Robust Large Covariance Matrix Estimation Under Huber Loss

Jialin Yu

***Abstract*—XXX**

***Index Terms*—XXX**

I. Introduction

XXX The estimation of covariance matrices is a fundamental problem in modern multivariate data analysis. It has broad applications in many fields such as statistics [1], biology [2], finance [3], [4], [5], signal processing [6], [7], [8], [9], [10], machine learning [11], etc. For example, many dimension reduction techniques, including principal component analysis [12] and linear and quadratic discriminant analysis [13], require the estimation of a covariance matrix in advance from a given collection of data points. Other prominent examples include portfolio optimization [14] and beamformer design [15]. When the dimension of the covariance matrix becomes large, the estimation problem is generally challenging. It is well-known that when the dimension is larger than the sample size, the commonly used sample covariance matrix (SCM) is singular, which may cause trouble in many applications. In addition, the number of parameters to be estimated grows quadratically with the dimension of the covariance matrix. Therefore, the problem of large (or high-dimensional) covariance matrix estimation has received considerable attention over the past decade.

In order to estimate large covariance matrices effectively, one of the most popular assumptions is sparsity, i.e., a majority of the off-diagonal elements are nearly zeros, which largely reduces the number of parameters to be estimated. The sparsity assumption is reasonable in real applications [16]. A commonly used method for estimating sparse covariance matrices is called thresholding [17], [18], [19], which is to set small elements in the SCM to zeros. Some statistical properties of the thresholding covariance estimator, including minimax lower bounds [20] and rates of convergence [21], have been established in the literature. The thresholding covariance estimator is proved only to be asymptotically positive definite [17], [18]. However, in practice it is more desirable to require the positive definiteness of the estimator under finite samples. Besides the thresholding methods for general sparse covariance matrix estimation, when covariances have sparsely banded structures, specialized covariance estimation methods include banding [22] and tapering [23], [24], [25] have been proposed. Nevertheless, this paper will focus on the general sparse cases.

Theoretical properties of large covariance estimators discussed in the literature often hinge heavily on the Gaussian or sub-Gaussian1 assumption ([26]). See, for example, Corollary 8 of [26]. Such an assumption is typically very restrictive in practice. For example, a recent fMRI study by Eklund et al. (2016) reported that most of the common software packages for fMRI analysis, such as SPM and FSL, can result in inflated false-positive rates up to 70% under 5% nominal levels, and questioned a number of fMRI studies among approximately 40,000 studies according to PubMed. Their results suggested that The principal cause of the invalid cluster inferences is spatial autocorrelation functions that do not follow the assumed Gaussian shape. Eklund et al. (2016) plotted the empirical versus theoretical spatial autocorrelation functions for several datasets. The empirical autocorrelation functions have much heavier tails compared to their theoretical counterparts under the commonly used assumption of a Gaussian random field, which causes the failure of fMRI inferences. Similar phenomenon has also been discovered in genomic studies (Liu et al., 2003; Purdom and Holmes, 2005) and in quantitative finance (Cont, 2001). It is therefore imperative to develop robust inferential procedures that are less sensitive to the distributional assumptions.

(Or we substitute the previous paragraph with the following explanation from Adaptive Huber Regression) The sub-Gaussian tails requirement, albeit being convenient for theoretical analysis, is not realistic in many practical applications since modern data are often collected with low quality. For example, a recent study on functional magnetic resonance imaging (fMRI) (Eklund, Nichols and Knutsson, 2016) shows that the principal cause of invalid fMRI inferences is that the data do not follow the assumed Gaussian shape, which speaks to the need of validating the statistical methods being used in the field of neuroimaging. In a microarray data example considered in Wang, Peng and Li (2015), it is observed that some gene expression levels have heavy tails as their kurtosises are much larger than 3, despite of the normalization methods used. In finance, the power-law nature of the distribution of returns has been validated as a stylized fact (Cont, 2001). Fan et al. (2016) argued that heavy-tailed distribution is a stylized feature for high dimensional data and proposed a shrinkage principle to attenuate the influence of outliers. Standard statistical procedures that are based on the method of least squares often behave poorly in the presence of heavy-tailed data2 (Catoni, 2012). It is therefore of ever-increasing interest to develop new statistical methods that are robust against heavy-tailed errors and other potential forms of contamination.)

There are two ways to adjust to heavy-tailed data in existing literature: [27] assumes the polynomial-tail condition and proposes a quadratic loss ℓ_1 penalized robust covariance estimator, and methods summarized in [28] that assume finite fourth moments condition and use robust (Huber) loss in their formulations. However, neither of them achieves the oracle rate

in [26]. It is then imperative to find a formulation that achieves the oracle rate in the presence of heavy-tailedness. Implementing the robust loss with ℓ_1 penality, however, fails to boost the performance significantly as we will demonstrate via (example or numerical analysis?). Given that the existing literature that achieves the oracle rate i.e.[26] possesses concave penality, we propose a formulation with Huber loss and concave penality. Under this formulation, the resulting estimator assumes only the polynomial-tail condition and achieves the oracle rate.

A. Notations

The following notation is adopted. Standard lower-case or upper-case letters stand for scalars and boldface lower-case (upper-case) letters denote vectors (matrices). Both X_{ij} and $[\mathbf{X}]_{ij}$ denote the (i,j)-th entry of the matrix $\mathbf{X}$. $\mathbb{R}_+$ denotes the set of non-negative real numbers, $\mathbb{R}^{m\times n}$ denotes the set of real $m\times n$ matrices. $\mathbf{0}$ and $\mathbf{1}$ stand for the all-zero and all-one vector/matrix, respectively. $\mathbf{I}$ stands for the identity matrix. $\mathbf{X}\succ\mathbf{0}$ ($\mathbf{X}\succeq\mathbf{0}$) means $\mathbf{X}$ is positive definite (semidefinite). $\mathbf{x}\geq\mathbf{0}$ denotes each element of $\mathbf{x}$ is non-negative.

Let $\|\mathbf{X}\|_\infty=\max_{k,l}|X_{kl}|$ and $\|\mathbf{X}\|_{\min}=\min_{k,l}|X_{kl}|$. Let $\|\mathbf{X}\|_{1,\text{off}}=\sum_{k\neq l}|X_{kl}|$ denote the sum-absolute-value norm for all entries and for off-diagonals. We write $[d]$ for the set $\{1,2,\ldots,d\}$ and $\lfloor x\rfloor$ for the largest integer not exceeding x. For an index set $\mathcal{E}$, we use $|\mathcal{E}|$ to denote its cardinality, $\overline{\mathcal{E}}$ to denote its complement. Use $\mathbf{X}_\mathcal{E}$ to denote the matrix whose (i,j)-th entry is equal to X_{ij} if $(i,j)\in\mathcal{E}$, and zero otherwise. Let $\mathbf{A}\circ\mathbf{B}$ denote the Hadamard product between matrix $\mathbf{A}$ and $\mathbf{B}$. Let $\partial f(\cdot)$ denote the subdifferential of a multivariate function f.

Let $\operatorname{sgn}(x)$ denote the sign of variable x, i.e., $\operatorname{sgn}(x)=x/|x|$. For functions $f(n)$ and $g(n)$, we denote $f(n)\lesssim g(n)$ if $f(n)\leq Cg(n)$, $f(n)\gtrsim g(n)$ if $f(n)\geq cg(n)$ and $f(n)\asymp g(n)$ if $cg(n)\leq f(n)\leq Cg(n)$ for some positive constants c and C.

II. Problem Formulation

Given samples $\mathbf{x}_i$, $i=1,\ldots,n$ from a heavy-tailed distribution with covariance matrix $\mathbf{\Sigma}^*$, let $N:=n(n-1)/2$ and define the paired data

$$\{\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N\}=\{\mathbf{x}_1-\mathbf{x}_2,\mathbf{x}_1-\mathbf{x}_3,\ldots,\mathbf{x}_{n-1}-\mathbf{x}_n\},$$

which are identically distributed from a random vector $\mathbf{y}$ with mean $\mathbf{0}$ and covariance matrix $\operatorname{cov}(\mathbf{y})=2\mathbf{\Sigma}^*$. Let

$$\begin{aligned}L_\alpha(\mathbf{\Sigma}):&=\sum_{k,\ell}\frac{1}{N}\sum_{1\leq i<j\leq n}\rho_\alpha(\Sigma_{k\ell}-(x_{ik}-x_{jk})(x_{il}-x_{jl})/2)\\&=\sum_{k,\ell}\frac{1}{N}\sum_{m=1}^{N}\rho_\alpha(\Sigma_{k\ell}-y_{mk}y_{m\ell}/2)\end{aligned}$$

with $\rho_\alpha:\mathbb{R}\to\mathbb{R}_+$ a Huber loss function defined as

$$\rho_\alpha(x)=\begin{cases}x^2/2 & \text{if } |x|\leq\alpha,\\ \alpha|x|-\alpha^2/2 & \text{if } |x|>\alpha,\end{cases}\tag{1}$$

where $\alpha>0$ is a robustification parameter. Compared with the squared error loss, large values of x are down-weighted in the Huber loss, yielding robustness. Generally speaking, minimizing Huber's loss produces a biased estimator of the mean, and parameter α can be chosen to control the bias. In other words, α quantifies the tradeoff between bias and robustness. As observed by Sun, Zhou and Fan (2018), in order to achieve an optimal tradeoff, α should adapt to the sample size, the dimension, and the noise level.

We will consider the following covariance estimation problem

$$\min_{\mathbf{\Sigma}\succ\mathbf{0}}\left\{L_\alpha(\mathbf{\Sigma})-\tau\log\det\mathbf{\Sigma}+\sum_{k\neq\ell}p_\lambda(|\Sigma_{k\ell}|)\right\},\tag{2}$$

where $\tau\log\det$ is a positive-definiteness penalty function with a regularization parameter $\tau>0$ and $p_\lambda:\mathbb{R}_+\to\mathbb{R}_+$ is a non-convex penalty function with a regularization parameter $\lambda>0$. We consider a class of non-convex penalty functions $p_\lambda(\cdot)$ satisfying the following assumptions.

Assumption 1. *The function $p_\lambda(x)$ defined on $[0,+\infty)$ satisfies:*

(a) *$p_\lambda(x)=\lambda^2p(x/\lambda)$ with some function $p(x)$ defined on $[0,+\infty)$.*
(b) *$p(x)$ is non-decreasing and concave on $[0,+\infty)$ with $p(0)=0$ and is differentiable almost everywhere on $(0,+\infty)$;*
(c) *$0\leq p'(x_1)\leq p'(x_2)\leq\lambda$ for all $x_1\geq x_2\geq 0$ and $\lim_{x\to 0}p'(x)=1$.*
(d) *There exists some $\phi_1>\phi_0>\sqrt{5}$ such that $p'(\phi_0)>0$ and $p'(x)=0,\forall x\geq\phi_1$.*

III. Algorithm

A. The MM Algorithmic Framework: A Brief Review

Consider the minimization of a continuous function $F(\mathbf{y})$. Initialized as $\mathbf{y}^{(0)}$, the MM algorithm [29], [30] generates a sequence of feasible points $\{\mathbf{y}^{(t)}\}_{t\geq 1}$ by the following induction. At point $\mathbf{y}^{(t-1)}$, in the majorization step, we design a surrogate function $\bar{F}(\mathbf{y}\mid\mathbf{y}^{(t-1)})$ that locally approximates the objective function $F(\mathbf{y})$, satisfying

$$\begin{cases}\bar{F}(\mathbf{y}\mid\mathbf{y}^{(t-1)})\geq F(\mathbf{y}),\\ \bar{F}(\mathbf{y}^{(t-1)}\mid\mathbf{y}^{(t-1)})=F(\mathbf{y}^{(t-1)}).\end{cases}$$

Then, in the minimization step, we update $\mathbf{y}^{(t)}$ as

$$\mathbf{y}^{(t)}\in\arg\min_{\mathbf{y}}\left\{\bar{F}(\mathbf{y}\mid\mathbf{y}^{(t-1)})\right\}.$$

B. Robust covariance matrix estimation via MM

We follow the MM framework to solve (2). In each iteration of MM, we find a weighted ℓ_1 surrogate function of $\sum_{i\neq j}p_\lambda(|\Sigma_{ij}|)$. Consequently, we consider a multistage procedure that solves a sequence of convex relaxation subproblems, which is also known as an iteratively re-weighted ℓ_1 algorithm [31] or a difference of convex algorithm [32]. Specifically, starting with an initial estimate $\widehat{\mathbf{\Sigma}}^{(0)}$, we consider a sequence of convex optimization problems

$$\min_{\mathbf{\Sigma}\succ\mathbf{0}}\left\{L_\alpha(\mathbf{\Sigma})-\tau\log\det\mathbf{\Sigma}+\sum_{k\neq l}p'_\lambda(|\widehat{\Sigma}^{(t-1)}_{kl}|)|\Sigma_{kl}|\right\},\tag{3}$$

where $t = 1, 2, \dots$, and $\widehat{\mathbf{\Sigma}}^{(t)}$ is the optimal solution to the t-th subproblem.

