Large Sparse Covariance Matrix Sensing from Quadratic Measurements

Wenbin Wang
School of Information Science and Technology
ShanghaiTech University
Shanghai, China
wangwb2023@shanghaitech.edu.cn

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

***Abstract*—Covariance matrix, characterizing the correlation property among multiple random variables, plays a pivotal role in many data analytics areas. A common approach is to directly estimate the covariance matrix from samples, yet it faces substantial challenges when data changes rapidly and the acquisition devices have limited processing power and storage capacity. To address these issues, a classical solution is covariance matrix sensing, which involves compressing signals and estimating the covariance matrix from the sensing measurements. In this paper, we consider a quadratic measurement model and adopt the sparsity assumption, which is prevalent in the high dimensional setting. We propose a regularized least squares type estimator, which can guarantee the positive definiteness and sparse property of the covariance matrices. To obtain the estimator, a multistage convex relaxation algorithm based on majorization-minimization (MM) is developed. We clearly establish the statistical performance of all the sequential approximate solutions produced by the MM-based algorithm, and prove after sufficient iterations the final estimator can actually attain the oracle statistical rate. The theoretical findings are supported by numerical simulations.**

***Index Terms*—Covariance matrix sensing, quadratic measurements, majorization-minimization, positive definiteness, sparsity, non-convex statistical optimization.**

I. Introduction

The covariance matrix, explaining the variability and relationship between different variables, plays a central role in statistical inference and information processing [1]–[3]. A common method for covariance matrix estimation is directly based on the samples. However, this approach often encounters significant challenges due to the rapidly changing data, and the limited processing capability and storage capacity of sampling equipment. Thus, it is desirable to consider recovering covariance matrices from a single data stream and a minimal number of storage measurements. Covariance matrix sensing, a.k.a., covariance matrix sketching, is a classical solution to address these challenging issues. It involves two key processes: sensing the signals, and estimating the covariance matrix from potentially noisy measurements [4].

In this paper, we consider a quadratic covariance matrix sensing model. Denote the ground truth covariance matrix as $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$. We obtain m measurements as follows:

$$y_i = \boldsymbol{a}_i^\top \mathbf{\Sigma} \boldsymbol{a}_i + \eta_i, \qquad i = 1, \ldots, m, \tag{1}$$

where $\{y_i\}_{i=1}^m$ denotes a sequence of measurements, $\left\{\boldsymbol{a}_i \in \mathbb{R}^d\right\}_{i=1}^m$ are the sensing vectors, and $\{\eta_i\}_{i=1}^m$ represent the noise. This model has applications in many fields such as wireless communications [5], optical imaging [6], environmental monitoring [7], etc. In wireless communications, the power spectrum characterizes the distribution of power of a signal across different frequencies, providing crucial insights into the signal properties [8], [9]. Consequently, estimating the power spectrum is vital. Power spectrum estimation enables the reconstruction of signals from far fewer samples than traditionally required by the Nyquist-Shannon sampling theorem, used to infer the spectral characteristics of a signal [10]. This technique focuses on capturing the spectral properties of random signals based on energy observations over time [5]. In this approach, sensing vectors $\left\{\boldsymbol{a}_i \in \mathbb{R}^d\right\}_{i=1}^m$ are employed to interact with the signal $\boldsymbol{x}_t$ at discrete times, where the vector $\boldsymbol{a}_i$ could be implemented using techniques such as random demodulators [11]. The average energy of the signal over N observations is calculated as follows:

$$y_i = \frac{1}{N}\sum_{t=1}^{N} \left(\boldsymbol{a}_i^\top \boldsymbol{x}_t\right)^2 = \boldsymbol{a}_i^\top \mathbf{\Sigma}_N \boldsymbol{a}_i, \text{ for } i = 1, \ldots, m, \tag{2}$$

where $\mathbf{\Sigma}_N = \frac{1}{N}\sum_{t=1}^{N} \boldsymbol{x}_t \boldsymbol{x}_t^\top$ symbolizes the sample covariance matrix. The power spectrum can then be inferred by analyzing the structure of the covariance matrix, typically through spectral analysis techniques like Fourier transforms. This sampling model ties the observed energy directly to the power spectrum, providing a robust way to estimate the spectral distribution of the signal based purely on energy measurements, making it particularly suitable for high-frequency applications where phase measurements may be difficult to obtain [5]. In high-dimensional data processing, covariance sketching emerges as a pivotal application of the quadratic model, particularly advantageous in settings where the data dimensions and volume surpass the available computational resources [4], [12]. This technique fundamentally entails the estimation of covariance matrices through the use of compressed or "sketched" data representations. Specifically, the quadratic nature of the model is employed to transform and reduce the dimensionality of data streams or samples, which are then used to approximate the covariance matrix efficiently. For

instance, considering a sequence of high-dimensional vectors $\boldsymbol{x}_t \in \mathbb{R}^d$, the process involves selecting a random sketching vector $\boldsymbol{a}_{i_t}$ at each time t and computing the quadratic sketch $y_i = \left(\boldsymbol{a}_i^\top \boldsymbol{x}_t\right)^2$. These sketches are subsequently aggregated and normalized to form an estimate of the covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}\left(\boldsymbol{x}_t \boldsymbol{x}_t^\top\right)$, with adjustments for error terms denoted by η_i. Such methodologies not only facilitate the processing of data with limited memory and computational capabilities but also enhance the ability to handle real-time data streams efficiently, thus exemplifying the practical applications of quadratic models in modern statistical data analysis [13]. Overall, each of these applications necessitates the recovery of covariance matrices from a minimal set of quadratic measurements (1).

The covariance matrix sensing technique is inspired by the success of compressed sensing [14], which asserts it is possible to sense and compress signals simultaneously without lose information. To achieve it effectively, exploiting low-dimensional structures inherent in high-dimensional data is a common approach [15], with sparsity being one of the widely adopted assumption [16], i.e., most off-diagonal elements are almost zero, significantly decreasing the number of parameters that need estimation. Under approximately sparse assumption, [8] and [17] investigated the recovery of second-order statistics for nearly sparse cyclostationary signals through random linear measurements, employing ℓ_1-minimization but without performance guarantees. Another pioneering work by Dasarathy et al. [4] have suggested estimating a nearly sparse covariance matrix via measurements formatted as $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top$, where $\boldsymbol{A}$ represents a sketching matrix based on expander graphs.

The ℓ_1 penalty has been utilized in covariance matrix sensing. [18] derived a fundamental guarantee on how many samples are sufficient to approximate the ground truth by ℓ_1 regularization and showed that covariance estimation from compressive measurements can be highly robust under sparse assumption. However, it is widely acknowledge that it introduces a non-negligible bias into the estimator. To mitigate this bias, alternatives like the non-convex smoothly clipped absolute deviation (SCAD) penalty [19] and minimax concave penalty (MCP) [20] have been suggested. The benefits of non-convex penalties can also be found in many other topics such as sparse covariance estimation [21], low-rank matrix recovery [22], high-dimensional graphical models [23], graph trend filtering [24], etc.

Based on these insights, we proposed to estimate covariance matrix from a small number of measurements using the non-convex penalty. And we adopts a sub-Gaussian error assumption, different from the error term in [18] which is assumed to be bounded in ℓ_1 or ℓ_2 norm. Analyzing the statistical optimality of oracle-style convergence is a complex task due to the non-convexity of the penalty function. To address this, we designed a multistage convex relaxation algorithm using a majorization-minimization (MM) approach that solves a series of convex subproblems. This ensures that the approximate local optimum shares the same optimal statistical properties as the unattainable global optimum. Our findings also confirm that the estimator produced by the MM-based algorithm reaches the oracle statistical rate under minimal assumptions. Theoretical analyses are supported by numerical experiments, which demonstrate the method's effectiveness.

Detailed proofs of the results in this paper are available at: https://www.ncvxopt.com/pubs/WangZhao-CovSensing.pdf.

II. Problem Formulation

We denote the matrix to be estimated as $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$. In general, it is NP-hard to estimate the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ from $m < \frac{d(d+1)}{2}$ measurements, unless appropriate structural assumptions, such as sparsity, are assumed to be known as a prior. In this paper, we concentrate on the sparse covariance matrices estimation. Additionally, we outline some standard assumptions on sensing vectors and noise in Assumption 1.

Assumption 1. *We operate under the assumption that $\{\eta_i\}_{i=1}^m$ and $\{\boldsymbol{a}_i\}_{i=1}^m$ consist of i.i.d sub-Gaussian random variables. Specifically, $\eta_i\,(1 \le i \le m)$ is drawn from a sub-Gaussian distribution characterized by a mean of 0 and a variance proxy σ^2. Similarly, $\boldsymbol{a}_i$'s $(1 \le i \le m)$ are considered i.i.d copies of $\boldsymbol{z} = [z_1, \cdots, z_d]^\top$, where each z_i is i.i.d drawn from a distribution satisfying*

$$\mathbb{E}(z_i) = 0, \mathbb{E}\left(z_i^2\right) = 1, \text{ and } \mathbb{E}\left(z_i^4\right) > 1. \tag{3}$$

We propose to estimate the sparse covariance matrices from quadratic measurements using the non-convex penalty. The idea is simple yet powerful. For notational simplicity, let $\boldsymbol{A}_i \coloneqq \boldsymbol{a}_i\boldsymbol{a}_i^\top$ represent the equivalent sensing matrix, $\boldsymbol{Y} \coloneqq [y_1, \cdots, y_m]^\top$ express the equivalent measurements vector, and define the linear operator $\mathcal{A}(\boldsymbol{\Sigma}) : \mathbb{R}^{d\times d} \mapsto \mathbb{R}^m$ that maps a matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ to $\{\langle \boldsymbol{\Sigma}, \boldsymbol{A}_i\rangle\}_{i=1}^m$. Thus, the measures $\boldsymbol{Y}$ obeys $y_i \coloneqq \langle \boldsymbol{A}_i, \boldsymbol{\Sigma}\rangle + \eta_i$ and we can express the estimator as

$$\min_{\boldsymbol{\Sigma}\succ\boldsymbol{0}} \left\{ \frac{1}{2m}\left\|\boldsymbol{Y} - \mathcal{A}(\boldsymbol{\Sigma})\right\|_2^2 - \tau \log\det \boldsymbol{\Sigma} + \sum_{j,k} p_\lambda\left(|\Sigma_{jk}|\right) \right\}. \tag{4}$$

In (4), the first term aims to minimize the least squares errors. The second term, a log-determinant barrier function, ensures positive definiteness with a barrier parameter $\tau \ge 0$. The third term p_λ represents a non-convex penalty function governed by a regularization parameter $\lambda > 0$. We impose certain restrictions on it, as illustrated in Assumption 2.

Assumption 2. *The function $p_\lambda(t)$ defined on $[0, +\infty)$ satisfies:*

- $p_\lambda(t)$ is non-decreasing on $[0, +\infty)$ with $p_\lambda(0) = 0$ and is differentiable almost everywhere on $(0, +\infty)$;
- $0 \le p'_\lambda(t_1) \le p'_\lambda(t_1) \le \lambda$ for all $t_1 \ge t_2 \ge 0$ and $\lim_{t\to 0} p'_t(t) = \lambda$;
- There exists an $\alpha > 0$ such that $p'_\lambda(t) = 0$ for $t \ge \alpha\lambda$;
- There exists some $c \in (0, \alpha)$ such that $p'_\lambda(c\lambda) \ge \frac{\lambda}{2}$.

Under Assumption 2, the initial trio of conditions guarantee sparsity and unbiasedness. The final criterion is invariably satisfied on account of $p'_\lambda(0) = \lambda$ and $p'_\lambda(\alpha\lambda) = 0$ which is listed for the convenience of subsequent theoretical analysis. Typical examples of $p_\lambda(\cdot)$ in Assumption (2) include SCAD [19] and MCP [19].

III. Optimization Algorithm

We now construct an MM-based algorithm to solve (4). The multistage convex relaxation is essentially a sequential optimization framework [25]. At each stage, we take weighted ℓ_1-norm as surrogate function of $\sum_{j,k} p_\lambda(|\Sigma_{jk}|)$. For notational simplicity, we define $f(\mathbf{\Sigma}) = \frac{1}{2m}\|\boldsymbol{Y} - \mathcal{A}(\mathbf{\Sigma})\|_2^2 - \tau \log\det \mathbf{\Sigma}$. Consequently, we consider to solve a sequence of convex relaxation subproblems as follows:

$$\min_{\mathbf{\Sigma}\succ\mathbf{0}} \left\{ f(\mathbf{\Sigma}) + \sum p'_\lambda\left(\left|\widehat{\Sigma}_{jk}^{(q-1)}\right|\right)|\Sigma_{jk}| \right\}, \quad 1 \le q \le Q, \tag{5}$$

where $\widehat{\mathbf{\Sigma}}^{(q)}$ is the optimal solution to the q-th subproblem.

Each subproblem corresponds to a weighted ℓ_1-penalized covariance estimation problem, which can generally be reformulated as follows:

$$\min_{\mathbf{\Sigma}\succ\mathbf{0}} \left\{ f(\mathbf{\Sigma}) + \|\mathbf{\Lambda}\odot\mathbf{\Sigma}\|_1 \right\}, \tag{6}$$

where $\mathbf{\Lambda}$ is the regularized parameter matrix with $\Lambda_{jk} = p'_\lambda\left(\left|\widehat{\Sigma}_{jk}\right|\right) \in [0,\lambda]$. According to the Karush-Kuhn-Tucker (KKT) conditions, the unique sparse global optimal solution $\widehat{\mathbf{\Sigma}}$ for (6) satisfies the first-order optimal condition:

$$\nabla f\left(\widehat{\mathbf{\Sigma}}\right) + \mathbf{\Lambda}\odot\widehat{\mathbf{\Xi}} = 0, \text{ with } \widehat{\mathbf{\Xi}} \in \partial\left\|\widehat{\mathbf{\Sigma}}\right\|_1,$$

in which $\nabla f(\mathbf{\Sigma}) = -\frac{1}{m}\mathcal{A}^*(\boldsymbol{Y} - \mathcal{A}(\mathbf{\Sigma})) - \tau\mathbf{\Sigma}^{-1}$. Obviously, it is impossible to find an exact solution of problem (6) which has no analytical solution. In practice, we settle for the second-best solution defined in 3 to (6), and terminate the iterations when the approximate KKT condition holds.

