Robust Sparse Large Covariance Matrix Estimation Under Huber Loss

XXX

***Abstract*—XXX**

Keywords: Heavy-tailed, fourth moment, non-adaptive,

I. Introduction

[Covariance estimation]

The estimation of covariance matrices is a fundamental problem in modern multivariate data analysis. It has broad applications in many fields such as statistics, biology, finance, signal processing, machine learning, etc. For example, many dimension reduction techniques, including principal component analysis [1] and linear and quadratic discriminant analysis [2], require the estimation of a covariance matrix in advance from a given collection of data points. Theoretical properties of large covariance estimators discussed in the literature often hinge heavily on the Gaussian or sub-Gaussian assumption. Given that data from fields including genomic studies and quantative finance usually do not follow the assumed Gaussian or sub-Gaussian (elliptical) shape, such an assumption is typically very restrictive in practice. It is therefore imperative to develop robust inferential procedures that are less sensitive to the distributional assumptions.

[why robustness: outliers]

Heavy-tailed distribution[1] is a viable model for data contaminated by outliers that are typically encountered in applications. The concept of tail-robustness was first introduced in [3]. "Due to heavy-tailedness, the probability that some observations are sampled far away from the "true" parameter of the population is nonnegligible. We refer to these outlying data points as stochastic outliers. A procedure that is robust against such outliers, evidenced by its better finite-sample performance than a nonrobust method, is called a tail-robust procedure." The tail robustness is different from the classical notion of robustness that is often characterized by the breakdown point. Nevertheless, as stated in [3], it does not provide any information on the convergence properties of an estimator, such as consistency and efficiency. Tail-robustness is a concept that combines robustness, consistency, and finite-sample error bounds.

A typical assumption on the heavy-tailed distribution is the fourth moment condition, the truncation methods and their M-Estimation counterparts from [3], the (adaptive) generalized thresholding methods from [4][5][6], or its variants including the polynomial-tail condition in ℓ^1-penalized estimators [7][8] and the finite kurtoses condition in [3][9][10][11]. The fourth moment assumption is justified in scenarios where the data is subject to heavy-tailed and asymmetric errors. For instance, it is widely known that financial returns typically exhibit heavy tails, and [12] provides further evidence showing that a Student's t-distribution with **four degrees of freedom** displays a tail behavior similar to many asset returns. Another method based on the median-of-means (MOM) technique [13][14][15] do not need the tuning of thresholding parameters but requires stronger assumptions, namely, existence of moments of order six. (should I further explain that results obtained under <4 moment assumptions do not achieve the minimax optimal rate?)

[why sparsity: high-dimensional]

In short: to deal with high-dimensionality, there are ways including the effective rank for sparsity in the spectral [papers] and the canonical definition of sparsity in coefficients.

[Robust and sparse covariance: existing methods]

Apart from the canonical definition of sparsity which assumes a small amount of nonzero elements in the covariance matrix, an alternative class of sparse covariances introduced first by [16] imposes an uniform ℓ^q norm bound on each row or column of the covariance matrix:

$$\mathcal{U}(q, s_0(p), M) = \left\{ \mathbf{\Omega} : \mathbf{\Omega} \succ 0, \max_j \omega_{jj} \le M, \max \sum_{j=1}^{p} |\omega_{ij}|^q \le s_0(p) \right\},$$

This class of covariances is popular in robust sparse covariance/precision matrix estimation, see the adaptive thresholding method based on pilot estimators in [5][6]. While this approach also characterizes a class of exactly sparse matrices when $q = 0$ under the convention that $0^0 = 0$, it has quite a different meaning from the canonical definition of sparsity in the sense that it imposes an uniform bound on the nonzero elements in each row, and thus might suffer from suboptimal statistical rates when nonzero elements in the covariance matrix are imblanced in its rows.

The square-loss ℓ^1-penalized covariance estimator [17][8][7] has been extensively studied for estimating large sparse covariance matrices with assumptions on heavy-tailedness. The proposed Frobenius deviation bound in [17] does not match the minimax optimal statistical rate obtained under gaussian/sub-gaussian assumptions. While [8] and [7] have obtained the minimax optimal statistical rate under heavy-tailedness, their results do not apply in the high-dimensional setting where dimension of data is exponential to the sample size, which is a sufficient setting when the data is gaussian/sub-gaussian.

[Robust scatter matrix estimation based on M-estimation, but they cannot exactly obtain a covariance matrix (in most

[1] The distribution of a random variable X is said to be heavy-tailed if the moment generating function of X is infinite for all $t > 0$, that is, $\mathbb{E}e^{tX} = \infty$ for all $t > 0$. Hence we can only assume weaker moment conditions (for instance, $\mathbb{E}X^4 < \infty$) for heavy-tailed distributions.

cases, up to a scalar)]

1. Cannot exactly obtain a covariance matrix and in most cases, can only obtain a covariance matrix up to a scalar under elliptical assumption for the distribution.

2. The method can be adapted for sparse covariance estimation where all nonzero covariates are contained in an $s \times s$ submatrix, but not applicable to covariance matrices with general sparsity pattern.

[in this paper, we XXX]

Notations. XXX

II. Problem Formulation

Given zero-mean samples $\mathbf{x}_i$, $i = 1, \ldots, n$ from a heavy-tailed distribution, define

$$L_\alpha(\mathbf{\Sigma}) := \sum_{k,\ell} \frac{1}{n} \sum_{i=1}^{n} \rho_\alpha(\Sigma_{k\ell} - x_{ik}x_{i\ell})$$

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

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

Further, define

$$\widehat{\mathbf{\Sigma}} \in \arg\min_{\mathbf{\Sigma}} \left\{ L_\alpha(\mathbf{\Sigma}) + \lambda \|\mathbf{\Sigma}\|_{1,\text{off}} \right\} \tag{1}$$

In this paper, we want to show $\widehat{\mathbf{\Sigma}}$ achieves the minimax optimal statistical rate for robust sparse covariance estimation.

