Achieving Oracle Property for Low-rank Covariance Matrix Estimation from Quadratic Measurements

Anonymous CPAL submission

The covariance matrix is a fundamental statistical quantity that characterizes the linear relationships among multiple random variables. The most common approach in estimating a covariance matrix is based on measurements assuming access to all the variables. However, this assumption is not practical for a dynamic and resource-constrained environment. Compressive covariance sketching offers a viable alternative to estimate the covariance matrix based on compressed measurements, which largely reduces the storage and computational requirements. In this paper, we introduce a novel method for low-rank covariance matrix estimation from quadratic measurements based on non-convex regularization and propose a proximal gradient homotopy method to solve the proposed estimator. The proposed method is theoretically proven to attain superior statistical rates to estimators based on the convex nuclear norm, achieving the oracle rate. The theoretical advancements are corroborated by extensive numerical experiments, highlighting the efficacy and robustness of our method.

1. Introduction

The covariance matrix, which quantifies linear correlations among multiple random variables, is a cornerstone of statistical analysis [1–4]. Despite its significance across numerous data and signal analysis applications [5–7], estimating the covariance matrix is challenging since it is not directly observable and can only be estimated from samples. The most naive approach to estimating the covariance matrix is the sample covariance matrix $\mathbf{\Sigma}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t^\top$, where $\left\{\mathbf{x}_t \in \mathbb{R}^d\right\}_{t=1}^{T}$ denote T samples [8–10]. In many situations, directly measuring $\mathbf{x}_t$ can be too costly, particularly when there are limitations on computational and storage resources. In these cases, a practical solution is to compress the measurements $\mathbf{x}_t$ into a lower-dimensional representation. The covariance can then be estimated using these compressed measurements.

The **C**ompressive **C**ovariance **S**ensing (or **C**ompressive **C**ovariance **S**ketching) (CCS), adapts the principles of compressive sensing for covariance matrix estimation, providing a viable approach to inferring the covariance matrix through compressed measurements [11–16]. In this paper, we study a quadratic measurement model [15, 16] as follows:

$$y_i = \frac{1}{T}\sum_{t=1}^{T} y_{it} + \eta_i, \quad \text{where } y_{it} = \left(\boldsymbol{a}_i^\top \mathbf{x}_t\right)^2,$$

where $\boldsymbol{a_i}$ is a random sensing vector. It is easy to verify that $y_i = \boldsymbol{a}_i^\top \mathbf{\Sigma}_T \boldsymbol{a}_i + \eta_i$. CCS is particularly advantageous when dealing with high-dimensional data, as it significantly reduces both storage and computational requirements [17]. The primary challenge in the CCS problem lies in its inherent ill-posedness: the number of available measurements is insufficient relative to the degrees of freedom of the covariance matrix [15]. To address this issue, exploiting structural information of the covariance matrix—where a small number of components capture most of the data variability—has proven highly effective, dramatically reducing the dimensionality of the estimation problem.

This paper studies the situation where the covariance matrix exhibits a low-rank property [18, 19]. This structural property is inspired by the advancements in phase retrieval [20–25], a field dedicated to recovering rank-one covariance matrices from quadratic measurements. Building on this

Submitted to Second Conference on Parsimony and Learning (CPAL 2025). Do not distribute.

foundation, our proposed recovery algorithm extends the PhaseLift [21] methodology to handle low-rank matrices effectively. A direct approach is through rank minimization to achieve accurate recovery of a low-rank matrix. However, this approach is computationally intractable, leading to the development of various surrogate loss functions, which include the nuclear norm [26], Schatten-p norm [27], and weighted nuclear norm [28], among others. The nuclear norm approach has been praised for its close convex approximation to matrix rank, facilitating global optimality [29, 30]. A more advanced variant, weighted nuclear norm, assigns different weights to different singular values, enhancing estimator stability and accuracy over conventional nuclear norm penalty [31]. Despite their effectiveness, these ℓ_1-homotopy penalties introduce a non-negligible bias into estimators, which inevitably compromises accuracy. In contrast, non-convex penalties, such as the **S**moothly **C**lipped **A**bsolute **D**eviation (SCAD) penalty [32] and the **M**inimax **C**oncave **P**enalty (MCP) [33], have demonstrated promising empirical results [34–36].

In this paper, we study the CCS problem by assuming the covariance matrix has a low-rank structure. We propose a novel method for low-rank CCS by incorporating a non-convex penalty. We propose a proximal gradient homotopy method to solve the proposed estimator. Our theoretical analysis confirms that the proposed estimator achieves superior statistical convergence rates compared to traditional nuclear norm-penalized estimators. Under reasonable assumptions, our approach demonstrates the oracle property, ensuring precise recovery of the true rank of the covariance matrix along with improved convergence rates. The efficacy of our proposed method is further substantiated by extensive simulation studies, showcasing its superior performance relative to existing methods. Our main contributions are summarized as follows:

- We propose a new approach for estimating low-rank covariance matrices from a quadratic measurement model using a non-convex penalty, which aids in obtaining unbiased estimates. A proximal gradient homotopy algorithm solves the proposed estimator.
- We rigorously demonstrate the statistical properties of the approximate solution generated by our proposed algorithm. Under weak assumptions, we prove that the proposed estimator achieves the oracle statistical rate in the Frobenius norm.
- Our theoretical findings are corroborated through extensive numerical experiments on synthetic datasets, showcasing the effectiveness and robustness of our proposed method.

2. Compressive Covariance Sketching

This section presents a unified framework for low-rank covariance matrix sketching. Subsequently, we formulate our estimator with the non-convex penalty and provide several practical illustrations of the covariance matrix sketching model.

2.1. The Quadratic Measurement Model

We consider the quadratic measurement model defined as follows:

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

where $\{y_i\}_{i=1}^m$ are the observed measurements, $\boldsymbol{\Sigma}_T$ denotes the sample covariance matrix , $\{\boldsymbol{a}_i\}_{i=1}^m$ are the sensing vectors, $\{\eta_i\}_{i=1}^m$ represents the noise components, and m is the total number of measurements. We assume that the noise term $\{\eta_i\}_{i=1}^m$ are independent and sampled from a sub-Gaussian distribution with zero mean and variance proxy δ^2. To facilitate further analysis, we decompose the sample covariance matrix $\boldsymbol{\Sigma}_T$ into the true covariance matrix $\boldsymbol{\Sigma}^*$ and a bias term $\mathbf{E}$, such that $\boldsymbol{\Sigma}_T = \boldsymbol{\Sigma}^* + \mathbf{E}$. For clarity, we define the sensing matrix $\mathbf{A}_i \coloneqq \boldsymbol{a}_i \boldsymbol{a}_i^\top$. Observing that $\boldsymbol{a}_i^\top \boldsymbol{\Sigma}_T \boldsymbol{a}_i = \langle \mathbf{A}_i, \boldsymbol{\Sigma}_T \rangle$, we rewrite the measurement model as $y_i = \langle \mathbf{A}_i, \boldsymbol{\Sigma}_T \rangle + \eta_i$. Let $\boldsymbol{y} \coloneqq [y_1, \cdots, y_m]^\top$ be the measurement vector and $\boldsymbol{\eta} \coloneqq [\eta_1, \cdots, \eta_m]^\top$ the noise vector. We introduce the linear operator $\mathcal{A}(\cdot) : \mathbb{R}^{d \times d} \mapsto \mathbb{R}^m$, defined by $\mathcal{A}(\boldsymbol{\Sigma}_T) = [\langle \mathbf{A}_1, \boldsymbol{\Sigma}_T \rangle, \ldots, \langle \mathbf{A}_m, \boldsymbol{\Sigma}_T \rangle]$ for any $\boldsymbol{\Sigma}_T \in \mathbb{R}_+^{d \times d}$. With these definitions, the measurement model can be compactly expressed as

$\boldsymbol{y} = \mathcal{A}(\boldsymbol{\Sigma}_T) + \boldsymbol{\eta}$. Notably, the quadratic measurement model—previously explored by [15, 21]—offers a significantly simpler implementation and greater computational efficiency than models that utilize full-rank measurement matrices with independently and identically distributed entries.