Definition 3. Given a tolerance level ε, $\widetilde{\mathbf{\Sigma}}^{(q)}$ qualifies as an ε-optimal solution to (6) if $\omega_{\mathbf{\Lambda}^{(q-1)}}\left(\widetilde{\mathbf{\Sigma}}^{(q)}\right) \le \varepsilon$, where

$$\omega_{\mathbf{\Lambda}^{(q-1)}}\left(\widetilde{\mathbf{\Sigma}}\right) = \min_{\mathbf{\Xi}\in\partial\|\mathbf{\Sigma}\|_1} \left\|\nabla f(\mathbf{\Sigma}) + \mathbf{\Lambda}^{(q-1)}\odot\mathbf{\Xi}\right\|_{\max},$$

where $\Lambda_{jk}^{(q-1)} = p'_\lambda\left(\left|\widehat{\Sigma}_{jk}^{(q-1)}\right|\right)$.

The MM algorithm is described in Algorithm 1, where we set $\mathbf{\Sigma}^{\{0\}} = \mathbf{I}$ as a trivial start. To obtain $\widetilde{\mathbf{\Sigma}}^{(q)}$ by solving (6), we applying a proximal Newton method with backtracking line search [26], [27]. We start with $\mathbf{\Sigma}_0^{(q)} = \widetilde{\mathbf{\Sigma}}^{(q-1)}$ and stop the proximal Newton algorithm when $\omega_{\mathbf{\Lambda}^{(q-1)}}\left(\widetilde{\mathbf{\Sigma}}\right)$ researches the prefixed optimization error ε, which is given in Definition on this page.

Algorithm 1: The MM-Based Multistage Convex Relaxation Algorithm for Solving (4).

Input: $\mathcal{S}$, τ, λ;
Initialize $\widetilde{\mathbf{\Sigma}}^{(0)} = \mathbf{I}$
for $q = 1, 2, \ldots, Q$ **do**
  $\Lambda_{jk}^{(q-1)} = p'_\lambda\left(\left|\widetilde{\Sigma}_{jk}^{(q-1)}\right|\right)$;
  obtain $\widetilde{\mathbf{\Sigma}}^{(q)}$ by solving (6);
  $q = q + 1$;
end
Output: $\widetilde{\mathbf{\Sigma}}^{(Q)}$

IV. Statistical Theories

In this section, we first introduce some important definitions and technical assumptions. Following this foundational setup, we establish the statistical convergence rate of our proposed covariance estimator.

A. Assumptions

Definition 4 (Sparse Eigenvalue)**.** Given any positive integer s, we define the largest and smallest s-sparse eigenvalues of $\nabla^2 f(\mathbf{\Sigma})$ as

$$\rho_s^+ = \sup\left\{\mathbf{v}^\top\nabla^2 f(\mathbf{\Sigma})\,\mathbf{v} \mid \|\mathbf{v}\|_2^2 = 1, \|\mathbf{v}\|_0 \le s\right\};$$
$$\rho_s^- = \inf\left\{\mathbf{v}^\top\nabla^2 f(\mathbf{\Sigma})\,\mathbf{v} \mid \|\mathbf{v}\|_2^2 = 1, \|\mathbf{v}\|_0 \le s\right\}.$$

Moreover, we define $\kappa_s = \rho_s^+/\rho_s^-$ as the s-sparse condition number.

The sparse eigenvalue condition is widely studied in the high-dimensional statistics. And it is closely related to the restricted strongly convex, restricted smoothness and restricted eigenvalues properties [28]. Such condition have been employed by [29]–[31].

Definition 5. Define a local region of $\mathbf{\Sigma}^*$ by

$$\mathcal{B}(\mathbf{\Sigma}^*, r) = \{\mathbf{\Sigma} \succ 0 \mid \|\mathbf{\Sigma} - \mathbf{\Sigma}^*\|_{\mathsf{F}} \le r\}.$$

In the following, we will show that $f(\mathbf{\Sigma})$ satisfies local strong convexity and local strong smoothness over a sparse domain. And in our analysis, we set the radius r as $\frac{\rho^-_{2s^*+2\widetilde{s}}}{4\tau\kappa}$.

Assumption 6. *Given $\mathbf{\Sigma} \in \mathcal{B}(\mathbf{\Sigma}^*, r)$ for a generic constant r, there exists a generic constant C_1 such that $\nabla^2 f(\mathbf{\Sigma})$ satisfies the sparse eigenvalue properties with parameters $\rho^-_{2s^*+2\widetilde{s}}$ and $\rho^+_{2s^*+2\widetilde{s}}$ satisfying*

$$0 < \rho^-_{2s^*+2\widetilde{s}} < \rho^+_{2s^*+2\widetilde{s}} < +\infty$$

with $\widetilde{s} \ge C_1\kappa^2_{2s^*+2\widetilde{s}}s^*$ and $\kappa_{2s^*+2\widetilde{s}} = \rho^+_{2s^*+2\widetilde{s}}/\rho^-_{2s^*+2\widetilde{s}}$.

Assumption 6 shows that $\nabla^2 f(\mathbf{\Sigma})$ has bounded largest and non-zero smallest sparse eigenvalues, under the condition that $\mathbf{\Sigma}$ is adequately sparse and proximate to $\mathbf{\Sigma}^*$. This employment of similar conditions has become a common practice

in tackling high-dimensional problems [32]–[35]. Under our design, we confirm the existence of sparse eigenvalue. We refer readers to [36]–[42] for further details.

Assumption 7 (Local Restricted Hessian Smoothness)**.** *Referencing $\widetilde{s}$ introduced in Assumption 6. Constant $L_{2s^*+2\widetilde{s}}$ and r exist such that for any $\mathbf{\Sigma}, \mathbf{\Sigma}' \in \mathcal{B}(\mathbf{\Sigma}^*, r)$ with $\|\mathbf{\Sigma}_{\overline{\mathcal{S}}}\|_0 \leq \widetilde{s}$ and $\|\mathbf{\Sigma}'_{\overline{\mathcal{S}}}\|_0 \leq \widetilde{s}$, the following inequality holds:*

$$\sup_{\mathbf{u}\in\Delta, \|\mathbf{u}\|_{\mathsf{F}}=1} \|\mathbf{u}\|^2_{\nabla^2 f(\mathbf{\Sigma}')-\nabla^2 f(\mathbf{\Sigma})} \leq L_{2s^*+2\widetilde{s}} \left\|\mathbf{\Sigma}-\mathbf{\Sigma}'\right\|_{\mathsf{F}}^2,$$

where $\Delta = \left\{\mathbf{u} \mid \mathsf{supp}(\mathbf{u}) \subseteq \left(\mathsf{supp}(\mathbf{\Sigma}) \cup \mathsf{supp}\left(\mathbf{\Sigma}'\right)\right)\right\}$ *and*

$$\begin{aligned}&\|\mathbf{u}\|^2_{\nabla^2 f(\mathbf{\Sigma}')-\nabla^2 f(\mathbf{\Sigma})}\\ =&\mathsf{vec}^\top(\mathbf{u})\left(\nabla^2 f\left(\mathbf{\Sigma}'\right)-\nabla^2 f(\mathbf{\Sigma})\right)\mathsf{vec}(\mathbf{u}).\end{aligned}$$

Assumption 7 ensures that $\nabla^2 f(\mathbf{\Sigma})$ exhibits Lipschitz continuity in the vicinity of $\mathbf{\Sigma}^*$ across a sparse domain.

In the following, we specify some mild conditions for the true covariance matrix. We represent the true covariance matrix as $\mathbf{\Sigma}^*$, and identify the support set of $\mathbf{\Sigma}^*$ as $\mathcal{S}^* = \left\{(j,k) \mid \mathbf{\Sigma}^*_{jk} \neq 0\right\}$, with s^* denoting its size, that is, $s^* = |\mathcal{S}^*|$.

Assumption 8. *Given the true covariance matrix $\mathbf{\Sigma}^*$, there exist generic constants α, c such that*

$$\|\mathbf{\Sigma}^*_{\mathcal{S}^*}\|_{\min} = \min_{(j,k)\in\mathcal{S}^*} \left|\mathbf{\Sigma}^*_{jk}\right| \geq (\alpha+c)\lambda \gtrsim \lambda,$$

where α and c being constants introduced in Assumption (2)*.*

Assumption 8 guarantees the minimum signal strength condition, which is frequently utilized in the study of non-convex penalized regression problems [19], [20], [43]. In our design case, the tuning parameter λ can be taken to be the order of $\sqrt{\frac{\log d}{m}}$ that could not be too large when the measurements m increases, which guarantees that the estimators are close enough to the true model parameter. Taking the signal strength into account, we can obtain the oracle rate of convergence.

Assumption 9. *For the true covariance matrix $\mathbf{\Sigma}^*$, there exists some constant $\kappa \geq 1$ such that*

$$0 < \frac{1}{\kappa} \leq \lambda_{\mathsf{min}}(\mathbf{\Sigma}^*) \leq \lambda_{\mathsf{max}}(\mathbf{\Sigma}^*) \leq \kappa < \infty.$$

The condition regarding the smallest and largest eigenvalues of the true covariance matrix in Assumption 9 is standard within the existing research on the sparse covariance matrix estimation problem [44].

Assumption 10. *At every stage of addressing the convex relaxed subproblem for all $q \geq 1$, we set*

$$\varepsilon = \frac{C_3}{\sqrt{m}} \leq \frac{\lambda}{8},$$

where C_3 is a predetermined small constant.

Assumption 10 ensures that the solution $\widetilde{\mathbf{\Sigma}}^{(q)}$ obtained at each stage, for all $q \geq 1$, achieves adequate precision. This level of accuracy is essential for the convergence analysis of multistage convex relaxation.

B. Statistical Guarantees and Consequences

Now we introduce the primary result, illustrating the contraction property of the solution path $\left\{\widetilde{\mathbf{\Sigma}}^{(q)}\right\}_{q\geq 1}$. For simplicity, we denote the following symbols. For functionals $f(n)$ and $g(n)$, $f(n) \gtrsim g(n)$ if $f(n) \geq cg(n)$, $f(n) \lesssim g(n)$ if $f(n) \leq Cg(n)$, and $f(n) \asymp g(n)$ if $cg(n) \leq f(n) \leq Cg(n)$ for some constant c and C. $\mathcal{O}_p(\cdot)$ is used to denote bounded in probability.

Theorem 11 (Contraction Property)**.** *Under the Assumptions 1, 6, 8 and 9, the ε-optimal solution $\widetilde{\mathbf{\Sigma}}^{(q)}$ adheres to the following bound:*

$$\begin{aligned}\left\|\widetilde{\mathbf{\Sigma}}^{(q)}-\mathbf{\Sigma}^*\right\|_{\mathsf{F}} \leq& \frac{1}{\rho^-_{2s^*+2\widetilde{s}}} \underbrace{\|(\nabla f(\mathbf{\Sigma}^*))_{\mathcal{S}^*}\|_{\mathsf{F}}}_{\textit{oracle rate}}\\ &+ \frac{1}{\rho^-_{2s^*+2\widetilde{s}}} \underbrace{\varepsilon\sqrt{s^*}}_{\textit{optimization error}} + \underbrace{\delta\left\|\widetilde{\mathbf{\Sigma}}^{(q-1)}-\mathbf{\Sigma}^*\right\|_{\mathsf{F}}}_{\textit{contraction}},\end{aligned} \tag{7}$$

for $1 \leq q \leq Q$, where $\delta \in (0,1)$ is the contraction factor.

Remark 12. The oracle rate describes the statistical convergence rate for an oracle estimator, which has prior knowledge the true support set $\mathcal{S}^*$. The oracle estimator $\widehat{\mathbf{\Sigma}}^O$ is defined as

$$\widehat{\mathbf{\Sigma}}^O = \arg\min_{\mathbf{\Sigma}:\mathbf{\Sigma}_{\overline{\mathcal{S}^*}}=\mathbf{0}} f(\mathbf{\Sigma}).$$

According to the definition, it follows quite straightforwardly that $\widehat{\mathbf{\Sigma}}^O$ fulfills the condition $\left\|\widehat{\mathbf{\Sigma}}^O-\mathbf{\Sigma}^*\right\|_{\mathsf{F}} \lesssim \|(\nabla f(\mathbf{\Sigma}^*))_{\mathcal{S}^*}\|_{\mathsf{F}}$, indicating a close approximation to $\mathbf{\Sigma}^*$ within the confines of the true support set's gradient norms.

Theorem 11 reveals the estimation discrepancy between the ε-optimal solution $\widetilde{\mathbf{\Sigma}}^{(q)}$ and the actual parameter $\mathbf{\Sigma}^*$ is bounded by three components: the oracle rate, the optimization error, and a contraction term. Following this, we detail the precise statistical convergence rate applicable within our framework.

Corollary 13. *Suppose that Assumptions 1, 6, 8, 9 and 10 hold. If $\lambda \asymp \sqrt{\frac{\log d}{m}}$, $\tau \lesssim \sqrt{\frac{1}{m}}\left\|(\mathbf{\Sigma}^*)^{-1}\right\|^{-1}_{\max}$, the ε-optimal solution $\widetilde{\mathbf{\Sigma}}^{(1)}$ satisfies*

$$\left\|\widetilde{\mathbf{\Sigma}}^{(1)}-\mathbf{\Sigma}^*\right\|_{\mathsf{F}} \lesssim \sqrt{\frac{s^*\log d}{m}}$$

with high probability.

Corollary 13 naturally follows from Theorem 11 when $q = 1$. Furthermore, the contraction property, as discussed in Theorem 11, is induced by the MM-based multistage convex relaxation algorithm. To attain the oracle rate, the optimization error must be strategically selected so that $\varepsilon \leq \frac{\|(\nabla f(\mathbf{\Sigma}^*))_{\mathcal{S}^*}\|_{\mathsf{F}}}{\sqrt{s^*}}$ and the parameter Q should be sufficiently large. This leads us to the subsequent conclusion.

Corollary 14. *Suppose that Assumptions 1, 6, 7, 8, 9, and 10 hold. If* $\lambda \asymp \sqrt{\frac{\log d}{m}}$, $\tau \lesssim \sqrt{\frac{1}{m}} \left\|(\mathbf{\Sigma}^*)^{-1}\right\|_{\max}^{-1}$, *and* $Q \gtrsim \log(\lambda\sqrt{m}) \gtrsim \log\log d$, *then the* ε*-optimal solution* $\widetilde{\mathbf{\Sigma}}^{(Q)}$ *satisfies*

$$\left\|\widetilde{\mathbf{\Sigma}}^{(Q)} - \mathbf{\Sigma}^*\right\|_{\mathsf{F}} = \mathcal{O}_p\left(\sqrt{\frac{s^*}{m}}\right).$$

Corollary 14 is a direct consequence of Theorem 11, which suggests that under minimal assumptions, solving no more than approximately $\log\log d$ convex problems is sufficient to reach the oracle rate $\sqrt{\frac{s^*}{m}}$.