III. Theoretical Results

We denote the underlying true covariance matrix by $\mathbf{\Sigma}^*$. Let $\mathcal{S} = \{(i,j) \mid \Sigma^*_{ij} \neq 0\}$ 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 $\mathbf{x}_i$, $i = 1, \ldots, n$.

Assumption 1. *$\mathbf{x}_i \in \mathbb{R}^d$ is a heavy-tailed random variable with zero mean, i.e. $\mathbb{E}[x_{ij}] = 0$ and $\mathbb{E}\left[|x_{ij}|^4\right] \le \sigma^2$ for all $1 \le j \le d$ with some positive σ.*

Remark 2. Assumption 1 immediately implies that there exists constant $K > 0$ that only depends on σ, such that $\mathbb{E}\left[\left(\Sigma^*_{kl} - x_{ik}x_{il}\right)^2\right] \le K$ for all $k, l \in [d]$. Also note that a scaling scheme of K with respect to d is explicitly assumed. In other words, K also depends on d.

Lemma 3. *Assume $\left\|\nabla L_\alpha(\widehat{\mathbf{\Sigma}})\right\|_\infty < \sqrt{K}\epsilon_{n,d}$ holds with $\epsilon_{n,d}$ to be a deterministic bounded sequence. Let $\alpha \asymp \sqrt{Kn/\log d}$. Take the sample size $n \gtrsim \log d$, then*

$$\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_\infty \lesssim \sqrt{K \log d/n} + \sqrt{K}\epsilon_{n,d} \tag{2}$$

holds with high probability.

Proof: For fixed k, l, let $\widehat{\theta} := (\widehat{\mathbf{\Sigma}})_{kl}$ and define

$$\Psi(\theta) := \frac{1}{n} \sum_{i=1}^{n} \rho'_\alpha(\theta - x_{ik}x_{il}), \qquad \theta \in \mathbb{R}.$$

Note that $\left|\Psi(\widehat{\theta})\right| = \left|\left(\nabla L_\alpha(\widehat{\mathbf{\Sigma}})\right)_{kl}\right| < \sqrt{K}\epsilon_{n,d}$ always hold. In addition, it is easy to verify the inequality that

$$-\log(1 - x + x^2) \le \rho'_1(x) \le \log(1 + x + x^2) \tag{3}$$

By (3) and the fact that $\alpha^{-1}\rho'_\alpha(t) = \rho'_1(t/\alpha)$,

$$\begin{aligned}
&\mathbb{E}e^{(n/\alpha)\cdot\Psi(\theta)} = \prod_{i=1}^{n} \mathbb{E}e^{\rho'_1((\theta - x_{ik}x_{il})/\alpha)} \\
&\le \prod_{i=1}^{n} \mathbb{E}\left\{1 + \alpha^{-1}(\theta - x_{ik}x_{il}) + \alpha^{-2}(\theta - x_{ik}x_{il})^2\right\} \\
&\le \prod_{i=1}^{n} \left[1 + \alpha^{-1}(\theta - \Sigma^*_{kl}) + \alpha^{-2}\left\{(\theta - \Sigma^*_{kl})^2 + K\right\}\right] \\
&\le \exp\left[n\alpha^{-1}(\theta - \Sigma^*_{kl}) + n\alpha^{-2}\left\{(\theta - \Sigma^*_{kl})^2 + K\right\}\right].
\end{aligned} \tag{4}$$

Similarly, it can be shown that

$$\begin{aligned}
&\mathbb{E}e^{-(n/\alpha)\cdot\Psi(\theta)} \\
&\le \exp\left[-n\alpha^{-1}(\theta - \Sigma^*_{kl}) + n\alpha^{-2}\left\{(\theta - \Sigma^*_{kl})^2 + K\right\}\right].
\end{aligned} \tag{5}$$

For $\eta \in (0,1)$, define

$$\begin{aligned}
B_-(\theta) &= (\theta - \Sigma^*_{kl}) + \left\{(\theta - \Sigma^*_{kl})^2 + K\right\}/\alpha - (\alpha/n)\log\eta \\
B_+(\theta) &= -(\theta - \Sigma^*_{kl}) + \left\{(\theta - \Sigma^*_{kl})^2 + K\right\}/\alpha + (\alpha/n)\log\eta
\end{aligned}$$

Together, (4), (5) and Markov's inequality imply

$$\begin{aligned}
&\Pr\left(\Psi(\theta) > B_-(\theta)\right) \le e^{-nB_-(\theta)/\alpha} \cdot \mathbb{E}e^{(n/\alpha)\cdot\Psi(\theta)} \le \eta, \\
\text{and}\quad &\Pr\left(\Psi(\theta) < B_+(\theta)\right) \le e^{-nB_+(\theta)/\alpha} \cdot \mathbb{E}e^{-(n/\alpha)\cdot\Psi(\theta)} \le \eta.
\end{aligned}$$