2.2. The Proposed Estimator

We denote the matrix to be estimated as $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ and propose the following regularized least squares type optimization estimator

$$\min_{\boldsymbol{\Sigma}\succeq\mathbf{0}} \left\{ \frac{1}{2m} \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{\Sigma})\|_2^2 + \sum_{i=1}^{d} p_\lambda(\sigma_i(\boldsymbol{\Sigma})) \right\}, \tag{1}$$

where $\sigma_i(\boldsymbol{\Sigma})$ denotes the i-th largest eigenvalue of $\boldsymbol{\Sigma}$, and $p_\lambda(\cdot)$ is a non-convex penalty function governed by a regularization parameter $\lambda > 0$. As detailed in Assumption 1, we impose specific conditions on it.

Assumption 1. *The function $p_\lambda(t)$ defined on $[0, +\infty)$ can be decomposed as $p_\lambda(t) = \lambda t + q_\lambda(t)$, and the functions $p_\lambda(t)$ and $q_\lambda(t)$ satisfy the following conditions:*

(*a*) *For all $t_1 \geq t_2 \geq 0$, the derivative satisfies $0 \leq p'_\lambda(t_1) \leq p'_\lambda(t_2) \leq \lambda$ and $\lim_{t\to 0} p'_\lambda(t) = \lambda$;*

(*b*) *$q'_\lambda(t)$ is monotonic and Lipschitz continuous. Specifically, for $t_1 \geq t$, there exists a constant $\zeta^- \geq 0$ such that $q'_\lambda(t_1) - q'_\lambda(t) \geq -\zeta^-(t_1 - t)$.*

(*c*) *$p_\lambda(t)$ is non-decreasing with $p_\lambda(0) = 0$ and is differentiable almost everywhere on $(0, +\infty)$;*

(*d*) *There exists an $\alpha > 0$ such that $p'_\lambda(t) = 0$ for $t \geq \alpha$;*

According to [32], a well-designed penalty function should induce an estimator exhibiting three desirable properties: unbiasedness, sparsity, and continuity. These properties align with the conditions outlined in Assumption 1. Several functions satisfying Assumption 1 have been explored. Below, we present a few representative examples:

- Smooth clipped absolute deviation penalty: This penalty, due to [32], is given by

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

  where $b > 2$ is an additional tuning parameter. The authors in [32] 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 [33], is defined as follows:

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

  for some $b > 0$.

2.3. Applications

Low-rank CCS serves as a foundational technique in a multitude of applications across diverse domains. These applications range from traffic data monitoring [37] and array signal processing [38] to collaborative filtering [39, 40] and metric learning [41, 42]. In these fields, the primary focus is often on extracting second-order or higher-order statistics—such as the power spectrum and cyclic power spectrum—rather than directly reconstructing random signals [43–45]. Here, we briefly review two representative examples: PhaseLift and spectrum estimation.

- **PhaseLift:** PhaseLift is a pivotal method for addressing the phase retrieval problem, which aims to recover a complex vector $\mathbf{x} \in \mathbb{C}^d$ from magnitude-only measurements. Specifically,

Algorithm 1: The Projection Proximal Gradient Algorithm

Input: λ, ε, $\boldsymbol{y}$, $\{\boldsymbol{a}_i\}_{i=1}^m$
Initialize η, δ, $\boldsymbol{\Sigma}_0 = \mathbf{0}$, $\phi_k = 0$;
for $k = 0, 1, 2, \ldots, K-1$ **do**
  $\lambda_{k+1} = \eta\lambda_k$, $\varepsilon_{k+1} = \frac{\lambda_k}{4}$;
  $(\boldsymbol{\Sigma}_{k+1}, \phi_{k+1}) \leftarrow \mathrm{ProxGrad}\left(\lambda_{k+1}, \varepsilon_{k+1}, \boldsymbol{\Sigma}_k, \phi_k\right)$;
  $\mathbf{U}_{k+1}\boldsymbol{\Psi}_{k+1}\mathbf{U}_{k+1}^\top \leftarrow \mathrm{EigDec}\left(\boldsymbol{\Sigma}_{k+1}\right)$;
  $\boldsymbol{\Psi}_{k+1} = \max\left(0, \boldsymbol{\Psi}_{k+1}\right)$; $\boldsymbol{\Sigma}_{k+1} = \mathbf{U}_{k+1}\boldsymbol{\Psi}_{k+1}\mathbf{U}_{k+1}^\top$;
end
Output: $\boldsymbol{\Sigma}_K$

phase retrieval involves reconstructing $\mathbf{x}$ using m intensity measurements of the form $b_i = |\langle \mathbf{x}, \mathbf{z}_i \rangle|^2$. PhaseLift reformulates this problem by lifting it into a high-dimensional space, transitioning from recovering the vector $\mathbf{x}$ to recovering the matrix $\mathbf{X} = \mathbf{x}\mathbf{x}^{\mathrm{H}}$. Here, $\mathbf{x}^{\mathrm{H}}$ denotes the conjugate transpose of $\mathbf{x}$, and $\mathbf{X}$ is a rank-1 complex Hermitian matrix. This lifting technique effectively transforms the original nonlinear problem into a tractable linear problem through convex optimization. For a comprehensive discussion on the equivalence conditions between the original and lifted formulations, we refer readers to [21].

- **Spectrum Estimation:** Spectrum estimation is an essential technique for analyzing the frequency characteristics of signals within stochastic processes. Its primary objective is to accurately recover the power spectrum (or power spectral density) of a signal, which is intrinsically linked to its autocorrelation function. Specifically, the task of spectrum sensing involves estimating the power spectrum of a wide-sense stationary analog signal $x(t)$. The system under consideration employs a sampling architecture composed of M parallel branches. Each i-th branch modulates the input signal $x(t)$ with a potentially complex-valued periodic waveform $p_i(t)$ that has a period of NT. This modulation is followed by an integrate-and-dump circuit operating with the same period NT, effectively reducing the sampling rate to $\frac{1}{N}$ of the Nyquist rate. The output of the i-th branch at the k-th sampling interval is given by

$$y_i[k] = \frac{1}{NT}\int_{kNT}^{(k+1)NT} p_i(t)x(t)dt.$$

  Leveraging this relationship, the goal is to reconstruct the power spectrum of $x(t)$ using the collected samples $\{y_i[k]\}_{i,k}$ [12]. This approach facilitates efficient spectrum estimation by exploiting the multi-branch sampling architecture and the properties of wide-sense stationary signals.

3. Optimization Algorithm

We propose solving problem (1) via a projection proximal gradient algorithm. Define $f\left(\boldsymbol{\Sigma}\right) = \frac{1}{2m}\left\|\boldsymbol{y} - \mathcal{A}\left(\boldsymbol{\Sigma}\right)\right\|_2^2$. Given Assumption 1, the penalty term $\sum_{i=1}^d p_\lambda\left(\sigma_i\left(\boldsymbol{\Sigma}\right)\right)$ can be expressed as a combination of $\lambda\left\|\boldsymbol{\Sigma}\right\|_*$ and $Q_\lambda\left(\boldsymbol{\Sigma}\right)$, where $Q_\lambda\left(\boldsymbol{\Sigma}\right) = \sum_{i=1}^d q_\lambda\left(\sigma_i\left(\boldsymbol{\Sigma}\right)\right)$ is the concave component. Define $\widetilde{f}\left(\boldsymbol{\Sigma}\right) = f\left(\boldsymbol{\Sigma}\right) + Q_\lambda\left(\boldsymbol{\Sigma}\right)$. Consequently, the optimization problem is reformulated as follows:

$$\min_{\boldsymbol{\Sigma}} \quad \widetilde{f}\left(\boldsymbol{\Sigma}\right) + \lambda\left\|\boldsymbol{\Sigma}\right\|_* + \mathbb{I}_{\{\boldsymbol{\Sigma}\succeq\mathbf{0}\}}\left(\boldsymbol{\Sigma}\right), \tag{2}$$

where $\mathbb{I}_{\{\boldsymbol{\Sigma}\succeq\mathbf{0}\}}\left(\boldsymbol{\Sigma}\right)$ is the indicator function for the semidefinite cone, taking value zero if $\boldsymbol{\Sigma}\succeq\mathbf{0}$ and infty otherwise.