V. Numerical Experiments

We compare our proposal with the existing estimator, focusing specifically on the estimator that incorporates an ℓ_1 penalty. We conduct a series of Monte Carlo simulations utilizing matrices of dimensions 50×50. The tuning parameters λ and τ were selected via five-fold cross-validation. Initially, positive symmetric sparse covariance matrices are generated utilizing the "sprandsym" function in MATLAB, defined by a sparsity value k, which represents the quantity of non-zero elements within the matrix. The non-convex penalty function chosen for these experiments is the MCP, with a constant setting of $b = 2$ across all trials.

For each scenario characterized by a specific (m, k) pair, we execute 10 independent iterations. The outcomes are depicted in Figure 1, which illustrates the mean relative error of the estimated covariance matrices as a function of m across varying levels of sparsity. Each measurement includes additive noise, achieved by generating η_i from a normal distribution $\sigma \cdot \mathcal{N}(0,1)$ with $\sigma = 10^{-4}$. From the simulation results, it is evident that the estimator with MCP performs better than the ℓ_1 penalty-based estimator in terms of Frobenius norm error. This supports our theoretical findings that non-convex penalty can reduce the covariance estimation error.

Fig. 2 illustrates the performance of oracle rate in estimating a covariance matrix from quadratic measurement model. The y-axis represents the deviation between an estimated covariance matrix and the true covariance matrix, measured using the Frobenius norm. The x-axis is a scaled measure related to the support size of true covariance matrix and the measurement complexity, indicated by $\sqrt{\frac{s^*}{m}}$. Two lines are plotted to show the estimation error for two different dimensions $(d = \{50, 60\})$ under the sparsity level of $\mathsf{k} = 80$. As the measurement size increase, the error between the estimated and true covariance matrix decreases. The black line $(d = 60)$ shows a slightly higher error trend than the blue line $(d = 50)$, suggesting that higher-dimensional covariance tends to have larger deviations from the true matrix. This visual representation provides insight into how the quality of covariance matrix estimation varies with the oracle rate.

VI. Conclusion and Discussion

This paper introduces a new sample strategy named covariance matrix sensing, including sensing and compressing signals and estimating the covariance matrix from a small number of measurements. Specifically, we estimate large sparse covariance matrices from quadratic samplings through a non-convex penalty, detailing both its theoretical results and empirical validations. We demonstrate that our estimators achieve superior statistical convergence rates when contrasted with current methodologies.

And all theoretical derivations and the suggested algorithm can be seamlessly applied to the bilinear measurement model where $y_i = \boldsymbol{a}_i^\top \mathbf{\Sigma} \boldsymbol{b}_i$, with $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ being independently produced sensing vectors. This demonstrates that our outcomes are equally applicable to this asymmetric sensing model too.

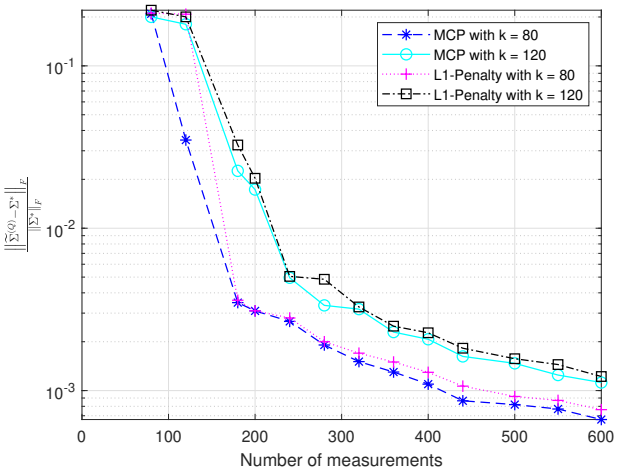

Fig. 1. The mean relative error of the estimated covariance matrices for different sparsity level with noise.

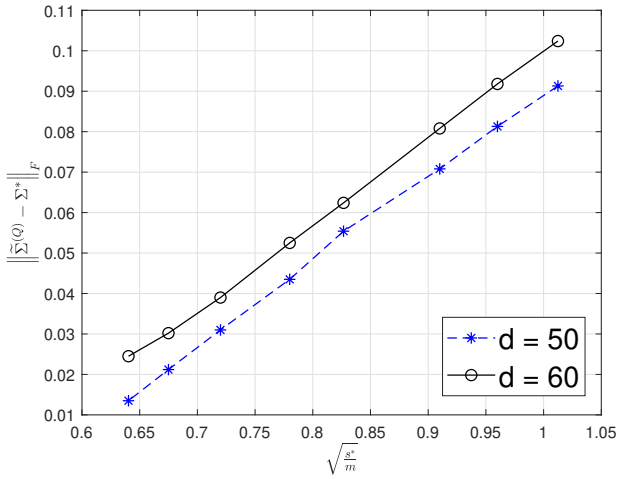

Fig. 2. The oracle rate for different measurements and different dimension of $\mathbf{\Sigma}^*$.

REFERENCES

[1] M. Abt and W. J. Welch, "Fisher information and maximum-likelihood estimation of covariance parameters in Gaussian stochastic processes," *Canadian Journal of Statistics*, vol. 26, no. 1, pp. 127–137, 1998.

[2] M. Pourahmadi, *High-dimensional covariance estimation: with high-dimensional data*. John Wiley & Sons, 2013, vol. 882.

[3] Z. Bao, X. Ding, J. Wang, and K. Wang, "Statistical inference for principal components of spiked covariance matrices," *The Annals of Statistics*, vol. 50, no. 2, pp. 1144–1169, 2022.

[4] G. Dasarathy, P. Shah, B. N. Bhaskar, and R. Nowak, "Covariance sketching," in *Proceedings of 2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2012, pp. 1026–1033.

[5] R. C. Daniels and R. W. Heath, "60 GHz wireless communications: Emerging requirements and design recommendations," *IEEE Vehicular Technology Magazine*, vol. 2, no. 3, pp. 41–50, 2007.

[6] M. Raymer, M. Beck, and D. McAlister, "Complex wave-field reconstruction using phase-space tomography," *Physical Review Letters*, vol. 72, no. 8, p. 1137, 1994.

[7] M. Behrendt, "Uncertainty modelling in power spectrum estimation of environmental processes," 2022.

[8] G. Leus and Z. Tian, "Recovering second-order statistics from compressive measurements," in *Proceedings of 2011 4th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2011, pp. 337–340.

[9] D. D. Ariananda and G. Leus, "Compressive wideband power spectrum estimation," *IEEE Transactions on signal processing*, vol. 60, no. 9, pp. 4775–4789, 2012.

[10] ——, "Compressive Wideband Power Spectrum Estimation," *IEEE Transactions on Signal Processing*, vol. 60, no. 9, pp. 4775–4789, 2012.

[11] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk, "Beyond nyquist: Efficient sampling of sparse bandlimited signals," *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 520–544, 2009.

[12] S. Muthukrishnan *et al.*, "Data streams: Algorithms and applications," *Foundations and Trends® in Theoretical Computer Science*, vol. 1, no. 2, pp. 117–236, 2005.

[13] G. Dasarathy, P. Shah, B. N. Bhaskar, and R. D. Nowak, "Sketching sparse matrices, covariances, and graphs via tensor products," *IEEE Transactions on Information Theory*, vol. 61, no. 3, pp. 1373–1388, 2015.

[14] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[15] J. Wright and Y. Ma, *High-dimensional data analysis with low-dimensional models: Principles, computation, and applications*. Cambridge University Press, 2022.

[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] Z. Tian, Y. Tafesse, and B. M. Sadler, "Cyclic feature detection with sub-Nyquist sampling for wideband spectrum sensing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 1, pp. 58–69, 2011.

[18] Y. Chen, Y. Chi, and A. J. Goldsmith, "Exact and Stable Covariance Estimation From Quadratic Sampling via Convex Programming," *IEEE Transactions on Information Theory*, vol. 61, no. 7, pp. 4034–4059, 2015.

[19] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[20] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Annals of Statistics*, vol. 38, pp. 894–942, 2010.

[21] Q. Wei and Z. Zhao, "Large Covariance Matrix Estimation With Oracle Statistical Rate via Majorization-Minimization," *IEEE Transactions on Signal Processing*, vol. 71, pp. 3328–3342, 2023.

[22] H. Gui, J. Han, and Q. Gu, "Towards faster rates and oracle property for low-rank matrix estimation," in *Proceedings of International Conference on Machine Learning*. PMLR, 2016, pp. 2300–2309.

[23] Q. Sun, K. M. Tan, H. Liu, and T. Zhang, "Graphical nonconvex optimization via an adaptive convex relaxation," in *Proceedings of International Conference on Machine Learning*. PMLR, 2018, pp. 4810–4817.

[24] R. Varma, H. Lee, J. Kovačević, and Y. Chi, "Vector-valued graph trend filtering with non-convex penalties," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 6, pp. 48–62, 2019.

[25] T. Zhang, "Analysis of multi-stage convex relaxation for sparse regularization." *Journal of Machine Learning Research*, vol. 11, no. 3, 2010.

[26] X. Li, L. Yang, J. Ge, J. Haupt, T. Zhang, and T. Zhao, "On quadratic convergence of DC proximal Newton algorithm in nonconvex sparse learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[27] C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, P. Ravikumar *et al.*, "QUIC: quadratic approximation for sparse inverse covariance estimation." *Journal of Machine Learning Research*, vol. 15, no. 83, pp. 2911–2947, 2014.

[28] S. Zhou, "Restricted eigenvalue conditions on subgaussian random matrices," *arXiv preprint arXiv:0912.4045*, 2009.

[29] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *The Annals of Statistics*, vol. 37, no. 4, pp. 1705 – 1732, 2009.

[30] P.-L. Loh and M. J. Wainwright, "Regularized m-estimators with nonconvexity: statistical and algorithmic theory for local optima," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 559–616, jan 2015.

[31] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Statistical Science*, vol. 27, no. 4, pp. 538 – 557, 2012.

[32] L. Xiao and T. Zhang, "A proximal-gradient homotopy method for the sparse least-squares problem," *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 1062–1091, 2013.

[33] Z. Wang, H. Liu, and T. Zhang, "Optimal computational and statistical rates of convergence for sparse nonconvex learning problems," *Annals of Statistics*, vol. 42, no. 6, p. 2164, 2014.

[34] X. Li, J. Haupt, R. Arora, H. Liu, M. Hong, and T. Zhao, "A first order free lunch for sqrt-lasso," *arXiv preprint arXiv:1605.07950*, 2016.

[35] X. Li, T. Zhao, R. Arora, H. Liu, and J. Haupt, "Stochastic variance reduced optimization for nonconvex sparse learning," in *Proceedings of International Conference on Machine Learning*. PMLR, 2016, pp. 917–925.

[36] V. Sivakumar, A. Banerjee, and P. Ravikumar, "Beyond sub-gaussian measurements: High-dimensional structured estimation with sub-exponential designs," *Advances in Neural Information Processing Systems*, vol. 28, pp. 2197–2205, 2015.

[37] V. Koltchinskii and S. Mendelson, "Bounding the Smallest Singular Value of a Random Matrix Without Concentration," *International Mathematics Research Notices*, vol. 2015, no. 23, pp. 12 991–13 008, 03 2015.

[38] R. I. Oliveira, "The lower tail of random quadratic forms with applications to ordinary least squares," *Probability Theory and Related Fields*, vol. 166, no. 3, pp. 1175–1194, Dec 2016.

[39] S. Mendelson and G. Paouris, "On generic chaining and the smallest singular value of random matrices with heavy tails," *Journal of Functional Analysis*, vol. 262, no. 9, pp. 3775–3811, 2012.

[40] A. K. Kuchibhotla and A. Chakrabortty, "Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression," *Information and Inference: A Journal of the IMA*, vol. 11, no. 4, pp. 1389–1456, 2022.

[41] M. Genzel and C. Kipp, "Generic error bounds for the generalized lasso with sub-exponential data," *Sampling Theory, Signal Processing, and Data Analysis*, vol. 20, no. 2, p. 15, 2022.

[42] S. van de Geer and A. Muro, "On higher order isotropy conditions and lower bounds for sparse quadratic forms," *Electronic Journal of Statistics*, vol. 8, no. 2, pp. 3031 – 3061, 2014.

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

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

[45] P. Bühlmann and S. Van De Geer, *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

APPENDIX A
PROOF OF STATISTICAL THEORY

In this appendix, we first provide essential lemmata and then provide the proofs of all the statistical theoretical results in Section IV-B. In Subsection A-A, we begin by the fact that the existence of SE within our design framework. Subsequently, we give the definition of Localized Restricted Eigenvalues (LRE), and then demonstrate the relationship between SE and LRE. Followed by a proposition that connects the SE condition to the localized version of the sparse strong convexity/sparse strong smoothness. Then we introduce Lemma 20, which sets bounds on the estimation error for the general problem **??**, followed by Lemma 21 that outlines the estimation error limited for approximate solutions derived using the MM-based algorithm. Subsection A-B elaborates on Theorem 11, leveraging Lemma 20 and 21 to illustrate the solution path's contraction characteristic. Subsection A-C focuses on key concentration inequalities, which are vital for deducing the explicit statistical convergence rate. In Subsections A-D and A-E, the document presents the proofs for Corollary 13 and Corollary 14, respectively.

A. Technical Lemmata

Lemma 15. *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ have independent isotropic sub-exponential rows. Let $\widehat{\boldsymbol{A}} \subseteq S^{p-1}$, $0 < \xi < 1$, and c is a constant that depends on the sub-exponential norm $K = \sup_{\boldsymbol{u}\in\widehat{\boldsymbol{A}}} \|\langle \boldsymbol{X}, \boldsymbol{u}\rangle\|_{\psi_1}$. Let $\omega_e\left(\widehat{\boldsymbol{A}}\right)$ denote the exponential width of the set. Then for some $\nu > 0$ with probability at least $\left(1 - \exp\left(-\nu^2/2\right)\right)$,*

$$\begin{aligned}&\inf_{\boldsymbol{u}\in\widehat{\boldsymbol{A}}} \|\boldsymbol{X}\boldsymbol{u}\|_2\\ \geq& c\xi\left(1-\xi^2\right)^2\sqrt{n} - 4\omega_e\left(\widehat{\boldsymbol{A}}\right) - \xi\nu.\end{aligned}$$

And if $\{V_j\}$ is $\mathcal{F}_0$-measurable we get for all $\|u\|_2 = 1$,

$$\begin{aligned}&\sup \|\boldsymbol{X}\boldsymbol{u}\|_2\\ \leq& 2^{1-1/h_0}\sqrt{2h}\mu_h + 2^{1-1/h_0}hK,\end{aligned}$$

where the definition of related letters involved is consistent with [36], [42].

