Oracle Sparse PCA via Adaptive Estimation

Wenfu Zhong and Ziping Zhao
School of Information Science and Technology, ShanghaiTech University, Shanghai, China
{wenfuzhong, zipingzhao}@shanghaitech.edu.cn

Abstract—**In this paper, we study a sparse principal component analysis (SPCA) estimator. The estimator is based on a non-convex regularized variance maximization formulation, which aims to estimate the principal subspace of a covariance matrix. We propose an efficient algorithm that employs the minorization-maximization (MM) algorithmic framework to decompose the non-convex problem into a sequence of cascaded convex subproblems. For each subproblem, we develop an alternating direction method of multipliers (ADMM) algorithm, with a novel optimality metric to explicitly controlling the computational error. Theoretically, we present a combined analysis of the computational and statistical properties of the proposed algorithm. The obtained non-convex estimator shares the same statistical convergence rate with the oracle estimator, i.e., an SPCA estimator designed with prior knowledge on the true support. The simulation results further support the theoretical findings and demonstrate the superiority of our method.**

Index Terms—**Principal component analysis, subspace estimation, sparsity, non-convex statistical optimization, minorization-maximization.**

I. Introduction

Principal component analysis (PCA) [1] is a popular technique for dimension reduction and feature extraction in multivariate data analysis. For a random vector $\mathbf{x} \in \mathbb{R}^d$ with zero mean and covariance $\boldsymbol{\Sigma}$, it aims to find $r \leq d$ weighted combinations of the variables in $\mathbf{x}$ such that the derived variables, called the principal components, capture maximal variation in $\mathbf{x}$. Specifically, the weight vectors (or the loadings) can be obtained by extracting r orthogonal eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_r$ of $\boldsymbol{\Sigma}$ associated with the r largest eigenvalues. Since the true population covariance matrix $\boldsymbol{\Sigma}$ is generally unknown, we often use the sample covariance matrix as an alternative. However, the classical PCA method faces difficulties, including lack of interpretability and inconsistent estimation in the high-dimensional settings where the dimension d is comparable to or larger than the sample size n [2]. To address these issues, sparsity structural assumptions are often presupposed on the loadings $\mathbf{u}_1, \ldots, \mathbf{u}_r$. Under the sparsity assumption, the loadings are supposed to be sparse, i.e., only a few elements in each loading vector remain non-zero, giving rise to the sparse PCA (SPCA) problem [3].

In the traditional SPCA, we typically aim to estimate the first loading vector $\mathbf{u}_1$ or individual loadings $\mathbf{u}_1, \ldots, \mathbf{u}_r$. Early methods for sparse estimates either involved simple truncation or rotation [4]–[6], or required solving a non-convex optimization problem [7]. They only apply to small or medium dimensional problems. For high-dimensional problems, [8], [9] proposed to apply an extra sparsity constraint directly on the loadings to the PCA formulation, and carry out a convex relaxation on it. The resulting convex problems can be efficiently handled with specialized algorithms [9], [10]. Later, [11] suggested using an alternating direction method of multipliers (ADMM) algorithm to solve the convex SPCA formulation proposed by [8]. A minorization-maximization (MM) algorithm was applied to address the non-convex nature in the sparsity constrained SPCA formulation [12].

However, none of the methods mentioned above include theoretical analysis for the estimation error. Moreover, when the there exists some pair of values among the largest r eigenvalues of $\boldsymbol{\Sigma}$ are identical or very close, the loadings become unidentifiable. Therefore, [13] suggested estimating the principal subspace spanned by $\mathbf{u}_1, \ldots, \mathbf{u}_r$ instead, which provided a more natural perspective of the SPCA, namely the sparse principal subspace estimation. They proposed a thresholding algorithm based on the orthogonal iteration to estimate the projection matrix of the principal subspace, and established a statistical rate of convergence for the estimation error in the spectral norm. At the same time, [14] introduced a formal definition of the sparse principal subspace. They proposed the notion of row-sparse subspace which is generated by a small subset of variables of size $s < d$. They further designed a subspace estimator with a row-sparsity and an orthogonality constraints, and proved that the estimator is minimax optimal in the Frobenius norm with milder assumptions than those in [13]. However, this formulation is NP-hard. Later, [15] proposed a convex relaxation approach to the formulation in [14] by introducing a row-sparsity ℓ_1 penalty term and a convex set, a relaxation set to the orthogonality set, called the Fantope [16]. Theoretically, they presented an $s\sqrt{\frac{\log d}{n}}$ statistical rate of convergence for the estimation error in the Frobenius norm. The convex relaxation enables the problem to be addressed using efficient optimization methods [15], [17]. It is worth mentioning that the SPCA can also be formulated as a bivariate regression optimization problem and solved using an alternating optimization strategy [3], [18], [19], although this kind of formulation will not be discussed in this paper.

In this paper, we propose a sparse principal subspace estimator based on a non-convex penalty term. It is well-known that the ℓ_1 penalty incurs an estimation bias [20], [21], and the non-convex penalty functions, such as the minimax concave penalty (MCP) [22] and the smoothly clipped absolute deviation (SCAD) penalty [20], can efficiently eliminate it. However, the non-convex penalty also lead to difficulties in optimization and analysis. We design a multistage convex optimization procedure under the MM algorithmic framework [23]–[25], which decomposes the nontrivial non-convex problem into a sequence of surrogate convex subproblems. Each subproblem is similar to the one proposed in [15], but with a weighted ℓ_1 penalty term. We apply an ADMM algorithm to each of the subproblems. To control the optimization error and simplify the theoretical analysis, we design a novel optimality metric based on the variational inequality for the ADMM subprocedure.

Theoretically, we first prove that the optimality metric exhibits a convergence rate of $O\left(t^{-1}\right)$ where t denotes the iteration number, and characterize the number of ADMM steps required to solve each subproblem. Then, we demonstrate the contraction property of the proposed MM-based algorithm, by which the estimation bias from the first subproblem (an ℓ_1-penalized problem) can be eliminated. We prove that by solving no more than approximately $\log\log d$ subproblems, our estimate is able to achieve an oracle rate $s\sqrt{\frac{1}{n}}$. We also note that there are two related works [26], [27]. Specifically, [26] proposed estimators with a quadratic and a non-convex penalty terms, whereas they merely analyzed the statistical properties of the exact local optimum. A truncated orthogonal iteration algorithm aiming at the NP-hard problem in [14] was proposed by [27], which used the estimator from [15] as the initialization. They simultaneously analyzed the computational and statistical properties of the algorithm. However, their method relies on prior knowledge on the true sparsity degree s which is hardly known in practice.

II. Row-sparse Principal Subspace Estimation

A. Problem Setting

Let $\mathbf{u}_1, \ldots, \mathbf{u}_d \in \mathbb{R}^d$ be a set of orthogonal eigenvectors of $\boldsymbol{\Sigma}$ corresponding to the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$. We assume $\lambda_r > \lambda_{r+1}$, so that the r-dimensional principal subspace spanned by $\mathbf{u}_1, \ldots, \mathbf{u}_r$ can be uniquely determined. We also assume the matrix $[\mathbf{u}_1, \ldots, \mathbf{u}_r]$ has at most s non-zero rows, ensuring the principal subspace is row-sparse and has a row-sparsity level of s. In these settings, we aim to estimate the projection matrix associated with the principal subspace, i.e. $\boldsymbol{\Pi}^* := \sum_{i=1}^{r} \mathbf{u}_i \mathbf{u}_i^\top$, which is sparse with no more than s^2 non-zero elements. Given n independent and identically distributed (i.i.d.) samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with underlying true covariance $\boldsymbol{\Sigma}$, we estimate $\boldsymbol{\Pi}^*$ based on the sample covariance matrix $\mathbf{S} := \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top$.

B. A Non-convex SPCA Estimator

We consider a family of non-convex penalty functions satisfying the following assumptions.

Assumption 1. *The non-convex function $p_\lambda \colon \mathbb{R}_+ \to \mathbb{R}_+$ with a regularization parameter $\lambda > 0$ satisfies:*

- *(a) $p_\lambda(z)$ is non-decreasing on $[0, +\infty)$ with $p_\lambda(0) = 0$ and is differentiable almost everywhere on $(0, +\infty)$;*
- *(b) $0 \leq p'_\lambda(z_1) \leq p'_\lambda(z_2) \leq \lambda$ for all $z_1 \geq z_2 \geq 0$ and $\lim_{z \to 0} p'_\lambda(z) = \lambda$;*
- *(c) There exists an $\alpha > 0$ such that $p'_\lambda(z) = 0$ for $z \geq \alpha\lambda$;*
- *(d) There exists some $c \in (0, \alpha)$ such that $p'_\lambda(c\lambda) \geq \frac{\lambda}{2}$.*

It worth mentioning that in Assumption 1, the first three conditions are key properties leading to sparsity and unbiasedness, while the last condition can actually be inferred from them and is listed here just for the convenience of theoretical analysis. It is not difficult to verify that Assumption 1 encompasses some commonly used folded concave penalties, like the minimax concave penalty (MCP) [22] and the smoothly clipped absolute deviation (SCAD) [20] penalty.

Based on Assumption 1, we propose a non-convex subspace estimator derived from the following optimization problem:

$$\underset{\boldsymbol{\Pi} \in \mathcal{F}}{\text{maximize}} \quad \langle \mathbf{S}, \boldsymbol{\Pi} \rangle - \sum_{i,j=1}^{d} p_\lambda(|\Pi_{ij}|) \quad , \tag{1}$$

in which $\langle \cdot, \cdot \rangle$ is the Frobenius inner product and

$$\mathcal{F} := \left\{ \boldsymbol{\Pi} \in \mathbb{R}^{d \times d} \colon \mathbf{I} \succeq \boldsymbol{\Pi} \succeq \mathbf{0}, \langle \boldsymbol{\Pi}, \mathbf{I} \rangle = r \right\}$$

is the Fantope [16] set.

III. Algorithm

A. MM Algorithmic Framework

Consider maximizing a continuous function $F(\mathbf{z})$. Starting at $\mathbf{z}^0$, we generate a sequence of feasible points $\left\{\mathbf{z}^k\right\}_{k \geq 1}$ through the following process. At point $\mathbf{z}^{k-1}$, during the minorization step, we construct a surrogate function $\underline{F}\left(\mathbf{z} \mid \mathbf{z}^{k-1}\right)$ that locally approximates the objective function $F(\mathbf{z})$, satisfying:

$$\begin{cases} \underline{F}\left(\mathbf{z} \mid \mathbf{z}^{k-1}\right) \leq F(\mathbf{z}), \\ \underline{F}\left(\mathbf{z}^{k-1} \mid \mathbf{z}^{k-1}\right) = F\left(\mathbf{z}^{k-1}\right). \end{cases}$$

In the maximization step, we update $\mathbf{z}^k$ by:

$$\mathbf{z}^k \in \arg\max_{\mathbf{z}} \underline{F}\left(\mathbf{z} \mid \mathbf{z}^{k-1}\right).$$

Algorithm 1: MM-Based Adaptive ℓ_1 Regularization Algorithm

```
Input: $\mathbf{S}, r, K$
Output: $\tilde{\boldsymbol{\Pi}}^K$
1 $\tilde{\boldsymbol{\Pi}}^0 \leftarrow \mathbf{0}$;
2 for $k \leftarrow 1, \ldots, K$ do
3     $\Lambda_{ij}^{k-1} = p'_\lambda\left(\left|\tilde{\Pi}_{ij}^{k-1}\right|\right)$;
4     obtain $\tilde{\boldsymbol{\Pi}}^k$ by solving (3);
```

B. Adaptive ℓ_1 Regularization

We apply the MM algorithmic framework to our formulation (1), where we minorize the penalty term $\sum_{i,j=1}^{d} p_\lambda(|\Pi_{ij}|)$ with its local linear approximation [28]. Specifically, starting from an initial estimate $\hat{\boldsymbol{\Pi}}^0$, we turn to iteratively solving a sequence of convex subproblems:

$$\underset{\boldsymbol{\Pi} \in \mathcal{F}}{\text{maximize}} \quad \langle \mathbf{S}, \boldsymbol{\Pi} \rangle - \sum_{i,j=1}^{d} p'_\lambda\left(\left|\hat{\Pi}_{ij}^{k-1}\right|\right) |\Pi_{ij}|, \ k = 1, 2, \ldots, \tag{2}$$

where $\hat{\boldsymbol{\Pi}}^k$ represents the solution to the k-th subproblem. For each k, (2) can be regarded as a weighted ℓ_1-penalized problem, and the weights are adaptively adjusted based on the solution from the previous iteration. However, the closed-form solutions to these subproblems do not exist. We require an additional iterative algorithm to solve each of them, which inevitably involves computational errors. Therefore, we denote $\tilde{\boldsymbol{\Pi}}^k$ as an approximate solution to (2) at the k-th iteration. Then, we have the following compact formulation:

$$\underset{\boldsymbol{\Pi} \in \mathcal{F}}{\text{maximize}} \quad \langle \mathbf{S}, \boldsymbol{\Pi} \rangle - \left\| \boldsymbol{\Lambda}^{k-1} \odot \boldsymbol{\Pi} \right\|_1, \ k = 1, 2, \ldots, \tag{3}$$

in which $\Lambda_{ij}^{k-1} := p'_\lambda\left(\left|\tilde{\Pi}_{ij}^{k-1}\right|\right)$, $\odot$ is the the Hadamard product, and $\|\cdot\|_1$ represents the sum of the absolute values of all matrix elements. The problem solving procedure presented above is summarized in Algorithm 1. We set $\tilde{\boldsymbol{\Pi}}^0 = \mathbf{0}$ as the initial estimate. Since $p'_\lambda\left(\left|\tilde{\Pi}_{ij}^{0}\right|\right) = p'_\lambda(0) = \lambda$, the first subproblem (i.e., $k = 1$) is an ℓ_1-penalized problem, exactly the same as the formulation in [15]. After running for K iterations, we extract the projection matrix from the r-dimensional principal subspace of the output $\tilde{\boldsymbol{\Pi}}^K$ as our final estimate.