Each subproblem in (3) corresponds to a weighted ℓ_1 penalized covariance estimation problem, which generally can be written in the following form:

$$\min_{\mathbf{\Sigma} \succ \mathbf{0}} \left\{ L_\alpha(\mathbf{\Sigma}) - \tau \log\det \mathbf{\Sigma} + \|\mathbf{\Lambda} \circ \mathbf{\Sigma}\|_{1,\text{off}} \right\} \tag{4}$$

where $\mathbf{\Lambda}$ is a $d \times d$ matrix of regularization parameters with $\Lambda_{ij} \in [0, \lambda]$. By convex optimization theory, any optimal solution $\widehat{\mathbf{\Sigma}}$ to (4) satisfies the following first-order optimality condition:

$$\nabla L_\alpha(\widehat{\mathbf{\Sigma}}) - \tau \widehat{\mathbf{\Sigma}}^{-1} + \mathbf{\Lambda} \circ \mathbf{\Xi} = \mathbf{0}, \text{ with } \mathbf{\Xi} \in \partial \|\widehat{\mathbf{\Sigma}}\|_{1,\text{off}}.$$

Since analytical solution does not exist for (4), the exact solution $\widehat{\mathbf{\Sigma}}$ can never be achieved. In practice, due to optimization error from iterative methods, we define the ϵ-optimal solution to problem (4) as follows:

Definition 2. For a pre-specified tolerance level $\epsilon > 0$, we say $\widetilde{\mathbf{\Sigma}}$ is an ϵ-optimal solution to (4) if

$$\min_{\mathbf{\Xi} \in \partial \|\widetilde{\mathbf{\Sigma}}\|_{1,\text{off}}} \left\| \nabla L_\alpha(\widetilde{\mathbf{\Sigma}}) - \tau \widetilde{\mathbf{\Sigma}}^{-1} + \mathbf{\Lambda} \circ \mathbf{\Xi} \right\|_\infty \le \epsilon.$$

We use $\widetilde{\mathbf{\Sigma}}^{(t)}$ to denote an ϵ-optimal solution to the t-th subproblem, which is given by

$$\min_{\mathbf{\Sigma} \succ \mathbf{0}} \left\{ L_\alpha(\mathbf{\Sigma}) - \tau \log\det \mathbf{\Sigma} + \left\| \mathbf{\Lambda}^{(t-1)} \circ \mathbf{\Sigma} \right\|_{1,\text{off}} \right\}, \tag{5}$$

where $\Lambda_{kl}^{(t-1)} = p'_\lambda(|\widetilde{\Sigma}_{kl}^{(t-1)}|)$ for all $k, l \in [d]$.

IV. Theoretical Results

Recall that the underlying true covariance matrix is denoted by $\mathbf{\Sigma}^*$. Let $\mathcal{S} = \{(k,l) \mid \Sigma^*_{kl} \neq 0\}$ be the support set of $\mathbf{\Sigma}^*$ and s be its cardinality, i.e., $s = |\mathcal{S}|$. In the following, we impose some mild conditions on the true covariance matrix $\mathbf{\Sigma}^*$ and the distribution of the i.i.d. samples $\mathbf{x}_i$, $i = 1, \dots, n$.

Assumption 3. *For the true covariance matrix, assume* $\mathbf{\Sigma}^* \succ \mathbf{0}$.

Assumption 4. $\mathbf{x}_i \in \mathbb{R}^d$ *is a heavy-tailed random variable, i.e.* $\mathbb{E}\left[|x_{ij}|^{4(1+\gamma)}\right] \le \sigma_x^{2(1+\gamma)}$ *for all* $1 \le j \le d$ *with some positive* σ_x.

We assume Assumptions 1, 3, and 4 hold for the rest of the paper. (declare this within every result.)

Definition 5. Given $\mathbb{B}^\infty(r) := \{\mathbf{\Delta} \in \mathbb{R}^{d \times d} : \|\mathbf{\Delta}\|_\infty \le r\}$. We define $\mathcal{E}_1(r, \kappa)$ for any $\mathbf{\Sigma} \in \mathbf{\Sigma}^* + \mathbb{B}^\infty(r)$,

$$\langle \nabla L_\alpha(\mathbf{\Sigma}) - \nabla L_\alpha(\mathbf{\Sigma}^*), \mathbf{\Sigma} - \mathbf{\Sigma}^* \rangle \ge \kappa \|\mathbf{\Sigma} - \mathbf{\Sigma}^*\|_{\text{F}}^2.$$

Proposition 6. *Suppose that Assumptions 1 and 3 hold. Let* $\tau, \lambda, r > 0$ *satisfy*

$$\tau \le \frac{\lambda s^{1/2}}{\|(\mathbf{\Sigma}^*)^{-1}\|_F}, \qquad r > 5\lambda s^{1/2}.$$

Then, conditioned on the event $\mathcal{E}_1(r, 1/2) \cap \{\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty + \epsilon \le 0.5\lambda\}$, *any* ϵ*-optimal solution to* (4) *satisfies*

$$\left\| \widetilde{\mathbf{\Sigma}}^{(1)} - \mathbf{\Sigma}^* \right\|_F \le 5\lambda s^{1/2}$$

If we take $\lambda \asymp \sqrt{\frac{\log d}{n}}$ in Proposition 6, then we can see that the rate of convergence $\sqrt{\frac{s \log d}{n}}$ is the same as that obtained by Wei's paper.

Definition 7. Let $\mathbb{C}(l) := \{\mathbf{\Delta} \in \mathbb{R}^{d \times d} : \|\mathbf{\Delta}\|_1 \le l \|\mathbf{\Delta}\|_{\text{F}}\}$.

The definition of $\mathbb{C}(l)$ converts a bound for $\left\|\widetilde{\mathbf{\Sigma}}^{(T)} - \mathbf{\Sigma}^*\right\|_{\text{F}}$ into a bound for $\left\|\widetilde{\mathbf{\Sigma}}^{(T)} - \mathbf{\Sigma}^*\right\|_1$. Assume that we have bounded $\left\|\widetilde{\mathbf{\Sigma}}^{(T)} - \mathbf{\Sigma}^*\right\|_{\text{F}}$, i.e. $\left\|\widetilde{\mathbf{\Sigma}}^{(T)} - \mathbf{\Sigma}^*\right\|_{\text{F}} \le M$ for some $M > 0$. Then, given that

$$\widetilde{\mathbf{\Sigma}}^{(T)} \in \mathbf{\Sigma}^* + \mathbb{C}(l), \tag{6}$$

we have $\left\|\widetilde{\mathbf{\Sigma}}^{(T)} - \mathbf{\Sigma}^*\right\|_1 \le lM$. To gaurantee (6), a trivial choice would be $l \asymp d$, but we will show that $l \asymp s^{1/2}$ is actually enough.

The following Proposition demonstrates the contraction property of the solution path $\left\{\widetilde{\mathbf{\Sigma}}^{(t)}\right\}_{t \ge 1}$.

Proposition 8. *Suppose that Assumptions 1 and 3 hold. By Assumption 1, there exists some* $\phi_0 > \sqrt{5}$ *such that* $p'(\phi_0) > 0$. *Let*

$$\tau \le \frac{\lambda s^{1/2}}{\|(\mathbf{\Sigma}^*)^{-1}\|_F} \tag{7}$$

and choose $c > 0$ *so that*

$$0.5p'(\phi_0)(c^2+1)^{1/2} + 2 = 0.5c\phi_0 \tag{8}$$

Set $l = (2 + \frac{2}{p'(\phi_0)})(c^2+1)^{1/2} s^{1/2} + \frac{2}{p'(\phi_0)} s^{1/2}$ *and let* $r > 0$ *satisfy*

$$c\phi_0 \lambda s^{1/2} \le r \tag{9}$$

Under the minimal signal strength condition $\|\mathbf{\Sigma}^*_{\mathcal{S}}\|_{\min} \ge \phi_0 \lambda$ *and conditioned on event* $\mathcal{E}_1(r, 1/2) \cap \{\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty + \epsilon \le 0.5p'(\phi_0)\lambda\}$, *any* ϵ*-optimal solution* $\widetilde{\mathbf{\Sigma}}^{(t)}$ *to* (5) *satisfies* $\widetilde{\mathbf{\Sigma}}^{(t)} \in \mathbf{\Sigma}^* + \mathbb{C}(l)$, *and we have*

$$\left\| \widetilde{\mathbf{\Sigma}}^{(t)} - \mathbf{\Sigma}^* \right\|_F \le \delta \left\| \widetilde{\mathbf{\Sigma}}^{(t-1)} - \mathbf{\Sigma}^* \right\|_F + 2\Big\{ \|p'_\lambda(|\mathbf{\Sigma}^*_{\mathcal{S}}| - \phi_0\lambda)\|_F + s^{1/2}\epsilon + \|\nabla L_\alpha(\mathbf{\Sigma}^*)_{\mathcal{S}}\|_F + \tau \left\|(\mathbf{\Sigma}^*)^{-1}\right\|_F \Big\} \tag{10}$$

where $\delta = \sqrt{5}/\phi_0 \in (0, 1)$.

Remark 9. Let

$$r^{\text{ora}} := 2\left\{ \|p'_\lambda(|\mathbf{\Sigma}^*_{\mathcal{S}}| - \phi_0\lambda)\|_{\text{F}} + \|\nabla L_\alpha(\mathbf{\Sigma}^*)_{\mathcal{S}}\|_{\text{F}} + s^{1/2}\epsilon + \tau \left\|(\mathbf{\Sigma}^*)^{-1}\right\|_{\text{F}} \right\},$$

then Proposition 2.2 states that

$$\left\| \widetilde{\mathbf{\Sigma}}^{(t)} - \mathbf{\Sigma}^* \right\|_{\text{F}} \le \delta \left\| \widetilde{\mathbf{\Sigma}}^{(t-1)} - \mathbf{\Sigma}^* \right\|_{\text{F}} + r^{\text{ora}},$$

where $\delta = \sqrt{5}/\phi_0 \in (0,1)$. And by induction we further have

$$\left\|\widetilde{\mathbf{\Sigma}}^{(t)} - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} \leq \delta^{t-1}\left\|\widetilde{\mathbf{\Sigma}}^{(1)} - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} + (1-\delta)^{-1} r^{\mathrm{ora}}.$$

In Proposition 8, we have demonstrated the exact choice of l in 6, which satisfies $l \asymp s^{1/2}$. Meanwhile, we have shown that the estimation error between the ϵ-optimal solution $\widetilde{\mathbf{\Sigma}}^{(t)}$ and the true covariance $\mathbf{\Sigma}^*$ is bounded by two terms, namely, r^{ora} and a contraction term.

Lemma 10. *Suppose that Assumption 4 holds. Define* $\epsilon_{kl} = \Sigma^*_{k\ell} - y_{mk}y_{m\ell}/2$ *for all* $k,l \in [d]$. *Then* $\max\left\{\mathbb{E}\left[\epsilon_{kl}^2\right], \mathbb{E}\left[\epsilon_{kl}^{2(1+\gamma)}\right]\right\} \leq K$ *for all* $k,l \in [d]$ *with some constant* K *that depends only on* σ_x *and* γ.

Note that for the sake of simplicity, in 10, $\mathbb{E}\left[\epsilon_{kl}^2\right]$ and $\mathbb{E}\left[\epsilon_{kl}^{2(1+\gamma)}\right]$ are bounded with the same constant K. Lemma 10 is a direct implication of Assumption 4. It uses the moment condition on $\mathbf{x}_i$ to achieve the moment conditions on ϵ_{kl}, which will play an important role in the following propositions.

