Optimal Compressive Covariance Sensing via Quadratic Sampling based on Nonconvex Learning

Wenbin Wang, *Student Member, IEEE* and Ziping Zhao, *Member, IEEE*

***Abstract*—Covariance matrix, capturing the degree of linear correlation between multiple variables, is a fundamental statistical quantity in diverse data analysis applications. Traditional method for estimation the covariance matrix typically assume full access to all available measurements. However, it becomes impractical when data evolves rapidly or the data acquisition devices have limited processing power and storage. To address these challenges, compressive covariance sensing (CCS) has emerged as a promising approach, enabling covariance matrix estimation from a reduced number of measurements, often significantly smaller than the dimension of the variables. In this paper, we focus on the quadratic (or rank-one) measurement model for CCS under the assumption of a sparse covariance structure, which minimizes memory requirements and computational complexity during the sampling process. We propose a least-squares estimator for the covariance matrix, regularized with positive-definiteness and non-convex sparsity-inducing penalties. To efficiently compute this estimator, we develop a multistage convex relaxation algorithm based on the majorization-minimization (MM) framework. This algorithm provides strong computational and statistical guarantees by leveraging local restricted strong convexity and Hessian smoothness. Our proposed algorithm achieves quadratic convergence to approximate solutions at each stage of convex relaxation. We further rigorously establish the statistical performance of all sequential approximate solutions generated by the MM-based algorithm and demonstrate that the optimal solution reached after sufficient iterations possesses oracle statistical properties in the Frobenius norm. Comprehensive numerical simulations are included to validate our theoretical results.**

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

I. Introduction

The covariance matrix is a fundamental statistical tool used to quantify the degree of linear correlation among multiple random variables [1]–[3]. It plays a pivotal role in multivariate data processing across science and engineering, including component analysis [4], factor analysis [5], beamformer design [6], portfolio optimization [7], among others. Despite its significance, the covariance matrix is not directly observable and must be estimated from empirical data, thus requiring the development of efficient estimation techniques.

Consider n independent observations $\left\{\boldsymbol{x}_i \in \mathbb{R}^d\right\}_{i=1}^n$ drawn from a zero-mean random vector $\boldsymbol{x}$. The sample covariance matrix (SCM), $\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top$, is a commonly adopted covariance estimator. However, in high-dimensional settings where the number of variables significantly exceeds the number of observations, the SCM becomes unreliable [8], [9]. To address this, it is crucial to exploit structural nature of the covariance matrix, with sparsity being one of the most favored assumptions. Sparsity suggests that many covariance entries are zero, thereby substantially reducing the number of parameters to be estimated. Thresholding techniques [10]–[12], which leverage sparse assumption, are widely used to adjust the SCM by eliminating elements below a specific threshold, thereby improving estimation accuracy. The statistical properties of these thresholding covariance estimators, including minimax lower bounds [13] and convergence rates [14], have been extensively investigated, validating both their theoretical foundations and practical applicability. Thresholding techniques have been theoretically validated to yield estimators that are asymptotically positive definite [10], [11]. However, practical applications often require that estimators be positive definite even with a finite number of samples. This property is critical in supervised learning contexts, where methods like quadratic discriminant analysis and covariance regularized regression depend on nonnegative definite covariance matrices for the convexity of their optimization landscapes. To ensure both sparsity and positive definiteness in the estimators, Rothman [15] proposed integrating a log-determinant barrier function into the soft-thresholding covariance matrix estimation problem.

Despite covariance estimation based on full measurements has been extensively studied in the literature, it faces great challenge in the scenario characterized by limited sampling capabilities and rapidly changing data environments, where the high dimensionality of the data renders it either impractical to obtain full data samples. A viable alternative approach is to estimate covariance matrix based on compressed measurements, a technique known as compressive covariance sensing or compressed covariance sketching (CCS). CCS offers significant advantages by enabling the recovery of covariance matrices from data streams while substantially reducing computation and storage overhead. Compressive sensing for covariance matrices generalizes the classical compressive sensing problem for signals [16], [17], focusing on extracting second-order statistics rather than reconstructing random vectors.

CCS is an advanced extension of the foundational principles of compressed sensing [16], [17]. In compressed sensing, the presumption of signal sparsity underpins many of the theoretical and practical developments [16], [18]. This assumption has been effectively adapted for CCS to focus on the covariance structure. [19] proposed methods for estimating sparse covariance from quadratic sampling via convex programming, providing guarantee on performance. Unlike the traditional Restricted Isometry Property (RIP-ℓ_2/ℓ_2) which does not directly apply to CCS [20], a novel mixed-norm RIP-ℓ_2/ℓ_1 has

been introduced. This criterion uses the Frobenius norm and the ℓ_1 norm to assess input and output strengths, respectively, ensuring that the dimensionality reduction projection retains signal integrity and enables perfect recovery from minimal samples. A similar criterion, RIP-ℓ_1/ℓ_1, is also applicable to sparse covariance recovery [20]. Unlike previous studies, this paper focuses on the sparse eigenvalue condition, a related but weaker concept compared to the restricted isometry property [21], [22]. Other works have explored the CCS problem under distributed sparsity assumption [23], and further extensions have considered other covariance structures such as banded, Toeplitz, circulant matrices [24] .

On the other hand, many practical applications involve random processes where directly reconstructing the random signal is not meaningful. In such case, the focus often shifts towards extracting second-order or higher-order statistics, such as the power spectrum, which contains valuable features and provide reliable information about the process. [25] developed robust compressive techniques for wideband spectrum sensing that leverage the sparsity of two-dimensional cyclic spectra in communications signals. A concurrent work [26] employed a non-negative least squares approach to estimate the signal power spectrum, based on the assumption of sparsity.

A. Contributions

This paper investigates the methodology of covariance matrix sensing, a prevalent technique for estimating high-dimensional sparse covariance matrices. Our primary focus is on the estimation of the covariance matrix utilizing a quadratic measurement model that integrates both a log-determinant penalty and a nonconvex penalty. The introduction of the nonconvex penalty aids in obtaining unbiased estimates; however, it simultaneously introduces considerable difficulties in solving the optimization problem and in performing theoretical analyses. The three questions of interest then are (a) is it feasible to accurately reconstruct the original high-dimensional covariance matrix $\mathbf{\Sigma}$ from the compressed measurements, (b) can a global optimum be achieved, and (c) how can the statistical properties of local optima be characterized. To answer the first question, we examine the sparse eigenvalue, which, to our knowledge, represents the weakest condition documented in the literature for ensuring exact recovery. For the second question, we utilize multiple Monte Carlo simulations to approximate the global optimum. In response to the third question, we propose a multistage convex relaxation technique within the majorization-minimization (MM) framework. This approach guarantees that the approximate local optimum not only converges but also possesses favorable statistical properties that approximate those of the global optimum.

B. Organization

The rest structure of this paper is organized as follows: In Section II, we establish the foundational background knowledge necessary for understanding the problem of covariance matrix sensing. Section III then transitions to the design of the algorithm, wherein we present the development of the proximal Newton method within the MM framework. Following this, Section IV provides an asymptotic statistical analysis that elucidates the reconstruction performance of the model outlined in (5), specifically addressing the convergence rate of the estimator and the iteration complexity of the algorithm. The findings from numerical experiments are subsequently detailed in Section V. The paper concludes with a discussion in Section VII. The appendices include proofs for all theoretical claims made throughout the paper.

C. Notation

This subsection provides a concise overview of the essential notations employed throughout the paper. Scalar values are denoted by conventional lower-case or upper-case letters, while boldface lower-case letters signify vectors and upper-case letters represent matrices. The symbols $\mathbf{0}$ and $\mathbf{I}$ are used to denote the all-zero matrix and the identity matrix, respectively, with dimensions that are contextually appropriate. The set of all $m \times n$ matrices with real entries is represented by $\mathbb{R}^{m\times n}$.

Let $\mathbf{X} \in \mathbb{R}^{m\times n}$ be a matrix. The elements of the matrix are denoted by X_{ij} and $[\mathbf{X}]_{ij}$, both referring to the entry in the (i,j)-th position. We denote $\mathbf{X} \succeq \mathbf{0}$ (resp. $\mathbf{X} \succ \mathbf{0}$) to indicate that $\mathbf{X}$ is positive semidefinite (resp. definite). The transpose, inverse, and determinant of the matrix $\mathbf{X}$ are represented by $\mathbf{X}^\top$, $\mathbf{X}^{-1}$, and $\det(\mathbf{X})$, respectively. The smallest and largest eigenvalues of $\mathbf{X}$ are denoted by $\lambda_{\min}(\mathbf{X})$ and $\lambda_{\max}(\mathbf{X})$. The p-norm of $\mathbf{X}$ is defined as $\|\mathbf{X}\|_p = \left(\sum_{i=1}^m \sum_{j=1}^n |X_{ij}|^p\right)^{\frac{1}{p}}$ for a real $p > 0$. For instance, the Frobenius norm and spectral norm are indicated by $\|\mathbf{X}\|_{\mathsf{F}}$ and $\|\mathbf{X}\|_2$, respectively, while $\|\mathbf{X}\|_1$ represents the sum of the absolute values of all entries of $\mathbf{X}$. Specifically, the maximum-absolute-value norm of $\mathbf{X}$ is expressed as $\|\mathbf{X}\|_{\max}$, and the minimum-absolute-value norm as $\|\mathbf{X}\|_{\min}$. The notation $\mathbf{X}_{\cdot j}$ ($\mathbf{X}_{k\cdot}$) is used to denote the j-th column (k-th row) of $\mathbf{X}$. The vectorization of $\mathbf{X}$, achieved by stacking its columns, is denoted as $\mathsf{vec}(\mathbf{X})$, while $\mathsf{mat}(\mathbf{X})$ represents the inverse operation. Furthermore, the Kronecker product and Hadamard product (also referred to as the entry-wise product) of matrices $\mathbf{X}$ and $\mathbf{Y}$ are denoted by $\mathbf{X}\otimes\mathbf{Y}$ and $\mathbf{X}\odot\mathbf{Y}$, respectively. The Euclidean inner product is defined as $\langle \mathbf{X}, \mathbf{Y}\rangle := \mathsf{tr}\left(\mathbf{X}\mathbf{Y}^\top\right)$, where $\mathsf{tr}(\cdot)$ denotes the trace of a matrix. For a real-valued function $f(\mathbf{X})$, the gradient $\nabla f(\mathbf{X})$ is a $d \times d$ matrix with the (i,j)-th element given by $\frac{\partial}{\partial X_{ij}} f(\mathbf{X})$, denoted by $\nabla_{ij} f(\mathbf{X})$, while $\nabla^2 f(\mathbf{X})$ represents the $d^2 \times d^2$ Hessian matrix. Additionally, we define $\|\mathbf{X}\|_{\mathbf{H}}^2 := \mathsf{vec}^\top(\mathbf{X})\,\mathbf{H}\mathsf{vec}(\mathbf{X})$ for the purpose of simplified representation.

For an index set $\mathcal{E}$, its cardinality is represented as $|\mathcal{E}|$, while its complement is denoted by $\overline{\mathcal{E}}$. The notation $\mathbf{X}_{\mathcal{E}}$ refers to the matrix comprising entries X_{ij} for $(i,j) \in \mathcal{E}$ and zero elsewhere. The function $\mathsf{sign}(x)$ is defined as $\mathsf{sign} = x/|x|$ when $x \neq 0$ and is equal to zero otherwise. For the functionals $f(n)$ and $g(n)$, we express $f(n) \gtrsim g(n)$ 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 positive constants c and C. Additionally, $\mathcal{O}_p(\cdot)$ is utilized to indicate boundedness in probability.

II. Preliminaries

In this section, we introduce a comprehensive framework for covariance matrix sensing, delineate its fundamental assumptions, and examine the sparse eigenvalue associated with the design sensing matrix. Subsequently, we provide several practical illustrations of the covariance matrix sensing model. Lastly, we formulate our estimator with nonconvex penalty.

A. The Measurement Model

We consider a conventional framework for indirect measurements, referred to as the quadratic measurement (or rank-one measurement) model, represented mathematically as

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

Here $\{y_i\}_{i=1}^m$ denotes the measurements, $\mathbf{S}$ is the SCM, $\left\{\boldsymbol{a}_i \in \mathbb{R}^d\right\}_{i=1}^m$ signifies the sensing vectors, $\{\eta_i\}_{i=1}^m$ indicates the noise components, and m represents the total number of measurements. The error term $\eta_i \, (1 \leq i \leq m)$ is presumed to be sampled from a sub-exponential distribution with mean 0 and variance proxy σ^2. We further analyze the decomposition of the measurements as follows:

$$y_i = \boldsymbol{a}_i^\top \mathbf{S} \boldsymbol{a}_i + \eta_i = \boldsymbol{a}_i^\top \left(\boldsymbol{\Sigma}^* + \mathbf{E}\right) \boldsymbol{a}_i + \eta_i,$$

where $\boldsymbol{\Sigma}^*$ is the true covariance matrix, and $\mathbf{E}$ is the bias term. For the sake of clarity, we define $\mathbf{A}_i \coloneqq \boldsymbol{a}_i \boldsymbol{a}_i^\top$ as the corresponding sensing matrix. Consequently, the sampling process can be articulated as $\{\langle \mathbf{A}_i, \mathbf{S} \rangle\}_{i=1}^m$. Thus, the measures $\{y_i\}_{i=1}^m$ can be expressed as $y_i \coloneqq \langle \mathbf{A}_i, \mathbf{S} \rangle + \eta_i$. Next, we define the vector of measurements as $\boldsymbol{y} \coloneqq [y_1, \cdots, y_m]^\top$ and the vector of noise as $\boldsymbol{\eta} \coloneqq [\eta_1, \cdots, \eta_m]^\top$. Additionally, we introduce the linear operator $\mathcal{A}(\mathbf{S}) : \mathbb{R}^{d \times d} \mapsto \mathbb{R}^m$, which maps a matrix $\mathbf{S} \in \mathbb{R}^{d \times d}$ to the set $\{\langle \mathbf{A}_i, \mathbf{S} \rangle\}_{i=1}^m$. Consequently, it can be reformulated as $\mathbf{y} = \mathcal{A}(\mathbf{S}) + \boldsymbol{\eta}$. It is noteworthy that the measurement model in (1) has been considered before by [19], [20]. It is essential to emphasize that the measurement model delineated in (1) is considerably more straightforward to implement and entails a reduced computational cost in comparison to full-rank measurement matrices characterized by independently and identically distributed entries.

B. The Design Matrix $\widetilde{\mathbf{A}}$

We first present several fundamental assumptions regarding the sensing vectors $\boldsymbol{a}_i$, as delineated in Assumption 1.

Assumption 1. *We assume that the sensing vectors $\boldsymbol{a}_i$'s $(1 \leq i \leq m)$ comprise independently and identically distributed (i.i.d.) sub-Gaussian random variables. Each vector $\boldsymbol{a}_i$ contains d elements, denoted as $(\boldsymbol{a}_i)_j \, (1 \leq j \leq d)$, which are characterized by the following conditions:*

$$\mathbb{E}\left[(\boldsymbol{a}_i)_j\right] = 0, \ \mathbb{E}\left[(\boldsymbol{a}_i)_j^2\right] = 1, \ \textit{and} \ \mathbb{E}\left[(\boldsymbol{a}_i)_j^4\right] > 1. \tag{2}$$

This indicates a significant heavy-tailed behavior in these components.

Rewriting Model (1), we equivalently have $y_i = (\boldsymbol{a}_i \otimes \boldsymbol{a}_i)^\top \operatorname{vec}(\mathbf{S}) + \eta_i$. Define the design matrix $\widetilde{\mathbf{A}}$ as follows:

$$\widetilde{\mathbf{A}} = \left[\ (\boldsymbol{a}_1 \otimes \boldsymbol{a}_1) \quad \cdots \quad (\boldsymbol{a}_m \otimes \boldsymbol{a}_m) \ \right]^\top \in \mathbb{R}^{m \times d^2}.$$

The Assumption 1 asserts that each row of the design matrix $\widetilde{\mathbf{A}}$ is composed of i.i.d. sub-exponential random variables. Existing research suggests that to adequately recover a sparse covariance matrix $\boldsymbol{\Sigma}$ from an underdetermined system of equations, the design matrix $\widetilde{\mathbf{A}}$ must exhibit specific favorable properties, specifically adhering to designated incoherence conditions. Various concepts of incoherence have been proposed within the sparse reconstruction literature, such as the Uniform Uncertainty Principle (UUP). Bickel et al. [27] introduced the Restricted Eigenvalue (RE) condition, demonstrating that it is one of the weakest and most general conditions imposed on sparse reconstruction problems in the literature, serving as a relaxation of the UUP. Here, we present one version of the RE condition. First, we outline a formulation of the RE condition as described by Bickel et al. [27].

Definition 2 (Restricted Eigenvalue)**.** For some integer $0 < s_0 < d$ and a positive constant c_0, the RE$(s_0, c_0, \mathbf{X})$ for the matrix $\mathbf{X} \in \mathbb{R}^{d \times d}$ necessitates that the following inequality holds:

$$\forall \mathbf{u} \neq 0, \min_{\substack{J \subset \{1,\ldots,d\}, \\ |J| \leq s_0}} \min_{\left\|\mathbf{u}_{\overline{J}}\right\|_1 \leq c_0 \left\|\mathbf{u}_J\right\|_1} \frac{\|\mathbf{X}\mathbf{u}\|_2}{\left\|\mathbf{u}_J\right\|_2} > 0, \tag{3}$$

where $\mathbf{v}_J$ denotes the subvector of $\mathbf{v} \in \mathbb{R}^d$ restricted to the subset J of $\{1, \ldots, d\}$.

In the following, we introduce the Sparse Eigenvalues (SE) condition as delineated by Fan et al. [28], which exhibits a significant correlation with the RE properties. For the detailed relationships between RE and SE, please refer to Appendix A.

Definition 3 (Sparse Eigenvalue)**.** For a positive integer $0 < s_0 < d$, the localized sparse eigenvalues are defined as

$$\rho_{s_0,r}^+ = \sup \left\{ \frac{\mathbf{u}^\top \nabla^2 f(\boldsymbol{\Sigma}) \mathbf{u}}{\mathbf{u}^\top \mathbf{u}} \mid \|\mathbf{u}\|_0 \leq s_0, \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^*\|_{\mathsf{F}} \leq r \right\};$$

$$\rho_{s_0,r}^- = \inf \left\{ \frac{\mathbf{u}^\top \nabla^2 f(\boldsymbol{\Sigma}) \mathbf{u}}{\mathbf{u}^\top \mathbf{u}} \mid \|\mathbf{u}\|_0 \leq s_0, \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^*\|_{\mathsf{F}} \leq r \right\}.$$

The SE condition has been extensively studies within the domain of high-dimensional sparse recovery or compressed sensing [29]–[33]. This condition serves as a criterion to ensure recovery for anisotropic measurements. For additional conditions closely related to the RE/SE condition, please refer to Van de Geer et al. [34].

The SE condition with parameter s_0 has been demonstrated to be valid for random Gaussian measurements, specifically for a design matrix consisting of $m = \mathcal{O}(s_0 \log d)$ independent instances of a d-dimensional Gaussian random vector $\mathbf{x}$ with covariance matrix $\boldsymbol{\Sigma}$ [32]. This holds under the assumption (3) is satisfied for the square root of $\boldsymbol{\Sigma}$. Further analysis by [35] extends these insights by establishing that the exponential width of any set does not exceed $\sqrt{\log d}$ times the Gaussian width of the set. This connection allows for the derivation of Gaussian width-based results in scenarios involving sub-exponential distributions through the application of generic chaining. Therefore, when our design matrix $\widetilde{\mathbf{A}}$ satisfies $m = \mathcal{O}\left(s_0 \log^{\frac{3}{2}} d\right)$, the SE condition is met with overwhelming probability at least $1 - \exp\left(-\epsilon^2/2\right)$ for some $\epsilon > 0$. We present the following assumption:

Assumption 4. *There exists a universal constant c_1 such that for $\widetilde{s} \geq c_1 s^*$, the SE property is guaranteed with parameters $\rho^-_{2s^*+2\widetilde{s},r}$ and $\rho^+_{2s^*+2\widetilde{s},r}$ satisfying*

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

with probability at least $1 - \exp\left(-\epsilon^2/2\right)$ for some $\epsilon > 0$.

Assumption 4 asserts that the sparse eigenvalues of the Hessian matrix $\nabla^2 f(\mathbf{\Sigma})$ are lower and upper bounded when $\mathbf{\Sigma}$ is adequate sparsity and in close proximity to $\mathbf{\Sigma}^*$ with high probability (w.h.p). In other words, $\nabla^2 f(\mathbf{\Sigma})$ possesses both bounded maximum and non-zero minimum sparse eigenvalues over a cone.

C. Practical Examples

Practical applications of this quadratic measurement model (1) are extensive and varied. One prominent example is its use in high-frequency communications systems, where noncoherent energy measurements are preferred over phase measurements. In these systems, the model is employed to estimate the energy spectral density of signals, where the focus lies solely on spectral distribution rather than signal recovery [36]. While naturally occurring signals possess infinite duration, the data available is typically limited to finite discrete-time signals. The practical implementation frequently employs random demodulators [37] to capture energy measurements via the sensing vector $\boldsymbol{a}_i$, with average energy over T observations given by:

$$\begin{aligned} y_i &= \frac{1}{T}\sum_{t=1}^{T}\left|\boldsymbol{a}_i^\top \mathbf{x}_t\right|^2 + \eta_i \\ &= \boldsymbol{a}_i^\top \mathbf{S}\boldsymbol{a}_i + \eta_i, \text{ for } i = 1, \ldots, m, \end{aligned} \tag{4}$$

where $\mathbf{S} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t^\top$ represents the sample covariance matrix[1]over T time periods. Furthermore, this model plays a crucial role in incoherent imaging within optical systems, offering a precise mathematical framework for describing the response of the imaging process to light intensity. In this scenario, the intensity y_i at each detector is modeled as the ensemble average of the squared projections of the scene state onto the detector's sensitivity pattern. This quadratic formulation is instrumental in determining the energy absorbed by each detector, aiding in the reconstruction of high-quality images from incoherent light measurements. Additional applications encountered in real-time financial [38], IoT sensor data [39] or direction of arrival estimation [40] also leverage quadratic measurement model to captures covariance information on the fly.

D. Proposed Estimator

A successful method for regression estimation is to minimize the least squares errors while adding regularization terms to ensure the low-dimensional structure of the estimator. We denote the matrix to be estimated as $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ and propose the following regularized least squares type optimization estimator

$$\min_{\mathbf{\Sigma}\succ\mathbf{0}}\left\{\frac{1}{2m}\left\|\boldsymbol{y} - \mathcal{A}(\mathbf{\Sigma})\right\|_2^2 - \tau\log\det\mathbf{\Sigma} + \sum_{i,j}p_\lambda\left(|\Sigma_{ij}|\right)\right\}. \tag{5}$$

In (5), the first term aims to minimize the empirical errors; the log-determinant barrier function in the second term ensures positive definiteness with a barrier parameter $\tau \geq 0$; p_λ in the third term represents a non-convex penalty function governed by a regularization parameter $\lambda > 0$. We impose certain restrictions on it, as illustrated in Assumption 5.

Assumption 5. *The function $p_\lambda(t) : \mathbb{R} \to \mathbb{R}$ satisfies:*

- $p_\lambda(t)$ is symmetric around zero with $p_\lambda(0) = 0$, nondecreasing on the nonnegative, differentiable almost everywhere on $(0, +\infty)$, and subdifferentiable at $t = 0$;
- $0 \leq p'_\lambda(t_1) \leq p'_\lambda(t_2) \leq \lambda$ for all $t_1 \geq t_2 \geq 0$ and $\lim_{t\to 0^+} p'_\lambda(t) = \lambda$;
- There exists an $\alpha > 0$ such that $p'_\lambda(t) = 0$ for $t \geq \alpha\lambda$;

It is suggested that a well-designed penalty function should yield an estimator with three properties: unbiasedness, sparsity, continuity [41], which coincides with the three conditions in Assumption 5. A variety of functions have been explored through past research on nonconvex regularization, of which we present a few representative examples here:

- Smooth clipped absolute deviation penalty: This penalty, due to [41], is given by

$$p'_\lambda(t) \coloneqq \begin{cases} \lambda, & \text{for } 0 < t \leq \lambda, \\ \frac{b\lambda - t}{b-1}, & \text{for } \lambda \leq t \leq b\lambda, \\ 0, & \text{for } t \geq b\lambda, \end{cases} \tag{6}$$

  where $b > 2$ is an additional tuning parameter. The authors in [41] proposed $b = 3.7$ through a Bayesian rationale, applicable when the variable dimension is less than 100.
- Minimax concave penalty regularizer: This penalty, due to [42], is defined as follows:

$$p_\lambda(t) \coloneqq \mathsf{sign}(t)\,\lambda \cdot \int_0^{|t|}\left(1 - \frac{z}{\lambda b}\right)_+ dz \tag{7}$$

  for some $b > 0$.

Note that the capped ℓ_1-penalty, takes the form $p_\lambda(t) \coloneqq \lambda \cdot \min\{|t|, b\lambda\}$, where $b > 0$ is a fixed parameter. The authors in [22] point out it is a simpler but less smooth version of the smooth clipped absolute deviation penalty.

III. Optimization Algorithm

In this section, we begin by revisiting the MM framework, which will be pivotal for the forthcoming theoretical analysis. Subsequently, we present an MM-based multistage convex relaxation algorithm to solve (5), along with explicit pseudocode. Additionally, we detail the specific solution method utilized at each stage of the MM-based algorithm —— proximal Newton homotopy algorithm.

