Strong Formulations and Algorithms for Regularized A-optimal Design

Yongchun Li

Department of Industrial and Systems Engineering, University of Tennessee, Knoxville, ycli@utk.edu

Abstract. We study the Regularized A-optimal Design (RAOD) problem, which selects a subset of k experiments to minimize the inverse of the Fisher information matrix, regularized with a scaled identity matrix. RAOD has broad applications in Bayesian experimental design, sensor placement, and cold-start recommendation. We prove its NP-hardness via a reduction from the independent set problem. By leveraging convex envelope techniques, we propose a new convex integer programming formulation for RAOD, whose continuous relaxation dominates those of existing formulations. More importantly, we demonstrate that our continuous relaxation achieves bounded optimality gaps for all k, whereas previous relaxations may suffer from unbounded gaps. This new formulation enables the development of an exact cutting-plane algorithm with superior efficiency, especially in high-dimensional and small-k scenarios. We also investigate scalable forward and backward greedy algorithms for solving RAOD, each with provable performance guarantees for different k ranges. Finally, our numerical results on synthetic and real data demonstrate the efficacy of the proposed exact and approximation algorithms. We further showcase the practical effectiveness of RAOD by applying it to a real-world user cold-start recommendation problem.

Key words : Regularized A-optimality, Mixed-integer nonlinear optimization, Convex relaxations, Cutting-plane algorithm, Forward and backward greedy algorithms, User cold-start recommendation

1. Introduction

The A-optimal Design (AOD) problem arises in statistical design for situations where, due to budget limitations, one can observe the outcomes for a small subset of experiments to optimize the parameter estimates. Specifically, AOD aims to minimize the trace of the inverse of the Fisher information matrix, resulting in minimizing the average variance of the parameter estimates. AOD has found broad applications in fields such as engineering (Ilzarbe et al. 2008, Mason et al. 2003), biology (Quinn and Keough 2002), and chemistry (Higgs et al. 1997, Leardi 2009), when dealing with large-scale and costly experiments. Linear regression models are one of the most widely used and studied models in this area (see, Jobson 2012, Nikolov et al. 2022, Winer et al. 1971). This paper focuses on the Regularized AOD (RAOD) problem in the context of linear regression models, which adds a scaled identity matrix to the Fisher information matrix, as shown below. This addition typically helps prevent over-fitting (Tantipongpipat 2020), improves numerical stability when the information matrix is rank-deficient (Dereziński and Warmuth 2017), and encodes prior information (Bian et al. 2017).

$$z_k := \min_{S \subseteq [n]} \left\{ \mathrm{tr}\left(\left(\sum_{i \in S} \boldsymbol{a}_i \boldsymbol{a}_i^\top + \lambda \boldsymbol{I}_d \right)^{-1} \right) : |S| = k \right\}, \tag{RAOD}$$

where $\{\boldsymbol{a}_i\}_{i\in[n]} \in \mathbb{R}^d$ represents a collection of n experiments parameterized by d-dimensional vectors with $d \leq n$, $k \leq n$ denotes the number of chosen experiments, $\lambda > 0$ is a constant, $\boldsymbol{I}_d$ is a $d \times d$ identity matrix, and for a subset $S \subseteq [n]$, let $|S|$ denote its cardinality. Note that the selection size k can be smaller or larger than the dimension d, depending on the specific problem setup.

The following outlines the two most relevant variants and an equivalent problem of RAOD, which will be used throughout to show the scope and limitations of existing research, facilitate our analysis, and highlight the general applicability of some of our results.

(i) *AOD*. As previously mentioned, λ is fixed at zero in AOD. It is noted that the condition $k \geq d$ must hold for AOD to ensure that the objective matrix is invertible;

(ii) *Bayesian AOD*. Since specific information is usually available prior to the experimental selection process, Bayesian AOD, which leverages prior knowledge, can play an important role in improving design efficiency (see the excellent survey by Chaloner and Verdinelli 1995, Rainforth et al. 2024). Mathematically, it generalizes the scaled identity matrix in RAOD to an arbitrary positive definite matrix representing prior information. Bayesian AOD is also referred to as the A-Fusion problem in Hendrych et al. (2023). When the prior is an isotropic Gaussian, Bayesian AOD reduces to RAOD as a special case (see, e.g., Bian et al. 2017, Chamon and Ribeiro 2017a); and

(iii) *A-optimal Maximum Entropy Sampling Problem (AMESP)*. As introduced in Li and Xie (2024), AMESP focuses on selecting the most informative principal submatrix from a given covariance matrix by minimizing the trace of its inverse. We are the first to demonstrate that RAOD and AMESP are equivalent, with an additive additional constant of $(d-k)/\lambda$, when the covariance matrix in AMESP is positive definite.

Although RAOD is a special case of Bayesian AOD, it remains broadly applicable. One key motivation comes from the user cold-start problem in recommendation systems. The problem arises when a new user joins a recommendation platform and has not rated any item, making recommendations challenging due to the lack of historical data (Lika et al. 2014).

Traditional methods, such as content-based filtering and collaborative filtering, exploit relationships and similarities between users to make recommendations (Bobadilla et al. 2012, Volkovs et al. 2017). Alternatively, experimental design approaches, including sequential strategies, tackle the problem at its root by explicitly asking new users to rate selected items upon joining a recommendation platform (see, e.g., Anava et al. 2015, Elahi et al. 2016, Gope and Jain 2017). These

items are selected carefully to capture new users' preferences. However, users are often reluctant to rate many items due to cognitive costs. If users rate too few items, the Fisher information matrix becomes rank-deficient, leading to ill-posed matrix inverse and determinant problems in experimental design. To mitigate this, regularized experimental design has emerged as a practical approach by adding a scaled identity matrix, as shown in prior research by Barz et al. (2015), Anava et al. (2015), Rubens et al. (2009). In particular, Chamon and Ribeiro (2017a) applied the RAOD approach and demonstrated its effectiveness in selecting movies for new users to rate.

1.1. Related Work

In this subsection, we review the relevant literature on RAOD and its variants, with a focus on their exact algorithms, convex relaxations, and approximation algorithms.

Bayesian D-optimal Design (Bayesian DOD) & Bayesian AOD. In experimental designs, the D- and A-optimality criteria are widely used for selecting experiments to achieve precise parameter estimation. A-optimality minimizes the estimation variance directly, while D-optimality seeks to reduce the volume of the confidence ellipsoid of the parameter estimates by maximizing the determinant of the Fisher information matrix. Notably, given that the determinant is a multiplicative operator, Bayesian DOD is equivalent to regularized DOD (Li et al. 2024); however, this equivalence does not hold between Bayesian AOD and RAOD.

A-optimality is known to be computationally more expensive than D-optimality due to its need to compute the inverse matrix (Ahipaşaoglu 2021). As a result, algorithms and software for Bayesian DOD are more developed than Bayesian AOD (see Jones et al. 2021). Unfortunately, these techniques cannot be directly applied to Bayesian AOD. Nevertheless, Bayesian AOD often outperforms Bayesian DOD, with improved variance and bias properties (see, e.g., Jones et al. 2021, Stallrich et al. 2023). Our numerical study on cold-start recommendation further shows the superior performance of RAOD compared to regularized DOD. In addition, DOD, AOD, and Bayesian DOD have been proven to be NP-hard, see Welch (1982), Nikolov et al. (2022), Li et al. (2024), respectively. In contrast, the complexity of Bayesian AOD remains an open question, though Hendrych et al. (2023) conjectured that it was NP-hard. The limited literature on Bayesian AOD motivates us to develop efficient algorithms for its special case, RAOD. While this paper does not address the general Bayesian AOD problem, we demonstrate its NP-hardness for the first time.

Exact Algorithms. Experimental design problems can be formulated as mixed-integer nonlinear programs (MINLPs) by introducing binary variables to represent subset selection. In view of this,

advancing exact solution methods for these MINLPs has been a longstanding research focus. Given the complexity of the A-optimality function, RAOD and its variants naturally reduce to mixed-integer semidefinite programs (MISDPs) (Duarte 2023). To solve these MISDPs, the branch-and-bound (B&B) method has been extensively studied in the literature (see, e.g., Ahipaşaoglu 2021, Hendrych et al. 2023, Liang and Yang 2024). These works leveraged the standard continuous relaxation of RAOD to provide a lower bound, referred to as RAOD-RI throughout this paper, which is also reviewed in detail below.

An alternative approach formulates RAOD and its variants as mixed-integer second-order cone programs (MISOCPs). It is worth noting that unlike MISDP, modern optimization solvers such as Gurobi and CPLEX are well-suited for solving MISOCP. Sagnol (2011) pioneered this direction by applying Elfving's Theorem to demonstrate the second-order representability of the continuous relaxation of AOD, and it is actually a consequence of Gauss–Markov Theorem. Duarte and Wong (2015) later extended the result to the Bayesian version. Building on these, a seminar work by Sagnol and Harman (2015) developed exact MISOCP reformulations of AOD and Bayesian AOD.

Convex Relaxation. A natural way to get rid of the integer challenge is relaxing the binary variables in the exact MISDP reformulation of RAOD to be continuous, which leads to the convex relaxation RAOD-RI. Much of the literature on the relaxation has focused on developing efficient convex optimization algorithms, see Ahipaşaoglu (2015), Liang and Yang (2024), and many references therein. However, despite these computational advancements, we demonstrate that RAOD-RI may perform poorly when k is within the range of $[d-1]$. This limitation motivates us to develop stronger relaxations, which are important for improving the efficiency of exact algorithms.

Approximation Algorithms. Avron and Boutsidis (2013) studied the classic AOD problem and proposed volume-sampling-based algorithms that achieved an approximation ratio of $(n-d+1)/(k-d+1)$ for $k \in [d, n]$. The result was later extended to RAOD by Dereziński and Warmuth (2017) in the same regime of k, who developed a faster $\mathcal{O}(nd^2)$-time sampling algorithm.

Another approach to developing sampling algorithms is based on convex relaxations. These algorithms start from solving a convex relaxation and then, based on the continuous solution, design randomized sampling strategies to obtain a discrete solution with approximation guarantees, see Nikolov et al. 2022, Wang et al. 2017 for the classic AOD problem, Tantipongpipat (2020), Li and Xie (2024) for RAOD, and Allen-Zhu et al. (2017), Derezinski et al. (2020) for Bayesian AOD, respectively. However, these convex relaxation-based sampling algorithms of RAOD often

scale poorly with the number of experiments n (see, e.g., the fourth and fifth rows of Table 2). It is important to note that when applying the approximation algorithms developed for AMESP by Li and Xie (2024) to RAOD, an additional constant $(d-k)/\lambda$ must be added, as shown in Proposition 1. Hence, their approximation ratios remain valid when $k \leq d$ after adding a nonnegative constant, since for any $c_1 \geq c_2$, it is easy to verify that $\frac{c_1+(d-k)/\lambda}{c_2+(d-k)/\lambda} \leq \frac{c_1}{c_2}$. However, these ratios do not hold for $k \in [d+1, n]$, where $(d-k)/\lambda$ becomes negative.

Natural combinatorial algorithms, such as the greedy and local search algorithms, have been widely used in experimental design (see, e.g., Krause et al. 2008, Madan et al. 2019). The forward greedy algorithm has also been extended to Bayesian criteria in Bian et al. (2017), Chamon and Ribeiro (2017b). However, A-optimality is not a supermodular function (Chamon and Ribeiro 2017a), which means that the typical approximation ratio for forward greedy (i.e., $(1-1/e)$) cannot apply. Instead, these greedy methods often achieve data-dependent guarantees, where the worst-case performance can vary significantly with the experimental setup. By data-independence, we mean a theoretical guarantee that does not rely on the experiments $\{\boldsymbol{a}_i\}_{i \in [n]}$ or λ but may depend on the parameters n, d, and k. We are unaware of any prior work on developing a backward greedy search method in the literature of RAOD. This gap is addressed in our paper. Notably, Li and Xie (2024) proposed a local search algorithm for AMESP with strong empirical performance, but its theoretical guarantee is data-independent. The performance guarantees and complexities of the approximation algorithms reviewed for RAOD are summarized in Table 2.

1.2. Summary of the Organization and Contributions

Below we list the major contributions and an outline of the remaining paper.

(i) **Complexity analysis.** In Section 2, we prove the NP-hardness of RAOD by establishing its equivalence to AMESP and reducing the independent set decision problem to AMESP;

(ii) **Optimality gaps of convex relaxations.** Section 3 presents existing (mixed-)integer convex formulations of RAOD and analyzes the optimality gaps of their convex relaxations, as summarized in Table 1. Notably, we show that the conventional relaxation RAOD-RI can have unbounded gaps for $k \in [d-1]$, while the relaxation AMESP-R of AMESP may perform worse than the trivial zero lower bound for $k \in [d+1, n]$. Finally, in Subsection 3.3, we revisit the formulation MISOCP from an optimization perspective and prove that its continuous relaxation coincides with RAOD-RI;

(iii) **New formulations.** Motivated by the weaknesses of existing formulations, we propose a novel convex integer programming formulation for RAOD in Section 4, whose continuous relaxation RAOD-RII is tighter than those presented in Section 3. More importantly, RAOD-RII achieves bounded optimality gaps for all $k \in [n]$, which addresses a key drawback of RAOD-RI and AMESP-R. Building on our new formulation, we develop an exact cutting-plane algorithm to solve RAOD to optimality;

(iv) **Greedy algorithms.** In Section 5, we first establish a data-independent approximation ratio for the forward greedy algorithm when $k \in [d-1]$. To complement this, we propose a backward greedy algorithm for RAOD and derive its data-independent performance guarantee for $k \in [d, n]$. These results, summarized in Table 2, present a comprehensive analysis of our greedy algorithms across different ranges of k; and

(v) **Numerical study.** Finally, in Section 6, our numerical study on synthetic and real data demonstrates the effectiveness of our exact algorithm and the scalability and high-quality outputs of our approximation algorithms. In particular, when applied to the user cold-start problem in movie recommendation systems, RAOD effectively identifies movie subsets that lead to more reliable recommendations for new users in most instances, outperforming its D-optimality counterpart.

Table 1 Summary of convex relaxations for RAOD

Convex relaxation	Optimality gap (i.e., z_k/relaxation value)	Value of k
RAOD-RI	Can be arbitrarily large	$k \in [d-1]$
	$\min\{d, n-d+1\}$	$k = d$
	$\frac{n-d+1}{k-d+1}$	$k \in [d+1, n]$
AMESP-R	$\min\left\{\frac{d-1}{d-k}, n-k+1\right\}$	$k \in [d-1]$
	$\min\{d, n-d+1\}$	$k = d$
	Can be negative	$k \in [d+1, n]$
RAOD-RII	$\min\left\{\frac{d-1}{d-k}, n-k+1\right\}$	$k \in [d-1]$
	$\min\{d, n-d+1\}$	$k = d$
	$\frac{n-d+1}{k-d+1}$	$k \in [d+1, n]$

Notations. We let $\mathcal{S}_+^n, \mathcal{S}_{++}^n$ denote the set of all the $n \times n$ symmetric positive semidefinite and positive definite matrices, respectively. We let $\mathbb{R}^n$, $\mathbb{R}_+^n$, and $\mathbb{R}_{++}^n$ denote the set of all the

Table 2   Summary of approximation algorithms for RAOD

Approximation algorithm	Time complexity	Approximation ratio
Sampling (Dereziński and Warmuth 2017)	$\mathcal{O}((n^2 + d^2)d)$	$\frac{n-d+1}{k-d+1}$ if $k \in [d, n]$
Fast sampling (Dereziński and Warmuth 2017)	$\mathcal{O}((n + d)d^2)$	$\frac{n-d+1}{k-d+1}$ if $k \in [d, n]$
Forward greedy (Chamon and Ribeiro 2017b)	$\mathcal{O}(knd^2)$	$-$[i]
Sampling (Tantipongpipat 2020)	$\mathcal{O}(n^4 dk^2 \log(dk))$	$-$
Sampling (Li and Xie 2024)	$\mathcal{O}(n^4 dk^2 \log(dk))$	$\min\{k, n-k+1\}$ if $k \in [d]$
Local search (Li and Xie 2024)	$\mathcal{O}(k(n-k)d^5)$	$-$
Forward greedy Algorithm 2	$\mathcal{O}(knd^2)$	$\frac{d-1}{d-k}$ if $k \in [d-1]$
Backward greedy Algorithm 3	$\mathcal{O}((n-k)nd^2)$	$\frac{n-d+1}{k-d+1}$ if $k \in [d, n]$

[i] "$-$": The approximation ratio depends on data $\{\boldsymbol{a}_i\}_{i\in[n]}$ and the constant λ

n-dimensional vectors, nonnegative vectors, and positive vectors, respectively. We use bold lower-case letters (e.g., $\boldsymbol{x}$) and bold upper-case letters (e.g., $\boldsymbol{X}$) to denote vectors and matrices, respectively, and use corresponding non-bold letters (e.g., x_i) to denote their components. Given two positive integers $s \leq n$, we let $[n] = \{1, 2, \cdots, n\}$, and let $[s, n] = \{s, s+1, \cdots, n\}$. We let $\boldsymbol{I}_n$ be the $n \times n$ identity matrix. For a vector $\boldsymbol{x} \in \mathbb{R}^n$, let $\mathrm{Diag}(\boldsymbol{x})$ denote a diagonal matrix with diagonal entries being $\boldsymbol{x}$ and let $\|\boldsymbol{x}\|_2$ denote its two norm. If $\boldsymbol{x} \in \mathbb{R}_+^n$ is nonnegative, we define $\sqrt{\boldsymbol{x}}$ as the vector where each entry is the square root of the corresponding element in $\boldsymbol{x}$. For a matrix $\boldsymbol{X} \in \mathbb{R}^{d \times n}$, we let $\|\boldsymbol{X}\|_F$ denote its Frobenius norm, let $\boldsymbol{X}_i \in \mathbb{R}^d$ denote its i-th column vector for each $i \in [n]$, and for a subset $S \subseteq [n]$, let $\boldsymbol{X}_S$ denote the submatrix consisting of the columns of $\boldsymbol{X}$ indexed by S. For a symmetric matrix $\boldsymbol{X} \in \mathcal{S}_+^n$, let $\mathrm{tr}(\boldsymbol{X})$ denote its trace, and for a subset $S \subseteq [n]$, we let $\boldsymbol{X}_{S,S}$ denote a principal submatrix of $\boldsymbol{X}$ indexed by S.

2. Complexity Analysis

We begin by showing the equivalence between RAOD and AMESP with an additional constant of $(d-k)/\lambda$. Proposition 1 extends Li et al. (2024, Theorems 1 and 2), where they demonstrated the equivalence between Bayesian DOD and MESP. Our Proposition 1 further connects experimental design problems with maximum entropy-based optimization.

For brevity, we define the objective function of RAOD as

$$f(S) := \mathrm{tr}\left(\left(\sum_{i \in S} \boldsymbol{a}_i \boldsymbol{a}_i^\top + \lambda \boldsymbol{I}_d\right)^{-1}\right)$$

for any subset $S \subseteq [n]$ throughout.

PROPOSITION 1. *RAOD can be converted into the following:*

$$z_k = \min_{S \subseteq [n]} \left\{ \operatorname{tr}\left((\boldsymbol{C}_{S,S})^{-1}\right) + \frac{d-k}{\lambda} : |S| = k \right\}, \qquad \text{(AMESP)}$$

where $\boldsymbol{C} \in \mathcal{S}_+^n := \boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{I}_n$, and $\boldsymbol{A} \in \mathbb{R}^{d \times n}$ is a matrix whose columns are vectors $\{\boldsymbol{a}_i\}_{i \in [n]}$.

Proof. See Appendix A.1. □

It is well known that the independent set decision problem is NP-hard (Cormen et al. 2022, Chapter 34). Next, we prove that the independent set decision problem can be reduced to AMESP, which, combined with Proposition 1, establishes the NP-hardness of RAOD.