Remark 16. Consider the matrix $\widetilde{\boldsymbol{A}}$ constructed as follows:

$$\widetilde{\boldsymbol{A}} = \begin{bmatrix} \mathsf{vec}\left(\boldsymbol{a}_1\boldsymbol{a}_1^\top\right) \\ \vdots \\ \mathsf{vec}\left(\boldsymbol{a}_m\boldsymbol{a}_m^\top\right) \end{bmatrix},$$

where $\widetilde{\boldsymbol{A}} \in \mathbb{R}^{m\times d^2}$ and its rows are independent isotropic sub-exponential vectors. Under the designed moment conditions, such a matrix $\widetilde{\boldsymbol{A}}$ satisfies the sparse eigenvalue condition with high probability. Specifically, it is straightforward to derive a non-trivial asymptotic lower and upper bound for $\widetilde{\boldsymbol{A}}$ with exponential probability. Moreover, the column of $\widetilde{\boldsymbol{A}}$ have ℓ_2 norms in the order of $\sqrt{m}$, which holds with overwhelming probability when $\widetilde{\boldsymbol{A}}$ is under our specified design. For more details on the proof, we refer readers to [36]–[42] for a comprehensive review.

Definition 17. We denote the local ℓ_1 cone as

$$\begin{aligned}&\mathcal{C}(s,\vartheta,r) = \{\boldsymbol{u},\boldsymbol{\Sigma} \mid \mathcal{S}^* \subseteq \mathcal{J}, |\mathcal{J}| \leq s\} \cup\\ &\left\{\boldsymbol{u},\boldsymbol{\Sigma} \mid \left\|\boldsymbol{u}_{\overline{\mathcal{J}}}\right\|_1 \leq \vartheta\left\|\boldsymbol{u}_{\mathcal{J}}\right\|_1, \|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}^*\|_{\mathsf{F}} \leq r\right\}.\end{aligned}$$

Then we define the largest and smallest LRE as

$$\begin{aligned}\psi^+_{s,\vartheta,r} &= \sup_{\boldsymbol{u},\boldsymbol{\Sigma}}\left\{\frac{\boldsymbol{u}^\top\nabla^2 f(\boldsymbol{\Sigma})\,\boldsymbol{u}}{\boldsymbol{u}^\top\boldsymbol{u}} \mid (\boldsymbol{u},\boldsymbol{\Sigma}) \in \mathcal{C}(s,\vartheta,r)\right\},\\ \psi^-_{s,\vartheta,r} &= \inf_{\boldsymbol{u},\boldsymbol{\Sigma}}\left\{\frac{\boldsymbol{u}^\top\nabla^2 f(\boldsymbol{\Sigma})\,\boldsymbol{u}}{\boldsymbol{u}^\top\boldsymbol{u}} \mid (\boldsymbol{u},\boldsymbol{\Sigma}) \in \mathcal{C}(s,\vartheta,r)\right\}.\end{aligned}$$

The following proposition demonstrates the relationships between SE and LRE. The proof can be found in [45], thus is omitted here.

Proposition 18. *Given any $\boldsymbol{\Sigma},\boldsymbol{\Sigma}' \in \mathcal{C}(s,\vartheta,r)\cap\mathcal{B}(\boldsymbol{\Sigma}^*,r)$, we have*

$$\begin{aligned}c_1\psi^-_{s,\vartheta,r} \leq \rho^-_s \leq c_2\psi^-_{s,\vartheta,r},\\ c_3\psi^+_{s,\vartheta,r} \leq \rho^+_s \leq c_4\psi^+_{s,\vartheta,r},\end{aligned}$$

where c_1, c_2, c_3 and c_4 are constant.

Lemma 19. *Let $D_f(\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2) = f(\boldsymbol{\Sigma}_1) - f(\boldsymbol{\Sigma}_2) - \langle\nabla f(\boldsymbol{\Sigma}_2), \boldsymbol{\Sigma}_1-\boldsymbol{\Sigma}_2\rangle$ and define the symmetrized Bregman divergence for the loss function $f(\cdot)$ as*

$$\begin{aligned}D^s_f(\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2) &= D_f(\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2) + D_f(\boldsymbol{\Sigma}_2,\boldsymbol{\Sigma}_1)\\ &= \langle\nabla f(\boldsymbol{\Sigma}_1) - \nabla f(\boldsymbol{\Sigma}_2), \boldsymbol{\Sigma}_1-\boldsymbol{\Sigma}_2\rangle.\end{aligned}$$

For any $\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2 \in \mathcal{B}(\boldsymbol{\Sigma}^*,r)$ such that $\|(\boldsymbol{\Sigma}_1)_{\overline{\mathcal{S}^*}}\|_0 \leq \widetilde{s}$ and $\|(\boldsymbol{\Sigma}_2)_{\overline{\mathcal{S}^*}}\|_0 \leq \widetilde{s}$, we have

$$\begin{aligned}\frac{1}{2}\rho^-_{2s^*+2\widetilde{s}}\|\boldsymbol{\Sigma}_1-\boldsymbol{\Sigma}_2\|_2^2 &\leq D_f(\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2)\\ &\leq \frac{1}{2}\rho^+_{2s^*+2\widetilde{s}}\|\boldsymbol{\Sigma}_1-\boldsymbol{\Sigma}_2\|_2^2,\\ \rho^-_{2s^*+2\widetilde{s}}\|\boldsymbol{\Sigma}_1-\boldsymbol{\Sigma}_2\|_2^2 &\leq D^s_f(\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2)\\ &\leq \rho^+_{2s^*+2\widetilde{s}}\|\boldsymbol{\Sigma}_1-\boldsymbol{\Sigma}_2\|_2^2.\end{aligned}$$

Proof: By the mean value theorem, there exists a $\theta \in [0,1]$ such that

$$\widetilde{\boldsymbol{\Sigma}} = \theta\boldsymbol{\Sigma}_1 + (1-\theta)\boldsymbol{\Sigma}_2 \in \mathcal{B}(\boldsymbol{\Sigma}^*,r),$$

$\left\|\widetilde{\boldsymbol{\Sigma}}\right\|_0 \leq 2\widetilde{s}$ and

$$\langle\nabla f(\boldsymbol{\Sigma}_1) - \nabla f(\boldsymbol{\Sigma}_2), \boldsymbol{\Sigma}_1-\boldsymbol{\Sigma}_2\rangle = \|\boldsymbol{\Sigma}_1-\boldsymbol{\Sigma}_2\|^2_{\nabla^2 f(\widetilde{\boldsymbol{\Sigma}})}.$$

By the definition of the sparse eigenvalue, we obtain the desired result. ∎

Lemma 20. *Suppose that Assumption 9 hold. Consider the general problem in* (6)*. Assume that there exists a set $\mathcal{E}$ such that*

$$\mathcal{S}^* \subseteq \mathcal{E}, |\mathcal{E}| \leq 2s^*, \text{ and } \|\boldsymbol{\Lambda}_{\overline{\mathcal{E}}}\|_{\min} \geq \frac{\lambda}{2}.$$

Then we have $\lambda \geq 2\left(\left\|\nabla f\left(\mathbf{\Sigma}^*\right)\right\|_{\max} + \varepsilon\right)$*, and the* ε*-optimal solution* $\widetilde{\mathbf{\Sigma}}$ *satisfies*

$$\begin{aligned}\left\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_{\mathsf{F}} &\leq \frac{1}{\rho^-_{2s^*+2\widetilde{s}}}\left(\left\|\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)_{\mathcal{E}}\right\|_{\mathsf{F}} + \varepsilon\sqrt{|\mathcal{E}|} + \left\|\mathbf{\Lambda}_{\mathcal{S}^*}\right\|_{\mathsf{F}}\right) \\ &\leq \frac{2+\sqrt{2}}{2\rho^-_{2s^*+2\widetilde{s}}}\lambda\sqrt{s^*}.\end{aligned}$$

Proof: We know that $\left\|\nabla f\left(\mathbf{\Sigma}^*\right)\right\|_{\max} \leq \frac{\lambda}{4}$ and $\varepsilon \leq \frac{\lambda}{8}$ based on assumptions 8 and 10, hence $\lambda \geq 2\left(\left\|\nabla f\left(\mathbf{\Sigma}^*\right)\right\|_{\max} + \varepsilon\right)$. Then we construct an intermediate estimator $\widetilde{\mathbf{\Sigma}}^* = \mathbf{\Sigma}^* + \theta\left(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right)$, where θ is taken such that $\left\|\widetilde{\mathbf{\Sigma}}^* - \mathbf{\Sigma}^*\right\|_{\mathsf{F}} = r$, if $\left\|\widetilde{\mathbf{\Sigma}}^* - \mathbf{\Sigma}^*\right\|_{\mathsf{F}} > r$; $\theta = 1$ otherwise. By the definition of $\widetilde{\mathbf{\Sigma}}^*$, we see that $\left\|\widetilde{\mathbf{\Sigma}}^* - \mathbf{\Sigma}^*\right\|_{\mathsf{F}} \leq r$. Using Lemma 26, we know that the approximate solution falls in the ℓ_1-cone. From the construction of $\widetilde{\mathbf{\Sigma}}^*$, we know $\widetilde{\mathbf{\Sigma}}^* - \mathbf{\Sigma}^* = \theta\left(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right)$. Thus, we have

$$\left\|\left(\widetilde{\mathbf{\Sigma}}^* - \mathbf{\Sigma}^*\right)_{\mathcal{E}}\right\|_1 \leq 5\left\|\left(\widetilde{\mathbf{\Sigma}}^* - \mathbf{\Sigma}^*\right)_{\overline{\mathcal{E}}}\right\|_1.$$

Combining the inequality above with the assumption $|\mathcal{E}| \leq 2s^*$ results that $\widetilde{\mathbf{\Sigma}}^*$ falls in the local ℓ_1-cone. Combing the Lemma 19, Definition 17 and Proposition 18, it implies the localized restricted strong convexity, i.e.

$$\rho^-_{2s^*+2\widetilde{s}}\left\|\widetilde{\mathbf{\Sigma}}^* - \mathbf{\Sigma}^*\right\|^2_{\mathsf{F}} \leq D^s_f\left(\widetilde{\mathbf{\Sigma}}^*, \mathbf{\Sigma}^*\right). \tag{8}$$

We use Lemma 27 to bound the right hand side of the above inequality such as

$$\begin{aligned}D^s_f\left(\widetilde{\mathbf{\Sigma}}^*, \mathbf{\Sigma}^*\right) &\leq \theta D^s_f\left(\widetilde{\mathbf{\Sigma}}, \mathbf{\Sigma}^*\right) \\ &= \theta\left\langle\nabla f\left(\widetilde{\mathbf{\Sigma}}\right) - \nabla f\left(\mathbf{\Sigma}^*\right), \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\rangle.\end{aligned} \tag{9}$$

Plugging (9) back into (8) yields

$$\begin{aligned}\rho^-_{2s^*+2\widetilde{s}}\left\|\widetilde{\mathbf{\Sigma}}^* - \mathbf{\Sigma}^*\right\|^2_{\mathsf{F}} &\leq \theta\left\langle\nabla f\left(\widetilde{\mathbf{\Sigma}}\right) - \nabla f\left(\mathbf{\Sigma}^*\right), \widetilde{\mathbf{\Delta}}\right\rangle \\ &= \theta\underbrace{\left\langle\nabla f\left(\widetilde{\mathbf{\Sigma}}\right) + \mathbf{\Lambda}\odot\widetilde{\mathbf{\Xi}}, \widetilde{\mathbf{\Delta}}\right\rangle}_{\mathbf{I}} \\ &\quad - \theta\underbrace{\left\langle\nabla f\left(\mathbf{\Sigma}^*\right), \widetilde{\mathbf{\Delta}}\right\rangle}_{\mathbf{II}} - \theta\underbrace{\left\langle\mathbf{\Lambda}\odot\widetilde{\mathbf{\Xi}}, \widetilde{\mathbf{\Delta}}\right\rangle}_{\mathbf{III}},\end{aligned} \tag{10}$$

where $\widetilde{\mathbf{\Delta}} = \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*$, $\mathbf{\Xi}^{(t+1)} \in \partial\left\|\mathbf{\Sigma}^{(t+1)}\right\|_1$. Then we establish limits for terms **I**, **II** and **III**, respectively.

