_{i=1}^p \sum_{j=1}^q w_i w_j$. □

Lemma 10 and (32) can be trivially extended to the case where $W = \bar{w}w^\top$, where $\bar{w} \in \mathbb{R}^m$, $w \in \mathbb{R}^n$, $\bar{w}, w \geq 0$, and $r = pq$, in which case the optimal value is $\sum_{i=1}^{p}\sum_{j=1}^{q} \bar{w}_{[i]} w_{[j]}$. However, this is not of direct relevance for our sparse PCA relaxations. Whereas $h(W)$ is nonlinear because the order of w is unknown, $h(U)$ is linear because it is the sum of the entries of the p-by-q upper-left submatrix of U. In particular, for $q = n$ and $p \in \{1, \ldots, n\}$, the inequalities $h(U) \geq h(X)$ are equivalent to the inequality parts of $R^U \geq_m R^X$, where R^U and R^X are the vectors of the row sums of U and X, respectively. The equality part of the majorization is equivalent to (30c). Therefore, $R^U \geq_m R^X$ is a special case of the aforementioned modeling technique.

In the remainder of the section, we introduce various strengthened semidefinite programming relaxations for sparse PCA.

5.1. An SDP Relaxation for Sparse PCA

Principal component analysis is a well-known dimension reduction technique in statistical analysis. A principal component is a linear combination of independent variables. It also typically stands for the coefficient vector of the linear combination. The first principal component is a unit principal component for which variance is maximized; it is the eigenvector corresponding to the largest eigenvalue of the covariance matrix. Even though the first principal component explains the most variance of the data, it is often hard to interpret because most of its coefficients are nonzero. Sparse PCA is a variant of the approach introduced to resolve this issue by finding linear combinations with few explanatory variables.

Formally, let $\Sigma \in \mathcal{S}^n$ be the covariance matrix of the data set. The following optimization problem, which we refer to as sparse PCA and as defined in Section 1, where $x \in \mathbb{R}^n$ is the coefficient vector of the principal component, finds a unit sparse vector with at most K nonzero entries that explains most of the variance of the data:

$$\begin{aligned} \max \quad & x^\top \Sigma x \\ \text{s.t.} \quad & \|x\| \leq 1, \\ & \operatorname{card}(x) \leq K, \end{aligned} \tag{Sparse PCA}$$

where K is a positive integer satisfying $1 < K < n$.

We studied the feasible set of sparse PCA in Section 3, where we denoted it by $N^K_{\|\cdot\|}$, assuming $\|\cdot\|$ is the ℓ_2-norm. The feasible region of sparse PCA is nonconvex because of the sparsity constraint. We established in Section 3 that

$$\operatorname{conv}(N^K_{\|\cdot\|}) = \left\{ x \;\middle|\; \begin{array}{l} \|u\| \leq 1, \\ u_1 \geq \cdots \geq u_K \geq 0, \\ u_{K+1} = \cdots = u_n = 0, \\ u \geq_{wm} |x| \end{array} \right\}. \tag{34}$$

Because sparse PCA maximizes a convex function, we can replace the feasible set with its convex hull, thereby obtaining a new problem formulation. This formulation, however, remains difficult to solve as it is a convex maximization problem. Next, we introduce new positive semidefinite relaxations for sparse PCA. The most commonly used (and, to the best of our knowledge, only) SDP relaxation for sparse PCA was introduced in d'Aspremont et al. [7] as follows:

$$\begin{aligned} \max \quad & \operatorname{trace}(\Sigma X) \\ \text{s.t.} \quad & \operatorname{trace}(X) \leq 1, \\ & \mathbb{1}^\top |X| \mathbb{1} \leq K, \\ & X \succeq 0. \end{aligned} \tag{35}$$

We refer to this as the *D-relaxation*.

We next present a strengthened SDP relaxation based on the convex hull description (34). First, we introduce the matrix variable X to model the relationship $X = xx^\top$. Then, we introduce variables y and Y to represent $|x|$ and $|X|$, respectively. Furthermore, we add the auxiliary variables v, w, V, and W to model the absolute values. The variables and the constraints are

$$\begin{cases} x = v - w, & y = v + w \\ v, w \in \mathbb{R}^n_+, & x, y \in \mathbb{R}^n \end{cases} \tag{36}$$

and

$$\begin{cases} X = V - W, & Y = V + W \\ V, W \in \mathcal{M}^n_{\geq 0}, & X, Y \in \mathcal{S}^n \end{cases} \tag{37}$$

where $\mathcal{M}^n_{\geq 0}$ is the set of n-by-n (entry-wise) nonnegative matrices, and $\mathcal{S}^n$ is the set of n-by-n symmetric matrices. Next, we introduce the vector u majorizing $y(=|x|)$ and the matrix U to model $uu^\top$. The constraint $u \in \Delta_+$ and the constraint (31) in the cardinality setting can be written as

$$\begin{cases} u_1 \geq \cdots \geq u_K & \\ u_i = 0, & i \geq K+1, \\ U_{i,1} \geq \cdots \geq U_{i,K}, & i = 1, \ldots, K, \\ U_{i,j} = 0, & i \geq K+1 \text{ or } j \geq K+1, \\ u \in \mathbb{R}^n_+, U \in \mathcal{S}^n_{\geq 0}, & \end{cases} \tag{38}$$

where $\mathcal{S}^n_{\geq 0}$ is the set of n-by-n (entry-wise) nonnegative symmetric matrices. In the following construction, we use the relationship between Y and U that $Y = PUP^\top$ for some $P \in \mathcal{P}_n$. (That is, the entries of Y and U equal up to row permutations and the corresponding column permutations.) We impose the constraints (30) as follows:

$$\begin{cases} \operatorname{trace}(U) \leq 1, & \text{(39a)} \\ \mathbb{1}^\top U \mathbb{1} = \mathbb{1}^\top Y \mathbb{1}. & \text{(39b)} \end{cases}$$

The nonconvex relationships $X = xx^\top$, $Y = yy^\top$, and $U = uu^\top$ can be relaxed using Schur complements as

$$\begin{bmatrix} X & x \\ \bar{x}^\top & 1 \end{bmatrix} \succeq 0, \quad \begin{bmatrix} Y & y \\ \bar{y}^\top & 1 \end{bmatrix} \succeq 0, \quad \begin{bmatrix} U & u \\ u^\top & 1 \end{bmatrix} \succeq 0. \tag{40}$$