[1]In this case, the matrix should be interpreted as the autocorrelation matrix. Nevertheless, assuming the mean is zero, it is equivalent to the covariance matrix. Therefore, the aforementioned trick can be utilized. In Section VI-B, we will discuss the approach to take when the mean is not zero.

Algorithm 1: The MM-Based Multistage Convex Relaxation Algorithm for Solving (5).

Input: $\{y_i, \boldsymbol{a}_i\}_{i=1}^m$, τ, λ;
Initialize $\widetilde{\boldsymbol{\Sigma}}^{(0)} = \mathbf{I}$
for $k = 1, 2, \ldots, K$ **do**
  $\Lambda_{ij}^{(k-1)} = p'_\lambda\left(\left|\widetilde{\Sigma}_{ij}^{(k-1)}\right|\right)$;
  $\widetilde{\boldsymbol{\Sigma}}^{(k)} = \arg\min F(\boldsymbol{\Sigma})$;
  $k = k + 1$;
end
Output: $\widetilde{\boldsymbol{\Sigma}}^{(K)}$

A. The MM Framework and Multistage Convex Relaxation Algorithm

The MM algorithm framework is an iterative process that encompasses two main steps: the Majorization step and the Minimization step. Suppose we want to minimize a real-valued function $F(\mathbf{x})$, below is the basic procedure of this algorithm framework:

- Majorization Step: In this step, the goal is to select or construct a surrogate function $\overline{F}\left(\mathbf{x} \mid \mathbf{x}^{(k-1)}\right)$ which equals the value of the original objective function $F(\mathbf{x})$ at the current iteration point $\mathbf{x}^{(k-1)}$, but is an upper bound for $F(\mathbf{x})$ at all other points. Specifically, $\overline{F}\left(\mathbf{x} \mid \mathbf{x}^{(k-1)}\right) \geq F(\mathbf{x})$ for all $\mathbf{x}$ and $\overline{F}\left(\mathbf{x}^{(k-1)} \mid \mathbf{x}^{(k-1)}\right) = F\left(\mathbf{x}^{(k-1)}\right)$.
- Minimization Step: After the surrogate function is determined, the next step is to find the point $\mathbf{x}^{(k)}$ that minimizes $\overline{F}\left(\mathbf{x} \mid \mathbf{x}^{(k-1)}\right)$, such that $\mathbf{x}^{(k)} \in \arg\min \overline{F}\left(\mathbf{x} \mid \mathbf{x}^{(k-1)}\right)$.

These two steps are executed in an alternating fashion, with each iteration aimed at reducing or at least not increasing the value of the objective function $F(\mathbf{x})$, ensuring $F\left(\mathbf{x}^{(k)}\right) \leq F\left(\mathbf{x}^{(k-1)}\right)$. Instead of minimizing $F(\mathbf{x})$ directly, the algorithm focuses on sequentially solving a series of simple optimization problem. The process begins with a feasible initial point $\mathbf{x}^{(0)}$ and continues iteratively through $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$ until a specified convergence criterion is satisfied.

We now introduce the multistage convex relaxation algorithm based on the MM approach to find the stationary solution of (5). The multistage convex relaxation algorithm is described in Algorithm 1, where we set $\boldsymbol{\Sigma}^{(0)} = \mathbf{I}$ as a trivial start. This algorithm, built on a sequential optimization framework, draws theoretical support from the analysis provided by Zhang et al. [22]. To simplify, we define

$$f(\boldsymbol{\Sigma}) = \frac{1}{2m}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{\Sigma})\|_2^2 - \tau \log\det \boldsymbol{\Sigma}.$$

Throughout the stages of the algorithm, a weighted ℓ_1-norm serves as the surrogate function for $\sum_{i,j} p_\lambda(\Sigma_{ij})$. Specifically, for each $1 \leq k \leq K$, we minimize the subsequent convex relaxation subproblems sequentially:

$$F(\boldsymbol{\Sigma}) = f(\boldsymbol{\Sigma}) + \sum_{i,j} p'_\lambda\left(\left|\widehat{\Sigma}_{ij}^{(k-1)}\right|\right)|\Sigma_{ij}|, \tag{8}$$

where $\widehat{\boldsymbol{\Sigma}}^{(k)}$ represents the optimal solution to the k-th subproblem. Each subproblem can further be reformulated as follows:

$$\min_{\boldsymbol{\Sigma} \succ \mathbf{0}} \left\{f(\boldsymbol{\Sigma}) + \|\boldsymbol{\Lambda} \odot \boldsymbol{\Sigma}\|_1\right\}, \tag{9}$$

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

$$\nabla f\left(\widehat{\boldsymbol{\Sigma}}\right) + \boldsymbol{\Lambda} \odot \widehat{\boldsymbol{\Xi}} = 0,$$

where $\widehat{\boldsymbol{\Xi}} \in \partial\left\|\widehat{\boldsymbol{\Sigma}}\right\|_1$ and the gradient

$$\nabla f(\boldsymbol{\Sigma}) = -\frac{1}{m}\mathcal{A}^*\left(\boldsymbol{y} - \mathcal{A}(\boldsymbol{\Sigma})\right) - \tau\boldsymbol{\Sigma}^{-1}.$$

Here, $\mathcal{A}^*(\cdot)$ is the conjugate operator of $\mathcal{A}(\cdot)$. Given that the problem (9) lacks an analytical solution, we aim for a suboptimal solution defined under a prespecified tolerance level 6, and terminate the iterations when the approximate KKT condition are satisfied.

Definition 6. For a specified tolerance level ε, the solution $\widetilde{\boldsymbol{\Sigma}}^{(k)}$ is deemed ε-optimal for the k-th subproblem 9 if the condition $\omega_{\boldsymbol{\Lambda}}\left(\widetilde{\boldsymbol{\Sigma}}^{(k)}\right) \leq \varepsilon$ is satisfied, where

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

In order to guarantee that the solution $\widetilde{\boldsymbol{\Sigma}}^{(k)}$ derived at each stage attains sufficient precision for all $k \geq 1$, we present the following assumption, which is essential for the convergence analysis of multistage convex relaxation.

Assumption 7. *For each iteration of the convex relaxation process* (9) *for all $k \geq 1$, we set the accuracy parameter ε as follows:*

$$\varepsilon = \frac{c_2}{\sqrt{mn}} \leq \frac{\lambda}{8},$$

where c_2 is a predetermined small constant.

B. Proximal Newton Algorithm

At each stage, the subproblem is obviously convex, enabling the application of a variety of solution techniques. However, exact analytical solutions are unattainable, iterative algorithms are required. Among these, first-order methods such as the proximal gradient algorithm and second-order method like the proximal Newton algorithm are particularly noteworthy. In this study, we adopt the proximal Newton method to enhance computational efficiency. The proximal Newton algorithm operates by solving a proximal subproblem in each iteration, which involves updating the current estimate using both gradient and Hessian information. Specifically, the algorithm computes a Newton direction to serve as the descent direction, followed by determining an appropriate step size that

ensures a sufficient decrease in the objective function, thereby facilitating convergence.

For the sake of notational simplicity, we denote the iteration index within the k-th stage as t and omit the stage index k. We build a quadratic approximation by considering the second-order Taylor expansion of $f(\mathbf{\Sigma})$:

$$\begin{aligned}\overline{f}(\mathbf{\Sigma} \mid \mathbf{\Sigma}_t) =& f(\mathbf{\Sigma}_t) + \langle \nabla f(\mathbf{\Sigma}_t), \mathbf{\Sigma} - \mathbf{\Sigma}_t \rangle \\ &+ \frac{1}{2} \|\mathbf{\Sigma} - \mathbf{\Sigma}_t\|^2_{\nabla^2 f(\mathbf{\Sigma}_t)}.\end{aligned} \tag{10}$$

We then solve

$$\mathbf{\Sigma}_{t+\frac{1}{2}} = \underset{\mathbf{\Sigma} \succ \mathbf{0}}{\arg\min}\, \overline{F}(\mathbf{\Sigma} \mid \mathbf{\Sigma}_t, \mathbf{\Lambda}), \tag{11}$$

where

$$\overline{F}(\mathbf{\Sigma} \mid \mathbf{\Sigma}_t, \mathbf{\Lambda}) = \overline{f}(\mathbf{\Sigma} \mid \mathbf{\Sigma}_t) + \|\mathbf{\Lambda} \odot \mathbf{\Sigma}\|_1.$$

Denote

$$\Delta \mathbf{\Sigma}_t = \mathbf{\Sigma}_{t+\frac{1}{2}} - \mathbf{\Sigma}_t$$

as the Newton direction for the function $F(\mathbf{\Sigma})$. Particularly, we adopt coordinate descent algorithms combined with active set strategy to get the Newton direction, which can improve the time complexity. The detailed computation will be introduced in the Appendix B.

An additional backtracking line searching procedure is included to guarantee the descent of the objective value. We try to find a step size $\beta \in (0, 1]$ such that the Armijo condition [43] holds. Specially, we start with a fixed constant $\mu \in (0.5, 1)$ and update $\beta = \mu^q$ from $q = 0$ with a constant decrease rate until we find the smallest $q \in \mathbb{N}$ such that

$$F(\mathbf{\Sigma}_t + \beta \Delta \mathbf{\Sigma}_t) \leq F(\mathbf{\Sigma}_t) + \alpha \beta \delta_t,$$

where $\alpha \in (0, 0.5)$ and

$$\begin{aligned}\delta_t =& \langle \nabla f(\mathbf{\Sigma}_t), \Delta \mathbf{\Sigma}_t \rangle - \|\mathbf{\Lambda} \odot \mathbf{\Sigma}_t\|_1 \\ &+ \|\mathbf{\Lambda} \odot (\mathbf{\Sigma}_t + \Delta \mathbf{\Sigma}_t)\|_1.\end{aligned} \tag{12}$$

Then $\mathbf{\Sigma}_{t+1}$ is set as $\mathbf{\Sigma}_{t+1} = \mathbf{\Sigma}_t + \beta \Delta \mathbf{\Sigma}_t$. The whole proximal Newton algorithm is summarized in Algorithm 2.

Algorithm 2: Proximal Newton Algorithm (ProxNewton) With Back-tracking Line Search for Solving (9).

Input: $\widetilde{\mathbf{\Sigma}}^{(k-1)}$, $\mathbf{\Lambda}^{(k-1)}$, ε;
Initialize $t = 0$, $\mathbf{\Sigma}_t^{(k)} = \widetilde{\mathbf{\Sigma}}^{(k-1)}$, $\mu = 0.8$, $\alpha = 0.3$
repeat
  $\mathbf{\Sigma}_{t+\frac{1}{2}}^{(k)} = \underset{\mathbf{\Sigma} \succ \mathbf{0}}{\arg\min} \overline{F}\left(\mathbf{\Sigma} \mid \mathbf{\Sigma}_t^{(k)}, \mathbf{\Lambda}^{(k-1)}\right)$
  $\Delta \mathbf{\Sigma}_t^{(k)} = \mathbf{\Sigma}_{t+\frac{1}{2}}^{(k)} - \mathbf{\Sigma}_t^{(k)}$
  $\delta_t = \left\langle \nabla f\left(\mathbf{\Sigma}_t^{(k)}\right), \Delta \mathbf{\Sigma}_t^{(k)} \right\rangle - \left\|\mathbf{\Lambda}^{(k-1)} \odot \mathbf{\Sigma}_t^{(k)}\right\|_1$
    $+ \left\|\mathbf{\Lambda}^{(k-1)} \odot \left(\mathbf{\Sigma}_t^{(k)} + \Delta \mathbf{\Sigma}_t^{(k)}\right)\right\|_1$
  $\beta = 1$, $q = 0$
  repeat
    $\beta = \mu^q$
    $q = q + 1$
    if $\mathbf{\Sigma}_t^{(k)} + \beta \Delta \mathbf{\Sigma}_t^{(k)} \preceq \mathbf{0}$ **then**
      continue;
    end
  until $F\left(\mathbf{\Sigma}_t^{(k)} + \beta \Delta \mathbf{\Sigma}_t^{(k)}\right) \leq F\left(\mathbf{\Sigma}_t^{(k)}\right) + \alpha \beta \delta_t$;
  $\mathbf{\Sigma}_{t+1}^{(k)} = \mathbf{\Sigma}_t^{(k)} + \beta \Delta \mathbf{\Sigma}_t^{(k)}$
  $t = t + 1$
until $\omega_{\mathbf{\Lambda}^{(k-1)}}\left(\mathbf{\Sigma}_t^{(k)}\right) \leq \varepsilon$;
Output: $\widetilde{\mathbf{\Sigma}}^{(k)} = \mathbf{\Sigma}_t^{(k)}$

IV. Statistical and Computational Theories

In this section, we first introduce some technical assumptions and important definitions. Following these, we establish the statistical convergence rate of our proposed covariance estimator and the iteration complexity of the proposed algorithm. The proofs are provided in the supplementary material.

A. Assumptions

We present several assumptions pertinent to the true covariance matrix, denoted as $\mathbf{\Sigma}^*$. We define the support set of $\mathbf{\Sigma}^*$ as $\mathcal{S}^* = \left\{(j,k) \mid \mathbf{\Sigma}^*_{jk} \neq 0\right\}$, with s^* representing its size, i.e, $s^* = |\mathcal{S}^*|$. These assumptions serve as essential parameters for the examination of the sparsity and structural characteristics of the true covariance matrix within the framework of our proposed statistical model.

Assumption 8. *For the ground truth covariance matrix $\mathbf{\Sigma}^*$, there exists finite upper and positive lower bounds on its eigenvalue. Specifically, there exists a constant $\kappa \geq 1$ satisfies*

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

Assumption 8 is a common premise in the literature concerning of the estimation of sparse covariance matrices [9], [15], [44], [45]. This assumption offers several advantages, and readers are encouraged to consult [15] for further details.

Assumption 9. *Given the true covariance matrix $\mathbf{\Sigma}^*$, there exist universal constants α, c_3 such that*

$$\|\mathbf{\Sigma}^*_{\mathcal{S}^*}\|_{\min} = \min_{(i,j) \in \mathcal{S}^*} \left|\mathbf{\Sigma}^*_{ij}\right| \geq (\alpha + c_3)\lambda,$$

where α is a constant introduced in Assumption 5, and $c_3 \in (0, \alpha)$ satisfies $p'_\lambda(c_3\lambda) \geq \frac{\lambda}{2}$.

Additionally, it is essential to select λ appropriately to ensure that the regularization is sufficiently large to eliminate irrelevant coordinates, thereby yielding the solution is adequately sparse.

Assumption 10. *There exists a generic constant c_4 such that*

$$\lambda = c_4 \sqrt{\frac{\log d}{m}} \geq 4\left(\|\nabla f(\mathbf{\Sigma}^*)\|_{\max} + \varepsilon\right).$$

Assumption 9 is referred to the minimum signal strength condition, commonly applied in the analysis of non-convex penalized regression problems [28], [41], [42]. Assumption 10 sets the tuning parameter λ in the order of $\sqrt{\frac{\log d}{m}}$. It prevents

λ from becoming overly large as the number of measurements m increases, thereby ensuring that the estimators maintain a close correspondence with the true model parameters. Taking these assumptions into account, we can obtain the oracle rate of the convergence.

Definition 11. Define a local region around $\mathbf{\Sigma}^*$ as

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

In our analysis, we set the radius r as $\frac{\rho^-_{2s^*+2\widetilde{s}}}{4\tau\kappa}$. Under Assumption 10, we prove that the solution produced by our algorithm will fall within the local region $\mathcal{B}\left(\mathbf{\Sigma}^*, \frac{\rho^-_{2s^*+2\widetilde{s}}}{4\tau\kappa}\right)$ after a finite number of iterations.

B. Statistical Guarantees and Consequences

The following will present the main result, which demonstrates the contraction property of the solution path $\left\{\widetilde{\mathbf{\Sigma}}^{(k)}\right\}_{k\geq 1}$.

Theorem 12 (Contraction Property)**.** *Consider the estimator in 5 and suppose the Assumptions 1, 4, 8, and 9 hold. Then with probability exceeding* $1 - \exp\left(-\epsilon^2/2\right)$ *for some* $\epsilon > 0$*, the* ε*-optimal solution* $\widetilde{\mathbf{\Sigma}}^{(k)}$ *is bounded by:*

$$\begin{aligned}\left\|\widetilde{\mathbf{\Sigma}}^{(k)} - \mathbf{\Sigma}^*\right\|_{\mathsf{F}} \leq \frac{1}{\rho^-_{2s^*+2\widetilde{s}}}\Bigg(&\underbrace{\left\|\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}}}_{\textit{oracle rate}} + \underbrace{\varepsilon\sqrt{s^*}}_{\textit{optimization error}}\Bigg) \\ &+ \underbrace{\delta\left\|\widetilde{\mathbf{\Sigma}}^{(k-1)} - \mathbf{\Sigma}^*\right\|_F}_{\textit{contraction}},\end{aligned} \tag{13}$$

for $1 \leq k \leq K$*, where* $\delta \in (0,1)$ *is the contraction factor, provided that* $m = \mathcal{O}\left((s^* + \widetilde{s})\log^{\frac{3}{2}} d\right)$.

Remark 13. The oracle estimator $\widehat{\mathbf{\Sigma}}^O$ is defined with prior knowledge of the true support set $\mathcal{S}^*$, and is given by

$$\begin{aligned}\widehat{\mathbf{\Sigma}}^O &= \arg\min_{\mathbf{\Sigma}} f\left(\mathbf{\Sigma}\right) \\ \text{s.t.} \quad &\mathbf{\Sigma}_{\overline{\mathcal{S}^*}} = \mathbf{0}.\end{aligned}$$

The statistical convergence rate of this oracle estimator, commonly referred to as the oracle rate, provides a benchmark for evaluating the theoretical upper limit of estimator performance when the true support set is known. Specifically, the distance between $\widehat{\mathbf{\Sigma}}^O$ and $\mathbf{\Sigma}^*$ is characterized by $\left\|\widehat{\mathbf{\Sigma}}^O - \mathbf{\Sigma}^*\right\|_{\mathsf{F}} \lesssim \left\|\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}}$. This inequality indicates that the oracle estimator can theoretically approximate the true parameter matrix $\mathbf{\Sigma}^*$very closely when the true support sets are known.

Theorem 12 elaborates the estimation discrepancy between the ε-optimal solution $\widetilde{\mathbf{\Sigma}}^{(k)}$ and the ground truth $\mathbf{\Sigma}^*$ is constrained by three primary factors: the oracle rate, the optimization error, and a contraction term. The oracle rate represents the error bound achievable under ideal conditions using the optimal strategy; the optimization error reflects the performance gap between the actual optimization algorithm and the optimal strategy; the contraction term indicates the rate at which the error converges to its minimum value as the number of iterations increases. The ensuing results specify the exact statistical convergence rate.

Corollary 14. *Let* $\mathbf{x}$ *be a sub-Gaussian random vector with zero mean and covariance* $\mathbf{\Sigma}^*$ *and* $\{\mathbf{x}_i\}_{i=1}^n$ *be a collection of independent and identically distributed (i.i.d) samples from* $\mathbf{x}$ *. Suppose that Assumptions 1, 4, 7, 8, and 9 hold, that is,* $\lambda \asymp \sqrt{\frac{\log d}{mn}}$ *and* $\varepsilon \lesssim \sqrt{\frac{1}{mn}}$*. If* $\tau \lesssim \sqrt{\frac{1}{mn}}\left\|\left(\mathbf{\Sigma}^*\right)^{-1}\right\|^{-1}_{\max}$*, then the* ε*-optimal solution* $\widetilde{\mathbf{\Sigma}}^{(1)}$ *satisfies*

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

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

Corollary 15. *Let* $\mathbf{x}$ *be a sub-Gaussian random vector with zero mean and covariance* $\mathbf{\Sigma}^*$ *and* $\{\mathbf{x}_i\}_{i=1}^n$ *be a collection of independent and identically distributed (i.i.d) samples from* $\mathbf{x}$ *. Suppose that Assumptions 1, 4, 7, 8, and 9 hold, that is,* $\lambda \asymp \sqrt{\frac{\log d}{mn}}$ *and* $\varepsilon \lesssim \sqrt{\frac{1}{mn}}$*. If* $\tau \lesssim \sqrt{\frac{1}{mn}}\left\|\left(\mathbf{\Sigma}^*\right)^{-1}\right\|^{-1}_{\max}$*, and* $K \gtrsim \log\left(\lambda\sqrt{mn}\right) \gtrsim \log\log d$*, then the* ε*-optimal solution* $\widetilde{\mathbf{\Sigma}}^{(K)}$ *satisfies*

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

Corollary 14 and Corollary 15 are direct consequence of Theorem 12. The former addresses the scenario where $k = 1$ and describes a contraction property resulting from the MM-based multistage convex relaxation algorithm. It demonstrates that to achieve the oracle rate, the optimization error ε must be chosen such that $\varepsilon \leq \frac{\left\|\left(\nabla f(\mathbf{\Sigma}^*)\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}}}{\sqrt{s^*}}$, and the parameter K must be sufficiently large. The latter implies that, with minimal assumptions, solving no more than approximately $\log\log d$ convex problems is enough to achieve the oracle rate $\sqrt{\frac{s^*}{mn}}$.

C. Computational Theory

Lemma 16. *Under Assumptions 4 and 8,* $f\left(\mathbf{\Sigma}\right)$ *is* $\left(\rho^-_{2s^*+2\widetilde{s}} + \frac{16\tau^3}{\kappa^2\left(4\tau+\rho^-_{2s^*+2\widetilde{s}}\right)^2}\right)$*-strongly convex and* $\left(\rho^+_{s^*+2\widetilde{s}} + \frac{16\tau^3\kappa^2}{\left(4\tau-\rho^-_{2s^*+2\widetilde{s}}\right)^2}\right)$*-smooth in the region* $\mathcal{B}\left(\mathbf{\Sigma}^*, \frac{\rho^-_{2s^*+2\widetilde{s}}}{4\tau\kappa}\right)$*, where*

- κ is defined in Assumption 8
- $\rho^-_{2s^*+2\widetilde{s}}$, $\rho^+_{2s^*+2\widetilde{s}}$ is defined in Assumption 4
- τ is the positive tuning hyper-parameter.

Lemma 16 ensures that within the specified region, the function exhibits strong convexity and smoothness properties, which are essential for stability and convergence analysis. Next, we present the explicit iteration complexity of the proposed algorithm. We begin by characterizing the convergence for the first stage.

Theorem 17. *For $k = 1$, suppose Assumption 1, 4, 7, 8, and 9 hold. After a sufficient number of iterations $T < \infty$, for all $t \geq T$, the following conditions are satisfied: $\left\|\mathbf{\Sigma}_t^{(1)} - \mathbf{\Sigma}^*\right\|_{\mathsf{F}} \leq r$, $\left\|\left(\mathbf{\Sigma}_t^{(1)}\right)_{\overline{\mathcal{S}^*}}\right\|_0 \leq \widetilde{s}$ and*

$$\left\|\mathbf{\Sigma}_{t+1}^{(1)} - \widehat{\mathbf{\Sigma}}^{(1)}\right\|_{\mathsf{F}} \leq \frac{\tau\kappa^3}{\rho^-_{2s^*+2\widetilde{s}}}\left\|\mathbf{\Sigma}_t^{(1)} - \widehat{\mathbf{\Sigma}}^{(1)}\right\|_{\mathsf{F}}^2.$$

Moreover, we need at most $T + \log\log\left(\frac{3\rho^+_{2s^*+2\widetilde{s}}}{\varepsilon}\right)$ iterations to terminate the proximal Newton algorithm.

Theorem 18. *For $k \geq 2$, suppose 1, 4, 7, 8, and 9 hold. For all iterations t, we have $\left\|\mathbf{\Sigma}_t^{(k)} - \mathbf{\Sigma}^*\right\|_{\mathsf{F}} \leq r$, $\left\|\left(\mathbf{\Sigma}_t^{(k)}\right)_{\overline{\mathcal{S}^*}}\right\| \leq \widetilde{s}$, which guarantee $\beta = 1$, and*

$$\left\|\mathbf{\Sigma}_{t+1}^{(k)} - \widehat{\mathbf{\Sigma}}^{(k)}\right\|_{\mathsf{F}} \leq \frac{\tau\kappa^3}{\rho^-_{2s^*+2\widetilde{s}}}\left\|\mathbf{\Sigma}_t^{(k)} - \widehat{\mathbf{\Sigma}}^{(k)}\right\|_{\mathsf{F}}^2.$$

Moreover, the proximal Newton algorithm requires at most $\log\log\left(\frac{3\rho^+_{2s^*+2\widetilde{s}}}{\varepsilon}\right)$ iterations to terminate.

Theorem 17 demonstrates that the solution enters the ball $\mathcal{B}\left(\mathbf{\Sigma}^*, \frac{\rho^-_{2s^*+2\widetilde{s}}}{4\tau\kappa}\right)$ following the initial stage. To satisfy the approximate KKT conditions, the number of iterations is capped at $\mathcal{O}\left(T + \log\log\left(\frac{3\rho^+_{2s^*+2\widetilde{s}}}{\varepsilon}\right)\right)$. The resulting solution $\widetilde{\mathbf{\Sigma}}^{(1)}$ exhibits notable qualities, laying the foundation for improved computational efficiency in the subsequent stages of our proximal Newton algorithm. Theorem 18 states that the sparsity of the solution is preserved throughout the iterations, and the solution remains within $\mathcal{B}\left(\mathbf{\Sigma}^*, \frac{\rho^-_{2s^*+2\widetilde{s}}}{4\tau\kappa}\right)$. Due to the sparsity of the solution, the algorithm exhibits quadratic convergence. Moreover, the algorithm achieves a logarithmic number of iterations, $\mathcal{O}\left(\log\log\left(\frac{3\rho^+_{2s^*+2\widetilde{s}}}{\varepsilon}\right)\right)$, to satisfy the approximate KKT condition.