Let θ_+ be the smallest solution of the quadratic equation $B_+(\theta_+) = \sqrt{K}\epsilon_{n,d}$, and θ_- be the largest solution of the quadratic equation $B_-(\theta_-) = -\sqrt{K}\epsilon_{n,d}$. We need to check that θ_- and θ_+ are well-defined. Let Δ_- and Δ_+ denote the discriminant of equation $B_-(\theta) = -\sqrt{K}\epsilon_{n,d}$ and $B_+(\theta) = \sqrt{K}\epsilon_{n,d}$, respectively. Since $\alpha \asymp \sqrt{Kn/\log d}$, $\epsilon_{n,d} = O(1)$ and by taking $n \gtrsim \log d$, $\eta = 1/d^3$, we have

$$\begin{aligned}
B_-(\Sigma^*_{kl} - \alpha/2) &= -\alpha/4 + K/\alpha - (\alpha/n)\log\eta < -\sqrt{K}\epsilon_{n,d} \\
B_-(\Sigma^*_{kl}) &= K/\alpha - (\alpha/n)\log\eta > -\sqrt{K}\epsilon_{n,d}
\end{aligned}$$

which implies that θ_- is well-defined as a solution to $B_-(\theta) = -\sqrt{K}\epsilon_{n,d}$ on $(\Sigma^*_{kl} - \alpha/2, \Sigma^*_{kl})$. Similarly, θ_+ is also well-defined. Then, with at least $1 - 2\eta$ probability,

$$\Psi(\theta_+) \ge B_+(\theta_+) = \sqrt{K}\epsilon_{n,d} \quad \text{and} \quad \Psi(\theta_-) \le B_-(\theta_-) = -\sqrt{K}\epsilon_{n,d}.$$

Recall that $\left|\Psi(\widehat{\theta})\right| < \sqrt{K}\epsilon_{n,d}$ always hold, and given that $\Psi(\theta)$ is nondecreasing, $\Psi(\theta_-) < \Psi(\widehat{\theta}) < \Psi(\theta_+)$ immediately implies $\theta_- \le \widehat{\theta} \le \theta_+$.

Now we estimate θ_-. Notice that by convexity, the following holds for all $\theta \in (\Sigma^*_{kl} - \alpha/2, \Sigma^*_{kl})$:

$$B_-(\theta) \le (1/2) \cdot (\theta - \Sigma^*_{kl}) + B_-(\Sigma^*_{kl}),$$

which immediately implies that

$$\theta_- - \Sigma^*_{kl} \ge -2\left(K/\alpha - (\alpha/n)\log\eta + \sqrt{K}\epsilon_{n,d}\right).$$

To estimate θ_+, it can be seen that assuming $B_+(\theta_+) - \sqrt{K}\epsilon_{n,d} = K/\alpha + (\alpha/n)\log\eta - \sqrt{K}\epsilon_{n,d} > 0$, we have $\theta_+ \in (\Sigma^*_{kl}, \Sigma^*_{kl} + \alpha/2)$, and similarly

$$\theta_+ - \Sigma^*_{kl} \le 2\left(K/\alpha + (\alpha/n)\log\eta - \sqrt{K}\epsilon_{n,d}\right). \tag{6}$$

Otherwise if $B_+(\theta_+) - \sqrt{K}\epsilon_{n,d} \le 0$, then $\theta_+ \le 0$. Combining this with (6), we have

$$\theta_+ - \Sigma^*_{kl} \le \max\left\{2\left(K/\alpha + (\alpha/n)\log\eta - \sqrt{K}\epsilon_{n,d}\right), 0\right\}.$$

Therefore, with $\theta_- \le \widehat{\theta} \le \theta_+$,

$$\left|\widehat{\theta} - \Sigma^*_{kl}\right| \le 2\left(K/\alpha - (\alpha/n)\log\eta + \sqrt{K}\epsilon_{n,d}\right).$$

With $\eta = 1/d^3$ and the union bound, we have that with at least $1 - 2/d$ probability, $\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_\infty \lesssim \sqrt{K\log d/n} + \sqrt{K}\epsilon_{n,d}$. ■

Proposition 4. *Let $\widetilde{\mathbf{\Sigma}}$ denote any solution to (1). Then, $\widetilde{\mathbf{\Sigma}} \in \mathbf{\Sigma}^* + \mathbb{C}(l)$, where $l = 4s^{1/2}$. Further, assume $\left\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_\infty \le \alpha/2$. Conditioned on the event $\mathcal{E}_1(\alpha/2, 1/2) \cap \{\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty \le 0.5\lambda\}$,*

$$\left\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_F \le 3\lambda s^{1/2} \quad \textit{and} \quad \left\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_I \le 12\lambda s.$$

Proposition 4 gives the deterministic interpretation of Theorem 7. In the following propositions we will analyze the probability of the conditioned event $\mathcal{E}_1(\alpha/2, 1/2) \cap \{\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty \le 0.5\lambda\}$ mentioned in Proposition 4.

Proposition 5. *Suppose that Assumption 1 hold. Recall that K is the constant defined in Remark 2. Assume $\alpha \asymp \sqrt{Kn/\log d}$ and take $n \gtrsim \log d$. Then, for any $\kappa \in (0, 1)$ and $C > 0$,*

$$\langle \nabla L_\alpha(\mathbf{\Sigma}) - \nabla L_\alpha(\mathbf{\Sigma}^*), \mathbf{\Sigma} - \mathbf{\Sigma}^*\rangle \ge \min\{\kappa, \kappa/2C\}\|\mathbf{\Sigma} - \mathbf{\Sigma}^*\|_F^2$$

holds uniformly for all $\mathbf{\Sigma} \in \mathbf{\Sigma}^* + \mathbb{B}^\infty(C\alpha)$ with high probability.