By constraint $X \succeq xx^\top$, it holds that $X \succeq 0$; hence, $\operatorname{diag}(X) \geq 0$. Therefore, the constraint $\operatorname{trace}(U) = \operatorname{trace}(Y)$ can be replaced with

$$\operatorname{trace}(U) = \operatorname{trace}(X). \tag{41}$$

We next present modeling details for the majorization constraints using the arguments presented earlier. First, the majorization relationship $u \geq_m y$ is represented as

$$\begin{cases} \sum_{i=1}^j u_i \geq j r_j + \sum_{j=1}^n t_{ij}, & j = 1, \ldots, n-1, \\ y_i \leq t_{ij} + r_j, & i = 1, \ldots, n,\ j = 1, \ldots, n-1 \\ r \in \mathbb{R}^{n-1},\ t \in \mathbb{R}^{n \times (n-1)}_+ & \end{cases} \tag{42}$$

as mentioned in (4). Even though the modeling techniques using (32) and (33) can potentially derive several inequalities, we only present certain representative inequalities in the proposed SDP relaxation for the sake of exposition. First, the row sum majorization $R^U \geq_m R^Y$ and the diagonal majorization $\operatorname{diag}(U) \geq_m \operatorname{diag}(Y)$ are represented as

$$\begin{cases} R^U_i = \sum_{j=1}^n U_{ij},\ R^Y_i = \sum_{j=1}^n Y_{ij} & i = 1, \ldots, n, \\ \sum_{i=1}^j R^U_i \geq j r^R_j + \sum_{j=1}^n t^R_{ij}, & j = 1, \ldots, n-1, \\ R^Y_i \leq t^R_{ij} + r^R_j, & i = 1, \ldots, n,\ j = 1, \ldots, n-1 \\ r^R \in \mathbb{R}^{n-1},\ t^R \in \mathbb{R}^{n \times (n-1)}_+,\ R^U, R^Y \in \mathbb{R}^n & \end{cases} \tag{43}$$

and

$$\begin{cases} \sum_{i=1}^j U_{ii} \geq j r^D_j + \sum_{j=1}^n t^D_{ij}, & j = 1, \ldots, n-1, \\ X_{ii} \leq t^D_{ij} + r^D_j, & i = 1, \ldots, n,\ j = 1, \ldots, n-1 \\ r^D \in \mathbb{R}^{n-1},\ t^D \in \mathbb{R}^{n \times (n-1)}_+, & \end{cases} \tag{44}$$

respectively. Last, we use (33) as follows to underestimate $\mathbb{1}^\top (U^{pq}) \mathbb{1}$ for $p \in \{1, \ldots, n\}$ and $q \in \{1, \ldots, n\}$, where for a matrix $A \in \mathbb{R}^{m \times n}$ and indices $p \leq m$ and $q \leq n$, A^{pq} is defined as the p-by-q submatrix whose sum of entries is

maximum:

$$\begin{cases} \sum_{i=1}^{p}\sum_{j=1}^{q} U_{ij} \geq \bar{q}\sum_{i=1}^{n}\alpha_i^{pq} + \bar{p}\sum_{j=1}^{q}\beta_j^{pq} + \bar{p}\bar{q}\gamma^{pq} + \sum_{i=1}^{n}\sum_{j=1}^{n}\delta_{ij}^{pq}, & p\in\{1,\dots,K\}, q\in\{p,\dots,K\}, \\ \alpha_i^{pq} + \beta_j^{pq} + \gamma^{pq} + \delta_{ij}^{pq} \geq Y_{ij}, & p\in\{1,\dots,K\}, q\in\{p,\dots,K\}, i,j\in\{1,\dots,n\}, \\ \alpha\in\mathbb{R}_+^{\frac{K(K-1)}{2}\times n}, \beta\in\mathbb{R}_+^{\frac{K(K-1)}{2}\times n}, \gamma\in\mathbb{R}_+^{\frac{K(K-1)}{2}}, \delta\in\mathbb{R}_+^{\frac{K(K-1)}{2}\times n\times n}, \end{cases} \tag{45}$$

with $\bar{p} = \begin{cases} p & \text{if } p \leq K-1 \\ n & \text{if } p = K \end{cases}$ and $\bar{q} = \begin{cases} q & \text{if } q \leq K-1 \\ n & \text{if } q = K. \end{cases}$

The proposed SDP relaxation, which we refer to as the *submatrix relaxation*, is

$$\begin{aligned} \max \quad & \operatorname{trace}(\Sigma X) \\ \text{s.t.} \quad & (36)-(44), (45). \end{aligned} \tag{46}$$

Theorem 12. *All constraints in* (35) *are implied by* (37), (39a), (39b), (40), *and* (41).

Proof. First, $\operatorname{trace}(X) = \operatorname{trace}(U) \leq 1$, where the equality and the inequality directly follow from (41) and (39a), respectively. We next show that $\mathbb{1}^\top|X|\mathbb{1} \leq K$ is implied. Let U_K be the upper-left K-by-K submatrix of U, and let $\mathbb{1}_K$ be the K-dimensional vector of ones. Define $f:\mathbb{R}^K \to \mathbb{R}$ as $f(x) := x^\top U_{K,K} x$. Because $U \succeq 0$, we have $U_{K,K} \succeq 0$, showing that f is convex. Furthermore, $f(\alpha x) = \alpha^2 f(x)$ for any scalar α. Therefore,

$$\begin{aligned} \mathbb{1}^\top U \mathbb{1} &= \mathbb{1}_K^\top U_{K,K}\mathbb{1}_K = f(\mathbb{1}_K) = f\left(\sum_{i=1}^{K} e_i\right) = f\left(K\sum_{i=1}^{K}\frac{1}{K}e_i\right) \\ &= K^2 f\left(\sum_{i=1}^{K}\frac{1}{K}e_i\right) \leq K^2\frac{1}{K}\sum_{i=1}^{K} f(e_i) = K\operatorname{trace}(U) \leq K, \end{aligned} \tag{47}$$

where the inequalities follows from the convexity of f and (39a). Therefore,

$$\mathbb{1}^\top|X|\mathbb{1} = \mathbb{1}^\top|V-W|\mathbb{1} \leq \mathbb{1}^\top(V+W)\mathbb{1} = \mathbb{1}^\top Y\mathbb{1} = \mathbb{1}^\top U\mathbb{1} \leq K, \tag{48}$$

where the first two equalities and the first inequality follow from (37), the third equality from (39b), and the last inequality from (47). The positive semidefiniteness of X follows from the first Schur complement condition in (40). □

Remark 3. We present another relaxation of (45) that improves computational efficiency compared with (45), although this approach provides a weaker bound. First, we introduce variables $(SR)_{iq}$ to define the sum of q-largest components of the ith row of Y as follows:

$$\begin{cases} (SR)_{iq} \geq q r_{qi}^b + \sum_{j=1}^{n} t_{qij}^b & i=1,\dots,n, q=1,\dots,K, \\ Y_{ij} \leq t_{qij}^b + r_{qi}^b & i=1,\dots,n, j=1,\dots,n, q=1,\dots,K, \\ t^b \geq 0, & \\ t^b \in \mathbb{R}^K\times\mathbb{R}^n\times\mathbb{R}^n, r^b\in\mathbb{R}^K\times\mathbb{R}^n. & \end{cases}$$

Using a specific feasible solution of (32), we relax (45) as follows:

$$\begin{cases} \sum_{i=1}^{p}\sum_{j=1}^{q} U_{ij} \geq p r_{pq}^B + \sum_{i=1}^{n} t_{pqi}^B & p=1,\dots,K, q=1,\dots,K, \\ (SR)_{iq} \leq t_{pqi}^B + r_{pq}^B & p=1,\dots,K, q=1,\dots,K, i=1,\dots,n, \\ t^B \geq 0, & \\ t^B \in \mathbb{R}^K\times\mathbb{R}^K\times\mathbb{R}^n, r^B\in\mathbb{R}^K\times\mathbb{R}^K. & \end{cases}$$

We refer to this relaxation as the *two-step relaxation*.