V. Numerical Experiments

In this section, we examine the practical recovery performance of proposed estimator and provide numerical results for the proposed algorithm for sparse covariance matrix sensing. Section V-A describe the generation of synthetic data and the measurement criteria used in the paper. In Section V-B and Section V-C, the statistical theoretical results of Section IV-B and the computational theoretical results of Section IV-C are demonstrated, respectively. Section V-D compares the relative estimation error of our proposed estimator and existing ℓ_1 estimator. The non-convex penalty function chosen for these experiments is the MCP, with a constant setting of $b = 2$ across all trials. The selection of the tuning parameters λ and τ is determined through the application of five-fold cross-validation. All methods are implemented in MATLAB, and run on an Intel $i7-10700$ 2.90 GHz $\times 8$ with 16 GB of RAM. All results are averaged on 100 Monte Carlo realizations.

A. Synthetic Datasets and Measurement Criteria

Synthetic data for this section is generated by the following methods. For different fixed d, n independent data points are generated by drawing i.i.d. samples from a $\mathcal{N}(0, \mathbf{\Sigma}^*)$ distribution. In this paper, we consider the following three typical covariance models and two randomly generated sparse covariance models.

- Banded matrix:
$$\Sigma_{ij}^* = \begin{cases} 1 - \frac{|i-j|}{10}, & |i-j| \leq 10, \\ 0, & \text{otherwise.} \end{cases}$$
- Block matrix: The indices $1, 2, \ldots, d$ are partitioned into 10 ordered groups of equal size with
$$\Sigma_{ij}^* = \begin{cases} 1, & i = j, \\ 0.5, & i \text{ and } j\,(i \neq j) \text{ belong to the same group,} \\ 0, & \text{otherwise.} \end{cases}$$
- Toeplitz matrix: $\Sigma_{ij}^* = 0.75^{|i-j|}$.
- Sprandsym matrix: $\mathbf{\Sigma}^*$ is generated utilizing the "sprandsym" built-in function in MATLAB with a sparsity value s^*.
- Probability matrix: $\mathbf{\Sigma}^*$ is generated with off-diagonal entries drawn $i.i.d.$ from a uniform $(-1, 1)$ distribution. These entries are set to zero with a fixed probability $\mathbf{P}$. Finally, a multiple of the identity was added to the resulting matrix.

The first two models are sparse, while the third one is approximately sparse. The fourth model is positive, symmetric, and sparse when s^* is small. The fifth model is positive definite, well-conditioned, and can be sparse when $\mathbf{P}$ is appropriate (in this case, either $\mathbf{P} = 0.97$ or $\mathbf{P} = 0.85$ to simulate a very sparse and a somewhat sparse model).

The estimation performance is measured by the absolute error and relative error under both the Frobenius norm and the spectral norm, respectively. Specifically, the Frobenius absolute error (FAE) is defined as $\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_{\mathsf{F}}$, where $\widehat{\mathbf{\Sigma}}$[2] is the estimated covariance matrix and $\mathbf{\Sigma}^*$ is the true value. Similarly, the Frobenius relative error (FRE) is defined as $\frac{\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_{\mathsf{F}}}{\|\mathbf{\Sigma}^*\|_{\mathsf{F}}}$. The selection performance is examined by the false positive rate (FPR) and the true positive rate (TPR), defined as, respectively

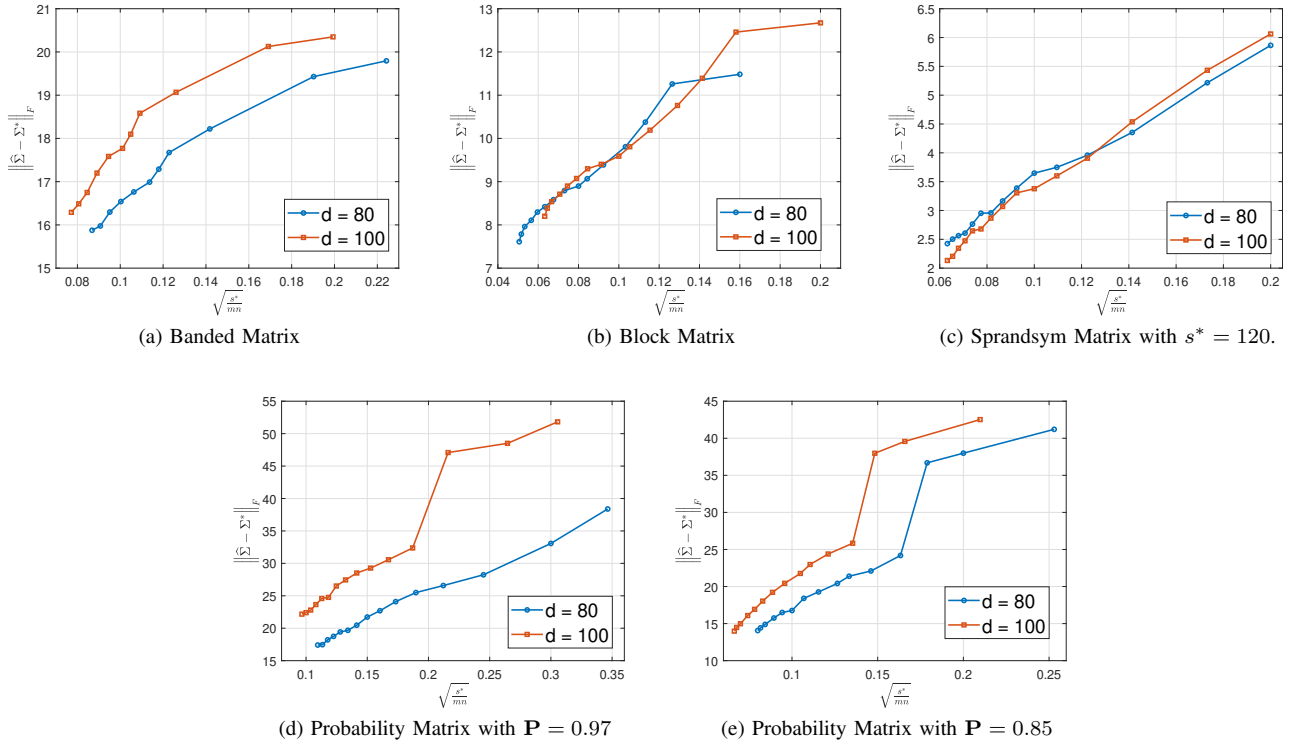
$$\text{FPR} = \frac{\#\left\{(i,j) : \widehat{\Sigma}_{ij} \neq 0 \;\&\; \Sigma_{ij}^* = 0\right\}}{\#\left\{(i,j) : \Sigma_{ij}^* = 0\right\}},$$

$$\text{TPR} = \frac{\#\left\{(i,j) : \widehat{\Sigma}_{ij} \neq 0 \;\&\; \Sigma_{ij}^* \neq 0\right\}}{\#\left\{(i,j) : \Sigma_{ij}^* \neq 0\right\}}.$$

B. Demonstration of Oracle Rate

The oracle rate derived for proposed estimator and algorithm in Section IV-B is shown to be dependent on the sparsity level s^*, the number of measurements m, and number of samples n.

[2]In the whole Section V and Appendix E, we use $\widehat{\mathbf{\Sigma}}$ to denote a generic covariance estimator, which can be the estimation results obtained by various algorithms and estimators.



(a) Banded Matrix (b) Block Matrix (c) Sprandsym Matrix with $s^* = 120$.

(d) Probability Matrix with $\mathbf{P} = 0.97$ (e) Probability Matrix with $\mathbf{P} = 0.85$

*The Toeplitz model is not reported due to the fact that it is not sparse but approximately sparse.

Fig. 1. The oracle rate for different dimension of different covariance model.

To demonstrate this theoretical result, we apply our proposed algorithm to synthetic datasets, as describe in Section V-A. Fig. 1 illustrates the estimation performance across various matrix configurations differentiated by dimensions d, sparsity level s^*, and measurement count m. The x-axis is $\sqrt{\frac{s^*}{mn}}$, and the y-axis represents FAE. Two distinct lines depicts the FAE for dimensions $d = \{80, 100\}$. We observe that an increase in measurement size m and sample count n corresponds to a reduction in FAE, visually demonstrating how the estimation quality of the covariance matrix varies with the oracle rate. According to Corollary 15, the Frobenius norm difference $\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_{\mathsf{F}}$ converges to the statistical errors as the stage index k increases, following the order of $\mathcal{O}\left(\sqrt{\frac{s^*}{mn}}\right)$. We can see that the estimation errors grow approximately linearly with the theoretical rate, which validates our theoretical guarantee.

C. Demonstration of Computational Rate

Fig. 2 illustrates a detailed analysis of the convergence rates observed during distinct phases of the computational process. In the initial phase, the algorithm display a sub-linear convergence, for that the optimization problem is characterized by a lack of strong convexity. As the computation progresses and the solution approaches the sparse region, a notable transformation occurs in the dynamics of convergence. In this region, the characteristics of the feasible solutions change markedly—they become increasingly sparse. Concurrently, the objective function undergoes a transformation into an essentially low-dimensional form. This new form is not only sparse but also exhibits properties of strong convexity and smoothness, which are crucial for faster convergence. Consequently, our algorithm exhibits a linear convergence rate for $k \geq 2$. Interestingly, this property extends even to the case of $k = 1$, wherein the convergence rate remains sub-linear until the algorithm transitions into the contraction region, at which point it achieves a quadratic rate of convergence.

D. Simulation Experiment Comparisons

We conduct a series of Monte Carlo trials to evaluate the performance of our proposed covariance matrix sensing algorithm using "Sprandsym matrix model" with dimension $d = 100$ and $n = 50$. The efficacy of the algorithm is assessed under varying sparsity levels and noise conditions, and comparisons are made against existing estimators employing different penalization techniques.

Initially, we examine the FRE of the estimated covariance matrices over various sparsity levels, represented by s^* values of $120, 180, 240$ and 300. The results, obtained in the absence of noise and while sensing the true covariance $\mathbf{\Sigma}^*$ with sample count n fixed at 50, are depicted in Figure 3a. This figure illustrates the inverse relationship between FRE and the number of measurements m, where an increase in m correlates with enhanced estimation accuracy. Concurrently, within the same number of measurements, a lower s^* value, which corresponds to higher sparsity, results in a reduced FRE, highlighting the influence of sparsity on estimation precision.

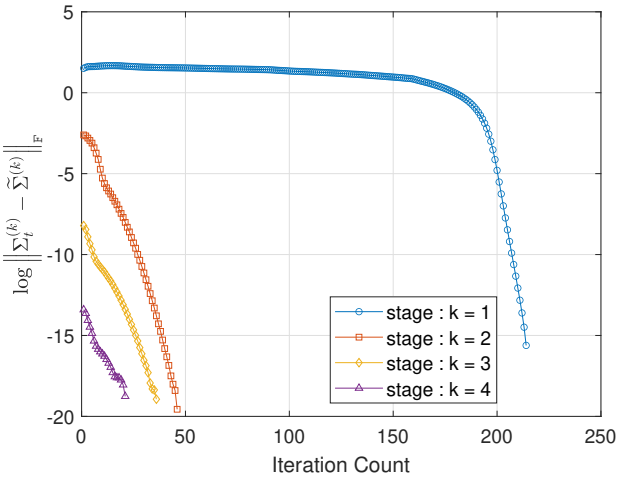


Fig. 2. Computational rate of convergence in each stage for simulation experiment. The $\mathbf{x}$-axis is the iteration count t within the k-th subproblem. A distinct phase transition in the algorithmic convergence rate from sub-linear to super-linear is observable as the iterations progress into the sparse region.

To reflect more realistic conditions, additive noise is introduced into each measurement, where the noise η_i is generated from a sub-exponential distributed random variable scaled by γ, specifically $\gamma \cdot \mathcal{M}(0,1)$. The influence of this noise, paticularly at a level of $\gamma = 10^{-1}$, on the average relative error is illustrated in Figure 3b. These results demonstrate the resilience of our estimator against external noise disturbances, with the trend of FRE across various measurements and sparsity levels remaining consistent even in noisy environments. Concurrently, we examine the variations in FRE relative to sample number as depicted in Figure 3c, while maintaining a constant measurement count of $m = 300$.

Morover, we explore the performance of our proposed algorithm under different noise intensities for the sparsity level $s^* = 300$. The impact of these noise intensities on the accuracy of the estimator is showcased in Figure 4, offering a comprehensive view of how noise intensity affects the reliability of the estimation process. The figure indicates that at lower numbers of measurements, the noise intensity marginally affects the accuracy of covariance matrix recovery. However, as the number of measurements increases, the influence of noise diminishes, leading to more precise recovery outcomes. This trend underscores the importance of sufficient data in mitigating the effects of noise and enhancing the reliability of the estimation process.

Lastly, we conduct a comparative analysis between our proposed covariance matrix sensing estimator and an existing estimator that incorporates an ℓ_1 penalty. Each test scenario incorporated a consistent level of additive noise $\left(\gamma = 10^{-1}\right)$, with the comparative results depicted in Figure 5. The comparison reveals that our MCP-based estimator consistently outperforms the ℓ_1-penalty-based estimator in terms of the Frobenius norm error. This supports our theoretical proposition that using a nonconvex penalty can significantly reduce the error in covariance matrix sensing. For rigorous statistical validation of these results, we have detailed the average metrics and corresponding standard errors (in parentheses) in the Appendix E.

VI. Discussion

A. Bilinear Rank-one Measurement Model

Our research findings provides evidence that covariance matrices can be effectively reconstructed from a limited set of quadratic measurements, contingent upon the sparsity of the underlying structure. This reconstruction process is particularly advantageous in terms of storage efficiency when the sensing vectors are i.i.d. from sub-Gaussian distribution, and when the number of measurements exceeds the basic sampling theoretical limit. Additionally, our analysis naturally extends to the bilinear rank-one measurement model, where measurements take the form $y_i = \boldsymbol{a}_i^\top \boldsymbol{\Sigma} \boldsymbol{b}_i$, with the sensing vectors $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ being independently generated. This extension substantiates our results within an asymmetric sensing framework, which has significant implications in various fields, including communication [46], imaging [47], and machine learning [48].

The bilinear model introduces additional complexity due to the interaction between two independent sensing vectors, $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$. Despite this, it is straightforward to demonstrate that this configuration meets the Assumption 3, a critical requirement for successful recovery. The derivation process closely parallels that of the quadratic model, with necessary modifications to accommodate the bilinear structure. Importantly, we also establish that within the context of the bilinear model, the Hessian matrix retains its symmetric properties, thereby enabling our proposed algorithm to sustain both efficiency and effectiveness. As a result, the bilinear measurement model inherits the oracle property, further enhancing the robustness and extensive applicability of our approach.

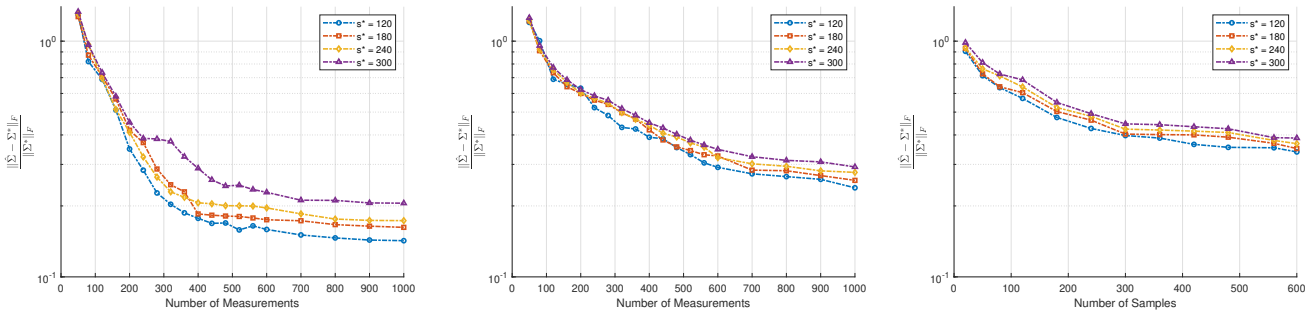
B. Covariance Sketching for variable with non-zero mean

In the preceding discussion, we consider the scenario where the mean signal vector is assumed to be zero. However, this assumption frequently does not hold true in numerous practical contexts. In case where the mean vector is not known, one might first utilize robust methodologies to estimate the mean vector, and subsequently apply this estimation to perform a robust calculation of the covariance. This bifurcated approach to estimating the mean and covariance not only introduces additional tuning parameters but also increases both statistical variability and computational complexity.

To streamline this process, we propose an end-to-end methodology. We recommend a pairwise difference approach that obviates the necessity for direct mean estimation. Specifically, we define $N \coloneqq n(n-1)/2$ and consider the pairwise differences

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

which are identically distributed with zero mean. The covariance of these differences is characterized by $\mathrm{cov}(\widetilde{\mathbf{x}}) = 2\boldsymbol{\Sigma}$. Furthermore, the sample covariance matrix can be expressed as a U-statistic. Normally, the sample covariance is calculated as $\boldsymbol{\Sigma}_{\text{sam}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top$ where $\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$. However, using the pairwise differences, it can be reformulated



(a) (b) (c)

Fig. 3. The FRE of the estimated covariance matrices is examined in two contexts: (a) the true covariance under varying levels of sparsity in the absence of noise; (b) the sample covariance under different sparsity levels when subjected to a noise parameter of $\gamma = 0.1$ and $n = 50$; (c) the sample covariance under different sparsity levels when subjected to a noise parameter of $\gamma = 0.1$ and $m = 300$;

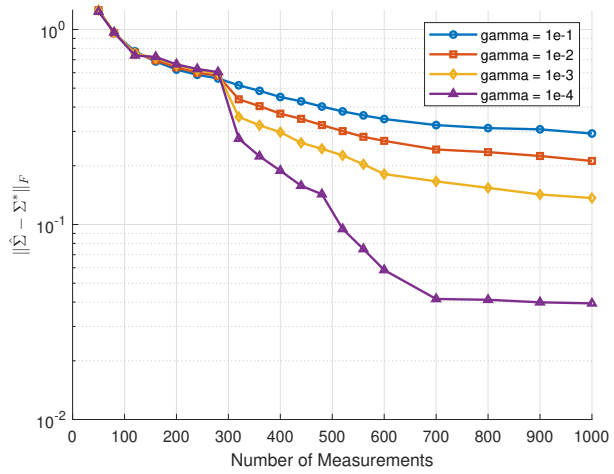


Fig. 4. The FRE of the estimated covariance matrices for different noise levels when $s^* = 300$.

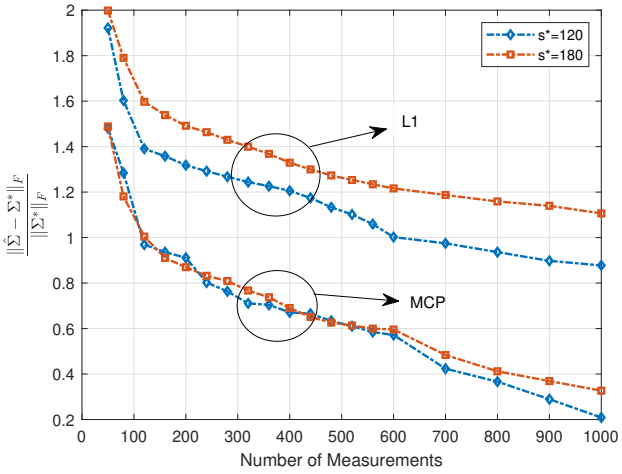


Fig. 5. The FRE of the estimated covariance matrices for different sparsity level with noise (ℓ_1 v.s. MCP).

equivalently as $\mathbf{\Sigma}_{\text{sam}} = \frac{1}{N}\sum_{i=1}^{N} \widetilde{\mathbf{x}}_i \widetilde{\mathbf{x}}_i^\top /2$, thereby simplifying the computation by avoiding the direct estimation of the mean. This reformulation provides a more streamlined approach to covariance estimation, reducing both the computational load and the potential statistical errors associated with separate mean and covariance estimations.

VII. Conclusion

In this paper, we present a comprehensive study on covariance matrix sensing, focusing on the high-dimensional sparse covariance matrices estimation through the quadratic measurements model. This model is particularly suited for scenarios with constrained processing capabilities and limited memory, such as real-time data acquisition devices. It encompasses various sampling strategies, primarily those that capture magnitude or energy samples. Our findings demonstrate that sparse covariance matrices can be effectively and accurately reconstructed using a minimal set of quadratic measurements with bottommost storage requirements. We provide rigorous theoretical backing and empirical evidence to support these claims. Notably, our proposed estimators exhibit superior statistical convergence rates compared to existing methods, highlighting their practical utility and efficiency in real-world applications.

REFERENCES

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

[2] A. Bose and M. Bhattacharjee, *Large covariance and autocovariance matrices*. Chapman and Hall/CRC, 2018.

[3] H. Tsukuma and T. Kubokawa, *Shrinkage estimation for mean and covariance matrices*. Springer, 2020.

[4] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[5] R. L. Gorsuch, *Factor analysis: Classic edition*. Routledge, 2014.

[6] R. A. Monzingo and T. W. Miller, *Introduction to adaptive arrays*. Scitech publishing, 2004.

[7] L. K. Chan, J. Karceski, and J. Lakonishok, "On portfolio optimization: Forecasting covariances and choosing the risk model," *The review of Financial studies*, vol. 12, no. 5, pp. 937–974, 1999.

[8] J. Fan, Y. Liao, and H. Liu, "An overview of the estimation of large covariance and precision matrices," *The Econometrics Journal*, vol. 19, no. 1, pp. C1–C32, 2016.

[9] C. Lam and J. Fan, "Sparsistency and rates of convergence in large covariance matrix estimation," *Annals of Statistics*, vol. 37, no. 6B, p. 4254, 2009.

[10] P. J. Bickel and E. Levina, "Covariance regularization by thresholding," *The Annals of Statistics*, vol. 36, no. 6, pp. 2577 – 2604, 2008.

[11] N. El Karoui, "Operator norm consistent estimation of large-dimensional sparse covariance matrices," 2008.

[12] A. J. Rothman, E. Levina, and J. Zhu, "Generalized thresholding of large covariance matrices," *Journal of the American Statistical Association*, vol. 104, no. 485, pp. 177–186, 2009.

[13] T. T. Cai and H. H. Zhou, "Minimax estimation of large covariance matrices under ℓ_1-norm," *Statistica Sinica*, pp. 1319–1349, 2012.

[14] C. T. Tony and Z. H. H, "Optimal rates of convergence for sparse covariance matrix estimation," *The Annals of Statistics*, vol. 40, no. 5, pp. 2389–2420, 2012.

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

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

[17] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[18] Y. C. Eldar and G. Kutyniok, *Compressed sensing: theory and applications*. Cambridge university press, 2012.

[19] 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.

[20] E. J. Candes, T. Strohmer, and V. Voroninski, "Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming," *Communications on Pure and Applied Mathematics*, vol. 66, no. 8, pp. 1241–1274, 2013.

[21] S. A. van de Geer and P. Bühlmann, "On the conditions used to prove oracle results for the Lasso," *Electronic Journal of Statistics*, vol. 3, no. none, pp. 1360 – 1392, 2009.

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

[23] 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.

[24] D. Romero, D. D. Ariananda, Z. Tian, and G. Leus, "Compressive Covariance Sensing: Structure-based compressive sensing beyond sparsity," *IEEE Signal Processing Magazine*, vol. 33, no. 1, pp. 78–93, 2016.

[25] 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.

[26] M. A. Lexa, M. E. Davies, J. S. Thompson, and J. Nikolic, "Compressive power spectral density estimation," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 3884–3887.

[27] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of lasso and dantzig selector," 2009.

[28] 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.

[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] G. Raskutti, M. J. Wainwright, and B. Yu, "Restricted eigenvalue properties for correlated gaussian designs," *The Journal of Machine Learning Research*, vol. 11, pp. 2241–2259, 2010.

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

[35] 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.

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

[37] 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.

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

[39] R. Krishnamurthi, A. Kumar, D. Gopinathan, A. Nayyar, and B. Qureshi, "An overview of iot sensor data processing, fusion, and analysis techniques," *Sensors*, vol. 20, no. 21, p. 6076, 2020.

[40] N. González-Prelcic and M. E. Domínguez-Jiménez, "Circular sparse rulers based on co-prime sampling for compressive power spectrum estimation," in *Proceedings of 2014 IEEE Global Communications Conference*. IEEE, 2014, pp. 3044–3050.

[41] 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.

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

[43] L. Armijo, "Minimization of functions having Lipschitz continuous first partial derivatives," *Pacific Journal of Mathematics*, vol. 16, no. 1, pp. 1–3, 1966.

[44] H. Liu, L. Wang, and T. Zhao, "Sparse covariance matrix estimation with eigenvalue constraints," *Journal of Computational and Graphical Statistics*, vol. 23, no. 2, pp. 439–459, 2014.

[45] 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.

[46] X. Wang and H. V. Poor, "Blind equalization and multiuser detection in dispersive cdma channels," *IEEE Transactions on Communications*, vol. 46, no. 1, pp. 91–103, 1998.

[47] P. Campisi and K. Egiazarian, *Blind image deconvolution: theory and applications*. CRC press, 2017.

[48] H. Bristow, A. Eriksson, and S. Lucey, "Fast convolutional sparse coding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 391–398.