Proof: Let $D_{kl} = (1/n)\sum_{i=1}^n 1\left(|\Sigma^*_{kl} - x_{ik}x_{il}| \le \alpha/2\right)$. By Chebyshev's inequality,

$$\mathrm{E}[D_{kl}] = \Pr\left(|\Sigma^*_{kl} - x_{ik}x_{il}| \le \alpha/2\right) \ge 1 - 4K/\alpha^2 > (1+\kappa)/2.$$

The last inequality holds because $4K/\alpha^2 < (1-\kappa)/2$, which follows from $\alpha \asymp \sqrt{Kn/\log d}$ and by taking $n \gtrsim \log d$.

For each fixed $k, l \in [d]$, let $X_i = 1(|\Sigma^*_{kl} - x_{ik}x_{il}| \le \alpha/2)$. With Hoeffding's inequality,

$$\begin{aligned}
&\Pr\left(\left|\sum_{i=1}^n \{X_i - \mathrm{E}[X_i]\}\right| \ge (1-\kappa)n/2\right) \\
\le & 2\cdot\exp\left(-(1-\kappa)^2 n^2/(2n)\right) = 2\cdot\exp\left(-(1-\kappa)^2 n/2\right)
\end{aligned}$$

and

$$\begin{aligned}
&\Pr\{D_{kl} < \kappa\} \\
\le & \Pr\{|D_{kl} - \mathrm{E}[D_{kl}]| \ge (1-\kappa)/2\} \\
= & \Pr\left\{\left|(1/n)\sum_{i=1}^n \{X_i - \mathrm{E}[X_i]\}\right| \ge (1-\kappa)/2\right\} \\
\le & 2\cdot\exp\left(-(1-\kappa)^2 n/2\right).
\end{aligned}$$

With union bound we have

$$\Pr\left[\min_{k,l} D_{kl} < \kappa\right] \le 2d^2\cdot\exp\left(-(1-\kappa)^2 n/2\right) < 2/d,$$

where the last inequality follows by taking $n \ge 6\log d/(1-\kappa)^2$. Let $\mathcal{G}_{kl} := \{i \in [n] : |\Sigma^*_{kl} - x_{ik}x_{il}| \le \alpha/2\}$. Under the event that $\min_{k,l} D_{kl} \ge \kappa$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \{\rho'_\alpha(\Sigma_{kl} - x_{ik}x_{il}) - \rho'_\alpha(\Sigma^*_{kl} - x_{ik}x_{il})\}\cdot(\Sigma_{kl} - \Sigma^*_{kl}) \\
\ge & \frac{1}{n}\sum_{i\in\mathcal{G}_{kl}} \{\rho'_\alpha(\Sigma_{kl} - x_{ik}x_{il}) - \rho'_\alpha(\Sigma^*_{kl} - x_{ik}x_{il})\}\cdot(\Sigma_{kl} - \Sigma^*_{kl}) \\
\ge & \frac{1}{n}\sum_{i\in\mathcal{G}_{kl}} \min\{|\Sigma_{kl} - \Sigma^*_{kl}|, \alpha/2\}\cdot|\Sigma_{kl} - \Sigma^*_{kl}| \\
\ge & \frac{1}{n}\sum_{i\in\mathcal{G}_{kl}} \min\{1, 1/2C\}(\Sigma_{kl} - \Sigma^*_{kl})^2 \\
\ge & \kappa\min\{1, 1/2C\}(\Sigma_{kl} - \Sigma^*_{kl})^2
\end{aligned}$$

The second last inequality holds since $\mathbf{\Sigma} \in \mathbf{\Sigma}^* + \mathbb{B}^\infty(C\alpha)$ implies $\alpha/2 \ge |\Sigma_{kl} - \Sigma^*_{kl}|/2C$, and the last inequality follows from $|\mathcal{G}_{kl}|/n = D_{kl}$. Therefore

$$\begin{aligned}
&\langle \nabla L_\alpha(\mathbf{\Sigma}) - \nabla L_\alpha(\mathbf{\Sigma}^*), \mathbf{\Sigma} - \mathbf{\Sigma}^*\rangle \\
= & \sum_{k,l}\frac{1}{n}\sum_{i=1}^n \{\rho'_\alpha(\Sigma_{kl} - x_{ik}x_{il}) - \rho'_\alpha(\Sigma^*_{kl} - x_{ik}x_{il})\}\cdot(\Sigma_{kl} - \Sigma^*_{kl}) \\
\ge & \kappa\cdot\min\{1, 1/2C\}\cdot\|\mathbf{\Sigma} - \mathbf{\Sigma}^*\|_{\mathrm{F}}^2
\end{aligned}$$

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

Proposition 5 implies that for any $\kappa \in (0, 1)$ and $C > 0$, with $n \gtrsim \log d$, event $\mathcal{E}_1(C, \min\{\kappa, \kappa/2C\})$ happens with high probability.

Proposition 6. *Suppose that Assumption 1 hold. Let K be the constant defined in Remark 2. Then,*

$$\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty \le \sqrt{6K\log d/n} + 6\alpha\log d/n + K/\alpha \tag{7}$$

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

In Proposition 6, (7) indicates that event $\{\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty \le 0.5\lambda\}$ happens with high probability if we take $\alpha \asymp \sqrt{Kn/\log d}$ and $\lambda \asymp \sqrt{K\log d/n}$.

Theorem 7. *(minimax-optimal rate) Suppose that Assumption 1 holds. Take $\lambda \asymp \sqrt{K\log d/n}$ and let $\alpha \asymp \sqrt{Kn/\log d}$. If the sample size satisfies $n \gtrsim \log d$, then*

$$\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_F \lesssim \sqrt{\frac{Ks\log d}{n}} \quad \textit{and} \quad \left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_1 \lesssim s\sqrt{\frac{K\log d}{n}}$$

hold simultaneously with high probability (w.h.p.).