Proposition 11. *Suppose that Assumption 4 holds. Let* K *be the constant defined in Lemma (10). Assume* α *satisfies* $4K/\alpha^2 < 1/6$. *Then, with at least* $1 - d^2\exp(-n/12)$ *probability, for all* $\mathbf{\Sigma} \in \mathbf{\Sigma}^* + \mathbb{B}^{\infty}(\alpha/2)$,

$$\langle \nabla L_\alpha(\mathbf{\Sigma}) - \nabla L_\alpha(\mathbf{\Sigma}^*), \mathbf{\Sigma} - \mathbf{\Sigma}^*\rangle \geq \frac{1}{2}\left\|\mathbf{\Sigma} - \mathbf{\Sigma}^*\right\|_F^2$$

Proposition 11 implies that with proper assumptions on α, $\mathcal{E}_1(\alpha/2, 1/2)$ happens with high probability. For instance, by taking $n \gtrsim (\log d)^{1+\frac{1}{2\gamma}}$, we have $1 - d^2\exp(-n/12) \geq 1 - 2/d$.

Proposition 12. *Suppose that Assumption 4 holds. Let* K *be the constant defined in Lemma (10). Assume* $\alpha = \sqrt{Kn/\log d}$, *then*

$$\left\|\nabla L_\alpha(\mathbf{\Sigma}^*)\right\|_\infty \leq 8\sqrt{\frac{K\log d}{n}} \tag{11}$$

with at least $1 - 2/d$ *probability. Furthermore, if we have* $n \geq K^{-1}(\log d)^{1+\frac{1}{2\gamma}}$, *then for any* $\beta > 0$,

$$\Pr\left\{\left\|\nabla L_\alpha(\mathbf{\Sigma}^*)_{\mathcal{S}}\right\|_F \geq (\beta+1)\sqrt{\frac{Ks}{n}}\right\} \leq \frac{2}{\beta}. \tag{12}$$

In Proposition 12, 11 indicates that $\{\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty + \epsilon \leq 0.5p'(\phi_0)\lambda\}$ happens with high probability if we take $\epsilon \leq \sqrt{1/n}$ and $\lambda \asymp \sqrt{\log d/n}$. 12 implies that for any $p \in [0,1)$, there exists β such that $\|\nabla L_\alpha(\mathbf{\Sigma}^*)_{\mathcal{S}}\|_{\mathrm{F}} \leq (\beta+1)\sqrt{Ks/n}$ holds with at least p probability as long as $n \geq K^{-1}(\log d)^{1+\frac{1}{2\gamma}} \asymp (\log d)^{1+\frac{1}{2\gamma}}$. Weakly speaking, $\|\nabla L_\alpha(\mathbf{\Sigma}^*)_{\mathcal{S}}\|_{\mathrm{F}} = O_P(\sqrt{s/n})$.

Theorem 13. *Suppose that Assumptions 1, 3 and 4 hold. By Assumption 1, there exists some* $\phi_1 > \phi_0 > \sqrt{5}$ *such that*

$$p'(\phi_0) > 0 \quad \textit{and} \quad p'(\phi) = 0, \forall \phi \geq \phi_1$$

Suppose the sample size satisfies $n \gtrsim (\log d)^{1+\frac{1}{2\gamma}}$. *Take* $\lambda \asymp \sqrt{\log d/n}$, $l \asymp s^{1/2}$ *and let* $\alpha = \sqrt{Kn/\log d}$, $r = \alpha/2$, $\tau \leq \left\|(\mathbf{\Sigma}^*)^{-1}\right\|_F^{-1} \cdot \sqrt{Ks/n}$, $\epsilon \leq \sqrt{K/n}$. *Then, under the minimal signal strength condition* $\|\mathbf{\Sigma}^*_{\mathcal{S}}\|_{\min} \geq (\phi_0 + \phi_1)\lambda$, *the multi-stage estimator* $\widetilde{\mathbf{\Sigma}}^{(T)}$ *with* $T \gtrsim \frac{\log(\log d)}{\log(1/\delta)}$ *satisfies the bound*

$$\left\|\widetilde{\mathbf{\Sigma}}^{(T)} - \mathbf{\Sigma}^*\right\|_F \lesssim \beta\sqrt{\frac{s}{n}}, \quad \left\|\widetilde{\mathbf{\Sigma}}^{(T)} - \mathbf{\Sigma}^*\right\|_1 \lesssim \beta\frac{s}{\sqrt{n}}$$

with at least $1 - 2/d - d^2\exp(-n/12) - 2/\beta$ *probability and* K *to be the constant defined in Lemma (10). This immediately implies the following weaker conclusion*

$$\left\|\widetilde{\mathbf{\Sigma}}^{(T)} - \mathbf{\Sigma}^*\right\|_F = O_P\left(\sqrt{\frac{s}{n}}\right), \quad \left\|\widetilde{\mathbf{\Sigma}}^{(T)} - \mathbf{\Sigma}^*\right\|_1 = O_P\left(\frac{s}{\sqrt{n}}\right)$$

with at least $1 - 2/d$ *probability.*

Theorem 13 is a direct consequence of the combination of Proposition 8, 11 and 12, which implies that under weak assumptions, we just need to solve no more than approximately $\log\log d$ convex problems to achieve the oracle rate $\sqrt{s/n}$. It is easy to see that the proposed estimator achieves the oracle statistical rate of convergence under weaker assumptions on the distribution of sample data $\mathbf{x}_i$. In existing literature, Wei's paper proposes a formulation with squared loss that requires sub-Gaussian assumption to achieve the oracle rate. With Huber lost in our formulation, we only need to assume the existence of $4(1+\gamma)$-th moment of $\mathbf{x}_i$. In other words, our estimator is capable of handling heavy-tailed data.

V. Numerical Simulations

VI. Conclusions

XXX

Appendix A
Proofs of Statistical Theory

In this appendix, we first provide some necessary lemmata, and then provide the proofs of all the statistical theoretical results in Section IV.

Lemma 14. *Let* $\mathcal{E}$ *be a subset of* $[d]$ *that contains* $\mathcal{S}$. *For any* $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ *satisfying* $\mathbf{\Sigma}_{\overline{\mathcal{E}}} = \mathbf{0}$ *and* $\epsilon > 0$, *provided* $\mathbf{\Lambda} = (\Lambda)_{kl}$ *satisfies* $\|\mathbf{\Lambda}_{\overline{\mathcal{E}}}\|_{\min} > \|\nabla L_\alpha(\mathbf{\Sigma})_{\overline{\mathcal{E}}}\|_\infty + \epsilon$, *any* ϵ*-optimal solution* $\widetilde{\mathbf{\Sigma}}$ *to* (4) *satisfies*

$$\begin{aligned}
&\left\|(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma})_{\overline{\mathcal{E}}}\right\|_1 \\
\leq&(\|\mathbf{\Lambda}_{\overline{\mathcal{E}}}\|_{\min} - \|\nabla L_\alpha(\mathbf{\Sigma})_{\overline{\mathcal{E}}}\|_\infty - \epsilon)^{-1} \\
&\cdot\left\{(\|\mathbf{\Lambda}_{\mathcal{E}}\|_\infty + \|\nabla L_\alpha(\mathbf{\Sigma})_{\mathcal{E}}\|_\infty + \epsilon)\cdot\left\|(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma})_{\mathcal{E}}\right\|_1\right. \\
&\left.+\left\|\tau\mathbf{\Sigma}^{-1}\right\|_F\cdot\left\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_F\right\}
\end{aligned}$$

Proof: For any $\mathbf{\Xi} \in \partial\left\|\widetilde{\mathbf{\Sigma}}\right\|_{1,\text{off}}$, define $\boldsymbol{U}(\mathbf{\Xi}) = \nabla L_\alpha(\widetilde{\mathbf{\Sigma}}) - \tau\widetilde{\mathbf{\Sigma}}^{-1} + \mathbf{\Lambda}\circ\mathbf{\Xi} \in \mathbb{R}^{d\times d}$. By convexity of $L_\alpha(\mathbf{\Sigma})$ and $-\log\det\mathbf{\Sigma}$:

$$\langle\nabla L_\alpha(\widetilde{\mathbf{\Sigma}}) - \nabla L_\alpha(\mathbf{\Sigma}), \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\rangle \geq 0 \text{ and } \langle\mathbf{\Sigma}^{-1} - \widetilde{\mathbf{\Sigma}}^{-1}, \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\rangle \geq 0.$$

Therefore,

$$\begin{aligned}
&\langle \boldsymbol{U}(\boldsymbol{\Xi}), \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \rangle \geq \|\boldsymbol{U}(\boldsymbol{\Xi})\|_\infty \left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_1 \\
=&\langle \nabla L_\alpha(\widetilde{\boldsymbol{\Sigma}}) - \nabla L_\alpha(\boldsymbol{\Sigma}), \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \rangle + \langle \nabla L_\alpha(\boldsymbol{\Sigma}), \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \rangle \\
&- \langle \tau \boldsymbol{\Sigma}^{-1}, \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \rangle + \langle \tau \boldsymbol{\Sigma}^{-1} - \tau \widetilde{\boldsymbol{\Sigma}}^{-1}, \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \rangle \\
&+ \langle \boldsymbol{\Lambda} \circ \boldsymbol{\Xi}, \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \rangle \\
\geq& 0 + \|\nabla L_\alpha(\boldsymbol{\Sigma})_{\mathcal{E}}\|_\infty \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{E}}\right\|_1 \\
&+ \|\nabla L_\alpha(\boldsymbol{\Sigma})_{\overline{\mathcal{E}}}\|_\infty \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\overline{\mathcal{E}}}\right\|_1 \\
&+ \left\|\tau \boldsymbol{\Sigma}^{-1}\right\|_{\mathrm{F}} \cdot \left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_{\mathrm{F}} + 0 + \langle \boldsymbol{\Lambda} \circ \boldsymbol{\Xi}, \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \rangle
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
&\langle \boldsymbol{\Lambda} \circ \boldsymbol{\Xi}, \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \rangle \\
=&\langle (\boldsymbol{\Lambda} \circ \boldsymbol{\Xi})_{\overline{\mathcal{E}}}, (\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\overline{\mathcal{E}}} \rangle + \langle (\boldsymbol{\Lambda} \circ \boldsymbol{\Xi})_{\mathcal{E}}, (\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{E}} \rangle \\
\geq& \|\boldsymbol{\Lambda}_{\overline{\mathcal{E}}}\|_{\min} \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\overline{\mathcal{E}}}\right\|_1 - \|\boldsymbol{\Lambda}_{\mathcal{E}}\|_\infty \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{E}}\right\|_1
\end{aligned}$$

Together, the last two displays imply

$$\begin{aligned}
&\|\boldsymbol{U}(\boldsymbol{\Xi})\|_\infty \left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_1 \\
\geq& \|\nabla L_\alpha(\boldsymbol{\Sigma})_{\mathcal{E}}\|_\infty \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{E}}\right\|_1 + \|\nabla L_\alpha(\boldsymbol{\Sigma})_{\overline{\mathcal{E}}}\|_\infty \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\overline{\mathcal{E}}}\right\|_1 \\
&+ \left\|\tau \boldsymbol{\Sigma}^{-1}\right\|_{\mathrm{F}} \cdot \left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_{\mathrm{F}} + \|\boldsymbol{\Lambda}_{\overline{\mathcal{E}}}\|_{\min} \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\overline{\mathcal{E}}}\right\|_1 \\
&- \|\boldsymbol{\Lambda}_{\mathcal{E}}\|_\infty \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{E}}\right\|_1
\end{aligned}$$