Define $\mathbf{\Omega} = \nabla f\left(\widetilde{\mathbf{\Sigma}}\right) + \mathbf{\Lambda}\odot\widetilde{\mathbf{\Xi}}$. For term **I**, partitioning the support of $\mathbf{\Omega}$ and $\widetilde{\mathbf{\Delta}}$ into $\mathcal{E}$ and its complement $\overline{\mathcal{E}}$, and then applying the matrix Hölder's inequality, we obtain

$$\begin{aligned}\mathbf{I} &= \left\langle\mathbf{\Omega}_{\mathcal{E}}, \widetilde{\mathbf{\Delta}}_{\mathcal{E}}\right\rangle + \left\langle\mathbf{\Omega}_{\overline{\mathcal{E}}}, \widetilde{\mathbf{\Delta}}_{\overline{\mathcal{E}}}\right\rangle \\ &\leq \left\|\mathbf{\Omega}_{\mathcal{E}}\right\|_{\mathsf{F}}\left\|\widetilde{\mathbf{\Delta}}_{\mathcal{E}}\right\|_{\mathsf{F}} + \left\|\mathbf{\Omega}_{\overline{\mathcal{E}}}\right\|_{\max}\left\|\widetilde{\mathbf{\Delta}}_{\overline{\mathcal{E}}}\right\|_1 \\ &\leq \sqrt{|\mathcal{E}|}\left\|\mathbf{\Omega}_{\mathcal{E}}\right\|_{\max}\left\|\widetilde{\mathbf{\Delta}}_{\mathcal{E}}\right\|_{\mathsf{F}} + \left\|\mathbf{\Omega}\right\|_{\max}\left\|\widetilde{\mathbf{\Delta}}_{\overline{\mathcal{E}}}\right\|_1 \\ &\underset{(i)}{\leq} \varepsilon\sqrt{|\mathcal{E}|}\left\|\widetilde{\mathbf{\Delta}}_{\mathcal{E}}\right\|_{\mathsf{F}} + \varepsilon\left\|\widetilde{\mathbf{\Delta}}_{\overline{\mathcal{E}}}\right\|_1,\end{aligned}$$

where (i) is from $\left\|\mathbf{\Omega}\right\|_{\max} \leq \varepsilon$ by Definition 3. For term **II**, dividing the support of $\nabla f\left(\mathbf{\Sigma}^*\right)$ and $\widetilde{\mathbf{\Delta}}$ into $\mathcal{E}$ and $\overline{\mathcal{E}}$. Applying the matrix Hölder's inequality, we obtain

$$\begin{aligned}\mathbf{II} &= \left\langle\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)_{\mathcal{E}}, \widetilde{\mathbf{\Delta}}_{\mathcal{E}}\right\rangle + \left\langle\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)_{\overline{\mathcal{E}}}, \widetilde{\mathbf{\Delta}}_{\overline{\mathcal{E}}}\right\rangle \\ &\geq -\left\|\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)_{\mathcal{E}}\right\|_{\mathsf{F}}\left\|\widetilde{\mathbf{\Delta}}_{\mathcal{E}}\right\|_{\mathsf{F}} - \left\|\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)_{\overline{\mathcal{E}}}\right\|_{\max}\left\|\widetilde{\mathbf{\Delta}}_{\overline{\mathcal{E}}}\right\|_1\end{aligned}$$

For term **III**,dividing the support of $\mathbf{\Lambda}\odot\widetilde{\mathbf{\Xi}}$ and $\widetilde{\mathbf{\Delta}}$ into $\mathcal{S}^*$ and $\overline{\mathcal{S}^*}$, and then applying the matrix Hölder's inequality, we obtain

$$\begin{aligned}\mathbf{III} &= \left\langle\left(\mathbf{\Lambda}\odot\widetilde{\mathbf{\Xi}}\right)_{\mathcal{S}^*}, \widetilde{\mathbf{\Delta}}_{\mathcal{S}^*}\right\rangle + \left\langle\left(\mathbf{\Lambda}\odot\widetilde{\mathbf{\Xi}}\right)_{\overline{\mathcal{S}^*}}, \widetilde{\mathbf{\Delta}}_{\overline{\mathcal{S}^*}}\right\rangle \\ &\underset{(i)}{=} \left\langle\left(\mathbf{\Lambda}\odot\widetilde{\mathbf{\Xi}}\right)_{\mathcal{S}^*}, \widetilde{\mathbf{\Delta}}_{\mathcal{S}^*}\right\rangle + \left\langle\mathbf{\Lambda}_{\overline{\mathcal{S}^*}}, \left|\widetilde{\mathbf{\Delta}}_{\overline{\mathcal{S}^*}}\right|\right\rangle \\ &\geq -\left\|\mathbf{\Lambda}_{\mathcal{S}^*}\right\|_{\mathsf{F}}\left\|\widetilde{\mathbf{\Delta}}_{\mathcal{S}^*}\right\|_{\mathsf{F}} + \left\langle\mathbf{\Lambda}_{\overline{\mathcal{E}}}, \left|\widetilde{\mathbf{\Delta}}_{\overline{\mathcal{E}}}\right|\right\rangle \\ &\underset{(ii)}{\geq} -\left\|\mathbf{\Lambda}_{\mathcal{S}^*}\right\|_{\mathsf{F}}\left\|\widetilde{\mathbf{\Delta}}_{\mathcal{E}}\right\|_{\mathsf{F}} + \left\|\mathbf{\Lambda}_{\overline{\mathcal{E}}}\right\|_{\min}\left\|\widetilde{\mathbf{\Delta}}_{\overline{\mathcal{E}}}\right\|_1,\end{aligned}$$

where (i) is from

$$\left\langle\left(\mathbf{\Lambda}\odot\widetilde{\mathbf{\Xi}}\right)_{\overline{\mathcal{S}^*}}, \widetilde{\mathbf{\Delta}}_{\overline{\mathcal{S}^*}}\right\rangle = \left\langle\mathbf{\Lambda}_{\overline{\mathcal{S}^*}}, \left|\widetilde{\mathbf{\Sigma}}_{\overline{\mathcal{S}^*}}\right|\right\rangle = \left\langle\mathbf{\Lambda}_{\overline{\mathcal{S}^*}}, \left|\widetilde{\mathbf{\Delta}}_{\overline{\mathcal{S}^*}}\right|\right\rangle,$$

and (ii) is from

$$\begin{aligned}\left\langle\mathbf{\Lambda}_{\overline{\mathcal{E}}}, \left|\widetilde{\mathbf{\Delta}}_{\overline{\mathcal{E}}}\right|\right\rangle &= \sum_{(j,k)\in\overline{\mathcal{E}}}\mathbf{\Lambda}_{jk}\left|\widetilde{\Delta}_{jk}\right| \geq \left\|\mathbf{\Lambda}_{\overline{\mathcal{E}}}\right\|_{\min}\sum_{(j,k)\in\overline{\mathcal{E}}}\left|\widetilde{\Delta}_{jk}\right| \\ &= \left\|\mathbf{\Lambda}_{\overline{\mathcal{E}}}\right\|_{\min}\left\|\widetilde{\mathbf{\Delta}}_{\overline{\mathcal{E}}}\right\|_1,\end{aligned}$$

and

$$\left\|\widetilde{\mathbf{\Delta}}_{\mathcal{S}^*}\right\|_{\mathsf{F}} \leq \left\|\widetilde{\mathbf{\Delta}}_{\mathcal{E}}\right\|_{\mathsf{F}}.$$

Combining the above results into 10 yields

$$\begin{aligned}&\rho^-_{2s^*+2\widetilde{s}}\left\|\widetilde{\mathbf{\Sigma}}^* - \mathbf{\Sigma}^*\right\|^2_{\mathsf{F}} \\ &\leq \left(\left\|\mathbf{\Lambda}_{\mathcal{S}^*}\right\|_{\mathsf{F}} + \left\|\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)_{\mathcal{E}}\right\|_{\mathsf{F}} + \varepsilon\sqrt{|\mathcal{E}|}\right)\times\theta\left\|\widetilde{\mathbf{\Delta}}_{\mathcal{E}}\right\|_{\mathsf{F}} \\ &\quad - \theta\left(\left\|\mathbf{\Lambda}_{\overline{\mathcal{E}}}\right\|_{\min} - \left(\left\|\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)_{\overline{\mathcal{E}}}\right\|_{\max} + \varepsilon\right)\right)\left\|\widetilde{\mathbf{\Delta}}_{\overline{\mathcal{E}}}\right\|_1 \\ &\underset{(i)}{\leq} \left(\left\|\mathbf{\Lambda}_{\mathcal{S}^*}\right\|_{\mathsf{F}} + \left\|\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)_{\mathcal{E}}\right\|_{\mathsf{F}} + \varepsilon\sqrt{|\mathcal{E}|}\right) \\ &\quad \times\theta\left\|\left(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right)_{\mathcal{E}}\right\|_{\mathsf{F}} \\ &\leq \left(\left(\left\|\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)_{\mathcal{E}}\right\|_{\max} + \varepsilon\right)\sqrt{|\mathcal{E}|} + \left\|\mathbf{\Lambda}_{\mathcal{S}^*}\right\|_{\max}\sqrt{|\mathcal{S}^*|}\right) \\ &\quad \times\theta\left\|\left(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right)_{\mathcal{E}}\right\|_{\mathsf{F}} \\ &\leq \left(\left(\left\|\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)_{\mathcal{E}}\right\|_{\max} + \varepsilon\right)\sqrt{2s^*} + \lambda\sqrt{s^*}\right) \\ &\quad \times\theta\left\|\left(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right)_{\mathcal{E}}\right\|_{\mathsf{F}} \\ &\leq \frac{2+\sqrt{2}}{2}\lambda\sqrt{s^*}\times\underbrace{\theta\left\|\left(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right)_{\mathcal{E}}\right\|_{\mathsf{F}}}_{\mathbf{IV}},\end{aligned}$$

where (i) is due to the fact that $\left\|\mathbf{\Lambda}_{\overline{\mathcal{E}}}\right\|_{\min} \geq \frac{\lambda}{2} \geq \left\|\nabla f\left(\mathbf{\Sigma}^*\right)\right\|_{\max} + \varepsilon$.

For **IV**, we have $\theta \left\| \left(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^* \right)_{\mathcal{E}} \right\|_{\mathsf{F}} = \left\| \left(\widetilde{\mathbf{\Sigma}}^* - \mathbf{\Sigma}^* \right)_{\mathcal{E}} \right\|_{\mathsf{F}}$. Thus, we obtain

$$\rho^-_{2s^*+2\widetilde{s}} \left\| \widetilde{\mathbf{\Sigma}}^* - \mathbf{\Sigma}^* \right\|_{\mathsf{F}} \leq \frac{2+\sqrt{2}}{2} \lambda \sqrt{s^*},$$

which is a contraction with the construction of $\widetilde{\mathbf{\Sigma}}^*$. This indicates that $\widetilde{\mathbf{\Sigma}}^* = \widetilde{\mathbf{\Sigma}}$. Therefore, the desired bound hold for $\widetilde{\mathbf{\Sigma}}$. ∎

Lemma 21. *Suppose that Assumptions 1 and 9 hold. Define the set $\mathcal{E}^{(q)}$ by*

$$\mathcal{E}^{(q)} = \mathcal{S}^* \cup \mathcal{S}^{(q)}, \text{ with } \mathcal{S}^{(q)} = \left\{ (j,k) \mid \mathbf{\Lambda}^{(q-1)}_{jk} \leq p'_\lambda(u) \right\},$$

where $u = c\lambda$ and $c = \frac{2+\sqrt{2}}{2\rho^-_{2s^*+2\widetilde{s}}}$ is the same given in the Assumption 2. Then we have $\lambda \geq 2\left(\|\nabla f(\mathbf{\Sigma}^*)\|_{\max} + \varepsilon\right)$, and for $q \geq 1$, $\left|\mathcal{E}^{(q)}\right| \leq 2s^*$, $\left\| \mathbf{\Lambda}^{(q-1)}_{\mathcal{E}^{(q)}} \right\|_{\min} \geq \frac{\lambda}{2}$, and

$$\begin{aligned}
\left\| \widetilde{\mathbf{\Sigma}}^{(q)} - \mathbf{\Sigma}^* \right\|_{\mathsf{F}} \leq & \frac{1}{\rho^-_{2s^*+2\widetilde{s}}} \left(\left\| \left(\nabla f(\mathbf{\Sigma}^*) \right)_{\mathcal{E}^{(q)}} \right\|_{\mathsf{F}} + \varepsilon \sqrt{\left|\mathcal{E}^{(q)}\right|} \right) \\
& + \frac{1}{\rho^-_{2s^*+2\widetilde{s}}} \left\| \mathbf{\Lambda}^{(q-1)}_{\mathcal{S}^*} \right\|_{\mathsf{F}} \\
\leq & \frac{2+\sqrt{2}}{2\rho^-_{2s^*+2\widetilde{s}}} \lambda \sqrt{s^*}.
\end{aligned}$$

Proof: We first prove $\left|\mathcal{E}^{(q)}\right| \leq 2s^*$ holds by induction. For $q = 1$, we have $\Lambda^{(0)}_{jk} = \lambda \geq p'_\lambda(u)$ for $j \neq k$ and thus $\left|\mathcal{S}^{(1)}\right| \leq s^*$ and $\mathcal{E}^{(1)} = \mathcal{S}^* \cup \mathcal{S}^{(1)}$, which implies $\left|\mathcal{E}^{(1)}\right| \leq 2s^*$ holds. Assume $\left|\mathcal{E}^{(q)}\right| \leq 2s^*$ holds at $q-1$, i.e., $\left|\mathcal{E}^{(q-1)}\right| \leq 2s^*$ holds for some $q \geq 2$. Next, we will prove $\left|\mathcal{E}^{(q)}\right| \leq 2s^*$ holds at q. For any $(j,k) \in \mathcal{S}^{(q)}$, we obtain $\left|\widetilde{\Sigma}^{(q-1)}_{jk}\right| \geq u$ and further have

$$\begin{aligned}
\sqrt{\left|\mathcal{S}^{(q)} \backslash \mathcal{S}^*\right|} &\leq \sqrt{\sum_{(j,k) \in \mathcal{S}^{(q)} \backslash \mathcal{S}^*} \left(u^{-1} \widetilde{\mathbf{\Sigma}}^{(q-1)}_{jk} \right)^2} \\
&= u^{-1} \left\| \widetilde{\mathbf{\Sigma}}^{(q-1)}_{\mathcal{S}^{(q)} \backslash \mathcal{S}^*} \right\|_{\mathsf{F}} \\
&= u^{-1} \left\| \left(\widetilde{\mathbf{\Sigma}}^{(q-1)} - \mathbf{\Sigma}^* \right)_{\mathcal{S}^{(q)} \backslash \mathcal{S}^*} \right\|_{\mathsf{F}} \\
&\leq u^{-1} \left\| \widetilde{\mathbf{\Sigma}}^{(q-1)} - \mathbf{\Sigma}^* \right\|_{\mathsf{F}}.
\end{aligned} \tag{11}$$

For any $(j,k) \in \overline{\mathcal{S}^{(q-1)}}$, we have $\Lambda^{(q-2)}_{jk} = p'_\lambda\left(\widetilde{\Sigma}^{(q-2)}_{jk}\right) \geq p'_\lambda(u) \geq \frac{\lambda}{2}$, which implies

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

One also has $\left|\mathcal{E}^{(q-1)}\right| \leq 2s^*$ and $\mathcal{S}^* \subseteq \mathcal{E}^{(q-1)}$. Applying Lemma 20 with $\widetilde{\mathbf{\Sigma}} = \widetilde{\mathbf{\Sigma}}^{(q-1)}$, $\mathcal{E} = \mathcal{E}^{(q-1)}$, and $\mathbf{\Lambda}_{\mathcal{S}^*} = \mathbf{\Lambda}^{(q-2)}_{\mathcal{S}^*}$ yields

$$\left\| \widetilde{\mathbf{\Sigma}}^{(q-2)} - \mathbf{\Sigma}^* \right\|_{\mathsf{F}} \leq \frac{2+\sqrt{2}}{2\rho^-_{2s^*+2\widetilde{s}}} \lambda \sqrt{s^*}.$$

Substituting the above result into the inequality 11 yields

$$\sqrt{\left|\mathcal{S}^{(q)} \backslash \mathcal{S}^*\right|} \leq \frac{2+\sqrt{2}}{2u\rho^-_{2s^*+2\widetilde{s}}} \lambda \sqrt{s^*} = \sqrt{s^*}.$$

Thus, we have

$$\left|\mathcal{E}^{(q)}\right| = \left|\mathcal{S}^* \cup \left(\mathcal{S}^{(q)} \backslash \mathcal{S}^*\right)\right| = \left|\mathcal{S}^*\right| + \left|\mathcal{S}^{(q)} \backslash \mathcal{S}^*\right| \leq 2s^*,$$

completing the induction.