C. An ADMM Algorithm For Problem (3)

For notational simplicity, we consider a general form of (3):

$$\underset{\boldsymbol{\Pi} \in \mathcal{F}}{\text{maximize}} \quad \langle \mathbf{S}, \boldsymbol{\Pi} \rangle - \|\boldsymbol{\Lambda} \odot \boldsymbol{\Pi}\|_1, \tag{4}$$

in which $\boldsymbol{\Lambda} \in \mathbb{R}^{d \times d}$ with $\Lambda_{ij} \in [0, \lambda]$ for $1 \leq i, j \leq d$. We denote $\tilde{\boldsymbol{\Pi}}$ as the approximate solution to (4). We consider the ADMM algorithm as a sub-procedure to handle the update in Line 4 of Algorithm 1.

To derive an ADMM algorithm, we introduce an auxiliary variable $\boldsymbol{\Psi} \in \mathbb{R}^{d \times d}$ to obtain an equivalent formulation of (4):

$$\begin{aligned} &\underset{\boldsymbol{\Pi}, \boldsymbol{\Psi} \in \mathbb{R}^{d \times d}}{\text{maximize}} && \langle \mathbf{S}, \boldsymbol{\Pi} \rangle - \|\boldsymbol{\Lambda} \odot \boldsymbol{\Psi}\|_1 \\ &\text{subject to} && \boldsymbol{\Pi} \in \mathcal{F}, \quad \boldsymbol{\Psi} = \boldsymbol{\Pi}. \end{aligned} \tag{5}$$

Then, we construct the augmented Lagrangian for (5) as

$$\mathcal{L}(\boldsymbol{\Pi}, \boldsymbol{\Psi}, \mathbf{Z}) = \langle \mathbf{S}, \boldsymbol{\Pi} \rangle - \|\boldsymbol{\Lambda} \odot \boldsymbol{\Psi}\|_1 + \langle \mathbf{Z}, \boldsymbol{\Pi} - \boldsymbol{\Psi} \rangle - \frac{\rho}{2} ||\boldsymbol{\Pi} - \boldsymbol{\Psi}||_{\mathsf{F}}^2,$$

in which $\mathbf{Z} \in \mathbb{R}^{d \times d}$ is the Lagrange multiplier associated with the equality constraint $\boldsymbol{\Psi} = \boldsymbol{\Pi}$, and $\rho > 0$ is a specified constant used to control penalization for violating the constraint. The ADMM algorithm has three iterative steps: maximizing $\mathcal{L}$ with respect to $\boldsymbol{\Pi}$, maximizing $\mathcal{L}$ with respect to $\boldsymbol{\Psi}$, and updating the multiplier $\mathbf{Z}$.

Algorithm 2: ADMM For Solving (4).

```
  Input: $\mathbf{S}, \boldsymbol{\Lambda}, r$
  Output: $\widetilde{\boldsymbol{\Pi}} = \text{average}\left\{\boldsymbol{\Pi}^1, \boldsymbol{\Pi}^2, \dots\right\}$
  parameter: $\varepsilon > 0, \rho > 0$
1 $t \leftarrow 0$;
2 $\boldsymbol{\Psi}^0 \leftarrow \mathbf{0}, \mathbf{Z}^0 \leftarrow \mathbf{0}$;
3 repeat
4     $\boldsymbol{\Pi}^{t+1} = \arg\max_{\boldsymbol{\Pi} \in \mathcal{F}} \mathcal{L}\left(\boldsymbol{\Pi}, \boldsymbol{\Psi}^t, \mathbf{Z}^t\right)$;
5     $\boldsymbol{\Psi}^{t+1} = \arg\max_{\boldsymbol{\Psi} \in \mathbb{R}^{d \times d}} \mathcal{L}\left(\boldsymbol{\Pi}^{t+1}, \boldsymbol{\Psi}, \mathbf{Z}^t\right)$;
6     $\mathbf{Z}^{t+1} = \mathbf{Z}^t - \rho\left(\boldsymbol{\Pi}^{t+1} - \boldsymbol{\Psi}^{t+1}\right)$;
7     $t \leftarrow t + 1$;
8 until $\omega^t \leq \varepsilon$;              /* Stopping criterion */
```

The overall algorithm is summarized in Algorithm 2. Both of the updates in Line 4 and Line 5 of Algorithm 2 have closed-form expressions. Specifically, $\boldsymbol{\Pi}^{t+1}$ is obtained by projecting $\rho^{-1}\left(\mathbf{S} + \mathbf{Z}^t\right) + \boldsymbol{\Psi}^t$ onto $\mathcal{F}$ using a Fantope projection operator defined in [15, Lemma 4.1]. $\boldsymbol{\Psi}^{t+1}$ is obtained by applying a weighted soft thresholding operator $\mathcal{T}_{\rho^{-1}\boldsymbol{\Lambda}}(\cdot)$ to $\boldsymbol{\Pi}^{t+1} - \rho^{-1}\mathbf{Z}^t$. For a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, the weighted soft thresholding operator is defined element-wise as

$$\left[\mathcal{T}_{\boldsymbol{\Lambda}}(\mathbf{A})\right]_{ij} := \left[\operatorname{sgn}(\mathbf{A})\right]_{ij} \max\left(\left|A_{ij}\right| - \Lambda_{ij}, 0\right), \quad 0 \leq i, j \leq d,$$

where $\operatorname{sgn}(\mathbf{A})$ is a $d \times d$ matrix whose (i,j)-th element $\left[\operatorname{sgn}(\mathbf{A})\right]_{ij} = \frac{A_{ij}}{\left|A_{ij}\right|}$ if $A_{ij} \neq 0$ and zero otherwise. In the stopping criterion, i.e., Line 8 of Algorithm 2, ω^t represents an optimality metric and $\varepsilon > 0$ is a tolerance factor to control the computational error. In this paper, we define the optimality metric based on a variational inequality associated with problem (5) [29] as

$$\begin{aligned}\omega^t = {} & \sum_{i=1}^{r} \lambda_i\left(\mathbf{S} + \mathring{\mathbf{Z}}^t\right) - \left\langle \mathbf{S}, \mathring{\boldsymbol{\Pi}}^t \right\rangle \\ & + \left\|\boldsymbol{\Lambda} \odot \mathring{\boldsymbol{\Psi}}^t\right\|_1 + \left\|\boldsymbol{\Lambda} \odot \left(\mathring{\boldsymbol{\Pi}}^t - \mathring{\boldsymbol{\Psi}}^t\right)\right\|_1,\end{aligned}$$

where $\lambda_i(\cdot)$ represents the i-th largest eigenvalue of a matrix. $\mathring{\boldsymbol{\Pi}}^t := \frac{1}{t}\sum_{i=1}^{t} \boldsymbol{\Pi}^i$, $\mathring{\boldsymbol{\Psi}}^t := \frac{1}{t}\sum_{i=1}^{t} \boldsymbol{\Psi}^i$, and $\mathring{\mathbf{Z}}^t := \frac{1}{t}\sum_{i=1}^{t} \mathbf{Z}^i$ represents the arithmetic means of the primal and dual variables over t iterations. The final output is stabilized by averaging the intermediate results. For details on our algorithm derivation, refer to [30].

IV. Theoretical Results[1]

A. Computational Analysis

We first give a proposition that guarantees the convergence of Algorithm 2.

Proposition 1. *The optimality metric ω^t approaches zero with an $O\left(t^{-1}\right)$ convergence rate:*

$$\omega^t \leq \frac{1}{t}\left(\frac{\rho r}{2} + \frac{\lambda^2 d^2}{2\rho}\right).$$

The convergence rate shown in Proposition 1 aligns with the result proven in [31] for general ADMM algorithms. According to Proposition 1, if we take $\rho = \frac{\lambda d}{\sqrt{r}}$, Algorithm 2 executes at most $\left\lceil \frac{\lambda d \sqrt{r}}{\varepsilon} \right\rceil$ iterations. We further validate this convergence rate by the simulation curve in Fig. (1b).

B. Statistical Analysis

The next assumption is related to the data distribution.

Assumption 2. *Assume $\mathbf{x} \in \mathbb{R}^d$ is a sub-Gaussian random vector, which satisfies*

$$\mathbb{P}\left(\left|\langle \mathbf{x}, \mathbf{v} \rangle\right| \geq z\right) \leq 2 \exp\left(\frac{-Lz^2}{\left\|\boldsymbol{\Sigma}^{1/2}\mathbf{v}\right\|_2^2}\right)$$

for all unit $\mathbf{v} \in \mathbb{R}^d$ and $z > 0$, with a constant $L > 0$.

Assumption 2 is a general setting widely adopted in many SPCA literatures [15], [26], [27], the most common example of which is the multivariate normal distribution with covariance $\boldsymbol{\Sigma}$.

Let $\mathcal{S}^*$ denote the support set of $\boldsymbol{\Pi}^*$. We make another mild assumption on the magnitude of the projection matrix.

Assumption 3. *The projection matrix $\boldsymbol{\Pi}^*$ satisfies*

$$\min_{(i,j) \in \mathcal{S}^*} \left|\Pi^*_{ij}\right| \geq (\alpha + c)\,\lambda \gtrsim \lambda,$$

in which α and c are defined in Assumption 1.

Assumption 3 is referred to as the minimum signal strength condition, which often arises in the field of the sparse optimization and analysis [32]–[34]. Specially, if we take $\lambda \asymp \sqrt{\frac{\log d}{n}}$, the requirement will become negligible as $\frac{\log d}{n} \to 0$.

Now, we are ready to present the theoretical results that characterize the contraction property of Algorithm 1 and the statistical convergence rate of our estimate. We take the estimation error in the Frobenius norm $\|\cdot\|_{\mathsf{F}}$ as the metric, and denote $\mathbf{W} := \mathbf{S} - \boldsymbol{\Sigma}$ and $\delta := \lambda_r - \lambda_{r+1}$ for notational simplicity. We also use $\mathbf{W}_{\mathcal{S}^*}$ to denote a matrix whose (i,j)-th entry is equal to W_{ij} if $(i,j) \in \mathcal{S}^*$ and zero otherwise.

Lemma 2. *Suppose that Assumptions 1 and 3 hold. If $\max_{1 \leq i,j \leq d} \left|W_{ij}\right| \leq \frac{\lambda}{2}$, we have*

$$\begin{aligned}\left\|\widetilde{\boldsymbol{\Pi}}^k - \boldsymbol{\Pi}^*\right\|_{\mathsf{F}} \leq {} & \underbrace{\frac{2}{\delta}\left\|\mathbf{W}_{\mathcal{S}^*}\right\|_{\mathsf{F}}}_{\text{oracle rate}} + \underbrace{\sqrt{\frac{2\varepsilon}{\delta}}}_{\text{optimization error}} \\ & + \underbrace{\tau\left\|\widetilde{\boldsymbol{\Pi}}^{k-1} - \boldsymbol{\Pi}^*\right\|_{\mathsf{F}}}_{\text{contraction}},\end{aligned}$$

for some constant $\tau \in (0,1)$.

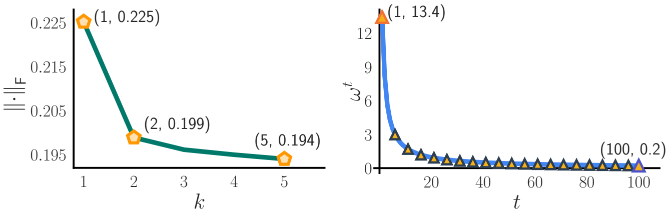
Lemma 2 shows the contraction property [32] of Algorithm 1 along the solution path $\left\{\widetilde{\boldsymbol{\Pi}}^k\right\}_{k \geq 2}$. On the right-hand side of the inequality, the *oracle rate* term represents the intrinsic statistical error of formulation (1), which coincides with that of the oracle estimator.[2] The *optimization error*, stemming from Algorithm 2, becomes negligible by setting an appropriate tolerance factor ε. The *contraction* term indicates the shrinkage of estimation error during the MM iterations, implying that we can perfectly eliminate the biased estimate from the first subproblem by solving a sufficient number of subproblems. In the following result, we show that the final output $\widetilde{\boldsymbol{\Pi}}^K$ achieves the same statistical convergence rate as the oracle estimator.