To derive the estimate $\widehat{\boldsymbol{\Sigma}}$ from problem (2), we develop a projection proximal gradient method with backtracking line search [46, 47]. Starting from $\boldsymbol{\Sigma}_0 = \mathbf{0}$ for simplicity, we generate a sequence $\{\boldsymbol{\Sigma}_k\}_{k\geq 0}$ via the proximal gradient update described below.

Algorithm 2: $(\mathbf{\Sigma}_j, \phi) \leftarrow \mathrm{ProxGrad}\,(\lambda, \varepsilon, \mathbf{\Sigma}, \phi)$

Input: λ, ε, $\mathbf{\Sigma}$, ϕ
Initialize $j = 0$, $\mathbf{\Sigma}_j = \mathbf{\Sigma}$;
while $\omega\left(\widetilde{\mathbf{\Sigma}}_j\right) > \varepsilon$ **do**
  $j = j + 1$;
  $\mathbf{\Sigma}_j \leftarrow \arg\min_{\mathbf{\Sigma}} \widetilde{g}\left(\mathbf{\Sigma} \mid \mathbf{\Sigma}_{j-1}\right)$;
  while $\left\langle \nabla \widetilde{f}\left(\mathbf{\Sigma}_{j-1}\right), \mathbf{\Sigma}_j - \mathbf{\Sigma}_{j-1} \right\rangle + \frac{\phi}{2} \left\|\mathbf{\Sigma} - \mathbf{\Sigma}_j\right\|_{\mathrm{F}}^2 < 0$ **do**
    $\mathbf{\Sigma}_j \leftarrow \arg\min_{\mathbf{\Sigma}} \widetilde{g}\left(\mathbf{\Sigma} \mid \mathbf{\Sigma}_{j-1}\right)$;
    $\phi = 2\phi$;
  end
  $\phi = \frac{\phi}{2}$;
end
Output: $(\mathbf{\Sigma}_j, \phi)$

At the $(k+1)$-th iteration, we construct an isotropic quadratic approximation of $\widetilde{f}\left(\mathbf{\Sigma}\right)$ at a given matrix $\mathbf{\Sigma}_k$ as follows:

$$\widetilde{g}\left(\mathbf{\Sigma} \mid \mathbf{\Sigma}_k\right) = \widetilde{f}\left(\mathbf{\Sigma}_k\right) + \left\langle \nabla \widetilde{f}\left(\mathbf{\Sigma}_k\right), \mathbf{\Sigma} - \mathbf{\Sigma}_k \right\rangle + \frac{\phi}{2} \left\|\mathbf{\Sigma} - \mathbf{\Sigma}_k\right\|_{\mathrm{F}}^2 + \lambda \left\|\mathbf{\Sigma}\right\|_* + \mathbb{I}_{\{\mathbf{\Sigma} \succeq \mathbf{0}\}}\left(\mathbf{\Sigma}\right), \tag{3}$$

where ϕ is an appropriately chosen quadratic coefficient. The next iterate $\mathbf{\Sigma}_{k+1}$ is then obtained by solving the optimization problem:

$$\mathbf{\Sigma}_{k+1} = \arg\min_{\mathbf{\Sigma}} \widetilde{g}\left(\mathbf{\Sigma} \mid \mathbf{\Sigma}_k\right)$$

Suppose $\widehat{\mathbf{\Sigma}}$ is an exact minimizer. It satisfies the optimality condition:

$$\left\langle \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}, \nabla \widetilde{f}(\widehat{\mathbf{\Sigma}}) + \lambda \widehat{\mathbf{W}} + \widehat{\mathbf{\Xi}} \right\rangle \leq 0,$$

where $\widehat{\mathbf{W}} \in \partial \|\widehat{\mathbf{\Sigma}}\|_*$ and $\widehat{\mathbf{\Xi}} \in \partial \mathbb{I}_{\{\mathbf{\Sigma} \succeq \mathbf{0}\}}(\widehat{\mathbf{\Sigma}})$ are sub-gradients of the nuclear norm and the indicator function, respectively. However, obtaining the exact solution $\widehat{\mathbf{\Sigma}}$ can be computationally intensive. Therefore, we define an approximate solution to (3).

Definition 2. *An approximate solution $\widetilde{\mathbf{\Sigma}}$ to (3) is said to be ε-optimal if its sub-optimal measure $\omega\left(\widetilde{\mathbf{\Sigma}}\right)$ satisfies $\omega\left(\widetilde{\mathbf{\Sigma}}\right) \leq \varepsilon$, where*

$$\begin{aligned}
\omega\left(\widetilde{\mathbf{\Sigma}}\right) &= \min_{\widetilde{\mathbf{W}} \in \partial \|\widetilde{\mathbf{\Sigma}}\|_*, \widetilde{\mathbf{\Xi}} \in \partial \mathbb{I}_{\{\mathbf{\Sigma} \succeq \mathbf{0}\}} \widetilde{\mathbf{\Sigma}}} \max_{\mathbf{\Sigma}'} \left\{ \frac{\left\langle \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}', \left(\nabla \widetilde{f}\left(\widetilde{\mathbf{\Sigma}}\right) + \lambda \widetilde{\mathbf{W}} + \widetilde{\mathbf{\Xi}}\right) \right\rangle}{\left\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}'\right\|_*} \right\} \\
&= \min_{\widetilde{\mathbf{W}} \in \partial \left\|\widetilde{\mathbf{\Sigma}}\right\|_*, \widetilde{\mathbf{\Xi}} \in \partial \mathbb{I}_{\{\mathbf{\Sigma} \succeq \mathbf{0}\}}(\widetilde{\mathbf{\Sigma}})} \left\{ \left\| \nabla \widetilde{f}\left(\widetilde{\mathbf{\Sigma}}\right) + \lambda \widetilde{\mathbf{W}} + \widetilde{\mathbf{\Xi}} \right\|_2 \right\},
\end{aligned}$$

and the second equality follows from the duality between the nuclear norm $\|\cdot\|_*$ and the spectral norm $\|\cdot\|_2$.

4. Main Theories

In this section, we first introduce some technical backgrounds and assumptions. 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.

4.1. Backgrounds and Assumptions

We consider a symmetric positive semi-definite matrix $\mathbf{\Sigma}^* \in \mathbb{R}^{d\times d}$ whose non-zero eigenvalues are collected in the vector $\boldsymbol{\sigma}^* \in \mathbb{R}^r$. The eigen-decomposition of $\mathbf{\Sigma}^*$ is given by $\mathbf{\Sigma}^* = \mathbf{U}^*\mathbf{\Gamma}^*\mathbf{U}^{*\top}$, where $\mathbf{U}^* \in \mathbb{R}^{d\times r}$ is a matrix whose columns are orthonormal eigenvectors corresponding to the non-zero eigenvalues in $\boldsymbol{\sigma}^*$, and $\mathbf{\Gamma}^* = \operatorname{diag}(\boldsymbol{\sigma}^*) \in \mathbb{R}^{r\times r}$ is a diagonal matrix containing these eigenvalues. Based on this decomposition, we define two complementary subspaces $\mathcal{S}$ and $\mathcal{S}^\perp$ as follows:

$$\begin{aligned}\mathcal{T}(\mathbf{U}^*) &\coloneqq \{\mathbf{\Delta} \mid \operatorname{row}(\mathbf{\Delta}) \subseteq \mathbf{U}^* \text{ and } \operatorname{col}(\mathbf{\Delta}) \subseteq \mathbf{U}^*\},\\ \mathcal{T}^\perp(\mathbf{U}^*) &\coloneqq \{\mathbf{\Delta} \mid \operatorname{row}(\mathbf{\Delta}) \perp \mathbf{U}^* \text{ and } \operatorname{col}(\mathbf{\Delta}) \perp \mathbf{U}^*\}.\end{aligned}$$