Then by the definition of $\mathcal{E}^{(q)}$ and $\mathcal{S}^{(q)}$, we have

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

Applying Lemma 20 with $\widetilde{\mathbf{\Sigma}} = \widetilde{\mathbf{\Sigma}}^{(q)}$, $\mathcal{E} = \mathcal{E}^{(q)}$, and $\mathbf{\Lambda}_{\mathcal{S}^*} = \mathbf{\Lambda}^{(q-1)}_{\mathcal{S}^*}$, the ε-optimal solution $\widetilde{\mathbf{\Sigma}}^{(q)}$ satisfies

$$\begin{aligned}
& \left\| \widetilde{\mathbf{\Sigma}}^{(q)} - \mathbf{\Sigma}^* \right\|_{\mathsf{F}} \\
\leq & \frac{1}{\rho^-_{2s^*+2\widetilde{s}}} \left(\left\| \left(\nabla f(\mathbf{\Sigma}^*) \right)_{\mathcal{E}^{(q)}} \right\|_{\mathsf{F}} + \varepsilon \sqrt{\left|\mathcal{E}^{(q)}\right|} + \left\| \mathbf{\Lambda}^{(q-1)}_{\mathcal{S}^*} \right\|_{\mathsf{F}} \right) \\
\leq & \frac{2+\sqrt{2}}{2\rho^-_{2s^*+2\widetilde{s}}} \lambda \sqrt{s^*}.
\end{aligned}$$

∎

B. Proof of Theorem 11

Proof: Based on Lemma 21, we have

$$\begin{aligned}
\left\| \widetilde{\mathbf{\Sigma}}^{(q)} - \mathbf{\Sigma}^* \right\|_{\mathsf{F}} \leq & \frac{1}{\rho^-_{2s^*+2\widetilde{s}}} \underbrace{\left(\left\| \left(\nabla f(\mathbf{\Sigma}^*) \right)_{\mathcal{E}^{(q)}} \right\|_{\mathsf{F}} + \varepsilon \sqrt{\left|\mathcal{E}^{(q)}\right|} \right)}_{\mathbf{V}} \\
& + \frac{1}{\rho^-_{2s^*+2\widetilde{s}}} \underbrace{\left\| \mathbf{\Lambda}^{(q-1)}_{\mathcal{S}^*} \right\|_{\mathsf{F}}}_{\mathbf{VI}}.
\end{aligned} \tag{12}$$

Then, we proceed to establish bounds for the term **V** and **VI**, respectively.

For term **V**, dividing the support set into $\mathcal{S}^*$ and $\mathcal{E}^{(q)} \backslash \mathcal{S}^*$, we obtain

$$\begin{aligned}
\mathbf{V} &\leq \left\| \left(\nabla f(\mathbf{\Sigma}^*) \right)_{\mathcal{S}^*} \right\|_{\mathsf{F}} + \varepsilon \sqrt{s^*} \\
&\quad + \left(\left\| \nabla f(\mathbf{\Sigma}^*) \right\|_{\max} + \varepsilon \right) \sqrt{\left|\mathcal{E}^{(q)} \backslash \mathcal{S}^*\right|} \\
&\leq \left\| \left(\nabla f(\mathbf{\Sigma}^*) \right)_{\mathcal{S}^*} \right\|_{\mathsf{F}} + \varepsilon \sqrt{s^*} + \frac{\lambda}{2u} \left\| \widetilde{\mathbf{\Sigma}}^{(q-1)} - \mathbf{\Sigma}^* \right\|_{\mathsf{F}},
\end{aligned} \tag{13}$$

where the second inequality is due to

$$\sqrt{\left|\mathcal{E}^{(q)} \backslash \mathcal{S}^*\right|} = \sqrt{\left|\mathcal{S}^{(q)} \backslash \mathcal{S}^*\right|} \leq u^{-1} \left\| \widetilde{\mathbf{\Sigma}}^{(q-1)} - \mathbf{\Sigma}^* \right\|_{\mathsf{F}}, \tag{14}$$

which follows from the inequality (11).

By Assumption 2 and 8, for any $\mathbf{\Sigma}$, if $\left|\Sigma_{jk} - \Sigma^*_{jk}\right| \geq u$, then $p'_\lambda(\Sigma_{jk}) \leq \lambda \leq \lambda u^{-1} \left|\Sigma_{jk} - \Sigma^*_{jk}\right|$; otherwise,

$p'_\lambda(\Sigma_{jk}) \le p'_\lambda\left(\left|\Sigma^*_{jk}\right| - u\right) = 0$. Therefore, for term **VI**, we have

$$\mathbf{VI} \le \lambda u^{-1}\left\|\widetilde{\mathbf{\Sigma}}^{(q-1)}_{\mathcal{S}^*} - \mathbf{\Sigma}^*_{\mathcal{S}^*}\right\|_{\mathsf{F}} \le \lambda u^{-1}\left\|\widetilde{\mathbf{\Sigma}}^{(q-1)} - \mathbf{\Sigma}^*\right\|_{\mathsf{F}}. \tag{15}$$

Substituting the above results into (12) yields

$$\begin{aligned}&\left\|\widetilde{\mathbf{\Sigma}}^{(q)} - \mathbf{\Sigma}^*\right\|_{\mathsf{F}}\\ &\le \frac{1}{\rho^-_{s^*+2\widetilde{s}}}\left(\left\|\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}} + \varepsilon\sqrt{s^*}\right) + \delta\left\|\widetilde{\mathbf{\Sigma}}^{(q-1)} - \mathbf{\Sigma}^*\right\|_{\mathsf{F}},\end{aligned} \tag{16}$$

where $\delta = \frac{3\lambda}{2u\rho^-_{s^*+2\widetilde{s}}} = \frac{3}{2+\sqrt{2}} \in (0,1)$. ∎

C. Concentration Inequality

Lemma 22. *Under Assumptions 8, there exists some constant c_1, the following result holds*

$$\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^m \boldsymbol{a}_i\left(y_i - \boldsymbol{a}_i^\top\mathbf{\Sigma}^*\boldsymbol{a}_i\right)\boldsymbol{a}_i^\top\right\|_{\max} \le \lambda\right) \ge 1 - \frac{c_1}{d}.$$

Proof: We start by bounding

$$\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^m \boldsymbol{a}_i\left(y_i - \boldsymbol{a}_i^\top\mathbf{\Sigma}^*\boldsymbol{a}_i\right)\boldsymbol{a}_i^\top\right\|_{\max} > t\right).$$

First of all, using the union bound, we obtain

$$\begin{aligned}&\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^m \boldsymbol{a}_i\left(y_i - \boldsymbol{a}_i^\top\mathbf{\Sigma}^*\boldsymbol{a}_i\right)\boldsymbol{a}_i^\top\right\|_{\max} > t\right)\\ =&\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^m \eta_i\boldsymbol{a}_i\boldsymbol{a}_i^\top\right\|_{\max} > t\right)\\ \le&\sum_{l=1}^{d^2}\mathbb{P}\left(\frac{1}{m}\left|\widetilde{\boldsymbol{A}}_{\cdot l}^\top\cdot\boldsymbol{\eta}\right| > t\right),\end{aligned} \tag{17}$$

where $\boldsymbol{\eta} = [\eta_1,\ldots,\eta_m]^\top$. Let $\varrho_l = \widetilde{\boldsymbol{A}}_{\cdot l}^\top\cdot\boldsymbol{\eta}$. Since η_i is sub-Gaussian $\left(0,\sigma^2\right)$ for $i = 1,\ldots,m$, we obtain

$$\begin{aligned}&\mathbb{E}\left(\exp\left\{t_0\varrho_l\right\} + \exp\left\{-t_0\varrho_l\right\}\right)\\ \le&2\exp\left\{m^{-2}\left\|\widetilde{\boldsymbol{A}}_{\cdot l}\right\|^2\sigma_\eta^2 t_0^2/2\right\},\end{aligned} \tag{18}$$

which implies

$$\mathbb{P}\left(|\varrho_l| \ge t\right)\exp\left\{t_0 t\right\} \le 2\exp\left\{m^{-2}\left\|\widetilde{\boldsymbol{A}}_{\cdot l}\right\|^2\sigma_\eta^2 t_0^2/2\right\}.$$

Taking $t_0 = t\left(m^{-2}\left\|\widetilde{\boldsymbol{A}}_{\cdot l}\right\|_2^2\sigma_\eta^2\right)^{-1}$ yields that

$$\mathbb{P}\left(|\varrho_l| \ge t\right) \le 2\exp\left\{-\frac{t^2}{2\sigma_\eta^2\left\|\widetilde{\boldsymbol{A}}_{\cdot l}\right\|_2^2/m^2}\right\}.$$

Plugging it into (17) results

$$\begin{aligned}&\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^m \boldsymbol{a}_i\left(y_i - \boldsymbol{a}_i^\top\mathbf{\Sigma}^*\boldsymbol{a}_i\right)\boldsymbol{a}_i^\top\right\|_{\max} > t\right)\\ \underset{(i)}{\le}&2d^2\exp\left\{-\frac{mt^2}{2\sigma_\eta^2\max_l\left\{\left\|\widetilde{\boldsymbol{A}}_{\cdot l}\right\|_2^2/m\right\}}\right\}\\ =&c_1\exp\left\{-c_2mt^2 + 2\log d\right\},\end{aligned}$$

where (i) is from Lemma 15 that the column of $\widetilde{\boldsymbol{A}}$ are normalized such that $\max_l\left\|\widetilde{\boldsymbol{A}}_{\cdot l}\right\|_2 \le \sqrt{m}$, and c_1, c_2 are constants. Then taking $\lambda = \sqrt{\frac{3\log d}{c_3 m}} \asymp \sqrt{\frac{\log d}{m}}$, we obtain

$$\begin{aligned}&\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^m \boldsymbol{a}_i\left(y_i - \boldsymbol{a}_i^\top\mathbf{\Sigma}^*\boldsymbol{a}_i\right)\boldsymbol{a}_i^\top\right\|_{\max} \le \lambda\right)\\ \ge&1 - c_1\exp\left(-c_2m\lambda^2 + 2\log d\right)\\ =&1 - \frac{c_1}{d}.\end{aligned}$$

∎

Lemma 23. *Under the same conditions in Lemma 22, the following result hold*

$$\left\|\left(\frac{1}{m}\sum_{i=1}^m \boldsymbol{a}_i\left(y_i - \boldsymbol{a}_i^\top\mathbf{\Sigma}^*\boldsymbol{a}_i\right)\boldsymbol{a}_i^\top\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}} = O_p\left(\sqrt{\frac{s^*}{m}}\right).$$

Proof: For simplicity, we define

$$\nabla f_1\left(\mathbf{\Sigma}^*\right) = -\frac{1}{m}\sum_{i=1}^m \boldsymbol{a}_i\left(y_i - \boldsymbol{a}_i^\top\mathbf{\Sigma}^*\boldsymbol{a}_i\right)\boldsymbol{a}_i^\top.$$

Similar to Lemma 22 analysis, for any M such that $0 < M\sqrt{\frac{1}{m}}$, we obtain

$$\begin{aligned}&\mathbb{P}\left(\left\|\left(-\nabla f_1\left(\mathbf{\Sigma}^*\right)\right)_{\mathcal{S}^*}\right\|_{\max} > M\sqrt{\frac{1}{m}}\right)\\ \le&c_1 s^*\exp\left(-c_2M\right)\\ =&c_1\exp\left(-c_2M + \log s^*\right).\end{aligned}$$

By taking M such that $\sqrt{\frac{2\log s^*}{c_3}} < M$ and $M \to \infty$ in the above inequality obtains

$$\lim_{M\to\infty}\sup_m\mathbb{P}\left(\left\|\left(-\nabla f_1\left(\mathbf{\Sigma}^*\right)\right)_{\mathcal{S}^*}\right\|_{\max} > M\sqrt{\frac{1}{m}}\right) = 0.$$

The proof is completed by applying

$$\begin{aligned}&\left\|\left(\frac{1}{m}\sum_{i=1}^m \boldsymbol{a}_i\left(y_i - \boldsymbol{a}_i^\top\mathbf{\Sigma}^*\boldsymbol{a}_i\right)\boldsymbol{a}_i^\top\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}}\\ \le&\sqrt{s^*}\left\|\left(\frac{1}{m}\sum_{i=1}^m \boldsymbol{a}_i\left(y_i - \boldsymbol{a}_i^\top\mathbf{\Sigma}^*\boldsymbol{a}_i\right)\boldsymbol{a}_i^\top\right)_{\mathcal{S}^*}\right\|_{\max}.\end{aligned}$$

∎

D. Proof of Corollary 13

Proof: One has

$$\begin{aligned}\left\|\nabla f\left(\boldsymbol{\Sigma}^*\right)\right\|_{\max} \leq & \left\|\frac{1}{m}\sum_{i=1}^m \boldsymbol{a}_i\left(y_i-\boldsymbol{a}_i^\top\boldsymbol{\Sigma}^*\boldsymbol{a}_i\right)\boldsymbol{a}_i^\top\right\|_{\max} \\ & +\tau\left\|\left(\boldsymbol{\Sigma}^*\right)^{-1}\right\|_{\max}.\end{aligned} \tag{19}$$

If τ satisfy $\tau \lesssim \sqrt{\frac{1}{m}}\left\|\left(\boldsymbol{\Sigma}^*\right)^{-1}\right\|_{\max}^{-1}$, then by Lemma 22, $\lambda \geq 2\left(\left\|\nabla f\left(\boldsymbol{\Sigma}^*\right)\right\|_{\mathsf{max}}+\varepsilon\right)$ holds w.h.p.