[49] I. S. Dhillon, C.-J. Hsieh, M. A. Sustik, and P. Ravikumar, "Sparse inverse covariance matrix estimation using quadratic approximation," in *Symposium on Machine Learning in Speech and Language Processing*, 2011.

[50] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, p. 1, 2010.

[51] T. T. Wu and K. L. Lange, "Coordinate descent algorithms for lasso penalized regression," *The Annals of Applied Statistics*, vol. 2, pp. 224–244, 2008.

[52] Y. Ning, T. Zhao, and A. Liu, "A likelihood ratio framework for high-dimensional semiparametric regression," *Annals of Statistics*, vol. 45, no. 6, pp. 2299–2327, Dec. 2017.

[53] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani, "Strong Rules for Discarding Predictors in Lasso-Type Problems," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 74, no. 2, pp. 245–266, 11 2011.
[54] T. Zhao, H. Liu, and T. Zhang, "Pathwise coordinate optimization for sparse learning: Algorithm and theory," *The Annals of Statistics*, vol. 46, no. 1, pp. 180 – 218, 2018.
[55] 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.
[56] 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.
[57] 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.
[58] I. E.-H. Yen, C.-J. Hsieh, P. K. Ravikumar, and I. S. Dhillon, "Constant Nullspace Strong Convexity and Fast Convergence of Proximal Methods under High-Dimensional Settings," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.

APPENDIX A
RESTRICTED EIGENVALUE & SPARSE EIGENVALUE

We first introduce an equivalent definition of definition 3.

Definition 19. For three positive integers s_0, ϑ, r, the largest and smallest localized restricted eigenvalues are defined as

$$\psi^{+}_{s_0,\vartheta,r} = \sup\left\{\frac{\mathbf{u}^\top \nabla^2 f(\mathbf{\Sigma})\,\mathbf{u}}{\mathbf{u}^\top\mathbf{u}} \mid (\mathbf{u},\mathbf{\Sigma}) \in \mathcal{I}(s_0,\vartheta,r)\right\};$$
$$\psi^{-}_{s_0,\vartheta,r} = \inf\left\{\frac{\mathbf{u}^\top \nabla^2 f(\mathbf{\Sigma})\,\mathbf{u}}{\mathbf{u}^\top\mathbf{u}} \mid (\mathbf{u},\mathbf{\Sigma}) \in \mathcal{I}(s_0,\vartheta,r)\right\},$$

where $\mathcal{I}(s_0,\vartheta,r)$ is defined as

$$\left\{(\mathbf{u},\mathbf{\Sigma}) \middle| \begin{array}{l} \mathcal{S}^* \subseteq \mathcal{J}, |\mathcal{J}| \le s_0, \\ \left\|\mathbf{u}_{\overline{\mathcal{J}}}\right\|_1 \le \vartheta \left\|\mathbf{u}_{\mathcal{J}}\right\|_1, \left\|\mathbf{\Sigma}-\mathbf{\Sigma}^*\right\|_{\mathsf{F}} \le r \end{array}\right\},$$

which represents a local ℓ_1 cone.

We now proceed to introduce the relationships between SE and LRE.

Proposition 20. *For any $\mathbf{\Sigma} \in \mathcal{I}(s,\vartheta,r) \cap \mathcal{B}(\mathbf{\Sigma}^*, r)$, the following inequalities hold:*

$$c_1\psi^{-}_{s,\vartheta,r} \le \rho^{-}_s \le c_2\psi^{-}_{s,\vartheta,r},$$
$$c_3\psi^{+}_{s,\vartheta,r} \le \rho^{+}_s \le c_4\psi^{+}_{s,\vartheta,r},$$

where c_1, c_2, c_3 and c_4 are constant.

For a detailed proof of Proposition 20, readers are referred to the foundational work in [34], which provides comprehensive mathematical derivations and discussions. This reference is omitted here for brevity.

APPENDIX B
COMPUTING THE NEWTON DIRECTION WITH ACTIVE SET STRATEGY

In this section, we elaborate on the computation of the Newton Direction within the context of our optimization framework. To simplify the exposition, we omit the stage index k and concentrate on a specific Newton iteration. The gradient and Hessian for $f(\mathbf{\Sigma})$ are given by

$$\nabla f(\mathbf{\Sigma}) = -\frac{1}{m}\mathcal{A}^*\left(\boldsymbol{y} - \mathcal{A}(\mathbf{\Sigma})\right) - \tau\mathbf{\Sigma}^{-1},$$

and

$$\nabla^2 f(\mathbf{\Sigma}) = \frac{1}{m}\sum_{i=1}^{m}\left(\mathbf{A}_i \otimes \mathbf{A}_i\right) + \tau\left(\mathbf{\Sigma}^{-1}\otimes\mathbf{\Sigma}^{-1}\right).$$

The direct computation of the Hessian matrix is computationally intensive, exhibiting a complexity of $\mathcal{O}\left(md^4\right)$, which render it impractical for high-dimensional data. However, the inherent symmetric structure of the Hessian permits certain computational optimizations. By utilizing a coordinate descent approach, it is possible to update all variables in $\mathcal{O}\left(md^3\right)$ time. This methodology has demonstrated efficacy in addressing Lasso-type problems [49]–[51] with theoretical justifications [52].

In order to precisely articulate the issue, we reformulate (10) as follows:

$$\begin{aligned} &\overline{f}(\mathbf{\Delta}) \\ =&f(\mathbf{\Sigma}_t) + \mathsf{tr}\left(\left(\mathbf{Q}-\tau\mathbf{W}\right)^\top\mathbf{\Delta}\right) \\ &+\frac{1}{2}\left(\frac{1}{m}\sum_{i=1}^{m}\mathsf{tr}\left(\mathbf{A}_i\mathbf{\Delta}\mathbf{A}_i\mathbf{\Delta}\right) + \tau\mathsf{tr}\left(\mathbf{W}\mathbf{\Delta}\mathbf{W}\mathbf{\Delta}\right)\right), \end{aligned} \tag{14}$$

where $\mathbf{Q} = -\frac{1}{m}\mathcal{A}^*\left(\boldsymbol{y}-\mathcal{A}(\mathbf{\Sigma})\right)$ and $\mathbf{W} = \mathbf{\Sigma}^{-1}$.

In the process of refining the Newton direction, we represent the existing approximation as $\mathbf{D}$ and the revised direction as $\mathbf{D}'$. When implementing a coordinate descent update for the variable Σ_{ij} (assuming $i<j$), the update is structured to preserve symmetry as follows:

$$\mathbf{D}' = \mathbf{D} + \mu\left(\boldsymbol{e}_i\boldsymbol{e}_j^\top + \boldsymbol{e}_j\boldsymbol{e}_i^\top\right),$$

where μ is the scalar update magnitude, and $\boldsymbol{e}_i$, $\boldsymbol{e}_j$ are the standard basis vectors. The resolution of the one-dimensional problem associated with the equation for (11) is expressed as follows:

$$\underset{\mu}{\operatorname{argmin}}\ \overline{f}\left(\mathbf{D}+\mu\left(\boldsymbol{e}_i\boldsymbol{e}_j^\top+\boldsymbol{e}_j\boldsymbol{e}_o^\top\right)\right) + 2\Lambda_{ij}\left|\Sigma_{ij}+D_{ij}+\mu\right|.$$

After substituting $\mathbf{D}' = \mathbf{D} + \mu\left(\boldsymbol{e}_i\boldsymbol{e}_j^\top + \boldsymbol{e}_j\boldsymbol{e}_i^\top\right)$ for $\mathbf{\Delta}$ in (14), the function $\overline{f}$ is refined to focus solely on terms dependent on μ, neglecting others. The contributions to the function from three terms are:

- The linear term $\mathsf{tr}\left(\left(\mathbf{Q}-\tau\mathbf{W}\right)^\top\mathbf{\Delta}\right)$ contributes $2\mu\left(Q_{ij}-\tau W_{ij}\right)$,
- The penalty term gives $2\Lambda_{ij}\left|\Sigma_{ij}+D_{ij}+\mu\right|$,
- The quadratic term yields

$$\begin{aligned} &\frac{1}{m}\sum_{l=1}^{m}4\mu\left(\mathbf{A}_l\right)_{\cdot i}^\top\mathbf{D}\left(\mathbf{A}_l\right)_{\cdot j} + 2\mu^2\tau\left(W_{ij}^2 + W_{ii}W_{jj}\right) \\ &+\frac{1}{m}\sum_{l=1}^{m}2\mu^2\left(\left(\mathbf{A}_l\right)_{ij}^2 + \left(\mathbf{A}_l\right)_{ii}\left(\mathbf{A}_l\right)_{jj}\right) + 4\mu\tau\mathbf{W}_{\cdot i}^\top\mathbf{D}\mathbf{W}_{\cdot j}. \end{aligned}$$

Combining all contributions yields a quadratic function of μ, expressed as (15) and the closed-form solution for μ is given

by

$$\mu = -e + \mathcal{T}\left(e - g/h, \Lambda_{ij}/h\right), \tag{16}$$

where $\mathcal{T}(u, v) = \mathsf{sign}(u)\max\{|u| - v, 0\}$ is the soft-thresholding function, and

$$h = \tau\left(W_{ij}^2 + W_{ii}W_{jj}\right) + \frac{\sum_{l=1}^m \left((\mathbf{A}_l)_{ij}^2 + (\mathbf{A}_l)_{ii}(\mathbf{A}_l)_{jj}\right)}{m},$$
$$g = Q_{ij} - \tau W_{ij} + \tau \mathbf{W}_{\cdot i}^\top \mathbf{D}\mathbf{W}_{\cdot j} + \frac{\sum_{l=1}^m (\mathbf{A}_l)_{\cdot i}^\top \mathbf{D}(\mathbf{A}_l)_{\cdot j}}{m},$$
$$e = \Sigma_{ij} + D_{ij}.$$

When $i = j$, the substitution of $\mathbf{D}' = \mathbf{D} + \mu \boldsymbol{e}_i \boldsymbol{e}_i^\top$ yields the update formula for D_{ii} as derived from (16). The components of the update formula are defined as follows:

$$h = \tau D_{ii}^2 + \frac{\sum_{l=1}^m (\mathbf{A}_l)_{ii}^2}{m},$$
$$g = Q_{ii} - \tau W_{ii} + \tau \mathbf{W}_{\cdot i}^\top \mathbf{D}\mathbf{W}_{\cdot i} + \frac{\sum_{l=1}^m (\mathbf{A}_l)_{\cdot i}^\top \mathbf{D}(\mathbf{A}_l)_{\cdot i}}{m},$$
$$e = \Sigma_{ii} + D_{ii}.$$

The primary computational challenge is encountered in the evaluation of the expressions $\mathbf{W}_{\cdot i}^\top \mathbf{D}\mathbf{W}_{\cdot j}$ and $\sum_{l=1}^m (\mathbf{A}_l)_{\cdot i}^\top \mathbf{D}(\mathbf{A}_l)_{\cdot j}$. To alleviate the computational burden, which typically necessitates $\mathcal{O}\left(md^2\right)$ time, we employ intermediary $d \times d$ matrices defined as $\mathbf{U}_l = \mathbf{D}\mathbf{A}_l$, $\mathbf{V} = \mathbf{D}\mathbf{W}$. This strategy facilitates the computation of $\mathbf{W}_{\cdot i}^\top \mathbf{V}_{\cdot j}$ and $\mathbf{W}_{\cdot j}^\top (\mathbf{A}_l)_{\cdot i}$ using $\mathcal{O}(d)$ floating-point operations. Furthermore, the maintenance $\mathbf{U}_l$ and $\mathbf{V}$ requires the adjustment of $\mathcal{O}(d)$ elements. For a comprehensive description of the methodology, please consult Algorithm 3.

Algorithm 3: Coordinate Descent Algorithms in Conjunction with the active set strategy.

Input: $\nabla f(\mathbf{\Sigma})$, $\mathbf{\Lambda}$, $\mathbf{\Sigma}$, $\{\boldsymbol{a}_l\}_{l=1}^m$
Initialize $\mathbf{W} = \mathbf{\Sigma}^{-1}$, $\mathbf{A}_l = \boldsymbol{a}_l \boldsymbol{a}_l^\top$, $\mathbf{D} = \mathbf{0}$, $\mathbf{U}_l = \mathbf{0}$, $\mathbf{V} = \mathbf{0}$
while *not converged* **do**
  `// Partition the variables into fixed and free sets`
  $\mathcal{C}_{\mathsf{fixed}} = \{(i,j) \mid |\nabla f_{ij}(\mathbf{\Sigma})| \le \Lambda_{ij} \text{ and } \Sigma_{ij} = 0\}$
  $\mathcal{C}_{\mathsf{free}} = \{(i,j) \mid |\nabla f_{ij}(\mathbf{\Sigma})| \le \Lambda_{ij} \text{ and } \Sigma_{ij} \neq 0\}$
  for $(i,j) \in \mathcal{A}_{\mathsf{free}}$ **do**
    $h = \tau\left(W_{ij}^2 + W_{ii}W_{jj}\right) + \frac{\sum_{l=1}^m \left((\mathbf{A}_l)_{ij}^2 + (\mathbf{A}_l)_{ii}(\mathbf{A}_l)_{jj}\right)}{m}$
    $g = \nabla_{ij} f(\mathbf{\Sigma}) + \tau \mathbf{W}_{\cdot i}^\top \mathbf{V}_{\cdot j} + \frac{\sum_{l=1}^m (\mathbf{A}_l)_{\cdot i}^\top (\mathbf{U}_l)_{\cdot j}}{m}$
    $e = \Sigma_{ij} + D_{ij}$
    $\mu = -e + \mathcal{T}\left(e - g/h, \mathbf{\Lambda}_{jk}/h\right)$
    $D_{ij} = D_{ij} + \mu$
    $(\mathbf{U}_l)_{i\cdot} = (\mathbf{U}_l)_{i\cdot} + \mu(\mathbf{A}_l)_{j\cdot}$
    $(\mathbf{U}_l)_{j\cdot} = (\mathbf{U}_l)_{j\cdot} + \mu(\mathbf{A}_l)_{i\cdot}$
    $\mathbf{V}_{i\cdot} = \mathbf{V}_{i\cdot} + \mu \mathbf{W}_{j\cdot}$, $\mathbf{V}_{j\cdot} = \mathbf{V}_{j\cdot} + \mu \mathbf{W}_{i\cdot}$
  end
end
Output: $\mathbf{D}$

A. Updating Only a Subset of Variables

We introduce an algorithm that leverages the active set update strategy to exploit the solution sparsity, thereby accelerating computation. The strategy involves two consecutive nested loops. In the outer loop, all coordinates are categorized into two sets based on a heuristic coordinate gradient thresholding rule (strong rule, [53]): the free set (active coordinates) and the fixed set (inactive coordinates). Subsequently, each iteration of the outer loop initiates an inner loop that performs coordinate optimization exclusively on the active set until convergence, while the inactive coordinates remain fixed at zero. Post each inner loop iteration, the algorithm employs a heuristic rule to reassess and update the active set, aiming to further reduce the objective value. This iterative process continues until the composition of the active set stabilizes across subsequent outer loop iterations. The efficacy of the active set strategy is well-documented in practical applications [50] and is underpinned by strong theoretical foundations [54].

The free set and fixed set are respectively defined as follows:

$$\Sigma_{ij} \in \mathcal{C}_{\mathsf{fixed}} \text{ if } |\nabla f_{ij}(\mathbf{\Sigma})| \le \Lambda_{ij} \text{ and } \Sigma_{ij} = 0,$$
$$\Sigma_{ij} \in \mathcal{C}_{\mathsf{free}} \text{ if } |\nabla f_{ij}(\mathbf{\Sigma})| > \Lambda_{ij} \text{ and } \Sigma_{ij} \neq 0.$$

Further, we ascertain that a Newton update confined to the variables within the fixed set does not alter any coordinates in that set. For a detailed discussion on this analysis, we refer readers to the work by [50], [55].

Remark 21. The computation of the Newton direction is streamlined when $\mathbf{\Sigma}$ is diagonal, which is encountered in the initial Newton iteration, wherein the Algorithm is started with an identity matrix. This simple scenario has a succinct closed-form optimal solution presented in (17). On this occasion, it only takes $\mathcal{O}(1)$ time to determine each variable, so the time complexity for solving the first Newton direction drops sharply to $\mathcal{O}\left(md^2\right)$. The core cause of this occurrence lies in the fact that, in this case, the Hessian matrix is diagonal, indicating that each one-variable sub-problem is detached from

$$\frac{1}{2}\left(\frac{\sum_{l=1}^m \left((\mathbf{A}_l)_{ij}^2 + (\mathbf{A}_l)_{ii}(\mathbf{A}_l)_{jj}\right)}{m} + \tau\left(W_{ij}^2 + W_{ii}W_{jj}\right)\right)\mu^2 + \left(Q_{ij} - \tau W_{ij} + \tau \mathbf{W}_{\cdot i}^\top \mathbf{D}\mathbf{W}_{\cdot j} + \frac{\sum_{l=1}^m (\mathbf{A}_l)_{\cdot i}^\top \mathbf{D}(\mathbf{A}_l)_{\cdot j}}{m}\right)\mu + 2\Lambda_{ij}|\Sigma_{ij} + D_{ij} + \mu| \tag{15}$$

$$\widehat{D}_{ij}=\begin{cases}\mathcal{T}\left(-\frac{Q_{ij}}{\tau W_{ii}W_{jj}+\sum_{l=1}^{m}\left((\mathbf{A}_l)_{ij}^2+(\mathbf{A}_l)_{ii}(\mathbf{A}_l)_{jj}\right)},\frac{\Lambda_{ij}}{\tau W_{ii}W_{jj}+\sum_{l=1}^{m}\left((\mathbf{A}_l)_{ij}^2+(\mathbf{A}_l)_{ii}(\mathbf{A}_l)_{jj}\right)}\right) & \text{if } i\neq j,\\ -\Sigma_{ii}+\mathcal{T}\left(\Sigma_{ii}-\frac{Q_{ii}-\tau W_{ii}}{\tau W_{ii}^2+\sum_{l=1}^{m}(\mathbf{A}_l)_{ii}^2},\frac{\Lambda_{ii}}{\tau W_{ii}^2+\sum_{l=1}^{m}(\mathbf{A}_l)_{ii}^2}\right) & \text{if } i=j.\end{cases} \tag{17}$$

the others. This means that updating each variable individually once suffices to achieves the optimal solution.

APPENDIX C
PROOF OF STATISTICAL THEORY

In this appendix, we provide the proofs of all the statistical theoretical results in Section IV-B. In Subsection C-A, we begin by Lemma 22 that connects the SE condition to the localized version of the sparse strong convexity/sparse strong smoothness. Subsequently, we introduce several preliminary lemmas that support Lemma 24. Following this, we present Lemma 25, which sets bounds on the estimation error for the general problem 9, followed by Lemma 26 that outlines the estimation error limited for approximate solutions derived using the MM-based algorithm. Subsection C-B elaborates on Theorem 12, leveraging Lemma 25 and 26 to illustrate the solution path's contraction characteristic. Subsection C-C focuses on key concentration inequalities, which are vital for deducing the explicit statistical convergence rate. In Subsections C-D and C-E, the document presents the proofs for Corollary 14 and Corollary 15, respectively.

A. Technical Lemmata

Lemma 22. *Let $f(\cdot)$ be a convex differentiable function, and let $G_f(\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)$ be defined as*

$$G_f(\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)=f(\mathbf{\Sigma}_1)-f(\mathbf{\Sigma}_2)-\langle\nabla f(\mathbf{\Sigma}_2),\mathbf{\Sigma}_1-\mathbf{\Sigma}_2\rangle.$$

Define the symmetrized Bregman divergence for $f(\cdot)$ as

$$\begin{aligned}G_f^s(\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)&=G_f(\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)+G_f(\mathbf{\Sigma}_2,\mathbf{\Sigma}_1)\\&=\langle\nabla f(\mathbf{\Sigma}_1)-\nabla f(\mathbf{\Sigma}_2),\mathbf{\Sigma}_1-\mathbf{\Sigma}_2\rangle.\end{aligned}$$

For any $\mathbf{\Sigma}_1,\mathbf{\Sigma}_2\in\mathcal{B}(\mathbf{\Sigma}^*,r)$ such that

$$\max\left\{\|(\mathbf{\Sigma}_1)_{\overline{\mathcal{S}^*}}\|_0,\|(\mathbf{\Sigma}_2)_{\overline{\mathcal{S}^*}}\|_0\right\}\leq\widetilde{s},$$

the following holds:

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

Proof: According to the mean value theorem and the convexity of the local region $\mathcal{B}(\mathbf{\Sigma}^*,r)$, there exists parameters $\theta_1,\theta_2,\theta_3\in[0,1]$ such that

$$\begin{aligned}\widetilde{\mathbf{\Sigma}}&=\theta_1\mathbf{\Sigma}_1+(1-\theta_1)\mathbf{\Sigma}_2\in\mathcal{B}(\mathbf{\Sigma}^*,r),\\ \widetilde{\mathbf{\Sigma}}'&=\theta_2\widetilde{\mathbf{\Sigma}}+(1-\theta_2)\mathbf{\Sigma}_2\in\mathcal{B}(\mathbf{\Sigma}^*,r),\\ \widetilde{\mathbf{\Sigma}}''&=\theta_3\mathbf{\Sigma}_1+(1-\theta_3)\mathbf{\Sigma}_2\in\mathcal{B}(\mathbf{\Sigma}^*,r),\end{aligned}$$

and satisfies $\max\left\{\left\|\widetilde{\mathbf{\Sigma}}\right\|_0,\left\|\widetilde{\mathbf{\Sigma}}'\right\|_0\right\}\leq 2\widetilde{s}$. The function $G_f(\mathbf{\Sigma}_1,\mathbf{\Sigma}_2)$ and the symmetrized Bregman divergence can be articulated as follows:

$$\begin{aligned}&f(\mathbf{\Sigma}_1)-f(\mathbf{\Sigma}_2)-\langle\nabla f(\mathbf{\Sigma}_2),\mathbf{\Sigma}_1-\mathbf{\Sigma}_2\rangle\\&=\left\langle\nabla f\left(\widetilde{\mathbf{\Sigma}}\right)-\nabla f(\mathbf{\Sigma}_2),\mathbf{\Sigma}_1-\mathbf{\Sigma}_2\right\rangle=\|\mathbf{\Sigma}_1-\mathbf{\Sigma}_2\|^2_{\nabla^2 f(\widetilde{\mathbf{\Sigma}}')},\end{aligned}$$

and

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

By applying the definition of the sparse eigenvalue, we arrive at the desired conclusion. ∎

Lemma 23. *Consider $\mathbf{\Sigma}(\theta)=\mathbf{\Sigma}^*+\theta(\mathbf{\Sigma}-\mathbf{\Sigma}^*)$ for $\theta\in(0,1]$. Then, we have*

$$G_f^s(\mathbf{\Sigma}(\theta),\mathbf{\Sigma}^*)\leq\theta G_f^s(\mathbf{\Sigma},\mathbf{\Sigma}^*).$$

Proof: Define $\varphi(\theta)$ as follows:

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

The derivative of $f(\mathbf{\Sigma}(\theta))$ with respect to θ is $\langle\nabla f(\mathbf{\Sigma}(\theta)),\mathbf{\Sigma}-\mathbf{\Sigma}^*\rangle$. Thus

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

Consequently, the symmetric Bregman divergence $G_f^s(\mathbf{\Sigma}(\theta),\mathbf{\Sigma}^*)$ is

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

for $0<\theta\leq 1$. Given $\varphi'(1)=G_f^s(\mathbf{\Sigma},\mathbf{\Sigma}^*)$, and assuming the convexity of $\varphi(\theta)$ based on the convexity of $f(\cdot)$ and the linearity of $\mathbf{\Sigma}(\theta)$, $\varphi'(\theta)$ is non-decreasing. Thus

$$G_f^s(\mathbf{\Sigma}(\theta),\mathbf{\Sigma}^*)=\theta\varphi'(\theta)\leq\theta\varphi'(1)=\theta G_f^s(\mathbf{\Sigma},\mathbf{\Sigma}^*),$$

completing the proof. For a more detailed proof of the convexity of $\varphi(\theta)$, see [28], [56]. ∎

Lemma 24. *Consider a set $\mathcal{E}$ such that $\mathcal{S}^*\subseteq\mathcal{E}$. Given the condition $\|\nabla f(\mathbf{\Sigma}^*)\|_{\max}+\varepsilon\leq\|\mathbf{\Lambda}_{\mathcal{E}}\|_{\min}$, 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

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

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

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

Given $\left\|\widetilde{\boldsymbol{\Delta}}\right\|_{H(\theta)}^2 \geq 0$, it follows that:

$$0 \leq \|\boldsymbol{\Omega}\|_{\max}\left\|\widetilde{\boldsymbol{\Delta}}\right\|_1 - \underbrace{\left\langle \nabla f\left(\boldsymbol{\Sigma}^*\right), \widetilde{\boldsymbol{\Delta}}\right\rangle}_{\mathbf{I}} - \underbrace{\left\langle \boldsymbol{\Lambda}\odot\widetilde{\boldsymbol{\Xi}}, \widetilde{\boldsymbol{\Delta}}\right\rangle}_{\mathbf{II}}. \quad (18)$$

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

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

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

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

Plugging the above results into 18 yields

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

where (i) is from $\|\boldsymbol{\Lambda}_{\overline{\mathcal{E}}}\|_{\min} \geq \frac{\lambda}{2}$, Assumption 10 and $\omega_{\boldsymbol{\Lambda}}\left(\widetilde{\boldsymbol{\Sigma}}\right) \leq \varepsilon$. ∎