Proof: The proof combines Proposition 4 with Lemma 3, Proposition 5 and Proposition 6. By Proposition 6, $\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty \lesssim \sqrt{K\log d/n}$ given $\alpha \asymp \sqrt{Kn/\log d}$.

By taking $\lambda \asymp \sqrt{K\log d/n}$, event $\{\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty \le 0.5\lambda\}$ happens with at least $1 - 2/d$ probability.

To invoke Proposition 4, we first notice that given $\nabla L_\alpha(\widehat{\mathbf{\Sigma}}) + \lambda\mathbf{\Xi} = \mathbf{0}$ for some $\mathbf{\Xi} \in \partial\left\|\widehat{\mathbf{\Sigma}}\right\|_{1,\text{off}}$, we must have

$\left\|\nabla L_\alpha(\widehat{\boldsymbol{\Sigma}})\right\|_\infty < 2\lambda$ always hold. Taking the deterministic sequence in Lemma 3 to be $\epsilon_{n,d} \coloneqq \lambda_{n,d}/\sqrt{K} \lesssim \sqrt{\log d/n}$, we conclude that

$$\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\right\|_\infty \lesssim \sqrt{K\log d/n} + 2\lambda \asymp \sqrt{K\log d/n} \le \alpha/2,$$

where the last inequality holds by taking $n \gtrsim \log d$.

With $n \gtrsim \log d$, Proposition 5 indicates that $\mathcal{E}_1(\alpha/2, 1/2)$ happens with high probability. With union bound, event $\mathcal{E}_1(\alpha/2,1/2)\cap\{\|\nabla L_\alpha(\boldsymbol{\Sigma}^*)\|_\infty \le 0.5\lambda\}$ holds with high probability. Under this event and by Proposition 4,

$$\left\|\widehat{\boldsymbol\Sigma}-\boldsymbol\Sigma^*\right\|_{\mathrm{F}} \le 3\lambda s^{1/2} \quad\text{and}\quad \left\|\widehat{\boldsymbol\Sigma}-\boldsymbol\Sigma^*\right\|_1 \le 12\lambda s.$$

Then it suffices to recall $\lambda \asymp \sqrt{K\log d/n}$. ■

IV. Numerical Simulation

XX

V. Conclusion

XX

Appendix

Lemma 8. *For any $\boldsymbol\Sigma \in \mathbb{R}^{d\times d}$ satisfying $\boldsymbol\Sigma_{\overline{\mathcal{S}}} = \mathbf{0}$ and $\epsilon > 0$, provided $\lambda > \|\nabla L_\alpha(\boldsymbol\Sigma)_{\overline{\mathcal{S}}}\|_\infty$, any solution $\widetilde{\boldsymbol\Sigma}$ to (1) satisfies*

$$\begin{aligned}&\left\|(\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\overline{\mathcal{S}}}\right\|_1\\ \le&(\lambda - \|\nabla L_\alpha(\boldsymbol\Sigma)_{\overline{\mathcal{S}}}\|_\infty)^{-1}\\ &\cdot(\lambda + \|\nabla L_\alpha(\boldsymbol\Sigma)_{\mathcal{S}}\|_\infty)\cdot\left\|(\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\mathcal{S}}\right\|_1\end{aligned}$$

Proof: For any $\Xi \in \partial\left\|\widetilde{\boldsymbol\Sigma}\right\|_{1,\text{off}}$, define $\boldsymbol{U}(\Xi) = \nabla L_\alpha(\widetilde{\boldsymbol\Sigma}) + \lambda\Xi \in \mathbb{R}^{d\times d}$. Optimality condition of (1) implies $\inf_\Xi \boldsymbol{U}(\Xi) = 0$. By convexity of $L_\alpha(\boldsymbol\Sigma)$:

$$\langle\nabla L_\alpha(\widetilde{\boldsymbol\Sigma}) - \nabla L_\alpha(\boldsymbol\Sigma), \widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma\rangle \ge 0.$$

Therefore,

$$\begin{aligned}&\|\boldsymbol{U}(\Xi)\|_\infty\left\|\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma\right\|_1 \ge \langle\boldsymbol{U}(\Xi), \widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma\rangle\\ =&\langle\nabla L_\alpha(\widetilde{\boldsymbol\Sigma}) - \nabla L_\alpha(\boldsymbol\Sigma), \widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma\rangle + \langle\nabla L_\alpha(\boldsymbol\Sigma), \widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma\rangle\\ &+\langle\lambda\Xi, \widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma\rangle\\ \ge&0 - \|\nabla L_\alpha(\boldsymbol\Sigma)_{\mathcal{S}}\|_\infty\left\|(\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\mathcal{S}}\right\|_1\\ &-\|\nabla L_\alpha(\boldsymbol\Sigma)_{\overline{\mathcal{S}}}\|_\infty\left\|(\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\overline{\mathcal{S}}}\right\|_1 + \langle\lambda\Xi, \widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma\rangle\end{aligned}$$

Moreover, we have

$$\begin{aligned}&\langle\lambda\Xi, \widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma\rangle\\ =&\lambda\langle\Xi_{\overline{\mathcal{S}}}, (\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\overline{\mathcal{S}}}\rangle + \lambda\langle\Xi_{\mathcal{S}}, (\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\mathcal{S}}\rangle\\ \ge&\lambda\left\|(\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\overline{\mathcal{S}}}\right\|_1 - \lambda\left\|(\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\mathcal{S}}\right\|_1\end{aligned}$$