[1] Due to space limitation, all the proofs of this paper are given in [30].

[2] The oracle estimator is a theoretical benchmark for evaluating practical algorithms. It assumes the support set $\mathcal{S}^*$ is known beforehand. Specifically, in the context of SPCA with Fantope constraint, the oracle estimator is given by $\hat{\boldsymbol{\Pi}}_O := \arg\max_{\boldsymbol{\Pi} \in \mathcal{F}, \operatorname{supp}(\boldsymbol{\Pi}) \subseteq \mathcal{S}^*} \langle \mathbf{S}, \boldsymbol{\Pi} \rangle$ where $\operatorname{supp}(\cdot)$ represents the support set of a matrix. It is straightforward to verify that the oracle estimator satisfies $\left\|\hat{\boldsymbol{\Pi}}_O - \boldsymbol{\Pi}^*\right\|_{\mathsf{F}} \leq \frac{2}{\delta}\left\|\mathbf{W}_{\mathcal{S}^*}\right\|_{\mathsf{F}}$.

TABLE I: Comparison between different principal subspace estimation methods. ($d = 100$, $s = 10$, $r = 5$; eigenvalues: $\lambda_1 = \cdots = \lambda_{r-1} = 100$, $\lambda_r = 10$, $\lambda_{r+1} = \cdots = \lambda_d = 1$)

Metrics	n	PCA	SOIP			cvx. ADMM	ncvx. ADMM	MM (ours)
			$\hat{s} = 1.2 \times s$	$\hat{s} = 1.5 \times s$	$\hat{s} = 2 \times s$			
$\|\cdot\|_{\mathsf{F}}$	60	0.7215 (0.0524)	0.3057 (0.0459)	0.3909 (0.0487)	0.4699 (0.0491)	0.3094 (0.0314)	0.3013 (0.0317)	**0.2830** (**0.0227**)
	80	0.6259 (0.0420)	0.2517 (0.0302)	0.3302 (0.0348)	0.4032 (0.0373)	0.2614 (0.0295)	0.2526 (0.0274)	**0.2376** (**0.0171**)
	120	0.5026 (0.0392)	0.2065 (0.0280)	0.2677 (0.0336)	0.3266 (0.0391)	0.2048 (0.0276)	0.1903 (0.0261)	**0.1819** (**0.0173**)
CPU Time (sec.)	60	–	16.64	16.31	15.71	21.31	22.57	21.13
	80	–	18.45	18.51	18.06	23.21	21.99	20.06
	120	–	17.65	16.83	17.54	22.17	21.16	20.16



(a) MM (b) ADMM

Fig. 1: Convergence of algorithms.

Theorem 3. *Suppose that Assumptions 1, 2, and 3 hold. If we take*

$$\lambda \asymp \sqrt{\frac{\log d}{n}}, \ \varepsilon \lesssim \frac{1}{n},$$

and $K \gtrsim \log \log d$, *then the final output* $\widetilde{\mathbf{\Pi}}^K$ *of Algorithm 1 satisfies*

$$\left\| \widetilde{\mathbf{\Pi}}^K - \mathbf{\Pi}^* \right\|_{\mathsf{F}} \lesssim \frac{\lambda_1}{\delta} s \sqrt{\frac{1}{n}},$$

with probability at least $1 - \frac{2}{d^2} - \frac{2}{e^s}$.

Theorem 3 shows that we just need to solve no more than approximately $\log \log d$ subproblems to attain the oracle rate $s\sqrt{\frac{1}{n}}$, which is sharper than the result reported in [15]. Fig. (1a) also illustrates that after only a few iterations, the estimation error significantly decreases.

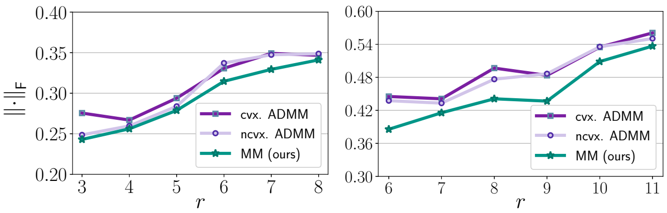
Recall that the final estimate is obtained by extracting the projection matrix $\mathbf{\Pi}^K$ from the r-dimensional principal subspace of $\widetilde{\mathbf{\Pi}}^K$. And the following relation readily follows [14], [15]:

$$\left\| \mathbf{\Pi}^K - \mathbf{\Pi}^* \right\|_{\mathsf{F}} \leq 2 \left\| \widetilde{\mathbf{\Pi}}^K - \mathbf{\Pi}^* \right\|_{\mathsf{F}}.$$

Therefore, the estimation error bound in Theorem 3 is also applicable to $\mathbf{\Pi}^K$.

V. Numerical Experiments

In this section, we compare the proposed method (MM) with three well-known principal subspace estimation approaches [15], [26], [27], all of which rely on the Fantope constraint. Specifically, [15] proposed a convex SPCA estimator based on ℓ_1 regularization, while [26] considers both quadratic and non-convex regularization terms in their SPCA formulation. Since both apply an ADMM algorithm to solve their respective problems, we refer to these methods as cvx. ADMM and ncvx. ADMM, respectively. The sparse orthogonal iteration pursuit (SOIP) algorithm, proposed by [27] , consists of two stages. In the first stage, an estimator from [15] is obtained, and in the second stage, this estimator is refined using a truncated orthogonal iteration procedure, where a sparsity parameter $\hat{s}$, dependent on the true sparsity level s, must be prespecified. In [27], it is recommended to set $\hat{s}$ as a multiple (≥ 1) of s. For comparison, we choose $\hat{s} = 1.2, 1.5, 2$ in the simulation. Additionally, we include the traditional PCA method as a baseline. We select MCP as the non-convex penalty function for both MM and ncvx. ADMM, and use cross-validation to tune the parameters. All data are generated from a multivariate normal distribution, and the reported results are averaged over 20 Monte Carlo runs.



(a) $s = 10$ (b) $s = 20$

Fig. 2: Comparison between cvx. ADMM, ncvx. ADMM, and MM on different rank r. ($d = 100$, $n = 60$; eigenvalues: $\lambda_1 = \cdots = \lambda_{r-1} = 100$, $\lambda_r = 10$, $\lambda_{r+1} = \cdots = \lambda_d = 1$)

In Fig. (2), we demonstrate that as r increases, the estimation errors of cvx. ADMM, ncvx. ADMM, and MM exhibit a similar upward trend.[3] However, MM consistently outperforms both cvx. ADMM and ncvx. ADMM. In some cases, ncvx. ADMM even performs worse than cvx. ADMM. In TABLE I, we present the estimation errors and CPU times on a server with 32 Intel 2.30GHz CPUs. The results clearly demonstrate that MM significantly outperforms the other methods in terms of both estimation accuracy and numerical stability. For SOIP, even a slight deviation of the parameter $\hat{s}$ from the true sparsity degree s can lead to significant estimation errors and numerical instability. Additionally, MM runs faster than both cvx. ADMM and ncvx. ADMM. Although SOIP takes less CPU time, this is because it only approximately solves the SPCA formulation in its first stage by using an early stopping strategy.

VI. Conclusion

In this paper, we have proposed a novel non-convex formulation for sparse principal component analysis. To solve this, we have developed an efficient multi-stage convex relaxation algorithm within the majorization-minimization algorithmic framework. A thorough computational and statistical analysis has been provided to evaluate the performance of the proposed algorithm. Our simulation results have not only validated the theoretical findings but also demonstrated the superiority of our method.

[3] In theoretical proofs, we generally consider r as a constant. Our experiments treat it as a variable for a more comprehensive comparison, thereby demonstrating the robustness of our method.

References

[1] H. Hotelling, "Analysis of a complex of statistical variables into principal components." *Journal of Educational Psychology*, vol. 24, no. 6, p. 417, 1933.

[2] I. M. Johnstone and A. Y. Lu, "On consistency and sparsity for principal components analysis in high dimensions," *Journal of the American Statistical Association*, vol. 104, no. 486, pp. 682–693, 2009.

[3] T. H. Hui Zou and R. Tibshirani, "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[4] I. T. Jolliffe, "Rotation of principal components: choice of normalization constraints," *Journal of Applied Statistics*, vol. 22, no. 1, pp. 29–35, 1995.

[5] J. Cadima and I. T. Jolliffe, "Loading and correlations in the interpretation of principle compenents," *Journal of Applied Statistics*, vol. 22, no. 2, pp. 203–214, 1995.

[6] S. Vines, "Simple principal components," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 49, no. 4, pp. 441–451, 2000.

[7] N. T. T. Ian T Jolliffe and M. Uddin, "A modified principal component technique based on the LASSO," *Journal of Computational and Graphical Statistics*, vol. 12, no. 3, pp. 531–547, 2003.

[8] A. d'Aspremont, L. Ghaoui, M. Jordan, and G. Lanckriet, "A direct formulation for sparse PCA using semidefinite programming," in *Advances in Neural Information Processing Systems*, vol. 17. MIT Press, 2004.

[9] A. d'Aspremont, F. Bach, and L. El Ghaoui, "Optimal solutions for sparse principal component analysis." *Journal of Machine Learning Research*, vol. 9, no. 7, 2008.

[10] A. Nemirovski, "Prox-method with rate of convergence $O\left(1/t\right)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems," *SIAM Journal on Optimization*, vol. 15, no. 1, pp. 229–251, 2004.

[11] S. Ma, "Alternating direction method of multipliers for sparse principal component analysis," *Journal of the Operations Research Society of China*, vol. 1, pp. 253–274, 2013.

[12] K. Benidis, Y. Sun, P. Babu, and D. P. Palomar, "Orthogonal sparse PCA and covariance estimation via procrustes reformulation," *IEEE Transactions on Signal Processing*, vol. 64, no. 23, pp. 6211–6226, 2016.

[13] Z. Ma, "Sparse principal component analysis and iterative thresholding," *The Annals of Statistics*, vol. 41, no. 2, pp. 772 – 801, 2013.

[14] V. Q. Vu and J. Lei, "Minimax sparse principal subspace estimation in high dimensions," *The Annals of Statistics*, vol. 41, no. 6, pp. 2905 – 2947, 2013.

[15] V. Q. Vu, J. Cho, J. Lei, and K. Rohe, "Fantope projection and selection: A near-optimal convex relaxation of sparse PCA," in *Advances in Neural Information Processing Systems*, vol. 26. Curran Associates, Inc., 2013.

[16] J. Dattorro, *Convex Optimization & Euclidean Distance Geometry*. Meboo Publishing, 2005.

[17] Y. Qiu, J. Lei, and K. Roeder, "Gradient-based sparse principal component analysis with extensions to online learning," *Biometrika*, vol. 110, no. 2, pp. 339–360, 2023.

[18] H. Shen and J. Z. Huang, "Sparse principal component analysis via regularized low rank matrix approximation," *Journal of Multivariate Analysis*, vol. 99, no. 6, pp. 1015–1034, 2008.

[19] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, 04 2009.

[20] F. Jianqing and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, p. 1348, 2001.

[21] H. Zou, "The adaptive LASSO and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[22] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, vol. 38, no. 2, pp. 894 – 942, 2010.

[23] K. Lange, D. R. Hunter, and I. Yang, "Optimization transfer using surrogate objective functions," *Journal of Computational and Graphical Statistics*, vol. 9, no. 1, pp. 1–20, 2000.

[24] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[25] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 794–816, 2016.

[26] Q. Gu, Z. Wang, and H. Liu, "Sparse PCA with oracle property," in *Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Inc., 2014.

[27] Z. Wang, H. Lu, and H. Liu, "Tighten after relax: Minimax-optimal sparse PCA in polynomial time," in *Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Inc., 2014.

[28] H. Zou and R. Li, "One-step sparse estimates in nonconcave penalized likelihood models," *Annals of Statistics*, vol. 36, no. 4, p. 1509, 2008.

[29] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*. SIAM, 2000.

[30] https://www.ncvxopt.com/pubs/ZhongZhao-SparsePCA-supp.pdf.

[31] B. He and X. Yuan, "On the $O\left(1/n\right)$ convergence rate of the douglas–rachford alternating direction method," *SIAM Journal on Numerical Analysis*, vol. 50, no. 2, pp. 700–709, 2012.