Lemma 25. *Suppose that Assumption 7, 8, 9 and 10 hold. Consider the general problem in* (9)*. Let 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 the ε-optimal solution $\widetilde{\boldsymbol{\Sigma}}$ satisfies

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

Proof: Given the assumption 10 , it follows that $\lambda \geq 2\left(\|\nabla f\left(\boldsymbol{\Sigma}^*\right)\|_{\max} + \varepsilon\right)$. We introduce an intermediate estimator $\widetilde{\boldsymbol{\Sigma}}^* = \boldsymbol{\Sigma}^* + \theta\left(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\right)$, where θ is chosen to ensure $\left\|\widetilde{\boldsymbol{\Sigma}}^* - \boldsymbol{\Sigma}^*\right\|_{\mathsf{F}} = r$ if $\left\|\widetilde{\boldsymbol{\Sigma}}^* - \boldsymbol{\Sigma}^*\right\|_{\mathsf{F}} > r$; otherwise $\theta = 1$. This construction guarantees that $\left\|\widetilde{\boldsymbol{\Sigma}}^* - \boldsymbol{\Sigma}^*\right\|_{\mathsf{F}} \leq r$. By leveraging Lemma 24, we deduce that the approximate solution is contained within the ℓ_1-cone. Given the structure of $\widetilde{\boldsymbol{\Sigma}}^*$, we find that $\widetilde{\boldsymbol{\Sigma}}^* - \boldsymbol{\Sigma}^* = \theta\left(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\right)$. This leads to the following inequality

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

By integrating the aforementioned inequality with the premise that $|\mathcal{E}| \leq 2s^*$, it follows that $\widetilde{\boldsymbol{\Sigma}}^*$ resides within the local ℓ_1-cone. Furthermore, by synthesizing Lemma 22, Definition 19 and Proposition 20, one can deduce the presence of localized restricted strong convexity, i.e.

$$\rho^-_{2s^*+2\widetilde{s}}\left\|\widetilde{\boldsymbol{\Sigma}}^* - \boldsymbol{\Sigma}^*\right\|_{\mathsf{F}}^2 \leq G_f^s\left(\widetilde{\boldsymbol{\Sigma}}^*, \boldsymbol{\Sigma}^*\right). \quad (19)$$

We use Lemma 23 to bound the right hand side of the above inequality such as

$$\begin{aligned}
G_f^s\left(\widetilde{\boldsymbol{\Sigma}}^*, \boldsymbol{\Sigma}^*\right) \leq& \theta G_f^s\left(\widetilde{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}^*\right) \\
=& \theta\left\langle \nabla f\left(\widetilde{\boldsymbol{\Sigma}}\right) - \nabla f\left(\boldsymbol{\Sigma}^*\right), \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\right\rangle.
\end{aligned} \quad (20)$$

Plugging (20) back into (19) yields

$$\begin{aligned}
\rho^-_{2s^*+2\widetilde{s}}\left\|\widetilde{\boldsymbol{\Sigma}}^* - \boldsymbol{\Sigma}^*\right\|_{\mathsf{F}}^2 \leq& \theta\left\langle \nabla f\left(\widetilde{\boldsymbol{\Sigma}}\right) - \nabla f\left(\boldsymbol{\Sigma}^*\right), \widetilde{\boldsymbol{\Delta}}\right\rangle \\
=& \theta\underbrace{\left\langle \nabla f\left(\widetilde{\boldsymbol{\Sigma}}\right) + \boldsymbol{\Lambda}\odot\widetilde{\boldsymbol{\Xi}}, \widetilde{\boldsymbol{\Delta}}\right\rangle}_{\mathbf{I}} \\
&- \theta\underbrace{\left\langle \nabla f\left(\boldsymbol{\Sigma}^*\right), \widetilde{\boldsymbol{\Delta}}\right\rangle}_{\mathbf{II}} - \theta\underbrace{\left\langle \boldsymbol{\Lambda}\odot\widetilde{\boldsymbol{\Xi}}, \widetilde{\boldsymbol{\Delta}}\right\rangle}_{\mathbf{III}},
\end{aligned} \quad (21)$$

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

Define $\boldsymbol{\Omega} = \nabla f\left(\widetilde{\boldsymbol{\Sigma}}\right) + \boldsymbol{\Lambda}\odot\widetilde{\boldsymbol{\Xi}}$. For term $\mathbf{I}$, we partition the support of $\boldsymbol{\Omega}$ and $\widetilde{\boldsymbol{\Delta}}$ into the set $\mathcal{E}$ and its complement $\overline{\mathcal{E}}$. By employing the matrix Hölder's inequality, we derive

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

where (i) is from $\|\boldsymbol{\Omega}\|_{\max} \leq \varepsilon$ by Definition 6. For term $\mathbf{II}$, we divide the support of $\nabla f\left(\boldsymbol{\Sigma}^*\right)$ and $\widetilde{\boldsymbol{\Delta}}$ into the set $\mathcal{E}$ and $\overline{\mathcal{E}}$. Applying the matrix Hölder's inequality, we obtain

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

For term $\mathbf{III}$, we separate the support of $\boldsymbol{\Lambda}\odot\widetilde{\boldsymbol{\Xi}}$ and $\widetilde{\boldsymbol{\Delta}}$ into the set $\mathcal{S}^*$ and $\overline{\mathcal{S}^*}$. And then applying the matrix Hölder's

inequality, we obtain

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

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

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

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 (21) yields

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

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

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

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

which serves as a contraction in relation to the construction of $\widetilde{\mathbf{\Sigma}}^*$. This leads to the conclusion that $\widetilde{\mathbf{\Sigma}}^* = \widetilde{\mathbf{\Sigma}}$. Therefore, the desired bound is satisfied for $\widetilde{\mathbf{\Sigma}}$. ∎

Lemma 26. *Suppose that Assumptions 1 and 8 hold. Consider the problem in* (9)*. Define the set* $\mathcal{E}^{(k)} = \mathcal{S}^* \cup \mathcal{S}^{(k)}$*, where* $\mathcal{S}^{(k)} = \left\{(i,j) \mid \mathbf{\Lambda}^{(k-1)}_{ij} \leq p'_\lambda(u)\right\}$ *with* $u = c\lambda$ *and* $c = \frac{2+\sqrt{2}}{2\rho^-_{2s^*+2\widetilde{s}}} \in (0, \alpha)$*, such that* $p'_\lambda(c\lambda) \geq \frac{\lambda}{2}$*. This can always hold due to Assumption 5. Then for* $k \geq 1$*, we establish that:*

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

Proof: We first show that $\left|\mathcal{E}^{(k)}\right| \leq 2s^*$ for all k by induction on k.

Base Case $(k = 1)$**:**

For $k = 1$, we have $\Lambda^{(0)}_{ij} = \lambda$ for $i \neq j$, where λ is a threshold parameter satisfying $\lambda \geq p'_\lambda(u)$. It follows that $\left|\mathcal{S}^{(1)}\right| \leq s^*$. And we see $\mathcal{E}^{(1)} = \mathcal{S}^* \cup \mathcal{S}^{(1)}$, which implies $\left|\mathcal{E}^{(1)}\right| \leq 2s^*$ holds.

Inductive Step $(k \geq 2)$**:**

Assume for some $k \geq 2$, $\left|\mathcal{E}^{(k-1)}\right| \leq 2s^*$ holds true. We aim to show that $\left|\mathcal{E}^{(k)}\right| \leq 2s^*$. For any $(i,j) \in \mathcal{S}^{(k)}$, we obtain $\left|\widetilde{\Sigma}^{(k-1)}_{ij}\right| \geq u$ and further have

$$\begin{aligned}\sqrt{\left|\mathcal{S}^{(k)} \backslash \mathcal{S}^*\right|} &\leq \sqrt{\sum_{(i,j)\in\mathcal{S}^{(k)}\backslash\mathcal{S}^*} \left(u^{-1}\widetilde{\Sigma}^{(k-1)}_{ij}\right)^2} \\ &= u^{-1} \left\|\widetilde{\mathbf{\Sigma}}^{(k-1)}_{\mathcal{S}^{(k)}\backslash\mathcal{S}^*}\right\|_{\mathsf{F}} \\ &= u^{-1} \left\|\left(\widetilde{\mathbf{\Sigma}}^{(k-1)} - \mathbf{\Sigma}^*\right)_{\mathcal{S}^{(k)}\backslash\mathcal{S}^*}\right\|_{\mathsf{F}} \\ &\leq u^{-1} \left\|\widetilde{\mathbf{\Sigma}}^{(k-1)} - \mathbf{\Sigma}^*\right\|_{\mathsf{F}}.\end{aligned} \tag{22}$$

For any $(i,j) \in \overline{\mathcal{S}^{(k-1)}}$, it follows that $\Lambda^{(k-2)}_{ij} = p'_\lambda\left(\widetilde{\Sigma}^{(k-2)}_{ij}\right) \geq p'_\lambda(u) \geq \frac{\lambda}{2}$. This chain of inequalities indicates that

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

Furthermore, it is established that $\left|\mathcal{E}^{(k-1)}\right| \leq 2s^*$ and $\mathcal{S}^* \subseteq \mathcal{E}^{(k-1)}$. By applying Lemma 25 with $\widetilde{\mathbf{\Sigma}} = \widetilde{\mathbf{\Sigma}}^{(k-1)}$, $\mathcal{E} = \mathcal{E}^{(k-1)}$, and $\mathbf{\Lambda}_{\mathcal{S}^*} = \mathbf{\Lambda}^{(k-2)}_{\mathcal{S}^*}$ yields

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

Substituting this result into the inequality 22 yields

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

Thus, we have

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

completing the induction process.

Then according to the definition of $\mathcal{E}^{(k)}$ and $\mathcal{S}^{(k)}$, it follows that

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

By employing Lemma 25 with $\widetilde{\mathbf{\Sigma}} = \widetilde{\mathbf{\Sigma}}^{(k)}$, $\mathcal{E} = \mathcal{E}^{(k)}$, and $\mathbf{\Lambda}_{\mathcal{S}^*} = \mathbf{\Lambda}^{(k-1)}_{\mathcal{S}^*}$, the ε-optimal solution $\widetilde{\mathbf{\Sigma}}^{(k)}$ to 9 satisfies

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

■

B. Proof of Theorem 12

Proof: Based on Lemma 26, we have

$$\begin{aligned}
\left\|\widetilde{\mathbf{\Sigma}}^{(k)} - \mathbf{\Sigma}^*\right\|_{\mathsf{F}} \leq& \frac{1}{\rho^-_{2s^*+2\widetilde{s}}}\underbrace{\left(\left\|\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)_{\mathcal{E}^{(k)}}\right\|_{\mathsf{F}} + \varepsilon\sqrt{\left|\mathcal{E}^{(k)}\right|}\right)}_{\mathbf{I}} \\
&+ \frac{1}{\rho^-_{2s^*+2\widetilde{s}}}\underbrace{\left\|\mathbf{\Lambda}^{(k-1)}_{\mathcal{S}^*}\right\|_{\mathsf{F}}}_{\mathbf{II}}.
\end{aligned} \tag{23}$$

Then, we proceed to establish bounds for the term $\mathbf{I}$ and $\mathbf{II}$, respectively.

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

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

where the second inequality is due to

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

which follows from the inequality (22).

By Assumption 5 and 9, for any $\mathbf{\Sigma}$, if $\left|\Sigma_{ij} - \Sigma^*_{ij}\right| \geq u$, then $p'_\lambda\left(\Sigma_{ij}\right) \leq \lambda \leq \lambda u^{-1}\left|\Sigma_{ij} - \Sigma^*_{ij}\right|$; otherwise, $p'_\lambda\left(\Sigma_{ij}\right) \leq p'_\lambda\left(\left|\Sigma^*_{ij}\right| - u\right) = 0$. Therefore, for term $\mathbf{II}$, we have

$$\mathbf{II} \leq \lambda u^{-1}\left\|\widetilde{\mathbf{\Sigma}}^{(k-1)}_{\mathcal{S}^*} - \mathbf{\Sigma}^*_{\mathcal{S}^*}\right\|_{\mathsf{F}} \leq \lambda u^{-1}\left\|\widetilde{\mathbf{\Sigma}}^{(k-1)} - \mathbf{\Sigma}^*\right\|_{\mathsf{F}}.$$

Substituting the above results into (23) yields

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

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

C. Concentration Inequality

Lemma 27 (Hanson-Wright Inequality)**.** *Let $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ be a random vector with independent components x_i which satisfy $\mathbb{E}(x_i) = 0$ and $\|x_i\|_{\psi_2} \leq L$. Let $\mathbf{B}$ be an $n \times n$ matrix. Then for any $t > 0$,*

$$\begin{aligned}
&\mathbb{P}\left\{\left|\mathbf{x}^\top\mathbf{B}\mathbf{x} - \mathbb{E}\left(\mathbf{x}^\top\mathbf{B}\mathbf{x}\right)\right| > t\right\} \\
\leq& 2\exp\left(-c'\min\left(\frac{t^2}{L^4\|\mathbf{B}\|_{\mathsf{F}}^2}, \frac{t}{L^2\|\mathbf{B}\|}\right)\right).
\end{aligned} \tag{24}$$

Lemma 28 (Lemma D.1 in [56])**.** *Let $\mathbf{x}$ be a sub-Gaussian random vector with zero mean and covariance $\mathbf{\Sigma}^*$ and $\{\mathbf{x}_i\}_{i=1}^n$ be a collection of i.i.d. samples from $\mathbf{x}$. There exists some constant c_1, c_2, and t_0 such that for all t with $0 < t < t_0$, the sample covariance matrix $\mathbf{S}$ satisfies the following tail bound*

$$\mathbb{P}\left(\left|\Sigma^*_{ij} - S_{ij}\right| > t\right) \leq c_1\exp\left(-c_2 n t^2\right).$$

Lemma 29. *Under Assumptions 9 and the same conditions in Lemma 28, the columns of $\widetilde{\mathbf{A}}$ are normalized such that $\max_l\left\|\widetilde{\mathbf{A}}_{\cdot l}\right\|_2 \leq \sqrt{m}$, there exists some constant c_1 such that*

$$\mathbb{P}\left(\left\|\frac{1}{m}\mathcal{A}^*\left(\boldsymbol{y} - \mathcal{A}\left(\mathbf{\Sigma}^*\right)\right)\right\|_{\max} \leq \lambda\right) \geq 1 - \frac{c_1}{d}.$$

Proof: We begin by establishing a bound on the probability that $\left\|\frac{1}{m}\mathcal{A}^*\left(\boldsymbol{y} - \mathcal{A}\left(\mathbf{\Sigma}^*\right)\right)\right\|_{\max} > t$ using a union bound approach.

First, consider the bound $\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^m \eta'_i\boldsymbol{a}_i\boldsymbol{a}_i^\top\right\|_{\max} > t\right)$, where $\eta'_i = \boldsymbol{a}_i^\top\left(\mathbf{\Sigma}_N - \mathbf{\Sigma}^*\right)\boldsymbol{a}_i + \eta_i$. By applying the Hanson-Wright inequality (24), we establish

$$\begin{aligned}
&\mathbb{P}\left(\boldsymbol{a}_i^\top\left(\mathbf{\Sigma}_N - \mathbf{\Sigma}^*\right)\boldsymbol{a}_i > t\right) \\
=&\mathbb{P}\left(\left|\left\langle\boldsymbol{a}_i\boldsymbol{a}_i^\top, \left(\mathbf{\Sigma}_N - \mathbf{\Sigma}^*\right)\right\rangle\right| > t\right) \\
\leq& 2\exp\left(-c'\min\left(\frac{t^2}{L^4\left\|\mathbf{\Sigma}_N - \mathbf{\Sigma}^*\right\|_{\mathsf{F}}^2}, \frac{t}{L^2\left\|\mathbf{\Sigma}_N - \mathbf{\Sigma}^*\right\|}\right)\right),
\end{aligned}$$

where c' is a positive constant, and L is a scaling factor. This indicates that $\left\langle\boldsymbol{a}_i\boldsymbol{a}_i^\top, \mathbf{\Sigma}_N - \mathbf{\Sigma}^*\right\rangle$ is a sub-exponential random variable. Additionally, given that η_i follows a sub-exponential distribution, η'_i is a also sub-exponential random variable. By summing over individual probabilities, we obtain

$$\begin{aligned}
&\mathbb{P}\left(\left\|\frac{1}{m}\mathcal{A}^*\left(\boldsymbol{y} - \mathcal{A}\left(\mathbf{\Sigma}^*\right)\right)\right\|_{\max} > t\right) \\
=&\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^m \eta'_i\boldsymbol{a}_i\boldsymbol{a}_i^\top\right\|_{\max} > t\right) \\
\leq&\sum_{l=1}^{d^2}\mathbb{P}\left(\frac{1}{m}\left|\widetilde{\mathbf{A}}_{\cdot l}^\top\cdot\boldsymbol{\eta}'\right| > t\right),
\end{aligned} \tag{25}$$

Define $\zeta_l = \widetilde{\mathbf{A}}_{\cdot l}^\top\cdot\boldsymbol{\eta}'$. Given that η'_i is sub-exponential$\left(0, \xi^2\right)$ for $i = 1, \ldots, m$, we apply Lemma 28 and employ concentration inequalities:

$$\mathbb{E}\left(\exp\left\{t_0\zeta_l\right\} + \exp\left\{-t_0\zeta_l\right\}\right) \leq 2\exp\left\{nm^{-2}\left\|\widetilde{\mathbf{A}}_{\cdot l}\right\|^2\xi^2 t_0^2/2\right\},$$

which implies

$$\mathbb{P}\left(|\zeta_l| \geq t\right) \exp\left\{t_0 t\right\} \leq 2\exp\left\{nm^{-2}\left\|\widetilde{\mathbf{A}}_{\cdot l}\right\|^2 \xi^2 t_0^2/2\right\}.$$

Selecting $t_0 = t\left(nm^{-2}\left\|\widetilde{\mathbf{A}}_{\cdot l}\right\|_2^2 \xi^2\right)^{-1}$ yields that

$$\mathbb{P}\left(|\zeta_l| \geq t\right) \leq 2\exp\left\{-nt^2 \cdot \left(2\xi^2\left\|\widetilde{\mathbf{A}}_{\cdot l}\right\|_2^2/m^2\right)^{-1}\right\}.$$

Plugging it into (25) results

$$\begin{aligned}
&\mathbb{P}\left(\left\|\frac{1}{m}\mathcal{A}^*\left(\boldsymbol{y}-\mathcal{A}\left(\boldsymbol{\Sigma}^*\right)\right)\right\|_{\max} > t\right)\\
\underset{(i)}{\leq}&2d^2\exp\left\{-nmt^2\cdot\left(2\xi^2\max_l\left\{\left\|\widetilde{\mathbf{A}}_{\cdot l}\right\|_2^2/m\right\}\right)^{-1}\right\}\\
=&c_1\exp\left\{-c_2 nmt^2+2\log d\right\},
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{P}\left(\left\|\frac{1}{m}\mathcal{A}^*\left(\boldsymbol{y}-\mathcal{A}\left(\boldsymbol{\Sigma}^*\right)\right)\right\|_{\max} \leq \lambda\right)\\
\geq&1-c_1\exp\left(-c_2 nm\lambda^2+2\log d\right)\\
=&1-\frac{c_1}{d}.
\end{aligned}$$

■

Lemma 30. *Under the same conditions as in Lemma 29, the following result hold*

$$\left\|\frac{1}{m}\mathcal{A}^*\left(\boldsymbol{y}-\mathcal{A}\left(\boldsymbol{\Sigma}^*\right)\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}} = \mathcal{O}_p\left(\sqrt{\frac{s^*}{mn}}\right).$$

Proof: Following the analysis presented in Lemma 29 analysis, consider any M such that $0 < M\sqrt{\frac{1}{mn}}$, we have

$$\begin{aligned}
&\mathbb{P}\left(\left\|\frac{1}{m}\mathcal{A}^*\left(\boldsymbol{y}-\mathcal{A}\left(\boldsymbol{\Sigma}^*\right)\right)_{\mathcal{S}^*}\right\|_{\max} > M\sqrt{\frac{1}{mn}}\right)\\
\leq&c_1 s^*\exp\left(-c_2 M\right)\\
=&c_1\exp\left(-c_2 M+\log s^*\right).
\end{aligned}$$

Setting M such that $\sqrt{\frac{2\log s^*}{c_3}} < M$ and letting $M\to\infty$, we obtain

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

The proof is completed by applying

$$\begin{aligned}
&\left\|\frac{1}{m}\mathcal{A}^*\left(\boldsymbol{y}-\mathcal{A}\left(\boldsymbol{\Sigma}^*\right)\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}}\\
\leq&\sqrt{s^*}\left\|\frac{1}{m}\mathcal{A}^*\left(\boldsymbol{y}-\mathcal{A}\left(\boldsymbol{\Sigma}^*\right)\right)_{\mathcal{S}^*}\right\|_{\max}.
\end{aligned}$$

■

D. Proof of Corollary 14

Proof: We begin by bound the gradient norm of $f\left(\boldsymbol{\Sigma}\right)$ at $\boldsymbol{\Sigma}^*$:

$$\begin{aligned}
\left\|\nabla f\left(\boldsymbol{\Sigma}^*\right)\right\|_{\max} \leq& \left\|\frac{1}{m}\mathcal{A}^*\left(\boldsymbol{y}-\mathcal{A}\left(\boldsymbol{\Sigma}^*\right)\right)\right\|_{\max}\\
&+\tau\left\|\left(\boldsymbol{\Sigma}^*\right)^{-1}\right\|_{\max}.
\end{aligned}$$

Assuming $\lambda \asymp \sqrt{\frac{\log d}{mn}}$ and $\tau \lesssim \sqrt{\frac{1}{mn}}\left\|\left(\boldsymbol{\Sigma}^*\right)^{-1}\right\|_{\max}^{-1}$, and invoking Lemma 29, we establish that $\lambda \geq 2\left(\left\|\nabla f\left(\boldsymbol{\Sigma}^*\right)\right\|_{\mathsf{max}}+\varepsilon\right)$ holds w.h.p.

Applying Lemma 26 with $k=1$, we obtain

$$\left\|\widetilde{\boldsymbol{\Sigma}}^{(1)}-\boldsymbol{\Sigma}^*\right\|_{\mathsf{F}} \leq \frac{2+\sqrt{2}}{2\rho^-_{2s^*+2\widetilde{s}}}\lambda\sqrt{s^*}. \tag{26}$$

Under the condition $\lambda \asymp \sqrt{\frac{\log d}{mn}}$, the Frobenius norm difference is asymptotically bounded as $\left\|\widetilde{\boldsymbol{\Sigma}}^{(1)}-\boldsymbol{\Sigma}^*\right\|_{\mathsf{F}} \lesssim \sqrt{\frac{s^*\log d}{mn}}$ w.h.p. ■

E. Proof of Corollary 15

Proof: We begin by bound the gradient norm of $f\left(\boldsymbol{\Sigma}\right)$ at $\boldsymbol{\Sigma}^*$:

$$\begin{aligned}
\left\|\nabla f\left(\boldsymbol{\Sigma}^*\right)\right\|_{\mathsf{max}} \leq& \left\|\frac{1}{m}\mathcal{A}^*\left(\boldsymbol{y}-\mathcal{A}\left(\boldsymbol{\Sigma}^*\right)\right)\right\|_{\mathsf{max}}\\
&+\tau\left\|\left(\boldsymbol{\Sigma}^*\right)^{-1}\right\|_{\mathsf{max}}.
\end{aligned}$$

Under Assumptions 7 and 9 condition, and with parameters λ and τ scaled as $\lambda \asymp \sqrt{\frac{\log d}{mn}}$ and $\tau \lesssim \sqrt{\frac{1}{mn}}\left\|\left(\boldsymbol{\Sigma}^*\right)^{-1}\right\|_{\max}^{-1}$, then by Lemma 29, $\lambda \geq 2\left(\left\|\nabla f\left(\boldsymbol{\Sigma}^*\right)\right\|_{\mathsf{max}}+\varepsilon\right)$ holds w.h.p.

Next, applying Theorem 12, we derive

$$\begin{aligned}
&\left\|\widetilde{\boldsymbol{\Sigma}}^{(1)}-\boldsymbol{\Sigma}^*\right\|_{\mathsf{F}}\\
\leq&\frac{1}{\rho^-_{2s^*+2\widetilde{s}}}\left(\left\|\left(\nabla f\left(\boldsymbol{\Sigma}^*\right)\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}}+\varepsilon\sqrt{s^*}\right)+\delta\left\|\widetilde{\boldsymbol{\Sigma}}^{(k-1)}-\boldsymbol{\Sigma}^*\right\|_{\mathsf{F}}\\
\leq&\frac{1}{\rho^-_{2s^*+2\widetilde{s}}}\left(\left\|\left(\nabla f\left(\boldsymbol{\Sigma}^*\right)\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}}+\varepsilon\sqrt{s^*}\right)+\delta^{k-1}\left\|\widetilde{\boldsymbol{\Sigma}}^{(1)}-\boldsymbol{\Sigma}^*\right\|_{\mathsf{F}}\\
\leq&\frac{1}{\rho^-_{2s^*+2\widetilde{s}}}\left(\left\|\left(\nabla f\left(\boldsymbol{\Sigma}^*\right)\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}}+\varepsilon\sqrt{s^*}\right)+\delta^{k-1}\frac{2+\sqrt{2}}{2\rho^-_{2s^*+2\widetilde{s}}}\lambda\sqrt{s^*},
\end{aligned} \tag{27}$$

where the last inequality is due to $\left\|\widetilde{\boldsymbol{\Sigma}}^{(1)}-\boldsymbol{\Sigma}^*\right\|_{\mathsf{F}} \leq \frac{2+\sqrt{2}}{2\rho^-_{2s^*+2\widetilde{s}}}\lambda\sqrt{s^*}$, which follows from Lemma 26 with $k=1$.