Corresponding to these subspaces, we define two projection operators $\mathcal{P}(\cdot)$ and $\mathcal{P}^\perp(\cdot)$ that project matrices onto $\mathcal{T}$ and $\mathcal{T}^\perp$, respectively. For any matrix $\Delta \in \mathbb{R}^{d\times d}$, these projections are explicitly given by:

$$\mathcal{P}(\mathbf{\Delta}) = \mathbf{U}^*\mathbf{U}^{*\top}\mathbf{\Delta}\mathbf{U}^*\mathbf{U}^{*\top}, \mathcal{P}^\perp(\mathbf{\Delta}) = \left(\mathbf{I} - \mathbf{U}^*\mathbf{U}^{*\top}\right)\mathbf{\Delta}\left(\mathbf{I} - \mathbf{U}^*\mathbf{U}^{*\top}\right),$$

where $\mathbf{I}$ denotes the identity matrix of the appropriate dimension. We then define a subset $\mathcal{B}$, which forms a cone representing a restricted set of directions:

$$\mathcal{B} = \left\{\mathbf{\Delta} \in \mathbb{R}^{d\times d} \mid \left\|\mathcal{P}^\perp(\mathbf{\Delta})\right\|_* \le 5\left\|\mathcal{P}(\mathbf{\Delta})\right\|_*\right\}.$$

Assumption 3 (Restricted Strong Convexity)**.** *For the empirical loss function $f(\cdot)$, there exists some $\rho^- > 0$ such that, for all $\mathbf{\Sigma}_2 - \mathbf{\Sigma}_1 \in \mathcal{B}$,*

$$f(\mathbf{\Sigma}_2) \ge f(\mathbf{\Sigma}_1) + \langle \nabla f(\mathbf{\Sigma}_1), \mathbf{\Sigma}_2 - \mathbf{\Sigma}_1\rangle + \frac{\rho^-}{2}\left\|\mathbf{\Sigma}_2 - \mathbf{\Sigma}_1\right\|_{\mathrm{F}}^2.$$

Assumption 4 (Restricted Strong Smoothness)**.** *For the empirical loss function $f(\cdot)$, there exists some $\infty > \rho^+ \ge \rho^- > 0$ such that, for all $\mathbf{\Sigma}_2 - \mathbf{\Sigma}_1 \in \mathcal{B}$,*

$$f(\mathbf{\Sigma}_1) + \langle \nabla f(\mathbf{\Sigma}_1), \mathbf{\Sigma}_2 - \mathbf{\Sigma}_1\rangle + \frac{\rho^+}{2}\left\|\mathbf{\Sigma}_2 - \mathbf{\Sigma}_1\right\|_{\mathrm{F}}^2 \ge f(\mathbf{\Sigma}_2).$$

Previous research has conclusively demonstrated that random Gaussian measurements satisfy the conditions of restricted strong convexity and restricted strong smoothness [48].

4.2. Statistical Properties

We now present our main theorem, which establishes a deterministic error bound for the estimator in the quadratic measurement model.

Theorem 5. *Suppose that Assumptions 1, 3, and 4 hold. If $\lambda \ge 2\left\|\mathcal{A}^*(\boldsymbol{y} - \mathcal{A}(\mathbf{\Sigma}))\right\|_2/m$ and $\rho^- > \zeta^-$, the optimal solution $\widehat{\mathbf{\Sigma}}$ satisfies:*

$$\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} \le \frac{\vartheta\sqrt{|\mathcal{S}_1|}}{\rho^- - \zeta^-} + \frac{3\lambda\sqrt{|\mathcal{S}_2|}}{\rho^- - \zeta^-},$$

where $|\mathcal{S}_1|$ and $|\mathcal{S}_2|$ denote the cardinality of the sets $|\mathcal{S}_1| = \{i \mid \sigma_i^* \ge \alpha\}$ and $|\mathcal{S}_2| = \{i \mid \alpha > \sigma_i^* > 0\}$ respectively, $\vartheta = \left\|\mathcal{P}_{S_1}(\nabla f(\mathbf{\Sigma}^*))\right\|_2$, and $\mathcal{P}_{S_1}$ is the projection onto the subspace of $\mathcal{T}$ associated with S_1.

Remark 6. *The oracle rate refers to the statistical convergence rate of the oracle estimator, which knows the true rank subspace $\mathcal{T}(\mathbf{U}^*)$ in advance. The oracle estimator $\widehat{\mathbf{\Sigma}}^O$ is defined as*

$$\widehat{\mathbf{\Sigma}}^O = \arg\min_{\mathbf{\Sigma}:\mathbf{\Sigma}\in\mathcal{T}(\mathbf{U}^*)} f(\mathbf{\Sigma}).$$

The upper bound on the Frobenius norm of the estimation error provided by Theorem 5 consists of two distinct components, each corresponding to different magnitudes of the eigenvalues of the true covariance matrix $\mathbf{\Sigma}^*$. Specifically, the set $\mathcal{S}_1$ corresponds to the indices of the larger eigenvalues, while $\mathcal{S}_2$ includes those of the smaller eigenvalues. By selecting an appropriate value for ζ^-, such as $\zeta^- = \rho^-/2$, we demonstrate that the proposed estimator achieves faster convergence than the nuclear norm-based estimator. Next, we give the explicit statistical rate of convergence.

Corollary 7 (Oracle Property)**.** *Suppose that Assumptions 1, 3, and 4 hold. If $\rho^- > \zeta^-$, λ is selected such that*

$$\lambda \geq \frac{2\left\|\mathcal{A}^*\left(\boldsymbol{y}-\mathcal{A}\left(\boldsymbol{\Sigma}\right)\right)\right\|_2}{m}+\frac{2\sqrt{r}\rho^+\left\|\mathcal{A}^*\left(\boldsymbol{y}-\mathcal{A}\left(\boldsymbol{\Sigma}\right)\right)\right\|_2}{m\rho^-}$$

and if $\boldsymbol{\sigma}^*$ satisfies that,

$$\min_{i\in\mathcal{S}}|\sigma_i^*| \geq \alpha+\frac{2\sqrt{r}\left\|\mathcal{A}^*\left(\boldsymbol{y}-\mathcal{A}\left(\boldsymbol{\Sigma}\right)\right)\right\|_2}{m\rho^-},$$

where $\mathcal{S}=\operatorname{supp}\left(\boldsymbol{\sigma}^*\right)$, then the optimal solution $\widehat{\boldsymbol{\Sigma}}$ satisfies:

$$\left\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}^*\right\|_{\mathrm{F}} \leq \frac{2\left\|\mathcal{P}(\nabla f(\boldsymbol{\Sigma}^*))\right\|_2\sqrt{r}}{\rho^-}.$$

As established in 7, by appropriately selecting the regularization parameter λ and ensuring that the smallest non-zero eigenvalue has a sufficiently large, the proposed estimator in 1 can closely approximate the oracle estimator, even in the absence of direct knowledge of the subspace $\mathcal{T}$. Consequently, the estimator in 1 is able to recover the rank of the true covariance matrix $\boldsymbol{\Sigma}^*$. It indicates the CCS algorithm exhibits the oracle property, enabling near-optimal recovery of the covariance matrix, which mirrors the performance of an "oracle" that has full knowledge of the true underlying data structure. In practical terms, this means that CCS can produce covariance estimates that are almost as accurate as those obtained from full data sampling, even in situations where the number of measurements is significantly reduced.

Next, we give the explicit statistical rate of convergence under the sub-Gaussian design. First, we make the essential assumption about the sensing vector.

Assumption 8. *We assume that the sensing vectors $\boldsymbol{a}_i$'s $(1\leq i\leq m)$ consist of independently and identically distributed (i.i.d.) sub-Gaussian random variables. Each vector $\boldsymbol{a}_i$ contains d elements, denoted as $\left(\boldsymbol{a}_i\right)_j$ $(1\leq j\leq d)$, which satisfy the following conditions:*

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

Assumption 8 indicates that each component of the sensing vector exhibits heavy-tailed behavior. As a result, the sensing matrix $\mathbf{A}_i$ follows a sub-exponential distribution. Let $\operatorname{vec}\left(\mathbf{A}_i\right)\in\mathbb{R}^{d^2}$ represent the vectorization of $\mathbf{A}_i$. Under Assumption 8, $\operatorname{vec}\left(\mathbf{A}_i\right)$ is distributed according to a sub-exponential distribution with mean 0 and variance proxy $\boldsymbol{\Theta}$.