DEFINITION 1 (INDEPENDENT SETS AND THE INDEPENDENT SET DECISION PROBLEM). In an undirected graph G, an independent set S is a set of vertices such that no two vertices in S are adjacent. Given an integer k, the independent set decision problem is to determine if the graph G contains an independent set of size k.

Let us introduce a known inequality based on the comparison between the harmonic mean and arithmetic mean, which is a key technical result used throughout the paper.

LEMMA 1 (**Sedrakyan and Sedrakyan 2018**). *Suppose $\boldsymbol{\sigma} \in \mathbb{R}_{++}^n$. Then, the following holds*

$$\sum_{i \in [n]} \frac{1}{\sigma_i} \geq \frac{n^2}{\sum_{i \in [n]} \sigma_i},$$

which becomes an equality if and only if all entries of $\boldsymbol{\sigma}$ are equal.

THEOREM 1 (**NP-hardness**). *RAOD is NP-hard.*

Proof. See Appendix A.2. □

Theorem 1 confirms the following conjecture from Hendrych et al. (2023) regarding Bayesian AOD. Since RAOD is a special case of Bayesian AOD, the conjecture is clearly true.

CONJECTURE 1 (**Hendrych et al. 2023**). *The Bayesian AOD problem is NP-hard.*

3. Optimality Gaps of Convex Relaxations

This section presents two widely-used (mixed-)integer convex formulations of RAOD and derives the optimality gaps for their continuous relaxations. Leveraging Proposition 1, we further analyze the quality of a convex relaxation for AMESP.

3.1. The Conventional Convex Integer Programming Formulation and its Convex Relaxation

By introducing a binary characterization of the subset $S \subseteq [n]$, where $x_i = 1$ if $i \in S$ and $x_i = 0$ otherwise, we can readily obtain a convex integer nonlinear formulation of RAOD

$$z_k = \min_{\boldsymbol{x} \in \{0,1\}^n} \left\{ \mathrm{tr} \left(\left(\sum_{i \in [n]} x_i \boldsymbol{a}_i \boldsymbol{a}_i^\top + \lambda \boldsymbol{I}_d \right)^{-1} \right) : \sum_{i \in [n]} x_i = k \right\}, \tag{1}$$

and its continuous relaxation provides a practical lower bound (Tantipongpipat 2020):

$$z_k \geq \nu_k^{\mathrm{C}} := \min_{\boldsymbol{x} \in [0,1]^n} \left\{ \mathrm{tr} \left(\left(\sum_{i \in [n]} x_i \boldsymbol{a}_i \boldsymbol{a}_i^\top + \lambda \boldsymbol{I}_d \right)^{-1} \right) : \sum_{i \in [n]} x_i = k \right\}. \tag{RAOD-RI}$$

It is evident that ν_k^{C} is always positive.

The problem (1) can be recast as a MISDP by nature. The conventional relaxation RAOD-RI has been widely incorporated into the B&B algorithms for solving (1) to optimality (e.g., Hendrych et al. 2023). For B&B algorithms, the quality of the relaxation bound used is crucial for pruning the search space. We are thus motivated to explore the quality of RAOD-RI. Before that, let us explore the properties of the objective function $f(S)$ in RAOD. First, the function $f(S)$ decreases monotonically with respect to S, that is, $f(S \setminus \{i\}) \geq f(S)$ for any $i \in S$. We also remark that

REMARK 1. The optimal value z_k decreases as the selection size k increases. $\diamond$

Lemma 2 establishes an upper bound for the smallest value of $f(S \setminus \{i\})$ over $i \in S$. Our derivation leverages eigen-decomposition techniques and the sampling probability of removing an element from S, as introduced by Dereziński and Warmuth (2017). In Dereziński and Warmuth (2017), this probability led to a regularized volume sampling algorithm with a theoretical guarantee for RAOD only when $k \in [d, n]$. A key difference from Dereziński and Warmuth (2017, Lemma 5) is that Lemma 2 provides a general treatment that applies to any subset S, without restricting its size at least d. Based on Lemma 2, we derive the optimality gap of RAOD-RI for $k \in [d, n]$, as shown in Theorem 2, and analyze the performance guarantees of our approximation algorithms in Section 5.

LEMMA 2. *For a nonempty subset $S \subseteq [n]$ of size $s \geq 1$, we have that*

$$\min_{i \in S} f(S \setminus \{i\}) \leq \begin{cases} \left(1 + \frac{1}{s\rho + d - s}\right) f(S), & \text{if } s \in [d-1]; \\ \left(1 + \frac{1}{d\rho + s - d}\right) f(S), & \text{if } s \in [d, n]. \end{cases},$$

where $\rho \in (0,1) := \frac{\lambda}{\lambda + \max\{1, s/d\} \max_{i \in [n]} \|\boldsymbol{a}_i\|_2^2}$.

Proof. See Appendix A.3. □

THEOREM 2 (**Optimality gap**). *The following hold for the lower bound ν_k^C in RAOD-RI:*

(i) For any $k \in [d, n]$, we have

$$1 \le \frac{z_k}{\nu_k^C} \le \frac{n - d + 1}{k - d + 1};$$

and

(ii) For $k \in [d-1]$, there is a case of RAOD-RI where $z_k/\nu_k^C \to \infty$ as $\lambda \to 0$.

Proof. See Appendix A.4. □

Theorem 2 shows that RAOD-RI provides a lower bound with a bounded data-independent optimality gap for $k \in [d, n]$. It is evident that this gap decreases in k but increases in d and n. Note that the gap can be further tightened at $k = d$, as shown in Corollary 3. However, for small values of $k \in [d-1]$, RAOD-RI may lead to an infinite optimality gap.

3.2. A Convex Relaxation of AMESP

Based on the equivalence between RAOD and AMESP, this subsection presents a convex relaxation of AMESP proposed by Li and Xie (2024). We establish a novel optimality gap of AMESP-R for $k \in [d]$ using Lemma 1. However, when $k \in [d+1, n]$, we demonstrate that AMESP-R may yield a negative optimal value due to the negativity of the constant $(d-k)/\lambda$.

DEFINITION 2 (LI AND XIE 2024). For a matrix $\boldsymbol{Y} \in \mathcal{S}_+^n$ with eigenvalues $\beta_1 \ge \cdots \ge \beta_n \ge 0$, we define a function $\Phi : \mathcal{S}_+^n \to \mathbb{R}$ as

$$\Phi(\boldsymbol{Y}) := \phi(\boldsymbol{\beta}) = \sum_{i \in [\iota]} \frac{1}{\beta_i} + \frac{(k - \iota)^2}{\sum_{i \in [\iota+1, n]} \beta_i},$$

where $0 \le \iota \le k - 1$ is a unique integer such that $\beta_\iota > \frac{1}{k-\iota} \sum_{i \in [\iota+1, n]} \beta_i \ge \beta_{\iota+1}$ with $\beta_0 = \infty$.

Note that $\Phi(\boldsymbol{Y})$ is a spectral function that relies on only the eigenvalues of $\boldsymbol{Y}$. Li and Xie (2024) demonstrated its convexity and explicitly described its subgradient.

Recall that in Proposition 1, we define $\boldsymbol{C} = \boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{I}_n$. Given that $\boldsymbol{C} \in \mathcal{S}_{++}^n$ is positive definite, we compute its Cholesky factorization $\boldsymbol{C} = \boldsymbol{H}^\top \boldsymbol{H}$, where $\boldsymbol{H} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{h}_i \in \mathbb{R}^n$ represents the i-th column vector of $\boldsymbol{H}$ for all $i \in [n]$. According to Li and Xie (2024, Theorem 9), the following formulation based on the function Φ is a convex relaxation of AMESP.

$$z_k \ge \nu_k^{\mathrm{M}} := \min_{\boldsymbol{x} \in [0,1]^n} \left\{ \Phi\left(\sum_{i \in [n]} x_i \boldsymbol{h}_i \boldsymbol{h}_i^\top \right) + \frac{d - k}{\lambda} : \sum_{i \in [n]} x_i = k \right\}. \qquad \text{(AMESP-R)}$$

It is important to note that AMESP-R is exact when $\boldsymbol{x}$ is binary.

Next, we present our main results about AMESP-R.

THEOREM 3 (**Optimality gap**). *The following hold for the lower bound ν_k^M in AMESP-R:*

(i) For any $k \in [d]$, we have

$$1 \le \frac{z_k}{\nu_k^M} \le \begin{cases} \min\left\{\frac{d-1}{d-k}, n-k+1\right\} & \text{if } k \in [d-1]; \\ \min\left\{d, n-d+1\right\} & \text{if } k = d. \end{cases};$$

and

(ii) If $k \ge d+1$, then there is a case of AMESP-R where $\nu_k^M < 0$.

Proof.   See Appendix A.5.   □

When $k \in [d-1]$, Part (i) of Theorem 3 guarantees an optimality gap for AMESP-R equal to the minimum of two bounds: the first bound $(d-1)/(d-k)$, is newly derived in this work, while the second $(n-k+1)$ is the prior result in Li and Xie (2024). When $k \le (n+1)/2$, it is easy to check that $n-k+1 \ge k \ge (d-1)/(d-k)$, where the second inequality is from the proof of Theorem 3. While not directly comparable, our bound $(d-1)/(d-k)$ offers a compelling advantage due to its independence of n and its scaling with d. Specifically, it becomes smaller as the dimension d increases, which is a somewhat counterintuitive yet interesting property. This suggests that for $k \in [d-1]$, AMESP-R achieves improved theoretical guarantees in high-dimensional settings.

According to Theorems 2 and 3, the two convex relaxations, RAOD-RI and AMESP-R, differ significantly in their theoretical guarantees. Overall, AMESP-R performs well for $k \in [d]$, while RAOD-RI is effective for larger values of $k \in [d, n]$. In addition, we would like to highlight that AMESP-R has two major limitations.

(i) First, AMESP-R is formulated in a higher-dimensional matrix space of size $n \times n$, as opposed to the $d \times d$ space used in RAOD-RI. This increases the computational cost of solving AMESP-R, making it less unfavorable for large-scale instances; and

(ii) For $k \in [d+1, n]$, according to Part (ii) of Theorem 3, the lower bound from AMESP-R may underperform even the trivial bound of zero. The numerical results in Subsection 6.1 further confirm this, showing that AMESP-R often yields negative lower bounds and performs much more poorly compared to RAOD-RI in this range of k.

3.3. A Mixed-integer Second-order Cone Programming Formulation

This subsection presents a MISOCP reformulation of RAOD, which was first introduced by Sagnol and Harman (2015) based on the Gauss–Markov Theorem from statistical estimation theory

(Pukelsheim 2006). We provide a different proof of this formulation through an optimization lens and further contribute by analyzing the optimality gap of its continuous relaxation.

Using the Woodbury matrix identity, we show that the objective function of RAOD-RI is SOCP representable for any (not necessarily binary) vector $\boldsymbol{x}$, as summarized below.

LEMMA 3. *The identity below holds for any vector $\boldsymbol{x} \in [0,1]^n$.*

$$\mathrm{tr}\left[\left(\sum_{i\in[n]} x_i \boldsymbol{a}_i \boldsymbol{a}_i^\top + \lambda \boldsymbol{I}_d\right)^{-1}\right] = \min_{\boldsymbol{X}\in\mathbb{R}^{d\times n}} \frac{1}{\lambda}\left\|\boldsymbol{X}\,\mathrm{Diag}\left(\sqrt{\boldsymbol{x}}\right)\boldsymbol{A}^\top - \boldsymbol{I}_d\right\|_F^2 + \|\boldsymbol{X}\|_F^2.$$

Proof. See Appendix A.6. □

Using the identity in Lemma 3 to replace the objective function of (1), we arrive at an equivalent MISCOP formulation.

PROPOSITION 2. *RAOD is equivalent to a MISOCP problem:*

$$z_k = \min_{\substack{\boldsymbol{x}\in\{0,1\}^n,\\ \boldsymbol{X}\in\mathbb{R}^{d\times n},\boldsymbol{\mu}\in\mathbb{R}^n}} \left\{\frac{1}{\lambda}\|\boldsymbol{X}\boldsymbol{A}^\top - \boldsymbol{I}_d\|_F^2 + \sum_{i\in[n]}\mu_i : \|\boldsymbol{X}_i\|_2^2 \le \mu_i x_i, \forall i \in [n], \sum_{i\in[n]} x_i = k\right\}.$$

(MISOCP)

In addition, the continuous relaxation of MISOCP matches that of RAOD, that is,

$$\nu_k^C = \min_{\substack{\boldsymbol{x}\in[0,1]^n,\\ \boldsymbol{X}\in\mathbb{R}^{d\times n},\boldsymbol{\mu}\in\mathbb{R}^n}} \left\{\frac{1}{\lambda}\|\boldsymbol{X}\boldsymbol{A}^\top - \boldsymbol{I}_d\|_F^2 + \sum_{i\in[n]}\mu_i : \|\boldsymbol{X}_i\|_2^2 \le \mu_i x_i, \forall i \in [n], \sum_{i\in[n]} x_i = k\right\}. \quad (2)$$

Proof. See Appendix A.7. □

As MISOCP attains the same continuous relaxation value as (1), we can directly apply the optimality gaps of RAOD-RI from Theorem 2 to (2), which are summarized below. However, analogous to RAOD-RI, (2) may provide a weak lower bound for $k \in [d-1]$.

COROLLARY 1. *The following hold for the lower bound ν_k^C in (2):*

(i) *For any $k \in [d,n]$, we have*

$$1 \le \frac{z_k}{\nu_k^C} \le \frac{n-d+1}{k-d+1};$$

and

(ii) *For $k \in [d-1]$, there is a case of RAOD-RI where $z_k/\nu_k^C \to \infty$ as $\lambda \to 0$.*

4. A Novel Convex Integer Formulation and its Convex Relaxation

The limitations of RAOD-RI and AMESP-R highlight the need for a stronger reformulation of RAOD. Motivated by this, this section introduces a novel convex integer program for RAOD, and we demonstrate that its continuous relaxation is stronger than RAOD-RI and AMESP-R. We also develop a cutting-plane algorithm to solve RAOD based on this formulation.

4.1. A Novel Convex Integer Programming Formulation

We observe that for a binary vector $\boldsymbol{x}$ with a cardinality of k, the matrix $\sum_{i\in[n]} x_i \boldsymbol{a}_i \boldsymbol{a}_i^\top$ in the objective of (1) has a rank of at most $\min\{k,d\}$, and becomes low-rank when $k < d$. This rank deficiency results in an arbitrarily large gap of RAOD-RI for $k < d$, as detailed in the proof of Part (ii) of Theorem 2. Inspired by this observation, we introduce a function based on the top $\min\{k,d\}$ eigenvalues to reformulate the objective function of (1). We define $\tilde{k} := \min\{k,d\}$ throughout for brevity.

DEFINITION 3. *For a matrix $\boldsymbol{X} \in \mathcal{S}_+^d$ with eigenvalues $\sigma_1 \geq \cdots \geq \sigma_d \geq 0$, we define a function $\Gamma(\boldsymbol{X};\lambda) : \mathcal{S}_+^d \to \mathbb{R}$ as*

$$\Gamma(\boldsymbol{X};\lambda) = \sum_{i\in[\tilde{k}]} \frac{1}{\sigma_i + \lambda}.$$

LEMMA 4. *For a binary vector $\boldsymbol{x} \in \{0,1\}^n$ with $\sum_{i\in[n]} x_i = k$, the identity below holds.*

$$\Gamma\left(\sum_{i\in[n]} x_i \boldsymbol{a}_i \boldsymbol{a}_i^\top; \lambda\right) + \frac{d-\tilde{k}}{\lambda} = \operatorname{tr}\left(\left(\sum_{i\in[n]} x_i \boldsymbol{a}_i \boldsymbol{a}_i^\top + \lambda \boldsymbol{I}_d\right)^{-1}\right).$$

Proof. See Appendix A.8. □

Although the function Γ equals the objective function of (1) based on Lemma 4, it is not convex, as noted below. To address this, we derive its convex envelope, denoted by $\operatorname{conv}\Gamma$, which provides the tightest convex underestimator of Γ (Rockafellar 1997). While Li and Xie (2024) derived the convex envelope $\operatorname{conv}\Gamma$ for the special case $\lambda = 0$ to obtain AMESP-R, our Proposition 3 below extends this result to the setting with $\lambda > 0$.

REMARK 2. *The function $\Gamma(\boldsymbol{X};\lambda) : \mathcal{S}_+^d \to \mathbb{R}$ is nonconvex.* ◇

Proof. See Appendix A.9. □

Next, we present a technical result of Nikolov (2015, Lemma 14). We will use the integer η throughout this section, with the vector $\boldsymbol{\sigma}$ being specified in context. We show that η remains unchanged under certain modifications to $\boldsymbol{\sigma}$. This invariance result plays a key role in deriving the convex envelope $\operatorname{conv}\Gamma$ and its subgradient.

LEMMA 5 (**The integer** η**, Nikolov (2015)**). *Let $\sigma_1 \geq \cdots \geq \sigma_d \geq 0$ be nonnegative reals. Then, there exists a unique integer η, $0 \leq \eta \leq \tilde{k} - 1$, such that $\sigma_\eta > \frac{1}{\tilde{k}-\eta} \sum_{i\in[\eta+1,d]} \sigma_i \geq \sigma_{\eta+1}$ with the convention $\sigma_0 = \infty$.*

LEMMA 6. *Let $s \in [d, n]$, and let $\sigma_1 \geq \cdots \geq \sigma_d \geq 0$ be nonnegative reals. Define a new vector $\boldsymbol{\sigma}^+ \in \mathbb{R}_+^s$ as*

$$\sigma_i^+ = \sigma_i + \lambda, \forall i \in [\tilde{k}], \ \ \sigma_i^+ = \sigma_i, \forall i \in [\tilde{k}+1, d], \ \ \sigma_i^+ = 0, \forall i \in [d+1, s].$$

Then, the unique index $0 \leq \eta \leq \tilde{k} - 1$, defined based on $\boldsymbol{\sigma}$ in Lemma 5, remains valid for $\boldsymbol{\sigma}^+$, i.e.,

$$\sigma_\eta^+ > \frac{1}{\tilde{k} - \eta} \sum_{i\in[\eta+1,s]} \sigma_i^+ \geq \sigma_{\eta+1}^+.$$

Proof. See Appendix A.10. □

PROPOSITION 3. *For a matrix $\boldsymbol{X} \in \mathcal{S}_+^d$ with eigenvalues $\sigma_1 \geq \cdots \geq \sigma_d \geq 0$, the following hold:*

(i) The convex envelope of the function Γ, denoted by $\mathrm{conv}\,\Gamma$, is defined as

$$\mathrm{conv}\,\Gamma(\boldsymbol{X}; \lambda) = \sum_{i\in[\eta]} \frac{1}{\sigma_i + \lambda} + \frac{(\tilde{k} - \eta)^2}{\sum_{i\in[\eta+1,d]} \sigma_i + (\tilde{k}-\eta)\lambda},$$

where $0 \leq \eta \leq \tilde{k} - 1$ is a unique integer from Lemma 5.

(ii) Let $\boldsymbol{X} = \boldsymbol{Q}\,\mathrm{Diag}(\boldsymbol{\sigma})\boldsymbol{Q}^\top$ denote the eigen-decomposition of $\boldsymbol{X}$, and let r denote its rank. Then, $\boldsymbol{Q}\,\mathrm{Diag}\,(\boldsymbol{\varsigma})\,\boldsymbol{Q}^\top$ is a subgradient of $\mathrm{conv}\,\Gamma(\boldsymbol{X}; \lambda)$ at $\boldsymbol{X}$, where

$$\varsigma_i = \frac{1}{\sigma_i + \lambda}, \forall i \in [\eta], \ \ \varsigma_i = \frac{\tilde{k} - \eta}{\sum_{j\in[\eta+1,d]} \sigma_j + (\tilde{k}-\eta)\lambda}, \forall i \in [\eta+1, r], \ \ \varsigma_i = \varsigma_{\eta+1}, \forall i \in [r+1, d].$$

Proof. See Appendix A.11. □

The functions $\mathrm{conv}\,\Gamma$ and Φ from Definition 2 share a similar form but differ notably in their domains and parameter settings.