Finally, considering the norm of the projected gradient within the support set $\mathcal{S}^*$, we obtain

$$\begin{aligned}
&\left\|\left(\nabla f\left(\boldsymbol{\Sigma}^*\right)\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}}\\
=&\left\|\left(-\frac{1}{m}\mathcal{A}^*\left(\boldsymbol{y}-\mathcal{A}\left(\boldsymbol{\Sigma}^*\right)\right)-\tau\left(\boldsymbol{\Sigma}^*\right)^{-1}\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}}\\
\leq&\left\|\frac{1}{m}\mathcal{A}^*\left(\boldsymbol{y}-\mathcal{A}\left(\boldsymbol{\Sigma}^*\right)\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}}+\tau\left\|\left(\left(\boldsymbol{\Sigma}^*\right)^{-1}\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}}\\
\leq&\left\|\frac{1}{m}\mathcal{A}^*\left(\boldsymbol{y}-\mathcal{A}\left(\boldsymbol{\Sigma}^*\right)\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}}+\tau\left\|\left(\boldsymbol{\Sigma}^*\right)^{-1}\right\|_{\mathsf{F}}
\end{aligned} \tag{28}$$

By Lemma 30, $\left\|\frac{1}{m}\mathcal{A}^*\left(\boldsymbol{y}-\mathcal{A}\left(\boldsymbol{\Sigma}^*\right)\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}}=\mathcal{O}_p\left(\sqrt{\frac{s^*}{mn}}\right)$. If $\tau\lesssim\sqrt{\frac{1}{mn}}\left\|\left(\boldsymbol{\Sigma}^*\right)^{-1}\right\|_{\max}^{-1}$, then $\left\|\left(\nabla f\left(\boldsymbol{\Sigma}^*\right)\right)_{\mathcal{S}^*}\right\|_{\mathsf{F}}=\mathcal{O}_p\left(\sqrt{\frac{s^*}{mn}}\right)$.

Furthermore, if $K\geq 1+\frac{\log\left(\lambda\sqrt{mn}\right)}{\log\delta^{-1}}\gtrsim\log\left(\lambda\sqrt{mn}\right)\gtrsim\log\log d$, we derive

$$\delta^{K-1}\lambda\sqrt{s^*}\leq\frac{1}{\lambda\sqrt{mn}}\lambda\sqrt{s^*}\leq\sqrt{\frac{s^*}{mn}}.$$

By integrating all findings with the condition $\varepsilon\lesssim\sqrt{\frac{1}{mn}}$, we conclude that $\left\|\widetilde{\boldsymbol{\Sigma}}^{(K)}-\boldsymbol{\Sigma}^*\right\|_{\mathsf{F}}=\mathcal{O}_p\left(\sqrt{\frac{s^*}{mn}}\right)$. ∎

APPENDIX D
PROOF OF COMPUTATIONAL THEORY 18

In this appendix, we first show $f\left(\boldsymbol{\Sigma}\right)$ exhibits properties of the local strong convexity , local strong smoothness and local restricted Hessian smoothness. Then we demonstrate the sparsity of the solution is preserved within a vicinity of the true model parameter $\boldsymbol{\Sigma}^*$, provided that the starting point is a sparse solution. Following this, it is established that proximal Newton updates achieve a quadratic rate of convergence towards a local minimizer at each step within each stage, assuming the initialization occurs within a suitably refined sparse region. Subsequently, Lemmas 38 and 35 confirm that each subproblem experiences a substantial reduction. Furthermore, we outline the necessary number of iterations required at each convex relaxation stage to satisfy the approximate KKT conditions.

A. Proof of Lemma 16

Proof: We first verify the convexity of the set $\mathcal{B}\left(\boldsymbol{\Sigma}^*,\frac{\rho^-_{2s^*+2\widetilde{s}}}{4\tau\kappa}\right)$. For any $\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2\in\mathcal{B}\left(\boldsymbol{\Sigma}^*,\frac{\rho^-_{2s^*+2\widetilde{s}}}{4\tau\kappa}\right)$, the linearity of the space implies that any convex combination of these points also resides within the set due to the properties of the norm and the definition of the set.

To establish the sparse strong convexity of $f\left(\boldsymbol{\Sigma}\right)$ within this set, consider two arbitrary points $\boldsymbol{\Sigma}_3,\boldsymbol{\Sigma}_4$ and convex combination $\boldsymbol{\Sigma}_\rho=\rho\boldsymbol{\Sigma}_3+\left(1-\rho\right)\boldsymbol{\Sigma}_4$ for $\rho\in[0,1]$. The function $f\left(\boldsymbol{\Sigma}\right)$ can be expanded using a Taylor series approximation around $\boldsymbol{\Sigma}_4$:

$$\begin{aligned}f\left(\boldsymbol{\Sigma}_3\right)=&f\left(\boldsymbol{\Sigma}_4\right)+\left\langle\nabla f\left(\boldsymbol{\Sigma}_4\right),\boldsymbol{\Sigma}_3-\boldsymbol{\Sigma}_4\right\rangle\\&+\frac{1}{2}\mathsf{vec}^\top\left(\boldsymbol{\Sigma}_3-\boldsymbol{\Sigma}_4\right)\nabla^2 f\left(\boldsymbol{\Sigma}_\rho\right)\mathsf{vec}\left(\boldsymbol{\Sigma}_3-\boldsymbol{\Sigma}_4\right).\end{aligned}$$

The next step is to estimate the lower bound of $\lambda_{\min}\left(\nabla^2 f\left(\boldsymbol{\Sigma}_\rho\right)\right)$. Using the fact that $\lambda_{\min}\left(\mathbf{X}^{-1}\right)=\left\|\mathbf{X}\right\|_2^{-1}$ and $\left\|\mathbf{X}\otimes\mathbf{Y}\right\|_2=\left\|\mathbf{X}\right\|_2\left\|\mathbf{Y}\right\|_2$ for any matrices $\mathbf{X}$, $\mathbf{Y}$, we deduce

$$\begin{aligned}&\lambda_{\min}\left(\nabla^2 f\left(\boldsymbol{\Sigma}_\rho\right)\right)\\=&\lambda_{\min}\left(\sum_{i=1}^m\mathbf{A}_i\otimes\mathbf{A}_i^\top+\tau\boldsymbol{\Sigma}_\rho^{-1}\otimes\boldsymbol{\Sigma}_\rho^{-1}\right)\\\geq&\lambda_{\min}\left(\sum_{i=1}^m\mathbf{A}_i\otimes\mathbf{A}_i^\top\right)+\tau\lambda_{\min}\left(\boldsymbol{\Sigma}_\rho^{-1}\otimes\boldsymbol{\Sigma}_\rho^{-1}\right)\\\underset{(i)}{\geq}&\rho^-_{2s^*+2\widetilde{s}}+\tau\left\|\boldsymbol{\Sigma}_\rho\right\|_2^{-2},\end{aligned}$$

where (i) is from Assumption 4. One has

$$\begin{aligned}\left\|\boldsymbol{\Sigma}_\rho\right\|_2&=\left\|\boldsymbol{\Sigma}^*-\boldsymbol{\Sigma}^*+\rho\boldsymbol{\Sigma}_1+\left(1-\rho\right)\boldsymbol{\Sigma}_2\right\|_2\\&\leq\left\|\boldsymbol{\Sigma}^*\right\|_2+\rho\left\|\boldsymbol{\Sigma}_1-\boldsymbol{\Sigma}^*\right\|_2+\left(1-\rho\right)\left\|\boldsymbol{\Sigma}_2-\boldsymbol{\Sigma}^*\right\|_2\\&\leq\left\|\boldsymbol{\Sigma}^*\right\|_2+\rho\left\|\boldsymbol{\Sigma}_1-\boldsymbol{\Sigma}^*\right\|_{\mathsf{F}}+\left(1-\rho\right)\left\|\boldsymbol{\Sigma}_2-\boldsymbol{\Sigma}^*\right\|_{\mathsf{F}}\\&\underset{(i)}{\leq}\kappa+\frac{\rho^-_{2s^*+2\widetilde{s}}}{4\tau\kappa}\leq\kappa+\frac{\kappa\rho^-_{2s^*+2\widetilde{s}}}{4\tau},\end{aligned}$$

where (i) holds due to $\left\|\boldsymbol{\Sigma}^*\right\|_2\leq\kappa$ and $\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2\in\mathcal{B}\left(\boldsymbol{\Sigma}^*,\frac{\rho^-_{2s^*+2\widetilde{s}}}{4\tau\kappa}\right)$; the last inequality holds due to $\kappa\geq 1$. Thus, $f\left(\boldsymbol{\Sigma}\right)$ is shown to be strongly convex over $\mathcal{B}\left(\boldsymbol{\Sigma}^*,\frac{\rho^-_{2s^*+2\widetilde{s}}}{4\tau\kappa}\right)$ with a convexity parameter of $\left(\rho^-_{2s^*+2\widetilde{s}}+\frac{16\tau^3}{\kappa^2\left(4\tau+\rho^-_{2s^*+2\widetilde{s}}\right)^2}\right)$.

Next, we establish the strong smoothness of $f\left(\boldsymbol{\Sigma}\right)$, which involves setting an upper bound for $\lambda_{\max}\left(\nabla^2 f\left(\boldsymbol{\Sigma}_\rho\right)\right)$. By the fact that $\lambda_{\max}\left(\mathbf{X}^{-1}\right)=\lambda_{\min}^{-1}\left(\mathbf{X}\right)$ and $\lambda_{\min}\left(\mathbf{X}\otimes\mathbf{Y}\right)$ for any matrices $\mathbf{X}$, $\mathbf{Y}$. Therefore,

$$\begin{aligned}&\lambda_{\max}\left(\nabla^2 f\left(\boldsymbol{\Sigma}_\rho\right)\right)\\=&\lambda_{\max}\left(\sum_{i=1}^m\mathbf{A}_i\otimes\mathbf{A}_i^\top+\tau\boldsymbol{\Sigma}_\rho^{-1}\otimes\boldsymbol{\Sigma}_\rho^{-1}\right)\\\leq&\lambda_{\max}\left(\sum_{i=1}^m\mathbf{A}_i\otimes\mathbf{A}_i^\top\right)+\tau\lambda_{\max}\left(\boldsymbol{\Sigma}_\rho^{-1}\otimes\boldsymbol{\Sigma}_\rho^{-1}\right)\\\leq&\rho^+_{2s^*+2\widetilde{s}}+\tau\lambda_{\min}^{-2}\left(\boldsymbol{\Sigma}_\rho\right).\end{aligned}$$

One has

$$\begin{aligned}&\lambda_{\min}\left(\boldsymbol{\Sigma}_\rho\right)\\=&\lambda_{\min}\left(\boldsymbol{\Sigma}^*-\boldsymbol{\Sigma}^*+\rho\boldsymbol{\Sigma}_1+\left(1-\rho\right)\boldsymbol{\Sigma}_2\right)\\\geq&\lambda_{\min}\left(\boldsymbol{\Sigma}^*\right)+\rho\lambda_{\min}\left(\boldsymbol{\Sigma}_1-\boldsymbol{\Sigma}^*\right)+\left(1-\rho\right)\lambda_{\min}\left(\boldsymbol{\Sigma}_2-\boldsymbol{\Sigma}^*\right)\\\geq&\lambda_{\min}\left(\boldsymbol{\Sigma}^*\right)-\rho\lambda_{\max}\left(\boldsymbol{\Sigma}_1-\boldsymbol{\Sigma}^*\right)-\left(1-\rho\right)\lambda_{\max}\left(\boldsymbol{\Sigma}_2-\boldsymbol{\Sigma}^*\right)\\\geq&\lambda_{\min}\left(\boldsymbol{\Sigma}^*\right)-\rho\left\|\boldsymbol{\Sigma}_1-\boldsymbol{\Sigma}^*\right\|_{\mathsf{F}}-\left(1-\rho\right)\left\|\boldsymbol{\Sigma}_2-\boldsymbol{\Sigma}^*\right\|_{\mathsf{F}}\\\geq&\frac{1}{\kappa}-\frac{\rho^-_{2s^*+2\widetilde{s}}}{4\tau\kappa},\end{aligned}$$

utilizing the bounds given by $\lambda_{\min}\left(\boldsymbol{\Sigma}^*\right)\geq\frac{1}{\kappa}$ and $\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2\in\mathcal{B}\left(\boldsymbol{\Sigma}^*,\frac{\rho^-_{2s^*+2\widetilde{s}}}{4\tau\kappa}\right)$. Thus, $f\left(\boldsymbol{\Sigma}\right)$ is proven to be $\left(\rho^+_{2s^*+2\widetilde{s}}+\frac{16\tau^3\kappa^2}{\left(4\tau-\rho^-_{2s^*+2\widetilde{s}}\right)^2}\right)$-smooth over $\mathcal{B}\left(\boldsymbol{\Sigma}^*,\frac{\rho^-_{2s^*+2\widetilde{s}}}{4\tau\kappa}\right)$. ∎

B. Preliminary Lemmas

Lemma 31 (Local Restricted Hessian Smoothness)**.** *Recalled $\widetilde{s}$ introduced in Assumption 4. We assert the existence of constant $L_{2s^*+2\widetilde{s}}$ and r 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 $\left\|\mathbf{\Sigma}'_{\overline{\mathcal{S}}}\right\|_0 \leq \widetilde{s}$, the following inequality is satisfied:*

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

where $\mathbf{\Omega} = \left\{\mathbf{U} \mid \mathsf{supp}\left(\mathbf{U}\right) \subseteq \left(\mathsf{supp}\left(\mathbf{\Sigma}\right) \cup \mathsf{supp}\left(\mathbf{\Sigma}'\right)\right)\right\}$.

Proof: To establish the Lipschitz continuity of $\nabla^2 f(\mathbf{\Sigma})$, consider the third derivative $\nabla^3 f(\mathbf{\Sigma})$, which is expressed as $-2\tau\mathbf{\Sigma}^{-1} \otimes \mathbf{\Sigma}^{-1} \otimes \mathbf{\Sigma}^{-1}$. Employing a proof technique analogous to that utilized in Lemma 16, we have

$$\lambda_{\max}\left(\left|\nabla^3 f(\mathbf{\Sigma})\right|\right) \leq 2\tau\lambda_{\min}^{-3}(\mathbf{\Sigma}) \leq 2\tau\kappa^3.$$

This finding allows us to conveniently define the Lipschitz constant for the Hessian matrix, represented as $L_{2s^*+2\widetilde{s}}$, to be $2\tau\kappa^3$. ∎

Lemma 32. *Under the same condition of Lemma 26, 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 26 with $k=1$, we obtain

$$\left\|\widetilde{\mathbf{\Sigma}}^{(1)} - \mathbf{\Sigma}^*\right\|_{\mathsf{F}} \leq \frac{2+\sqrt{2}}{2\rho^-_{2s^*+2\widetilde{s}}}\lambda\sqrt{s^*}. \tag{29}$$

Moreover, by invoking Lemma 24 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 $\mathcal{S}^*$. Combining these results with 29, we conclude

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

Therefore, we obtain

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

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

$$\begin{aligned}&\left\langle \nabla f\left(\widetilde{\mathbf{\Sigma}}^{(1)}\right) - \nabla f\left(\mathbf{\Sigma}^*\right), \widetilde{\mathbf{\Sigma}}^{(1)} - \mathbf{\Sigma}^* \right\rangle \\ \leq &\left\|\nabla f\left(\widetilde{\mathbf{\Sigma}}^{(1)}\right) + \mathbf{\Lambda}\odot\widetilde{\mathbf{\Xi}}^{(1)}\right\|_{\max}\left\|\widetilde{\mathbf{\Sigma}}^{(1)} - \mathbf{\Sigma}^*\right\|_1 \\ &+ \left\|\mathbf{\Lambda}\odot\widetilde{\mathbf{\Xi}}^{(1)} + \nabla f\left(\mathbf{\Sigma}^*\right)\right\|_{\max}\left\|\widetilde{\mathbf{\Sigma}}^{(1)} - \mathbf{\Sigma}^*\right\|_1 \\ \leq &c_0\lambda\left\|\widetilde{\mathbf{\Sigma}}^{(1)} - \mathbf{\Sigma}^*\right\|_1 \leq \frac{c_1\lambda^2 s^*}{\rho^-_{2s^*+2\widetilde{s}}},\end{aligned}$$

for some constants c_0 and c_1. ∎

Lemma 33. *Given $\omega_{\mathbf{\Lambda}^{(k-1)}}\left(\widetilde{\mathbf{\Sigma}}^{(k)}\right) \leq \frac{\lambda}{8}$, we have that for all $t \geq 1$ at the $(k+1)$-th stage,*

$$\omega_{\mathbf{\Lambda}^{(k)}}\left(\mathbf{\Sigma}_t^{(k+1)}\right) \leq \frac{\lambda}{4}$$

and

$$F_{\mathbf{\Lambda}^{(k)}}\left(\mathbf{\Sigma}_t^{(k+1)}\right) \leq F_{\mathbf{\Lambda}^{(k)}}\left(\mathbf{\Sigma}^*\right) + \frac{\lambda}{4}\left\|\mathbf{\Sigma}_t^{(k)} - \mathbf{\Sigma}^*\right\|_1.$$

Proof: We begin by establishing the foundational inequality at the $(k+1)$-th stage. Note that $\mathbf{\Sigma}_0^{(k+1)} = \widetilde{\mathbf{\Sigma}}^{(k)}$. By definition, we have

$$\omega_{\mathbf{\Lambda}^{(k)}}(\mathbf{\Sigma}_0) = \min_{\mathbf{\Xi}\in\partial\|\mathbf{\Sigma}_0\|_1}\left\|\nabla f(\mathbf{\Sigma}_0) + \mathbf{\Lambda}^{(k)}\odot\mathbf{\Xi}\right\|_{\max}.$$

Using the triangle inequality, we can expand the above expression as follows:

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

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

Continuing, for any $t \geq 0$, consider

$$\mathbf{\Xi}_t = \operatorname*{arg\,min}_{\mathbf{\Xi}\in\partial\|\mathbf{\Sigma}_t\|_1}\left\|\nabla f(\mathbf{\Sigma}_t) + \mathbf{\Lambda}^{(k)}\odot\mathbf{\Xi}\right\|_{\max}.$$

By the convexity of F, we have

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

where (ii) is from the fact that $\left\|\nabla f(\mathbf{\Sigma}_t) + \mathbf{\Lambda}^{(k)}\odot\mathbf{\Xi}_t\right\|_{\max} \leq \frac{\lambda}{4}$. ∎

Lemma 34. *Suppose $\|(\mathbf{\Sigma}_t)_{\overline{\mathcal{S}}}\|_0 \leq \widetilde{s}$ and $\omega_{\mathbf{\Lambda}^{(k)}}(\mathbf{\Sigma}_t) \leq \frac{\lambda}{4}$. Then, there exists a generic constant c_1 such that*

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

Proof: The proof is analogous to that of Lemma 26. Therefore, we omit it for brevity. For more details, we refer readers to [28]. ∎

Lemma 35. *Recall $\mathbf{\Sigma}_{t+\frac{1}{2}}$ defined in* (11) *and δ_t as in* (12)*. Denote $\Delta\mathbf{\Sigma}_t = \mathbf{\Sigma}_{t+\frac{1}{2}} - \mathbf{\Sigma}_t$. Then we have*

$$\delta_t \leq -\|\Delta\mathbf{\Sigma}_t\|^2_{\nabla^2 f(\mathbf{\Sigma}_t)}.$$

Proof: We note $\Delta\mathbf{\Sigma}_t$ is the solution for $\overline{F}\left(\mathbf{\Sigma} \mid \mathbf{\Sigma}_t, \mathbf{\Lambda}^{(k-1)}\right)$. For any $\eta_t \in (0,1]$, we have formula

$$\begin{aligned}&\eta_t \left\langle \nabla f\left(\boldsymbol{\Sigma}_t\right), \Delta\boldsymbol{\Sigma}_t\right\rangle + \frac{\eta_t^2}{2}\left\|\Delta\boldsymbol{\Sigma}_t\right\|^2_{\nabla^2 f(\boldsymbol{\Sigma}_t)} + \left\|\boldsymbol{\Lambda}^{(k-1)} \odot \left(\boldsymbol{\Sigma}_t + \eta_t \Delta\boldsymbol{\Sigma}_t\right)\right\|_1 \\ \geq &\left\langle \nabla f\left(\boldsymbol{\Sigma}_t\right), \Delta\boldsymbol{\Sigma}_t\right\rangle + \frac{1}{2}\left\|\Delta\boldsymbol{\Sigma}_t\right\|^2_{\nabla^2 f(\boldsymbol{\Sigma}_t)} + \left\|\boldsymbol{\Lambda}^{(k-1)} \odot \left(\boldsymbol{\Sigma}_t + \Delta\boldsymbol{\Sigma}_t\right)\right\|_1.\end{aligned} \tag{30}$$

$$\begin{aligned}&\eta_t \left\langle \nabla f\left(\boldsymbol{\Sigma}_t\right), \Delta\boldsymbol{\Sigma}_t\right\rangle + \frac{\eta_t^2}{2}\left\|\Delta\boldsymbol{\Sigma}_t\right\|^2_{\nabla^2 f(\boldsymbol{\Sigma}_t)} + \eta_t\left\|\boldsymbol{\Lambda}^{(k-1)} \odot \left(\boldsymbol{\Sigma}_t + \eta_t \Delta\boldsymbol{\Sigma}_t\right)\right\|_1 + (1-\eta_t)\left\|\boldsymbol{\Lambda}^{(k-1)} \odot \boldsymbol{\Sigma}_t\right\|_1 \\ \geq &\left\langle \nabla f\left(\boldsymbol{\Sigma}_t\right), \Delta\boldsymbol{\Sigma}_t\right\rangle + \frac{1}{2}\left\|\Delta\boldsymbol{\Sigma}_t\right\|^2_{\nabla^2 f(\boldsymbol{\Sigma}_t)} + \left\|\boldsymbol{\Lambda}^{(k-1)} \odot \left(\boldsymbol{\Sigma}_t + \Delta\boldsymbol{\Sigma}_t\right)\right\|_1.\end{aligned} \tag{31}$$

(30). By the convexity of ℓ_1-norm, we have formula (31). Rearranging the terms, canceling the $(1-\eta_t)$ factor from both sides and let $\eta_t \to 1$, we obtain the desired inequality $\delta_t \leq -\left\|\Delta\boldsymbol{\Sigma}_t\right\|^2_{\nabla^2 f(\boldsymbol{\Sigma}_t)}$. ∎

C. Proof of Theorem 18

Lemma 36 (Sparsity Preserving Lemma)**.** *Suppose that Assumptions 8 and 9 hold. Given $\boldsymbol{\Sigma}_t^{(k)} \in \mathcal{B}\left(\boldsymbol{\Sigma}^*, r\right)$ and $\left\|\left(\boldsymbol{\Sigma}_t^{(k)}\right)_{\overline{\mathcal{S}}}\right\|_0 \leq \widetilde{s}$, there exists a generic constant c_1 such that*

$$\left\|\left(\boldsymbol{\Sigma}_{t+1}^{(k)}\right)_{\overline{\mathcal{S}}}\right\|_0 \leq \widetilde{s} \quad \textit{and} \quad \left\|\boldsymbol{\Sigma}_{t+1}^{(k)} - \boldsymbol{\Sigma}^*\right\|_{\mathsf{F}} \leq \frac{c_1 \lambda \sqrt{s^*}}{\rho^-_{2s^*+2\widetilde{s}}}.$$

Proof: To simplify notation, we omit the stage index (k). Considering $\boldsymbol{\Sigma}_{t+1}$ as the proximal Newton update, the system satisfies the following equation:

$$\mathsf{mat}\left(\nabla^2 f\left(\boldsymbol{\Sigma}_t\right)\mathsf{vec}\left(\boldsymbol{\Sigma}_{t+1}-\boldsymbol{\Sigma}_t\right)\right)+\nabla f\left(\boldsymbol{\Sigma}_t\right)+\boldsymbol{\Lambda}\odot\boldsymbol{\Xi}_{t+1}=\mathbf{0},$$

where $\boldsymbol{\Xi}_{t+1} \in \partial\left\|\boldsymbol{\Sigma}_{t+1}\right\|_1$.

From Lemma 25, under identical conditions, it holds that $\left\|\boldsymbol{\Lambda}_{\overline{\mathcal{E}}}\right\|_{\mathsf{min}} \geq \frac{\lambda}{2}$ and $|\mathcal{E}| \leq 2s^*$ for some set $\mathcal{E} \supseteq \mathcal{S}^*$. In the following, we decompose

$$\mathsf{mat}\left(\nabla^2 f\left(\boldsymbol{\Sigma}_t\right)\mathsf{vec}\left(\boldsymbol{\Sigma}_{t+1}-\boldsymbol{\Sigma}_t\right)\right)+\nabla f\left(\boldsymbol{\Sigma}_t\right)$$

into four parts:

- $\mathbf{V}_1 \coloneqq \mathsf{mat}\left(\nabla^2 f\left(\boldsymbol{\Sigma}_t\right)\mathsf{vec}\left(\boldsymbol{\Sigma}_{t+1}-\boldsymbol{\Sigma}^*\right)\right)$,
- $\mathbf{V}_2 \coloneqq \mathsf{mat}\left(\nabla^2 f\left(\boldsymbol{\Sigma}_t\right)\mathsf{vec}\left(\boldsymbol{\Sigma}^*-\boldsymbol{\Sigma}_t\right)\right)$,
- $\mathbf{V}_3 \coloneqq \nabla f\left(\boldsymbol{\Sigma}_t\right)-\nabla f\left(\boldsymbol{\Sigma}^*\right)$,
- $\mathbf{V}_4 \coloneqq \nabla f\left(\boldsymbol{\Sigma}^*\right)$.