[32] J. Fan, H. Liu, Q. Sun, and T. Zhang, "I-lamm for sparse learning: Simultaneous control of algorithmic complexity and statistical error," *The Annals of Statistics*, vol. 46, no. 2, pp. 814 – 841, 2018.

[33] J. Ying, J. V. de Miranda Cardoso, and D. Palomar, "Nonconvex sparse graph learning under laplacian constrained graphical model," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 7101–7113.

[34] Q. Wei and Z. Zhao, "Large covariance matrix estimation with oracle statistical rate via majorization-minimization," *IEEE Transactions on Signal Processing*, 2023.

Supplementary Material to "Oracle Sparse PCA via Adaptive Estimation"

Wenfu Zhong
School of Information Science and Technology
ShanghaiTech University
Shanghai, China
wenfuzhong@shanghaitech.edu.cn

Ziping Zhao
School of Information Science and Technology
ShanghaiTech University
Shanghai, China
zipingzhao@shanghaitech.edu.cn

This supplementary material contains the design details of the optimality metric mentioned in Section III and the proofs for Proposition 1, Lemma 2, and Theorem 3 in Section IV.

Appendix A
Optimality Metric

In this appendix, we design an optimality metric used in the stopping criterion of Algorithm 2. Firstly, we need to identify the variational inequality associated with problem (5), which characterizes the optimality of the solution. Suppose $\hat{\mathbf{\Pi}}$ and $\hat{\mathbf{\Psi}}$ are the primal optima of problem (5), and $\hat{\mathbf{Z}}$ is the dual optimum. Let

$$\mathcal{L}_0\left(\mathbf{\Pi},\mathbf{\Psi},\mathbf{Z}\right) := \langle \mathbf{S},\mathbf{\Pi}\rangle - \|\mathbf{\Lambda}\odot\mathbf{\Psi}\|_1 + \langle \mathbf{Z},\mathbf{\Pi}-\mathbf{\Psi}\rangle$$

be the Lagrangian function of problem (5). Then according to convex optimization theory [1], $\left(\hat{\mathbf{\Pi}},\hat{\mathbf{\Psi}},\hat{\mathbf{Z}}\right)$ should be a saddle point of $\mathcal{L}_0\left(\cdot\right)$, and the following inequalities hold:

$$\mathcal{L}_0\left(\hat{\mathbf{\Pi}},\hat{\mathbf{\Psi}},\mathbf{Z}\right) \geq \mathcal{L}_0\left(\hat{\mathbf{\Pi}},\hat{\mathbf{\Psi}},\hat{\mathbf{Z}}\right) \geq \mathcal{L}_0\left(\mathbf{\Pi},\mathbf{\Psi},\hat{\mathbf{Z}}\right),\ \forall \mathbf{\Pi}\in\mathcal{F},\mathbf{\Psi}\in\mathbb{R}^{d\times d},\mathbf{Z}\in\mathbb{R}^{d\times d}.$$

From the first inequality,

$$\left\langle \hat{\mathbf{Z}}-\mathbf{Z},\hat{\mathbf{\Pi}}-\hat{\mathbf{\Psi}}\right\rangle \leq 0. \tag{6}$$

From the second inequality,

$$\left\langle \mathbf{S},\mathbf{\Pi}-\hat{\mathbf{\Pi}}\right\rangle + \left\|\mathbf{\Lambda}\odot\hat{\mathbf{\Psi}}\right\|_1 - \|\mathbf{\Lambda}\odot\mathbf{\Psi}\|_1 + \left\langle \hat{\mathbf{Z}},\mathbf{\Pi}-\hat{\mathbf{\Pi}}\right\rangle - \left\langle \hat{\mathbf{Z}},\mathbf{\Psi}-\hat{\mathbf{\Psi}}\right\rangle \leq 0. \tag{7}$$

Summing up (6) and (7), we obtain the variational inequality:

$$\left\langle \mathbf{S},\mathbf{\Pi}-\hat{\mathbf{\Pi}}\right\rangle + \left\|\mathbf{\Lambda}\odot\hat{\mathbf{\Psi}}\right\|_1 - \|\mathbf{\Lambda}\odot\mathbf{\Psi}\|_1 - \left\langle \begin{bmatrix}\mathbf{\Pi}-\hat{\mathbf{\Pi}}\\ \mathbf{\Psi}-\hat{\mathbf{\Psi}}\\ \mathbf{Z}-\hat{\mathbf{Z}}\end{bmatrix}, \begin{bmatrix}-\hat{\mathbf{Z}}\\ \hat{\mathbf{Z}}\\ \hat{\mathbf{\Pi}}-\hat{\mathbf{\Psi}}\end{bmatrix}\right\rangle \leq 0,\ \forall \mathbf{\Pi}\in\mathcal{F},\mathbf{\Psi}\in\mathbb{R}^{d\times d},\mathbf{Z}\in\mathbb{R}^{d\times d}. \tag{8}$$

Now solving (5) is equivalent to finding a set of $\left(\hat{\mathbf{\Pi}},\hat{\mathbf{\Psi}},\hat{\mathbf{Z}}\right)$ satisfying (8).

Based on the variational inequality (8), we define an auxiliary function to measure the optimality of $\left(\mathring{\mathbf{\Pi}}^t,\mathring{\mathbf{\Psi}}^t,\mathring{\mathbf{Z}}^t\right)$ at iteration t:

$$V^t\left(\mathbf{\Pi},\mathbf{\Psi},\mathbf{Z}\right) := \left\langle \mathbf{S},\mathbf{\Pi}-\mathring{\mathbf{\Pi}}^t\right\rangle + \left\|\mathbf{\Lambda}\odot\mathring{\mathbf{\Psi}}^t\right\|_1 - \|\mathbf{\Lambda}\odot\mathbf{\Psi}\|_1 - \left\langle \begin{bmatrix}\mathbf{\Pi}-\mathring{\mathbf{\Pi}}^t\\ \mathbf{\Psi}-\mathring{\mathbf{\Psi}}^t\\ \mathbf{Z}-\mathring{\mathbf{Z}}^t\end{bmatrix}, \begin{bmatrix}-\mathring{\mathbf{Z}}^t\\ \mathring{\mathbf{Z}}^t\\ \mathring{\mathbf{\Pi}}^t-\mathring{\mathbf{\Psi}}^t\end{bmatrix}\right\rangle.$$

Our design approach is to identify the maximum value of the function V^t within a subset of set $\mathcal{F}\times\mathbb{R}^{d\times d}\times\mathbb{R}^{d\times d}$ and use it as the optimality metric, i.e.,

$$\omega^t := \underset{\mathbf{\Pi}\in\mathcal{F},\mathbf{\Psi}\in\mathbb{R}^{d\times d},\mathbf{Z}=\widetilde{\mathbf{Z}}^t}{\text{maximize}}\ V^t\left(\mathbf{\Pi},\mathbf{\Psi},\mathbf{Z}\right), \tag{9}$$

where $\widetilde{\mathbf{Z}}^t := -\mathbf{\Lambda}\odot\operatorname{sgn}\left(\mathring{\mathbf{\Pi}}^t-\mathring{\mathbf{\Psi}}^t\right)$. If there exists a small $\varepsilon>0$ such that $\omega^t\leq\varepsilon$, then $\left(\mathring{\mathbf{\Pi}}^t,\mathring{\mathbf{\Psi}}^t,\mathring{\mathbf{Z}}^t\right)$ can be regarded as an approximate solution of (8) with tolerance ε [2].

Since there is no coupling between variables $\mathbf{\Pi}$, $\mathbf{\Psi}$, and $\mathbf{Z}$ in problem (9), we can analyze their respective optimization processes separately. For the variable $\mathbf{\Pi}$, we get the following maximization subproblem:

$$\underset{\mathbf{\Pi}\in\mathcal{F}}{\text{maximize}}\quad \left\langle \mathbf{S}+\mathring{\mathbf{Z}}^t,\mathbf{\Pi}\right\rangle \tag{10}$$

The optimal value of (10) takes $\sum_{i=1}^{r}\lambda_i\left(\mathbf{S}+\mathring{\mathbf{Z}}^t\right)$ by applying the basic properties of Fantope projection (see Lemma 1 in [3]). For the variable $\mathbf{\Psi}$, we have

$$\underset{\mathbf{\Psi}\in\mathbb{R}^{d\times d}}{\text{maximize}} \quad -\left\|\mathbf{\Lambda}\odot\mathbf{\Psi}\right\|_1-\left\langle\mathring{\mathbf{Z}}^t,\mathbf{\Psi}\right\rangle. \tag{11}$$

By the first order optimality condition derived from Line 5 of Algorithm 2 and the update of dual variable in Line 6, there exists

$$-\mathbf{\Lambda}\odot\mathbf{\Xi}^{t'}=\mathbf{Z}^{t'-1}-\rho\left(\mathbf{\Pi}^{t'}-\mathbf{\Psi}^{t'}\right)=\mathbf{Z}^{t'},\ \exists\mathbf{\Xi}^{t'}\in\partial\left\|\mathbf{\Psi}^{t'}\right\|_1,\ t'\geq 1,$$

and thus the following relation holds:

$$\left|\mathring{Z}_{ij}^t\right|\leq\frac{1}{t}\sum_{t'=1}^{t}\left|Z_{ij}^{t'}\right|=\frac{1}{t}\sum_{t'=1}^{t}\left|\Lambda_{ij}\right|\left|\Xi_{ij}^{t'}\right|\leq\frac{1}{t}\sum_{t'=1}^{t}\left|\Lambda_{ij}\right|=\Lambda_{ij},\ 1\leq i,j\leq d,$$

in which the second inequality is due to $\Xi_{ij}^{t'}\in[-1,1]$. Consequently, it is straightforward to deduce that problem (11) must have a zero solution. For the variable $\mathbf{Z}$, as the feasible set contains only a single point, it is trivial to calculate the optimal value by substituting $\widetilde{\mathbf{Z}}^t$. Based on the above analysis, we can derive the explicit expression of the optimality metric:

$$\omega^t=\sum_{i=1}^{r}\lambda_i\left(\mathbf{S}+\mathring{\mathbf{Z}}^t\right)-\left\langle\mathbf{S},\mathring{\mathbf{\Pi}}^t\right\rangle+\left\|\mathbf{\Lambda}\odot\mathring{\mathbf{\Psi}}^t\right\|_1+\left\|\mathbf{\Lambda}\odot\left(\mathring{\mathbf{\Pi}}^t-\mathring{\mathbf{\Psi}}^t\right)\right\|_1.$$

Appendix B
Proof of Main Results

This appendix contains the proofs of main theoretical results, including Proposition 1 and Theorem 3.

A. Proof of Proposition 1

Proof: Let

$$\left(\mathbf{\Pi}_M,\mathbf{\Psi}_M,\widetilde{\mathbf{Z}}^t\right)=\underset{\mathbf{\Pi}\in\mathcal{F},\mathbf{\Psi}\in\mathbb{R}^{d\times d},\mathbf{Z}=\widetilde{\mathbf{Z}}^t}{\arg\max}\ V^t\left(\mathbf{\Pi},\mathbf{\Psi},\mathbf{Z}\right).$$

