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

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

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

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

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

I. 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 $K > 0$, such that $\mathbb{E}\left[(\Sigma^*_{kl} - x_{ik}x_{il})^2\right] \le K$ for all $k, l \in [d]$.

Lemma 3. *Let $\widehat{\mathbf{\Sigma}} \in \mathbb{R}^{d\times d}$ be any estimator to the true covariance matrix $\mathbf{\Sigma}^*$. Assume $\left\|\nabla L_\alpha(\widehat{\mathbf{\Sigma}})\right\|_\infty < \beta$ always hold, with some $\beta = O(1)$. Take $\alpha = \sqrt{Kn/\log d}$. If the sample size satisfies $n \gtrsim \log d$, then*

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

holds with high probability.

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

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

Note that $\left|\Psi(\widehat{\theta})\right| = \left|\left(\nabla L_\alpha(\widehat{\mathbf{\Sigma}})\right)_{kl}\right| < \beta$ 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

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

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

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

$$\Pr(\Psi(\theta) > B_-(\theta)) \le e^{-nB_-(\theta)/\alpha}\cdot\mathbb{E}e^{(n/\alpha)\cdot\Psi(\theta)} \le \eta,$$

$$\text{and} \quad \Pr(\Psi(\theta) < B_+(\theta)) \le e^{-nB_+(\theta)/\alpha}\cdot\mathbb{E}e^{-(n/\alpha)\cdot\Psi(\theta)} \le \eta.$$

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

$$\Delta_- = 1 - (4/\alpha)\cdot(K/\alpha - (\alpha/n)\cdot\log\eta + \beta) > 0,$$

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

$$\Psi(\theta_+) \ge B_+(\theta_+) = \beta \quad \text{and} \quad \Psi(\theta_-) \le B_-(\theta_-) = -\beta.$$

Recall that $\left|\Psi(\widehat{\theta})\right| < \beta$ 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_- and 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

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

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

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

Otherwise if $B_+(\theta_+) - \beta \le 0$, then $\theta_+ \le 0$. Combining this with (6), we have

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

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

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

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

Proposition 4. *Let $\widetilde{\mathbf{\Sigma}}$ denote an ϵ-optimal 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* $\mathcal{E}_1(\alpha/2, 1/2) \cap \{\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty + \epsilon \le 0.5\lambda\}$,

$$\left\|\widetilde{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\right\|_F \le 3\lambda s^{1/2} \quad \textit{and} \quad \left\|\widetilde{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\right\|_1 \le 12\lambda s.$$

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

Proposition 5. *Suppose that Assumption 1 holds. Recall that K is the constant defined in Remark 2. Assume $n \gtrsim \log d$. 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 $n \gtrsim \log d$.

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

$$\begin{aligned}\mathrm{Var}[X_i] &= \Pr(|\Sigma^*_{kl} - x_{ik}x_{il}| \le \alpha/2) \\ &\qquad \cdot (1 - \Pr(|\Sigma^*_{kl} - x_{ik}x_{il}| \le \alpha/2)) \\ &\le 1/4\end{aligned}$$

and with $|X_i - \mathrm{E}[X_i]| \le 1$,

$$\mathrm{E}\left|X_i - \mathrm{E}[X_i]\right|^l \le \mathrm{E}\left|X_i - \mathrm{E}[X_i]\right|^2 \cdot 1 \le 1/4.$$

Therefore, with Bernstein's inequality

$$\begin{aligned}&\Pr\left(\left|\sum_{i=1}^n \{X_i - \mathrm{E}[X_i]\}\right| \ge (1-\kappa)n/2\right) \\ \le& 2\cdot\exp\left(-\frac{(1-\kappa)^2 n^2/8}{n/4 + (1-\kappa)n/2}\right) = 2\cdot\exp\left(-\frac{(1-\kappa)^2 n}{6-4\kappa}\right)\end{aligned}$$

and

$$\begin{aligned}&\Pr\{D_{kl} < \kappa\} \\ \le& \Pr\{|D_{kl} - \mathrm{E}[D_{kl}]| \ge (1-\kappa)/2\} \\ =& \Pr\left\{\left|(1/n)\sum_{i=1}^n \{X_i - \mathrm{E}[X_i]\}\right| \ge (1-\kappa)/2\right\} \\ \le& 2\cdot\exp\left(-\frac{(1-\kappa)^2 n}{6-4\kappa}\right).\end{aligned}$$