Together, the last two displays imply

$$\begin{aligned}&\|\boldsymbol{U}(\Xi)\|_\infty\left\|\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma\right\|_1\\ \ge&-\|\nabla L_\alpha(\boldsymbol\Sigma)_{\mathcal{S}}\|_\infty\left\|(\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\mathcal{S}}\right\|_1 - \|\nabla L_\alpha(\boldsymbol\Sigma)_{\overline{\mathcal{S}}}\|_\infty\left\|(\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\overline{\mathcal{S}}}\right\|_1\\ &+\lambda\left\|(\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\overline{\mathcal{S}}}\right\|_1 - \lambda\left\|(\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\mathcal{S}}\right\|_1\end{aligned}$$

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

$$\begin{aligned}0 \ge &- \|\nabla L_\alpha(\boldsymbol\Sigma)_{\mathcal{S}}\|_\infty\left\|(\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\mathcal{S}}\right\|_1 - \|\nabla L_\alpha(\boldsymbol\Sigma)_{\overline{\mathcal{S}}}\|_\infty\left\|(\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\overline{\mathcal{S}}}\right\|_1\\ &+\lambda\left\|(\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\overline{\mathcal{S}}}\right\|_1 - \lambda\left\|(\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\mathcal{S}}\right\|_1\end{aligned}$$

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

Lemma 9. *Conditioned on event $\{\|\nabla L_\alpha(\boldsymbol\Sigma)\|_\infty \le 0.5\lambda\}$, any solution $\widetilde{\boldsymbol\Sigma}$ to (1) satisfies $\widetilde{\boldsymbol\Sigma} \in \boldsymbol\Sigma + \mathbb{C}(l)$, where $l = 4s^{1/2}$. Moreover, assume $\widetilde{\boldsymbol\Sigma} \in \boldsymbol\Sigma + \mathbb{B}^\infty(C\alpha)$. Then, conditioned on the event $\mathcal{E}_1(C\alpha, \kappa) \cap \{\|\nabla L_\alpha(\boldsymbol\Sigma)\|_\infty \le 0.5\lambda\}$,*

$$\begin{aligned}\left\|\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma\right\|_F &\le \kappa^{-1}\left(\lambda s^{1/2} + \|\nabla L_\alpha(\boldsymbol\Sigma)_{\mathcal{S}}\|_F\right)\\ &\le 1.5\kappa^{-1}\lambda s^{1/2}.\end{aligned}$$

Proof: Conditioned on the stated event, Lemma 8 indicates

$$\left\|(\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\overline{\mathcal{S}}}\right\|_1 \le 3\left\|(\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\mathcal{S}}\right\|_1.$$

Therefore,

$$\left\|\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma\right\|_1 \le 4s^{1/2}\left\|\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma\right\|_{\mathrm{F}},$$

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

Now we prove the second statement. Since $\widetilde{\boldsymbol\Sigma} - \boldsymbol\Sigma \in \mathbb{B}^\infty(C\alpha)$, conditioned on event $\mathcal{E}_1(C\alpha,\kappa)$, we have

$$\langle\nabla L_\alpha(\widetilde{\boldsymbol\Sigma}) - \nabla L_\alpha(\boldsymbol\Sigma), \widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma\rangle \ge \kappa\left\|\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma\right\|_{\mathrm{F}}^2 \tag{8}$$

Now we upper bound the right-hand side of (8). For any $\Xi \in \partial\left\|\widetilde{\boldsymbol\Sigma}\right\|_{1,\text{off}}$, write

$$\begin{aligned}&\langle\nabla L_\alpha(\widetilde{\boldsymbol\Sigma}) - \nabla L_\alpha(\boldsymbol\Sigma), \widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma\rangle\\ =&\underbrace{\langle\boldsymbol{U}(\Xi), \widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma\rangle}_{:=\Pi_1} - \underbrace{\langle\nabla L_\alpha(\boldsymbol\Sigma), \widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma\rangle}_{:=\Pi_2} - \underbrace{\langle\lambda\Xi, \widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma\rangle}_{:=\Pi_3}\end{aligned} \tag{9}$$

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

$$\begin{aligned}|\Pi_1| &\le \|\boldsymbol{U}(\Xi)\|_\infty\left\|(\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\overline{\mathcal{S}}}\right\|_1 + \|(\boldsymbol{U}(\Xi))_{\mathcal{S}}\|_{\mathrm{F}}\left\|(\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\mathcal{S}}\right\|_{\mathrm{F}}\\ |\Pi_2| &\le \|\nabla L_\alpha(\boldsymbol\Sigma)_{\mathcal{S}}\|_{\mathrm{F}}\left\|(\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\mathcal{S}}\right\|_{\mathrm{F}}\\ &\qquad+\|\nabla L_\alpha(\boldsymbol\Sigma)_{\overline{\mathcal{S}}}\|_\infty\left\|(\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\overline{\mathcal{S}}}\right\|_1\end{aligned}$$

Turning to Π_3, decompose $\lambda\Xi$ and $\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma$ according to $\mathcal{S}\cup\overline{\mathcal{S}}$ to reach

$$\Pi_3 = \langle(\lambda\Xi)_{\mathcal{S}}, (\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\mathcal{S}}\rangle + \langle(\lambda\Xi)_{\overline{\mathcal{S}}}, (\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\overline{\mathcal{S}}}\rangle$$