Since the right-hand side of this inequality does not depend on $\boldsymbol{\Xi}$, taking the infimum with respect to $\boldsymbol{\Xi} \in \partial \left\|\widetilde{\boldsymbol{\Sigma}}\right\|_{1,\mathrm{off}}$ on both sides to reach

$$\begin{aligned}
&\epsilon \left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_1 \\
\geq& \|\nabla L_\alpha(\boldsymbol{\Sigma})_{\mathcal{E}}\|_\infty \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{E}}\right\|_1 + \|\nabla L_\alpha(\boldsymbol{\Sigma})_{\overline{\mathcal{E}}}\|_\infty \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\overline{\mathcal{E}}}\right\|_1 \\
&+ \left\|\tau \boldsymbol{\Sigma}^{-1}\right\|_{\mathrm{F}} \cdot \left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_{\mathrm{F}} + \|\boldsymbol{\Lambda}_{\overline{\mathcal{E}}}\|_{\min} \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\overline{\mathcal{E}}}\right\|_1 \\
&- \|\boldsymbol{\Lambda}_{\mathcal{E}}\|_\infty \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{E}}\right\|_1
\end{aligned}$$

Decompose $\left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_1$ as $\left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{E}}\right\|_1 + \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\overline{\mathcal{E}}}\right\|_1$, the stated result follows immediately. ■

Lemma 15. *Consider some $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ satisfying $\boldsymbol{\Sigma}_{\mathcal{S}^c} = 0$, and let $\mathcal{E} \subseteq [d]$ be a subset that contains $\mathcal{S}$ that has cardinality $|\mathcal{E}| = k$. Assume that $\boldsymbol{\Lambda} = (\Lambda)_{kl}$ satisfies $\|\boldsymbol{\Lambda}\|_\infty \leq \lambda$ and $\|\boldsymbol{\Lambda}_{\overline{\mathcal{E}}}\|_{\min} \geq \rho\lambda > 0$ for some $\rho \in (0,1]$, $\tau \leq \frac{\lambda s^{1/2}}{\left\|\boldsymbol{\Sigma}^{-1}\right\|_F}$. Conditioned on event $\{\|\nabla L_\alpha(\boldsymbol{\Sigma})\|_\infty + \epsilon \leq 0.5\rho\lambda\}$, any ϵ-optimal solution $\widetilde{\boldsymbol{\Sigma}}$ to (4) satisfies $\widetilde{\boldsymbol{\Sigma}} \in \boldsymbol{\Sigma} + \mathbb{C}(l)$, where $l = \left(2 + \frac{2}{\rho}\right) k^{1/2} + \frac{2}{\rho} s^{1/2}$. Moreover, let $r, \kappa > 0$ satisfy*

$$r > \kappa^{-1} \left\{2\lambda s^{1/2} + 0.5\rho\lambda k^{1/2}\right\}.$$

Then, conditioned on the event $\mathcal{E}_1(r,\kappa) \cap \{\|\nabla L_\alpha(\boldsymbol{\Sigma})\|_\infty + \epsilon \leq 0.5\rho\lambda\}$,

$$\begin{aligned}
\left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_F &\leq \kappa^{-1} \left\{\|\boldsymbol{\Lambda}_{\mathcal{S}}\|_F + \|\nabla L_\alpha(\boldsymbol{\Sigma})_{\mathcal{E}}\|_F + k^{1/2}\epsilon + \left\|\tau\boldsymbol{\Sigma}^{-1}\right\|_F\right\} \\
&\leq \kappa^{-1} \left\{2\lambda s^{1/2} + 0.5\rho\lambda k^{1/2}\right\} < r
\end{aligned}$$

Proof: Conditioned on the stated event, Lemma 14 indicates

$$\left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\overline{\mathcal{E}}}\right\|_1 \leq \left(1 + \frac{2}{\rho}\right) \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{E}}\right\|_1 + \frac{2}{\rho}\sqrt{s} \left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_{\mathrm{F}}.$$

Therefore,

$$\left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_1 \leq \left(1 + \frac{2}{\rho}\right) \sqrt{k} \left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_{\mathrm{F}} + \frac{2}{\rho}\sqrt{s} \left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_{\mathrm{F}},$$

which implies that $\widetilde{\boldsymbol{\Sigma}} \in \boldsymbol{\Sigma} + \mathbb{C}(l)$.

Now we prove the second statement. Define $\eta = \sup\{u \in [0,1] : (1-u)\boldsymbol{\Sigma} + u\widetilde{\boldsymbol{\Sigma}} \in \mathbb{B}(r)\}$, where $\mathbb{B}(r) = \{\boldsymbol{\Delta} \in \mathbb{R}^{d\times d} : \|\boldsymbol{\Delta}\|_{\mathrm{F}} \leq r\}$. Note that $\eta = 1$ if $\widetilde{\boldsymbol{\Sigma}} \in \boldsymbol{\Sigma} + \mathbb{B}(r)$ and $\eta \in (0,1)$ otherwise. Let $\widetilde{\boldsymbol{\Sigma}}_\eta \coloneqq (1-\eta)\boldsymbol{\Sigma} + \eta\widetilde{\boldsymbol{\Sigma}}$. Notice that if $\left\|\widetilde{\boldsymbol{\Sigma}}_\eta - \boldsymbol{\Sigma}\right\|_{\mathrm{F}} < r$, then $\widetilde{\boldsymbol{\Sigma}}_\eta = \widetilde{\boldsymbol{\Sigma}}$. By the convexity of Huber loss, we have

$$\begin{aligned}
&\langle \nabla L_\alpha(\widetilde{\boldsymbol{\Sigma}}_\eta) - \nabla L_\alpha(\boldsymbol{\Sigma}), \widetilde{\boldsymbol{\Sigma}}_\eta - \boldsymbol{\Sigma} \rangle \\
&\qquad \leq \eta \langle \nabla L_\alpha(\widetilde{\boldsymbol{\Sigma}}) - \nabla L_\alpha(\boldsymbol{\Sigma}), \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \rangle \qquad (13)
\end{aligned}$$

Recall that $\mathbb{B}^\infty(r) \coloneqq \left\{\boldsymbol{\Delta} \in \mathbb{R}^{d\times d} : \|\boldsymbol{\Delta}\|_\infty \leq r\right\}$. Since $\widetilde{\boldsymbol{\Sigma}}_\eta - \boldsymbol{\Sigma} \in \mathbb{B}(r) \subseteq \mathbb{B}^\infty(r)$, conditioned on event $\mathcal{E}_1$, we have

$$\langle \nabla L_\alpha(\widetilde{\boldsymbol{\Sigma}}_\eta) - \nabla L_\alpha(\boldsymbol{\Sigma}), \widetilde{\boldsymbol{\Sigma}}_\eta - \boldsymbol{\Sigma} \rangle \geq \kappa \left\|\widetilde{\boldsymbol{\Sigma}}_\eta - \boldsymbol{\Sigma}\right\|_{\mathrm{F}}^2 \qquad (14)$$

Now we upper bound the right-hand side of (13). For any $\boldsymbol{\Xi} \in \partial \left\|\widetilde{\boldsymbol{\Sigma}}\right\|_{1,\mathrm{off}}$, write

$$\begin{aligned}
&\langle \nabla L_\alpha(\widetilde{\boldsymbol{\Sigma}}) - \nabla L_\alpha(\boldsymbol{\Sigma}), \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \rangle \\
=& \underbrace{\langle \boldsymbol{U}(\boldsymbol{\Xi}), \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \rangle}_{\coloneqq \Pi_1} - \underbrace{\langle \nabla L_\alpha(\boldsymbol{\Sigma}), \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \rangle}_{\coloneqq \Pi_2} \\
&- \underbrace{\langle \boldsymbol{\Lambda} \circ \boldsymbol{\Xi}, \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \rangle}_{\coloneqq \Pi_3} + \underbrace{\langle \tau\boldsymbol{\Sigma}^{-1}, \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \rangle}_{\coloneqq \Pi_4} - \underbrace{\tau \langle \boldsymbol{\Sigma}^{-1} - \widetilde{\boldsymbol{\Sigma}}^{-1}, \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \rangle}_{\geq 0} \qquad (15)
\end{aligned}$$

where $\boldsymbol{U}(\boldsymbol{\Xi}) \coloneqq \nabla L_\alpha(\widetilde{\boldsymbol{\Sigma}}) - \tau\widetilde{\boldsymbol{\Sigma}}^{-1} + \boldsymbol{\Lambda} \circ \boldsymbol{\Xi} \in \mathbb{R}^{d\times d}$. We have

$$\begin{aligned}
|\Pi_1| &\leq \|\boldsymbol{U}(\boldsymbol{\Xi})\|_\infty \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\overline{\mathcal{E}}}\right\|_1 + \|(\boldsymbol{U}(\boldsymbol{\Xi}))_{\mathcal{E}}\|_{\mathrm{F}} \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{E}}\right\|_{\mathrm{F}} \\
|\Pi_2| &\leq \|\nabla L_\alpha(\boldsymbol{\Sigma})_{\mathcal{E}}\|_{\mathrm{F}} \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{E}}\right\|_{\mathrm{F}} + \|\nabla L_\alpha(\boldsymbol{\Sigma})_{\overline{\mathcal{E}}}\|_\infty \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\overline{\mathcal{E}}}\right\|_1 \\
|\Pi_4| &\leq \left\|\tau\boldsymbol{\Sigma}^{-1}\right\|_{\mathrm{F}} \cdot \left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_{\mathrm{F}}
\end{aligned}$$

Turning to Π_3, decompose $\boldsymbol{\Lambda} \circ \boldsymbol{\Xi}$ and $\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}$ according to $\mathcal{S} \cup (\mathcal{E}/\mathcal{S}) \cup \mathcal{E}$ to reach

$$\begin{aligned}
\Pi_3 =& \langle (\boldsymbol{\Lambda} \circ \boldsymbol{\Xi})_{\mathcal{S}}, (\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{S}} \rangle \\
&+ \langle (\boldsymbol{\Lambda} \circ \boldsymbol{\Xi})_{\mathcal{E}\setminus\mathcal{S}}, (\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{E}\setminus\mathcal{S}} \rangle + \langle (\boldsymbol{\Lambda} \circ \boldsymbol{\Xi})_{\overline{\mathcal{E}}}, (\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\overline{\mathcal{E}}} \rangle
\end{aligned}$$

Since $\boldsymbol{\Sigma}_{\overline{\mathcal{E}}} = \mathbf{0}$ and $\boldsymbol{\Xi} \in \partial \left\|\widetilde{\boldsymbol{\Sigma}}\right\|_{1,\mathrm{off}}$, we have $\langle (\boldsymbol{\Lambda} \circ \boldsymbol{\Xi})_{\overline{\mathcal{E}}}, (\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\overline{\mathcal{E}}} \rangle = \langle \boldsymbol{\Lambda}_{\overline{\mathcal{E}}}, \left|\widetilde{\boldsymbol{\Sigma}}_{\overline{\mathcal{E}}}\right| \rangle = \langle \boldsymbol{\Lambda}_{\overline{\mathcal{E}}}, \left|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\overline{\mathcal{E}}}\right| \rangle$. Also, $\langle (\boldsymbol{\Lambda} \circ \boldsymbol{\Xi})_{\mathcal{E}\setminus\mathcal{S}}, (\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{E}\setminus\mathcal{S}} \rangle = \langle (\boldsymbol{\Lambda} \circ \boldsymbol{\Xi})_{\mathcal{E}\setminus\mathcal{S}}, \widetilde{\boldsymbol{\Sigma}}_{\mathcal{E}\setminus\mathcal{S}} \rangle \geq 0$. Therefore,

$$\Pi_3 \geq \|\boldsymbol{\Lambda}_{\overline{\mathcal{E}}}\|_{\min} \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\overline{\mathcal{E}}}\right\|_1 - \|\boldsymbol{\Lambda}_{\mathcal{S}}\|_{\mathrm{F}} \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{S}}\right\|_{\mathrm{F}}$$

Combining (15) with our estimation for Π_1, Π_2, Π_3 and Π_4, we have

$$\begin{aligned}
&\langle \nabla L_\alpha(\widetilde{\mathbf{\Sigma}}) - \nabla L_\alpha(\mathbf{\Sigma}), \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma} \rangle \\
\leq & - \left\{ \|\mathbf{\Lambda}_{\overline{\mathcal{E}}}\|_{\min} - \|\nabla L_\alpha(\mathbf{\Sigma})\|_\infty - \|\boldsymbol{U}(\mathbf{\Xi})\|_\infty \right\} \left\| (\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma})_{\overline{\mathcal{E}}} \right\|_1 \\
& + \|\nabla L_\alpha(\mathbf{\Sigma})_{\mathcal{E}}\|_{\mathrm{F}} \left\| (\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma})_{\mathcal{E}} \right\|_{\mathrm{F}} + \|(\boldsymbol{U}(\mathbf{\Xi}))_{\mathcal{E}}\|_{\mathrm{F}} \left\| (\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma})_{\mathcal{E}} \right\|_{\mathrm{F}} \\
& + \|\mathbf{\Lambda}_{\mathcal{S}}\|_{\mathrm{F}} \left\| (\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma})_{\mathcal{E}} \right\|_{\mathrm{F}} + \left\| \tau \mathbf{\Sigma}^{-1} \right\|_{\mathrm{F}} \left\| \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma} \right\|_{\mathrm{F}}
\end{aligned}$$