(i) Unlike Φ, which is defined over $n \times n$ matrices and relies on the condition $k \leq n$, the function $\mathrm{conv}\,\Gamma$ builds on $d \times d$ matrices and the integer $\tilde{k} \leq d$;

(ii) In addition, $\mathrm{conv}\,\Gamma$ includes a regularization parameter $\lambda > 0$, which Φ does not have; and

(iii) When using Φ to relax RAOD, the resulting relaxation AMESP-R incorporates a constant term $(d - k)/\lambda$ in its objective function. If $d > k$, that term becomes negative, which may result in a negative lower bound, as demonstrated in Theorem 3. In contrast, the relaxation RAOD-RII, which is based on $\mathrm{conv}\,\Gamma$, always incorporates a nonnegative constant $(d - \tilde{k})/\lambda$. Hence, $\mathrm{conv}\,\Gamma$ can prevent the relaxation from yielding negative lower bounds. More importantly, we later demonstrate that RAOD-RII dominates AMESP-R.

We establish below a connection between the function $\operatorname{conv}\Gamma$ and the trace of the matrix inverse, which lays the foundation for our subsequent results.

LEMMA 7. *If the rank of $\boldsymbol{X}$ is bounded by $\tilde{k}$, then we have*

$$\operatorname{conv}\Gamma(\boldsymbol{X};\lambda) + \frac{d - \tilde{k}}{\lambda} = \operatorname{tr}((\boldsymbol{X} + \lambda\boldsymbol{I}_d)^{-1}).$$

Proof. See Appendix A.12. □

Using Lemma 7, we are now ready to introduce a novel convex integer program for RAOD.

THEOREM 4. *RAOD is equivalent to the following convex integer program:*

$$z_k := \min_{\boldsymbol{x}\in\{0,1\}^n} \left\{ \operatorname{conv}\Gamma\bigg(\sum_{i\in[n]} x_i\boldsymbol{a}_i\boldsymbol{a}_i^\top ; \lambda\bigg) + \frac{d - \tilde{k}}{\lambda} : \sum_{i\in[n]} x_i = k \right\}. \tag{3}$$

Proof. Let $\boldsymbol{X} = \sum_{i\in[n]} x_i\boldsymbol{a}_i\boldsymbol{a}_i^\top$. For any binary vector $\boldsymbol{x}\in\{0,1\}^n$ with $\sum_{i\in[n]} x_i = k$, the rank of $\boldsymbol{X}$ must be bounded by $\tilde{k}$. According to Lemma 7, the objective functions of (1) and (3) must be equal in this context. □

4.2. A Novel Convex Relaxation

By relaxing the binary variables $\boldsymbol{x}$ of (3), we therefore obtain a new lower bound of RAOD:

$$z_k \geq \nu_k^{\mathrm{N}} := \min_{\boldsymbol{x}\in[0,1]^n} \left\{ \operatorname{conv}\Gamma\bigg(\sum_{i\in[n]} x_i\boldsymbol{a}_i\boldsymbol{a}_i^\top ; \lambda\bigg) + \frac{d - \tilde{k}}{\lambda} : \sum_{i\in[n]} x_i = k \right\}. \tag{RAOD-RII}$$

By leveraging Part (ii) of Proposition 3, we can readily obtain the (sub)gradient of the objective function of RAOD-RII, which enables us to use first-order methods, such as the Frank-Wolfe algorithm, to solve RAOD-RII efficiently.

REMARK 3. For any $\boldsymbol{x}\in[0,1]^n$ with $\sum_{i\in[n]} x_i = k$, define the matrix $\boldsymbol{X}(\boldsymbol{x}) = \sum_{i\in[n]} x_i\boldsymbol{a}_i\boldsymbol{a}_i^\top$. Suppose $\boldsymbol{W}(\boldsymbol{x})\in\mathcal{S}_+^n$ is a subgradient of $\operatorname{conv}\Gamma(\boldsymbol{X};\lambda)$ at $\boldsymbol{X}(\boldsymbol{x})$, as introduced in Proposition 3. Then, the vector $\boldsymbol{g}(\boldsymbol{x})\in\mathbb{R}_+^n$, defined by

$$g_i(\boldsymbol{x}) := \boldsymbol{a}_i^\top \boldsymbol{W}(\boldsymbol{x})\boldsymbol{a}_i, \ \ \forall i\in[n],$$

is a subgradient of $\operatorname{conv}\Gamma(\sum_{i\in[n]} x_i\boldsymbol{a}_i\boldsymbol{a}_i^\top;\lambda)$ with respect to $\boldsymbol{x}$. ◇

In the following, we compare the three convex relaxations of RAOD and analyze their relationships by leveraging Lemmas 6 and 7 and a technical result from Li (2024, Lemma 1).

THEOREM 5 (**Comparison of convex relaxations**). *RAOD-RII outperforms both RAOD-RI and AMESP-R. More precisely, the optimal values of these formulations satisfy*

(i) $\nu_k^N = \nu_k^C$ *for any* $k \in [d, n]$;

(ii) $\nu_k^N \geq \nu_k^C$ *for any* $k \in [d-1]$*; and*

(iii) $\nu_k^N \geq \nu_k^M$ *for any* $k \in [n]$.

Proof. See Appendix A.13. □

Based on Theorem 5, we note that RAOD-RII offers several key advantages:

(i) It provides a stronger lower bound than RAOD-RI and AMESP-R, effectively combining their strengths. Thus, we can directly combine Theorems 2 and 3 to guarantee theoretical gaps of RAOD-RII for all $k \in [n]$, as summarized in Corollary 2;

(ii) For $k \in [d, n]$, Parts (i) and (iii) of Theorem 5 imply that RAOD-RI outperforms AMESP-R. This allows us to apply Theorem 3 to refine the gap of RAOD-RI at $k = d$ in Corollary 3; and

(iii) It mitigates the computational limitations of AMESP-R by operating in a $d \times d$ matrix space, achieving similar efficiency as RAOD-RI.

COROLLARY 2 (**Optimality gap**). *RAOD-RII admits the following optimality gap*

$$1 \leq \frac{z_k}{\nu_k^N} \leq \min\left\{\frac{d-1}{d-k}, n-k+1\right\}, \quad \forall k \in [d-1], \quad 1 \leq \frac{z_k}{\nu_k^N} \leq \min\{d, n-d+1\}, \quad \textit{if } k = d,$$

$$\textit{and} \quad 1 \leq \frac{z_k}{\nu_k^N} \leq \frac{n-d+1}{k-d+1}, \quad \forall k \in [d+1, n].$$

To better understand the performance of RAOD-RII, we illustrate its optimality gaps in Figure 1. We see that the gap reaches its maximum when k approaches d from either side and tends to decrease as d increases for small values of k.
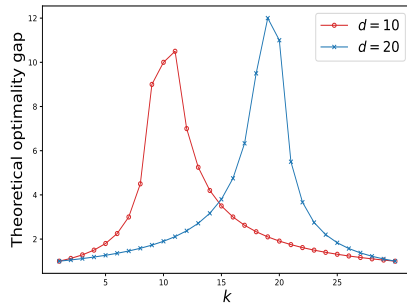
(a) $n = 30$
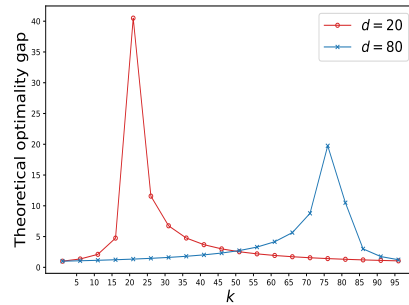
(b) $n = 100$

Figure 1 The optimality gaps of RAOD-RII in Corollary 2

COROLLARY 3 (**Optimality gap**). *When $k = d$, RAOD-RI admits the following optimality gap*

$$1 \leq \frac{z_k}{\nu_k^C} \leq \min\{d, n - d + 1\}.$$

4.3. The Cutting-Plane Algorithm

Solving RAOD to optimality is computationally expensive. To mitigate this challenge, this subsection presents an exact algorithm designed to enhance existing B&B methods, which typically use the classic lower bound from RAOD-RI (see, e.g., Hendrych et al. 2023). Instead, our approach leverages the formulation (3) with a stronger relaxation, RAOD-RII. Motivated by the convexity of the function $\mathrm{conv}\,\Gamma$, we can recast (3) as a mixed-integer linear programming (MILP) formulation (4). This reformulation is a well-established technique for mixed-integer nonlinear optimization problems (Duran and Grossmann 1986, Quesada and Grossmann 1992).

$$z_k = \min_{w \in \mathbb{R}^+, \boldsymbol{x} \in \{0,1\}^n} \left\{ w : \sum_{i \in [n]} x_i = k, w \geq \mathrm{conv}\,\Gamma\left(\sum_{i \in [n]} y_i \boldsymbol{a}_i \boldsymbol{a}_i^\top; \lambda\right) + \frac{d - \tilde{k}}{\lambda} \right.$$
$$\left. + \boldsymbol{g}(\boldsymbol{y})^\top (\boldsymbol{x} - \boldsymbol{y}), \ \ \forall \boldsymbol{y} \in \{0,1\}^n, \sum_{i \in [n]} y_i = k \right\}, \tag{4}$$

where $\boldsymbol{g}(\boldsymbol{y}) \in \mathbb{R}^n$ denotes a (sub)gradient of $\mathrm{conv}\,\Gamma(\sum_{i \in [n]} y_i \boldsymbol{a}_i \boldsymbol{a}_i^\top; \lambda)$ at $\boldsymbol{y}$, as defined in Remark 3.

However, the problem (4) incorporates an exponentially large number of linear inequalities, which causes computational challenges. To address this, we design a customized cutting plane algorithm for (4) that iteratively adds linear inequalities. The pseudo-code is provided in Algorithm 1. This outer-approximation scheme, as introduced in the seminal work of Duran and Grossmann (1986), was proven to converge optimally for solving MINLPs.

At each iteration of Algorithm 1, a MILP problem over $(w, \boldsymbol{x})$, called the *master* problem, is solved. This subproblem of (4) is based on a subset of linear inequalities. As the number of linear cuts increases, the solution values of w^t form a non-decreasing sequence of lower bounds for RAOD. Algorithm 1 terminates when the relative gap between this lower bound (i.e., w^t) and the best-known upper bound (i.e., UB) reaches a fixed tolerance ϵ. We choose $\epsilon = 10^{-4}$ in our experiments. According to Bertsimas et al. (2020), Algorithm 1 can be efficiently implemented in advanced MILP packages (e.g., CPLEX, Gurobi) with the use of *lazy* constraints and *callback* functions. However, a time limit is necessary to impose on Algorithm 1, especially when solving large-sized instances. Indeed, as often in mixed-integer optimization, the cutting plane algorithm can quickly find a high-quality solution, but proving its optimality can take much longer. In our experiments, we set a time limit of one hour for Algorithm 1.

Algorithm 1 The cutting-plane algorithm for (4)

Input: Vectors $\{\boldsymbol{a}_i \in \mathbb{R}^d\}_{i \in [n]}$, a constant $\lambda > 0$, an integer $k \in [n]$, and target optimality $\epsilon > 0$

Initialize $t = 1$, $\boldsymbol{x}^t \in \{0,1\}^n$ with $\sum_{i \in [n]} x_i^t = k$ and UB $= \text{conv}\,\Gamma\left(\sum_{i \in [n]} x_i^t \boldsymbol{a}_i \boldsymbol{a}_i^\top; \lambda\right) + \frac{d - \tilde{k}}{\lambda}$

repeat

Solve the following optimization problem, where $(w^{t+1}, \boldsymbol{x}^{t+1})$ denotes an optimal solution

$$\min_{\substack{w \in \mathbb{R}_+, \boldsymbol{x} \in \{0,1\}^n \\ \sum_{i \in [n]} x_i = k}} \left\{ w : w \geq \text{conv}\,\Gamma\left(\sum_{i \in [n]} x_i^\ell \boldsymbol{a}_i \boldsymbol{a}_i^\top; \lambda\right) + \frac{d - \tilde{k}}{\lambda} + \boldsymbol{g}\left(\boldsymbol{x}^\ell\right)^\top \left(\boldsymbol{x} - \boldsymbol{x}^\ell\right), \forall \ell \in [t] \right\}$$

Update UB $= \min\left\{\text{UB}, \text{conv}\,\Gamma\left(\sum_{i \in [n]} x_i^{t+1} \boldsymbol{a}_i \boldsymbol{a}_i^\top; \lambda\right) + \frac{d - \tilde{k}}{\lambda}\right\}$

Update $t = t + 1$

until $(w^t - \text{UB})/\text{UB} \leq \epsilon$

Output: $\boldsymbol{x}^t$ and w^t

We employ several strategies to accelerate the computation of Algorithm 1 in practice. First, we enhance the algorithm using a high-quality warm start obtained from our approximation algorithms in Section 5, instead of randomly initializing a feasible solution. Second, a variable-fixing technique, proposed by Li (2024), is applied to identify some binary variables in (3) that take values of 1 or 0 at optimality. Fixing these variables can reduce the problem size of RAOD while preserving the optimal solution. The effectiveness of this variable-fixing process depends on the strength of the convex relaxations used. Notably, the use of RAOD-RII allows for the fixing of more binary variables. Finally, we solve RAOD-RII before running Algorithm 1 to obtain an initial lower bound. This helps reduce the root gap.

5. The Forward and Backward Greedy Algorithms

In this section, we investigate the scalable forward and backward greedy search strategies and derive their approximation ratios tailored for two different ranges of k in RAOD, specifically for $k \in [d-1]$ and $k \in [d, n]$, respectively.

The forward greedy algorithm begins with an empty set and then iteratively selects the best new data point to add to the set. This process continues until the set reaches a size of k. Using the Sherman–Morrison formula, Chamon and Ribeiro (2017b) designed an efficient implementation of forward greedy for solving RAOD, as detailed in Algorithm 2.

The backward greedy algorithm, in contrast to the forward greedy approach, starts with a solution of selecting all data points and then removes an element that results in the smallest objective

---

Algorithm 2 The forward greedy algorithm for RAOD (Chamon and Ribeiro 2017a)

Input: Vectors $\{\boldsymbol{a}_i\}_{i\in[n]} \in \mathbb{R}^d$, a constant $\lambda > 0$, and an integer $k \in [n]$

Initialize $S_F = \emptyset$ and $\boldsymbol{\Lambda} = \frac{1}{\lambda}\boldsymbol{I}_d$

for $j \in [k]$ **do**

    Compute $i^* = \arg\max_{i\in[n]\backslash S_F} \frac{\boldsymbol{a}_i^\top \boldsymbol{\Lambda}\boldsymbol{\Lambda}\boldsymbol{a}_i}{1+\boldsymbol{a}_i^\top \boldsymbol{\Lambda}\boldsymbol{a}_i}$

    Update $S_F = S_F \cup \{i^*\}$ and $\boldsymbol{\Lambda} = \boldsymbol{\Lambda} - \frac{\boldsymbol{\Lambda}\boldsymbol{a}_{i^*}\boldsymbol{a}_{i^*}^\top\boldsymbol{\Lambda}}{1+\boldsymbol{a}_{i^*}^\top\boldsymbol{\Lambda}\boldsymbol{a}_{i^*}}$

end for

Output: S_F

---

value at each iteration. The procedure repeats until only k data points are left, which is presented in Algorithm 3. Specifically, at each iteration of Algorithm 3, we have that $\boldsymbol{\Lambda} = (\sum_{j\in S_B} \boldsymbol{a}_j\boldsymbol{a}_j^\top + \lambda\boldsymbol{I}_d)^{-1}$, and the goal is to solve

$$\min_{i\in S_B} f(S_B \backslash \{i\}) = \min_{i\in S_B} \operatorname{tr}\left(\boldsymbol{\Lambda} + \frac{\boldsymbol{\Lambda}\boldsymbol{a}_i\boldsymbol{a}_i^\top\boldsymbol{\Lambda}}{1-\boldsymbol{a}_i^\top\boldsymbol{\Lambda}\boldsymbol{a}_i}\right) = f(S_B) + \min_{i\in S_B}\frac{\boldsymbol{a}_i^\top\boldsymbol{\Lambda}\boldsymbol{\Lambda}\boldsymbol{a}_i}{1-\boldsymbol{a}_i^\top\boldsymbol{\Lambda}\boldsymbol{a}_i},$$

where the first equation results from the Sherman–Morrison formula. Similarly, we update the matrix $\boldsymbol{\Lambda}$ by the Sherman–Morrison formula.

Algorithm 3 The backward greedy algorithm for RAOD

Input: Vectors $\{\boldsymbol{a}_i\}_{i\in[n]} \in \mathbb{R}^d$, a constant $\lambda > 0$, and an integer $k \in [n]$

Initialize $S_B = [n]$ and $\boldsymbol{\Lambda} = (\sum_{j\in S_B} \boldsymbol{a}_j\boldsymbol{a}_j^\top + \lambda\boldsymbol{I}_d)^{-1}$

for $j \in [n-k]$ **do**

    Compute $i^* = \arg\min_{i\in S_B} \frac{\boldsymbol{a}_i^\top \boldsymbol{\Lambda}\boldsymbol{\Lambda}\boldsymbol{a}_i}{1-\boldsymbol{a}_i^\top \boldsymbol{\Lambda}\boldsymbol{a}_i}$

    Update $S_B = S_B \backslash \{i^*\}$ and $\boldsymbol{\Lambda} = \boldsymbol{\Lambda} + \frac{\boldsymbol{\Lambda}\boldsymbol{a}_{i^*}\boldsymbol{a}_{i^*}^\top\boldsymbol{\Lambda}}{1-\boldsymbol{a}_{i^*}^\top\boldsymbol{\Lambda}\boldsymbol{a}_{i^*}}$

end for

Output: S_B

Using Lemma 2, we guarantee the theoretical performance for Algorithms 2 and 3.

THEOREM 6 (**Approximation ratios and complexities**). *For $k \in [d-1]$, Algorithm 2 enjoys a-$\frac{d-1}{d-k}$ approximation ratio, that is,*

$$f(S_F) \leq \frac{d-1}{d-k} z_k.$$

For $k \in [d, n]$, Algorithm 3 enjoys a-$\frac{n-d+1}{k-d+1}$ approximation ratio, that is,

$$f(S_B) \leq \frac{n-d-1}{k-d+1} z_k.$$

In addition, the running time complexities of Algorithms 2 and 3 are $\mathcal{O}(knd^2)$ and $\mathcal{O}((n-k)nd^2)$, respectively.

Proof. See Appendix A.14. □

We remark on Theorem 6 that

(i) The forward greedy Algorithm 2 has been extensively studied in the Bayesian AOD literature (see, Chamon and Ribeiro 2017a, Bian et al. 2017). Theorem 6 offers the first-known data-independent approximation ratio for Algorithm 2 when applied to RAOD with $k \in [d-1]$;

(ii) We develop the novel backward greedy Algorithm 3 that delivers an approximation ratio when $k \in [d, n]$;

(iii) As a classical intuition, forward greedy is effective when selecting a small number of data points, while backward greedy becomes more efficient as the selection size grows. Theorem 6 rigorously validates this intuition, as it offers theoretical guarantees for Algorithms 2 and 3 in the small- and large-k ranges, respectively, as illustrated in Figure 2; and

(iv) Finally, we consider Algorithm 4 to select the better solution between the outputs of Algorithms 2 and 3, which yields a high-quality solution to RAOD for all values of k.