Since $\boldsymbol\Sigma_{\overline{\mathcal{S}}} = \mathbf{0}$ and $\Xi \in \partial\left\|\widetilde{\boldsymbol\Sigma}\right\|_{1,\text{off}}$, we have $\langle(\lambda\Xi)_{\overline{\mathcal{S}}}, (\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\overline{\mathcal{S}}}\rangle = \langle(\lambda\Xi)_{\overline{\mathcal{S}}}, \widetilde{\boldsymbol\Sigma}_{\overline{\mathcal{S}}}\rangle = \lambda\left\|\widetilde{\boldsymbol\Sigma}_{\overline{\mathcal{S}}}\right\|_1 = \lambda\left\|(\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\overline{\mathcal{S}}}\right\|_1$. Therefore,

$$\Pi_3 \ge \lambda\left\|(\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\overline{\mathcal{S}}}\right\|_1 - \lambda s^{1/2}\left\|(\widetilde{\boldsymbol\Sigma}-\boldsymbol\Sigma)_{\mathcal{S}}\right\|_{\mathrm{F}}$$

Combining (9) with our estimation for Π_1, Π_2 and Π_3, we have

$$\begin{aligned}&\langle \nabla L_\alpha(\widetilde{\boldsymbol{\Sigma}}) - \nabla L_\alpha(\boldsymbol{\Sigma}), \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\rangle \\ &\leq -\left\{\lambda - \|\nabla L_\alpha(\boldsymbol{\Sigma})\|_\infty - \|\boldsymbol{U}(\boldsymbol{\Xi})\|_\infty\right\} \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\right\|_1 \\ &\quad + \|\nabla L_\alpha(\boldsymbol{\Sigma})_{\mathcal{S}}\|_{\mathrm{F}} \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{S}}\right\|_{\mathrm{F}} + \|(\boldsymbol{U}(\boldsymbol{\Xi}))_{\mathcal{S}}\|_{\mathrm{F}} \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{S}}\right\|_{\mathrm{F}} \\ &\quad + \lambda s^{1/2} \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{S}}\right\|_{\mathrm{F}}\end{aligned}$$

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

$$\begin{aligned}&\langle \nabla L_\alpha(\widetilde{\boldsymbol{\Sigma}}) - \nabla L_\alpha(\boldsymbol{\Sigma}), \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\rangle \\ &\leq -\left\{\lambda - \|\nabla L_\alpha(\boldsymbol{\Sigma})\|_\infty\right\} \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\right\|_1 \\ &\quad + \|\nabla L_\alpha(\boldsymbol{\Sigma})_{\mathcal{S}}\|_{\mathrm{F}} \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{S}}\right\|_{\mathrm{F}} \\ &\quad + \lambda s^{1/2} \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{S}}\right\|_{\mathrm{F}}\end{aligned} \tag{10}$$

It follows from $\widetilde{\boldsymbol{\Sigma}} \in \boldsymbol{\Sigma} + \mathbb{B}^\infty(C\alpha)$, (8) and (10) that conditioned on $\mathcal{E}_1(C\alpha, \kappa) \cap \{\|\nabla L_\alpha(\boldsymbol{\Sigma})\|_\infty \leq 0.5\lambda\}$,

$$\begin{aligned}&\kappa \left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_{\mathrm{F}}^2 \leq \\ &\left\{\lambda s^{1/2} + \|\nabla L_\alpha(\boldsymbol{\Sigma})_{\mathcal{S}}\|_{\mathrm{F}}\right\} \left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_{\mathrm{F}}\end{aligned}$$

Therefore,

$$\begin{aligned}&\left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_{\mathrm{F}} \\ &\leq \kappa^{-1}\left\{\lambda s^{1/2} + \|\nabla L_\alpha(\boldsymbol{\Sigma})_{\mathcal{S}}\|_{\mathrm{F}}\right\} \\ &\leq \kappa^{-1}\{\lambda s^{1/2} + 0.5\lambda s^{1/2}\} = 1.5\kappa^{-1}\lambda s^{1/2}\end{aligned} \tag{11}$$

■

A. Proof of Proposition 4

Proof: $\left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\right\|_{\mathrm{F}} \leq 3\lambda s^{1/2}$ follows immediately from Lemma 9 with $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^*$ and $C = \kappa = 1/2$. Combining this with $\widetilde{\boldsymbol{\Sigma}} \in \boldsymbol{\Sigma}^* + \mathbb{C}(l)$, where $l = 4s^{1/2}$, yields $\left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\right\|_1 \leq 12\lambda s$. ■

B. Proof of Proposition 5

We adopt the following notations for the next stage of proof. Recall that $L_\alpha(\boldsymbol{\Sigma}) = \sum_{k,\ell} \frac{1}{n}\sum_{i=1}^n \rho_\alpha(\Sigma_{k\ell} - x_{ik}x_{i\ell})$. Define $\mathbf{B}^* := \mathbb{E}[\nabla L_\alpha(\boldsymbol{\Sigma}^*)]$, and $\mathbf{W}^* := \nabla L_\alpha(\boldsymbol{\Sigma}^*) - \mathbb{E}[\nabla L_\alpha(\boldsymbol{\Sigma}^*)]$.