Taking the infimum with respect to $\mathbf{\Xi} \in \partial \left\| \widetilde{\mathbf{\Sigma}} \right\|_{1,\text{off}}$ on both sides, it follows that

$$\begin{aligned}
&\langle \nabla L_\alpha(\widetilde{\mathbf{\Sigma}}) - \nabla L_\alpha(\mathbf{\Sigma}), \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma} \rangle \\
\leq & - \left\{ \|\lambda_{\overline{\mathcal{E}}}\|_{\min} - \|\nabla L_\alpha(\mathbf{\Sigma})\|_\infty - \epsilon \right\} \left\| (\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma})_{\overline{\mathcal{E}}} \right\|_1 \\
& + \left\{ \|\nabla L_\alpha(\mathbf{\Sigma})\|_{\mathrm{F}} + k^{1/2}\epsilon \right\} \left\| (\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma})_{\mathcal{E}} \right\|_{\mathrm{F}} \\
& + \|\mathbf{\Lambda}_{\mathcal{S}}\|_{\mathrm{F}} \left\| (\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma})_{\mathcal{E}} \right\|_{\mathrm{F}} + \left\| \tau \mathbf{\Sigma}^{-1} \right\|_{\mathrm{F}} \left\| \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma} \right\|_{\mathrm{F}}
\end{aligned} \tag{16}$$

With $\widetilde{\mathbf{\Sigma}}_\eta - \mathbf{\Sigma} = \eta(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma})$, it follows from (13), (14) and (16) that conditioned on $\mathcal{E}_1(r,\kappa) \cap \{\|\nabla L_\alpha(\mathbf{\Sigma})\|_\infty + \epsilon \leq 0.5\rho\lambda\}$,

$$\begin{aligned}
&\kappa \left\| \widetilde{\mathbf{\Sigma}}_\eta - \mathbf{\Sigma} \right\|_{\mathrm{F}}^2 \leq \\
&\left\{ \|\mathbf{\Lambda}_{\mathcal{S}}\|_{\mathrm{F}} + \|\nabla L_\alpha(\mathbf{\Sigma})_{\mathcal{E}}\|_{\mathrm{F}} + k^{1/2}\epsilon + \left\| \tau \mathbf{\Sigma}^{-1} \right\|_{\mathrm{F}} \right\} \left\| \widetilde{\mathbf{\Sigma}}_\eta - \mathbf{\Sigma} \right\|_{\mathrm{F}}
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left\| \widetilde{\mathbf{\Sigma}}_\eta - \mathbf{\Sigma} \right\|_{\mathrm{F}} \\
&\leq \kappa^{-1} \left\{ \|\mathbf{\Lambda}_{\mathcal{S}}\|_{\mathrm{F}} + \|\nabla L_\alpha(\mathbf{\Sigma})_{\mathcal{E}}\|_{\mathrm{F}} + k^{1/2}\epsilon + \left\| \tau \mathbf{\Sigma}^{-1} \right\|_{\mathrm{F}} \right\} \\
&\leq \kappa^{-1} \{ \lambda s^{1/2} + 0.5\rho\lambda k^{1/2} + \lambda s^{1/2} \} < r
\end{aligned} \tag{17}$$

Since $\widetilde{\mathbf{\Sigma}}_\eta - \mathbf{\Sigma}$ falls in the interior of $\mathbb{B}(r)$, we must have $\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma} = \widetilde{\mathbf{\Sigma}}_\eta - \mathbf{\Sigma} \in \mathbb{B}(r)$. Consequently, (17) also holds for $\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}$. ■

A. Proof of Proposition 6

Proof: With the initial estimate $\widetilde{\mathbf{\Sigma}}^{(0)} = \mathbf{I}$, we have $\Lambda_{kl}^{(0)} = p'_\lambda(0) = \lambda$ for all $k, l \in [d]$. Then the result follows immediately from Lemma 15 with $\mathbf{\Sigma} = \mathbf{\Sigma}^*$, $\mathcal{E} = \mathcal{S}$ and $\rho = 1$. ■

B. Proof of Proposition 8

Proof: With the initial estimate $\widetilde{\mathbf{\Sigma}}^{(0)} = \mathbf{0}$, we have $\Lambda_{kl}^{(0)} = p'_\lambda(0) = \lambda$ for all $k, l \in [d]$. Then, applying Lemma 15 with $\mathbf{\Sigma} = \mathbf{\Sigma}^*$, $\mathcal{E} = \mathcal{S}$ and $\rho = p'(\phi_0)$ we obtain that, conditioned on the event $\mathcal{E}_1(r, 1/2) \cap \{\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty + \epsilon \leq 0.5p'(\phi_0)\lambda\}$,

$$\begin{aligned}
&\left\| \widetilde{\mathbf{\Sigma}}^{(1)} - \mathbf{\Sigma}^* \right\|_{\mathrm{F}} \\
&\leq 2 \left\{ \left\| \mathbf{\Lambda}_{\mathcal{S}}^{(0)} \right\|_{\mathrm{F}} + \|\nabla L_\alpha(\mathbf{\Sigma}^*)_{\mathcal{S}}\|_{\mathrm{F}} + s^{1/2}\epsilon + \tau \left\| (\mathbf{\Sigma}^*)^{-1} \right\|_{\mathrm{F}} \right\} \\
&\leq 2 \left\{ 2 + 0.5p'(\phi_0) \right\} s^{1/2}\lambda
\end{aligned} \tag{18}$$

For $t \geq 1$, define the augmented set

$$\mathcal{E}_t = \mathcal{S} \cup \left\{ (k,l) : \Lambda_{kl}^{(t-1)} < p'(\phi_0)\lambda \right\}$$

which depends on the solution $\widetilde{\mathbf{\Sigma}}^{(t-1)}$ from the previous step. We claim that the above constructed sets satisfy

$$|\mathcal{E}_t| < (c^2+1)s \tag{19}$$

If (19) is true, it follows from Lemma 15 with $\mathbf{\Sigma} = \mathbf{\Sigma}^*$, $\mathcal{E} = \mathcal{E}_t$, $k = (c^2+1)$ and $\rho = p'(\phi_0)$ that conditioned on the event $\mathcal{E}_1(r, 1/2) \cap \{\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty + \epsilon \leq 0.5p'(\phi_0)\lambda\}$, we have $\widetilde{\mathbf{\Sigma}}^{(t)} \in \mathbf{\Sigma}^* + \mathbb{C}(l)$ and

$$\begin{aligned}
&\left\| \widetilde{\mathbf{\Sigma}}^{(t)} - \mathbf{\Sigma}^* \right\|_{\mathrm{F}} \\
&\leq 2 \left\{ \left\| \mathbf{\Lambda}_{\mathcal{S}}^{(t-1)} \right\|_{\mathrm{F}} + \|\nabla L_\alpha(\mathbf{\Sigma}^*)_{\mathcal{E}_t}\|_{\mathrm{F}} + |\mathcal{E}_t|^{1/2} \epsilon + \tau \left\| (\mathbf{\Sigma}^*)^{-1} \right\|_{\mathrm{F}} \right\} \\
&\leq 2 \left\{ 2\lambda s^{1/2} + 0.5p'(\phi_0)\lambda (c^2+1)^{1/2} s^{1/2} \right\} = c\phi_0 \lambda s^{1/2} \leq r
\end{aligned} \tag{20}$$

where the last two steps follow from (8) and (9).We will prove our claim (19) by induction. For $t = 1$, we have $\Lambda_{kl}^{(0)} = p'_\lambda(0) = \lambda$ for all $k, l \in [d]$, so $\mathcal{E}_1 = \mathcal{S}$ and (19) holds.

Next, assume (19) holds for some $t \geq 1$, from which (20) follows. To bound $|\mathcal{E}_{t+1}|$, note that for any $(k,l) \in \mathcal{E}_{t+1} \backslash \mathcal{S}$, $p'_\lambda(\left| \widetilde{\Sigma}_{kl}^{(t)} \right|) = \Lambda_{kl}^{(t)} < p'(\phi_0)\lambda = p'_\lambda(\phi_0\lambda)$. Together with the monotonicity of p'_λ, this implies that $\left| \widetilde{\Sigma}_{kl}^{(t)} \right| > \phi_0\lambda$. Recalling that $\Sigma_{kl}^* = 0$ for $(k,l) \in \mathcal{E}_{t+1} \backslash \mathcal{S}$, we obtain

$$\begin{aligned}
|\mathcal{E}_{t+1} \backslash \mathcal{S}|^{1/2} &< \left\{ \sum_{(k,l) \in \mathcal{E}_{t+1} \backslash \mathcal{S}} \left(\left| \widetilde{\Sigma}_{kl}^{(t)} \right| / \phi_0 \lambda \right)^2 \right\}^{1/2} \\
&= \left\| (\widetilde{\mathbf{\Sigma}}^{(t+1)} - \mathbf{\Sigma}^*)_{\mathcal{E}_{t+1} \backslash \mathcal{S}} \right\|_{\mathrm{F}} / (\phi_0 \lambda) \leq c s^{1/2}
\end{aligned} \tag{21}$$

The last inequality follows from (20). Therefore $|\mathcal{E}_{t+1}| \leq |\mathcal{S}| + |\mathcal{E}_{t+1} \backslash \mathcal{S}| < (c^2+1)^{1/2} s$, which completes the induction step.

Now we start to prove (10). Recall that for each (k,l), $\Lambda_{kl}^{(t-1)} = p'_\lambda(\left| \widetilde{\Sigma}_{kl}^{(t-1)} \right|)$. If $\left| \widetilde{\Sigma}_{kl}^{(t-1)} - \Sigma_{kl}^* \right| \geq \phi_0\lambda$, then $\Lambda_{kl}^{(t-1)} \leq \lambda \leq \left| \widetilde{\Sigma}_{kl}^{(t-1)} - \Sigma_{kl}^* \right| / \phi_0$; otherwise if $\left| \widetilde{\Sigma}_{kl}^{(t)} - \Sigma_{kl}^* \right| < \phi_0\lambda$, then $\Lambda_{kl}^{(t-1)} = p'_\lambda(\left| \widetilde{\Sigma}_{kl}^{(t-1)} \right|) \leq p'_\lambda(|\Sigma_{kl}^*| - \phi_0\lambda)$ due to the monotonicity of p'_λ. Putting together the pieces, we conclude that

$$\left\| \mathbf{\Lambda}_{\mathcal{S}}^{(t-1)} \right\|_{\mathrm{F}} \leq \|p'_\lambda(|\mathbf{\Sigma}_{\mathcal{S}}^*| - \phi_0\lambda)\|_{\mathrm{F}} + \phi_0^{-1} \left\| (\widetilde{\mathbf{\Sigma}}^{(t-1)} - \mathbf{\Sigma}^*)_{\mathcal{S}} \right\|_{\mathrm{F}} \tag{22}$$