Applying Lemma 21 with $q=1$, we obtain

$$\left\|\widetilde{\boldsymbol{\Sigma}}^{(1)}-\boldsymbol{\Sigma}^*\right\|_{\mathsf{F}} \leq \frac{2+\sqrt{2}}{2\rho^-_{2s^*+2\widetilde{s}}}\lambda\sqrt{s^*}. \tag{20}$$

If Assumption 8 holds, that is, $\lambda \asymp \sqrt{\frac{\log d}{m}}$, then $\left\|\widetilde{\boldsymbol{\Sigma}}^{(1)}-\boldsymbol{\Sigma}^*\right\|_{\mathsf{F}} \lesssim \sqrt{\frac{s^*\log d}{m}}$ w.h.p. ∎

E. Proof of Corollary 14

Proof: One has

$$\begin{aligned}\left\|\nabla f\left(\boldsymbol{\Sigma}^*\right)\right\|_{\mathsf{max}} \leq & \left\|\sum_{i=1}^m \boldsymbol{a}_i\left(y_i-\boldsymbol{a}_i^\top\boldsymbol{\Sigma}^*\boldsymbol{a}_i\right)\boldsymbol{a}_i^\top\right\|_{\mathsf{max}} \\ & +\tau\left\|\left(\boldsymbol{\Sigma}^*\right)^{-1}\right\|_{\mathsf{max}}.\end{aligned}$$

If Assumptions 8 and 10 hold, and τ satisfy

$$\tau \lesssim \sqrt{\frac{1}{m}}\left\|\left(\boldsymbol{\Sigma}^*\right)^{-1}\right\|_{\max}^{-1},$$

then by Lemma 22, $\lambda \geq 2\left(\left\|\nabla f\left(\boldsymbol{\Sigma}^*\right)\right\|_{\mathsf{max}}+\varepsilon\right)$ holds w.h.p.

Applying Theorem 11, we obtain

$$\begin{aligned}&\left\|\widetilde{\boldsymbol{\Sigma}}^{(q)}-\boldsymbol{\Sigma}^*\right\|_{\mathsf{F}} \\ &\leq \frac{1}{\rho^-_{s^*+2\widetilde{s}}}\left(\left\|\left(\nabla f\left(\boldsymbol{\Sigma}^*\right)\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}}+\varepsilon\sqrt{s^*}\right)+\delta\left\|\widetilde{\boldsymbol{\Sigma}}^{(q-1)}-\boldsymbol{\Sigma}^*\right\|_{\mathsf{F}} \\ &\leq \frac{1}{\rho^-_{s^*+2\widetilde{s}}}\left(\left\|\left(\nabla f\left(\boldsymbol{\Sigma}^*\right)\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}}+\varepsilon\sqrt{s^*}\right)+\delta^{k-1}\left\|\widetilde{\boldsymbol{\Sigma}}^{(1)}-\boldsymbol{\Sigma}^*\right\|_{\mathsf{F}} \\ &\leq \frac{1}{\rho^-_{s^*+2\widetilde{s}}}\left(\left\|\left(\nabla f\left(\boldsymbol{\Sigma}^*\right)\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}}+\varepsilon\sqrt{s^*}\right)+\delta^{k-1}\frac{2+\sqrt{2}}{2\rho^-_{s^*+2\widetilde{s}}}\lambda\sqrt{s^*},\end{aligned} \tag{21}$$

where the last inequality is due to $\left\|\widetilde{\boldsymbol{\Sigma}}^{(1)}-\boldsymbol{\Sigma}^*\right\|_{\mathsf{F}} \leq \frac{2+\sqrt{2}}{2\rho^-_{s^*+2\widetilde{s}}}\lambda\sqrt{s^*}$, which follows from Lemma 21 with $q=1$.

One has

$$\begin{aligned}&\left\|\left(\nabla f\left(\boldsymbol{\Sigma}^*\right)\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}} \\ =&\left\|\left(-\sum_{i=1}^m \boldsymbol{a}_i\left(y_i-\boldsymbol{a}_i^\top\boldsymbol{\Sigma}^*\boldsymbol{a}_i\right)\boldsymbol{a}_i^\top-\tau\left(\boldsymbol{\Sigma}^*\right)^{-1}\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}} \\ \leq&\left\|\left(\sum_{i=1}^m \boldsymbol{a}_i\left(y_i-\boldsymbol{a}_i^\top\boldsymbol{\Sigma}^*\boldsymbol{a}_i\right)\boldsymbol{a}_i^\top\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}}+\tau\left\|\left(\left(\boldsymbol{\Sigma}^*\right)^{-1}\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}} \\ \leq&\left\|\left(\sum_{i=1}^m \boldsymbol{a}_i\left(y_i-\boldsymbol{a}_i^\top\boldsymbol{\Sigma}^*\boldsymbol{a}_i\right)\boldsymbol{a}_i^\top\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}}+\tau\left\|\left(\boldsymbol{\Sigma}^*\right)^{-1}\right\|_{\mathsf{F}}\end{aligned} \tag{22}$$

By Lemma 23, $\left\|\left(\sum_{i=1}^m \boldsymbol{a}_i\left(y_i-\boldsymbol{a}_i^\top\boldsymbol{\Sigma}^*\boldsymbol{a}_i\right)\boldsymbol{a}_i^\top\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}} \lesssim \sqrt{\frac{s^*}{m}}$ holds w.h.p. ∎

APPENDIX B
FURTHER INTERMEDIATE RESULTS

Lemma 24. *Given* $\omega_{\boldsymbol{\Lambda}^{(q-1)}}\left(\widetilde{\boldsymbol{\Sigma}}^{(q)}\right) \leq \frac{\lambda}{8}$*, we have that for all* $t \geq 1$ *at the* $(q+1)$*-th stage,*

$$\omega_{\boldsymbol{\Lambda}^{(q)}}\left(\boldsymbol{\Sigma}_t\right) \leq \frac{\lambda}{4}$$

and

$$F_{\boldsymbol{\Lambda}^{(q)}}\left(\boldsymbol{\Sigma}_t\right) \leq F_{\boldsymbol{\Lambda}^{(q)}}\left(\boldsymbol{\Sigma}^*\right)+\frac{\lambda}{4}\left\|\boldsymbol{\Sigma}_t-\boldsymbol{\Sigma}^*\right\|_1.$$

Proof: Note that at the $(q+1)$-th stage, $\boldsymbol{\Sigma}_0=\widetilde{\boldsymbol{\Sigma}}^{(q)}$. Then we have

$$\begin{aligned}\omega_{\boldsymbol{\Lambda}^{(q)}}\left(\boldsymbol{\Sigma}_t\right) =& \min_{\boldsymbol{\Xi}\in\partial\left\|\boldsymbol{\Sigma}_0\right\|_1}\left\|\nabla f\left(\boldsymbol{\Sigma}_0\right)+\boldsymbol{\Lambda}^{(q)}\odot\boldsymbol{\Xi}\right\|_{\max} \\ \underset{(i)}{\leq}& \min_{\boldsymbol{\Xi}\in\partial\left\|\boldsymbol{\Sigma}_0\right\|_1}\left\|\nabla f\left(\boldsymbol{\Sigma}_0\right)+\boldsymbol{\Lambda}^{(q-1)}\odot\boldsymbol{\Xi}\right\|_{\max} \\ &+\left\|\left(\boldsymbol{\Lambda}^{(q)}-\boldsymbol{\Lambda}^{(q-1)}\right)\odot\boldsymbol{\Xi}\right\|_{\max} \\ \underset{(ii)}{\leq}& \omega_{\boldsymbol{\Lambda}^{(q-1)}}\left(\boldsymbol{\Sigma}_0\right)+\left\|\boldsymbol{\Lambda}^{(q)}-\boldsymbol{\Lambda}^{(q-1)}\right\|_{\max} \\ \underset{(iii)}{\leq}& \frac{\lambda}{8}+\frac{\lambda}{8} \leq \frac{\lambda}{4},\end{aligned}$$

where (i) is from triangle inequality, (ii) is from the definition of the approximate KKT condition and $\boldsymbol{\Xi}$, and (iii) is from $\omega_{\boldsymbol{\Lambda}^{(q-1)}}\left(\boldsymbol{\Sigma}_0\right) = \omega_{\boldsymbol{\Lambda}^{(q-1)}}\left(\widetilde{\boldsymbol{\Sigma}}^{(q)}\right) \leq \frac{\lambda}{8}$ and $\left\|\boldsymbol{\Lambda}^{(q)}-\boldsymbol{\Lambda}^{(q-1)}\right\|_{\max} \leq \frac{\lambda}{8}$.

For some $\boldsymbol{\Xi}_t = \underset{\boldsymbol{\Xi}\in\partial\left\|\boldsymbol{\Sigma}_t\right\|_1}{\arg\min}\left\|\nabla f\left(\boldsymbol{\Sigma}_t\right)+\boldsymbol{\Lambda}^{(q)}\odot\boldsymbol{\Xi}\right\|_{\max}$, we have

$$\begin{aligned}&F_{\boldsymbol{\Lambda}^{(q)}}\left(\boldsymbol{\Sigma}^*\right) \\ \underset{(i)}{\geq}& F_{\boldsymbol{\Lambda}^{(q)}}\left(\boldsymbol{\Sigma}_t\right)-\left\langle\nabla f\left(\boldsymbol{\Sigma}_t\right)+\boldsymbol{\Lambda}^{(q)}\odot\boldsymbol{\Xi}_t,\boldsymbol{\Sigma}_t-\boldsymbol{\Sigma}^*\right\rangle \\ \geq& F_{\boldsymbol{\Lambda}^{(q)}}\left(\boldsymbol{\Sigma}_t\right)- \\ &\left\|\nabla f\left(\boldsymbol{\Sigma}_t\right)+\boldsymbol{\Lambda}^{(q)}\odot\boldsymbol{\Xi}_t\right\|_{\max}\left\|\boldsymbol{\Sigma}_t-\boldsymbol{\Sigma}^*\right\|_1 \\ \underset{(ii)}{\geq}& F_{\boldsymbol{\Lambda}^{(q)}}\left(\boldsymbol{\Sigma}_t\right)-\frac{\lambda}{4}\left\|\boldsymbol{\Sigma}_t-\boldsymbol{\Sigma}^*\right\|_1,\end{aligned}$$

where (i) is from the convexity of F and (ii) is from the fact that for all $t \geq 0$, $\left\|\nabla f\left(\mathbf{\Sigma}_t\right) + \mathbf{\Lambda}^{(q)} \odot \mathbf{\Xi}_t\right\|_{\max} \leq \frac{\lambda}{4}$. ■

Lemma 25. *Suppose* $\left\|\mathbf{\Sigma}^t_{\overline{\mathcal{S}}}\right\|_0 \leq \widetilde{s}$ *and* $\omega_{\mathbf{\Lambda}^{(q)}}\left(\mathbf{\Sigma}_t\right) \leq \frac{\lambda}{4}$. *Then there exists a generic constant* c_1 *such that*

$$\left\|\mathbf{\Sigma}_t - \mathbf{\Sigma}^*\right\|_{\mathsf{F}} \leq \frac{c_1 \lambda \sqrt{s^*}}{\rho^-_{2s^*+2\widetilde{s}}}.$$

Proof: It is similar to the proof of Lemma 21, we omit it for simplicity. For more details, we refer readers to [43]. ■

Appendix C
Preliminary Lemmas

Lemma 26. *Consider a set* $\mathcal{E}$ *such that* $\mathcal{S}^* \subseteq \mathcal{E}$, *if* $\left\|\nabla f\left(\mathbf{\Sigma}^*\right)\right\|_{\max} + \varepsilon \leq \left\|\mathbf{\Lambda}_{\mathcal{E}}\right\|_{\min}$, *we have*

$$\left\|\left(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right)_{\overline{\mathcal{E}}}\right\|_1 \leq 5\left\|\left(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right)_{\mathcal{E}}\right\|_1.$$

Proof: Define $\mathbf{\Omega} = \nabla f\left(\widetilde{\mathbf{\Sigma}}\right) + \mathbf{\Lambda} \odot \widetilde{\mathbf{\Xi}}$ and $\widetilde{\mathbf{\Delta}} = \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*$. By the mean value theorem, there exist a $\theta \in [0,1]$, such that

$$\nabla f\left(\widetilde{\mathbf{\Sigma}}\right) - \nabla f\left(\mathbf{\Sigma}^*\right) = \mathsf{mat}\left(H\left(\theta\right)\mathsf{vec}\left(\widetilde{\mathbf{\Delta}}\right)\right),$$

where $H\left(\theta\right) = \nabla^2 f\left(\theta\mathbf{\Sigma}^* + \left(1-\theta\right)\widetilde{\mathbf{\Sigma}}\right)$. Then we have

$$\begin{aligned}
&\left\langle \nabla f\left(\widetilde{\mathbf{\Sigma}}\right) + \mathbf{\Lambda} \odot \widetilde{\mathbf{\Xi}}, \widetilde{\mathbf{\Delta}}\right\rangle \\
=&\left\langle \nabla f\left(\mathbf{\Sigma}^*\right) + \mathsf{mat}\left(H\left(\theta\right)\mathsf{vec}\left(\widetilde{\mathbf{\Delta}}\right)\right) + \mathbf{\Lambda} \odot \widetilde{\mathbf{\Xi}}, \widetilde{\mathbf{\Delta}}\right\rangle \\
\leq& \left\|\mathbf{\Omega}\right\|_{\max}\left\|\widetilde{\mathbf{\Delta}}\right\|_1.
\end{aligned}$$

Using the fact $\left\|\widetilde{\mathbf{\Delta}}\right\|^2_{H(\theta)} \geq 0$, we have

$$0 \leq \left\|\mathbf{\Omega}\right\|_{\max}\left\|\widetilde{\mathbf{\Delta}}\right\|_1 - \underbrace{\left\langle \nabla f\left(\mathbf{\Sigma}^*\right), \widetilde{\mathbf{\Delta}}\right\rangle}_{\mathbf{I}} - \underbrace{\left\langle \mathbf{\Lambda} \odot \widetilde{\mathbf{\Xi}}, \widetilde{\mathbf{\Delta}}\right\rangle}_{\mathbf{II}}. \quad (23)$$