Corollary 9. *Let $\mathbf{x}$ be a sub-Gaussian random vector with zero mean and covariance $\boldsymbol{\Sigma}^*$. Consider a collection of T i.i.d. samples $\{\mathbf{x}_i\}_{i=1}^T$ drawn from $\mathbf{x}$. Suppose that Assumptions 1, 3, 4 and 8 hold. There exist universal constants $C_1, C_2, \cdots, C_{11}$ such that if $\lambda\geq 2C_1\delta\varpi\sqrt{d/m}$, where $\varpi=\sqrt{\sup_{\|\mathbf{u}\|_2=1,\|\mathbf{v}\|_2=1}\mathbf{u}^\top\boldsymbol{a}_i\boldsymbol{a}_i^\top\mathbf{v}}$ and $\rho^-=C_2\lambda_{\min}\left(\boldsymbol{\Theta}\right)>\zeta^-$, it holds with probability at least $1-C_3\exp\left(-2dC_4\right)$ that*

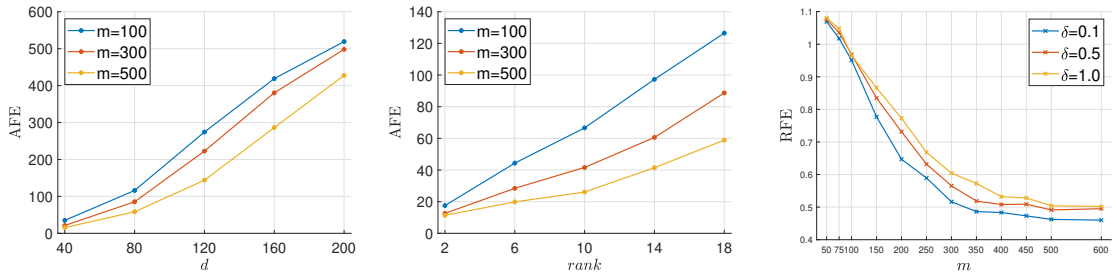
$$\left\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}^*\right\|_{\mathrm{F}} \leq \frac{\delta\varpi}{\sqrt{mT}\lambda_{\min}\left(\boldsymbol{\Theta}\right)}\left[C_5r_1+C_6\sqrt{r_2d}\right].$$

Further, if $\boldsymbol{\sigma}^*$ satisfies that

$$\min_{i\in\mathcal{S}}|\boldsymbol{\sigma}_i^*| \geq \nu+\frac{C_7\delta\varpi\sqrt{rd/m}}{\lambda_{\min}\left(\boldsymbol{\Theta}\right)},$$

where $\mathcal{S}=\operatorname{supp}\left(\boldsymbol{\sigma}^*\right)$. And if $\lambda\geq C_8\delta\varpi\sqrt{d/m}\left(1+\frac{\sqrt{r}\lambda_{\max}(\boldsymbol{\Theta})}{\lambda_{\min}(\boldsymbol{\Theta})}\right)$, it holds with probability at least $1-C_9\exp\left(-C_{10}d\right)$ that

$$\left\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}^*\right\|_{\mathrm{F}} \leq \frac{C_{11}r\delta\varpi}{\sqrt{mT}\lambda_{\min}\left(\boldsymbol{\Theta}\right)}.$$



(a) (b) (c)

Figure 1: Absolute Frobenius norm error of the proposed estimator changing d and $rank$ and Relative Frobenius norm error changing the noise variance proxy δ.

5. Numerical Experiments

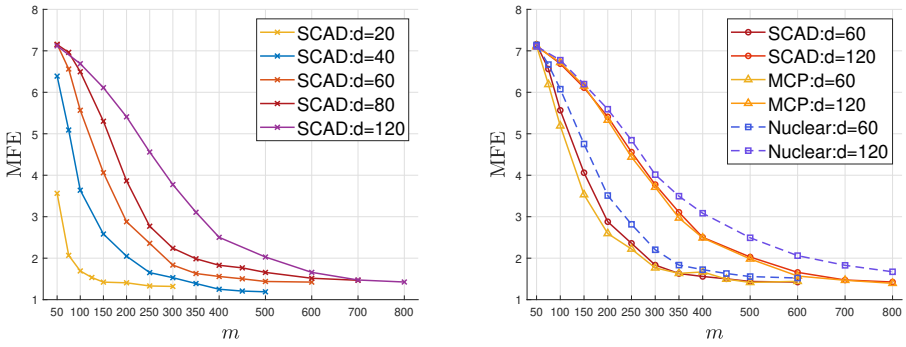
In this section, we perform several numerical experiments to evaluate the performance of the CCS with proposed regularization techniques. All experiments are implemented in MATLAB 2023b, run on an Intel $i7-10700$ 2.90 GHz $\times 8$ with 16 GB of RAM, and all the results reported represent averages over 100 Monte Carlo simulations.

Data Generation and Hyperparameter Setup: The synthetic datasets are generated as follows. We generate the true covariance matrix as $\mathbf{\Sigma}^* = \mathbf{L}\mathbf{L}^\top$, where $\mathbf{L}$ is a $d \times r$ matrix with independent Gaussian components. The sensing vector $\{\boldsymbol{a}_i\}_{i=1}^m$ are obtained by drawing i.i.d. samples from a sub-Gaussian $(\mathbf{0}, \mathbf{\Upsilon})$ distribution. The noise $\{\eta_i\}_{i=1}^m$ obeys the sub-Gaussian distribution with mean $\mathbf{0}$ and a variance proxy of δ^2. For all methods, the tuning parameter λ is determined through the application of five-fold cross-validation. Unless otherwise specified, we employ the SCAD as the non-convex penalty function, setting the parameter b to 3.7, and the variance proxy δ is 0.1.

Measurement Criteria: To evaluate the performance of the proposed estimator, we utilize several metrics based on the Frobenius norm. Let $\widehat{\mathbf{\Sigma}}$ denote the generic covariance matrix obtained by various algorithms and estimators, $\mathbf{\Sigma}^*$ represent the true covariance matrix. We define the **A**bsolute **F**robenius **E**rror (AFE) as $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{\mathrm{F}}$. Additionally, we introduce the **M**ean squared **F**robenius **E**rror (MFE) as $\frac{\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{\mathrm{F}}}{d^2}$, where d is the dimension of the covariance matrix. Similarly, the **R**elative **F**robenius **E**rror (RFE) is defined as $\frac{\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{\mathrm{F}}}{\|\mathbf{\Sigma}^*\|_{\mathrm{F}}}$.

Figure 1 shows the performance of our estimator under different parameter settings. Figure 1a shows the variation of the AFE as the dimension d of the covariance matrix increases, with a fixed rank of 10. Three curves, in different colors, correspond to different sample numbers $m = \{100, 300, 500\}$. From the plot, we observe that the estimation error increases with the dimension. A larger sample number is required for convergence as the dimension grows. Figure 1b presents the results when the matrix dimension $d = 60$, with three lines corresponding to different sample numbers. The results display that the augmentation of the rank of the true covariance matrix leads to a rise in the estimation error. For figure 1c, the dimension is fixed at $d = 60$. The experiment is implemented by changing the variance proxy δ of the noise component, which shows higher noise intensity adds up to estimation error.

Figures 2a and 2b illustrate experiments with synthetic data by varying the number of the samples while maintaining a fixed matrix rank of 10. Figure 2a shows the mean squared Frobenius norm error of our proposed non-convex estimator. The estimation error converges to the same value as the dimension increases. Figure 2b compares the performance of the non-convex method to the convex method (Nuclear Norm). The non-convex method demonstrates a faster rate of error reduction as the sample number increases.



(a) (b)

Figure 2: Mean Squared Frobenius norm error of the proposed estimator and the comparison between convex and nonconvex methods.

Table 1: The estimation error of non-convex and convex methods $\pm$ standard variance under 100 Monte Carlo Simulations.