For the remaining terms that involve $\mathcal{E}_t$ in (20), by the triangle inequality

$$\begin{aligned}
&\|\nabla L_\alpha(\mathbf{\Sigma}^*)_{\mathcal{E}_t}\|_{\mathrm{F}} + |\mathcal{E}_t|^{1/2} \epsilon \\
&\leq \|\nabla L_\alpha(\mathbf{\Sigma}^*)_{\mathcal{S}}\|_{\mathrm{F}} + s^{1/2}\epsilon + \left\| \nabla L_\alpha(\mathbf{\Sigma}^*)_{\mathcal{E}_t \backslash \mathcal{S}} \right\|_{\mathrm{F}} + |\mathcal{E}_t \backslash \mathcal{S}|^{1/2} \epsilon \\
&\leq \|\nabla L_\alpha(\mathbf{\Sigma}^*)_{\mathcal{S}}\|_{\mathrm{F}} + s^{1/2}\epsilon \\
&\qquad + (\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty + \epsilon) \cdot \left\| (\widetilde{\mathbf{\Sigma}}^{(t-1)} - \mathbf{\Sigma}^*)_{\mathcal{E}_t \backslash \mathcal{S}} \right\|_{\mathrm{F}} / (\phi_0 \lambda) \\
&\leq \|\nabla L_\alpha(\mathbf{\Sigma}^*)_{\mathcal{S}}\|_{\mathrm{F}} + s^{1/2}\epsilon + \frac{p'(\phi_0)}{2\phi_0} \left\| (\widetilde{\mathbf{\Sigma}}^{(t-1)} - \mathbf{\Sigma}^*)_{\mathcal{E}_t \backslash \mathcal{S}} \right\|_{\mathrm{F}}
\end{aligned} \tag{23}$$

where the second equality follows from (21) and the last inequality follows from $\{\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty + \epsilon \leq 0.5p'(\phi_0)\lambda\}$.Plugging (22) and (23) into (20) yields

$$\begin{aligned}
&\left\|\widetilde{\mathbf{\Sigma}}^{(t)} - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} \\
\leq& 2\Bigg\{\left\|p'_\lambda(|\mathbf{\Sigma}^*_{\mathcal{S}}| - \phi_0\lambda)\right\|_{\mathrm{F}} + \phi_0^{-1}\left\|(\widetilde{\mathbf{\Sigma}}^{(t-1)} - \mathbf{\Sigma}^*)_{\mathcal{S}}\right\|_{\mathrm{F}} \\
&+ \left\|\nabla L_\alpha(\mathbf{\Sigma}^*)_{\mathcal{S}}\right\|_{\mathrm{F}} + s^{1/2}\epsilon \\
&+ \frac{p'(\phi_0)}{2\phi_0}\left\|(\widetilde{\mathbf{\Sigma}}^{(t-1)} - \mathbf{\Sigma}^*)_{\mathcal{E}_t\setminus\mathcal{S}}\right\|_{\mathrm{F}} + \tau\left\|(\mathbf{\Sigma}^*)^{-1}\right\|_{\mathrm{F}}\Bigg\} \\
\leq& 2\Bigg\{\left\|p'_\lambda(|\mathbf{\Sigma}^*_{\mathcal{S}}| - \phi_0\lambda)\right\|_{\mathrm{F}} + \left\|\nabla L_\alpha(\mathbf{\Sigma}^*)_{\mathcal{S}}\right\|_{\mathrm{F}} + s^{1/2}\epsilon \\
&+ \frac{\sqrt{5}}{2\phi_0}\left\|(\widetilde{\mathbf{\Sigma}}^{(t-1)} - \mathbf{\Sigma}^*)_{\mathcal{E}_t}\right\|_{\mathrm{F}} + \tau\left\|(\mathbf{\Sigma}^*)^{-1}\right\|_{\mathrm{F}}\Bigg\} \\
=& 2\Bigg\{\left\|p'_\lambda(|\mathbf{\Sigma}^*_{\mathcal{S}}| - \phi_0\lambda)\right\|_{\mathrm{F}} + \left\|\nabla L_\alpha(\mathbf{\Sigma}^*)_{\mathcal{S}}\right\|_{\mathrm{F}} + s^{1/2}\epsilon \\
&+ \tau\left\|(\mathbf{\Sigma}^*)^{-1}\right\|_{\mathrm{F}}\Bigg\} + \frac{\sqrt{5}}{\phi_0}\left\|(\widetilde{\mathbf{\Sigma}}^{(t-1)} - \mathbf{\Sigma}^*)_{\mathcal{E}_t}\right\|_{\mathrm{F}}
\end{aligned}$$

The second inequality follows because $p'(\phi_0) \leq 1$ and for any $a, b \geq 0$, $\sqrt{a} + \sqrt{b/4} \leq \sqrt{5(a+b)/4}$. The proof of (10) is then completed. ∎

C. Proof of Lemma 10

Proof: Let $\sigma_y \coloneqq 4\sigma_x$. Given that $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$ are identically distributed and with the c_r-inequality,

$$\begin{aligned}
\mathbb{E}\left[|y_{mj}|^{4(1+\gamma)}\right] &= \mathbb{E}\left[|x_{1j} - x_{2j}|^{4(1+\gamma)}\right] \\
&\leq 2^{4(1+\gamma)-1}\left\{\mathbb{E}\left[|x_{1j}|^{4(1+\gamma)}\right] + \mathbb{E}\left[|x_{2j}|^{4(1+\gamma)}\right]\right\} \\
&\leq 2^{4(1+\gamma)}\sigma_x^{2(1+\gamma)} = \sigma_y^{2(1+\gamma)}, \qquad \forall j \in [d].
\end{aligned}$$

The second inequality follows from Assumption (4). Using the c_r-inequality again, $\mathbb{E}\left[\epsilon_{kl}^{2(1+\gamma)}\right]$ is bounded by a constant K as follows:

$$\begin{aligned}
&\mathbb{E}\left[\epsilon_{kl}^{2(1+\gamma)}\right] \\
&\leq 2^{2(1+\gamma)-1}\left(|\Sigma^*_{kl}|^{2(1+\gamma)} + \mathbb{E}\left[|y_{mk}y_{ml}/2|^{2(1+\gamma)}\right]\right) \\
&= 2^{2(1+\gamma)-1}\left(|\mathbb{E}\left[y_{mk}y_{ml}/2\right]|^{2(1+\gamma)} + (\sigma_y/2)^{2(1+\gamma)}\right) \\
&\leq 2^{2(1+\gamma)-1}\left\{\left(\sigma_y^{2(1+\gamma)} + 1\right)^{1+\gamma}/2^{2(1+\gamma)} + (\sigma_y/2)^{2(1+\gamma)}\right\} \\
&= 2^{-1}\left\{\left(\sigma_y^{2(1+\gamma)} + 1\right)^{1+\gamma} + \sigma_y^{2(1+\gamma)}\right\} \coloneqq K.
\end{aligned}$$

The last inequality uses the fact that $\mathbb{E}\left[y_{mj}^4\right] \leq \mathbb{E}\left[|y_{mj}|^{4(1+\gamma)}\right] + 1$ for $j \in [d]$. By similar reasoning,

$$\mathbb{E}\left[\epsilon_{kl}^2\right] = \mathrm{Var}\left(y_{mk}y_{m\ell}/2\right) \leq \mathbb{E}\left[y_{mk}^2 y_{m\ell}^2\right]/4 \leq (\sigma_y^{2(1+\gamma)}+1)/4 < K.$$

∎

D. Proof of Proposition 11

Proof: For fixed $k, l \in [d]$, $\Sigma^*_{k\ell} - y_{mk}y_{m\ell}/2$ are identically distributed for $m \in [N]$. Let $D_{kl} = (1/N)\sum_{m=1}^N 1\left(|\Sigma^*_{k\ell} - y_{mk}y_{m\ell}/2| \leq \alpha/2\right)$. By Chebyshev's inequality,

$$\mathbb{E}[D_{kl}] = \Pr\left(|\Sigma^*_{k\ell} - y_{mk}y_{m\ell}/2| \leq \alpha/2\right) \geq 1 - 4K/\alpha^2 > 5/6.$$

The last inequality holds with $4K/\alpha^2 < 1/6$.

Let $h_{kl}(\mathbf{x}_i, \mathbf{x}_j) \coloneqq 1\left(|\Sigma^*_{k\ell} - (x_{ik} - x_{jk})(x_{il} - x_{jl})/2| \leq \alpha/2\right)$. Then it's easy to see that

$$D_{kl} = (1/N)\sum_{1\leq i<j\leq n} h_{kl}(\mathbf{x}_i, \mathbf{x}_j)$$

Let $q = \lfloor n/2 \rfloor$ and $\sum_{\mathcal{P}}$ denote the summation over all $n!$ perumtations $(i_1, \ldots, i_n)$ of $[n] \coloneqq \{1, \ldots, n\}$. Using the same technique used in the Proof of Theorem 3.1 in [28], it can be shown that

$$\begin{aligned}
\Pr\left(D_{kl} \leq 1/2\right) &\leq \Pr\left(e^{-nD_{kl}} \geq e^{-n/2}\right) \\
&\leq e^{n/2}\frac{1}{n!}\sum_{\mathcal{P}}\prod_{j=1}^{q}\mathbb{E}e^{-(n/q)\cdot h_{kl}(\mathbf{x}_{i_{2j-1}}, \mathbf{x}_{i_{2j}})}.
\end{aligned} \tag{24}$$

Considering the fact that $e^{-(n/q)t} \leq 1 - (1 - e^{-n/q})t$ for $0 \leq t \leq 1$,

$$\begin{aligned}
&\mathbb{E}e^{-(n/q)\cdot h_{kl}(\mathbf{x}_{i_{2j-1}}, \mathbf{x}_{i_{2j}})} \\
&\quad\leq 1 - (1 - e^{-n/q})\mathbb{E}\left[h_{kl}(\mathbf{x}_{i_{2j-1}}, \mathbf{x}_{i_{2j}})\right] \\
&\quad\leq 1 - (5/6)\cdot(1 - e^{-n/q}).
\end{aligned} \tag{25}$$

With (24) and (25),

$$\Pr\left(D_{kl} \leq 1/2\right) \leq e^{n/2}\left[1/6 + (5/6)\cdot e^{-n/q}\right]^q. \tag{26}$$

Given $q = \lfloor n/2 \rfloor$, it is easy to verify that $1/6 + (5/6)\cdot e^{-n/q} \leq e^{-(7/12)\cdot(n/q)}$ for $n \geq 11$. Combining this with (26) yields

$$\Pr\left(D_{kl} \leq 1/2\right) \leq e^{-n/12}.$$

With union bound we have

$$\Pr\left[\min_{k,l} D_{kl} < \frac{1}{2}\right] \leq d^2\exp(-n/12).$$

Let $\mathcal{G}_{kl} = \{m \in [N] : |\Sigma^*_{k\ell} - y_{mk}y_{m\ell}/2| \leq \alpha/2\}$. Under the event that $\min_{k,l} D_{kl} \geq 1/2$,

$$\begin{aligned}
&\frac{1}{N}\sum_{m=1}^{N}\{\rho'_\alpha(\Sigma_{k\ell} - y_{mk}y_{m\ell}/2) - \rho'_\alpha(\Sigma^*_{k\ell} - y_{mk}y_{m\ell}/2)\}\cdot(\Sigma_{kl} - \Sigma^*_{kl}) \\
&\geq \frac{1}{N}\sum_{m\in\mathcal{G}_{kl}}\{\rho'_\alpha(\Sigma_{k\ell} - y_{mk}y_{m\ell}/2) - \rho'_\alpha(\Sigma^*_{k\ell} - y_{mk}y_{m\ell}/2)\} \\
&\qquad\cdot(\Sigma_{kl} - \Sigma^*_{kl}) \\
&= \frac{|\mathcal{G}_{kl}|}{N}\cdot(\Sigma_{kl} - \Sigma^*_{kl})^2 \geq \frac{1}{2}(\Sigma_{kl} - \Sigma^*_{kl})^2,
\end{aligned}$$

where the last equality uses $\mathbf{\Sigma} \in \mathbf{\Sigma}^* + \mathbb{B}^\infty(\alpha/2)$ and the last inequality follows from $|\mathcal{G}_{kl}|/N = D_{kl}$. Therefore,

$$\begin{aligned}
&\langle \nabla L_\alpha(\mathbf{\Sigma}) - \nabla L_\alpha(\mathbf{\Sigma}^*), \mathbf{\Sigma} - \mathbf{\Sigma}^* \rangle \\
=&\sum_{k,l} \frac{1}{N} \sum_{m=1}^{N} \{\rho'_\alpha(\Sigma_{k\ell} - y_{mk}y_{m\ell}/2) - \\
&\qquad \rho'_\alpha(\Sigma^*_{k\ell} - y_{mk}y_{m\ell}/2)\} \cdot (\Sigma_{kl} - \Sigma^*_{kl}) \\
\geq& \frac{1}{2} \|\mathbf{\Sigma} - \mathbf{\Sigma}^*\|_{\mathrm{F}}^2
\end{aligned}$$

with at least $1 - d^2 \exp(-n/12)$ probability. ∎

We adopt the following notations before the next stage of proof. Recall that $L_\alpha(\mathbf{\Sigma}) = \sum_{k,\ell} \frac{1}{N} \sum_{m=1}^{N} \rho_\alpha(\Sigma_{k\ell} - y_{mk}y_{m\ell}/2)$. Define $\mathbf{B}(\mathbf{\Sigma}) := \mathbb{E}[\nabla L_\alpha(\mathbf{\Sigma})]$, and $\mathbf{W}^* := \nabla L_\alpha(\mathbf{\Sigma}^*) - \mathbb{E}[\nabla L_\alpha(\mathbf{\Sigma})]$.