Then we have

$$\begin{aligned}
V^t\left(\mathbf{\Pi}_M,\mathbf{\Psi}_M,\widetilde{\mathbf{Z}}^t\right)&=\left\langle\mathbf{S},\mathbf{\Pi}_M-\mathring{\mathbf{\Pi}}^t\right\rangle+\left\|\mathbf{\Lambda}\odot\mathring{\mathbf{\Psi}}^t\right\|_1-\left\|\mathbf{\Lambda}\odot\mathbf{\Psi}_M\right\|_1-\left\langle\begin{bmatrix}\mathbf{\Pi}_M-\mathring{\mathbf{\Pi}}^t\\ \mathbf{\Psi}_M-\mathring{\mathbf{\Psi}}^t\\ \widetilde{\mathbf{Z}}^t-\mathring{\mathbf{Z}}^t\end{bmatrix},\begin{bmatrix}-\mathring{\mathbf{Z}}^t\\ \mathring{\mathbf{Z}}^t\\ \mathring{\mathbf{\Pi}}^t-\mathring{\mathbf{\Psi}}^t\end{bmatrix}\right\rangle\\
&=\left\langle\mathbf{S},\mathbf{\Pi}_M-\mathring{\mathbf{\Pi}}^t\right\rangle+\left\|\mathbf{\Lambda}\odot\mathring{\mathbf{\Psi}}^t\right\|_1-\left\|\mathbf{\Lambda}\odot\mathbf{\Psi}_M\right\|_1-\left\langle\begin{bmatrix}\mathbf{\Pi}_M\\ \mathbf{\Psi}_M\\ \widetilde{\mathbf{Z}}^t\end{bmatrix},\begin{bmatrix}-\mathring{\mathbf{Z}}^t\\ \mathring{\mathbf{Z}}^t\\ \mathring{\mathbf{\Pi}}^t-\mathring{\mathbf{\Psi}}^t\end{bmatrix}\right\rangle\\
&\leq\frac{1}{t}\sum_{i=1}^{t}\left\{\left\langle\mathbf{S},\mathbf{\Pi}_M-\mathbf{\Pi}^i\right\rangle+\left\|\mathbf{\Lambda}\odot\mathbf{\Psi}^i\right\|_1-\left\|\mathbf{\Lambda}\odot\mathbf{\Psi}_M\right\|_1-\left\langle\begin{bmatrix}\mathbf{\Pi}_M\\ \mathbf{\Psi}_M\\ \widetilde{\mathbf{Z}}^t\end{bmatrix},\begin{bmatrix}-\mathbf{Z}^i\\ \mathbf{Z}^i\\ \mathbf{\Pi}^i-\mathbf{\Psi}^i\end{bmatrix}\right\rangle\right\}\\
&=\frac{1}{t}\sum_{i=1}^{t}\left\{\left\langle\mathbf{S},\mathbf{\Pi}_M-\mathbf{\Pi}^i\right\rangle+\left\|\mathbf{\Lambda}\odot\mathbf{\Psi}^i\right\|_1-\left\|\mathbf{\Lambda}\odot\mathbf{\Psi}_M\right\|_1-\left\langle\begin{bmatrix}\mathbf{\Pi}_M-\mathbf{\Pi}^i\\ \mathbf{\Psi}_M-\mathbf{\Psi}^i\\ \widetilde{\mathbf{Z}}^t-\mathbf{Z}^i\end{bmatrix},\begin{bmatrix}-\mathbf{Z}^i\\ \mathbf{Z}^i\\ \mathbf{\Pi}^i-\mathbf{\Psi}^i\end{bmatrix}\right\rangle\right\},
\end{aligned}\tag{12}$$

in which the inequality is due to the convexity of $\|\cdot\|_1$. For each term $i\in[1,t]$ in the summation of the last equality in (12), there should be

$$\begin{aligned}
&\left\langle\mathbf{S},\mathbf{\Pi}_M-\mathbf{\Pi}^i\right\rangle+\left\|\mathbf{\Lambda}\odot\mathbf{\Psi}^i\right\|_1-\left\|\mathbf{\Lambda}\odot\mathbf{\Psi}_M\right\|_1-\left\langle\begin{bmatrix}\mathbf{\Pi}_M-\mathbf{\Pi}^i\\ \mathbf{\Psi}_M-\mathbf{\Psi}^i\\ \widetilde{\mathbf{Z}}^t-\mathbf{Z}^i\end{bmatrix},\begin{bmatrix}-\mathbf{Z}^i\\ \mathbf{Z}^i\\ \mathbf{\Pi}^i-\mathbf{\Psi}^i\end{bmatrix}\right\rangle\\
=&\underbrace{\left\langle\mathbf{S},\mathbf{\Pi}_M-\mathbf{\Pi}^i\right\rangle+\left\langle\mathbf{Z}^i,\mathbf{\Pi}_M-\mathbf{\Pi}^i\right\rangle}_{\text{I}}+\underbrace{\left\|\mathbf{\Lambda}\odot\mathbf{\Psi}^i\right\|_1-\left\|\mathbf{\Lambda}\odot\mathbf{\Psi}_M\right\|_1+\left\langle\mathbf{Z}^i,\mathbf{\Psi}^i-\mathbf{\Psi}_M\right\rangle}_{\text{II}}+\underbrace{\left\langle\mathbf{\Pi}^i-\mathbf{\Psi}^i,\mathbf{Z}^i-\widetilde{\mathbf{Z}}^t\right\rangle}_{\text{III}}.
\end{aligned}\tag{13}$$

We bound the terms I, II and III respectively.

For term I in (13), we have

$$\begin{aligned}
\mathrm{I} =& \left\langle \mathbf{S}, \boldsymbol{\Pi}_M - \boldsymbol{\Pi}^i \right\rangle + \left\langle \mathbf{Z}^{i-1} - \rho\left(\boldsymbol{\Pi}^i - \boldsymbol{\Psi}^i\right), \boldsymbol{\Pi}_M - \boldsymbol{\Pi}^i \right\rangle \\
=& \left\langle \mathbf{S}, \boldsymbol{\Pi}_M - \boldsymbol{\Pi}^i \right\rangle + \left\langle \mathbf{Z}^{i-1} + \rho\left(\boldsymbol{\Psi}^{i-1} - \boldsymbol{\Pi}^i\right), \boldsymbol{\Pi}_M - \boldsymbol{\Pi}^i \right\rangle + \left\langle \rho\left(\boldsymbol{\Psi}^i - \boldsymbol{\Psi}^{i-1}\right), \boldsymbol{\Pi}_M - \boldsymbol{\Pi}^i \right\rangle \\
=& \left\langle \mathbf{S} + \mathbf{Z}^{i-1} + \rho\left(\boldsymbol{\Psi}^{i-1} - \boldsymbol{\Pi}^i\right), \boldsymbol{\Pi}_M - \boldsymbol{\Pi}^i \right\rangle + \left\langle \rho\left(\boldsymbol{\Psi}^i - \boldsymbol{\Psi}^{i-1}\right), \boldsymbol{\Pi}_M - \boldsymbol{\Pi}^i \right\rangle \\
\leq& \left\langle \rho\left(\boldsymbol{\Psi}^i - \boldsymbol{\Psi}^{i-1}\right), \boldsymbol{\Pi}_M - \boldsymbol{\Pi}^i \right\rangle \\
=& \frac{\rho}{2}\left(\left\|\boldsymbol{\Pi}_M - \boldsymbol{\Psi}^{i-1}\right\|_{\mathsf{F}}^2 - \left\|\boldsymbol{\Pi}_M - \boldsymbol{\Psi}^i\right\|_{\mathsf{F}}^2\right) + \frac{\rho}{2}\left(\left\|\boldsymbol{\Pi}^i - \boldsymbol{\Psi}^i\right\|_{\mathsf{F}}^2 - \left\|\boldsymbol{\Pi}^i - \boldsymbol{\Psi}^{i-1}\right\|_{\mathsf{F}}^2\right) \\
\leq& \frac{\rho}{2}\left(\left\|\boldsymbol{\Pi}_M - \boldsymbol{\Psi}^{i-1}\right\|_{\mathsf{F}}^2 - \left\|\boldsymbol{\Pi}_M - \boldsymbol{\Psi}^i\right\|_{\mathsf{F}}^2\right) + \frac{\rho}{2}\left\|\boldsymbol{\Pi}^i - \boldsymbol{\Psi}^i\right\|_{\mathsf{F}}^2 \\
=& \frac{\rho}{2}\left(\left\|\boldsymbol{\Pi}_M - \boldsymbol{\Psi}^{i-1}\right\|_{\mathsf{F}}^2 - \left\|\boldsymbol{\Pi}_M - \boldsymbol{\Psi}^i\right\|_{\mathsf{F}}^2\right) + \frac{1}{2\rho}\left\|\mathbf{Z}^{i-1} - \mathbf{Z}^i\right\|_{\mathsf{F}}^2,
\end{aligned}$$

in which the first and last equality follows from the update of dual variable in Line 6 of Algorithm 2, and the first inequality is derived from the first-order optimality condition for Line 4.

For term II in (13), we have

$$\begin{aligned}
\mathrm{II} =& \left\|\boldsymbol{\Lambda} \odot \boldsymbol{\Psi}^i\right\|_1 - \left\|\boldsymbol{\Lambda} \odot \boldsymbol{\Psi}_M\right\|_1 + \left\langle \mathbf{Z}^{i-1} - \rho\left(\boldsymbol{\Pi}^i - \boldsymbol{\Psi}^i\right), \boldsymbol{\Psi}^i - \boldsymbol{\Psi}_M \right\rangle \\
=& \left\|\boldsymbol{\Lambda} \odot \boldsymbol{\Psi}^i\right\|_1 - \left\|\boldsymbol{\Lambda} \odot \boldsymbol{\Psi}_M\right\|_1 - \left\langle \boldsymbol{\Lambda} \odot \boldsymbol{\Xi}^i, \boldsymbol{\Psi}^i - \boldsymbol{\Psi}_M \right\rangle, \ \exists \boldsymbol{\Xi}^i \in \partial\left\|\boldsymbol{\Psi}^i\right\|_1 \\
\leq& 0,
\end{aligned}$$

where the first equality follows from Line 6 of Algorithm 2, the second equality holds due to the first-order optimality condition for Line 5, and the last inequality is derived from the convexity of $\|\cdot\|_1$.

For term III in (13), utilizing the update of dual variable in Line 6 of Algorithm 2, we can get

$$\begin{aligned}
\mathrm{III} =& \frac{1}{\rho}\left\langle \mathbf{Z}^{i-1} - \mathbf{Z}^i, \mathbf{Z}^i - \widetilde{\mathbf{Z}}^t \right\rangle \\
=& \frac{1}{2\rho}\left(\left\|\widetilde{\mathbf{Z}}^t - \mathbf{Z}^{i-1}\right\|_{\mathsf{F}}^2 - \left\|\widetilde{\mathbf{Z}}^t - \mathbf{Z}^i\right\|_{\mathsf{F}}^2 - \left\|\mathbf{Z}^i - \mathbf{Z}^{i-1}\right\|_{\mathsf{F}}^2\right).
\end{aligned}$$

Finally, by substituting I, II, and III back into (12), we obtain

$$\begin{aligned}
\omega^t =& V^t\left(\boldsymbol{\Pi}_M, \boldsymbol{\Psi}_M, \widetilde{\mathbf{Z}}^t\right) \\
\leq& \frac{1}{t}\sum_{i=1}^{t}\left\{\frac{\rho}{2}\left(\left\|\boldsymbol{\Pi}_M - \boldsymbol{\Psi}^{i-1}\right\|_{\mathsf{F}}^2 - \left\|\boldsymbol{\Pi}_M - \boldsymbol{\Psi}^i\right\|_{\mathsf{F}}^2\right) + \frac{1}{2\rho}\left(\left\|\widetilde{\mathbf{Z}}^t - \mathbf{Z}^{i-1}\right\|_{\mathsf{F}}^2 - \left\|\widetilde{\mathbf{Z}}^t - \mathbf{Z}^i\right\|_{\mathsf{F}}^2\right)\right\} \\
=& \frac{1}{t}\left(\frac{\rho}{2}\left(\left\|\boldsymbol{\Pi}_M - \boldsymbol{\Psi}^0\right\|_{\mathsf{F}}^2 - \left\|\boldsymbol{\Pi}_M - \boldsymbol{\Psi}^t\right\|_{\mathsf{F}}^2\right) + \frac{1}{2\rho}\left(\left\|\widetilde{\mathbf{Z}}^t - \mathbf{Z}^0\right\|_{\mathsf{F}}^2 - \left\|\widetilde{\mathbf{Z}}^t - \mathbf{Z}^t\right\|_{\mathsf{F}}^2\right)\right) \\
\leq& \frac{1}{t}\left(\frac{\rho}{2}\left\|\boldsymbol{\Pi}_M - \boldsymbol{\Psi}^0\right\|_{\mathsf{F}}^2 + \frac{1}{2\rho}\left\|\widetilde{\mathbf{Z}}^t - \mathbf{Z}^0\right\|_{\mathsf{F}}^2\right) \\
=& \frac{1}{t}\left(\frac{\rho}{2}\left\|\boldsymbol{\Pi}_M\right\|_{\mathsf{F}}^2 + \frac{1}{2\rho}\left\|\widetilde{\mathbf{Z}}^t\right\|_{\mathsf{F}}^2\right) \\
\leq& \frac{1}{t}\left(\frac{\rho r}{2} + \frac{\lambda^2 d^2}{2\rho}\right),
\end{aligned}$$

where we use the settings $\boldsymbol{\Psi}^0 = \mathbf{0}$ and $\mathbf{Z}^0 = \mathbf{0}$ for the last equality. The last inequality follows from two facts:1) $\forall \boldsymbol{\Pi} \in \mathcal{F}$, $\|\boldsymbol{\Pi}\|_{\mathsf{F}}^2 \leq r$, and 2) $\left\|\widetilde{\mathbf{Z}}^t\right\|_{\mathsf{F}}^2 = \left\|\boldsymbol{\Lambda} \odot \operatorname{sgn}\left(\mathring{\boldsymbol{\Pi}}^t - \mathring{\boldsymbol{\Psi}}^t\right)\right\|_{\mathsf{F}}^2 \leq \lambda^2 d^2$.