5.2. Computational Experiments for Sparse PCA

We next report our computational results on sparse PCA. First, we summarize the following standard result.

Proposition 12. *The following is a correct formulation for sparse* PCA:

$$\begin{aligned} \max \quad & \operatorname{trace}(\Sigma X) && \text{(49a)}\\ \text{s.t.} \quad & \operatorname{trace}(X) \le 1, && \text{(49b)}\\ & X \succeq 0, && \text{(49c)}\\ & \operatorname{diag}(X) \le z, && \text{(49d)}\\ & \mathbb{1}^\top z = K, && \text{(49e)}\\ & z \in \{0,1\}^n. && \text{(49f)} \end{aligned}$$

Let (X^*, z^*) be an optimal solution to (49) and $\sum_{i=1}^n \lambda_i x^i x^{i\top}$ be an eigenvalue decomposition of X^* so that each x^i is a unit eigenvector. Then, for any i' such that $\lambda_{i'} > 0$, $x^{i'}$ is an optimal solution to sparse PCA.

Theorem 13. *The following constraints are valid for sparse PCA:*

$$\begin{aligned} & Y_{ij}^2 \le T_{ij} T_{ji} \quad i = 1,\dots,n, j = i+1,\dots,n, && \text{(50a)}\\ & T_{ii} = Y_{ii} \qquad i = 1,\dots,n, && \text{(50b)}\\ & \sum_{j=1}^n T_{ij} = z_i \qquad i = 1,\dots,n, && \text{(50c)}\\ & \sum_{i=1}^n T_{ij} = K Y_{jj} \quad j = 1,\dots,n, && \text{(50d)}\\ & 0 \le T_{ij} \le Y_{jj} \quad i = 1,\dots,n, j = 1,\dots,n. && \text{(50e)} \end{aligned}$$

In particular, T_{ij} may be regarded as a linearization of $z_i y_j^2$. Using that $Y_{ij} = Y_{ji}$ for all (i, j), the above constraints imply

1. *for all (i, j), $T_{ij} + T_{ji} \ge 2Y_{ij}$;*
2. *for all i, $Y_{ii} \le z_i$ and, for all (i, j) with $i \ne j$, $2Y_{ij} \le z_i$;*
3. *for all $S \subseteq \{1,\dots,n\}$, $\sum_{i,j\in S} Y_{ij} \le \sum_{i\in S} z_i$;*
4. *for all i, $\sum_{j=1}^n Y_{ij}^2 \le z_i Y_{ii}$;*
5. *for all i, $\sum_{j=1}^n Y_{ij}^2 \le T_{ii}$ as long as, for all j', $Y_{ij'}^2 \le Y_{ii} Y_{j'j'}$ or, more specifically, $Y \succeq 0$;*
6. *$\mathbb{1}^\top |X| \mathbb{1} \le K$, as long as (37) holds;*
7. *$\sum_{j=1}^n X_{ij}^2 \le z_i X_{ii}$ as long as (37), and $\operatorname{trace}(Y) = \operatorname{trace}(X)$ hold;*
8. *$\sum_{j=1}^n X_{ij}^2 \le T_{ii}$ as long as (37) holds, $\operatorname{trace}(Y) = \operatorname{trace}(X)$, and $X \succeq 0$.*

Proof. To show that the constraints are valid, we only need to show that T_{ij} can be chosen to be $z_i y_j^2$. We may assume that for all (i, j), $Y_{ij} = y_i y_j$. Constraint (50a) follows because for $i' \in \{i, j\}$, $y_{i'} = z_{i'} y_{i'}$ implies that $(y_i y_j)^2 \le z_i y_i^2 z_j y_j^2$. Constraint (50b) follows because $z_i Y_{ii} = Y_{ii}$. Constraint (50c) follows because $\operatorname{trace}(Y) = 1$ implies $\sum_{j=1}^n z_i Y_{jj} = z_i$. Constraint (50d) is valid because (49e) implies that $\sum_{i=1}^n z_i Y_{jj} = K Y_{jj}$. Finally, (50e) follows because it relaxes $y_i^2 = z_i y_i^2 = Y_{ii}$.

We now show the implications claimed. First, we show statement 1. To see this, observe that we can write (50a) as $(2Y_{ij})^2 + (T_{ij} - T_{ji})^2 \le (T_{ij} + T_{ji})^2$. Then,

$$2Y_{ij} \le \sqrt{(2Y_{ij})^2 + (T_{ij} - T_{ji})^2} \le \sqrt{T_{ij} + T_{ji}} = T_{ij} + T_{ji}, \tag{51}$$

where the first inequality is because $(T_{ij} - T_{ji})^2 \ge 0$, the second inequality is because of (50a), and the equality is because, by (50e), T_{ij} and T_{ji} are nonnegative. Second, we show statement 2. Clearly, $Y_{ii} = T_{ii} \le z_i$, where the first equality is by (50b) and the inequality follows from (50c). If $i \ne j$, we have

$$2Y_{ij} \le T_{ij} + T_{ji} \le T_{ij} + Y_{ii} = T_{ij} + T_{ii} \le z_i, \tag{52}$$

where the first inequality is by statement 1, the second inequality is because of (50e), the equality is because of (50b), and the inequality is because $i \ne j$ and (50c). Third, we show statement 3. This is because

$$0 \le \sum_{i,j\in S} (T_{ij} - Y_{ij}) \le \sum_{i\in S} \sum_{j=1}^n T_{ij} - \sum_{i,j\in S} Y_{ij} = \sum_{i\in S} z_i - \sum_{i,j\in S} Y_{ij}, \tag{53}$$

where the first inequality is because of statement 1 and $Y_{ij} = Y_{ji}$, the second inequality is because of (50e), and the equality is by (50c). Now, we show statement 4 as follows:

$$\sum_{j=1}^{n} Y_{ij}^2 \le \sum_{j=1}^{n} T_{ji}T_{ij} \le Y_{ii}\sum_{j=1}^{n} T_{ij} = z_i Y_{ii}, \tag{54}$$

where the first inequality follows from (50a), the second inequality from (50e), and the equality from (50c). To see statement 5, observe that

$$\sum_{j=1}^{n} Y_{ij}^2 \le Y_{ii}\sum_{j=1}^{n} Y_{jj} = Y_{ii} = T_{ii}, \tag{55}$$

where the first inequality is because, for all j', we have $Y_{ij'}^2 \le Y_{ii}Y_{j'j'}$, the first equality is because $\operatorname{trace}(Y) = 1$, and the third equality is by (50b). Next, we show statement 6. We write

$$\mathbb{1}^\top |X| \mathbb{1} = \mathbb{1}^\top |V - W| \mathbb{1} \le \mathbb{1}^\top (|V| + |W|)\mathbb{1} = \mathbb{1}^\top Y \mathbb{1} \le \mathbb{1}^\top z = K,$$

where the second equality is because $Y = V + W$, $V \ge 0$, and $W \ge 0$; the first inequality is because of (53) with $S = \{1, \dots, n\}$; and the last equality is because of (49e). To show statement 7, we write

$$\sum_{j=1}^{n} X_{ij}^2 = \sum_{j=1}^{n} (V_{ij} - W_{ij})^2 \le \sum_{j=1}^{n} (V_{ij} + W_{ij})^2 = \sum_{j=1}^{n} Y_{ij}^2 \le z_i Y_{ii} = z_i X_{ii}, \tag{56}$$

where the first two equalities and first inequality are by (37), and the second equality is because of statement 4. The last equality is because (37) implies $\operatorname{diag}(Y - X) \ge 0$. Together with $\operatorname{trace}(Y - X) = 0$, this implies that for all i, $Y_{ii} = X_{ii}$. Finally, the proof of statement 8 is similar to that for statement 5, where Y is replaced with X, where we also utilize that $\operatorname{trace}(Y - X) = 0$, $\operatorname{diag}(Y - X) \ge 0$, and (50b) imply that $X_{ii} = T_{ii}$ for all i. □

We use the following formulation, which we refer to as the *T-formulation*, to find the optimal solution for sparse PCA. We chose to develop this formulation around the diagonal relaxation obtained using constraints (44) as it is simple to implement but, as we will show later, it is also strong:

$$\begin{aligned}
(T): \max \quad & \operatorname{trace}(\Sigma X) \\
\text{s.t.} \quad & (37), (38), (39), (44), (49e), (49f), (50), \\
& \operatorname{trace}(U) = \operatorname{trace}(Y) = \operatorname{trace}(X) = 1, \\
& Y_{ij}^2 \le Y_{ii}Y_{jj} \qquad i = 1, \dots, n, j = i+1, \dots, n. \\
& X \succeq 0, U \succeq 0.
\end{aligned}$$

To prove that this formulation is correct, we need to show only that (49d) is satisfied. To see this, observe that $X_{ii} \le Y_{ii} \le z_i$, where the first inequality follows from (37) and the second inequality from (52). Instead of requiring that $Y \succeq 0$, we relax it so that $Y_{ij}^2 \le Y_{ii}Y_{jj}$ for all (i, j). Notice that the constraints $Y_{ij}^2 \le T_{ij}T_{ji}$ and $Y_{ij}^2 \le Y_{ii}Y_{jj}$ can be written as second-order cone programming (SOCP) constraints $\|(2Y_{ij}, T_{ji} - T_{ij})\|_2 \le T_{ji} + T_{ij}$ and $\|(2Y_{ij}, Y_{ii} - Y_{jj})\|_2 \le Y_{ii} + Y_{jj}$, respectively. We refer to the relaxation obtained by dropping (49f) as the *T-relaxation*.

After the initial draft of this paper (Kim et al. [17]), the following relaxation for sparse PCA was proposed in Bertsimas et al. [6]:

$$\begin{aligned}
\max \quad & \operatorname{trace}(\Sigma X) \\
\text{s.t.} \quad & (37), (49e), (49f), \\
& \operatorname{trace}(X) = 1, \\
& Y_{ii} \le z_i \qquad i = 1, \dots, n, && \text{(57a)} \\
& 2Y_{ij} \le z_i \qquad i = 1, \dots, n,\ j = i+1, \dots, n, && \text{(57b)} \\
& \sum_{j=1}^{n} X_{ij}^2 \le z_i X_{ii} \quad i = 1, \dots, n, && \text{(57c)} \\
& \mathbb{1}^\top Y \mathbb{1} = K, \\
& X \succeq 0.
\end{aligned}$$

Table 3. Optimal values and gaps closed for the test problem `pitprops`.

	T-formulation	D-relaxation	B-relaxation			Rowsum relaxation			Diagonal relaxation			Two-step relaxation			Submatrix relaxation			T-relaxation		
K	z^*	z_D^*	z_B^*	Time (sec)	Gap closed	z_{rowsum}^*	Time (sec)	Gap closed	z_{diag}^*	Time (sec)	Gap closed	z_{2step}^*	Time (sec)	Gap closed	z_{submat}^*	Time (sec)	Gap closed	z_T^*	Time (sec)	Gap closed
3	2.4753	2.5218	2.4987	0.20	49.60	2.5033	0.26	39.78	2.4949	0.24	57.86	2.4753	0.68	100.00	2.4753	0.80	100.00	2.4753	0.22	100.00
4	2.9375	3.0172	2.9917	0.22	31.98	2.9766	0.27	50.89	2.9671	0.25	62.83	2.9477	0.82	87.15	2.9477	1.85	87.15	2.9375	0.22	100.00
5	3.4062	3.4581	3.4303	0.22	53.60	3.4103	0.27	91.92	3.4072	0.26	97.96	3.4062	1.23	100.00	3.4062	4.26	100.00	3.4062	0.23	100.00
6	3.7710	3.8137	3.7886	0.22	58.80	3.7710	0.26	100.00	3.7710	0.27	100.00	3.7710	1.39	100.00	3.7710	8.81	100.00	3.7710	0.22	100.00
7	3.9962	4.0316	3.9967	0.22	98.69	3.9962	0.24	100.00	3.9962	0.26	100.00	3.9962	1.52	100.00	3.9962	15.94	100.00	3.9962	0.24	100.00
8	4.0686	4.1448	4.0839	0.22	79.93	4.0770	0.24	88.99	4.0721	0.29	95.48	4.0721	2.53	95.48	4.0721	36.79	95.47	4.0686	0.23	99.99
9	4.1386	4.2063	4.1398	0.20	98.28	4.1399	0.26	98.20	4.1386	0.28	100.00	4.1386	3.32	100.00	4.1386	70.24	100.00	4.1386	0.23	100.00
10	4.1726	4.2186	4.1778	0.23	88.87	4.1807	0.26	82.42	4.1766	0.26	91.32	4.1766	3.75	91.32	4.1766	117.66	91.32	4.1733	0.23	98.49
Average				0.21	69.97		0.26	81.53		0.26	88.18		1.91	96.74		32.04	96.74		0.23	99.81

We refer to the relaxation as the *B-relaxation*. Theorem 13 shows that the *B*-relaxation is dominated by the *T*-relaxation because (57a)–(57c) are implied in the *T*-relaxation.