Lemma 10. *Recall that K is the constant defined in Remark 2. We have $|(\mathbf{B}^*)_{kl}| = |\mathrm{E}[\rho'_\alpha(\epsilon_{kl})]| < \frac{K}{\alpha}$ for all $k, l \in [d]$.*

Proof: For fixed $k, l \in [d]$, let $\epsilon_{kl} := \Sigma^*_{k\ell} - x_{ik}x_{i\ell}$, then

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

Therefore, for all k, l

$$|(\mathbf{B}^*)_{kl}| = \frac{1}{n}\left|\sum_{i=1}^n \mathrm{E}[\rho'_\alpha(\Sigma^*_{k\ell} - x_{ik}x_{i\ell})]\right| < \frac{K}{\alpha}.$$

■

C. Proof of Proposition 6

Proof: $W^*_{kl} = \frac{1}{n}\sum_{i=1}^n \{\rho'_\alpha(\Sigma^*_{k\ell} - x_{ik}x_{i\ell}) - \mathrm{E}\left[\rho'_\alpha(\Sigma^*_{k\ell} - x_{ik}x_{i\ell})\right]\}$. Given that $|\rho'_\alpha(\Sigma^*_{k\ell} - x_{ik}x_{i\ell})| \leq \alpha$, for all $m \geq 2$:

$$\begin{aligned}&\mathrm{E}[\rho'_\alpha(\Sigma^*_{k\ell} - x_{ik}x_{i\ell})]^m \\ &\leq \alpha^{m-2} \cdot \mathrm{Var}[\rho'_\alpha(\Sigma^*_{k\ell} - x_{ik}x_{i\ell})] \\ &\leq \alpha^{m-2} \cdot \mathrm{Var}[\Sigma^*_{k\ell} - x_{ik}x_{i\ell}] \\ &\leq \alpha^{m-2}K \leq \alpha^{m-2}K \cdot m!/2\end{aligned}$$

The second inequality follows given $\rho'_\alpha(\cdot)$ is 1-Lipschitz. With Bernstein's inequality, for any $t \geq 0$,

$$\begin{aligned}&\Pr\left(\left|\sum_{i=1}^n \{\rho'_\alpha(\Sigma^*_{k\ell} - x_{ik}x_{i\ell}) - \mathrm{E}[\rho'_\alpha(\Sigma^*_{k\ell} - x_{ik}x_{i\ell})]\}\right| \right. \\ &\qquad\qquad \left. \geq \sqrt{2Knt} + 2\alpha t\right) \\ &\leq 2 \cdot \exp\left(-\frac{(\sqrt{2Knt} + 2\alpha t)^2/2}{Kn + \alpha \cdot \sqrt{2Knt} + 2\alpha^2 t}\right) \\ &= 2 \cdot \exp\left(-\frac{Kn + 2\alpha \cdot \sqrt{2Knt} + 2\alpha^2 t}{Kn + \alpha \cdot \sqrt{2Knt} + 2\alpha^2 t} \cdot t\right) < e^{-t}.\end{aligned}$$

Taking $t = 3\log d$ and in conjunction with the union bound,

$$\Pr\left(\|\mathbf{W}^*\|_\infty \geq \sqrt{6K\log d/n} + 6\alpha\log d/n\right) < d^{-1}.$$

Recall that $\nabla L_\alpha(\boldsymbol{\Sigma}^*) = \mathbf{B}^* + \mathbf{W}^*$. With Lemma 10, we have $\|\mathbf{B}^*\|_\infty < K/\alpha$. Combing the two parts together and with the union bound, we have

$$\|\nabla L_\alpha(\boldsymbol{\Sigma}^*)\|_\infty < \sqrt{6K\log d/n} + 6\alpha\log d/n + K/\alpha$$

with at least $1 - d^{-1}$ probability. ■

References

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

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

[3] 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.

[4] 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.

[5] M. Avella-Medina, H. S. Battey, J. Fan, and Q. Li, "Robust estimation of high-dimensional covariance and precision matrices." *Biometrika*, vol. 105 2, pp. 271–284, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49237226

[6] Y. He, P. Liu, X. Zhang, and W. Zhou, "Robust covariance estimation for high-dimensional compositional data with application to microbial communities analysis," *Statistics in Medicine*, vol. 40, no. 15, pp. 3499–3515, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8979

[7] Y. Cui, C. Leng, and D. Sun, "Sparse estimation of high-dimensional correlation matrices," *Comput. Stat. Data Anal.*, vol. 93, pp. 390–403, 2016.

[8] 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.

[9] S. Mendelson and N. Zhivotovskiy, “Robust covariance estimation under $L_4 - L_2$ norm equivalence,” *The Annals of Statistics*, vol. 48, no. 3, pp. 1648 – 1664, 2020. [Online]. Available: https://doi.org/10.1214/19-AOS1862

[10] S. Minsker and X. Wei, “Robust modifications of U-statistics and applications to covariance estimation problems,” *Bernoulli*, vol. 26, no. 1, pp. 694 – 727, 2020. [Online]. Available: https://doi.org/10.3150/19-BEJ1149

[11] S. Minsker and L. Wang, “Robust estimation of covariance matrices: Adversarial contamination and beyond,” 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247292226

[12] R. Cont, “Empirical properties of asset returns: stylized facts and statistical issues,” *Quantitative Finance*, vol. 1, no. 2, pp. 223–236, 2001. [Online]. Available: https://doi.org/10.1080/713665670

[13] C. Blair, “Problem complexity and method efficiency in optimization (a. s. nemirovsky and d. b. yudin),” *SIAM Review*, vol. 27, no. 2, pp. 264–265, 1985. [Online]. Available: https://doi.org/10.1137/1027074

[14] L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira, “Sub-gaussian mean estimators,” *arXiv: Statistics Theory*, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:26805883

[15] S. Minsker and N. Strawn, “Distributed statistical estimation and rates of convergence in normal approximation,” *ArXiv*, vol. abs/1704.02658, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:13393096

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

[17] A. J. Rothman, “Positive definite estimators of large covariance matrices,” *Biometrika*, vol. 99, no. 3, pp. 733–740, 2012.