(a) $d = 10$　　　　　　　　　　(b) $d = 20$
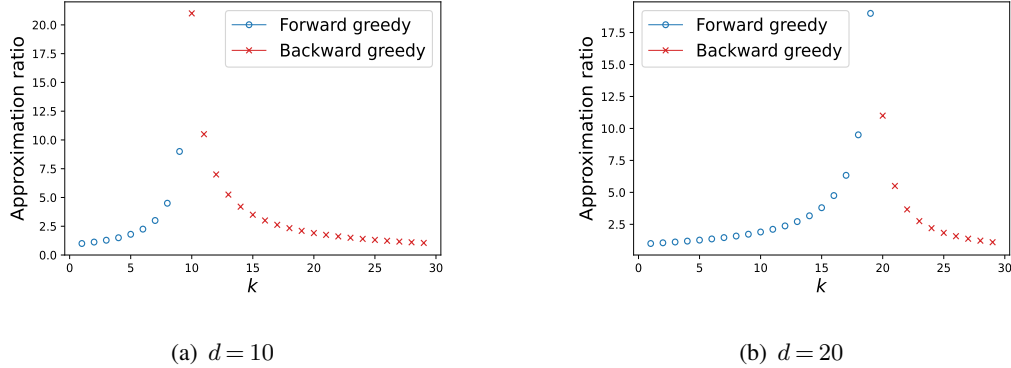
Figure 2　The approximation ratios of Algorithms 2 and 3 in Theorem 6 with $n = 30$

Algorithm 4 The combination of the forward and backward greedy algorithms for RAOD

Input: Vectors $\{\boldsymbol{a}_i\}_{i \in [n]} \in \mathbb{R}^d$, a constant $\lambda > 0$, and an integer $k \in [n]$

Denote by S_F the output of Algorithm 2

Denote by S_B the output of Algorithm 3

Output: The better solution between S_F and S_B

COROLLARY 4. *Algorithm 4 has an approximation ratio of $\frac{d-1}{d-k}$ for $k \in [d-1]$, and $\frac{n-d+1}{k-d+1}$ for $k \in [d, n]$. In addition, its time complexity is $\mathcal{O}(n^2 d^2)$.*

6. Numerical Results

This section evaluates the numerical performance of our formulations and algorithms developed for solving RAOD. The computational results not only demonstrate the efficiency of our approaches but also provide empirical support for our theoretical findings. All the experiments enforce a 1-hour timeout, and are conducted in Python 3.6 with calls to Gurobi 9.5.2 on a PC with 10-core CPU, 16-core GPU, and 16GB of memory. We begin by introducing the datasets, both synthetic and real, used in our computational experiments.

Synthetic data. Given (n, d), we randomly generate a data matrix $\boldsymbol{A} \in \mathbb{R}^{d \times n}$ by a normal distribution $\mathcal{N}(0, 1)$, and let $\boldsymbol{a}_i$ be the i-th column of $\boldsymbol{A}$ for each $i \in [n]$.

UCI data (Blake 1998). We further conduct experiments on five UCI datasets- *autos*, *breast-cancer*, *sml*, *gas*, and *song*, with dimensions (n, d) of $(159, 24)$, $(194, 33)$, $(200, 22)$, $(500, 128)$, and $(1000, 90)$, respectively. We begin by removing all-zero columns from the data matrix. For datasets with a huge number of observations, we randomly sample a manageable subset of n data points to ensure computational feasibility when testing exact methods. Finally, we apply min-max normalization to each feature (i.e., each row) of the resulting matrix $\boldsymbol{A} \in \mathbb{R}^{d \times n}$, so that it falls within the range $[0, 1]$.

Movie rating data (Dooms et al. 2013). The movie rating dataset is used to demonstrate an application of RAOD to the new user cold-start problem in recommendation systems. The dataset consists of ratings provided by 71,707 users for a total of 38,018 movies. Ratings are shifted from the range [0, 10] to [1, 11] by adding one to each rating to avoid zero vectors.

6.1. Evaluations of Convex Relaxations

This subsection evaluates the three convex relaxations: RAOD-RI, AMESP-R, and RAOD-RII, by comparing their optimality gaps and computation times with varying-scale instances. We use the Frank-Wolfe algorithm to solve them. The results are summarized in Figures 3 and 4. For each convex relaxation, we compute "Gap (%)" by $100 \times (\text{upper bound} - \text{relaxation bound})/\text{upper bound}$, where the upper bound of RAOD is obtained from Algorithm 4. Our results verify the effectiveness of RAOD-RII and its dominance over RAOD-RI and AMESP-R.

We first test three synthetic datasets with $n = 100$ and different values of d, λ, where for each dataset, we consider $k \in \{5, 10, \cdots, 95\}$. As Figure 3 below shows, RAOD-RII delivers the tightest

lower bound. It is worth noting that the gap curves for RAOD-RII in Figures 3(a) to 3(c) closely match the theoretical gaps shown in Figure 1, where the maximum occurs near $k = d$. Besides, we would like to comment on the gaps of RAOD-RI and AMESP-R in two ranges of k based on Figures 3(a) to 3(c). We see that for $k < d$, AMESP-R outperforms RAOD-RI. When $k \geq d+1$, RAOD-RI and RAOD-RII are identical, and they dominate AMESP-R, as demonstrated in Theorem 5. In fact, AMESP-R produces extremely large gaps for $k \in [d+1, n]$, often exceeding 100%, which indicates that it returns a negative lower bound. This phenomenon aligns with our result in Part (ii) of Theorem 3. As seen from the differences between Figures 3(a) and 3(b), increasing d tends to improve AMESP-R and RAOD-RII but worsen RAOD-RI, especially for small-k instances. Finally, we find that AMESP-R requires the largest computation time since it builds on the $n \times n$ matrix space.

In Figure 4, we test the relaxations of RAOD on three larger-sized real datasets ($n = 159, 200, 1000$) with $\lambda = 1$. The values of k are tested in increments of 5 for $n = 159, 200$, and larger increments of 50 for $n = 1000$. As before, RAOD-RII wins on all test instances. We observe in Figures 4(a) to 4(c) that as $k \to d$ from below, RAOD-RII strictly outperforms RAOD-RI and AMESP-R even if we combine them and take the minimum. It is still obvious that the gap pf RAOD-RII tends to be larger when k approaches d from either side. Interestingly, as we observe in Figures 4(a) and 4(b), RAOD-RI and AMESP-R are no longer directly comparable for values of $k < d$. For the largest test case in Figure 4(f), where $n = 1000$, AMESP-R hits the time limit, highlighting its scalability limitations.

6.2. Evaluations of Approximation Algorithms

In this subsection, we evaluate the performance of several approximation algorithms for solving RAOD on the same data instances. Specifically, we compare our Algorithms 2 and 3 with the following methods: "LocalSearch," referring to the local search algorithm proposed by Li and Xie (2024) for AMESP; "RegVol," the regularized volume sampling algorithm introduced by Dereziński and Warmuth (2017); and its more efficient implementation "FastRegVol." Local search starts from a random solution, and sampling algorithms are run 20 times per instance, with the best result reported. Notably, we omit comparisons with the randomized sampling algorithms from Li and Xie (2024), Tantipongpipat (2020), as they were computationally intensive and failed to finish within one hour on these instances.

(a) Gap for $(n, d, \lambda) = (100, 20, 1)$

(b) Gap for $(n, d, \lambda) = (100, 80, 1)$

(c) Gap for $(n, d, \lambda) = (100, 20, 10)$

(d) Time for $(n, d, \lambda) = (100, 20, 1)$

(e) Time for $(n, d, \lambda) = (100, 80, 1)$

(f) Time for $(n, d, \lambda) = (100, 20, 10)$
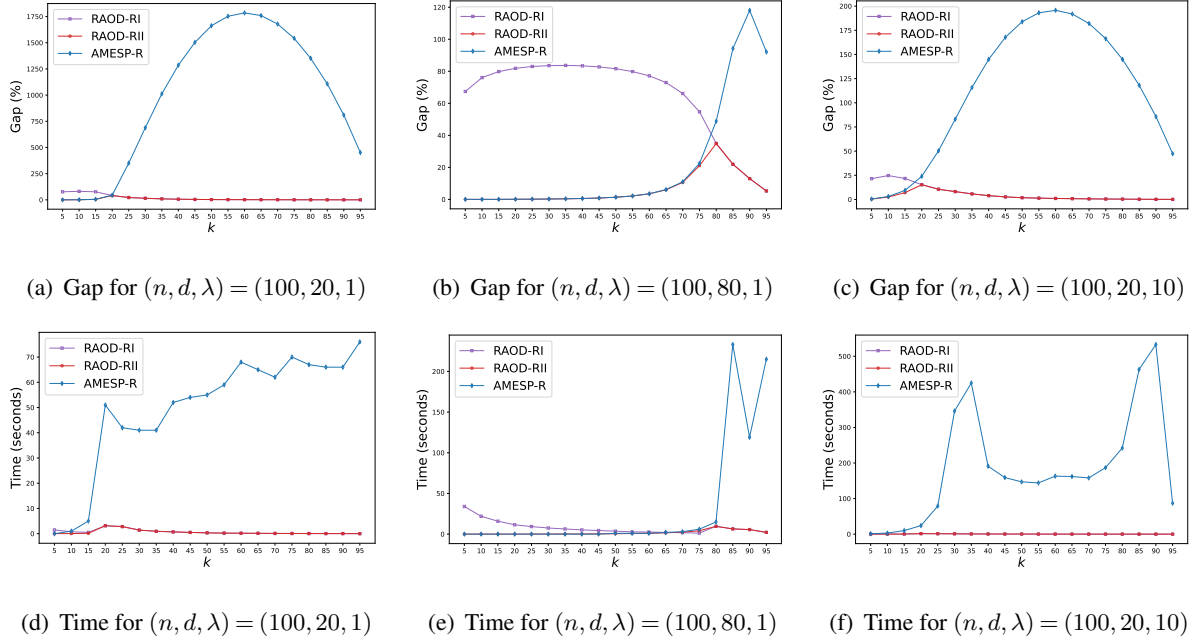
Figure 3   Convex relaxations of RAOD: Gaps and computation times on synthetic data

(a) Gap for $(n, d, \lambda) = (159, 24, 1)$

(b) Gap for $(n, d, \lambda) = (200, 22, 1)$

(c) Gap for $(n, d, \lambda) = (1000, 90, 1)$

(d) Time for $(n, d, \lambda) = (159, 24, 1)$

(e) Time for $(n, d, \lambda) = (200, 22, 1)$

(f) Time for $(n, d, \lambda) = (1000, 90, 1)$
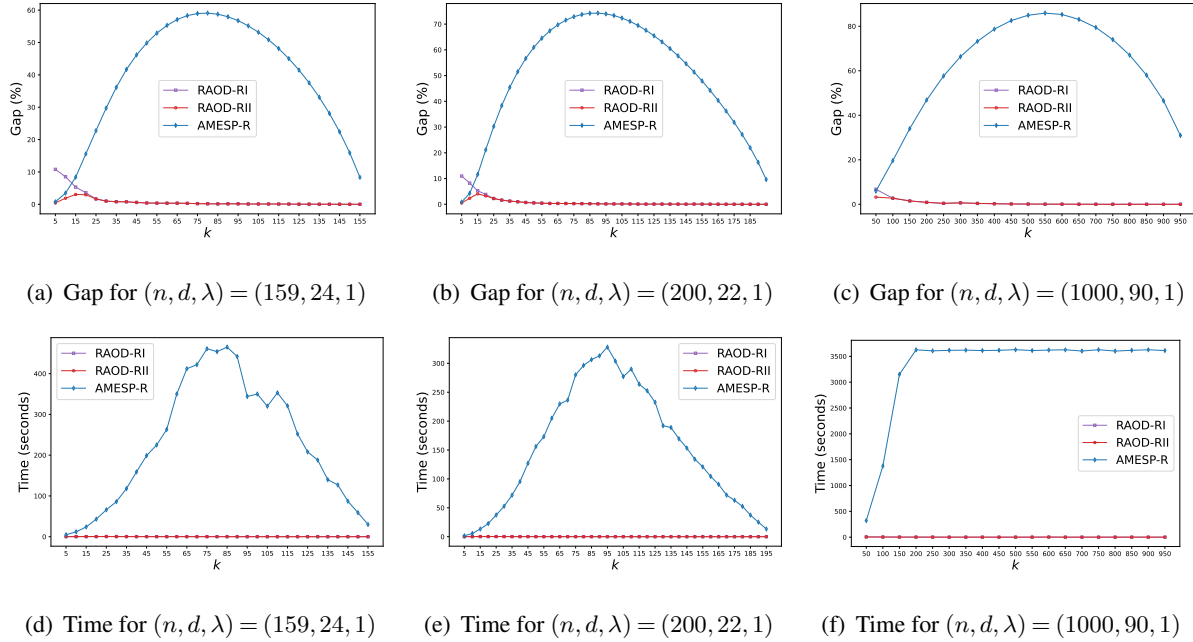
Figure 4   Convex relaxations of RAOD: Gaps and computation times on real data

The approximation gaps and computation times for the evaluated algorithms are summarized in Figures 5 and 6. We track the best lower bound ν_k^N found from the previous subsection, and compute "Gap (%)" by $100 \times$ (upper bound $- \nu_k^N$)/upper bound, where the upper bound is the objective value returned by each approximation method. We observe that combinatorial algorithms

outperform the two sampling algorithms, with significant improvements when $k > d$ and d is relatively small compared to n. The comparison of Figures 5(d) and 5(e) shows that the runtime of FastRegVol increases more rapidly as d increases, compared to the other algorithms. Across all algorithms, the largest approximation gaps tend to occur near $k = d$. This may result from using a lower bound ν_k^N rather than the true optimal value in computing the gap, as RAOD-RII deviates most from the optimum around $k = d$ (see Figures 3 and 4). In general, the three combinatorial algorithms are not directly comparable, but the local search algorithm outperforms our greedy Algorithms 2 and 3 in more cases. However, for the largest instance with $n = 1000$ in Figures 6(c) and 6(f), it takes considerably longer but achieves nearly identical gaps to our algorithms. In summary, our Algorithms 2 and 3 strike the best balance between scalability and solution quality among all the algorithms. As seen in Figure 6(b), Algorithm 2 yields a smaller gap when k is small, while Algorithm 3 becomes more effective when $k \geq d$. This trend aligns with their theoretical guarantees discussed following Theorem 6.

(a) Gap for $(n, d, \lambda) = (100, 20, 1)$ (b) Gap for $(n, d, \lambda) = (100, 80, 1)$ (c) Gap for $(n, d, \lambda) = (100, 20, 10)$

(d) Time for $(n, d, \lambda) = (100, 20, 1)$ (e) Time for $(n, d, \lambda) = (100, 80, 1)$ (f) Time for $(n, d, \lambda) = (100, 20, 10)$
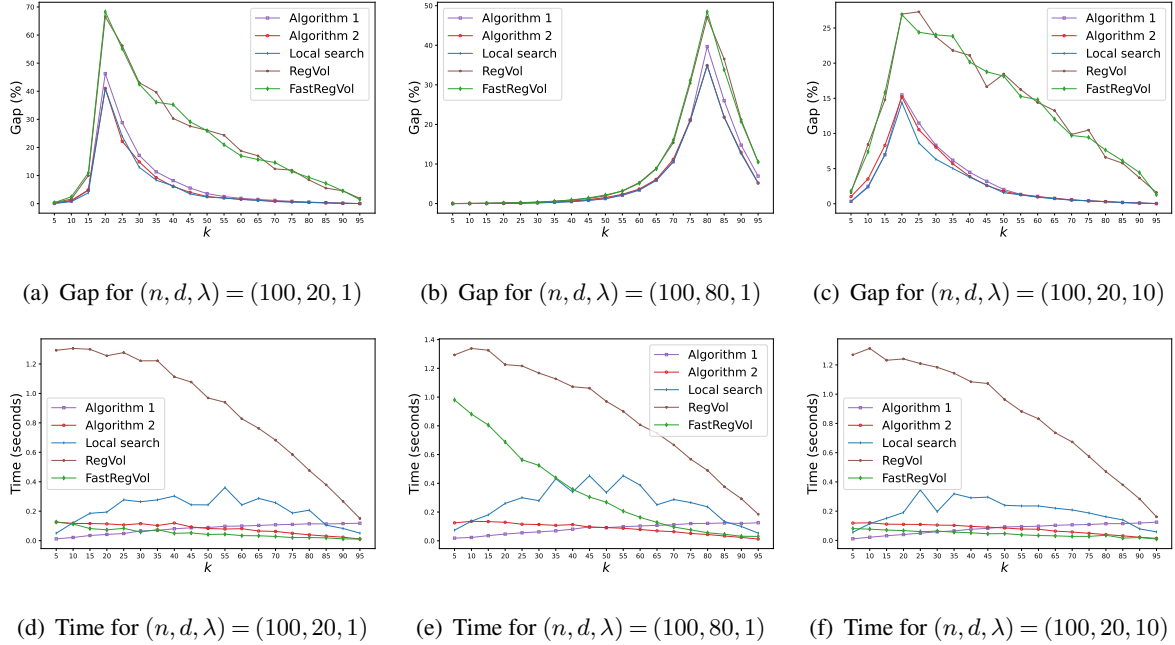
Figure 5   Approximation algorithms of RAOD: Gaps and computation times on synthetic data

6.3. Performance of Exact Algorithms

We now compare the three (mixed-)integer convex formulations introduced in Sections 3 and 4 using both synthetic and real datasets. To solve our proposed formulation (3), we develop the
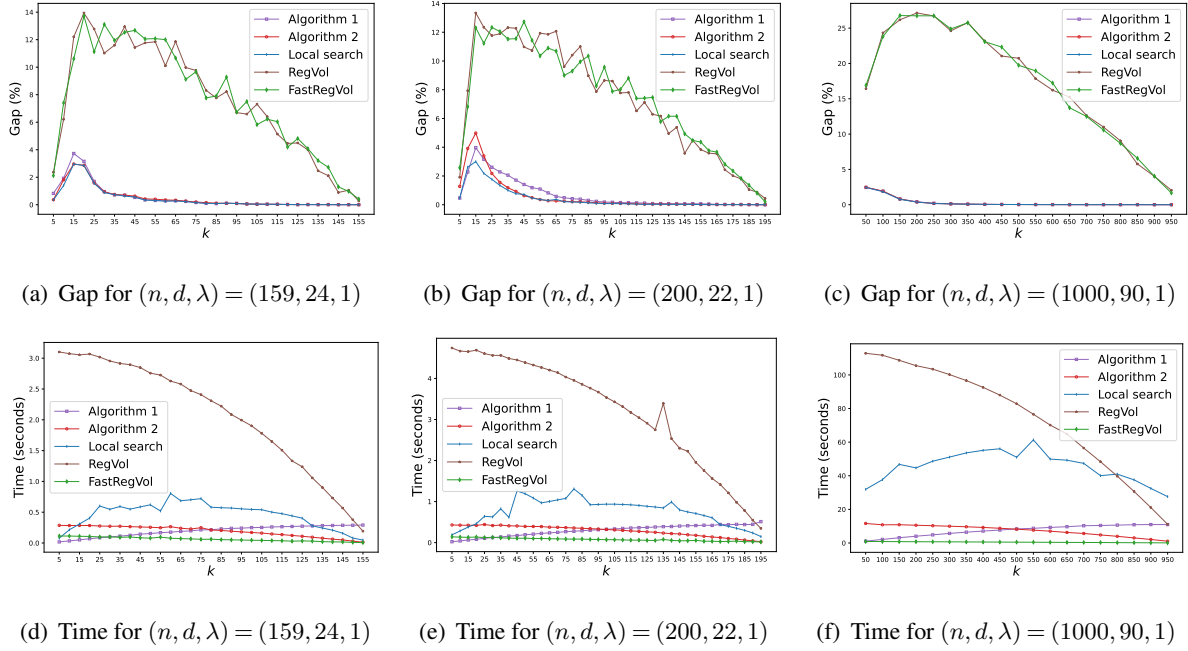
(a) Gap for $(n, d, \lambda) = (159, 24, 1)$ (b) Gap for $(n, d, \lambda) = (200, 22, 1)$ (c) Gap for $(n, d, \lambda) = (1000, 90, 1)$

(d) Time for $(n, d, \lambda) = (159, 24, 1)$ (e) Time for $(n, d, \lambda) = (200, 22, 1)$ (f) Time for $(n, d, \lambda) = (1000, 90, 1)$

Figure 6   Approximation Algorithms of RAOD: Gaps and computation times on real data

cutting-plane Algorithm 1, which we implement using Gurobi's *lazy* constraints. The same cutting-plane scheme can be readily extended to (1), since its objective function is also convex; we refer to this variant as "Cutting plane via (1)." For fairness, both algorithms employ the same computational enhancements and termination criterion, as detailed in Subsection 4.3. In addition, we use Gurobi to solve the formulation MISOCP directly. Although AMESP-R offers an exact formulation when $\boldsymbol{x}$ is binary, we do not include it in our comparison due to its high computational cost and the fact that RAOD-RII theoretically dominates AMESP-R (see Theorem 5, Part (iii)).