(1) For $\mathbf{V}_1$, we define set

$$\mathcal{H}_1 = \left\{(i,j) \in \overline{\mathcal{E}} \mid \left|\left(\mathbf{V}_1\right)_{ij}\right| \geq \frac{\lambda}{4}\right\}.$$

Referring to Lemma 33, we establish that

$$F_{\boldsymbol{\Lambda}}\left(\boldsymbol{\Sigma}_{t+1}\right) \leq F_{\boldsymbol{\Lambda}}\left(\boldsymbol{\Sigma}^*\right) + \frac{\lambda}{4}\left\|\boldsymbol{\Sigma}_{t+1}-\boldsymbol{\Sigma}^*\right\|_1.$$

This relationship leads to the following inequality:

$$\begin{aligned}&f\left(\boldsymbol{\Sigma}_{t+1}\right) - f\left(\boldsymbol{\Sigma}^*\right) \\ \leq &\lambda\left(\left\|\boldsymbol{\Sigma}^*\right\|_1 - \left\|\boldsymbol{\Sigma}_{t+1}\right\|_1\right) + \frac{\lambda}{4}\left\|\boldsymbol{\Sigma}_{t+1}-\boldsymbol{\Sigma}^*\right\|_1 \\ = &\lambda\left(\left\|\boldsymbol{\Sigma}^*_{\mathcal{E}}\right\|_1 - \left\|\left(\boldsymbol{\Sigma}_{t+1}\right)_{\mathcal{E}}\right\|_1 - \left\|\left(\boldsymbol{\Sigma}_{t+1}\right)_{\overline{\mathcal{E}}}\right\|_1\right) + \frac{\lambda}{4}\left\|\boldsymbol{\Sigma}_{t+1}-\boldsymbol{\Sigma}^*\right\|_1 \\ \leq &\frac{5\lambda}{4}\left\|\left(\boldsymbol{\Sigma}_{t+1}\right)_{\mathcal{E}} - \boldsymbol{\Sigma}^*_{\mathcal{E}}\right\|_1 - \frac{3\lambda}{4}\left\|\left(\boldsymbol{\Sigma}_{t+1}\right)_{\overline{\mathcal{E}}} - \boldsymbol{\Sigma}^*_{\overline{\mathcal{E}}}\right\|_1,\end{aligned} \tag{32}$$

where the equality holds since $\boldsymbol{\Sigma}^*_{\overline{\mathcal{E}}} = 0$. On the other hand, we have

$$\begin{aligned}&f\left(\boldsymbol{\Sigma}_{t+1}\right) - f\left(\boldsymbol{\Sigma}^*\right) \\ \underset{(i)}{\geq} &\left\langle \nabla f\left(\boldsymbol{\Sigma}^*\right), \boldsymbol{\Sigma}_{t+1}-\boldsymbol{\Sigma}^*\right\rangle \\ \geq &-\left\|\nabla f\left(\boldsymbol{\Sigma}^*\right)\right\|_{\mathsf{max}}\left\|\boldsymbol{\Sigma}_{t+1}-\boldsymbol{\Sigma}^*\right\|_1 \\ \underset{(ii)}{\geq} &-\frac{\lambda}{4}\left\|\boldsymbol{\Sigma}_{t+1}-\boldsymbol{\Sigma}^*\right\|_1 \\ = &-\frac{\lambda}{4}\left\|\left(\boldsymbol{\Sigma}_{t+1}\right)_{\mathcal{E}} - \boldsymbol{\Sigma}^*_{\mathcal{E}}\right\|_1 - \frac{\lambda}{4}\left\|\left(\boldsymbol{\Sigma}_{t+1}\right)_{\overline{\mathcal{E}}} - \boldsymbol{\Sigma}^*_{\overline{\mathcal{E}}}\right\|_1,\end{aligned} \tag{33}$$

where (i) is from the convexity of f and (ii) is from Assumption 9. Combining (32) and (33), we conclude

$$\left\|\left(\boldsymbol{\Sigma}_{t+1}\right)_{\overline{\mathcal{E}}} - \boldsymbol{\Sigma}^*_{\overline{\mathcal{E}}}\right\|_1 \leq 3\left\|\left(\boldsymbol{\Sigma}_{t+1}\right)_{\mathcal{E}} - \boldsymbol{\Sigma}^*_{\mathcal{E}}\right\|_1.$$

Consider a subset $\mathcal{S}' \subset \mathcal{H}_1$ with $|\mathcal{S}'| = s' \leq \widetilde{s}$. Select a vector $\mathbf{v} \in \mathbb{R}^{d^2}$ such that $\|\mathbf{v}\|_{\mathsf{max}} = 1$ and $\|\mathbf{v}\|_0 = s'$, satisfying the condition $s'\frac{\lambda}{4} \leq \mathbf{v}^\top \nabla^2 f\left(\boldsymbol{\Sigma}_t\right)\mathsf{vec}\left(\boldsymbol{\Sigma}_{t+1}-\boldsymbol{\Sigma}^*\right)$. Then, we have

$$\begin{aligned}s'\frac{\lambda}{4} &\leq \mathbf{v}^\top \nabla^2 f\left(\boldsymbol{\Sigma}_t\right)\mathsf{vec}\left(\boldsymbol{\Sigma}_{t+1}-\boldsymbol{\Sigma}^*\right) \\ &\leq \left\|\mathbf{v}^\top\left(\nabla^2 f\left(\boldsymbol{\Sigma}_t\right)\right)^{\frac{1}{2}}\right\|_2 \left\|\left(\nabla^2 f\left(\boldsymbol{\Sigma}_t\right)\right)^{\frac{1}{2}}\mathsf{vec}\left(\boldsymbol{\Sigma}_{t+1}-\boldsymbol{\Sigma}^*\right)\right\|_2 \\ &\underset{(i)}{\leq} c_1\sqrt{\rho^+_{2s^*+2\widetilde{s}}\rho^+_{s'}}\left\|\mathbf{v}\right\|_2\left\|\boldsymbol{\Sigma}^*-\boldsymbol{\Sigma}_{t+1}\right\|_{\mathsf{F}} \\ &\underset{(ii)}{\leq} c_1\sqrt{s'\rho^+_{2s^*+2\widetilde{s}}\rho^+_{s'}}\left\|\boldsymbol{\Sigma}^*-\boldsymbol{\Sigma}_{t+1}\right\|_{\mathsf{F}} \\ &\underset{(iii)}{\leq} \frac{c_1\sqrt{s'\rho^+_{2s^*+2\widetilde{s}}\rho^+_{s'}}}{\rho^-_{2s^*+2\widetilde{s}}}\lambda\sqrt{s^*},\end{aligned}$$

where (i) is from SE condition, (ii) is from the definition of $\mathbf{v}$, and (iii) is from Lemma 34. From above inequalities, it follows that

$$s' \leq \frac{c_1\rho^+_{2s^*+2\widetilde{s}}\rho^+_{s'}s^*}{\left(\rho^-_{2s^*+2\widetilde{s}}\right)^2} \leq c_1\varpi^2_{2s^*+2\widetilde{s}}s^*,$$

where $\varpi_{2s^*+2\widetilde{s}} = \rho^+_{2s^*+2\widetilde{s}}/\rho^-_{2s^*+2\widetilde{s}}$ is the condition number. Given that $s' = |\mathcal{S}'|$ achieves the maximum possible value such that $s' \leq \widetilde{s}$ for any subset $\mathcal{S}'$ of $\mathcal{H}_1$. We conclude $\mathcal{S}' = \mathcal{H}_1$ to attain the maximum and thus $|\mathcal{H}_1| = s' \leq c_1\varpi^2_{2s^*+2\widetilde{s}}s^*$.

(2) For $\mathbf{V}_2$, we define set

$$\mathcal{H}_2 = \left\{(i,j) \in \overline{\mathcal{E}} \mid \left|\left(\mathbf{V}_2\right)_{ij}\right| \geq \frac{\lambda}{4}\right\}.$$

Similar to the argument of $\mathcal{H}_1$, we have $|\mathcal{H}_2| \leq c_2 \varpi^2_{2s^*+2\widetilde{s}} s^*$.

(3) For $\mathbf{V}_3$, we define set

$$\mathcal{H}_3 = \left\{ (i,j) \in \overline{\mathcal{E}} \mid \left| (\mathbf{V}_3)_{ij} \right| \geq \frac{\lambda}{4} \right\}.$$

Consider a vector $\mathbf{V} \in \mathbb{R}^{d\times d}$ such that $\|\mathbf{V}\|_{\max} = 1$, $\|\mathbf{V}\|_0 = |\mathcal{H}_3|$, and

$$\begin{aligned} &\langle \mathbf{V}, \nabla f(\mathbf{\Sigma}_t) - \nabla f(\mathbf{\Sigma}^*) \rangle \\ =& \sum_{(i',j')\in\mathcal{H}_3} \mathbf{V}_{i'j'} \left(\nabla f(\mathbf{\Sigma}_t) - \nabla f(\mathbf{\Sigma}^*) \right)_{i'j'} \\ =& \sum_{(i',j')\in\mathcal{H}_3} \left(\nabla f(\mathbf{\Sigma}_t) - \nabla f(\mathbf{\Sigma}^*) \right)_{ij} \\ \geq& \frac{\lambda}{4}. \end{aligned} \tag{36}$$

Then we have

$$\begin{aligned} &\langle \mathbf{V}, \nabla f(\mathbf{\Sigma}_t) - \nabla f(\mathbf{\Sigma}^*) \rangle \\ \leq& \|\mathbf{V}\|_{\mathsf{F}} \left\| \nabla f(\mathbf{\Sigma}_t) - \nabla f(\mathbf{\Sigma}^*) \right\|_{\mathsf{F}} \\ \underset{(i)}{\leq}& \sqrt{|\mathcal{H}_3|} \left\| \nabla f(\mathbf{\Sigma}_t) - \nabla f(\mathbf{\Sigma}^*) \right\|_{\mathsf{F}} \\ \underset{(ii)}{\leq}& \rho^+_{2s^*+2\widetilde{s}} \sqrt{|\mathcal{H}_3|} \left\| \mathbf{\Sigma}_t - \mathbf{\Sigma}^* \right\|_{\mathsf{F}}, \end{aligned} \tag{37}$$

where (i) is from the definition of $\mathbf{V}$, and (ii) is from the mean value theorem and analogous argument for $\mathcal{H}_3$.

Combining (36) and (37), we have

$$\begin{aligned} \lambda |\mathcal{H}_3| &\leq 4\rho^+_{2s^*+2\widetilde{s}} \sqrt{|\mathcal{H}_3|} \left\| \mathbf{\Sigma}_t - \mathbf{\Sigma}^* \right\|_{\mathsf{F}} \\ &\underset{(i)}{\leq} 8\lambda \varpi_{2s^*+2\widetilde{s}} \sqrt{3s^* |\mathcal{H}_3|}, \end{aligned}$$

where (i) is from Lemma 34. This implies $|\mathcal{H}_3| \leq c_3 \varpi^2_{2s^*+2\widetilde{s}} s^*$.

(4) For $\mathbf{V}_4$, we define set

$$\mathcal{H}_4 = \left\{ (i,j) \in \overline{\mathcal{E}} \mid \left| (\boldsymbol{V}_4)_{ij} \right| \geq \frac{\lambda}{4} \right\}.$$

By assumption 9, we have

$$\begin{aligned} 0 \leq |\mathcal{H}_4| &\leq \sum_{(i,j)\in\overline{\mathcal{E}}} \frac{4}{\lambda} \left| \left(\nabla f(\mathbf{\Sigma}^*) \right)_{ij} \right| \cdot \mathbb{I}\left(\left| \left(\nabla f(\mathbf{\Sigma}^*) \right)_{ij} \right| > \frac{\lambda}{4} \right) \\ &= \sum_{(i,j)\in\overline{\mathcal{E}}} \frac{4}{\lambda} \left| \left(\nabla f(\mathbf{\Sigma}^*) \right)_{ij} \right| \cdot 0 = 0. \end{aligned} \tag{38}$$

Combining the results for Set $\mathcal{H}_1 \sim \mathcal{H}_4$, we have that there exists some constant c_0 such that

$$\left\| \left(\mathbf{\Sigma}_{t+\frac{1}{2}} \right)_{\overline{\mathcal{S}^*}} \right\|_0 \leq c_0 \kappa^2_{2s^*+2\widetilde{s}} s^* \leq \widetilde{s}.$$

■

Lemma 37. *Suppose Assumptions 8 and 9 hold. If $\mathbf{\Sigma}^{(t)} \in \mathcal{B}(\mathbf{\Sigma}^*, r)$ and $\left\| \mathbf{\Sigma}^{(t+1)}_{\overline{\mathcal{S}^*}} \right\|_0 \leq \widetilde{s}$, then for each stage $k \geq 2$, we have*

$$\left\| \mathbf{\Sigma}^{(k)}_{t+1} - \widehat{\mathbf{\Sigma}}^{(k)} \right\|_{\mathsf{F}} \leq \frac{\tau\kappa^3}{\rho^-_{2s^*+2\widetilde{s}}} \left\| \mathbf{\Sigma}^{(k)}_{t} - \widehat{\mathbf{\Sigma}}^{(k)} \right\|^2_{\mathsf{F}}.$$

Proof: First, we reformulate the prox-Newton update (11) as formula (34). By the Lemma 36, we have $\left\| \left(\mathbf{\Sigma}^{(k)}_{t+1} \right)_{\overline{\mathcal{S}^*}} \right\|_0 \leq \widetilde{s}$. And by the KKT condition, we have formula (35). By the strictly non-expansive property of the proximal operator, we obtain formula (39).

Note that both $\|\mathbf{\Sigma}_{t+1}\|_0 \leq \widetilde{s}$ and $\left\| \widehat{\mathbf{\Sigma}}^{(k)} \right\|_0 \leq \widetilde{s}$. On the other hand, from the SE properties, we have

$$\left\| \mathbf{\Sigma}^{(k)}_{t+1} - \widehat{\mathbf{\Sigma}}^{(k)} \right\|^2_{\nabla^2 f\left(\widehat{\mathbf{\Sigma}}^{(k)}\right)} \geq \rho^-_{2s^*+2\widetilde{s}} \left\| \mathbf{\Sigma}^{(k)}_{t+1} - \widehat{\mathbf{\Sigma}}^{(k)} \right\|^2_{\mathsf{F}}. \tag{40}$$

Combining (39) and (40), we have formula (41). Then the proof is finished. ■

In the subsequent analysis, it is essential to utilize the property that the iterates $\mathbf{\Sigma}_t \in \mathcal{B}\left(\widehat{\mathbf{\Sigma}}^{(k)}, 2r \right)$ rather than $\mathbf{\Sigma}_t \in \mathcal{B}(\mathbf{\Sigma}^*, r)$ for the convergence analysis of the proximal Newton method. This property is valid due to the simultaneous inclusion of $\mathbf{\Sigma}_t \in \mathcal{B}(\mathbf{\Sigma}^*, r)$ and $\widehat{\mathbf{\Sigma}}^{(k)} \in \mathcal{B}(\mathbf{\Sigma}^*, r)$. Consequently, it follows that $\mathbf{\Sigma}_t \in \mathcal{B}\left(\widehat{\mathbf{\Sigma}}^{(k)}, 2r \right)$, where $2r = \frac{\rho^-_{2s^*+2\widetilde{s}}}{2\tau\kappa}$ represents the radius of the quadratic convergence region for the proximal Newton algorithm.

Lemma 38. *Suppose that Assumptions 1, 4, 7, 8, and 7 hold. If $\mathbf{\Sigma}_t \in \mathcal{B}\left(\widehat{\mathbf{\Sigma}}^{(k)}, 2r \right)$ and $\|(\mathbf{\Sigma}_t)_{\overline{\mathcal{S}^*}}\|_0 \leq \widetilde{s}$ at each stage $k \geq 2$ with $\alpha = 0.3$, then $\eta_t = 1$. Further we have*

$$F\left(\mathbf{\Sigma}^{(k)}_{t+1} \right) \leq F\left(\mathbf{\Sigma}^{(k)}_{t} \right) + \frac{1}{4}\delta_t.$$

$$\mathbf{\Sigma}^{(k)}_{t+1} = \underset{\mathbf{\Sigma}}{\operatorname{argmin}} \left\langle \nabla f\left(\mathbf{\Sigma}^{(k)}_t \right), \mathbf{\Sigma} - \mathbf{\Sigma}^{(k)}_t \right\rangle + \frac{1}{2} \left\| \mathbf{\Sigma} - \mathbf{\Sigma}^{(k)}_t \right\|^2_{\nabla^2 f\left(\mathbf{\Sigma}^{(k)}_t\right)} + \left\| \mathbf{\Lambda}^{(k-1)} \odot \mathbf{\Sigma} \right\|_1 \tag{34}$$

$$\widehat{\mathbf{\Sigma}}^{(k)} = \underset{\mathbf{\Sigma}}{\operatorname{argmin}} \left\langle \nabla f\left(\widehat{\mathbf{\Sigma}}^{(k)} \right), \mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}^{(k)} \right\rangle + \frac{1}{2} \left\| \mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}^{(k)} \right\|^2_{\nabla^2 f\left(\mathbf{\Sigma}^{(k)}_t\right)} + \left\| \mathbf{\Lambda}^{(k-1)} \odot \widehat{\mathbf{\Sigma}}^{(k)} \right\|_1 \tag{35}$$

$$\begin{aligned}\left\|\boldsymbol{\Sigma}_{t+1}^{(k)}-\widehat{\boldsymbol{\Sigma}}^{(k)}\right\|_{\nabla^2 f\left(\widehat{\boldsymbol{\Sigma}}^{(k)}\right)}^2 &\leq \mathsf{vec}^\top\left(\boldsymbol{\Sigma}_{t+1}^{(k)}-\widehat{\boldsymbol{\Sigma}}^{(k)}\right)\left[\nabla^2 f\left(\boldsymbol{\Sigma}_t^{(k)}\right)\mathsf{vec}\left(\boldsymbol{\Sigma}_t^{(k)}-\widehat{\boldsymbol{\Sigma}}^{(k)}\right)+\left(\nabla f\left(\widehat{\boldsymbol{\Sigma}}^{(k)}\right)-\nabla f\left(\boldsymbol{\Sigma}_t^{(k)}\right)\right)\right]\\ &\leq \left\|\boldsymbol{\Sigma}_{t+1}^{(k)}-\widehat{\boldsymbol{\Sigma}}^{(k)}\right\|_{\mathsf{F}}\left\|\nabla^2 f\left(\boldsymbol{\Sigma}_t^{(k)}\right)\mathsf{vec}\left(\boldsymbol{\Sigma}_t^{(k)}-\widehat{\boldsymbol{\Sigma}}^{(k)}\right)+\left(\nabla f\left(\widehat{\boldsymbol{\Sigma}}^{(k)}\right)-\nabla f\left(\boldsymbol{\Sigma}_t^{(k)}\right)\right)\right\|_2.\end{aligned} \tag{39}$$

$$\begin{aligned}&\left\|\boldsymbol{\Sigma}_{t+1}^{(k)}-\widehat{\boldsymbol{\Sigma}}^{(k)}\right\|_{\mathsf{F}}\\ \leq&\frac{1}{\underline{\rho}_{2s^*+2\widetilde{s}}^-}\left\|\nabla^2 f\left(\boldsymbol{\Sigma}_t^{(k)}\right)\mathsf{vec}\left(\boldsymbol{\Sigma}_t^{(k)}-\widehat{\boldsymbol{\Sigma}}^{(k)}\right)+\left(\nabla f\left(\widehat{\boldsymbol{\Sigma}}^{(k)}\right)-\nabla f\left(\boldsymbol{\Sigma}_t^{(k)}\right)\right)\right\|_2\\ =&\frac{1}{\underline{\rho}_{2s^*+2\widetilde{s}}^-}\left\|\int_0^1\left[\nabla^2 f\left(\boldsymbol{\Sigma}_t^{(k)}+\theta\left(\widehat{\boldsymbol{\Sigma}}^{(k)}-\boldsymbol{\Sigma}_t^{(k)}\right)\right)-\nabla^2 f\left(\boldsymbol{\Sigma}_t^{(k)}\right)\right]\cdot\mathsf{vec}\left(\widehat{\boldsymbol{\Sigma}}^{(k)}-\boldsymbol{\Sigma}_t^{(k)}\right)d\theta\right\|_2\\ \leq&\frac{1}{\underline{\rho}_{2s^*+2\widetilde{s}}^-}\int_0^1\left\|\left[\nabla^2 f\left(\boldsymbol{\Sigma}_t^{(k)}+\theta\left(\widehat{\boldsymbol{\Sigma}}^{(k)}-\boldsymbol{\Sigma}_t^{(k)}\right)\right)-\nabla^2 f\left(\boldsymbol{\Sigma}_t^{(k)}\right)\right]\cdot\mathsf{vec}\left(\widehat{\boldsymbol{\Sigma}}^{(k)}-\boldsymbol{\Sigma}_t^{(k)}\right)\right\|_2 d\theta\\ \underset{(i)}{\leq}&\frac{\tau\kappa^3}{\underline{\rho}_{2s^*+2\widetilde{s}}^-}\left\|\boldsymbol{\Sigma}_t^{(k)}-\widehat{\boldsymbol{\Sigma}}^{(k)}\right\|_{\mathsf{F}}^2,\end{aligned} \tag{41}$$

where (i) is from the local restricted Hessian smoothness of f.

Proof: Recall $\boldsymbol{\Sigma}_{t+\frac{1}{2}}$ defined in (11) and denote $\Delta\boldsymbol{\Sigma}_t=\boldsymbol{\Sigma}_{t+\frac{1}{2}}-\boldsymbol{\Sigma}_t$. Then we have

$$\begin{aligned}&\left\|\Delta\boldsymbol{\Sigma}_t^{(k)}\right\|_{\mathsf{F}}^2\\ \underset{(i)}{\leq}&\left\|\boldsymbol{\Sigma}_t^{(k)}-\widehat{\boldsymbol{\Sigma}}^{(k)}\right\|_{\mathsf{F}}+\left\|\boldsymbol{\Sigma}_{t+\frac{1}{2}}^{(k)}-\widehat{\boldsymbol{\Sigma}}^{(k)}\right\|_{\mathsf{F}}\\ \underset{(ii)}{\leq}&\left\|\boldsymbol{\Sigma}_t^{(k)}-\widehat{\boldsymbol{\Sigma}}^{(k)}\right\|_{\mathsf{F}}+\frac{\tau\kappa^3}{\underline{\rho}_{2s^*+2\widetilde{s}}^-}\left\|\boldsymbol{\Sigma}_t^{(k)}-\widehat{\boldsymbol{\Sigma}}^{(k)}\right\|_{\mathsf{F}}^2\\ \underset{(iii)}{\leq}&\frac{3}{2}\left\|\boldsymbol{\Sigma}_t^{(k)}-\widehat{\boldsymbol{\Sigma}}^{(k)}\right\|_{\mathsf{F}},\end{aligned} \tag{42}$$

where (i) is from triangle inequality, (ii) is from Lemma 37, and (iii) is from $\left\|\boldsymbol{\Sigma}_t^{(k)}-\widehat{\boldsymbol{\Sigma}}^{(k)}\right\|_{\mathsf{F}}\leq R\leq\frac{\underline{\rho}_{2s^*+2\widetilde{s}}^-}{2\tau\kappa^3}$.