Lemma 16. *Recall that K is the constant defined in Lemma 10. $|(\mathbf{B}(\mathbf{\Sigma}^*))_{kl}| < \frac{K}{\alpha^{1+2\gamma}}$. Let $\alpha = \sqrt{\frac{Kn}{\log d}}$. If we have $n \geq K^{-1}(\log d)^{1+\frac{1}{2\gamma}}$, then $\|\mathbf{B}(\mathbf{\Sigma}^*)\|_\infty \leq \sqrt{K/n}$ and $\|\mathbf{B}(\mathbf{\Sigma}^*)_{\mathcal{S}}\|_F \leq \sqrt{Ks/n}$.*

Proof: For fixed $k, l \in [d]$, $\Sigma^*_{k\ell} - y_{mk}y_{m\ell}/2$ are identically distributed for $m \in [N]$. Let $\epsilon_{kl} = \Sigma^*_{k\ell} - y_{mk}y_{m\ell}/2$, then

$$\begin{aligned}
|\mathbb{E}[\rho'_\alpha(\epsilon_{kl})]| &= |\mathbb{E}[\epsilon_{kl} I(|\epsilon_{kl}| \leq \alpha) + \alpha \mathrm{sgn}(\epsilon_{kl}) I(|\epsilon_{kl}| > \alpha)]| \\
&= |\mathbb{E}[\epsilon_{kl} + (\alpha \mathrm{sgn}(\epsilon_{kl}) - \epsilon_{kl}) I(|\epsilon_{kl}| > \alpha)]| \\
&= |\mathbb{E}\{[\epsilon_{kl} - \alpha \mathrm{sgn}(\epsilon_{kl})] I(|\epsilon_{kl}| > \alpha)\}| \\
&\leq |\mathbb{E}[(|\epsilon_{kl}| - \alpha \mathrm{sgn}(\epsilon_{kl})) I(|\epsilon_{kl}| > \alpha)]| \\
&\leq \frac{\left|\mathbb{E}[(\epsilon_{kl}^{2(1+\gamma)} - \alpha^{2(1+\gamma)}) I(|\epsilon_{kl}| > \alpha)]\right|}{\alpha^{1+2\gamma}} \\
&< \frac{K}{\alpha^{1+2\gamma}}.
\end{aligned}$$

Therefore, for all k, l,

$$|(\mathbf{B}(\mathbf{\Sigma}^*))_{kl}| = \frac{1}{N} \left| \sum_{m=1}^{N} \mathbb{E}\left[\rho'_\alpha(\Sigma^*_{k\ell} - y_{mk}y_{m\ell}/2)\right] \right| < \frac{K}{\alpha^{1+2\gamma}}.$$

If we have $\alpha = \sqrt{Kn/\log d}$ and $n \geq K^{-1}(\log d)^{1+\frac{1}{2\gamma}}$, then

$$\|\mathbf{B}(\mathbf{\Sigma^*})\|_\infty \leq \frac{K}{(Kn/\log d)^{1/2+\gamma}} \leq \sqrt{\frac{K}{n}},$$

and

$$\|\mathbf{B}(\mathbf{\Sigma^*})_{\mathcal{S}}\|_{\mathrm{F}} \leq \sqrt{s} \cdot \|\mathbf{B}(\mathbf{\Sigma^*})\|_\infty \leq \sqrt{\frac{Ks}{n}}.$$

∎

E. Proof of Proposition 12

Proof: For fixed k, l, denote $\Sigma^*_{kl} - (x_{ik} - x_{jk})(x_{il} - x_{jl})/2$ as $\epsilon_{i,j}$. Let $h_{kl}(\mathbf{x}_i, \mathbf{x}_j) := \rho'_\alpha(\epsilon_{i,j})$. Then it's easy to see that

$$(\nabla L_\alpha(\mathbf{\Sigma}^*))_{kl} = (1/N) \sum_{1 \leq i < j \leq n} h_{kl}(\mathbf{x}_i, \mathbf{x}_j).$$

Let $q = \lfloor n/2 \rfloor$ and $\sum_{\mathcal{P}}$ denote the summation over all $n!$ perumtations $(i_1, \ldots, i_n)$ of $[n] := \{1, \ldots, n\}$. Using the same technique used in the Proof of Theorem 3.1 in [28], it can be shown that

$$\begin{aligned}
\Pr\left((\nabla L_\alpha(\mathbf{\Sigma}^*))_{kl} \geq y\right) &\leq \Pr\left(e^{(q/\alpha)\cdot(\nabla L_\alpha(\mathbf{\Sigma}^*))_{kl}} \geq e^{(q/\alpha)\cdot y}\right) \\
&\leq e^{-(q/\alpha)\cdot y} \frac{1}{n!} \sum_{\mathcal{P}} \prod_{j=1}^{q} \mathbb{E} e^{(1/\alpha)\cdot h_{kl}(\mathbf{x}_{i_{2j-1}}, \mathbf{x}_{i_{2j}})}.
\end{aligned} \tag{27}$$

Let $\rho_1(\cdot)$ denote the Huber loss function defined in (1) with $\alpha = 1$. Note that $h_{kl}(\mathbf{x}_i, \mathbf{x}_j) = \rho'_\alpha(\epsilon_{i,j}) = \alpha \rho'_1(\epsilon_{i,j}/\alpha)$. In addition, it is easy to verify the inequality that

$$-\log(1 - x + x^2) \leq \rho'_1(x) \leq \log(1 + x + x^2)$$

Therefore,

$$\begin{aligned}
\mathbb{E} e^{(1/\alpha)\cdot h_{kl}(\mathbf{x}_{i_{2j-1}}, \mathbf{x}_{i_{2j}})} &\leq 1 + \mathbb{E}\left[\epsilon_{i_{2j-1}, i_{2j}}/\alpha\right] + \mathbb{E}\left[(\epsilon_{i_{2j-1}, i_{2j}}/\alpha)^2\right] \\
&\leq 1 + K/\alpha^2 \leq e^{K/\alpha^2},
\end{aligned} \tag{28}$$

where the second inequality holds because $\mathbb{E}\left[\epsilon_{i_{2j-1}, i_{2j}}\right] = 0$ and $\mathbb{E}\left[\epsilon^2_{i_{2j-1}, i_{2j}}\right] \leq K$ by Lemma 10. Combining (27) and (28) yields

$$\Pr\left((\nabla L_\alpha(\mathbf{\Sigma}^*))_{kl} \geq y\right) \leq e^{-(q/\alpha)\cdot y + qK/\alpha^2}.$$

Similarly, it can be shown that $\Pr\left((\nabla L_\alpha(\mathbf{\Sigma}^*))_{kl} \leq -y\right) \leq e^{-(q/\alpha)\cdot y + qK/\alpha^2}$. Since $\alpha = \sqrt{Kn/\log d}$, by taking $y = 8\sqrt{K \log d/n}$, we conclude that

$$\begin{aligned}
&\Pr\left(|(\nabla L_\alpha(\mathbf{\Sigma}^*))_{kl}| \geq 8\sqrt{K \log d/n}\right) \\
&\leq 2e^{-(q/n)\cdot 7 \log d} \leq 2/d^3.
\end{aligned}$$

The last inequality follows from $q = \lfloor n/2 \rfloor$. With the union bound, we have

$$\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty \leq 8\sqrt{\frac{K \log d}{n}}$$

with at least $1 - 2/d$ probability.

Now we prove the second statement. We have

$$\mathbb{E}\|\mathbf{W}^*_{\mathcal{S}}\|_{\mathrm{F}} = \mathbb{E}\sqrt{\sum_{(k,l)\in\mathcal{S}} (W^*_{kl})^2} \leq \sqrt{\sum_{(k,l)\in\mathcal{S}} \mathbb{E}(W^*_{kl})^2} \tag{29}$$

The inequality holds because $f(t_1, t_2 \ldots, t_s) = \sqrt{t_1 + t_2 + \cdots + t_s}$ is concave. Recall that for fixed k, l, denote $\Sigma^*_{kl} - (x_{ik} - x_{jk})(x_{il} - x_{jl})/2$ as $\epsilon_{i,j}$. Note that with Lemma 10, $\mathbb{E}\left[\epsilon^2_{i,j}\right] \leq K$. Recall that

$$W^*_{kl} = \frac{1}{N} \sum_{1 \leq i < j \leq n} \{\rho'_\alpha(\epsilon_{i,j}) - \mathbb{E}\rho'_\alpha(\epsilon_{i,j})\}.$$

Then

$$\begin{aligned}
&N^2 \mathbb{E}(W^*_{kl})^2 \\
&= \sum_{(i,j,i',j')\in\mathcal{G}} \mathbb{E}\left[\{\rho'_\alpha(\epsilon_{i,j}) - \mathbb{E}\rho'_\alpha(\epsilon_{i,j})\}\right. \\
&\qquad \left. \cdot \{\rho'_\alpha(\epsilon_{i',j'}) - \mathbb{E}\rho'_\alpha(\epsilon_{i',j'})\}\right],
\end{aligned}$$

where $\mathcal{G} = \{(i,j,i',j') : 1 \leq i < j \leq n, 1 \leq i' < j' \leq n\}$ denote the set of pairs of indices. Let $\mathcal{G}_d = \{(i,j,i',j') \in \mathcal{G} : i, j, i', j' \text{ are not equal to each other}\}$. Notice that $(i,j,i',j') \in \mathcal{G}_d$ indicates that $\epsilon_{i,j}$ and $\epsilon_{i',j'}$ are

independent because $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_{i'}, \mathbf{x}_{j'}$ are independent samples, and consequently,

$$\sum_{(i,j,i',j')\in\mathcal{G}_d} \mathbb{E}\left[\{\rho'_\alpha(\epsilon_{i,j}) - \mathbb{E}\rho'_\alpha(\epsilon_{i,j})\} \cdot \{\rho'_\alpha(\epsilon_{i',j'}) - \mathbb{E}\rho'_\alpha(\epsilon_{i',j'})\}\right] = 0.$$

Therefore,

$$\begin{aligned} & N^2\mathbb{E}(W^*_{kl})^2 \\ &= \sum_{(i,j,i',j')\in\overline{\mathcal{G}_d}} \mathbb{E}\left[\{\rho'_\alpha(\epsilon_{i,j}) - \mathbb{E}\rho'_\alpha(\epsilon_{i,j})\} \cdot \{\rho'_\alpha(\epsilon_{i',j'}) - \mathbb{E}\rho'_\alpha(\epsilon_{i',j'})\}\right]. \end{aligned} \tag{30}$$

A more basic bound yields

$$\begin{aligned} & \mathbb{E}\left[\{\rho'_\alpha(\epsilon_{i,j}) - \mathbb{E}\rho'_\alpha(\epsilon_{i,j})\} \cdot \{\rho'_\alpha(\epsilon_{i',j'}) - \mathbb{E}\rho'_\alpha(\epsilon_{i',j'})\}\right] \\ & \le \left[\mathrm{Var}\left\{\rho'_\alpha(\epsilon_{i,j})\right\} \cdot \mathrm{Var}\left\{\rho'_\alpha(\epsilon_{i,j})\right\}\right]^{1/2} \\ & \le \left[\mathrm{Var}\left\{\epsilon_{i,j}\right\} \cdot \mathrm{Var}\left\{\epsilon_{i,j}\right\}\right]^{1/2} \le K \end{aligned} \tag{31}$$