We remark that the relaxations we develop here do not intrinsically depend on $\text{trace}(X) = 1$. In particular, when the constraint (50c) is replaced with $\sum_{j=1}^{n} T_{ij} \le z_i \text{trace}(T)$ and the constraint $\text{trace}(U) = \text{trace}(Y) = \text{trace}(X) = 1$ is replaced with $\text{trace}(U) = \text{trace}(Y) = \text{trace}(X)$, our relaxations can be used more generally. More specifically, they are valid whenever we seek a rank-one matrix X that is symmetric and positive definite, and is such that only K rows and columns have nonzero values.

We next define the notations used in the computational experiments. We refer to the relaxation obtained by dropping constraints (44) and (45) from (46) as the *rowsum relaxation*. Also, we refer to the relaxation obtained by dropping constraints (43) and (45) from (46) as the *diagonal relaxation*. We denote the optimal values of the *D*-relaxation, *B*-relaxation, rowsum relaxation, diagonal relaxation, two-step relaxation, submatrix relaxation, and *T*-relaxation by $z^*_D, z^*_B, z^*_{rowsum}, z^*_{diag}, z^*_{2step}, z^*_{submat}$, and z^*_T, respectively.

We use CVX (Grant and Boyd [10, 11]) version 2.2 or YALMIP (Löfberg [22]) version R20210331 to solve SDPs in the experiments. MOSEK (MOSEK ApS [26]) version 9.2.47 was selected as the SDP solver in both cases. To measure the relative tightness of a relaxation when compared with (35), we calculate *gap closed* as

$$\left(\frac{z^*_D - z^*_{SDP}}{z^*_D - z^*}\right) \times 100.$$

where z^*_{SDP} is the target SDP relaxation on which we calculate the gap closed. Here, z^* denotes the optimal value of sparse PCA and we obtain this value by solving the *T*-formulation of sparse PCA, (*T*), to optimality using the **bnb** solver of YALMIP with MOSEK version 9.1.9.

In addition, we conducted experiments with an alternate formulation. This formulation is based on a compact extended formulation of the permutahedron proposed by Goemans [9], where the construction and the size of the reformulation depend on the choice of a sorting network. Although the optimal Ajtai–Komlós–Szemerédi sorting network (Ajtai et al. [1]) gives an extended formulation for the permutahedron with $\Theta(n \log n)$ variables and inequalities, it has limited practical value because the constant hidden in the $\Theta(\cdot)$ notation is very large. Instead, we used Batcher's bitonic sorting network that gives an extended formulation with $\Theta(n \log^2 n)$ variables and inequalities. When it comes to the time comparison between these two reformulations, we use the diagonal relaxation among the aforementioned SDP relaxations because, as we show later, this relaxation produces competitive bounds and is simple to implement. Because the formulation based on a sorting network is equivalent to the duality-based formulation based on (4), the quality of bounds is the same, and we compare only their computation times.

The experiments were performed with an Intel Core i5-10400 machine containing a 2.90 GHz central processing unit with 32 GB of random access memory and running Windows 10.

5.2.1. `Pitprops` Problem. The pitprops problem (Jeffers [15]) is one of the most commonly used problems for sparse PCA algorithms. The instance has 13 variables and 180 observations. Table 3 shows the test results for cardinality $K = 3, \ldots, 10$.

Observe that the diagonal (respectively, submatrix) relaxation reduces the gaps of (35) by more than 88% (respectively, 96%), returning global optimal solutions for three (respectively, five) problems. The *T*-relaxation attains the optima except for the instance $K = 10$. On average, it reduces the gaps of (35) by 99.81%.

For all computational times reported in Table 3, we used (44) to model the majorization constraints. However, as we discussed earlier, these constraints can also be formulated using Batcher's bitonic sorting network, which requires only $\Theta(n \log^2 n)$ variables. Our primary intention here is to compare the formulation of the permutahedron based on (44) with that obtained via Batcher's bitonic sorting network. For this comparison, we use the diagonal relaxation and simplify it by dropping the requirement that $Y \succeq 0$. We refer to this simplified relaxation as the

Table 4. Computation time (in seconds) comparison with the diagonal relaxation for pitprops problems.

K	3	4	5	6	7	8	9	10	Average
Diagonal'-relaxation	0.23	0.22	0.23	0.22	0.23	0.23	0.23	0.22	0.23
Diagonal'-relaxation-sort	0.20	0.23	0.23	0.22	0.22	0.21	0.22	0.23	0.22

Table 5. Test results for sparse PCA with $n \in \{10, 15, \ldots, 40, 45, 50\}$ and $K = \text{round}(n/6)$.

		B-relaxation		Rowsum relaxation		Diagonal relaxation		Two-step relaxation		Submatrix relaxation		*T*-relaxation	
n	K	Gap closed	Time (sec)	Gap closed	Time (sec)	Gap closed	Time (sec)	Gap closed	Time (sec)	Gap closed	Time (sec)	Gap closed	Time (sec)
10	2	66.18	0.37	81.28	0.34	88.02	0.35	93.65	0.35	93.65	0.34	100.00	0.36
15	3	58.27	0.37	71.98	0.36	81.98	0.36	91.15	0.39	91.15	0.40	99.80	0.37
20	3	58.48	0.36	68.63	0.44	82.83	0.41	90.61	0.53	90.61	0.55	99.90	0.39
25	4	51.28	0.37	68.01	0.63	79.40	0.55	86.43	0.94	86.44	1.13	98.95	0.50
30	5	43.76	0.39	54.83	1.11	72.26	0.85	84.01	1.83	84.01	3.02	97.57	0.75
35	6	44.18	0.47	55.36	2.04	69.19	1.59	81.31	3.42	81.30	6.23	97.98	1.22
40	7	39.74	0.52	53.26	3.58	65.70	2.62	81.09	6.79	81.09	12.63	98.05	1.33
45	8	32.52	0.65	51.06	5.63	65.35	4.44	77.60	11.23	77.60	22.79	94.81	2.09
50	8	25.92	0.94	40.61	9.47	54.94	7.54	70.93	18.62	70.93	34.73	94.29	3.43
Average		46.70	0.49	60.56	2.62	73.30	2.08	84.09	4.90	84.09	9.09	97.93	1.16