		$d = 40$				$d = 80$			
		$m = 100$	$m = 200$	$m = 300$	$m = 500$	$m = 100$	$m = 300$	$m = 500$	$m = 600$
SCAD	AFE	76.2997 ± 13.0608	57.2471 ± 14.4645	49.4734 ± 15.8468	43.6211 ± 18.3060	203.8819 ± 14.3887	119.7443 ± 31.1828	102.9800 ± 36.4549	98.5062 ± 46.9165
	RFE	0.7281 ± 0.1261	0.5541 ± 0.1024	0.4796 ± 0.1516	0.4087 ± 0.1627	0.9846 ± 0.0275	0.5686 ± 0.1460	0.4883 ± 0.1725	0.46850 ± 0.2120
	MFE	3.6385 ± 1.1104	2.0483 ± 1.1008	1.5298 ± 1.2044	1.1893 ± 1.4276	6.4950 ± 0.7281	2.2404 ± 1.2968	1.6570 ± 1.4817	1.5162 ± 1.2715
MCP	AFE	72.3491 ± 16.2992	52.5658 ± 18.5107	43.5744 ± 16.5545	40.0282 ± 18.4621	200.6326 ± 13.0460	120.2025 ± 33.7358	101.3771 ± 37.7358	96.7045 ± 40.9204
	RFE	0.6823 ± 0.1459	0.5050 ± 0.1631	0.4164 ± 0.1694	0.4730 ± 0.1763	0.9483 ± 0.0377	0.5705 ± 0.1579	0.4851 ± 0.1927	0.4634 ± 0.2031
	MFE	3.2715 ± 1.2436	1.7271 ± 1.3924	1.1867 ± 1.4667	1.0001 ± 1.7633	6.2896 ± 0.8240	2.2576 ± 1.3099	1.5992 ± 1.8947	1.4612 ± 1.4458
Nuclear	AFE	79.6284 ± 12.5129	58.1602 ± 14.0585	50.1646 ± 19.4004	47.1060 ± 19.2026	209.4597 ± 12.3611	122.0572 ± 36.6691	103.4399 ± 47.9068	106.7587 ± 65.1436
	RFE	0.7729 ± 0.1113	0.5656 ± 0.1397	0.4824 ± 0.1879	0.4458 ± 0.2601	0.9966 ± 0.0542	0.6408 ± 0.1837	0.4465 ± 0.2279	0.5132 ± 0.3205
	MFE	3.9629 ± 1.2688	2.1141 ± 1.0384	1.5728 ± 1.5795	1.3869 ± 2.1682	7.2163 ± 0.7761	2.4620 ± 1.3350	1.8642 ± 2.0745	1.78080 ± 3.3428
		$d = 120$				$d = 160$			
		$m = 100$	$m = 300$	$m = 500$	$m = 700$	$m = 100$	$m = 300$	$m = 500$	$m = 800$
SCAD	AFE	290.3662 ± 13.0428	233.1275 ± 34.3399	160.9030 ± 39.5538	145.6902 ± 44.8508	416.1755 ± 13.1164	371.2838 ± 20.1159	270.8453 ± 40.8805	183.1849 ± 68.6043
	RFE	0.9958 ± 0.0215	0.7469 ± 0.1097	0.54620 ± 0.1885	0.4508 ± 0.1429	1.0001 ± 0.0122	0.8987 ± 0.0435	0.6989 ± 0.0990	0.4396 ± 0.1666
	MFE	6.6894 ± 0.5615	3.7742 ± 1.0867	2.0283 ± 1.7619	1.4740 ± 0.9685	6.7657 ± 0.4275	5.4722 ± 0.5948	3.0043 ± 0.8789	1.3108 ± 1.2793
MCP	AFE	291.0940 ± 11.3085	231.1940 ± 34.4239	168.9255 ± 45.8193	145.3658 ± 52.7231	417.0620 ± 13.0938	$379.877326.4551$	273.8305 ± 41.9579	184.9737 ± 72.2468
	RFE	0.9915 ± 0.0234	0.7325 ± 0.1089	0.5424 ± 0.1502	0.4648 ± 0.2616	0.9966 ± 0.0121	0.9033 ± 0.0543	0.7293 ± 0.1200	0.4698 ± 0.1680
	MFE	6.7208 ± 0.4928	3.7119 ± 1.0580	1.9817 ± 1.0836	1.4674 ± 1.4120	6.7946 ± 0.4241	5.6370 ± 0.7717	3.0960 ± 0.9044	1.4850 ± 1.3038
Nuclear	AFE	292.3386 ± 8.9425	240.4273 ± 35.7671	189.4290 ± 50.5312	162.2422 ± 64.2535	416.5137 ± 15.3655	374.2606 ± 24.9383	282.5598 ± 58.3104	192.6122 ± 94.5904
	RFE	1.0010 ± 0.0174	0.7728 ± 0.1072	0.6043 ± 0.2912	0.5190 ± 0.2086	1.0014 ± 0.0130	0.9004 ± 0.0548	0.6810 ± 0.1414	0.46160 ± 0.1530
	MFE	6.7747 ± 0.3910	4.0143 ± 1.2636	2.4919 ± 3.6096	1.8280 ± 1.6649	6.7767 ± 0.4999	5.4715 ± 0.7326	3.1188 ± 1.3402	1.4492 ± 1.1249

Table 1 presents the performance of our proposed methods with different penalties with different dimensions of the covariance matrix. The statistical results of SCAD, MCP, and Nuclear Norm methods are shown in the table with different estimation standards and the standard variance. For each experiment, the rank of the covariance matrix is set to 10, and the variance proxy $\delta = 0.1$. The dimension of the matrix to be estimated changes in the range of $d = 40, 80, 120, 160$. And different sample numbers for convergence of each dimension can be observed from the results in the table.

6. Conclusion and Discussion

In this paper, we investigate compressive covariance sensing and introduce a quadratic measurement model tailored for estimating high-dimensional low-rank covariance matrices. This model effectively addresses the inherent challenges of limited measurements, as well as the constraints on storage and computational resources typically in high-dimensional settings. We propose a novel non-convex penalization approach and employ a proximal gradient homotopy algorithm to solve the resulting estimator. Notably, our proposed estimator achieves superior statistical convergence rates compared to traditional nuclear norm penalized estimators and attains the oracle property. These advancements are corroborated through comprehensive theoretical analysis and empirical experiments. As a limitation, we note that our study is the absence of performance assessments on real-world datasets, such as those encountered in spectrum analysis. Future work will focus on applying our estimator to publicly available real-world datasets to further validate its effectiveness in practical scenarios.

References

[1] Wei-Liem Loh. *Estimating covariance matrices*. Stanford University, 1988.

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

[3] Arup Bose and Monika Bhattacharjee. *Large covariance and autocovariance matrices*. Chapman and Hall/CRC, 2018.

[4] Aygul Zagidullina. *High-dimensional covariance matrix estimation: an introduction to random matrix theory*. Springer Nature, 2021.

[5] Ted H Szatrowski. Necessary and sufficient conditions for explicit solutions in the multivariate normal estimation problem for patterned means and covariances. *The Annals of Statistics*, pages 802–810, 1980.

[6] Hsien-sen Hung and Mostafa Kaveh. On the statistical sufficiency of the coherently averaged covariance matrix for the estimation of the parameters of wideband sources. In *ICASSP'87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 12, pages 33–36. IEEE, 1987.

[7] Jianfeng Yao, Shurong Zheng, and ZD Bai. Sample covariance matrices and high-dimensional data analysis. *Cambridge UP, New York*, 2015.

[8] Emile Richard, Pierre-André Savalle, and Nicolas Vayatis. Estimation of simultaneously sparse and low rank matrices. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 51–58, 2012.

[9] Azer P Shikhaliev, Lee C Potter, and Yuejie Chi. Low-rank structured covariance matrix estimation. *IEEE Signal Processing Letters*, 26(5):700–704, 2019.

[10] Quan Wei and Ziping Zhao. Large covariance matrix estimation with oracle statistical rate via majorization-minimization. *IEEE Transactions on Signal Processing*, 2023.

[11] Dyonisius Dony Ariananda and Geert Leus. Wideband power spectrum sensing using sub-Nyquist sampling. In *2011 IEEE 12th International Workshop on Signal Processing Advances in Wireless Communications*, pages 101–105. IEEE, 2011.