Furthermore, if we set

$$\rho = \frac{\lambda d}{\sqrt{r}} = \arg\min_{z \in \mathbb{R}} \frac{zr}{2} + \frac{\lambda^2 d^2}{2z},$$

the stopping criterion is satisfied as $t \geq \frac{\lambda d \sqrt{r}}{\varepsilon}$. ■

B. Technical Lemmata

Before proving the next theorem, we first introduce several lemmata that will be used. For notational simplicity, we define $\|\mathbf{A}\|_{\max} := \max_{1 \leq i,j \leq d} |A_{ij}|$ and $\|\mathbf{A}\|_{\min} := \min_{1 \leq i,j \leq d} |A_{ij}|$ for a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$. We also use $\mathbf{A}_{\mathcal{E}}$ to denote a matrix whose (i,j)-th entry is equal to A_{ij} if (i,j) is contained within an index set $\mathcal{E}$ and zero otherwise. Let $\overline{\mathcal{E}}$ denote the complement of $\mathcal{E}$.

Lemma 4. *Suppose that Assumption 1 holds. Let*

$$\mathcal{E}^k := \mathcal{S}^* \cup \mathcal{S}^k \text{ and } \mathcal{S}^k := \left\{(i,j) : \Lambda_{ij}^{k-1} < p'_\lambda(u)\right\}, \text{ with } u := c\lambda \text{ and } k \geq 1,$$

where $c = \frac{2+\sqrt{2}}{\delta}$ *is specified in Assumption 1. If* $\lambda \geq 2\left\|\mathbf{W}\right\|_{\max} + \sqrt{2\delta\varepsilon}$ *holds, then there must be*

$$\left|\mathcal{E}^k\right| \leq 2s^2, \left\|\mathbf{\Lambda}_{\mathcal{E}^k}^{k-1}\right\|_{\min} \geq \frac{\lambda}{2}, \tag{14}$$

and

$$\left\|\widetilde{\mathbf{\Pi}}^k - \mathbf{\Pi}^*\right\|_{\mathsf{F}} \leq \frac{2\left(\|\mathbf{W}_{\mathcal{E}^k}\|_{\mathsf{F}} + \left\|\mathbf{\Lambda}_{\mathcal{S}^*}^{k-1}\right\|_{\mathsf{F}}\right) + \sqrt{2\delta\varepsilon}}{\delta} \leq \frac{2+\sqrt{2}}{\delta}\lambda s. \tag{15}$$

Lemma 4 provides deterministic estimation error bounds for the approximate solutions.

Lemma 5. *Under Assumption 2, there exists a constant* $C > 0$ *depending on* L *such that*

$$\max_{i,j} \mathbb{P}\left(|W_{ij}| \geq z\right) \leq 2\exp\left(-\frac{4nz^2}{\left(C\lambda_1\right)^2}\right)$$

for $0 \leq z \leq C\lambda_1$.

Lemma 6. *Under Assumption 2, the following relation holds:*

$$\|\mathbf{W}_{\mathcal{S}^*}\|_2 \lesssim \lambda_1\sqrt{\frac{s}{n}},$$

with probability at least $1 - \frac{2}{e^s}$.

Lemma 5 and Lemma 6 are concentration inequalities, serving as the foundations for the statistical analysis. They bound $\mathbf{W}$ in probability in terms of its elements magnitude and operator norm, respectively.

C. Proof of Theorem 3

Proof: We take

$$\lambda = 2C\lambda_1\sqrt{\frac{\log d}{n}} + \sqrt{\frac{\delta}{n}} \asymp \sqrt{\frac{\log d}{n}} \quad \text{and} \quad \varepsilon \leq \frac{1}{2n},$$

in which C is defined in Lemma 5. By Lemma 5, we have $\lambda \geq 2\left\|\mathbf{W}\right\|_{\max} + \sqrt{2\delta\varepsilon}$ holds with probability at least $1 - \frac{2}{d^2}$. Applying Lemma 2, we obtain

$$\begin{aligned}\left\|\widetilde{\mathbf{\Pi}}^K - \mathbf{\Pi}^*\right\|_{\mathsf{F}} &\leq \frac{2}{\delta}\left(\|\mathbf{W}_{\mathcal{S}^*}\|_{\mathsf{F}} + \frac{\sqrt{2\delta\varepsilon}}{2}\right) + \tau\left\|\widetilde{\mathbf{\Pi}}^{K-1} - \mathbf{\Pi}^*\right\|_{\mathsf{F}} \\ &\leq \frac{2}{(1-\tau)\,\delta}\left(\|\mathbf{W}_{\mathcal{S}^*}\|_{\mathsf{F}} + \frac{\sqrt{2\delta\varepsilon}}{2}\right) + \tau^{K-1}\left\|\widetilde{\mathbf{\Pi}}^1 - \mathbf{\Pi}^*\right\|_{\mathsf{F}} \\ &\leq \frac{2}{(1-\tau)\,\delta}\left(\|\mathbf{W}_{\mathcal{S}^*}\|_{\mathsf{F}} + \frac{\sqrt{2\delta\varepsilon}}{2}\right) + \frac{2+\sqrt{2}}{\delta}\tau^{K-1}\lambda s,\end{aligned}$$

where the last inequality follows from Lemma 4 for the special case $k = 1$. If $K \geq 1 + \frac{\log\frac{\lambda\sqrt{n}}{\lambda_1}}{\log\tau^{-1}} \gtrsim \log\log d$, we have

$$\tau^{K-1}\lambda s \leq \lambda_1 s\sqrt{\frac{1}{n}}.$$

We also have $\frac{\sqrt{2\delta\varepsilon}}{2} \lesssim \sqrt{\frac{1}{n}}$ since $\varepsilon \leq \frac{1}{2n}$, and $\|\mathbf{W}_{\mathcal{S}^*}\|_{\mathsf{F}} \leq \sqrt{s}\,\|\mathbf{W}_{\mathcal{S}^*}\|_2 \lesssim \lambda_1 s\sqrt{\frac{1}{n}}$ with probability at least $1 - \frac{2}{e^s}$ by Lemma 6. With the above results, we complete the proof. ■

APPENDIX C
PROOFS OF TECHNICAL LEMMATA

In this appendix, we will prove all lemmata, including those already presented and some newly introduced fundamental lemmata. In these newly introduced lemmata, Lemma 7, proposed in [4], bounds the curvature of the objective function along the Fantope and away from the truth. Lemma 8 and Lemma 9 characterize estimation error of the approximate solution $\tilde{\mathbf{\Pi}}$ to the reduced problem (4).

Lemma 7 (Lemma 3.1 in [4])**.** *For any $\mathbf{\Pi} \in \mathcal{F}$, the following relation holds:*

$$\frac{\delta}{2}\left\|\mathbf{\Pi}-\mathbf{\Pi}^*\right\|_{\mathsf{F}}^2 \le \left\langle \mathbf{\Sigma}, \mathbf{\Pi}^* - \mathbf{\Pi}\right\rangle .$$

Lemma 8. *For an approximate solution $\tilde{\mathbf{\Pi}}$ to problem (4), there should be*

$$\frac{\delta}{2}\left\|\tilde{\mathbf{\Pi}}-\mathbf{\Pi}^*\right\|_{\mathsf{F}}^2 \le \left\langle \mathbf{W}, \tilde{\mathbf{\Pi}} - \mathbf{\Pi}^*\right\rangle + \left\|\mathbf{\Lambda}\odot\mathbf{\Pi}^*\right\|_1 - \left\|\mathbf{\Lambda}\odot\tilde{\mathbf{\Pi}}\right\|_1 + \varepsilon.$$

Proof: Suppose Algorithm 2 executes T iterations in total, i.e., $\tilde{\mathbf{\Pi}} = \mathring{\mathbf{\Pi}}^T$. From Line 8 of Algorithm 2, we know that

$$\omega^T \le \varepsilon. \tag{16}$$

By the definition, we have

$$V^T\left(\mathbf{\Pi}^*, \mathbf{\Pi}^*, \widetilde{\mathbf{Z}}^T\right) \le \underset{\mathbf{\Pi}\in\mathcal{F}, \mathbf{\Psi}\in\mathbb{R}^{d\times d}, \mathbf{Z}=\widetilde{\mathbf{Z}}^T}{\text{maximize}} V^T\left(\mathbf{\Pi}, \mathbf{\Psi}, \mathbf{Z}\right) = \omega^T. \tag{17}$$

Combining (16) with (17), and expanding $V^T\left(\mathbf{\Pi}^*, \mathbf{\Pi}^*, \widetilde{\mathbf{Z}}^T\right)$, we obtain

$$\underbrace{\left\langle \mathbf{S}, \mathbf{\Pi}^* - \mathring{\mathbf{\Pi}}^T\right\rangle}_{\text{I}} + \underbrace{\left\|\mathbf{\Lambda}\odot\mathring{\mathbf{\Psi}}^T\right\|_1 - \left\|\mathbf{\Lambda}\odot\mathbf{\Pi}^*\right\|_1}_{\text{II}} \underbrace{- \left\langle \begin{bmatrix}\mathbf{\Pi}^* - \mathring{\mathbf{\Pi}}^T\\ \mathbf{\Pi}^* - \mathring{\mathbf{\Psi}}^T\\ \widetilde{\mathbf{Z}}^T - \mathring{\mathbf{Z}}^T\end{bmatrix}, \begin{bmatrix}-\mathring{\mathbf{Z}}^T\\ \mathring{\mathbf{Z}}^T\\ \mathring{\mathbf{\Pi}}^T - \mathring{\mathbf{\Psi}}^T\end{bmatrix}\right\rangle}_{\text{III}} \le \varepsilon. \tag{18}$$

Now we bound terms $\text{I}, \text{II}, \text{III}$ in (18) respectively.

For term I,

$$\begin{aligned}\text{I} =& \left\langle \mathbf{W}, \mathbf{\Pi}^* - \mathring{\mathbf{\Pi}}^T\right\rangle + \left\langle \mathbf{\Sigma}, \mathbf{\Pi}^* - \mathring{\mathbf{\Pi}}^T\right\rangle\\ \ge& \left\langle \mathbf{W}, \mathbf{\Pi}^* - \mathring{\mathbf{\Pi}}^T\right\rangle + \frac{\delta}{2}\left\|\mathring{\mathbf{\Pi}}^T - \mathbf{\Pi}^*\right\|_{\mathsf{F}}^2,\end{aligned} \tag{19}$$

in which the inequality holds by applying Lemma 7 and setting $\mathbf{\Pi} = \mathring{\mathbf{\Pi}}^T \in \mathcal{F}$.

For term II,

$$\begin{aligned}\text{II} =& \left\|\mathbf{\Lambda}\odot\left(\mathring{\mathbf{\Pi}}^T + \mathring{\mathbf{\Psi}}^T - \mathring{\mathbf{\Pi}}^T\right)\right\|_1 - \left\|\mathbf{\Lambda}\odot\mathbf{\Pi}^*\right\|_1\\ \ge& \left\|\mathbf{\Lambda}\odot\mathring{\mathbf{\Pi}}^T\right\|_1 - \left\|\mathbf{\Lambda}\odot\mathbf{\Pi}^*\right\|_1 - \left\|\mathbf{\Lambda}\odot\left(\mathring{\mathbf{\Psi}}^T - \mathring{\mathbf{\Pi}}^T\right)\right\|_1,\end{aligned} \tag{20}$$

where the inequality follows from the triangle inequality.

For term III,

$$\begin{aligned}\text{III} =& \left\langle \mathring{\mathbf{\Pi}}^T - \mathbf{\Pi}^*, \mathring{\mathbf{Z}}^T\right\rangle + \left\langle \mathbf{\Pi}^* - \mathring{\mathbf{\Psi}}^T, \mathring{\mathbf{Z}}^T\right\rangle + \left\langle \widetilde{\mathbf{Z}}^T - \mathring{\mathbf{Z}}^T, \mathring{\mathbf{\Pi}}^T - \mathring{\mathbf{\Psi}}^T\right\rangle\\ =& \left\langle \widetilde{\mathbf{Z}}^T, \mathring{\mathbf{\Pi}}^T - \mathring{\mathbf{\Psi}}^T\right\rangle\\ =& -\left\|\mathbf{\Lambda}\odot\left(\mathring{\mathbf{\Pi}}^T - \mathring{\mathbf{\Psi}}^T\right)\right\|_1,\end{aligned} \tag{21}$$

where the last equality holds by substituting $\widetilde{\mathbf{Z}}^T$.