With union bound we have

$$\Pr\left[\min_{k,l} D_{kl} < \kappa\right] \le 2d^2\cdot\exp\left(-\frac{(1-\kappa)^2 n}{6-4\kappa}\right) < 1/d,$$

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

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

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 $1 - 1/d$ 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 holds. Let K be the constant defined in Remark 2. Assume $\alpha = \sqrt{Kn/\log d}$, then*

$$\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty \le 8\sqrt{\frac{K\log d}{n}} \tag{7}$$

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

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

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

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

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

Proof: The proof combines Proposition 4 with Lemma 3, Proposition 5 and Proposition 6. To invoke Proposition 4, we first notice that given $\left\|\nabla L_\alpha(\widehat{\mathbf{\Sigma}}^+) + \lambda\mathbf{\Xi}\right\|_\infty \le \epsilon$ for some $\mathbf{\Xi} \in \partial\left\|\widehat{\mathbf{\Sigma}}^+\right\|_{1,\mathrm{off}}$, we must have $\left\|\nabla L_\alpha(\widehat{\mathbf{\Sigma}}^+)\right\|_\infty < 2\lambda + \epsilon$ always hold. Lemma 3 indicates that

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

where the last inequality hold with $n \gtrsim \log d$.

By Proposition 6, $\|\nabla L_\alpha(\boldsymbol{\Sigma}^*)\|_\infty \le 8\sqrt{K\log d/n}$. With $\epsilon \lesssim \sqrt{\log d/n}$ and $\lambda \asymp \sqrt{\log d/n}$, event $\{\|\nabla L_\alpha(\boldsymbol{\Sigma}^*)\|_\infty + \epsilon \le 0.5\lambda\}$ happens with at least $1-2/d$ probability. Still, with $n \gtrsim \log d$, Proposition 5 indicates that $\mathcal{E}_1(\alpha/2, 1/2)$ happens with high probability. With union bound, event $\mathcal{E}_1(\alpha/2,1/2) \cap \{\|\nabla L_\alpha(\boldsymbol{\Sigma}^*)\|_\infty + \epsilon \le 0.5\lambda\}$ holds with high probability. Under this event and by Proposition 4,

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

■

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 $\lambda > \|\nabla L_\alpha(\boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\|_\infty + \epsilon$, any ϵ-optimal solution $\widetilde{\boldsymbol{\Sigma}}$ to (1) satisfies*

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

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

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

Therefore,

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

Moreover, we have

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

Together, the last two displays imply

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

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

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

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 $\{\|\nabla L_\alpha(\boldsymbol{\Sigma})\|_\infty + \epsilon \le 0.5\lambda\}$, any ϵ-optimal 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 $\mathcal{E}_1(C\alpha,\kappa) \cap \{\|\nabla L_\alpha(\boldsymbol{\Sigma})\|_\infty + \epsilon \le 0.5\lambda\}$,*

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

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 $\boldsymbol{\Xi} \in \partial\left\|\widetilde{\boldsymbol{\Sigma}}\right\|_{1,\mathrm{off}}$, write

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

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

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

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

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

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

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

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

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

Therefore,

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

■

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: $W^*_{kl} = \frac{1}{n}\sum_{i=1}^n \{\rho'_\alpha(\Sigma^*_{k\ell} - x_{ik}x_{i\ell}) - \mathrm{E}[\rho'_\alpha(\Sigma^*_{k\ell} - x_{ik}x_{i\ell})]\}$. 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,

$$\begin{aligned}&\Pr\left(\left|\sum_{i=1}^n \{\rho'_\alpha(\Sigma^*_{k\ell} - x_{ik}x_{i\ell}) - \mathrm{E}[\rho'_\alpha(\Sigma^*_{k\ell} - x_{ik}x_{i\ell})]\}\right| \geq 7\sqrt{Kn\log d}\right) \\ &\leq 2 \cdot \exp\left(-\frac{(7\sqrt{Kn\log d})^2/2}{Kn + \alpha \cdot 7\sqrt{Kn\log d}}\right) \\ &= 2 \cdot \exp\left(-\frac{49\log d}{16}\right) < \frac{2}{d^3}\end{aligned}$$

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

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

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