We restrict our experiments to the case $k < d$. From a theoretical perspective, the objective functions of (1) and (3) are equivalent when $k \geq d$, as indicated by Lemma 7, making the comparison trivial. From a practical perspective, this choice is also well-motivated. In high-dimensional application problems such as user cold-start recommendation systems, constraints such as limited budgets or user cognitive costs naturally enforce small values of k (Chamon and Ribeiro 2017a).

For each instance with the fixed pair (n, d), we vary the parameters $\lambda \in \{1, 10\}$ and $k \in \{5, 10\}$, thus creating four testing cases. The numerical results are presented in Tables 3 and 4, where "MIPGap(%)" represents the relative optimality gap in percentage at which Gurobi terminates. A MIPGap of 0.00 indicates a certified optimal solution. Among the three exact algorithms, the cutting-plane algorithm based on (1) attains the largest MIPGaps in most instances and only occasionally reaches optimality when $\lambda = 10$, as seen in Table 4. MISOCP also struggles with large

MIPGaps and timeouts. In particular, for instances with $n \geq 500$, MISOCP may even exhaust available memory, marked by "–" in Table 4. In contrast, our proposed Algorithm 1 achieves significantly smaller MIPGaps and solves most cases to optimality within 1 hour, especially for $k = 5$. These results show the superiority of Algorithm 1 in both computational efficiency and MIPGap reduction for the tested instances, which can be primarily attributed to the strength of the continuous relaxation of (3). For small $k < d$, its relaxation RAOD-RII is strictly tighter than RAOD-RI, the continuous relaxation of (1) and MISOCP.

Next, we would like to comment on the effect of parameters d, k, and λ. What is interesting in Table 3 is that, for fixed n, k, λ, the MIPGap for Algorithm 1 decreases as the experimental dimension d increases. For instance, Algorithm 1 solves the case with $(n, d, k, \lambda) = (100, 80, 10, 1)$ in just 7 seconds, whereas the lower-dimensional case $(100, 20, 10, 1)$ takes longer than 1 hour. A likely explanation is that, according to Corollary 2 and the discussion following Theorem 3, RAOD-RII admits an optimality gap of $(d - 1)/(d - k)$ when $k \leq (n + 1)/2$. Since this theoretical gap is decreasing with d (for fixed k), RAOD-RII becomes tighter in higher dimensions, thus enhancing the performance of Algorithm 1. This observation highlights the advantage of our Algorithm 1 in complex, high-dimensional problems. As expected, instances with $k = 10$ are harder to solve than those with $k = 5$. We also observe that "Cutting plane via (1)" and MISOCP benefit from higher λ but still fail to close the gap in many instances. In contrast, the effect of λ on the performance of Algorithm 1 is instance-dependent. Table 3 shows that the case with $\lambda = 10$ poses larger MIPGaps for Algorithm 1 than the case with $\lambda = 1$, whereas in Table 4, the opposite holds.

Finally, we evaluate the approximation quality of Algorithm 4 by comparing its outputs with the best lower bounds returned by Algorithm 1. As shown in the "Gap(%)" column of Tables 3 and 4, this gap remains within 2.45% and is computed in under 11 seconds, which demonstrates both the scalability and near-optimality of Algorithm 4. Moreover, in cases where this gap is smaller than the MIPGap of Algorithm 1, it indicates that Algorithm 1 failed to find a better feasible solution within the time limit than the one produced by Algorithm 4.

6.4. Application to User Cold-Start Recommendation

This subsection presents an application of RAOD to the user cold-start recommendation problem introduced in Section 1. Recall that the system struggles with predicting preferences for new users without prior ratings. A common solution is to select a small, representative set of items for new users to rate. The criterion for selection usually relies on a statistical measure aimed at minimizing

Table 3   Exact algorithms for RAOD on synthetic data

Dataset	Params		Cutting plane via (1)		MISOCP		Algorithm 1 via (3)		Algorithm 4	
(n, d)	λ	k	MIPGap(%)	Time	MIPGap(%)	Time	MIPGap(%)	Time	Gap(%)	Time
(100, 20)	1	5	77.92	3600	76.08	3600	0.00	589	0.02	1
	1	10	82.54	3600	80.14	3600	0.82	3600	0.88	1
	10	5	21.17	3600	18.15	3600	0.00	6	0.04	1
	10	10	27.70	3600	22.63	3600	2.65	3600	2.45	1
(100, 50)	1	5	75.43	3600	73.18	3600	0.00	1	0.00	1
	1	10	83.57	3600	81.80	3600	0.04	3600	0.03	1
	10	5	25.80	3600	22.25	3600	0.00	1	0.00	1
	10	10	36.68	3600	31.40	3600	0.25	3600	0.23	1
(100, 80)	1	5	69.19	3600	66.73	3600	0.00	1	0.00	1
	1	10	78.01	3600	75.20	3600	0.00	7	0.00	1
	10	5	25.78	3600	23.60	3600	0.00	1	0.00	1
	10	10	36.04	3600	32.57	3600	0.07	3600	0.06	1
(150, 20)	1	5	79.95	3600	78.40	3600	0.00	8	0.00	1
	1	10	84.63	3600	82.14	3600	0.47	3600	0.59	1
	10	5	23.41	3600	21.88	3600	0.00	2	0.00	1
	10	10	30.58	3600	24.48	3600	1.79	3600	1.65	1
(150, 50)	1	5	78.67	3600	76.91	3600	0.00	1	0.00	1
	1	10	86.36	3600	85.03	3600	0.02	3600	0.02	1
	10	5	28.07	3600	25.79	3600	0.00	1	0.00	1
	10	10	39.70	3600	35.77	3600	0.11	3600	0.11	1
(150, 80)	1	5	75.18	3600	73.05	3600	0.00	1	0.00	1
	1	10	83.70	3600	82.13	3600	0.00	1	0.00	1
	10	5	28.69	3600	25.63	3600	0.00	1	0.00	1
	10	10	40.13	3600	36.23	3600	0.04	3600	0.04	1

the model mean squared error, most notably A-optimality and D-optimality (see, e.g., Chamon and Ribeiro 2017a, Rubens et al. 2009, Zhao and Wang 2015). These correspond to RAOD and regularized DOD optimization problems. In the following, we compare the performance of RAOD, regularized DOD, and a random selection baseline in addressing the cold-start problem. We apply the widely-used forward greedy algorithm to solve both RAOD and regularized DOD for a fair comparison (Krause et al. 2008).

Table 4   Exact algorithms for RAOD on real data

Dataset	Params		Cutting plane via (1)		MISOCP		Algorithm 1 via (3)		Algorithm 4	
(n, d)	λ	k	MIPGap(%)	time	MIPGap(%)	time	MIPGap(%)	time	Gap(%)	time
(159, 24)	1	5	13.10	3600	8.49	3600	0.00	3548	0.02	1
	1	10	16.14	3600	6.86	3600	1.64	3600	1.80	1
	10	5	0.00	19	0.27	3600	0.00	4	0.04	1
	10	10	0.12	3600	0.27	3600	0.06	3600	0.06	1
(194, 33)	1	5	6.65	3600	5.93	3600	0.00	119	0.12	1
	1	10	9.64	3600	3.92	3600	1.15	3600	1.23	1
	10	5	0.00	2	0.01	3600	0.00	1	0.00	1
	10	10	0.00	28	0.11	3600	0.00	3	0.00	1
(200, 22)	1	5	16.08	3600	0.46	3600	0.00	416	0.08	1
	1	10	19.70	3600	2.27	3600	2.43	3600	2.27	1
	10	5	0.00	47	0.10	3600	0.00	1	0.14	1
	10	10	0.24	3600	0.24	3600	0.08	3600	0.11	1
(500, 128)	1	5	2.90	3600	–	–	0.00	48	0.02	3
	1	10	1.91	3600	–	–	0.00	441	0.01	3
	10	5	0.00	30	–	–	0.00	1	0.00	3
	10	10	0.00	26	–	–	0.00	1	0.00	3
(1000, 90)	1	5	9.15	3600	–	–	0.00	192	0.00	11
	1	10	20.05	3600	–	–	0.10	3600	0.10	11
	10	5	0.00	2181	–	–	0.00	4	0.01	11
	10	10	0.37	3600	–	–	0.00	26	0.00	11

To understand a new user's movie preferences, we leverage the intuition that their tastes are related to how existing (training) users have rated the same movies (Deldjoo et al. 2019). For each movie $i \in [n]$, the vector $\boldsymbol{a}_i \in \mathbb{R}^d$ contains its ratings from d training users. As done in Chamon and Ribeiro (2017a), we assume that the new user's ratings can be expressed as a linear combination of the past rating data, modeled as $\boldsymbol{A}^\top \boldsymbol{\theta}$, where $\boldsymbol{\theta} \in \mathbb{R}^d$ captures how similar the new user is to each training user. Since $\boldsymbol{\theta}$ is unknown, we select a subset $\widehat{S} \subseteq [n]$ of k movies for the new user to rate, aiming to provide maximal information for estimating $\boldsymbol{\theta}$. With the estimate $\widehat{\boldsymbol{\theta}}$, we predict the new user's ratings for the remaining $(n - k)$ movies as $\boldsymbol{a}_i^\top \widehat{\boldsymbol{\theta}}$ for all $i \in [n] \setminus \widehat{S}$. Finally, we evaluate the quality of the selected subset $\widehat{S}$ by computing the mean squared error (MSE) between predicted

and actual ratings. A lower MSE indicates that the selected movies more effectively capture the new user's preferences.

To facilitate testing, we truncate the movie rating data from Dooms et al. (2013) by randomly selecting a subset of ratings. Specifically, for a fixed number of movies, training users, and new users, we repeat the truncation process 15 times with different random seeds. Tables 5 and 6 compare the average MSE across all 20 and 50 new users using the three selection methods. To be specific, the average MSE is computed by averaging per-seed MSEs of all new users and then averaging these values across 15 random seeds. We refer to the MSE values based on our RAOD method, regularized DOD, and random selection as "A-MSE", "D-MSE", and "R-MSE", respectively. For each dataset, we vary the regularization parameter $\lambda \in \{0.5, 1, 1.5\}$. The random selection method is clearly unaffected by the choice of λ. Across all experimental settings, RAOD consistently yields the lowest or comparable MSE. In contrast, while the D-optimality-based method occasionally outperforms RAOD in Table 5, it tends to generate extremely high MSE in larger-scale settings with 50 new users, as shown in Table 6 where $n = 700$. The random selection method performs the worst overall, particularly in Table 6. These results highlight the robustness and effectiveness of the RAOD-based selection for cold-start recommendations, particularly as the number of new users grows.

7. Conclusion

We study the regularized A-optimal design (RAOD) problem and prove its NP-hardness for the first time. A key insight from our results is that the performance of both formulations and algorithms for RAOD hinges on whether the selection size k exceeds the data dimension d. Specifically, we demonstrate that the two existing relaxations perform well in only one of the two ranges ($k \leq d - 1$ or $k \geq d$), but poorly in the other. To address this, we propose a novel convex integer formulation that yields a stronger relaxation with provable performance guarantees for all k. This formulation significantly accelerates the exact cutting-plane algorithm in small-k and high-dimensional settings. We further investigate the complementary forward and backward greedy algorithms, tailored respectively to the ranges $k \leq d - 1$ and $k \geq d$. Our numerical experiments demonstrate the effectiveness and efficiency of the proposed algorithms. Our work can be seen as a non-sequential experimental design scheme, where the data points are selected at once. A possible future direction is to explore adaptive strategies that select points sequentially based on prior selections.

Table 5 The average MSE of 20 new users

(n, d, k)	λ	A-MSE	D-MSE	R-MSE
(500, 100, 20)	0.5	**0.0346**	0.2430	0.0460
	1	**0.0346**	0.2430	0.0460
	1.5	**0.0346**	0.2435	0.0460
(600, 100, 20)	0.5	**0.0368**	0.0517	0.0463
	1	**0.0368**	0.1864	0.0463
	1.5	**0.0386**	0.1950	0.0463
(700, 100, 20)	0.5	0.0364	**0.0360**	0.0450
	1	0.0361	**0.0357**	0.0450
	1.5	0.0353	**0.0351**	0.0450
(800, 100, 20)	0.5	**0.0366**	0.0382	0.0466
	1	**0.0367**	0.0383	0.0466
	1.5	**0.0378**	0.0454	0.0466
(900, 100, 20)	0.5	**0.0383**	0.0385	0.0483
	1	0.0390	**0.0384**	0.0483
	1.5	0.0391	**0.0382**	0.0483

Table 6 The average MSE of 50 new users

(n, d, k)	λ	A-MSE	D-MSE	R-MSE
(500, 150, 50)	0.5	**0.0323**	0.0355	1.65e+26
	1	**0.0343**	0.0355	1.65e+26
	1.5	**0.0343**	0.0381	1.65e+26
(600, 150, 50)	0.5	**0.0388**	0.0578	2.35e+26
	1	**0.0398**	0.0737	2.35e+26
	1.5	**0.0392**	0.0752	2.35e+26
(700, 150, 50)	0.5	**0.0354**	1.72e+21	2.00e+26
	1	**0.0354**	1.72e+21	2.00e+26
	1.5	**0.0356**	1.72e+21	2.00e+26
(800, 150, 50)	0.5	**0.0366**	0.0993	3.54e+26
	1	**0.0367**	0.1179	3.54e+26
	1.5	**0.0367**	0.1182	3.54e+26
(900, 150, 50)	0.5	0.0361	**0.0355**	1.23e+26
	1	**0.0361**	0.0425	1.23e+26
	1.5	**0.0361**	0.0480	1.23e+26

References

Ahipaşaoglu SD (2021) A branch-and-bound algorithm for the exact optimal experimental design problem. *Statistics and Computing* 31(5):65.

Ahipaşaoglu SD (2015) A first-order algorithm for the a-optimal experimental design problem: a mathematical programming approach. *Statistics and Computing* 25:1113–1127.

Allen-Zhu Z, Li Y, Singh A, Wang Y (2017) Near-optimal design of experiments via regret minimization. *International Conference on Machine Learning*, 126–135 (PMLR).

Anava O, Golan S, Golbandi N, Karnin Z, Lempel R, Rokhlenko O, Somekh O (2015) Budget-constrained item cold-start handling in collaborative filtering recommenders via optimal design. *Proceedings of the 24th international conference on world wide web*, 45–54.

Avron H, Boutsidis C (2013) Faster subset selection for matrices and applications. *SIAM Journal on Matrix Analysis and Applications* 34(4):1464–1499.

Barz T, Körkel S, Wozny G, et al. (2015) Nonlinear ill-posed problem analysis in model-based parameter estimation and experimental design. *Computers & Chemical Engineering* 77:24–42.

Bertsimas D, Pauphilet J, Van Parys B (2020) Rejoinder: Sparse regression: Scalable algorithms and empirical performance. *Statistical Science* 35(4):623–624.

Bian AA, Buhmann JM, Krause A, Tschiatschek S (2017) Guarantees for greedy maximization of non-submodular functions with applications. *International conference on machine learning*, 498–507 (PMLR).

Blake CL (1998) Uci repository of machine learning databases. *http://www. ics. uci. edu/~ mlearn/MLRepository. html* .

Bobadilla J, Ortega F, Hernando A, Bernal J (2012) A collaborative filtering approach to mitigate the new user cold start problem. *Knowledge-based systems* 26:225–238.

Chaloner K, Verdinelli I (1995) Bayesian experimental design: A review. *Statistical science* 273–304.

Chamon L, Ribeiro A (2017a) Approximate supermodularity bounds for experimental design. *Advances in Neural Information Processing Systems* 30.

Chamon LF, Ribeiro A (2017b) Greedy sampling of graph signals. *IEEE Transactions on Signal Processing* 66(1):34–47.

Constantine GM (1983) Schur convex functions on the spectra of graphs. *Discrete Mathematics* 45(2-3):181–188.

Cormen TH, Leiserson CE, Rivest RL, Stein C (2022) *Introduction to algorithms* (MIT press).

Deldjoo Y, Dacrema MF, Constantin MG, Eghbal-Zadeh H, Cereda S, Schedl M, Ionescu B, Cremonesi P (2019) Movie genome: alleviating new item cold start in movie recommendation. *User Modeling and User-Adapted Interaction* 29:291–343.

Derezinski M, Liang F, Mahoney M (2020) Bayesian experimental design using regularized determinantal point processes. *International Conference on Artificial Intelligence and Statistics*, 3197–3207 (PMLR).

Dereziński M, Warmuth MK (2017) Subsampling for ridge regression via regularized volume sampling. *arXiv preprint arXiv:1710.05110* .

Dooms S, De Pessemier T, Martens L (2013) Movietweetings: a movie rating dataset collected from twitter. *Workshop on Crowdsourcing and Human Computation for Recommender Systems, CrowdRec at RecSys 2013* .

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

Duarte BP (2023) Exact optimal designs of experiments for factorial models via mixed-integer semidefinite programming. *Mathematics* 11(4):854.

Duarte BP, Wong WK (2015) Finding bayesian optimal designs for nonlinear models: a semidefinite programming-based approach. *International Statistical Review* 83(2):239–262.

Duran MA, Grossmann IE (1986) An outer-approximation algorithm for a class of mixed-integer nonlinear programs. *Mathematical programming* 36:307–339.

Elahi M, Ricci F, Rubens N (2016) A survey of active learning in collaborative filtering recommender systems. *Computer Science Review* 20:29–50.

Gope J, Jain SK (2017) A survey on solving cold start problem in recommender systems. *2017 International Conference on Computing, Communication and Automation (ICCCA)*, 133–138 (IEEE).

Hendrych D, Besançon M, Pokutta S (2023) Solving the optimal experiment design problem with mixed-integer convex methods. *arXiv preprint arXiv:2312.11200* .

Higgs RE, Bemis KG, Watson IA, Wikel JH (1997) Experimental designs for selecting molecules from large chemical databases. *Journal of chemical information and computer sciences* 37(5):861–870.

Ilzarbe L, Álvarez MJ, Viles E, Tanco M (2008) Practical applications of design of experiments in the field of engineering: a bibliographical review. *Quality and Reliability Engineering International* 24(4):417–428.

Jobson JD (2012) *Applied multivariate data analysis: regression and experimental design* (Springer Science & Business Media).

Jones B, Allen-Moyer K, Goos P (2021) A-optimal versus d-optimal design of screening experiments. *Journal of Quality Technology* 53(4):369–382.

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

Krause A, Singh A, Guestrin C (2008) Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research* 9(2).

Leardi R (2009) Experimental design in chemistry: A tutorial. *Analytica chimica acta* 652(1-2):161–172.

Lewis AS (1995) The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis* 2(1):173–183.

Li Y (2024) The augmented factorization bound for maximum-entropy sampling. *arXiv preprint arXiv:2410.10078* .