Plugging (19), (20) and (21) into (18), we conclude the proof. ∎

Lemma 9. *Assume there exists a set $\mathcal{E}$ such that*

$$\mathcal{S}^* \subseteq \mathcal{E}, \left|\mathcal{E}\right| \le 2\left|\mathcal{S}^*\right| \le 2s^2 \text{ and } \left\|\mathbf{\Lambda}_{\overline{\mathcal{E}}}\right\|_{\min} \ge \frac{\lambda}{2}.$$

If $\lambda \geq 2\|\mathbf{W}\|_{\max} + \sqrt{2\delta\varepsilon}$*, then we have*

$$\left\|\widetilde{\boldsymbol{\Pi}} - \boldsymbol{\Pi}^*\right\|_{\mathsf{F}} \leq \frac{2\left(\|\mathbf{W}_{\mathcal{E}}\|_{\mathsf{F}} + \|\boldsymbol{\Lambda}_{\mathcal{S}^*}\|_{\mathsf{F}}\right) + \sqrt{2\delta\varepsilon}}{\delta} \leq \frac{2+\sqrt{2}}{\delta}\lambda s.$$

Proof: By Lemma 8, we have

$$\frac{\delta}{2}\left\|\widetilde{\boldsymbol{\Pi}} - \boldsymbol{\Pi}^*\right\|_{\mathsf{F}}^2 \leq \underbrace{\left\langle \mathbf{W}, \widetilde{\boldsymbol{\Pi}} - \boldsymbol{\Pi}^* \right\rangle}_{\mathrm{I}} + \underbrace{\|\boldsymbol{\Lambda} \odot \boldsymbol{\Pi}^*\|_1 - \left\|\boldsymbol{\Lambda} \odot \widetilde{\boldsymbol{\Pi}}\right\|_1}_{\mathrm{II}} + \varepsilon. \tag{22}$$

For term I in (22),

$$\begin{aligned}
\mathrm{I} &= \left\langle \mathbf{W}_{\mathcal{E}}, \left(\widetilde{\boldsymbol{\Pi}} - \boldsymbol{\Pi}^*\right)_{\mathcal{E}} \right\rangle + \left\langle \mathbf{W}_{\overline{\mathcal{E}}}, \left(\widetilde{\boldsymbol{\Pi}} - \boldsymbol{\Pi}^*\right)_{\overline{\mathcal{E}}} \right\rangle \\
&\leq \|\mathbf{W}_{\mathcal{E}}\|_{\mathsf{F}} \left\|\left(\widetilde{\boldsymbol{\Pi}} - \boldsymbol{\Pi}^*\right)_{\mathcal{E}}\right\|_{\mathsf{F}} + \|\mathbf{W}_{\overline{\mathcal{E}}}\|_{\max} \left\|\left(\widetilde{\boldsymbol{\Pi}} - \boldsymbol{\Pi}^*\right)_{\overline{\mathcal{E}}}\right\|_1 \\
&\leq \|\mathbf{W}_{\mathcal{E}}\|_{\mathsf{F}} \left\|\widetilde{\boldsymbol{\Pi}} - \boldsymbol{\Pi}^*\right\|_{\mathsf{F}} + \|\mathbf{W}\|_{\max} \left\|\left(\widetilde{\boldsymbol{\Pi}} - \boldsymbol{\Pi}^*\right)_{\overline{\mathcal{E}}}\right\|_1,
\end{aligned} \tag{23}$$

where the first inequality follows from the Hölder's inequality.

For II in (22),

$$\begin{aligned}
\mathrm{II} &= \|(\boldsymbol{\Lambda} \odot \boldsymbol{\Pi}^*)_{\mathcal{S}^*}\|_1 - \left\|\boldsymbol{\Lambda} \odot \widetilde{\boldsymbol{\Pi}}\right\|_1 \\
&= \|(\boldsymbol{\Lambda} \odot \boldsymbol{\Pi}^*)_{\mathcal{S}^*}\|_1 - \left\|\left(\boldsymbol{\Lambda} \odot \widetilde{\boldsymbol{\Pi}}\right)_{\mathcal{S}^*}\right\|_1 - \left\|\left(\boldsymbol{\Lambda} \odot \widetilde{\boldsymbol{\Pi}}\right)_{\overline{\mathcal{S}^*}}\right\|_1 \\
&\leq \left\|\left(\boldsymbol{\Lambda} \odot \left(\widetilde{\boldsymbol{\Pi}} - \boldsymbol{\Pi}^*\right)\right)_{\mathcal{S}^*}\right\|_1 - \left\|\left(\boldsymbol{\Lambda} \odot \widetilde{\boldsymbol{\Pi}}\right)_{\overline{\mathcal{S}^*}}\right\|_1 \\
&\leq \left\|\left(\boldsymbol{\Lambda} \odot \left(\widetilde{\boldsymbol{\Pi}} - \boldsymbol{\Pi}^*\right)\right)_{\mathcal{S}^*}\right\|_1 - \left\|\left(\boldsymbol{\Lambda} \odot \widetilde{\boldsymbol{\Pi}}\right)_{\overline{\mathcal{E}}}\right\|_1 \\
&\leq \|\boldsymbol{\Lambda}_{\mathcal{S}^*}\|_{\mathsf{F}} \left\|\left(\widetilde{\boldsymbol{\Pi}} - \boldsymbol{\Pi}^*\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}} - \left\|\left(\boldsymbol{\Lambda} \odot \widetilde{\boldsymbol{\Pi}}\right)_{\overline{\mathcal{E}}}\right\|_1 \\
&\leq \|\boldsymbol{\Lambda}_{\mathcal{S}^*}\|_{\mathsf{F}} \left\|\left(\widetilde{\boldsymbol{\Pi}} - \boldsymbol{\Pi}^*\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}} - \|\boldsymbol{\Lambda}_{\overline{\mathcal{E}}}\|_{\min} \left\|\widetilde{\boldsymbol{\Pi}}_{\overline{\mathcal{E}}}\right\|_1 \\
&= \|\boldsymbol{\Lambda}_{\mathcal{S}^*}\|_{\mathsf{F}} \left\|\left(\widetilde{\boldsymbol{\Pi}} - \boldsymbol{\Pi}^*\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}} - \|\boldsymbol{\Lambda}_{\overline{\mathcal{E}}}\|_{\min} \left\|\left(\widetilde{\boldsymbol{\Pi}} - \boldsymbol{\Pi}^*\right)_{\overline{\mathcal{E}}}\right\|_1 \\
&\leq \|\boldsymbol{\Lambda}_{\mathcal{S}^*}\|_{\mathsf{F}} \left\|\widetilde{\boldsymbol{\Pi}} - \boldsymbol{\Pi}^*\right\|_{\mathsf{F}} - \|\boldsymbol{\Lambda}_{\overline{\mathcal{E}}}\|_{\min} \left\|\left(\widetilde{\boldsymbol{\Pi}} - \boldsymbol{\Pi}^*\right)_{\overline{\mathcal{E}}}\right\|_1.
\end{aligned} \tag{24}$$

The first equality and the last equality are due to the fact that $\boldsymbol{\Pi}^*$ has non-zero elements only on its support $\mathcal{S}^*$. The first inequality follows from the triangular inequality and the third inequality follows from the Hölder's inequality.

Plugging (23) and (24) into (22), we obtain

$$\begin{aligned}
\frac{\delta}{2}\left\|\widetilde{\boldsymbol{\Pi}} - \boldsymbol{\Pi}^*\right\|_{\mathsf{F}}^2 &\leq \left(\|\mathbf{W}_{\mathcal{E}}\|_{\mathsf{F}} + \|\boldsymbol{\Lambda}_{\mathcal{S}^*}\|_{\mathsf{F}}\right)\left\|\widetilde{\boldsymbol{\Pi}} - \boldsymbol{\Pi}^*\right\|_{\mathsf{F}} + \left(\|\mathbf{W}\|_{\max} - \|\boldsymbol{\Lambda}_{\overline{\mathcal{E}}}\|_{\min}\right)\left\|\left(\widetilde{\boldsymbol{\Pi}} - \boldsymbol{\Pi}^*\right)_{\overline{\mathcal{E}}}\right\|_1 + \varepsilon \\
&\leq \left(\|\mathbf{W}_{\mathcal{E}}\|_{\mathsf{F}} + \|\boldsymbol{\Lambda}_{\mathcal{S}^*}\|_{\mathsf{F}}\right)\left\|\widetilde{\boldsymbol{\Pi}} - \boldsymbol{\Pi}^*\right\|_{\mathsf{F}} + \varepsilon,
\end{aligned}$$

in which we use the condition that $\|\boldsymbol{\Lambda}_{\overline{\mathcal{E}}}\|_{\min} \geq \frac{\lambda}{2} \geq \|\mathbf{W}\|_{\max}$ for the last inequality. Finally, by solving the quadratic inequality, we have

$$\begin{aligned}
\left\|\widetilde{\boldsymbol{\Pi}} - \boldsymbol{\Pi}^*\right\|_{\mathsf{F}} &\leq \frac{2\left(\|\mathbf{W}_{\mathcal{E}}\|_{\mathsf{F}} + \|\boldsymbol{\Lambda}_{\mathcal{S}^*}\|_{\mathsf{F}}\right) + \sqrt{2\delta\varepsilon}}{\delta} \\
&\leq \frac{2}{\delta}\left(\sqrt{|\mathcal{E}|}\,\|\mathbf{W}\|_{\max} + \sqrt{|\mathcal{E}|}\frac{\sqrt{2\delta\varepsilon}}{2} + \sqrt{|\mathcal{S}^*|}\lambda\right) \\
&\leq \frac{2+\sqrt{2}}{\delta}\lambda s,
\end{aligned}$$

in which the last two inequalities are derived from the conditions that $\sqrt{2|\mathcal{S}^*|} \geq \sqrt{|\mathcal{E}|} \geq \sqrt{|\mathcal{S}^*|} = s \geq 1$ and $\lambda \geq 2\|\mathbf{W}\|_{\max} + \sqrt{2\delta\varepsilon}$. ■

Proof of Lemma 4: Firstly, we prove (14) by induction. For $k = 1$, since $\Lambda_{ij}^0 = p'_\lambda\left(\left|\widetilde{\Pi}_{ij}^0\right|\right) = p'_\lambda(0) = \lambda \geq p'_\lambda(u)$, we have $\mathcal{S}^1 = \emptyset$ and $\mathcal{E}^1 = \mathcal{S}^*$ such that

$$\left|\mathcal{E}^1\right| \leq 2s^2 \quad \text{and} \quad \left\|\boldsymbol{\Lambda}_{\overline{\mathcal{E}^1}}^0\right\|_{\min} \geq \lambda.$$

Assume (14) holds for some $k-1$ with $k \geq 2$, we show that (14) also holds for k. According to the monotonicity of $p'_{\lambda}(\cdot)$, i.e., property (b) in Assumption 1, for any $(i,j) \in \mathcal{S}^k$, we have $\left|\widetilde{\Pi}_{ij}^{k-1}\right| \geq u$. Then there should be

$$\begin{aligned}\sqrt{\left|\mathcal{S}^k \backslash \mathcal{S}^*\right|} \leq & \sqrt{\sum_{(i,j)\in\mathcal{S}^k\backslash\mathcal{S}^*}\left(u^{-1}\widetilde{\Pi}_{ij}^{k-1}\right)^2} \\ \leq & u^{-1}\left\|\left(\widetilde{\mathbf{\Pi}}^{k-1}\right)_{\mathcal{S}^k\backslash\mathcal{S}^*}\right\|_{\mathsf{F}} \\ \leq & u^{-1}\left\|\left(\widetilde{\mathbf{\Pi}}^{k-1}-\mathbf{\Pi}^*\right)_{\mathcal{S}^k\backslash\mathcal{S}^*}\right\|_{\mathsf{F}} \\ \leq & u^{-1}\left\|\widetilde{\mathbf{\Pi}}^{k-1}-\mathbf{\Pi}^*\right\|_{\mathsf{F}},\end{aligned} \tag{25}$$

where the third inequality is due to the fact that $\mathbf{\Pi}^*_{\mathcal{S}^k\backslash\mathcal{S}^*} = \mathbf{0}$. Applying Lemma 9 with $\mathcal{E} = \mathcal{E}^{k-1}$, $\mathbf{\Lambda} = \mathbf{\Lambda}^{k-2}$ and $\widetilde{\mathbf{\Pi}} = \widetilde{\mathbf{\Pi}}^{k-1}$, we obtain

$$\left\|\widetilde{\mathbf{\Pi}}^{k-1}-\mathbf{\Pi}^*\right\|_{\mathsf{F}} \leq \frac{2+\sqrt{2}}{\delta}\lambda s. \tag{26}$$

Combining (25) with (26),

$$\sqrt{\left|\mathcal{S}^k\backslash\mathcal{S}^*\right|} \leq u^{-1}\frac{2+\sqrt{2}}{\delta}\lambda s = s,$$

which further implies that

$$\left|\mathcal{E}^k\right| = \left|\mathcal{S}^*\right| + \left|\mathcal{S}^k\backslash\mathcal{S}^*\right| \leq 2s^2.$$

By the property (d) in Assumption 1, we have

$$\left\|\mathbf{\Lambda}^{k-1}_{\mathcal{E}^k}\right\|_{\min} \geq \left\|\mathbf{\Lambda}^{k-1}_{\mathcal{S}^k}\right\|_{\min} \geq p'_{\lambda}(u) \geq \frac{\lambda}{2}.$$

As a consequence, (15) can be proven by directly applying Lemma 9 with $\mathcal{E} = \mathcal{E}^k$, $\mathbf{\Lambda} = \mathbf{\Lambda}^{k-1}$ and $\widetilde{\mathbf{\Pi}} = \widetilde{\mathbf{\Pi}}^k$. ■

Proof of Lemma 2: By Lemma 4, we have

$$\left\|\widetilde{\mathbf{\Pi}}^k-\mathbf{\Pi}^*\right\|_{\mathsf{F}} \leq \frac{2}{\delta}\left(\underbrace{\left\|\mathbf{W}_{\mathcal{E}^k}\right\|_{\mathsf{F}}}_{\text{I}} + \underbrace{\left\|\mathbf{\Lambda}^{k-1}_{\mathcal{S}^*}\right\|_{\mathsf{F}}}_{\text{II}} + \frac{\sqrt{2\delta\varepsilon}}{2}\right). \tag{27}$$

We then bound I and II respectively.