[12] Dyonisius Dony Ariananda and Geert Leus. Compressive Wideband Power Spectrum Estimation. *IEEE Transactions on Signal Processing*, 60(9):4775–4789, 2012.

[13] Dyonisius Dony Ariananda, Daniel Romero, and Geert Leus. Compressive angular and frequency periodogram reconstruction for multiband signals. In *2013 5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, pages 440–443, 2013.

[14] Daniel Romero, Dyonisius Dony Ariananda, Zhi Tian, and Geert Leus. Compressive covariance sensing: Structure-based compressive sensing beyond sparsity. *IEEE Signal Processing Magazine*, 33(1):78–93, 2015.

[15] Yuxin Chen, Yuejie Chi, and Andrea J. Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7): 4034–4059, 2015.

[16] Kai Zhong, Prateek Jain, and Inderjit S Dhillon. Efficient matrix sensing using rank-1 gaussian measurements. In *Algorithmic Learning Theory: 26th International Conference, ALT 2015*, pages 3–18. Springer, 2015.

[17] Xixian Chen, Michael R Lyu, and Irwin King. Toward efficient and accurate covariance matrix estimation on compressed data. In *International Conference on Machine Learning*, pages 767–776. PMLR, 2017.

[18] Angelika Rohde and Alexandre B. Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

[19] Karim Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014.

[20] Kishore Jaganathan, Samet Oymak, and Babak Hassibi. Recovery of sparse 1-d signals from the magnitudes of their fourier transform. In *2012 IEEE International Symposium on Information Theory Proceedings*, pages 1473–1477. IEEE, 2012.

[21] Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[22] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. *Advances in Neural Information Processing Systems*, 26, 2013.

[23] Amir Beck and Yonina C Eldar. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM Journal on Optimization*, 23(3):1480–1509, 2013.

[24] Emmanuel J Candès and Xiaodong Li. Solving quadratic equations via phaselift when there are about as many equations as unknowns. *Foundations of Computational Mathematics*, 5(14): 1017–1026, 2014.

[25] Irene Waldspurger, Alexandre d'Aspremont, and Stéphane Mallat. Phase recovery, maxcut and complex semidefinite programming. *Mathematical Programming*, 149:47–81, 2015.

[26] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[27] Feiping Nie, Hua Wang, Xiao Cai, Heng Huang, and Chris Ding. Robust matrix completion via joint schatten p-norm and lp-norm minimization. In *2012 IEEE 12th International Conference on Data Mining*, pages 566–574. IEEE, 2012.

[28] Kun Chen, Hongbo Dong, and Kung-Sik Chan. Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100(4):901–920, 2013.

[29] Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.

[30] Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302, 2011.

[31] Yan Huang, Guisheng Liao, Yijian Xiang, Lei Zhang, Jie Li, and Arye Nehorai. Low-rank approximation via generalized reweighted iterative nuclear and frobenius norms. *IEEE Transactions on Image Processing*, 29:2244–2257, 2020.

[32] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[33] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

[34] Dehua Liu, Tengfei Zhou, Hui Qian, Congfu Xu, and Zhihua Zhang. A nearly unbiased matrix completion approach. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013*, pages 210–225. Springer, 2013.

[35] Zhaoran Wang, Han Liu, and Tong Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *The Annals of Statistics*, 42(6):2164, 2014.

[36] Huan Gui, Jiawei Han, and Quanquan Gu. Towards faster rates and oracle property for low-rank matrix estimation. In *International Conference on Machine Learning*, pages 2300–2309. PMLR, 2016.

[37] Hu Shao, William HK Lam, Agachai Sumalee, Anthony Chen, and Martin L Hazelton. Estimation of mean and covariance of peak hour origin–destination demands from day-to-day traffic counts. *Transportation Research Part B: Methodological*, 68:52–75, 2014.

[38] Björn Ottersten, Peter Stoica, and Richard Roy. Covariance matching estimation techniques for array signal processing applications. *Digital Signal Processing*, 8(3):185–210, 1998.

[39] Kai Yu, Shenghuo Zhu, John Lafferty, and Yihong Gong. Fast nonparametric matrix factorization for large-scale collaborative filtering. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 211–218, 2009.

[40] Yu Zhang, Bin Cao, and Dit Yan Yeung. Multi-domain collaborative filtering. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, page 725, 2010.

[41] Ravishankar Sivalingam, Vassilios Morellas, Daniel Boley, and Nikolaos Papanikolopoulos. Metric learning for semi-supervised clustering of region covariance descriptors. In *2009 Third ACM/IEEE International Conference on Distributed Smart Cameras*, pages 1–8. IEEE, 2009.

[42] Aurélien Bellet, Amaury Habrard, and Marc Sebban. *Metric learning*. Morgan & Claypool Publishers, 2015.

[43] William A Gardner. Exploitation of spectral redundancy in cyclostationary signals. *IEEE Signal Processing Magazine*, 8(2):14–36, 1991.

[44] Kyouwoong Kim, Ihsan A Akbar, Kyung K Bae, Jung-Sun Um, Chad M Spooner, and Jeffrey H Reed. Cyclostationary approaches to signal detection and classification in cognitive radio. In *2007 2nd IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks*, pages 212–215. IEEE, 2007.

[45] Georgios B Giannakis. Cyclostationary signal analysis. In *Digital Signal Processing Fundamentals*, pages 433–464. CRC press, 2017.

[46] Benjamin Rolfs, Bala Rajaratnam, Dominique Guillot, Ian Wong, and Arian Maleki. Iterative thresholding algorithm for sparse inverse covariance estimation. *Advances in Neural Information Processing Systems*, 25, 2012.

[47] Wenfu Xia, Ziping Zhao, and Ying Sun. C-ISTA: Iterative shrinkage-thresholding algorithm for sparse covariance matrix estimation. In *IEEE Statistical Signal Processing Workshop*, pages 215–219, 2023.

[48] Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. 2012.

[49] Qiang Sun, Kean Ming Tan, Han Liu, and Tong Zhang. Graphical nonconvex optimization via an adaptive convex relaxation. In *International Conference on Machine Learning*, pages 4810–4817. PMLR, 2018.

A. Appendices

A.1. Technical Lemmata

Lemma 10. *Under Assumption 3, and assuming that $\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2 \in \mathcal{B}$, we establish*

$$\widetilde{f}\left(\mathbf{\Sigma}_2\right) \geq \widetilde{f}\left(\mathbf{\Sigma}_1\right) + \left\langle \nabla \widetilde{f}\left(\mathbf{\Sigma}_1\right), \mathbf{\Sigma}_2 - \mathbf{\Sigma}_1 \right\rangle + \frac{\rho^- - \zeta^-}{2} \left\| \mathbf{\Sigma}_2 - \mathbf{\Sigma}_1 \right\|_{\mathrm{F}}^2 .$$

Proof. Considering the eigen-decompositions of covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ given by

$$\mathbf{\Sigma}_1 = \mathbf{U}_1 \mathbf{\Lambda}_1 \mathbf{U}_1^\top, \mathbf{\Sigma}_2 = \mathbf{U}_2 \mathbf{\Lambda}_2 \mathbf{U}_2^\top,$$

where $\mathbf{\Lambda}_1 = \operatorname{diag}\left(\boldsymbol{\sigma}_1\right)$ and $\mathbf{\Lambda}_2 = \operatorname{diag}\left(\boldsymbol{\sigma}_2\right)$ are diagonal matrices containing eigenvalues $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$ of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, respectively.