Li Y, Fampa M, Lee J, Qiu F, Xie W, Yao R (2024) D-optimal data fusion: Exact and approximation algorithms. *INFORMS Journal on Computing* 36(1):97–120.

Li Y, Xie W (2024) Best principal submatrix selection for the maximum entropy sampling problem: scalable algorithms and performance guarantees. *Operations Research* 72(2):493–513.

Liang L, Yang H (2024) Pnod: An efficient projected newton framework for exact optimal experimental designs. *arXiv preprint arXiv:2409.18392* .

Lika B, Kolomvatsos K, Hadjiefthymiades S (2014) Facing the cold start problem in recommender systems. *Expert systems with applications* 41(4):2065–2073.

Madan V, Singh M, Tantipongpipat U, Xie W (2019) Combinatorial algorithms for optimal design. *Conference on Learning Theory*, 2210–2258 (PMLR).

Marshall A (1979) Inequalities: Theory of majorization and its applications.

Mason RL, Gunst RF, Hess JL (2003) *Statistical design and analysis of experiments: with applications to engineering and science* (John Wiley & Sons).

Nikolov A (2015) Randomized rounding for the largest simplex problem. *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, 861–870.

Nikolov A, Singh M, Tantipongpipat U (2022) Proportional volume sampling and approximation algorithms for a-optimal design. *Mathematics of Operations Research* 47(2):847–877.

Pukelsheim F (2006) *Optimal design of experiments* (SIAM).

Quesada I, Grossmann IE (1992) An lp/nlp based branch and bound algorithm for convex minlp optimization problems. *Computers & chemical engineering* 16(10-11):937–947.

Quinn GP, Keough MJ (2002) *Experimental design and data analysis for biologists* (Cambridge university press).

Rainforth T, Foster A, Ivanova DR, Bickford Smith F (2024) Modern bayesian experimental design. *Statistical Science* 39(1):100–114.

Rockafellar RT (1997) *Convex analysis*, volume 28 (Princeton university press).

Rubens N, Tomioka R, Sugiyama M (2009) Output divergence criterion for active learning in collaborative settings. *IPSJ Online Transactions* 2:240–249.

Sagnol G (2011) Computing optimal designs of multiresponse experiments reduces to second-order cone programming. *Journal of Statistical Planning and Inference* 141(5):1684–1708.

Sagnol G, Harman R (2015) Computing exact d-optimal designs by mixed integer second-order cone programming. *The Annals of Statistics* 43(5):2198–2224.

Sedrakyan H, Sedrakyan N (2018) *Algebraic inequalities* (Springer).

Stallrich J, Allen-Moyer K, Jones B (2023) D-and a-optimal screening designs. *Technometrics* 65(4):492–501.

Tantipongpipat U (2020) λ-regularized a-optimal design and its approximation by λ-regularized proportional volume sampling. *arXiv preprint arXiv:2006.11182* .

Volkovs M, Yu GW, Poutanen T (2017) Content-based neighbor models for cold start in recommender systems. *Proceedings of the Recommender Systems Challenge 2017*, 1–6.

Wang Y, Yu AW, Singh A (2017) On computationally tractable selection of experiments in measurement-constrained regression models. *Journal of Machine Learning Research* 18(143):1–41.

Welch WJ (1982) Algorithmic complexity: three np-hard problems in computational statistics. *Journal of Statistical Computation and Simulation* 15(1):17–25.

Winer BJ, Brown DR, Michels KM, et al. (1971) *Statistical principles in experimental design*, volume 2 (Mcgraw-hill New York).

Zhao X, Wang J (2015) A theoretical analysis of two-stage recommendation for cold-start collaborative filtering. *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, 71–80.

Appendix A: Additional proof

A.1. Proof of Proposition 1

Proof. We prove the result by showing that the objective functions of RAOD and AMESP are equivalent. For any solution S of RAOD, let $\sigma_1 \geq \cdots \geq \sigma_d \geq 0$ and $\beta_1 \geq \cdots \geq \beta_k \geq 0$ denote the eigenvalues of $\boldsymbol{A}_S \boldsymbol{A}_S^\top \in \mathcal{S}_+^d$ and $\boldsymbol{A}_S^\top \boldsymbol{A}_S \in \mathcal{S}_+^k$, respectively. Then, our analysis is split into two parts depending on whether $k \leq d$ holds or not.

(i) $k \in [d]$. Because $\boldsymbol{A}_S \boldsymbol{A}_S^\top \in \mathcal{S}_+^d$ and $\boldsymbol{A}_S^\top \boldsymbol{A}_S \in \mathcal{S}_+^k$ admit the same nonzero eigenvalues, their eigenvalues satisfy $\sigma_i = \beta_i$ for any $i \in [k]$ and $\sigma_i = 0$ for all $i \in [k+1, d]$. Given this, we have that

$$f(S) = \operatorname{tr}\left(\left(\boldsymbol{A}_S \boldsymbol{A}_S^\top + \lambda \boldsymbol{I}_d\right)^{-1}\right) = \sum_{i \in [d]} \frac{1}{\sigma_i + \lambda} = \sum_{i \in [k]} \frac{1}{\beta_i + \lambda} + \frac{d-k}{\lambda}$$
$$= \operatorname{tr}\left(\left(\boldsymbol{A}_S^\top \boldsymbol{A}_S + \lambda \boldsymbol{I}_k\right)^{-1}\right) + \frac{d-k}{\lambda} = \operatorname{tr}\left(\left(\boldsymbol{C}_{S,S}\right)^{-1}\right) + \frac{d-k}{\lambda},$$

where the first and the last equations follow from the definitions of $\boldsymbol{A}$ and $\boldsymbol{C}$, respectively.

(ii) $k \in [d+1, n]$. Analogously, according to the relation of the eigenvalues between $\boldsymbol{A}_S \boldsymbol{A}_S^\top \in \mathcal{S}_+^d$ and $\boldsymbol{A}_S^\top \boldsymbol{A}_S \in \mathcal{S}_+^k$, we get $\sigma_i = \beta_i$ for any $i \in [d]$ and $\beta_i = 0$ for all $i \in [d+1, k]$. Then, we have that

$$f(S) = \operatorname{tr}\left(\left(\boldsymbol{A}_S \boldsymbol{A}_S^\top + \lambda \boldsymbol{I}_d\right)^{-1}\right) = \sum_{i \in [d]} \frac{1}{\sigma_i + \lambda} = \sum_{i \in [k]} \frac{1}{\beta_i + \lambda} - \frac{k-d}{\lambda}$$
$$= \operatorname{tr}\left(\left(\boldsymbol{A}_S^\top \boldsymbol{A}_S + \lambda \boldsymbol{I}_k\right)^{-1}\right) + \frac{d-k}{\lambda} = \operatorname{tr}\left(\left(\boldsymbol{C}_{S,S}\right)^{-1}\right) + \frac{d-k}{\lambda}.$$

We thus complete the proof. $\qquad\square$

A.2. Proof of Theorem 1

Proof. In order to prove the result, we reduce the NP-hard independent set decision problem to AMESP. Given a simple undirected graph $G([n], E)$, we define a symmetric matrix $\boldsymbol{C}$ as

$$C_{ij} = \begin{cases} n & \text{if } i = j; \\ 1 & \text{if } (i, j) \in E; \\ 0 & \text{Otherwise.} \end{cases}$$

For all $i \in [n]$, it is easy to verify that $C_{ii} = n > n - 1 \geq \sum_{j \neq i} |C_{ij}|$, implying that $\boldsymbol{C}$ is a strictly diagonally dominant matrix. Hence, the matrix $\boldsymbol{C}$ is positive definite, and so is any principal submatrix of $\boldsymbol{C}$. Besides, we can express $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{A}^\top + \lambda \boldsymbol{I}_n$ in the form of the objective matrix in AMESP, where λ equals the smallest eigenvalue of $\boldsymbol{C}$, and $\boldsymbol{A}$ denotes the Cholesky factor of $\boldsymbol{C} - \lambda \boldsymbol{I}_n$. Therefore, the resulting pair $(\boldsymbol{A}, \lambda)$ can be used as an input instance of RAOD such that RAOD is equivalent to AMESP.

For any subset $S \subseteq [n]$, the subgragh G_S comprises the vertices S and edges in the set $E \cap (S \times S)$, and then, the subgragh G_S can be represented by the principal submatrix $\boldsymbol{C}_{S,S}$.

Next, for a subset S of size k, we compute the trace of the inverse of $\boldsymbol{C}_{S,S}$ in two cases.

(i) If S is a sized-k independent set for the graph G, according to the property of independent sets, the resulting subgraph G_S has no edges, that is, $(i,j) \notin E$ for any two vertices i, j in S. Therefore, the corresponding submatrix $\boldsymbol{C}_{S,S}$ to G_S must be diagonal, and we can easily compute $\mathrm{tr}((\boldsymbol{C}_{S,S})^{-1}) = \sum_{i \in S} 1/C_{ii} = k/n$.

(ii) If S is not a sized-k independent set, then, unlike Part (i), there is at least one nonzero off-diagonal entry in $\boldsymbol{C}_{S,S}$. Suppose that the vector $\boldsymbol{\sigma} \in \mathbb{R}_{++}^k$ contains the eigenvalues of $\boldsymbol{C}_{S,S}$. It follows that

$$\|\boldsymbol{C}_{S,S}\|_F^2 = \sum_{i \in [k]} \sigma_i^2 = \sum_{i \in S} \sum_{j \in S} C_{ij}^2 \geq kn^2 + 2.$$

In addition, we have that $\sum_{i \in [k]} \sigma_i = \mathrm{tr}(\boldsymbol{C}_{S,S}) = kn$. As a result, not all entries of $\boldsymbol{\sigma}$ can be equal. Otherwise, we would find that $\sum_{i \in [k]} \sigma_i^2 = kn^2$, a contradiction with the inequality above.

According to Lemma 1, we get

$$\frac{k}{\mathrm{tr}((\boldsymbol{C}_{S,S})^{-1})} = \frac{k}{\sum_{i \in [k]} \frac{1}{\sigma_i}} < \frac{\sum_{i \in [k]} \sigma_i}{k} = n \implies \mathrm{tr}((\boldsymbol{C}_{S,S})^{-1}) > \frac{k}{n},$$

where the first inequality is strict because the entries of $\boldsymbol{\sigma}$ are not all equal.

By combining the results in Parts (i) and (ii), we conclude that if AMESP yields an optimal value of $k/n + (d-k)/\lambda$, then there exists an independent set of size k in the graph G. On the other hand, if the optimal value is larger than $k/n + (d-k)/\lambda$, the independent set decision problem is false. We thus complete the proof. □

A.3. Proof of Lemma 2

Proof. For each $i \in S$, let us consider the following probability to sample i out of S, proposed by Dereziński and Warmuth (2017):

$$\mathbb{P}(i|S) = \frac{1 - \boldsymbol{a}_i^\top \left(\boldsymbol{A}_S \boldsymbol{A}_S^\top + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{a}_i}{s - \mathrm{tr}\left(\boldsymbol{A}_S^\top \left(\boldsymbol{A}_S \boldsymbol{A}_S^\top + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{A}_S\right)}.$$

It is easy to verify that $\sum_{i \in S} \mathbb{P}(i|S) = 1$. Then, the expected function value of removing an element from S is equal to

$$
\begin{aligned}
\mathbb{E}\left[f(S \setminus \{i\})\right] &= \sum_{i \in S} \mathbb{P}(i|S) f(S \setminus \{i\}) = \sum_{i \in S} \mathbb{P}(i|S) \, \mathrm{tr}\left(\left(\boldsymbol{A}_S \boldsymbol{A}_S^\top - \boldsymbol{a}_i \boldsymbol{a}_i^\top + \lambda \boldsymbol{I}_d\right)^{-1}\right) \\
&= \sum_{i \in S} \mathbb{P}(i|S) \left(f(S) + \frac{\boldsymbol{a}_i^\top \left(\boldsymbol{A}_S \boldsymbol{A}_S^\top + \lambda \boldsymbol{I}_d\right)^{-2} \boldsymbol{a}_i}{1 - \boldsymbol{a}_i^\top \left(\boldsymbol{A}_S \boldsymbol{A}_S^\top + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{a}_i}\right) \\
&= f(S) + \frac{\mathrm{tr}\left(\boldsymbol{A}_S^\top \left(\boldsymbol{A}_S \boldsymbol{A}_S^\top + \lambda \boldsymbol{I}_d\right)^{-2} \boldsymbol{A}_S\right)}{s - \mathrm{tr}\left(\boldsymbol{A}_S^\top \left(\boldsymbol{A}_S \boldsymbol{A}_S^\top + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{A}_S\right)} \\
&= f(S) \left(1 + \frac{\mathrm{tr}\left(\left(\boldsymbol{A}_S \boldsymbol{A}_S^\top + \lambda \boldsymbol{I}_d\right)^{-2} \boldsymbol{A}_S \boldsymbol{A}_S^\top\right)}{\mathrm{tr}\left(\left(\boldsymbol{A}_S \boldsymbol{A}_S^\top + \lambda \boldsymbol{I}_d\right)^{-1}\right) \left(s - \mathrm{tr}\left(\left(\boldsymbol{A}_S \boldsymbol{A}_S^\top + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{A}_S \boldsymbol{A}_S^\top\right)\right)}\right),
\end{aligned}
\tag{5}
$$

where the second equation is based on the Sherman–Morrison formula, the third equation is from the expression of $\mathbb{P}(i|S)$, and the last one results from the cyclic property of trace operator.

Observe that the matrix $\boldsymbol{A}_S\boldsymbol{A}_S^\top \in \mathcal{S}_+^d$ has a rank of at most $\min\{s,d\}$; hence, it has at least $d - \min\{s,d\}$ zero eigenvalues. Let $\boldsymbol{A}_S\boldsymbol{A}_S^\top = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$ be the eigen-decomposition of $\boldsymbol{A}_S\boldsymbol{A}_S^\top$, where $\boldsymbol{\Lambda} \in \mathcal{S}_+^d$ is a diagonal matrix comprising $\min\{s,d\}$ nonnegative eigenvalues $\sigma_1, \cdots, \sigma_{\min\{s,d\}}$ and zero eigenvalues $\sigma_{\min\{s,d\}+1} = \cdots = \sigma_d = 0$ of $\boldsymbol{A}_S\boldsymbol{A}_S^\top$. Given the eigen-decomposition of $\boldsymbol{A}_S\boldsymbol{A}_S^\top$, we can show that

$$M_1 := \operatorname{tr}\left(\left(\boldsymbol{A}_S\boldsymbol{A}_S^\top + \lambda\boldsymbol{I}_d\right)^{-2}\boldsymbol{A}_S\boldsymbol{A}_S^\top\right) = \operatorname{tr}\left(\left(\boldsymbol{\Lambda} + \lambda\boldsymbol{I}_d\right)^{-2}\boldsymbol{\Lambda}\right) = \sum_{i\in[\min\{s,d\}]}\frac{\sigma_i}{(\lambda+\sigma_i)^2},$$

$$M_2 := \operatorname{tr}\left(\left(\boldsymbol{A}_S\boldsymbol{A}_S^\top + \lambda\boldsymbol{I}_d\right)^{-1}\right) = \operatorname{tr}\left(\left(\boldsymbol{\Lambda} + \lambda\boldsymbol{I}_d\right)^{-1}\right) = \sum_{i\in[\min\{s,d\}]}\frac{1}{\lambda+\sigma_i} + (d - \min\{s,d\})\frac{1}{\lambda}, \text{ and} \quad (6)$$

$$M_3 := \operatorname{tr}\left(\left(\boldsymbol{A}_S\boldsymbol{A}_S^\top + \lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{A}_S\boldsymbol{A}_S^\top\right) = \operatorname{tr}\left(\left(\boldsymbol{\Lambda} + \lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{\Lambda}\right) = \sum_{i\in[\min\{s,d\}]}\frac{\sigma_i}{\lambda+\sigma_i}.$$

Replacing the expressions on the right-hand side of (5) with M_1, M_2, M_3 leads to

$$\min_{i\in S} f(S \setminus \{i\}) \leq \mathbb{E}\left[f(S \setminus \{i\})\right] = f(S)\left(1 + \frac{M_1}{M_2(s - M_3)}\right). \quad (7)$$

To prove the result, we will bound the value of $\frac{M_1}{M_2(s-M_3)}$ from above in two cases.

Case I: $s \leq d - 1$. To begin, let us establish a lower bound:

$$\sum_{i\in[s]}\frac{\lambda}{\lambda+\sigma_i} \geq \frac{s^2\lambda}{\sum_{i\in[s]}(\sigma_i+\lambda)} = \frac{s^2\lambda}{\sum_{i\in S}\|\boldsymbol{a}_i\|_2^2 + s\lambda} \geq \frac{s^2\lambda}{s\max_{i\in[n]}\|\boldsymbol{a}_i\|_2^2 + s\lambda} = s\rho, \quad (8)$$

where the first inequality is from Lemma 1, and the second equality is from the fact that $\sigma_{s+1} = \cdots = \sigma_d = 0$. Next, based on (6) and (8), we have that

$$\frac{M_1}{M_2(s - M_3)} = \frac{\sum_{i\in[s]}\frac{\sigma_i}{(\lambda+\sigma_i)^2}}{\left(\sum_{i\in[s]}\frac{1}{\lambda+\sigma_i} + (d-s)\frac{1}{\lambda}\right)\left(\lambda\sum_{i\in[s]}\frac{1}{\lambda+\sigma_i}\right)} \leq \frac{1}{\sum_{i\in[s]}\frac{\lambda}{\lambda+\sigma_i} + d - s} \leq \frac{1}{s\rho + d - s}, \quad (9)$$

where the first equation is obtained from plugging the expressions of M_1, M_2, and M_3 in (6), the second inequality is because $\frac{\sigma_i}{(\lambda+\sigma_i)^2} \leq \frac{1}{\lambda+\sigma_i}$ for each $i \in [s]$, and the last one is from (8).

Case II: $s \geq d$. Analogous to (8) and (9), we can show that

$$\sum_{i\in[d]}\frac{\lambda}{\lambda+\sigma_i} \geq \frac{d^2\lambda}{\sum_{i\in[d]}(\sigma_i+\lambda)} = \frac{d^2\lambda}{\sum_{i\in S}\|\boldsymbol{a}_i\|_2^2 + d\lambda} \geq \frac{d^2\lambda}{s\max_{i\in[n]}\|\boldsymbol{a}_i\|_2^2 + d\lambda} = d\rho,$$

and

$$\frac{M_1}{M_2(s - M_3)} = \frac{\sum_{i\in[d]}\frac{\sigma_i}{(\lambda+\sigma_i)^2}}{\left(\sum_{i\in[d]}\frac{1}{\lambda+\sigma_i}\right)\left(s - d + d - \sum_{i\in[d]}\frac{\sigma_i}{\lambda+\sigma_i}\right)} \leq \frac{1}{s - d + \sum_{i\in[d]}\frac{\lambda}{\lambda+\sigma_i}} \leq \frac{1}{s - d + d\rho}. \quad (10)$$

By plugging the results of (9) and (10) into (7), we complete the proof. $\qquad\square$

A.4. Proof of Theorem 2

Proof. Part (i). Let S be an optimal solution of RAOD with the selection size $s \in [d, n]$. Then, $|S| = s$ and $f(S) = z_s$. Based on Lemma 2, we have

$$z_{s-1} \leq \min_{i \in S} f(S \setminus \{i\}) \leq \left(1 + \frac{1}{d\rho + s - d}\right) f(S) = \frac{d\rho + s - d + 1}{d\rho + s - d} z_s, \tag{11}$$

where the first inequality is because $S \setminus \{i\}$ is a feasible solution to RAOD at $s - 1$.