By Lemma 36, we have $\|(\boldsymbol{\Sigma}_t)_{\overline{\mathcal{S}^*}}\|_0\leq 2\widetilde{s}$. Then by expanding F, we have formula (43). The $\eta_t=1$ can be demonstrated analogously from the analysis in [55] and [57], thus we omit it. ∎

Lemma 39. *Suppose that Assumptions 1, 4, 7, 8, and 9 hold. To achieve the approximate KKT condition $\omega(\boldsymbol{\Sigma}_t)\leq\varepsilon$ for any $\varepsilon>0$ at each stage $k\geq 2$, the number of iteration for proximal Newton updates is at most*

$$\log\log\left(3\rho_{2s^*+2\widetilde{s}}^+/\varepsilon\right).$$

Proof: Based on the solution $\boldsymbol{\Sigma}_{t-1}$ obtained from the $(t-1)$-th iteration, we consider the optimal solution at t-th iteration. By the KKT condition, we have

$$\mathsf{mat}\left(\nabla^2 f(\boldsymbol{\Sigma}_{t-1})\,\mathsf{vec}\left(\boldsymbol{\Sigma}_t-\boldsymbol{\Sigma}_{t-1}\right)\right)+\nabla f(\boldsymbol{\Sigma}_{t-1})+\boldsymbol{\Lambda}\odot\boldsymbol{\Xi}_t=\mathbf{0},$$

where $\boldsymbol{\Xi}_t\in\partial\|\boldsymbol{\Sigma}_t\|_1$. Given any $\mathbf{V}\in\mathbb{R}^{d\times d}$ satisfying $\|\mathbf{V}\|_2\leq\|\mathbf{V}\|_1=1$ and $\|\mathbf{V}\|_0\leq 2s^*+2\widetilde{s}$, we derive formula (44). By taking the supremum of the L.H.S of (44) with respect to $\mathbf{V}$, we have

$$\|\nabla f(\boldsymbol{\Sigma}_t)+\boldsymbol{\Lambda}\odot\boldsymbol{\Xi}_t\|_{\max}\leq 2\rho_{2s^*+2\widetilde{s}}^+\|\boldsymbol{\Sigma}_t-\boldsymbol{\Sigma}_{t-1}\|_{\mathsf{F}}. \tag{45}$$

Then from Lemma (37), we have

$$\begin{aligned}&\left\|\boldsymbol{\Sigma}_{t+1}^{(k)}-\widehat{\boldsymbol{\Sigma}}^{(k)}\right\|_{\mathsf{F}}\\ \leq&\left(\tau\kappa^3/\underline{\rho}_{2s^*+2\widetilde{s}}^-\right)^{1+2+4+\ldots+2^{t-1}}\left\|\boldsymbol{\Sigma}_0^{(k)}-\widehat{\boldsymbol{\Sigma}}^{(k)}\right\|_{\mathsf{F}}^2\\ \leq&\left(\left(\tau\kappa^3/\underline{\rho}_{2s^*+2\widetilde{s}}^-\right)\left\|\boldsymbol{\Sigma}_0^{(k)}-\widehat{\boldsymbol{\Sigma}}^{(k)}\right\|_{\mathsf{F}}^2\right)^{2^t}.\end{aligned}$$

By (44) and (42) by taking $\Delta\boldsymbol{\Sigma}_{t-1}=\boldsymbol{\Sigma}_t-\boldsymbol{\Sigma}_{t-1}$, we obtain

$$\begin{aligned}&\omega_{\boldsymbol{\Lambda}^{(k-1)}}\left(\boldsymbol{\Sigma}_t^{(k)}\right)\\ \leq&2\rho_{2s^*+2\widetilde{s}}^+\left\|\boldsymbol{\Sigma}_t^{(k)}-\boldsymbol{\Sigma}_{t-1}^{(k)}\right\|_{\mathsf{F}}\\ \leq&3\rho_{2s^*+2\widetilde{s}}^+\left\|\boldsymbol{\Sigma}_{t-1}^{(k)}-\widehat{\boldsymbol{\Sigma}}^{(k)}\right\|_{\mathsf{F}}\\ \leq&3\rho_{2s^*+2\widetilde{s}}^+\left(\left(\tau\kappa^3/\underline{\rho}_{2s^*+2\widetilde{s}}^-\right)\left\|\boldsymbol{\Sigma}_0^{(k)}-\widehat{\boldsymbol{\Sigma}}^{(k)}\right\|_{\mathsf{F}}\right)^{2^t}.\end{aligned}$$

$$
\begin{aligned}
&F\left(\mathbf{\Sigma}_t^{(k)}+\Delta\mathbf{\Sigma}_t^{(k)}\right)-F\left(\mathbf{\Sigma}_t^{(k)}\right)\\
=&f\left(\mathbf{\Sigma}_t^{(k)}+\Delta\mathbf{\Sigma}_t^{(k)}\right)-f\left(\mathbf{\Sigma}_t^{(k)}\right)+\left\|\mathbf{\Lambda}^{(k-1)}\odot\left(\mathbf{\Sigma}_t^{(k)}+\Delta\mathbf{\Sigma}_t^{(k)}\right)\right\|_1-\left\|\mathbf{\Lambda}^{(k-1)}\odot\mathbf{\Sigma}_t^{(k)}\right\|_1\\
\underset{(i)}{\leq}&\left\langle\nabla f\left(\mathbf{\Sigma}_t^{(k)}\right),\Delta\mathbf{\Sigma}_t^{(k)}\right\rangle+\frac{1}{2}\left\|\Delta\mathbf{\Sigma}_t^{(k)}\right\|^2_{\nabla^2 f\left(\mathbf{\Sigma}_t^{(k)}\right)}+\frac{\tau\kappa^3}{3}\left\|\Delta\mathbf{\Sigma}_t^{(k)}\right\|_{\mathsf{F}}^3+\left\|\mathbf{\Lambda}^{(k-1)}\odot\left(\mathbf{\Sigma}_t^{(k)}+\Delta\mathbf{\Sigma}_t^{(k)}\right)\right\|_1-\left\|\mathbf{\Lambda}^{(k-1)}\odot\mathbf{\Sigma}_t^{(k)}\right\|_1\\
\underset{(ii)}{\leq}&\delta_t-\frac{1}{2}\delta_t+\frac{\tau\kappa^3}{3}\left\|\Delta\mathbf{\Sigma}_t^{(k)}\right\|_{\mathsf{F}}^3\underset{(iii)}{\leq}\frac{1}{2}\delta_t+\frac{\tau\kappa^3}{3\rho^-_{2s^*+2\widetilde{s}}}\left\|\Delta\mathbf{\Sigma}_t^{(k)}\right\|^2_{\nabla^2 f(\mathbf{\Sigma})}\left\|\Delta\mathbf{\Sigma}_t^{(k)}\right\|_{\mathsf{F}}\\
\underset{(iv)}{\leq}&\left(\frac{1}{2}-\frac{\tau\kappa^3}{3\rho^-_{2s^*+2\widetilde{s}}}\left\|\Delta\mathbf{\Sigma}_t^{(k)}\right\|_{\mathsf{F}}\right)\delta_t\underset{(v)}{\leq}\frac{1}{4}\delta_t,
\end{aligned}
\tag{43}
$$

where (i) is from the restricted Hessian smooth condition, (ii) and (iv) are from Lemma 35, (iii) is from the same argument of (40), and (v) is from (42), $\delta_t<0$, and $\left\|\mathbf{\Sigma}_t^{(k)}-\widehat{\mathbf{\Sigma}}^{(k)}\right\|_{\mathsf{F}}\leq r\leq\frac{\rho^-_{2s^*+2\widetilde{s}}}{2\tau\kappa^3}$.

$$
\begin{aligned}
&\left\langle\nabla f\left(\mathbf{\Sigma}_t\right)+\mathbf{\Lambda}\odot\mathbf{\Xi}_t,\mathbf{V}\right\rangle\\
=&\left\langle\nabla f\left(\mathbf{\Sigma}_t\right)-\nabla f\left(\mathbf{\Sigma}_{t-1}\right)-\mathsf{mat}\left(\nabla^2 f\left(\mathbf{\Sigma}_{t-1}\right)\mathsf{vec}\left(\mathbf{\Sigma}_t-\mathbf{\Sigma}_{t-1}\right)\right),\mathbf{V}\right\rangle\\
=&\left\langle\nabla f\left(\mathbf{\Sigma}_t\right)-\nabla f\left(\mathbf{\Sigma}_{t-1}\right),\mathbf{V}\right\rangle-\left(\nabla^2 f\left(\mathbf{\Sigma}_{t-1}\right)\mathsf{vec}\left(\mathbf{\Sigma}_t-\mathbf{\Sigma}_{t-1}\right)\right)^\top\mathsf{vec}\left(\mathbf{V}\right)\\
\underset{(i)}{\leq}&\left\|\left(\nabla^2 f\left(\breve{\mathbf{\Sigma}}\right)\right)^{\frac{1}{2}}\mathsf{vec}\left(\mathbf{\Sigma}_t-\mathbf{\Sigma}_{t-1}\right)\right\|_2\cdot\left\|\mathsf{vec}^\top\left(\mathbf{V}\right)\left(\nabla^2 f\left(\breve{\mathbf{\Sigma}}\right)\right)^{\frac{1}{2}}\right\|_2\\
&+\left\|\left(\nabla^2 f\left(\mathbf{\Sigma}_{t-1}\right)\right)^{\frac{1}{2}}\mathsf{vec}\left(\mathbf{\Sigma}_t-\mathbf{\Sigma}_{t-1}\right)\right\|_2\cdot\left\|\mathsf{vec}^\top\left(\mathbf{V}\right)\left(\nabla^2 f\left(\mathbf{\Sigma}_{t-1}\right)\right)^{\frac{1}{2}}\right\|_2\\
\underset{(ii)}{\leq}&2\rho^+_{2s^*+2\widetilde{s}}\left\|\mathbf{\Sigma}_t-\mathbf{\Sigma}_{t-1}\right\|_{\mathsf{F}},
\end{aligned}
\tag{44}
$$

where (i) is from mean value theorem with some $\breve{\mathbf{\Sigma}}=(1-\theta)\,\mathbf{\Sigma}_{t-1}+\theta\mathbf{\Sigma}_t$ for some $\theta\in[0,1]$ and Cauchy-Schwarz inequality, and (ii) is from the SE properties.

By requiring the R.H.S equal to ε we obtain

$$
\begin{aligned}
t=&\log\left(\log\left(\frac{3\rho^+_{2s^*+2\widetilde{s}}}{\varepsilon}\right)/\log\left(\frac{\rho^-_{2s^*+2\widetilde{s}}}{\tau\kappa^3\left\|\mathbf{\Sigma}_0^{(k)}-\widehat{\mathbf{\Sigma}}^{(k)}\right\|_{\mathsf{F}}}\right)\right)\\
=&\log\log\left(\frac{3\rho^+_{2s^*+2\widetilde{s}}}{\varepsilon}\right)-\log\log\left(\frac{\rho^-_{2s^*+2\widetilde{s}}}{\tau\kappa^3\left\|\mathbf{\Sigma}_0^{(k)}-\widehat{\mathbf{\Sigma}}^{(k)}\right\|_{\mathsf{F}}}\right)\\
\underset{(i)}{\leq}&\log\log\left(\frac{3\rho^+_{2s^*+2\widetilde{s}}}{\varepsilon}\right)-\log\log 4\\
\leq&\log\log\left(\frac{3\rho^+_{2s^*+2\widetilde{s}}}{\varepsilon}\right),
\end{aligned}
$$

where (i) is from the fact that $\left\|\mathbf{\Sigma}_0^{(k)}-\widehat{\mathbf{\Sigma}}^{(k)}\right\|_{\mathsf{F}}\leq r=\frac{\rho^-_{2s^*+2\widetilde{s}}}{4\tau\kappa^3}$. ∎

D. Proof of Theorem 17

Lemma 40. *Suppose that Assumptions 1, 4, 7, 8, and 9 hold. After sufficiently many iterations $T<\infty$, the following results hold for all $t\geq T$:*

$$
\left\|(\mathbf{\Sigma}_t)_{\overline{\mathcal{S}^*}}\right\|_0\leq\widetilde{s}\ \textit{and}\ F_{\mathbf{\Lambda}^{(0)}}\left(\mathbf{\Sigma}_t^{(1)}\right)\leq F_{\mathbf{\Lambda}^{(0)}}\left(\mathbf{\Sigma}^*\right)+\frac{15\lambda^2 s^*}{4\rho^-_{2s^*+2\widetilde{s}}}.
$$

Proof: Building upon the framework established in Lemma 35, Lemma 38, it can be concluded that the objective $F_{\mathbf{\Lambda}^{(0)}}$ exhibits a sufficient decrease in each iteration of the proximal Newton step, as also discussed in [58]. Consequently, there exists a constant T such that for all $t\geq T$, we have

$$
F_{\mathbf{\Lambda}^{(0)}}\left(\mathbf{\Sigma}_t^{(1)}\right)\leq F_{\mathbf{\Lambda}^{(0)}}\left(\mathbf{\Sigma}^*\right)+\frac{\lambda}{4}\left\|\mathbf{\Sigma}_t^{(1)}-\mathbf{\Sigma}^*\right\|_1,
$$

where $\left\|\mathbf{\Sigma}_t^{(1)}-\mathbf{\Sigma}^*\right\|_1\leq\frac{c_1\lambda\sqrt{s^*}}{\rho^-_{2s^*+2\widetilde{s}}}$ derived from a similar argument presented in the proof of Lemma 25. The subsequent analysis follows a pattern analogous to that of Lemma 36. Let $S_n=\left\{(i,j)\mid\left|\nabla f\left(\widetilde{\mathbf{\Sigma}}\right)_{ij}\right|=\lambda\right\}$ and consider the set $\left\{(i,j)\mid\left(\widetilde{\mathbf{\Sigma}}\right)_{ij}\neq 0\right\}\subseteq S_n$. It is required to show $|S_n|\leq s^*+\widetilde{s}$. To demonstrate this, we decompose S_n into two distinct parts:

- $S_n^1:\left\{(i,j)\notin\overline{\mathcal{S}^*}\mid\left|\left(\nabla f\left(\widetilde{\mathbf{\Sigma}}\right)-\nabla f\left(\mathbf{\Sigma}^*\right)\right)_{ij}\right|\geq\frac{\lambda}{2}\right\}$,
- $S_n^2:\left\{(i,j)\notin\overline{\mathcal{S}^*}\mid\left|\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)_{ij}\right|>\frac{\lambda}{2}\right\}$.

We have $S_n \subseteq \mathcal{S}^* \cup S_n^1 \cup S_n^2$ and establish upper bounds for each.

(1) For S_n^1, consider S' with maximum size $s' = |S'| \leq \widetilde{s}$ such that $S' \subseteq S_n^1$. Then there exists a matrix $\mathbf{V}$ such that $\|\mathbf{V}\|_{\max} = 1$ and $\|\mathbf{V}\|_0 = s'$, satisfying

$$\frac{\lambda s'}{2} \leq \left\langle \nabla f\left(\widetilde{\mathbf{\Sigma}}\right) - \nabla f\left(\mathbf{\Sigma}^*\right), \mathbf{V} \right\rangle.$$

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

$$\begin{aligned}&\nabla f\left(\widetilde{\mathbf{\Sigma}}\right) - \nabla f\left(\mathbf{\Sigma}^*\right)\\ =&\mathsf{mat}\left(\nabla^2 f\left(\theta\widetilde{\mathbf{\Sigma}} + (1-\theta)\mathbf{\Sigma}^*\right)\mathsf{vec}\left(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right)\right).\end{aligned}$$

For simplicity, we define $\mathbf{H} = \nabla^2 f\left(\rho\widetilde{\mathbf{\Sigma}} + (1-\rho)\mathbf{\Sigma}^*\right)$. Applying the Cauchy-Schwartz inequality, we have

$$\begin{aligned}\frac{\lambda s'}{2} &\leq \mathsf{vec}^\top(\mathbf{V})\,\mathbf{H}\mathsf{vec}\left(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right)\\ &\leq \underbrace{\left\|\mathsf{vec}^\top(\mathbf{V})\,(\mathbf{H})^{\frac{1}{2}}\right\|_2}_{\mathbf{I}}\underbrace{\left\|(\mathbf{H})^{\frac{1}{2}}\,\mathsf{vec}\left(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right)\right\|_2}_{\mathbf{II}}.\end{aligned} \tag{46}$$

We now establish upper bounds for terms $\mathbf{I}$ and $\mathbf{II}$ respectively. Let $\widetilde{\mathbf{\Sigma}}, \mathbf{\Sigma}^* \in \mathcal{B}(\mathbf{\Sigma}^*, r)$, any convex combination of $\widetilde{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}^*$ also falls in $\mathcal{B}(\mathbf{\Sigma}^*, r)$. The localized sparse eigenvalue condition can be applied to $\mathbf{H}$.

i) For term $\mathbf{I}$, based on Definition (3) and Assumption 4, we have

$$\begin{aligned}\left\|\mathsf{vec}^\top(\mathbf{U})\,(\mathbf{H})^{\frac{1}{2}}\right\|_2 &\leq \sqrt{\rho^+_{2s^*+2\widetilde{s}}}\,\|\mathbf{U}\|_{\mathsf{F}}^2\\ &\leq \sqrt{\rho^+_{2s^*+2\widetilde{s}}}\left(\|\mathbf{U}\|_1\|\mathbf{U}\|_{\max}\right)^{\frac{1}{2}}\\ &\leq \sqrt{\rho^+_{2s^*+2\widetilde{s}}}\sqrt{s'}.\end{aligned}$$

ii) For $\mathbf{II}$, using Lemma 32, we have

$$\begin{aligned}&\left\|(\mathbf{H})^{\frac{1}{2}}\,\mathsf{vec}\left(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right)\right\|_2\\ \leq&\left\langle \nabla f\left(\widetilde{\mathbf{\Sigma}}\right) - \nabla f\left(\mathbf{\Sigma}^*\right), \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\rangle\\ \leq&\frac{c_1\lambda^2 s^*}{\rho^-_{2s^*+2\widetilde{s}}}.\end{aligned}$$

By substituting the bounds for $\mathbf{I}$ and $\mathbf{II}$ back into (46), we obtain

$$\frac{\lambda s'}{2} \leq \sqrt{\rho^+_{2s^*+2\widetilde{s}}}\sqrt{s'} \times \frac{c_1\lambda\sqrt{s^*}}{\rho^-_{2s^*+2\widetilde{s}}}.$$

Multiplying both sides of this inequality by $\left(\frac{\lambda}{2}\right)^{\frac{1}{2}}$ and squaring gives us

$$s' \leq \frac{4\rho^+_{2s^*+2\widetilde{s}}c_2 s^*}{\left(\rho^-_{2s^*+2\widetilde{s}}\right)^2} < \widetilde{s}, \tag{47}$$

where the last inequality follows from our assumption. Since $s' = |S'|$ achieves the maximum possible value such that $s' \leq \widetilde{s}$ for any subset S' of S_n^1 and 47 shows that $s' < \widetilde{s}$, we conclude that $S' = S_n^1$, and thus

$$\left|S_n^1\right| = s' \leq \left\lfloor \frac{4\rho^+_{2s^*+2\widetilde{s}}c_2 s^*}{\left(\rho^-_{2s^*+2\widetilde{s}}\right)^2} \right\rfloor < \widetilde{s}.$$

(2) For S_n^2, Given $\|\nabla f(\mathbf{\Sigma}^*)\|_{\max} + \varepsilon \leq \frac{\lambda}{4}$, it follows that $S_n^2 = \emptyset$ and thus $\left|S_n^2\right| = 0$.

Combining all the results, we have $\left\|\mathbf{\Sigma}^{(t)}_{\overline{\mathcal{S}^*}}\right\|_0 \leq \widetilde{s}$. ∎

Lemma 41. *Suppose that Assumptions 4, 8 and 9 hold. If* $\left\|\left(\mathbf{\Sigma}_t^{(1)}\right)_{\overline{\mathcal{S}^*}}\right\|_0 \leq \widetilde{s}$, *and* $F_{\mathbf{\Lambda}^{(0)}}\left(\mathbf{\Sigma}_t^{(1)}\right) \leq F_{\mathbf{\Lambda}^{(0)}}(\mathbf{\Sigma}^*) + \frac{15\lambda^2 s^*}{4\rho^-_{2s^*+2\widetilde{s}}}$, *we have*

$$\left\|\mathbf{\Sigma}_t^{(1)} - \mathbf{\Sigma}^*\right\|_2 \leq \frac{c_1\lambda\sqrt{s^*}}{\rho^-_{2s^*+2\widetilde{s}}}$$

for some constant c_1 *and*

$$\left\|\mathbf{\Sigma}_{t+1}^{(1)} - \widehat{\mathbf{\Sigma}}^{(1)}\right\|_2 \leq \frac{\tau\kappa^3}{\rho^-_{2s^*+2\widetilde{s}}}\left\|\mathbf{\Sigma}_t^{(1)} - \widehat{\mathbf{\Sigma}}^{(1)}\right\|_2^2.$$

Proof: The estimation error is derived analogously from [28] and [57], thus we omit it here. The claim of quadratic convergence follows directly from Lemma (37) given sparse solutions. ∎

Lemma 42. *Suppose that Assumptions 4 8, and 9 hold. If* $\left\|\left(\mathbf{\Sigma}_T^{(1)}\right)_{\overline{\mathcal{S}^*}}\right\|_0 \leq \widetilde{s}$, *and* $F_{\mathbf{\Lambda}^{(0)}}\left(\mathbf{\Sigma}_T^{(1)}\right) \leq F_{\mathbf{\Lambda}^{(1)}}(\mathbf{\Sigma}^*) + \frac{15\lambda^2 s^*}{4\rho^-_{2s^*+2\widetilde{s}}}$ *at some iteration* T, *we need at most*

$$T_1 \leq \log\log\left(\frac{3\rho^+_{2s^*+2\widetilde{s}}}{\varepsilon}\right)$$

extra iterations of the proximal Newton updates such that $\omega_{\mathbf{\Lambda}^{(0)}}\left(\mathbf{\Sigma}_{T+T_1}^{(1)}\right) \leq \frac{\lambda}{8}$.

Proof: The maximum of the number of iterations for the proximal Newton update is determined by integrating Lemma (40) and Lemma (39). Note that

$$T_1 \leq \log\frac{\log\left(\frac{3\rho^+_{2s^*+2\widetilde{s}}}{\varepsilon}\right)}{\log\left(\frac{\rho^-_{2s^*+2\widetilde{s}}}{\tau\kappa^3\left\|\mathbf{\Sigma}^{(T+1)} - \widehat{\mathbf{\Sigma}}^{\{1\}}\right\|_{\mathsf{F}}}\right)}.$$

Then we obtain the result from $\left\|\mathbf{\Sigma}_{T+1}^{(1)} - \widehat{\mathbf{\Sigma}}^{(1)}\right\|_{\mathsf{F}} \leq r = \frac{\rho^-_{2s^*+2\widetilde{s}}}{4\tau\kappa}$. ∎

APPENDIX E
SUPPLEMENTARY TABLES

TABLE I
QUANTITATIVE COMPARISON AMONG DIFFERENT PENALTIES FOR THE BANDED MATRIX

	ℓ_1				MCP				ℓ_1				MCP			
	$m=100$	$m=500$	$m=1500$	$m=3000$	$m=100$	$m=500$	$m=1500$	$m=3000$	$m=100$	$m=500$	$m=1500$	$m=3000$	$m=100$	$m=500$	$m=1500$	$m=3000$
	$d=80, n=50$								$d=80, n=100$							
FAE	27.7825 (0.2616)	24.1565 (0.1648)	22.1561 (0.1246)	19.9998 (0.1051)	24.3152 (0.2551)	20.1542 (0.1566)	19.4295 (0.1047)	18.6415 (0.1015)	26.5923 (0.2436)	23.8546 (0.1534)	21.4165 (0.1132)	19.4620 (0.1052)	23.7515 (0.2435)	19.5612 (0.1530)	18.5164 (0.1106)	17.8962 (0.0945)
FRE	1.3876 (0.2546)	1.2065 (0.1591)	1.1066 (0.1198)	0.9989 (0.9895)	1.2144 (0.2357)	1.0066 (0.1296)	0.9704 (0.1210)	0.9311 (0.0890)	1.3282 (0.1973)	1.1914 (0.1275)	1.0697 (0.1016)	0.9720 (0.0999)	1.1863 (0.1773)	0.9770 (0.1176)	0.9248 (0.1012)	0.8938 (0.1001)
FPR	0.1615 (0.0702)	0.1528 (0.0765)	0.1137 (0.0686)	0.1053 (0.0552)	0.1256 (0.0699)	0.1038 (0.0438)	0.0681 (0.0423)	0.0311 (0.0217)	0.1572 (0.0425)	0.1315 (0.0407)	0.1152 (0.0319)	0.0873 (0.0287)	0.1137 (0.0366)	0.0852 (0.0287)	0.0489 (0.0209)	0.0241 (0.0174)
TPR	0.8072 (0.0368)	0.8491 (0.0382)	0.8774 (0.0397)	0.8955 (0.0399)	0.8413 (0.0397)	0.8618 (0.0407)	0.8957 (0.0412)	0.9214 (0.0420)	0.8117 (0.0371)	0.8546 (0.0392)	0.8810 (0.0401)	0.9016 (0.0408)	0.8455 (0.0400)	0.8821 (0.0412)	0.9016 (0.0418)	0.9425 (0.0431)
	$d=100, n=50$								$d=100, n=100$							
FAE	28.1276 (0.2492)	26.5312 (0.2151)	23.1562 (0.1813)	20.1532 (0.1513)	24.8612 (0.2263)	21.0017 (0.2052)	20.1562 (0.1673)	18.8964 (0.1233)	27.8613 (0.2161)	24.2361 (0.2111)	22.2613 (0.1562)	19.9131 (0.1162)	24.2315 (0.2023)	21.1156 (0.1762)	19.8216 (0.1230)	18.1566 (0.1055)
FRE	1.1012 (0.2435)	1.0387 (0.2051)	0.9066 (0.1736)	0.7890 (0.1435)	0.9734 (0.2315)	0.8223 (0.2184)	0.7892 (0.1763)	0.7398 (0.1256)	1.0908 (0.2251)	0.9489 (0.2056)	0.8716 (0.1813)	0.7796 (0.1562)	0.9487 (0.2130)	0.8267 (0.1504)	0.7761 (0.1261)	0.7109 (0.1023)
FPR	0.1511 (0.0851)	0.1341 (0.0793)	0.1056 (0.0712)	0.0915 (0.0663)	0.1105 (0.0714)	0.1004 (0.0627)	0.0557 (0.0487)	0.0303 (0.0226)	0.1462 (0.0487)	0.1227 (0.0415)	0.1005 (0.0326)	0.0905 (0.0291)	0.1126 (0.0392)	0.0756 (0.0311)	0.0448 (0.0265)	0.0131 (0.0199)
TPR	0.8157 (0.0337)	0.8581 (0.0391)	0.9167 (0.0401)	0.9294 (0.0413)	0.8780 (0.0402)	0.8860 (0.0412)	0.9070 (0.0422)	0.9421 (0.0430)	0.8307 (0.0380)	0.8706 (0.0401)	0.8913 (0.0409)	0.9212 (0.0411)	0.8605 (0.0416)	0.8915 (0.0423)	0.9172 (0.0427)	0.9512 (0.0436)