By definition of $\mathcal{G}_d$, it's easy to verify that $\left|\overline{\mathcal{G}_d}\right| = (2n-3)N$. Combining this with (29), (30) and (31), we have

$$\mathbb{E}\left\|\mathbf{W}^*_{\mathcal{S}}\right\|_{\mathrm{F}} \le \sqrt{\frac{(2n-3)Ks}{N}} < \sqrt{\frac{4Ks}{n}} \tag{32}$$

With Markov's inequality, we have

$$\Pr\left\{\left\|\mathbf{W}^*_{\mathcal{S}}\right\|_{\mathrm{F}} \ge \beta\sqrt{\frac{Ks}{n}}\right\} \le \frac{\mathbb{E}\left\|\mathbf{W}^*_{\mathcal{S}}\right\|_{\mathrm{F}}}{\beta\sqrt{Ks/n}} < \frac{2}{\beta}$$

Recall that $\nabla L_\alpha(\mathbf{\Sigma}^*) = \mathbf{B}(\mathbf{\Sigma}^*) + \mathbf{W}^*$. With Lemma 16, we have $\|\mathbf{B}(\mathbf{\Sigma}^*)_{\mathcal{S}}\|_{\mathrm{F}} \le \sqrt{Ks/n}$. Combing the this with (32), we have

$$\Pr\left\{\left\|\nabla L_\alpha(\mathbf{\Sigma}^*)_{\mathcal{S}}\right\|_{\mathrm{F}} \ge (\beta+1)\sqrt{\frac{Ks}{n}}\right\} \le \frac{2}{\beta}.$$

■

F. Proof of Theorem 13

Proof: Choose $c > 0$ so that

$$0.5p'(\phi_0)\left(c^2+1\right)^{1/2} + 2 = 0.5c\phi_0$$

Set $l = \left(2 + \frac{2}{p'(\phi_0)}\right)\left(c^2+1\right)^{1/2} s^{1/2} + \frac{2}{p'(\phi_0)}s^{1/2}$. Then it's easy to verify that given $r = \alpha/2$,

$$c\phi_0\lambda s^{1/2} \le r \quad \text{and} \quad 5\lambda s^{1/2} \le r.$$

Apply Proposition 8 to conclude that, for any positive constant β, conditioned on event $\mathcal{E}_1(r,1/2) \cap \{\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty + \epsilon \le 0.5p'(\phi_0)\lambda\} \cap \left\{\|\nabla L_\alpha(\mathbf{\Sigma}^*)_{\mathcal{S}}\|_{\mathrm{F}} \le (\beta+1)\sqrt{Ks/n}\right\}$, we have $\widetilde{\mathbf{\Sigma}}^{(t)} \in \mathbf{\Sigma}^* + \mathbb{C}(l)$ and

$$\begin{aligned} \left\|\widetilde{\mathbf{\Sigma}}^{(t)} - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} \le \delta\left\|\widetilde{\mathbf{\Sigma}}^{(t-1)} - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} + 2\{\|p'_\lambda(|\mathbf{\Sigma}^*_{\mathcal{S}}| - \phi_0\lambda)\|_{\mathrm{F}} \\ + s^{1/2}\epsilon + \|\nabla L_\alpha(\mathbf{\Sigma}^*)_{\mathcal{S}}\|_{\mathrm{F}} + \tau\left\|(\mathbf{\Sigma}^*)^{-1}\right\|_{\mathrm{F}}\} \end{aligned} \tag{33}$$

By $\|\mathbf{\Sigma}^*_{\mathcal{S}}\|_{\min} \ge (\phi_0+\phi_1)\lambda$, we have $\|p'_\lambda(|\mathbf{\Sigma}^*_{\mathcal{S}}| - \phi_0\lambda)\|_{\mathrm{F}} = 0$, so

$$\begin{aligned} \left\|\widetilde{\mathbf{\Sigma}}^{(t)} - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} \le \delta\left\|\widetilde{\mathbf{\Sigma}}^{(t-1)} - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} \\ + 2\{s^{1/2}\epsilon + \|\nabla L_\alpha(\mathbf{\Sigma}^*)_{\mathcal{S}}\|_{\mathrm{F}} + \tau\left\|(\mathbf{\Sigma}^*)^{-1}\right\|_{\mathrm{F}}\} \end{aligned}$$

Therefore by Remark 9,

$$\begin{aligned} \left\|\widetilde{\mathbf{\Sigma}}^{(T)} - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} &\le \delta^{T-1}\left\|\widetilde{\mathbf{\Sigma}}^{(1)} - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} \\ &\quad + \frac{2}{1-\delta}\{s^{1/2}\epsilon + \|\nabla L_\alpha(\mathbf{\Sigma}^*)_{\mathcal{S}}\|_{\mathrm{F}} + \tau\left\|(\mathbf{\Sigma}^*)^{-1}\right\|_{\mathrm{F}}\} \\ &\le \sqrt{s/n} + \frac{2}{1-\delta}(\beta{+}3)\sqrt{Ks/n} \lesssim \beta\sqrt{s/n}. \end{aligned}$$

The second inequality uses the bound of $\left\|\widetilde{\mathbf{\Sigma}}^{(1)} - \mathbf{\Sigma}^*\right\|_{\mathrm{F}}$ given in Proposition 6. Provided that $\widetilde{\mathbf{\Sigma}}^{(T)} - \mathbf{\Sigma}^* \in \mathbb{C}(l)$ and $l \asymp s^{1/2}$,

$$\left\|\widetilde{\mathbf{\Sigma}}^{(T)} - \mathbf{\Sigma}^*\right\|_1 \lesssim s^{1/2}\cdot\beta\sqrt{\frac{s}{n}}.$$

Take $n \ge K^{-1}(\log d)^{1+\frac{1}{2\gamma}}$. By Proposition 11, Proposition 12 and the union bound, event $\mathcal{E}_1(r,1/2) \cap \{\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty + \epsilon \le 0.5p'(\phi_0)\lambda\} \cap \left\{\|\nabla L_\alpha(\mathbf{\Sigma}^*)_{\mathcal{S}}\|_{\mathrm{F}} \le (\beta+1)\sqrt{Ks/n}\right\}$ happens with at least $1 - 2/d - d^2\exp(-n/12) - 2/\beta$ probability.

To prove the weaker conclusion, condition only on the event $\{\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty + \epsilon \le 0.5p'(\phi_0)\lambda\}$, which holds with at least $1 - 2/d$ probability. Compute:

$$\begin{aligned} & \lim_{M\to\infty}\limsup_n \Pr\left\{\left\|\widetilde{\mathbf{\Sigma}}^{(T)} - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} \ge M\sqrt{s/n}\right\} \\ &= \lim_{\beta\to\infty}\limsup_n \Pr\left\{\left\|\widetilde{\mathbf{\Sigma}}^{(T)} - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} \ge \sqrt{s/n} + \frac{2}{1-\delta}(\beta+3)\sqrt{s/n}\right\} \\ &\le \lim_{\beta\to\infty}\limsup_n d^2\exp(-n/12) + 2/\beta \\ &= \lim_{\beta\to\infty} 2/\beta = 0. \end{aligned}$$

That is,

$$\left\|\widetilde{\mathbf{\Sigma}}^{(T)} - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} = O_P(\sqrt{s/n}).$$

Similarly, $\left\|\widetilde{\mathbf{\Sigma}}^{(T)} - \mathbf{\Sigma}^*\right\|_1 = O_P(s/\sqrt{n})$. ■

REFERENCES

[1] A. P. Dempster, "Covariance selection," *Biometrics*, vol. 28, no. 1, pp. 157–175, 1972.

[2] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Stat. Appl. Genet. Mol. Biol.*, vol. 4, no. 1, 2005.

[3] O. Ledoit and M. Wolf, "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection," *J. Empir. Finance.*, vol. 10, no. 5, pp. 603–621, 2003.

[4] Z. Zhao and D. P. Palomar, "Mean-reverting portfolio with budget constraint," *IEEE Trans. Signal Process.*, vol. 66, no. 9, pp. 2342–2357, 2018.

[5] Z. Zhao, R. Zhou, and D. P. Palomar, "Optimal mean-reverting portfolio with leverage constraint for statistical arbitrage in finance," *IEEE Trans. Signal Process.*, vol. 67, no. 7, pp. 1681–1695, 2019.

[6] Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero, "Shrinkage algorithms for MMSE covariance estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5016–5029, 2010.

[7] E. Ollila, D. E. Tyler, V. Koivunen, and H. V. Poor, "Complex elliptically symmetric distributions: Survey, new results and applications," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5597–5625, 2012.

[8] Y. Sun, P. Babu, and D. P. Palomar, "Robust estimation of structured covariance matrix for heavy-tailed elliptical distributions," *IEEE Trans. Signal Process.*, vol. 64, no. 14, pp. 3576–3590, 2016.

[9] A. Aubry, A. De Maio, and L. Pallotta, "A geometric approach to covariance matrix estimation and its applications to radar problems," *IEEE Trans. Signal Process.*, vol. 66, no. 4, pp. 907–922, 2018.

[10] E. Ollila, D. P. Palomar, and F. Pascal, "Shrinking the eigenvalues of M-estimators of covariance matrix," *IEEE Trans. Signal Process.*, vol. 69, pp. 256–269, 2021.

[11] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[12] I. T. Jolliffe, *Principal Component Analysis*. Springer, 2002.

[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[14] H. Markowitz, "Portfolio selection," *J. Finance.*, vol. 7, no. 1, pp. 77–91, 1952.

[15] R. A. Monzingo and T. W. Miller, *Introduction to Adaptive Arrays*. SciTech Publishing, 2004.

[16] W. B. Wu and M. Pourahmadi, "Nonparametric estimation of large covariance matrices of longitudinal data," *Biometrika*, vol. 90, no. 4, pp. 831–844, 2003.

[17] P. J. Bickel and E. Levina, "Covariance regularization by thresholding," *Ann. Statist.*, vol. 36, no. 6, pp. 2577–2604, 2008.

[18] N. El Karoui, "Operator norm consistent estimation of large-dimensional sparse covariance matrices," *Ann. Statist.*, vol. 36, no. 6, pp. 2717–2756, 2008.

[19] A. J. Rothman, E. Levina, and J. Zhu, "Generalized thresholding of large covariance matrices," *J. Am. Stat. Assoc.*, vol. 104, no. 485, pp. 177–186, 2009.

[20] T. T. Cai and H. H. Zhou, "Minimax estimation of large covariance matrices under ℓ_1-norm," *Stat. Sin.*, vol. 22, pp. 1319–1378, 2012.

[21] ——, "Optimal rates of convergence for sparse covariance matrix estimation," *Ann. Statist.*, vol. 40, no. 5, pp. 2389–2420, 2012.

[22] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *Ann. Statist.*, vol. 36, no. 1, pp. 199–227, 2008.

[23] T. T. Cai, C.-H. Zhang, and H. H. Zhou, "Optimal rates of convergence for covariance matrix estimation," *Ann. Statist.*, vol. 38, no. 4, pp. 2118–2144, 2010.

[24] X. Chen, Z. J. Wang, and M. J. McKeown, "Shrinkage-to-tapering estimation of large covariance matrices," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5640–5656, 2012.

[25] E. Ollila and A. Breloy, "Regularized tapered sample covariance matrix," *IEEE Trans. Signal Process.*, vol. 70, pp. 2306–2320, 2022.

[26] Q. Wei and Z. Zhao, "Large covariance matrix estimation with oracle statistical rate," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 2023.

[27] L. Xue, S. Ma, and H. Zou, "Positive-definite ℓ_1-penalized estimation of large covariance matrices," *J. Am. Stat. Assoc.*, vol. 107, no. 500, pp. 1480–1491, 2012.

[28] Y. Ke, S. Minsker, Z. Ren, Q. Sun, and W.-X. Zhou, "User-friendly covariance estimation for heavy-tailed distributions," *Statistical Science*, vol. 34, no. 3, pp. 454–471, 2019.

[29] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Am. Stat.*, vol. 58, no. 1, pp. 30–37, 2004.

[30] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, 2017.

[31] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier. Anal. Appl.*, vol. 14, no. 5, pp. 877–905, 2008.

[32] H. A. Le Thi, T. P. Dinh, H. M. Le, and X. T. Vo, "DC approximation approaches for sparse optimization," *Eur. J. Oper. Res.*, vol. 244, no. 1, pp. 26–46, 2015.