For term $\mathbf{I}$, separating the support of $\nabla f\left(\mathbf{\Sigma}^*\right)$ and $\widetilde{\mathbf{\Delta}}$ to $\mathcal{E}$ and $\overline{\mathcal{E}}$, and then using the matrix Hölder's inequality, we obtain

$$\begin{aligned}
\mathbf{I} =& \left\langle \left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)_{\mathcal{E}}, \widetilde{\mathbf{\Delta}}_{\mathcal{E}}\right\rangle + \left\langle \left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)_{\overline{\mathcal{E}}}, \widetilde{\mathbf{\Delta}}_{\overline{\mathcal{E}}}\right\rangle \\
\geq& -\left\|\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)_{\mathcal{E}}\right\|_{\mathsf{F}}\left\|\widetilde{\mathbf{\Delta}}_{\mathcal{E}}\right\|_{\mathsf{F}} - \left\|\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)_{\overline{\mathcal{E}}}\right\|_{\max}\left\|\widetilde{\mathbf{\Delta}}_{\overline{\mathcal{E}}}\right\|_1 \\
\geq& -\left\|\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)_{\mathcal{E}}\right\|_{\mathsf{F}}\left\|\widetilde{\mathbf{\Delta}}\right\|_{\mathsf{F}} - \left\|\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)\right\|_{\max}\left\|\widetilde{\mathbf{\Delta}}\right\|_1.
\end{aligned} \quad (24)$$

For term $\mathbf{II}$, separating the support of $\mathbf{\Lambda} \odot \widetilde{\mathbf{\Xi}}$ and $\widetilde{\mathbf{\Delta}}$ to $\mathcal{E}$ and $\overline{\mathcal{E}}$, and then using the matrix Hölder's inequality, we obtain

$$\begin{aligned}
\mathbf{II} =& \left\langle \left(\mathbf{\Lambda} \odot \widetilde{\mathbf{\Xi}}\right)_{\mathcal{E}}, \widetilde{\mathbf{\Delta}}_{\mathcal{E}}\right\rangle + \left\langle \left(\mathbf{\Lambda} \odot \widetilde{\mathbf{\Xi}}\right)_{\overline{\mathcal{E}}}, \widetilde{\mathbf{\Delta}}_{\overline{\mathcal{E}}}\right\rangle \\
\geq& -\left\|\mathbf{\Lambda}_{\mathcal{E}}\right\|_{\max}\left\|\widetilde{\mathbf{\Delta}}_{\mathcal{E}}\right\|_1 + \left\|\mathbf{\Lambda}_{\overline{\mathcal{E}}}\right\|_{\min}\left\|\widetilde{\mathbf{\Delta}}_{\overline{\mathcal{E}}}\right\|_1.
\end{aligned} \quad (25)$$

Plugging the above results into 23 yields

$$\begin{aligned}
\left\|\widetilde{\mathbf{\Delta}}_{\overline{\mathcal{E}}}\right\|_1 \leq& \frac{\lambda + \left\|\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)\right\|_{\max} + \omega_{\mathbf{\Lambda}}\left(\widetilde{\mathbf{\Sigma}}\right)}{\left\|\mathbf{\Lambda}_{\overline{\mathcal{E}}}\right\|_{\min} - \left(\left\|\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)\right\|_{\max} + \omega_{\mathbf{\Lambda}}\left(\widetilde{\mathbf{\Sigma}}\right)\right)}\left\|\widetilde{\mathbf{\Delta}}_{\mathcal{E}}\right\|_1 \\
\underset{(i)}{\leq}& 5\left\|\widetilde{\mathbf{\Delta}}_{\mathcal{E}}\right\|_1,
\end{aligned}$$

where (i) is from $\left\|\mathbf{\Lambda}_{\overline{\mathcal{E}}}\right\|_{\min} \geq \frac{\lambda}{2}$, $\left\|\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)\right\|_{\max} \leq \frac{\lambda}{4}$ and $\omega_{\mathbf{\Lambda}}\left(\widetilde{\mathbf{\Sigma}}\right) \leq \varepsilon$. ■

Lemma 27. *Recall* $D_f\left(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2\right)$ *and* $D^s_f\left(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2\right)$ *defined in Lemma 19. For For* $\mathbf{\Sigma}\left(\theta\right) = \mathbf{\Sigma}^* + \theta\left(\mathbf{\Sigma} - \mathbf{\Sigma}^*\right)$ *with* $\theta \in (0,1]$, *we have*

$$D^s_f\left(\mathbf{\Sigma}\left(\theta\right), \mathbf{\Sigma}^*\right) \leq \theta D^s_f\left(\mathbf{\Sigma}, \mathbf{\Sigma}^*\right).$$

Proof: Let

$$\begin{aligned}
\varphi\left(\theta\right) =& D_f\left(\mathbf{\Sigma}\left(\theta\right), \mathbf{\Sigma}^*\right) \\
=& f\left(\mathbf{\Sigma}\left(\theta\right)\right) - f\left(\mathbf{\Sigma}^*\right) - \left\langle \nabla f\left(\mathbf{\Sigma}^*\right), \mathbf{\Sigma}\left(\theta\right) - \mathbf{\Sigma}^*\right\rangle.
\end{aligned}$$

Since the derivative of $f\left(\mathbf{\Sigma}\left(\theta\right)\right)$ with respect to θ is $\left\langle \nabla f\left(\mathbf{\Sigma}\left(\theta\right)\right), \mathbf{\Sigma} - \mathbf{\Sigma}^*\right\rangle$, it follows that

$$\varphi'\left(\theta\right) = \left\langle \nabla f\left(\mathbf{\Sigma}\left(\theta\right)\right) - \nabla f\left(\mathbf{\Sigma}^*\right), \mathbf{\Sigma} - \mathbf{\Sigma}^*\right\rangle.$$

Therefore, the symmetric Bregman divergence $D^s_f\left(\mathbf{\Sigma}\left(\theta\right), \mathbf{\Sigma}^*\right)$ can be written as

$$\begin{aligned}
&D^s_f\left(\mathbf{\Sigma}\left(\theta\right), \mathbf{\Sigma}^*\right) \\
=&\left\langle \nabla f\left(\mathbf{\Sigma}\left(\theta\right)\right) - \nabla f\left(\mathbf{\Sigma}^*\right), \theta\left(\mathbf{\Sigma} - \mathbf{\Sigma}^*\right)\right\rangle \\
=&\theta\varphi'\left(\theta\right)
\end{aligned}$$

for $0 < \theta \leq 1$. Plugging $\theta = 1$ in the equation above, we have $\varphi'\left(1\right) = D^s_f\left(\mathbf{\Sigma}, \mathbf{\Sigma}^*\right)$ as a special case. If we assume that $\varphi\left(\theta\right)$ is convex , then $\varphi'\left(\theta\right)$ is non-decreasing and thus

$$D^s_f\left(\mathbf{\Sigma}\left(\theta\right), \mathbf{\Sigma}^*\right) = \theta\varphi'\left(\theta\right) \leq \theta\varphi'\left(1\right) = \theta D^s_f\left(\mathbf{\Sigma}, \mathbf{\Sigma}^*\right).$$

It remains to show the convexity of $\varphi\left(\theta\right)$ with respect to θ, i.e.

$$\mathbf{\Sigma}\left(\alpha_1\theta_1 + \alpha_2\theta_2\right) \leq \alpha_1\mathbf{\Sigma}\left(\theta_1\right) + \alpha_2\mathbf{\Sigma}\left(\theta_2\right),$$

for $\theta_1, \theta_2 \in (0,1]$, $\alpha_1, \alpha_2 \geq 0$ and $\alpha_1 + \alpha_2 = 1$; or equivalently, the convexity of $f\left(\mathbf{\Sigma}\left(\theta\right)\right)$ and $\left\langle \nabla f\left(\mathbf{\Sigma}^*\right), \mathbf{\Sigma}^* - \mathbf{\Sigma}\left(\theta\right)\right\rangle$, respectively.

For $\forall \alpha_1, \alpha_2 \geq 0$ such that $\alpha_1 + \alpha_2 = 1$, and $\theta_1, \theta_2 \in (0,1)$, we have $\mathbf{\Sigma}\left(\alpha_1\theta_1 + \alpha_2\theta_2\right) = \alpha_1\mathbf{\Sigma}\left(\theta_1\right) + \alpha_2\mathbf{\Sigma}\left(\theta_2\right)$. By the bi-linearity property of the inner product function $\left\langle \cdot, \cdot\right\rangle$, and using the linearity property of $\mathbf{\Sigma}\left(\cdot\right)$, we have the following equality hold

$$\begin{aligned}
&-\left\langle \nabla f\left(\mathbf{\Sigma}^*\right), \mathbf{\Sigma}\left(\alpha_1\theta_1 + \alpha_2\theta_2\right) - \mathbf{\Sigma}^*\right\rangle \\
=&-\alpha_1\left\langle \nabla f\left(\mathbf{\Sigma}^*\right), \mathbf{\Sigma}\left(\theta_1\right) - \mathbf{\Sigma}^*\right\rangle - \alpha_2\left\langle \nabla f\left(\mathbf{\Sigma}^*\right), \mathbf{\Sigma}\left(\theta_2\right) - \mathbf{\Sigma}^*\right\rangle.
\end{aligned} \quad (26)$$

On the other side, by the convexity of the loss function $f\left(\cdot\right)$, we obtain

$$\begin{aligned}
&f\left(\mathbf{\Sigma}\left(\alpha_1\theta_1 + \alpha_2\theta_2\right)\right) \\
=&f\left(\alpha_1\mathbf{\Sigma}\left(\theta_1\right) + \alpha_2\mathbf{\Sigma}\left(\theta_2\right)\right) \\
\leq&\alpha_1 f\left(\mathbf{\Sigma}\left(\theta_1\right)\right) + \alpha_2 f\left(\mathbf{\Sigma}\left(\theta_2\right)\right).
\end{aligned} \quad (27)$$

By adding 26 and 27 together and using the definition of function $\mathbf{\Sigma}\left(\cdot\right)$, we obtain

$$\mathbf{\Sigma}\left(\alpha_1\theta_1 + \alpha_2\theta_2\right) \leq \alpha_1\mathbf{\Sigma}\left(\theta_1\right) + \alpha_2\mathbf{\Sigma}\left(\theta_2\right),$$

which indicates $\mathbf{\Sigma}\left(\theta\right)$ is a convex function. Thus we complete our proof. ■

Lemma 28. *Under the same condition of Lemma 21, we have the following basic inequality*

$$\left\langle \nabla f\left(\widetilde{\mathbf{\Sigma}}^{(1)}\right)-\nabla f\left(\mathbf{\Sigma}^{*}\right),\widetilde{\mathbf{\Sigma}}^{(1)}-\mathbf{\Sigma}^{*}\right\rangle \leq \frac{c_1\lambda^2 s^*}{\rho^{-}_{2s^*+2\widetilde{s}}}.$$

Proof: Applying Lemma 21 with $q=1$, we obtain

$$\left\|\widetilde{\mathbf{\Sigma}}^{(1)}-\mathbf{\Sigma}^{*}\right\|_{\mathsf{F}} \leq \frac{2+\sqrt{2}}{2\rho^{-}_{2s^*+2\widetilde{s}}}\lambda\sqrt{s^*}. \tag{28}$$

On the other side, applying Lemma 26 yields that

$$\begin{aligned}\left\|\widetilde{\mathbf{\Sigma}}^{(1)}-\mathbf{\Sigma}^{*}\right\|_1 &\leq \left\|\left(\widetilde{\mathbf{\Sigma}}^{(1)}-\mathbf{\Sigma}^{*}\right)_{\overline{\mathcal{E}^{\{1\}}}}\right\|_1+\left\|\left(\widetilde{\mathbf{\Sigma}}^{(1)}-\mathbf{\Sigma}^{*}\right)_{\mathcal{E}^{(1)}}\right\|_1\\ &\leq 6\left\|\left(\widetilde{\mathbf{\Sigma}}^{(1)}-\mathbf{\Sigma}^{*}\right)_{\mathcal{E}^{(1)}}\right\|_1,\end{aligned}$$

where $\mathcal{E}^{(1)}$ can be taken as $\mathcal{S}^*$. This, combined with 28, results

$$\begin{aligned}\left\|\left(\widetilde{\mathbf{\Sigma}}^{(1)}-\mathbf{\Sigma}^{*}\right)_{\mathcal{S}^*}\right\|_1 &\leq \sqrt{s^*}\left\|\widetilde{\mathbf{\Sigma}}^{(1)}-\mathbf{\Sigma}^{*}\right\|_{\mathsf{F}}\\ &\leq \frac{2+\sqrt{2}}{2\rho^{-}_{2s^*+2\widetilde{s}}}\lambda s^*.\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}\left\|\widetilde{\mathbf{\Sigma}}^{(1)}-\mathbf{\Sigma}^{*}\right\|_1 &\leq 6\left\|\left(\widetilde{\mathbf{\Sigma}}^{(1)}-\mathbf{\Sigma}^{*}\right)_{\mathcal{S}^*}\right\|_1\\ &\leq \frac{3\left(2+\sqrt{2}\right)}{\rho^{-}_{2s^*+2\widetilde{s}}}\lambda s^*.\end{aligned}$$

Because $\widetilde{\mathbf{\Sigma}}^{(1)}$ is a ε-optimal solution, we have

$$\begin{aligned}&\left\langle \nabla f\left(\widetilde{\mathbf{\Sigma}}^{(1)}\right)-\nabla f\left(\mathbf{\Sigma}^{*}\right),\widetilde{\mathbf{\Sigma}}^{(1)}-\mathbf{\Sigma}^{*}\right\rangle\\ &\leq \left\|\nabla f\left(\widetilde{\mathbf{\Sigma}}^{(1)}\right)+\mathbf{\Lambda}\odot\widetilde{\mathbf{\Xi}}^{(1)}-\mathbf{\Lambda}\odot\widetilde{\mathbf{\Xi}}^{(1)}-\nabla f\left(\mathbf{\Sigma}^{*}\right)\right\|_{\max}\times\\ &\quad\left\|\widetilde{\mathbf{\Sigma}}^{(1)}-\mathbf{\Sigma}^{*}\right\|_1\\ &\leq \left(1+\frac{1}{4}\right)\lambda\left\|\widetilde{\mathbf{\Sigma}}^{(1)}-\mathbf{\Sigma}^{*}\right\|_1 \leq \frac{c_1\lambda^2 s^*}{\rho^{-}_{2s^*+2\widetilde{s}}},\end{aligned}$$

for some constant c_1. ∎