Table 6. Comparison between the B-relaxation and the T-relaxation with $n \in \{55, 60, \ldots, 95, 100\}$ and $K = \text{round}(n/6)$.

n	55	60	65	70	75	80	85	90	95	100
K	9	10	11	12	13	13	14	15	16	17
B-relaxation										
Gap closed	32.88	29.28	17.18	19.07	18.32	17.01	11.82	13.56	10.68	11.61
Time (min)	0.02	0.03	0.05	0.07	0.09	0.13	0.17	0.23	0.30	0.40
T-relaxation										
Gap closed	96.19	91.93	87.11	90.97	84.79	87.12	85.44	79.70	79.97	76.65
Time (min)	0.09	0.14	0.21	0.31	0.45	0.66	0.90	1.17	1.60	2.14

Diagonal′-relaxation. To differentiate the relaxation based on a sorting network, we will refer to it as *Diagonal′-relaxation-sort*. Clearly, both relaxations yield the same bound. Therefore, we only report on the solution times when comparing these relaxations. As n and K are small, no significant difference between the reformulations is observed.

5.2.2. Experiments with Randomly Generated Matrices. We next report test results for randomly generated covariance matrices. Random matrices are generated using the following procedure. First, we choose a random integer $m \in \{1, \ldots, n\}$ for the number of nonzero eigenvalues of the matrix by setting $m = \lceil nU \rceil$, where U is randomly drawn from the uniform distribution $\mathcal{U}(0,1)$. Second, we generate m random vectors $v_i \in \mathbb{R}^n \sim \mathcal{N}(0, I_n)$, for $i = 1, \ldots, m$ for rank-1 matrices. Third, we generate m positive random eigenvalues $\lambda_i \sim \mathcal{U}(0,1)$, for $i = 1, \ldots, m$. Finally, we construct the desired random covariance matrix as $\Sigma = \sum_{i=1}^{m} \lambda_i v_i v_i^\top$.

The tests are performed for problems with size $n \in \{5 + 5i \mid i = 1, \ldots, 9\}$, and the cardinality $K = \text{round}(n/6)$ is chosen to reflect the motivation of sparse principal components analysis to produce sparse vectors, where $\text{round}(x)$ represents the integer closest to x (i.e., $\text{round}(x) = \lfloor x \rfloor$ as long as the fractional part of x is strictly less than 0.5 and $\lceil x \rceil$ otherwise). For each n and the associated K, 30 instances generated from random covariance matrices are tested. The reported computation times and gaps closed are the averages for the 30 instances. In Table 5, we present average gap closed for the computation times in seconds and gap closed of the relaxations. The computational results show that our SDP relaxations improve the gaps of the SDP relaxation (35) significantly.

Among the relaxations in Table 5, the T-relaxation yields the tightest bound on our instances. We remark that the T-relaxation bound improves when the constraints for the two-step relaxation and/or the submatrix relaxation are also imposed. However, we do not report on the performance of these relaxations because they are more complex and more computationally expensive to solve. For larger-dimensional instances, we choose the B-relaxation and the T-relaxation to compare. For $n \in \{5 + 5i \mid i = 10, \ldots, 19\}$, we generate 30 random covariance matrices for each n and summarize the results of the comparison in Table 6 and Figure 2, where the computation times are in minutes.

We next show that replacing (44) with a sorting network helps reduce the solution time for the diagonal relaxation by comparing the Diagonal′-relaxation and Diagonal′-relaxation-sort on the synthetic data sets. For this demonstration, we consider dimensions $n \in \{10i \mid i = 1, \ldots, 15\}$ and choose $K = \text{round}(n/6)$. For each choice of n and the associated K, 30 randomly generated instances are tested; see Table 7 and Figure 3 for comparison of the

Figure 2. Comparison between the B-relaxation and the T-relaxation with $n \in \{10, 15, \ldots, 95, 100\}$ and $K = \text{round}(n/6)$.

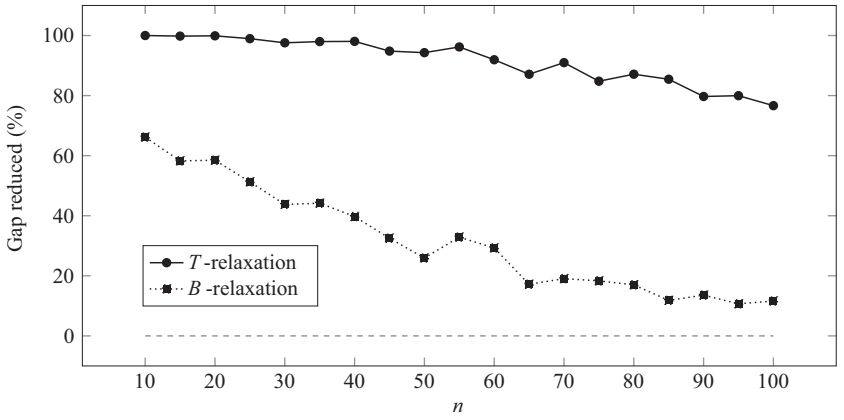
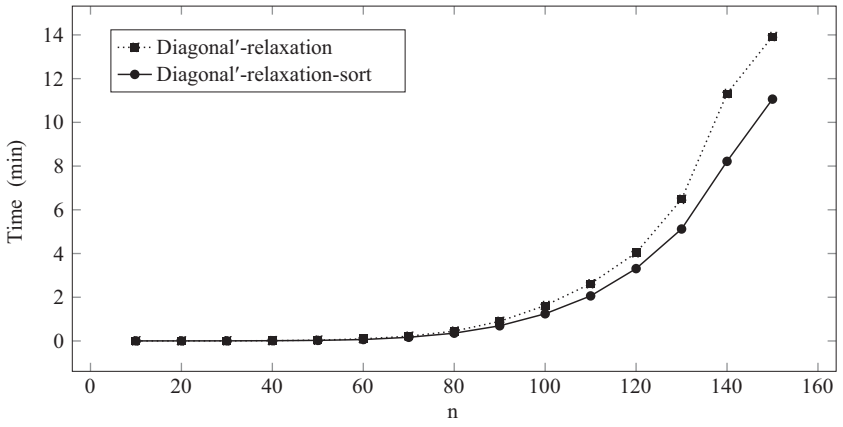


Table 7. Computation time comparison between the Diagonal′-relaxation and Diagonal′-relaxation-sort for sparse PCA with a random covariance matrix, $n = 10, 20, \ldots, 150$ and $K = \text{round}(n/6)$.