It is clear that ν_k^C must be at least as large as the objective value of selecting all data points, that is, $\nu_k^C \geq z_n$. Then, sequentially applying (11) with $s = k + 1, \cdots, n$ leads to

$$z_k \leq \prod_{s \in [k+1, n]} \frac{s - d + 1 + d\rho}{s - d + d\rho} z_n = \frac{n - d + 1 + d\rho}{k + 1 - d + d\rho} z_n \leq \frac{n - d + 1 + d\rho}{k - d + 1 + d\rho} \nu_k^C \leq \frac{n - d + 1}{k - d + 1} \nu_k^C,$$

where the last inequality follows immediately because of $\rho > 0$.

Part (ii). We construct a worst-case example of RAOD-RI to demonstrate how an unbounded optimality gap occurs for $k \in [d - 1]$.

EXAMPLE 1. Fix $d = n$. Let $\boldsymbol{a}_i = \boldsymbol{e}_i$ for all $i \in [n]$, where $\boldsymbol{e}_i$ represents the i-th column of $\boldsymbol{I}_d$. ◇

In Example 1, for a given $k \in [d - 1]$, any feasible solution is optimal for RAOD, since

$$\text{tr}\left(\left(\sum_{i \in S} \boldsymbol{a}_i \boldsymbol{a}_i^\top + \lambda \boldsymbol{I}_d\right)^{-1}\right) = \text{tr}\left(\left(\sum_{i \in S} \boldsymbol{e}_i \boldsymbol{e}_i^\top + \lambda \boldsymbol{I}_d\right)^{-1}\right) = \frac{k}{1 + \lambda} + \frac{d - k}{\lambda},$$

holds for any subset $S \subseteq [n]$ with $|S| = k$. Hence, the optimal value of RAOD is $z_k = k/(1 + \lambda) + (d - k)/\lambda$. On the other hand, the vector $\widehat{\boldsymbol{x}}$ is feasible for RAOD-RI, where let $\widehat{x}_i = k/n$ for any $i \in [n]$. Plugging $\widehat{\boldsymbol{x}}$ into RAOD-RI, we get

$$\nu_k^C \leq \text{tr}\left(\left(\sum_{i \in [n]} \frac{k}{n} \boldsymbol{a}_i \boldsymbol{a}_i^\top + \lambda \boldsymbol{I}_d\right)^{-1}\right) = \text{tr}\left(\frac{k}{n} \boldsymbol{I}_d + \lambda \boldsymbol{I}_d\right) = \frac{dn}{k + n\lambda} \leq \frac{dn}{k}.$$

It follows that

$$\frac{z_k}{\nu_k^C} \geq \frac{\frac{k}{1+\lambda} + \frac{d-k}{\lambda}}{\frac{dn}{k}} \geq \frac{\frac{d-k}{\lambda}}{\frac{dn}{k}} = \frac{k(d-k)}{dn\lambda} \to \infty, \text{ if } \lambda \to 0.$$

We thus complete the proof. □

A.5. Proof of Theorem 3

Proof. The proof includes two parts.

(i) We first establish a-$(d - 1)/(d - k)$ optimality gap for any $k \in [d - 1]$. The proof includes two steps-deriving a lower bound of ν_k^M and an upper bound of z_k.

Step I. Let $\widehat{\boldsymbol{x}}$ be an optimal solution to AMESP-R. Suppose that $\beta_1 \geq \cdots \geq \beta_n \geq 0$ denote the eigenvalues of $\sum_{i \in [n]} \widehat{x}_i \boldsymbol{h}_i \boldsymbol{h}_i^\top$. Then, we have that

$$
\begin{aligned}
\sum_{i \in [n]} \beta_i &= \operatorname{tr}\left(\sum_{i \in [n]} \widehat{x}_i \boldsymbol{h}_i \boldsymbol{h}_i^\top\right) = \operatorname{tr}\left(\boldsymbol{H} \operatorname{Diag}(\boldsymbol{x}) \boldsymbol{H}^\top\right) = \operatorname{tr}\left(\operatorname{Diag}(\sqrt{\widehat{\boldsymbol{x}}}) \boldsymbol{H}^\top \boldsymbol{H} \operatorname{Diag}(\sqrt{\widehat{\boldsymbol{x}}})\right) \\
&= \operatorname{tr}\left(\operatorname{Diag}(\sqrt{\widehat{\boldsymbol{x}}})(\boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{I}_n) \operatorname{Diag}(\sqrt{\widehat{\boldsymbol{x}}})\right) = \operatorname{tr}\left(\sum_{i \in [n]} \widehat{x}_i \boldsymbol{a}_i \boldsymbol{a}_i^\top\right) + \lambda \operatorname{tr}\left(\operatorname{Diag}(\widehat{\boldsymbol{x}})\right) \quad (12) \\
&= \sum_{i \in [n]} \widehat{x}_i \|\boldsymbol{a}_i\|_2^2 + k\lambda,
\end{aligned}
$$

where the third and the fifth equations are from the cyclic property of the trace operator, and the fourth equation is from the definition of $\boldsymbol{H}^\top \boldsymbol{H}$.

Next, we show that

$$
\begin{aligned}
\nu_k^{\mathrm{M}} - \frac{d-k}{\lambda} &= \Phi\left(\sum_{i \in [n]} \widehat{x}_i \boldsymbol{h}_i \boldsymbol{h}_i^\top\right) = \sum_{i \in [\iota]} \frac{1}{\beta_i} + \frac{k-\iota}{\frac{1}{k-\iota}\sum_{i \in [\eta+1,n]} \beta_i} \geq \frac{k^2}{\sum_{i \in [n]} \beta_i} = \frac{k^2}{\sum_{i \in [n]} \widehat{x}_i \|\boldsymbol{a}_i\|_2^2 + k\lambda} \\
&\geq \frac{k^2}{k \max_{i \in [n]} \|\boldsymbol{a}_i\|_2^2 + k\lambda} = \frac{k}{\max_{i \in [n]} \|\boldsymbol{a}_i\|_2^2 + \lambda},
\end{aligned}
$$

where the second equation is from Definition 2 of Φ, the first inequality is from Lemma 1, the third equation is because of (12), and the second inequality stems from the fact that $\sum_{i \in [n]} \widehat{x}_i = k$.

Step II. As noted in Remark 1, $z_k \leq z_1$ holds. We thus get

$$
z_k \leq z_1 = \min_{i \in [n]} \operatorname{tr}\left(\left(\boldsymbol{a}_i \boldsymbol{a}_i^\top + \lambda \boldsymbol{I}_d\right)^{-1}\right) = \min_{i \in [n]} \frac{1}{\|\boldsymbol{a}_i\|_2^2 + \lambda} + \frac{d-1}{\lambda} = \frac{1}{\max_{i \in [n]} \|\boldsymbol{a}_i\|_2^2 + \lambda} + \frac{d-1}{\lambda}.
$$

Combing Steps I and II, we obtain that

$$
\frac{z_k}{\nu_k^{\mathrm{M}}} \leq \frac{\frac{1}{\max_{i \in [n]} \|\boldsymbol{a}_i\|_2^2 + \lambda} + \frac{d-1}{\lambda}}{\frac{k}{\max_{i \in [n]} \|\boldsymbol{a}_i\|_2^2 + \lambda} + \frac{d-k}{\lambda}} = \frac{\lambda + (d-1)(\max_{i \in [n]} \|\boldsymbol{a}_i\|_2^2 + \lambda)}{k\lambda + (d-k)(\max_{i \in [n]} \|\boldsymbol{a}_i\|_2^2 + \lambda)} \leq \frac{d-1}{d-k},
$$

where the last step is because for any $c_1 \geq c_2$, it is easy to verify that $\frac{c_1 + \lambda}{c_2 + k\lambda} \leq \frac{c_1}{c_2}$.

Furthermore, according to Li and Xie (2024, Theorem 11), we have that

$$
1 \leq \frac{z_k - \frac{d-k}{\lambda}}{\nu_k^{\mathrm{M}} - \frac{d-k}{\lambda}} \leq \min\{k, n-k+1\}.
$$

For any $k \in [d]$, since $d - k \geq 0$, it is evident that the inequality still holds after removing the term $-\frac{d-k}{\lambda}$. We thus obtain a $\min\{k, n-k+1\}$ optimality gap for AMESP-R when $k \in [d]$.

Combining the optimality gaps for any $k \in [d-1]$, we obtain that

$$
\frac{z_k}{\nu_k^{\mathrm{M}}} \leq \min\left\{\frac{d-1}{d-k}, k, n-k+1\right\}.
$$

Notably, the ratio of k can be omitted because

$$
k - \frac{d-1}{d-k} = \frac{1}{d-k}(kd - k^2 - d + 1) = \frac{1}{d-k}(k-1)(d-k-1) \geq 0.
$$

(ii) To prove the result, we construct an example where $\nu_k^{\mathrm{M}} < 0$, as shown below.

> EXAMPLE 2. Suppose $d = 1$, $k = 2$, $n = 3$, $\lambda = 1$, $\boldsymbol{a}_1 = 1$, $\boldsymbol{a}_2 = -1$, and $\boldsymbol{a}_3 = 2$.   ◇

In Example 2, we observe that $\boldsymbol{A}\boldsymbol{A}^\top = \sum_{i \in [n]} \boldsymbol{a}_i \boldsymbol{a}_i^\top = 6$. Consequently, the matrix $\boldsymbol{A}^\top \boldsymbol{A} \in \mathcal{S}_+^n$ has eigenvalues: $6, 0, 0$, and $\boldsymbol{C} = \boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{I}_n$ has eigenvalues: $7, 1, 1$.

We construct a feasible solution $\widehat{\boldsymbol{x}}$ to [AMESP-R](#) that yields a negative objective value for the particular instance in Example 2. Specifically, we set $\widehat{x}_i = k/n = 2/3$ for any $i \in [n]$. Thus, $\sum_{i \in [n]} \widehat{x}_i \boldsymbol{h}_i \boldsymbol{h}_i^\top = 2/3\boldsymbol{H}\boldsymbol{H}^\top$. For any matrix $\boldsymbol{X}$, it is established that $\boldsymbol{X}\boldsymbol{X}^\top$ and $\boldsymbol{X}^\top \boldsymbol{X}$ have the same nonzero eigenvalues. According to this result, given $\boldsymbol{H}^\top \boldsymbol{H} = \boldsymbol{C}$, it is easy to verify that $2/3\boldsymbol{H}\boldsymbol{H}^\top$ has eigenvalues: $14/3, 2/3, 2/3$. Clearly, $14/3 > 4/3$ holds. From Definition 2, we get $\iota = 1$ in this case, and we thus have that

$$\Phi\left(\sum_{i \in [n]} \widehat{x}_i \boldsymbol{h}_i \boldsymbol{h}_i^\top\right) + \frac{d - k}{\lambda} = \frac{3}{14} + \frac{3}{4} - 1 = -\frac{1}{28} < 0.$$

At optimality of [AMESP-R](#), we must have $\nu_k^{\mathrm{M}} < 0$ in Example 2.

Thus, we complete the proof.   □

A.6. Proof of Lemma 3

Proof. First, define $\boldsymbol{B} = \boldsymbol{A}\operatorname{Diag}(\sqrt{\boldsymbol{x}}) \in \mathbb{R}^{d \times n}$. The minimization problem in Lemma 3 can be written as

$$\min_{\boldsymbol{X} \in \mathbb{R}^{d \times n}} \frac{1}{\lambda} \operatorname{tr}\left[\boldsymbol{X}\left(\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{I}_n\right) \boldsymbol{X}^\top\right] - \frac{2}{\lambda} \operatorname{tr}\left(\boldsymbol{X}\boldsymbol{B}^\top\right) + \frac{d}{\lambda},$$

which is a strongly convex and unconstrained problem over $\boldsymbol{X}$. Its first-order optimality condition gives an optimal solution $\boldsymbol{X}^*$. We get

$$\frac{2}{\lambda}\boldsymbol{X}^*\left(\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{I}_n\right) = \frac{2}{\lambda}\boldsymbol{B},$$

which leads to $\boldsymbol{X}^* = \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{I}_n\right)^{-1}$. Plugging $\boldsymbol{X}^*$ into the minimization problem above, the optimal value is equal to

$$-\frac{1}{\lambda} \operatorname{tr}\left[\boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{I}_n\right)^{-1} \boldsymbol{B}^\top\right] + \frac{d}{\lambda} = \operatorname{tr}\left[\left(\boldsymbol{B}\boldsymbol{B}^\top + \lambda \boldsymbol{I}_d\right)^{-1}\right] = \operatorname{tr}\left[\left(\sum_{i \in [n]} x_i \boldsymbol{a}_i \boldsymbol{a}_i^\top + \lambda \boldsymbol{I}_d\right)^{-1}\right],$$

where the first equation is due to the Woodbury matrix identity, i.e.,

$$\left(\boldsymbol{B}\boldsymbol{B}^\top + \lambda \boldsymbol{I}_d\right)^{-1} = \frac{1}{\lambda}\boldsymbol{I}_d - \frac{1}{\lambda}\boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{I}_n)^{-1}\boldsymbol{B}^\top,$$

and the second equation is because of the definition of $\boldsymbol{B}$. This completes the proof.   □

A.7. Proof of Proposition 2

Proof. According to the identity in Lemma 3, we can reformulate (1) as

$$z_k = \min_{\boldsymbol{x} \in \{0,1\}^n, \boldsymbol{X} \in \mathbb{R}^{d \times n}} \left\{ \frac{1}{\lambda} \left\| \boldsymbol{X} \operatorname{Diag}\left(\sqrt{\boldsymbol{x}}\right) \boldsymbol{A}^\top - \boldsymbol{I}_d \right\|_F^2 + \|\boldsymbol{X}\|_F^2 : \sum_{i \in [n]} x_i = k \right\}, \tag{13}$$

Next, we show that (13) and MISOCP can be transformed into each other in an equivalent manner.

(i) Let $(\boldsymbol{X}^*, \boldsymbol{x}^*, \boldsymbol{\mu}^*)$ be an optimal solution of MISOCP. Then, $\|\boldsymbol{X}_i^*\|_2^2 = \mu_i^* x_i^*$ must hold for all $i \in [n]$. Next, let us construct a matrix $\widehat{\boldsymbol{X}}$. For each $i, j \in [n]$, if $x_i^* = 0$, let $\widehat{X}_{ij} = 0$, and if $x_i^* \neq 0$, let $\widehat{X}_{ij} = X_{ij}^* / \sqrt{x_i}$. Then it is easy to check that

$$\|\widehat{\boldsymbol{X}}_i\|_2^2 = \frac{\|\boldsymbol{X}_i^*\|_2^2}{x_i^*} = \mu_i^*, \forall i \in [n], \quad \text{and}$$

$$\frac{1}{\lambda} \left\| \widehat{\boldsymbol{X}} \operatorname{Diag}\left(\sqrt{\boldsymbol{x}^*}\right) \boldsymbol{A}^\top - \boldsymbol{I}_d \right\|_F^2 + \|\widehat{\boldsymbol{X}}\|_F^2 = \frac{1}{\lambda} \|\boldsymbol{X}^* \boldsymbol{A}^\top - \boldsymbol{I}_d\|_F^2 + \sum_{i \in [n]} \mu_i^*,$$

which implies that $(\widehat{\boldsymbol{X}}, \boldsymbol{x}^*)$ is a feasible solution to (13) with the same objective value as MISOCP.

(ii) Let $(\widehat{\boldsymbol{X}}, \widehat{\boldsymbol{x}})$ be an optimal solution of (13). If $\widehat{x}_i = 0$ for any $i \in [n]$, then $\widehat{\boldsymbol{X}}_i = \boldsymbol{0}$, because in this context, $\widehat{\boldsymbol{X}}_i$ only affects the Frobenius norm of $\widehat{\boldsymbol{X}}$ in the objective function of (13). Then, we define $\boldsymbol{X}^*$ and $\boldsymbol{\mu}^*$ as $\boldsymbol{X}^* = \widehat{\boldsymbol{X}} \operatorname{Diag}(\sqrt{\widehat{\boldsymbol{x}}})$ and $\mu_i^* = \|\widehat{\boldsymbol{X}}_i\|_2^2$ for all $i \in [n]$. We get $\|\boldsymbol{X}_i^*\|_2^2 = \|\widehat{\boldsymbol{X}}_i\|_2^2 \widehat{x}_i = \mu_i^* \widehat{x}_i$ for each $i \in [n]$, implying that $(\boldsymbol{X}^*, \widehat{\boldsymbol{x}}, \boldsymbol{\mu}^*)$ is feasible for MISOCP. Analogous to Part (i), we can demonstrate that $(\boldsymbol{X}^*, \widehat{\boldsymbol{x}}, \boldsymbol{\mu}^*)$ gives the same objective value for MISOCP as (13).

Hence, MISOCP and (13) are equivalent. The analysis above is independent of the binary property of $\boldsymbol{x}$, and thus can be readily extended to show the equivalence of RAOD-RI and (2). $\square$

A.8. Proof of Lemma 4

Proof. Let $\sigma_1 \geq \cdots \geq \sigma_d \geq 0$ denote the eigenvalues of $\sum_{i \in [n]} x_i \boldsymbol{a}_i \boldsymbol{a}_i^\top$. Since the rank of $\sum_{i \in [n]} x_i \boldsymbol{a}_i \boldsymbol{a}_i^\top \in \mathcal{S}_+^d$ must be bounded by $\tilde{k}$, we have that $\sigma_{\tilde{k}+1} = \cdots = \sigma_d = 0$. Accordingly, the objective matrix $(\sum_{i \in [n]} x_i \boldsymbol{a}_i \boldsymbol{a}_i^\top + \lambda \boldsymbol{I}_d)$ in (1) has eigenvalues $\sigma_1 + \lambda, \cdots \sigma_{\tilde{k}} + \lambda, \lambda, \cdots, \lambda$. Then, we have that

$$\operatorname{tr}\left(\left(\sum_{i \in [n]} x_i \boldsymbol{a}_i \boldsymbol{a}_i^\top + \lambda \boldsymbol{I}_d \right)^{-1} \right) = \sum_{i \in [\tilde{k}]} \frac{1}{\sigma_i + \lambda} + \frac{d - \tilde{k}}{\lambda} = \Gamma\left(\sum_{i \in [n]} x_i \boldsymbol{a}_i \boldsymbol{a}_i^\top; \lambda \right) + \frac{d - \tilde{k}}{\lambda}.$$

We thus complete the proof. $\square$

A.9. Proof of Remark 2

Proof. Suppose $d = 2$, $k = 1$, and $\lambda = 1$. Given $\boldsymbol{X}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ and $\boldsymbol{X}_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$, we have that

$$\Gamma\left(\frac{1}{2} \boldsymbol{X}_1 + \frac{1}{2} \boldsymbol{X}_2; \lambda \right) = \frac{1}{\frac{1}{2} + 1} = \frac{2}{3} > \frac{1}{2} \Gamma(\boldsymbol{X}_1; \lambda) + \frac{1}{2} \Gamma(\boldsymbol{X}_2; \lambda) = \frac{1}{2},$$

where the first equation is because $1/2$ is the largest eigenvalue of the matrix $1/2 \boldsymbol{X}_1 + 1/2 \boldsymbol{X}_2$. The strict inequality violates the function convexity. Thus, Γ is nonconvex. $\square$

A.10. Proof of Lemma 6

Proof. According to Lemma 5, the conditions that define η are $\sigma_\eta > \frac{1}{\tilde{k}-\eta} \sum_{i \in [\eta+1,d]} \sigma_i \geq \sigma_{\eta+1}$. By adding $\lambda > 0$ to all terms, the inequalities still hold. It follows that $\sigma_\eta^+ > \frac{1}{\tilde{k}-\eta} \sum_{i \in [\eta+1,d]} \sigma_i^+ \geq \sigma_{\eta+1}^+$. Given $\sigma_i^+ = 0$ for all $i \in [d+1, s]$, the result follows immediately from replacing d with s. $\square$

A.11. Proof of Proposition 3

Proof. First, we define a new vector $\boldsymbol{\sigma}^+ \in \mathbb{R}_+^d$ where $\sigma_i^+ = \sigma_i + \lambda > 0$ for any $i \in [\tilde{k}]$ and $\sigma_i^+ = \sigma_i$ for any $i \in [\tilde{k}+1, d]$. Given $\sigma_1^+ \geq \cdots \geq \sigma_d^+ \geq 0$, we define a function $\gamma(\boldsymbol{\sigma}^+) : \mathbb{R}_+^d \to \mathbb{R}_+$ as

$$\gamma(\boldsymbol{\sigma}^+) = \sum_{i \in [\tilde{k}]} \frac{1}{\sigma_i^+} = \sum_{i \in [\tilde{k}]} \frac{1}{\sigma_i + \lambda}.$$

By Definition 3, the function $\Gamma(\boldsymbol{X}; \lambda)$ depends only on the eigenvalues of $\boldsymbol{X}$ and can be characterized by γ, that is, $\Gamma(\boldsymbol{X}; \lambda) = \gamma(\boldsymbol{\sigma}^+)$. Then, Kim et al. (2022, Theorem 8) implies that

$$\text{conv} \, \Gamma(\boldsymbol{X}; \lambda) = \text{conv} \, \gamma(\boldsymbol{\sigma}^+).$$

This shifts our attention to deriving the convex envelope of $\text{conv} \, \gamma$.