For I, we divide the set $\mathcal{E}^k$ into $\mathcal{S}^*$ and $\mathcal{E}^k\backslash\mathcal{S}^*$ to get

$$\begin{aligned}\text{I} \leq & \left\|\mathbf{W}_{\mathcal{S}^*}\right\|_{\mathsf{F}} + \left\|\mathbf{W}_{\mathcal{E}^k\backslash\mathcal{S}^*}\right\|_{\mathsf{F}} \\ \leq & \left\|\mathbf{W}_{\mathcal{S}^*}\right\|_{\mathsf{F}} + \sqrt{\left|\mathcal{E}^k\backslash\mathcal{S}^*\right|}\left\|\mathbf{W}\right\|_{\max} \\ \leq & \left\|\mathbf{W}_{\mathcal{S}^*}\right\|_{\mathsf{F}} + u^{-1}\left\|\widetilde{\mathbf{\Pi}}^{k-1}-\mathbf{\Pi}^*\right\|_{\mathsf{F}}\left\|\mathbf{W}\right\|_{\max} \\ \leq & \left\|\mathbf{W}_{\mathcal{S}^*}\right\|_{\mathsf{F}} + \frac{\lambda}{2}u^{-1}\left\|\widetilde{\mathbf{\Pi}}^{k-1}-\mathbf{\Pi}^*\right\|_{\mathsf{F}},\end{aligned} \tag{28}$$

where the third inequality follows from (25) and the last inequality follows from the condition that $\left\|\mathbf{W}\right\|_{\max} \leq \frac{\lambda}{2}$.

For II, under Assumption 1, if $\left|\widetilde{\Pi}_{ij}^{k-1}-\Pi_{ij}^*\right| \geq u$ we have

$$\left|\Lambda_{ij}^{k-1}\right| \leq \lambda \leq \lambda u^{-1}\left|\widetilde{\Pi}_{ij}^{k-1}-\Pi_{ij}^*\right|, \tag{29}$$

otherwise by Assumption 3 and the triangular inequality we have

$$\left|\widetilde{\Pi}_{ij}^{k-1}\right| \geq \left|\Pi_{ij}^*\right| - \left|\widetilde{\Pi}_{ij}^{k-1}-\Pi_{ij}^*\right| \geq \alpha\lambda,$$

which further implies

$$\left|\Lambda_{ij}^{k-1}\right| = p'_{\lambda}\left(\left|\widetilde{\Pi}_{ij}^{k-1}\right|\right) = 0. \tag{30}$$

Combining (29) with (30), we obtain

$$\begin{aligned}\mathrm{II} \leq& \lambda u^{-1}\left\|\left(\widetilde{\boldsymbol{\Pi}}^{k-1}-\boldsymbol{\Pi}^*\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}} \\ \leq& \lambda u^{-1}\left\|\widetilde{\boldsymbol{\Pi}}^{k-1}-\boldsymbol{\Pi}^*\right\|_{\mathsf{F}}.\end{aligned} \tag{31}$$

Finally, substituting (28) and (31) back into (27), we have

$$\begin{aligned}\left\|\widetilde{\boldsymbol{\Pi}}^{k}-\boldsymbol{\Pi}^*\right\|_{\mathsf{F}} \leq& \frac{2}{\delta}\left(\left\|\mathbf{W}_{\mathcal{S}^*}\right\|_{\mathsf{F}}+\frac{\sqrt{2\delta\varepsilon}}{2}+\frac{3\lambda}{2}u^{-1}\left\|\widetilde{\boldsymbol{\Pi}}^{k-1}-\boldsymbol{\Pi}^*\right\|_{\mathsf{F}}\right) \\ =&\frac{2}{\delta}\left(\left\|\mathbf{W}_{\mathcal{S}^*}\right\|_{\mathsf{F}}+\frac{\sqrt{2\delta\varepsilon}}{2}\right)+\tau\left\|\widetilde{\boldsymbol{\Pi}}^{k-1}-\boldsymbol{\Pi}^*\right\|_{\mathsf{F}},\end{aligned}$$

where $\tau=\frac{3}{2+\sqrt{2}}\in(0,1)$. ■

Proof of Lemma 5: See Corollary 3.3 in [4] for details. ■

Proof of Lemma 6: We first list two useful properties implied from Assumption 2. For $\forall \mathbf{v}\in\mathbb{R}^d\colon \mathbf{v}^\top\mathbf{v}=1$ and $z>0$, we have: 1) $\mathbb{P}\left(\left|\langle\mathbf{x},\mathbf{v}\rangle\right|\geq z\right)\leq 2\exp\left(-\frac{L}{\lambda_1}z^2\right)$, and 2) $\mathbb{E}\left[\exp\left(z\langle\mathbf{v},\mathbf{x}\rangle\right)\right]\leq\exp\left(C_1\frac{\lambda_1}{L}z^2\right)$ for some constant $C_1>0$. The first property is based on the observation that $\left\|\boldsymbol{\Sigma}^{1/2}\mathbf{v}\right\|_2^2\leq\lambda_1$, and the second property is an equivalent form of definition for sub-Gaussian variables [5].

Define

$$\mathcal{B}\left(d,\mathcal{S}^*\right):=\left\{\mathbf{v}\in\mathbb{R}^d\colon \mathbf{v}^\top\mathbf{v}=1,\operatorname{supp}\left(\mathbf{v}\mathbf{v}^\top\right)=\mathcal{S}^*\right\},$$

then we have

$$\begin{aligned}\left\|\mathbf{W}_{\mathcal{S}^*}\right\|_2=&\left\|\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbf{x}_i\mathbf{x}_i^\top-\mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i^\top\right]\right\}\right]_{\mathcal{S}^*}\right\|_2 \\ =&\sup_{\mathbf{v}\in\mathcal{B}(d,\mathcal{S}^*)}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\langle\mathbf{x}_i,\mathbf{v}\rangle^2-\mathbb{E}\langle\mathbf{x}_i,\mathbf{v}\rangle^2\right\}\right|.\end{aligned}$$

Define $\mu_{\mathbf{v}}:=\mathbb{E}\langle\mathbf{x}_i,\mathbf{v}\rangle^2$ for notational simplicity, and fix some $\mathbf{v}\in\mathcal{N}_{\frac{1}{8}}$ where $\mathcal{N}_{\frac{1}{8}}$ is an $\frac{1}{8}$-net of $\mathcal{B}\left(d,\mathcal{S}^*\right)$. The next proof follows from Proposition 1 in [6], and we present the details here for completeness. Specifically, since as $z\to 0$,

$$\begin{aligned}1+\frac{1}{2}\mu_{\mathbf{v}}z^2+o\left(z^2\right)=&\mathbb{E}\left[\exp\left(z\langle\mathbf{v},\mathbf{x}_i\rangle\right)\right] \\ \leq&\exp\left(C_1\frac{\lambda_1}{L}z^2\right) \\ =&1+\frac{1}{2}C_1\frac{\lambda_1}{L}z^2+o\left(z^2\right),\end{aligned}$$

we obtain $\mu_{\mathbf{v}}\frac{L}{\lambda_1}\leq C_1$. Then for any integer $m\geq 2$,

$$\begin{aligned}&\mathbb{E}\left[\left|\langle\mathbf{x}_i,\mathbf{v}\rangle^2-\mu_{\mathbf{v}}\right|^m\right] \\ \leq&\int_0^\infty\mathbb{P}\left\{\langle\mathbf{x}_i,\mathbf{v}\rangle^2-\mu_{\mathbf{v}}\geq z^{\frac{1}{m}}\right\}dz+\mu_{\mathbf{v}}^m \\ =&\int_0^\infty\mathbb{P}\left\{\left|\langle\mathbf{x}_i,\mathbf{v}\rangle\right|\geq\left(z^{\frac{1}{m}}+\mu_{\mathbf{v}}\right)^{\frac{1}{2}}\right\}dz+\mu_{\mathbf{v}}^m \\ \leq&2\int_0^\infty\exp\left(-\frac{L}{\lambda_1}\left(z^{\frac{1}{m}}+\mu_{\mathbf{v}}\right)\right)dz+\mu_{\mathbf{v}}^m \\ =&m!\left(\frac{\lambda_1}{L}\right)^m\left(2\exp\left(-\frac{L}{\lambda_1}\mu_{\mathbf{v}}\right)+\frac{1}{m!}\left(\frac{L\mu_{\mathbf{v}}}{\lambda_1}\right)^m\right) \\ \leq&C_2m!\left(\frac{\lambda_1}{L}\right)^m,\end{aligned}$$

in which C_2 is a constant depending on C_1. The last inequality follows from the fact that $\exp\left(-\frac{L}{\lambda_1}\mu_{\mathbf{v}}\right) \leq 1$ and $\frac{1}{m!}C_1^m$ is bounded above. By the Bernstein's inequality (see Lemma 2.2.11 in [7]),

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\langle\mathbf{x}_i,\mathbf{v}\rangle^2-\mu_{\mathbf{v}}\right\}\right|>z\right)\\
=&\mathbb{P}\left(\left|\sum_{i=1}^{n}\left\{\langle\mathbf{x}_i,\mathbf{v}\rangle^2-\mu_{\mathbf{v}}\right\}\right|>nz\right)\\
\leq&2\exp\left(-\frac{nz^2}{4C_2\left(\frac{\lambda_1}{L}\right)^2+\frac{2\lambda_1}{L}z}\right).
\end{aligned}$$

There exists a constant C_3 satisfying $C_3\lambda_1 \geq 4\left(1+\sqrt{1+C_2}\right)\frac{\lambda_1}{L}$ such that for $0 \leq z \leq C_3\lambda_1$,

$$\exp\left(-\frac{nz^2}{4C_2\left(\frac{\lambda_1}{L}\right)^2+\frac{2\lambda_1}{L}z}\right) \leq \exp\left(-\frac{4nz^2}{\left(C_3\lambda_1\right)^2}\right).$$

Now we unfix $\mathbf{v}$, and consider the event

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\langle\mathbf{x}_i,\mathbf{v}\rangle^2-\mathbb{E}\left[\langle\mathbf{x}_i,\mathbf{v}\rangle^2\right]\right\}\right| \leq z, \quad \forall\mathbf{v}\in\mathcal{N}_{\frac{1}{8}}.$$

It holds at least

$$1-2\left|\mathcal{N}_{\frac{1}{8}}\right|\exp\left(-\frac{4nz^2}{\left(C_3\lambda_1\right)^2}\right) \geq 1-2\exp\left(3s-\frac{4nz^2}{\left(C_3\lambda_1\right)^2}\right),$$

in which we utilize the fact that $\left|\mathcal{N}_{\frac{1}{8}}\right| \leq 17^s$ (see Example 5.8 in [8]). We finish the proof by using Lemma 2.2 in [9] and taking $z = C_3\lambda_1\sqrt{\frac{s}{n}} \leq C_3\lambda_1$. ■

References

[1] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[2] B. He, H. Liu, Z. Wang, and X. Yuan, "A strictly contractive peaceman–rachford splitting method for convex programming," *SIAM Journal on Optimization*, vol. 24, no. 3, pp. 1011–1040, 2014.

[3] J. Lei and V. Q. Vu, "Sparsistency and agnostic inference in sparse PCA," *The Annals of Statistics*, vol. 43, no. 1, pp. 299 – 322, 2015.

[4] V. Q. Vu, J. Cho, J. Lei, and K. Rohe, "Fantope projection and selection: A near-optimal convex relaxation of sparse PCA," in *Advances in Neural Information Processing Systems*, vol. 26. Curran Associates, Inc., 2013.

[5] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018, vol. 47.

[6] T. Wang, Q. Berthet, and R. J. Samworth, "Statistical and computational trade-offs in estimation of sparse principal components," *The Annals of Statistics*, vol. 44, no. 5, pp. 1896 – 1930, 2016.

[7] J. Wellner *et al.*, *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 2013.

[8] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge university press, 2019, vol. 48.

[9] R. Vershynin, "How close is the sample covariance matrix to the actual covariance matrix?" *Journal of Theoretical Probability*, vol. 25, no. 3, pp. 655–686, 2012.