	n														
	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150
Diagonal′-relaxation															
# vars	832	3,261	7,289	12,917	20,146	28,974	39,402	51,431	65,059	80,287	97,116	115,544	135,572	157,201	180,429
# cons	384	1,469	3,254	5,739	8,924	12,809	17,394	22,679	28,664	35,349	42,734	50,819	59,604	69,089	79,274
Time (min)	0.00	0.00	0.01	0.01	0.04	0.10	0.22	0.45	0.90	1.62	2.6184	4.30	6.48	11.33	13.93
Diagonal′-relaxation-sort															
# vars	848	3,087	6,125	11,451	16,890	23,528	34,214	43,253	53,491	64,929	77,568	91,406	113,676	129,915	147,353
# cons	411	1,452	2,737	5,226	7,511	10,296	15,437	19,222	23,507	28,292	33,577	39,362	50,383	57,168	64,453
Time (min)	0.00	0.00	0.01	0.01	0.03	0.07	0.17	0.36	0.69	1.24	2.06	3.31	5.12	8.21	11.06

Figure 3. Computation time comparison between the Diagonal′-relaxation and Diagonal′-relaxation-sort for sparse PCA with random covariance matrices; $n = 10, 20, \ldots, 150$; and $K = \text{round}(n/6)$.



computation times (in minutes). Our results show that the sorting-network-based formulation reduces the computational time for the relaxation.

6. Conclusion

In this paper, we present an explicit convex hull description of permutation-invariant sets and applications of the results to various important sets/functions in optimization. The construction of the convex hull is based on the fact that a permutation-invariant set is a union of permutahedra with generating vectors in $\Delta = \{x \in \mathbb{R}^n \,|\, x_1 \geq \cdots \geq x_n\}$ and the convex hull of this union can be described in closed form. We then applied this result to derive various convexification results. First, we presented an extended formulation for the convex hull of permutation-invariant norm balls constrained by a cardinality requirement. Second, we convexified sets of matrices that are characterized using functions of their singular values. Third, we derived convex/concave envelopes of various nonlinear functions and convex hulls of sets defined using nonlinear functions when bounds for variables are congruent. Fourth, we studied sets of rank-one matrices whose generating vectors lie in a permutation-invariant set. We use majorization inequalities in the space of generating vectors to construct valid inequalities for the convex hull in the matrix space. As a motivating example, we construct tight semidefinite programming relaxations for sparse principal component analysis and report computational results that show that our tightest relaxation reduces more than 99.8% (respectively, 97.9%) of the gaps for the pitprops data set (respectively, synthetic data sets with the dimensions up to $n = 50$). The concept of permutation invariance can be used to study a variety of other sets, including those arising from logical requirements in 0-1 mixed integer programming; see Kim et al. [17] for additional descriptions.

Acknowledgments

The authors thank the review team for their suggestions, which helped improve the presentation and clarify the contributions of this paper.

Appendix. Table of Notations

Notation	Definition	Defined in
$\text{card}(x)$	Number of nonzero components of x	Section 1
$\|\cdot\|_1$	l_1-norm of vectors	Section 1
$\mathcal{M}^{m,n}(\mathbb{R})$	Set of m-by-n real matrices	Sections 1 and 3.1
$\|\cdot\|_{sp}$	Spectral norm of matrices	Sections 1 and 3.1
Δ_π	$\{x \in \mathbb{R}^n \,\|\, x_{\pi(1)} \geq \cdots \geq x_{\pi(n)}\}$	Section 1
$\Delta^n (= \Delta)$	$\{x \in \mathbb{R}^n \,\|\, x_1 \geq \cdots \geq x_n\}$	Section 1, (1)
$\text{conv}(X)$	Convex hull of the set X	Section 1
$\mathcal{P}_k$	k-by-k permutation matrices	Section 2
$x_{[i]}$	ith largest element of x	Section 2, Definition 1
$x \geq_m y$	x majorizes y	Section 2, Definition 1
$x \geq_{wm} y$	x weakly majorizes y from below	Section 2, Definition 1

Appendix. (Continued)

Notation	Definition	Defined in
$\|\cdot\|_s$	An arbitrary sign- and permutation-invariant norm of vectors	Section 3, (10)
$N^K_{\|\cdot\|_s}$	$\{x \in \mathbb{R}^n \mid \|x\|_s \le 1, \mathrm{card}(x) \le K\}$	Section 3, (10)
$B_s(r)$	$\{x \in \mathbb{R}^n \mid \|x\|_s \le r\}$	Section 3
$\Delta^n_+ (= \Delta_+)$	$\Delta^n \cap \mathbb{R}^n_+$	Section 3, (12)
$x_{\Delta^n} (= x_\Delta)$	$(x_{\Delta^n})_i := x_{[i]}$	Section 3, (12)
$x_{\Delta^n_+} (= x_{\Delta_+})$	$(x_{\Delta^n_+})_i := \|x\|_{[i]}$	Section 3, (12)
$\mathrm{vert}(P)$	Set of vertices of the polyhedron P	Section 3, Proposition 1
$\mathrm{int}(X)$	Set of interior points of X	Section 3, Proposition 1
$\|\cdot\|_c$	The norm corresponding to the convex body $N^K_{\|\cdot\|_s}$	Section 3
$s(x)$	$s(x)_i = \frac{\sum_{j=i}^n \|x\|_{[j]}}{K-i+1}$, $i = 1, \dots, K$, $s(x)_0 = s(x)_{K+1} = \infty$	Section 3
i_x	$\arg\min \{s(x)_i \mid i = 1, \dots, K\}$	Section 3
$\delta(x)$	$s(x)_{i_x}$	Section 3
$u(x)$	$u(x)_i := \begin{cases} \|x\|_{[i]}, & i \in \{1, \dots, i_x - 1\} \\ \delta(x), & i \in \{i_x, \dots, K\} \\ 0, & \text{otherwise.} \end{cases}$	Section 3
$\|\cdot\|^{sp}_K$	K-support norm (K-overlap norm)	Section 3
$\sigma(M)$	Vector of singular value of the matrix M	Section 3.1
$\|\cdot\|_*$	Nuclear norm of matrices	Section 3.1
$\mathrm{diag}(v)$	The diagonal matrix whose diagonal is v	Section 3.1
$\mathcal{S}^p$	The set of p-by-p symmetric matrices	Section 3.1
$\mathcal{S}^p_+$	The set of p-by-p positive semidefinite matrices	Section 3.1
$\lambda(M)$	The vector of eigenvalues of the matrix M	Section 3.1
$\mathrm{conv}_C(\phi)$	The convex envelope of ϕ over C	Section 4
$\phi\|_X$	$\phi\|_X(x) = \phi(x)$ for any $x \in X$ and $+\infty$ otherwise	Section 4
$S(Z, a, b)$	$\{(x, z) \in [a, b]^n \times \mathbb{R}^m \mid (x, z) \in Z\}$	Section 4
$X(Z, a, b, \{F_i\}_{i=1}^r)$	$\{(x, z) \in [a, b]^n \times \mathbb{R}^m \mid (x, z) \in Z, x \in \cup_{i=1}^r F_i\}$	Section 4
$\mathcal{M}^n$	$\mathcal{M}^{n,n}(\mathbb{R})$	Section 5
M_S	$\{(x, X) \in \mathbb{R}^n \times \mathcal{M}^n \mid X = xx^\top, x \in S\}$	Section 5
$\mathbb{1}_n (= \mathbb{1})$	n-dimensional vector of ones	Section 5
$\mathbb{1}_{n \times n}$	k-by-k matrix of ones	Section 5
$\mathrm{trace}(M)$	The trace of the matrix M	Section 5, (30b)
$\mathrm{diag}(M)$	The diagonal vector of the matrix M	Section 5
R^M	The vector of row sums of the matrix M	Section 5
$\mathcal{M}^n_{\ge 0}$	The set of n-by-n nonnegative matrices	Section 5.1
$\mathcal{S}^n_{\ge 0}$	The set of n-by-n nonnegative symmetric matrices	Section 5.1