According to Li and Xie (2024, Theorems 8 and 9), the convex envelope of γ is defined as

$$\text{conv} \, \gamma(\boldsymbol{\sigma}^+) = \sum_{i \in [\ell]} \frac{1}{\sigma_i^+} + \frac{(\tilde{k}-\ell)^2}{\sum_{i \in [\ell+1,d]} \sigma_i^+},$$

where $0 \leq \ell \leq \tilde{k}-1$ denotes the unique integer satisfying $\sigma_\ell^+ > \frac{1}{\tilde{k}-\ell} \sum_{i \in [\ell+1,d]} \sigma_i^+ \geq \sigma_{\ell+1}^+$ with the convention $\sigma_0^+ = \infty$. Applying Lemma 6 with $s = d$, we get $\ell = \eta$. Hence, we can rewrite $\text{conv} \, \gamma$ as

$$\text{conv} \, \gamma(\boldsymbol{\sigma}^+) = \sum_{i \in [\eta]} \frac{1}{\sigma_i^+} + \frac{(\tilde{k}-\eta)^2}{\sum_{i \in [\eta+1,d]} \sigma_i^+} = \sum_{i \in [\eta]} \frac{1}{\sigma_i + \lambda} + \frac{(\tilde{k}-\eta)^2}{\sum_{i \in [\eta+1,d]} \sigma_i + (\tilde{k}-\eta)\lambda},$$

where the second equation is because of the definition of $\boldsymbol{\sigma}^+$.

Observe that $\text{conv} \, \Gamma(\boldsymbol{X}; \lambda)$ is a spectral function, since it can be characterized by $\text{conv} \, \gamma(\boldsymbol{\sigma}^+)$, a function defined on the eigenvalues of $\boldsymbol{X}$. According to standard results on spectral functions (Drusvyatskiy and Kempton 2015, Lewis 1995), any subgradient of $\text{conv} \, \Gamma(\boldsymbol{X})$ at $\boldsymbol{X}$ takes the form $\boldsymbol{Q} \, \text{Diag}(\varsigma) \boldsymbol{Q}^\top$, where ς is a subgradient of underlying function (i.e., $\text{conv} \, \gamma(\boldsymbol{\sigma}^+)$) on eigenvalues $\boldsymbol{\sigma}$. The fact that $\varsigma \in \partial \, \text{conv} \, \gamma(\boldsymbol{\sigma}^+)$ directly follows from Li and Xie (2024, Proposition 8). $\square$

A.12. Proof of Lemma 7

Proof. Our proof is split into two parts depending on whether the rank of $\boldsymbol{X}$ attains $\tilde{k}$ or not. To begin, we let $\sigma_1 \geq \cdots \geq \sigma_d \geq 0$ denote the eigenvalues of $\boldsymbol{X}$.

(i) The rank of $\boldsymbol{X}$ is equal to $\tilde{k}$. Without the loss of generality, suppose that $0 \leq \eta \leq \tilde{k} - 1$ is the largest integer satisfying $\sigma_\eta > \sigma_{\eta+1} = \cdots = \sigma_{\tilde{k}}$, where $\sigma_0 = \infty$. Then, given $\sigma_{\tilde{k}} > \sigma_{\tilde{k}+1} = \cdots = \sigma_d = 0$, it is easy to verify that

$$\sigma_\eta > \frac{1}{\tilde{k} - \eta} \sum_{i \in [\eta+1, d]} \sigma_i = \frac{1}{\tilde{k} - \eta} \sum_{i \in [\eta+1, \tilde{k}]} \sigma_i = \sigma_{\eta+1},$$

which implies that η satisfies the conditions in Lemma 5. Then, according to Proposition 3, $\operatorname{conv} \Gamma(\boldsymbol{X}; \lambda)$ is equal to

$$\operatorname{conv} \Gamma(\boldsymbol{X}; \lambda) = \sum_{i \in [\eta]} \frac{1}{\sigma_i + \lambda} + \frac{(\tilde{k} - \eta)^2}{\sum_{i \in [\eta+1, d]} \sigma_i + (\tilde{k} - \eta)\lambda} = \sum_{i \in [\eta]} \frac{1}{\sigma_i + \lambda} + \frac{\tilde{k} - \eta}{\sigma_{\eta+1} + \lambda} = \sum_{i \in [\tilde{k}]} \frac{1}{\sigma_i + \lambda}$$

$$= \sum_{i \in [d]} \frac{1}{\sigma_i + \lambda} - \frac{d - \tilde{k}}{\lambda} = \operatorname{tr}((\boldsymbol{X} + \lambda \boldsymbol{I}_d)^{-1}) - \frac{d - \tilde{k}}{\lambda},$$

where the second and third equations are because $\sigma_{\eta+1} = \cdots = \sigma_{\tilde{k}} > 0 = \sigma_{\tilde{k}+1} = \cdots = \sigma_d$.

(ii) The rank of $\boldsymbol{X}$ is less than $\tilde{k}$. Let η denote its rank, and thus $\eta \leq \tilde{k} - 1$. It is clear that $\sigma_\eta > 0 = \sigma_{\eta+1} = \cdots = \sigma_d = 0$. Given this, we can readily check that $\sigma_\eta > \frac{1}{\tilde{k}-\eta} \sum_{i \in [\eta+1, d]} \sigma_i = 0 = \sigma_{\eta+1}$, indicating that the integer η exactly satisfies the conditions in Lemma 5. Then, following the analysis of Part (i), we conclude that

$$\operatorname{conv} \Gamma(\boldsymbol{X}; \lambda) = \sum_{i \in [\eta]} \frac{1}{\sigma_i + \lambda} + \frac{(\tilde{k} - \eta)^2}{\sum_{i \in [\eta+1, d]} \sigma_i + (\tilde{k} - \eta)\lambda} = \sum_{i \in [\eta]} \frac{1}{\sigma_i + \lambda} + \frac{\tilde{k} - \eta}{\lambda} = \sum_{i \in [\tilde{k}]} \frac{1}{\sigma_i + \lambda}$$

$$= \operatorname{tr}((\boldsymbol{X} + \lambda \boldsymbol{I}_d)^{-1}) - \frac{d - \tilde{k}}{\lambda}.$$

We thus complete the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

A.13. Proof of Theorem 5

We begin by introducing *Schur-convex*, which are critical to proving our results.

DEFINITION 4 (CONSTANTINE 1983). A function $g : \mathbb{R}^n \to \mathbb{R}$ is Schur-convex if for all $\boldsymbol{\nu}, \boldsymbol{\mu} \in \mathbb{R}^n$ such that $\boldsymbol{\mu}$ majorizes $\boldsymbol{\nu}$, one has that $g(\boldsymbol{\mu}) \geq g(\boldsymbol{\nu})$.

Now we are ready to prove Theorem 5.

   Proof. To begin, for a vector $\widehat{\boldsymbol{x}} \in [0, 1]^n$ with $\sum_{i \in [n]} \widehat{x}_i = k$, we let $\boldsymbol{X} = \sum_{i \in [n]} \widehat{x}_i \boldsymbol{a}_i \boldsymbol{a}_i^\top$, and let $\sigma_1 \geq \cdots \geq \sigma_d \geq 0$ denote the eigenvalues of $\boldsymbol{X}$. Let $0 \leq \eta \leq \tilde{k} - 1$ denote the unique integer from Lemma 5, that is, $\sigma_\eta > \frac{1}{\tilde{k}-\eta} \sum_{i \in [\eta+1, d]} \sigma_i \geq \sigma_{\eta+1}$. The following of the proof includes three parts.

(i) For any $k \in [d, n]$, the rank of $\boldsymbol{X}$ must be bounded by $\tilde{k}$. Therefore, the identity $\operatorname{conv} \Gamma(\boldsymbol{X}; \lambda) + (d - \tilde{k})/\lambda = \operatorname{tr}((\boldsymbol{X} + \lambda \boldsymbol{I}_d)^{-1})$ always holds based on Lemma 7, implying that the optimal values of RAOD-RII and RAOD-RI are the same.

(ii) For any $k \in [d - 1]$, we begin by establishing a technical result on Schur-convexity.

CLAIM 1. *The function $\xi(\boldsymbol{y}) = \sum_{i \in [d]} \frac{1}{y_i + \lambda} : \mathbb{R}_+^d \to \mathbb{R}$ is Schur-convex.*

Proof. It is recognized that every convex and symmetric function is Schur-convex (see Marshall 1979). We observe that the function ξ is convex and permutation-invariant with the entries of $\boldsymbol{y}$. Hence, it is Schur-convex. ◇

Applying Part (i) of Proposition 3 with $\tilde{k} = k$, we can transform $\operatorname{conv}\Gamma(\boldsymbol{X};\lambda)$ into the function ξ:

$$\operatorname{conv}\Gamma(\boldsymbol{X};\lambda) = \sum_{i \in [\eta]} \frac{1}{\sigma_i + \lambda} + \frac{k - \eta}{\frac{1}{k-\eta}\sum_{i \in [\eta+1,d]}\sigma_i + \lambda} = \sum_{i \in [d]} \frac{1}{\widehat{\sigma}_i + \lambda} - \frac{d-k}{\lambda} = \xi(\widehat{\boldsymbol{\sigma}}) - \frac{d-k}{\lambda},$$

where $\widehat{\boldsymbol{\sigma}} \in \mathbb{R}_+^d$ is defined as

$$\widehat{\sigma}_i = \sigma_i, \forall i \in [\eta], \quad \widehat{\sigma}_i = \frac{1}{k - \eta} \sum_{i \in [\eta+1,d]} \sigma_i, \forall i \in [\eta+1,k], \quad \widehat{\sigma}_i = 0, \forall i \in [k+1,d].$$

Given $\sigma_\eta > \frac{1}{k-\eta}\sum_{i \in [\eta+1,d]}\sigma_i \geq \sigma_{\eta+1}$, we can verify that

$$\sum_{i \in [\ell]} \widehat{\sigma}_i = \sum_{i \in [\ell]} \sigma_i, \forall \ell \in [\eta], \quad \sum_{i \in [\ell]} \widehat{\sigma}_i \geq \sum_{i \in [\eta]} \widehat{\sigma}_i + (\ell - \eta)\sigma_{\eta+1} \geq \sum_{i \in [\ell]} \sigma_i, \forall \ell \in [\eta+1, k-1],$$

$$\sum_{i \in [\ell]} \widehat{\sigma}_i = \sum_{i \in [d]} \sigma_i \geq \sum_{i \in [\ell]} \sigma_i, \forall \ell \in [k, d-1], \quad \text{and} \quad \sum_{i \in [d]} \widehat{\sigma}_i = \sum_{i \in [d]} \sigma_i,$$

which means that $\widehat{\boldsymbol{\sigma}}$ majorizes $\boldsymbol{\sigma}$. By the Schur-convexity of ξ in Claim 1, we get

$$\operatorname{tr}((\boldsymbol{X} + \lambda\boldsymbol{I}_d)^{-1}) = \sum_{i \in [d]} \frac{1}{\sigma_i + \lambda} = \xi(\boldsymbol{\sigma}) \leq \xi(\widehat{\boldsymbol{\sigma}}) = \operatorname{conv}\Gamma(\boldsymbol{X};\lambda) + \frac{d-k}{\lambda}.$$

We thus conclude that the objective value of RAOD-RII is no less than that of RAOD-RI.

(iii) Finally, we prove the dominance of RAOD-RII over AMESP-R. Given the vector $\widehat{\boldsymbol{x}}$, let $\beta_1 \geq \cdots \geq \beta_n$ be the eigenvalues of $\sum_{i \in [n]} \widehat{x}_i \boldsymbol{h}_i \boldsymbol{h}_i$. By Definition 2, the objective function of AMESP-R can be represented as

$$\Phi\left(\sum_{i \in [n]} \widehat{x}_i \boldsymbol{h}_i \boldsymbol{h}_i\right) + \frac{d-k}{\lambda} = \phi(\boldsymbol{\beta}) + \frac{d-k}{\lambda}.$$

We construct an alternative variant, $\boldsymbol{\beta}^+ \in \mathbb{R}_+^n$, of the vector $\boldsymbol{\sigma} \in \mathbb{R}_+^d$ as shown below.

$$\beta_i^+ = \sigma_i + \lambda, \forall i \in [k], \quad \beta_i^+ = \sigma_i, \forall i \in [k+1,d], \beta_i^+ = 0, \forall i \in [d+1,n].$$

In fact, $\boldsymbol{\beta}^+$ majorizes $\boldsymbol{\beta}$, as demonstrated by Li (2024, Lemma 1). In addition, following Claim 1, we can show that the function ϕ from Definition 2 is also Schur-convex, implying that

$$\phi(\boldsymbol{\beta}) \leq \phi(\boldsymbol{\beta}^+) = \sum_{i \in [\iota]} \frac{1}{\beta_i^+} + \frac{(k-\iota)^2}{\sum_{i \in [\iota+1,n]} \beta_i^+},$$

where $0 \leq \iota \leq k-1$ is a unique integer such that $\beta_\iota^+ > \frac{1}{k-\iota}\sum_{i \in [\iota+1,n]}\beta_i^+ \geq \beta_{\iota+1}^+$ with $\beta_0 = \infty$. Next, to prove the result, it suffices to show that the objective function of RAOD-RII (i.e., $\operatorname{conv}\Gamma(\boldsymbol{X};\lambda) + (d-\tilde{k})/\lambda$) is exactly equal to $\phi(\boldsymbol{\beta}^+) + (d-k)/\lambda$.

(a) $k \in [d]$. As $\tilde{k} = k$ holds in this context, we can verify that the construction $\boldsymbol{\beta}^+$ follows from the variant of $\boldsymbol{\sigma}$ in Lemma 6 with $s = n$, implying that $\eta = \iota$. Then, it follows that

$$\phi(\boldsymbol{\beta}^+) = \sum_{i \in [\eta]} \frac{1}{\beta_i^+} + \frac{(k-\eta)^2}{\sum_{i \in [\eta+1,n]} \beta_i^+} = \sum_{i \in [\eta]} \frac{1}{\sigma_i + \lambda} + \frac{(k-\eta)^2}{\sum_{i \in [\eta+1,d]} \sigma_i + (k-\eta)\lambda} = \operatorname{conv}\Gamma(\boldsymbol{X}; \lambda),$$

where the second equation is obtained by substituting the expression of $\boldsymbol{\beta}^+$ and omitting the zero-sum term, i.e., $\sum_{i \in [d+1,n]} \beta_i^+ = 0$.

(b) $k \in [d+1, n]$. Let $r \in [d]$ denote the rank of $\boldsymbol{X}$. In this case, by the construction of $\boldsymbol{\beta}^+$, it is easy to verify that $\beta_r^+ > \frac{1}{k-r} \sum_{i \in [r+1,n]} \beta_i^+ = \frac{1}{k-r}(k-r)\lambda = \beta_{r+1}^+$. Since ι is the unique index satisfying $\beta_\iota^+ > \frac{1}{k-\iota} \sum_{i \in [\iota+1,n]} \beta_i^+ \geq \beta_{\iota+1}^+$, $\iota = r$ must hold. Thus, the function value $\phi(\boldsymbol{\beta}^+) + (d-k)/\lambda$ reduces to

$$\phi(\boldsymbol{\beta}^+) + \frac{d-k}{\lambda} = \sum_{i \in [r]} \frac{1}{\beta_i^+} + \frac{(k-r)^2}{(k-r)\lambda} + \frac{d-k}{\lambda} = \sum_{i \in [r]} \frac{1}{\sigma_i + \lambda} + \frac{d-r}{\lambda} = \operatorname{tr}((\boldsymbol{X} + \lambda\boldsymbol{I}_d)^{-1})$$
$$= \operatorname{conv}\Gamma(\boldsymbol{X}; \lambda),$$

where the last equation is from the result of Part (i).

This concludes the proof. $\qquad\qquad\square$

A.14. Proof of Theorem 6

Proof. We will prove the approximation ratios and time complexities of Algorithm 2 and Algorithm 3, respectively.

(i) $k \in [d-1]$. For any sized-$s \in [d-1]$ subset S, according to Lemma 2, we have that

$$\min_{i \in S} f(S \setminus \{i\}) \leq \frac{d-s+s\rho+1}{d-s+s\rho} f(S) \leq \frac{d-s+1}{d-s} f(S), \tag{14}$$

where the second inequality is from $\rho \in (0, 1)$. Suppose that S^* is an optimal solution of RAOD. Here, $|S^*| = k \leq d-1$. By applying (14) with $S = S^*$ and $s = k$, we have that

$$\min_{i \in S^*} f(S^* \setminus \{i\}) \leq \frac{d-k+1}{d-k} f(S^*) = \frac{d-k+1}{d-k} z_k.$$

Let $i^* = \arg\min_{i \in S^*} f(S^* \setminus \{i\})$. Then, we can apply (14) again with $S = S^* \setminus \{i^*\}$ and $s = k-1$. By doing this with $s = k, k-1, \cdots, 2$ in a sequential manner, we get

$$\min_{i \in S^*} f(\{i\}) \leq \prod_{s \in [2,k]} \frac{d-s+1}{d-s} f(S^*) = \frac{d-1}{d-k} z_k.$$

Clearly, the first iteration of Algorithm 2 selects $i^* \in \arg\min_{i \in [n]} f(\{i\})$. Thus, the approximation ratio is immediate because $f(S_F) \leq \min_{i \in S^*} f(\{i\})$ must hold.

In Algorithm 2, there are k iterations in total, and at each iteration, finding the next point i^* and updating $\boldsymbol{\Lambda}$ take $\mathcal{O}(nd^2)$ and $\mathcal{O}(d^2)$ operations, respectively. Hence, the overall time complexity of Algorithm 2 is $\mathcal{O}(knd^2)$.

(ii) $k \in [d]$. For the output S_B of Algorithm 3, using Lemma 2 with $[n]$ and $s = n, n-1, \cdots, k+1$, we have that

$$f(S_B) \leq \prod_{s \in [k+1,n]} \frac{s-d+d\rho+1}{s-d+d\rho} f([n]) = \frac{n-d+1+d\rho}{k-d+1+d\rho} z_n \leq \frac{n-d+1}{k-d+1} z_k,$$

where the last inequality is because $\rho > 0$ and $z_n \leq z_k$.

Algorithm 3 begins with computing the sum of n rank-one matrices of size $d \times d$ and inverting this matrix, which requires $\mathcal{O}(nd^2 + d^3)$ operations. Then, Algorithm 3 performs $(n-k)$ iterations in its *for* loop, where each iteration takes $\mathcal{O}(nd^2)$ and $\mathcal{O}(d^2)$ operations, similar to Algorithm 2. Hence, the time complexity of Algorithm 3 is $\mathcal{O}((n-k)nd^2)$. $\square$