References

[1] Ajtai M, Komlós J, Szemerédi E (1983) An $O(n \log n)$ sorting network. *Proc. 15th Annual ACM Sympos. Theory Comput.* (Association for Computing Machinery, New York), 1–9.
[2] Argyriou A, Foygel R, Srebro N (2012) Sparse prediction with the k-support norm. Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, NY), 25:1457–1465.
[3] Balas E (2018) *Disjunctive Programming* (Springer, Cham, Switzerland).
[4] Barvinok AI (2002) *A Course in Convexity* (American Mathematical Society, Providence, RI).
[5] Ben-Tal A, Nemirovski A (2001) *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, vol. 2 (SIAM, Philadelphia).
[6] Bertsimas D, Cory-Wright R, Pauphilet J (2020) Solving large-scale sparse PCA to certifiable (near) optimality. Preprint, submitted May 11, https://arxiv.org/abs/2005.05195.
[7] d'Aspremont A, El Ghaoui L, Jordan MI, Lanckriet GR (2007) A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.* 49(3):434–448.
[8] Fiorini S, Massar S, Pokutta S, Tiwary HR, de Wolf R (2015) Exponential lower bounds for polytopes in combinatorial optimization. *J. ACM* 62(2):17.
[9] Goemans MX (2015) Smallest compact formulation for the permutahedron. *Math. Programming* 153(1):5–11.
[10] Grant M, Boyd S (2008) Graph implementations for nonsmooth convex programs. Blondel V, Boyd S, Kimura H, eds. *Recent Advances in Learning and Control*. Lecture Notes in Control and Information Sciences, vol. 371 (Springer-Verlag, London), 95–110.
[11] Grant M, Boyd S (2014) CVX: Matlab software for disciplined convex programming. V. 2.1. http://cvxr.com/cvx.
[12] Hardy G, Littlewood J, Pólya G (1952) *Inequalities* (Cambridge University Press, Cambridge, UK).
[13] Hiriart-Urruty JB, Le HY (2012) Convexifying the set of matrices of bounded rank: Applications to the quasiconvexification and convexification of the rank function. *Optim. Lett.* 6(5):841–849.

[14] Horn RA, Johnson CR (2012) *Matrix Analysis*, 2nd ed. (Cambridge University Press, Cambridge, UK).
[15] Jeffers J (1967) Two case studies in the application of principal component analysis. *J. Royal Statist. Soc. Ser. C. Appl. Statist.* 16(3):225–236.
[16] Kaibel V, Pashkovich K (2011) Constructing extended formulations from reflection relations. Günlük O, Woeginger GJ, eds. *Proc. 15th Internat. Conf.* Integer Programming Combin. Optim. Lecture Notes in Computer Science, vol. 6655 (Springer, Berlin), 287–300.
[17] Kim J, Tawarmalani M, Richard JPP (2019) Convexification of permutation-invariant sets and applications. Preprint, submitted October 7, https://arxiv.org/abs/1910.02573.
[18] Lewis AS (1995) The convex analysis of unitarily invariant matrix functions. *J. Convex Anal.* 2(1):173–183.
[19] Li CK, Mehta PP (1995) Permutation invariant norms. *Linear Algebra Appl.* 219:93–110.
[20] Lim CH (2017) A note on extended formulations for cardinality-based sparsity. *Proc. 10th Neural Inform. Processing Systems Workshop Optim. Machine Learn.* (Long Beach, CA).
[21] Lim CH, Wright S (2017) *k*-support and ordered weighted sparsity for overlapping groups: Hardness and algorithms. Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, eds. *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., Red Hook, NY), 284–292.
[22] Löfberg J (2004) YALMIP: A toolbox for modeling and optimization in MATLAB. Hara S, Fu LC, Ge SS, Samad T, Sebek M, Engell S, eds. *Proc. IEEE Internat. Sympos. Comput. Aided Control System Design* (Institute of Electrical and Electronics Engineers, Piscataway, NJ), 284–289.
[23] Luedtke J, Namazifar M, Linderoth J (2012) Some results on the strength of relaxations of multilinear functions. *Math. Programming* 136(2):325–351.
[24] Marshall AW, Olkin I, Arnold B (2010) *Inequalities: Theory of Majorization and Its Applications* (Springer-Verlag, New York).
[25] Matoušek J (2002) *Lectures on Discrete Geometry*, vol. 212 (Springer-Verlag, New York).
[26] MOSEK ApS (2017) MOSEK optimization suite. Accessed June 11, 2021, https://www.mosek.com/.
[27] Nemirovski A (2012) Introduction to linear optimization, ISYE 6661. Lecture notes, Georgia Institute of Technology, Atlanta.
[28] Nguyen TT, Richard JPP, Tawarmalani M (2018) Deriving convex hulls through lifting and projection. *Math. Programming* 169(2):377–415.
[29] Rado R (1952) An inequality. *J. London Math. Soc.* 27(1):1–6.
[30] Rikun AD (1997) A convex envelope formula for multilinear functions. *J. Global Optim.* 10(4):425–437.
[31] Rockafellar RT (1970) *Convex Analysis* (Princeton University Press, Princeton, NJ), 284–289.
[32] Tawarmalani M, Richard JPP (2017) Decomposition techniques in convexification of sets. Working paper, Krannert School of Management, Purdue University, West Lafayette, IN.
[33] Tillmann AM, Pfetsch ME (2013) The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Trans. Inform. Theory* 60(2):1248–1259.
[34] Zou H, Hastie T, Tibshirani R (2006) Sparse principal component analysis. *J. Comput. Graphical Statist.* 15(2):265–286.