Under (b) of Assumption 1, for each component $i = 1, 2, \ldots, d$ in $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$, we have

$$q_\lambda'\left(\left(\boldsymbol{\sigma}_1\right)_i\right) - q_\lambda'\left(\left(\boldsymbol{\sigma}_2\right)_i\right) \geq -\zeta^-\left(\left(\boldsymbol{\sigma}_1\right)_i - \left(\boldsymbol{\sigma}_2\right)_i\right),$$

which leads to

$$\left\langle \left(-\nabla Q_\lambda\left(\mathbf{\Sigma}_1\right)\right) - \left(-\nabla Q_\lambda\left(\mathbf{\Sigma}_2\right)\right), \mathbf{\Sigma}_1 - \mathbf{\Sigma}_2 \right\rangle \leq \zeta^- \left\| \mathbf{\Sigma}_1 - \mathbf{\Sigma}_2 \right\|_{\mathrm{F}}^2 .$$

This inequality illustrates the strong smoothness of $-Q\left(\cdot\right)$ in the region $\mathcal{B}$. As a consequence, we establish

$$Q_\lambda\left(\mathbf{\Sigma}_1\right) + \left\langle \nabla Q_\lambda\left(\mathbf{\Sigma}_1\right), \mathbf{\Sigma}_2 - \mathbf{\Sigma}_1 \right\rangle - \frac{\zeta^-}{2} \left\| \mathbf{\Sigma}_1 - \mathbf{\Sigma}_2 \right\|_{\mathrm{F}}^2 \leq Q_\lambda\left(\mathbf{\Sigma}_2\right). \tag{4}$$

Additionally, by Assumption 3, we obtain

$$f\left(\mathbf{\Sigma}_2\right) \geq f\left(\mathbf{\Sigma}_1\right) + \left\langle \nabla f\left(\mathbf{\Sigma}_1\right), \mathbf{\Sigma}_2 - \mathbf{\Sigma}_1 \right\rangle + \frac{\rho^-}{2} \left\| \mathbf{\Sigma}_2 - \mathbf{\Sigma}_1 \right\|_{\mathrm{F}}^2 . \tag{5}$$

Given that $\widetilde{f}\left(\mathbf{\Sigma}\right)$ is the sum of $f\left(\mathbf{\Sigma}\right)$ and $Q_\lambda\left(\mathbf{\Sigma}\right)$, we can combine 4 and 5 to achieve

$$\widetilde{f}\left(\mathbf{\Sigma}_2\right) \geq \widetilde{f}\left(\mathbf{\Sigma}_1\right) + \left\langle \nabla \widetilde{f}\left(\mathbf{\Sigma}_1\right), \mathbf{\Sigma}_2 - \mathbf{\Sigma}_1 \right\rangle + \frac{\rho^- - \zeta^-}{2} \left\| \mathbf{\Sigma}_2 - \mathbf{\Sigma}_1 \right\|_{\mathrm{F}}^2 .$$

This result demonstrates the restricted strong convexity of the objective function $\widetilde{f}\left(\mathbf{\Sigma}\right)$. □

Lemma 11. *Suppose Assumption 3 holds. If $\rho^- > \zeta^-$, and $\lambda \geq \frac{2}{m} \left\| \mathcal{A}^*\left(\boldsymbol{y} - \mathcal{A}\left(\mathbf{\Sigma}^*\right)\right) \right\|_2$, we have*

$$\left\| \mathcal{P}^\perp\left(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right) \right\|_* \leq 5 \left\| \mathcal{P}\left(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right) \right\|_* .$$

Proof. Given Lemma 10, we obtain

$$\underbrace{\left(\widetilde{f}(\widehat{\mathbf{\Sigma}}) + \lambda \|\widehat{\mathbf{\Sigma}}\|_*\right) - \left(\widetilde{f}\left(\mathbf{\Sigma}^*\right) + \lambda \left\|\mathbf{\Sigma}^*\right\|_*\right)}_{\mathrm{I}} \geq \underbrace{\left\langle \nabla \widetilde{f}\left(\mathbf{\Sigma}^*\right), \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^* \right\rangle}_{\mathrm{II}} + \underbrace{\lambda \|\widehat{\mathbf{\Sigma}}\|_* - \lambda \left\|\mathbf{\Sigma}^*\right\|_*}_{\mathrm{III}}. \tag{6}$$

In the following, we analyze the term II. Applying Hölder's inequality, we have

$$\begin{aligned}
&\left\langle \nabla \widetilde{f}\left(\mathbf{\Sigma}^*\right), \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^* \right\rangle \\
&= \left\langle \nabla \widetilde{f}\left(\mathbf{\Sigma}^*\right), \mathcal{P}\left(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right) \right\rangle + \left\langle \nabla \widetilde{f}\left(\mathbf{\Sigma}^*\right), \mathcal{P}^\perp\left(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right) \right\rangle \\
&\geq - \underbrace{\left\| \mathcal{P}\left(\nabla \widetilde{f}\left(\mathbf{\Sigma}^*\right)\right) \right\|_2}_{\mathrm{II}_1} \cdot \left\| \mathcal{P}\left(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right) \right\|_* - \underbrace{\left\| \mathcal{P}^\perp\left(\nabla \widetilde{f}\left(\mathbf{\Sigma}^*\right)\right) \right\|_2}_{\mathrm{II}_2} \cdot \left\| \mathcal{P}^\perp\left(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right) \right\|_* .
\end{aligned} \tag{7}$$

It remains to bound the terms I, II_1, II_2 and III, respectively.

For the term I, since $\widehat{\mathbf{\Sigma}}$ represents the optimal solution to the problem (1). This leads to

$$\left(\widetilde{f}\left(\widehat{\mathbf{\Sigma}}\right)+\lambda\left\|\widehat{\mathbf{\Sigma}}\right\|_*\right)-\left(\widetilde{f}\left(\mathbf{\Sigma}^*\right)+\lambda\left\|\mathbf{\Sigma}^*\right\|_*\right)\leq 0. \tag{8}$$

For the term II_1, considering the choice of λ, we have

$$\left\|\nabla f\left(\mathbf{\Sigma}^*\right)\right\|_2=\left\|\frac{1}{m}\mathcal{A}^*\left(\boldsymbol{y}-\mathcal{A}\left(\mathbf{\Sigma}^*\right)\right)\right\|_2\leq\frac{\lambda}{2}.$$

Based on (a) in Assumption 1 and $\lambda\geq\frac{2}{m}\left\|\mathcal{A}^*\left(\boldsymbol{y}-\mathcal{A}\left(\mathbf{\Sigma}^*\right)\right)\right\|_2$, we have $\left\|\nabla Q_\lambda\left(\mathbf{\Sigma}^*\right)\right\|_2\leq\lambda$, which leads to

$$\left\|\mathcal{P}\left(\nabla\widetilde{f}\left(\mathbf{\Sigma}^*\right)\right)\right\|_2=\left\|\mathcal{P}\left(\nabla f\left(\mathbf{\Sigma}^*\right)+\nabla Q_\lambda\left(\mathbf{\Sigma}^*\right)\right)\right\|_2\leq\frac{3\lambda}{2}. \tag{9}$$

For term II_2, we first note from the definition of $\mathcal{P}^\perp(\cdot)$ that $\mathcal{P}^\perp\left(\mathbf{\Sigma}^*\right)=\mathbf{0}$. Based on (c) in Assumption (1), it follows that $Q_\lambda\left(\mathcal{P}^\perp\left(\mathbf{\Sigma}^*\right)\right)=0$. This leads to

$$\left\|\mathcal{P}^\perp\left(\nabla\widetilde{f}\left(\mathbf{\Sigma}^*\right)\right)\right\|_2=\left\|\mathcal{P}^\perp\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)\right\|_2\leq\frac{\lambda}{2}. \tag{10}$$

Then we substitute (9) and (10) into (7), we have

$$\left\langle\nabla\widetilde{f}\left(\mathbf{\Sigma}^*\right),\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\right\rangle\geq-\frac{3\lambda}{2}\left\|\mathcal{P}\left(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\right)\right\|_*-\frac{\lambda}{2}\left\|\mathcal{P}^\perp\left(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\right)\right\|_*. \tag{11}$$

For the term III, projecting onto the complementary subspaces $\mathcal{T}$ and $\mathcal{T}^\perp$ yields the following decomposition

$$\begin{aligned}\lambda\|\widehat{\mathbf{\Sigma}}\|_*-\lambda\left\|\mathbf{\Sigma}^*\right\|_*&=\lambda\|\mathcal{P}(\widehat{\mathbf{\Sigma}})\|_*-\lambda\|\mathcal{P}\left(\mathbf{\Sigma}^*\right)\|_*+\lambda\|\mathcal{P}^\perp(\widehat{\mathbf{\Sigma}})\|_*-\lambda\|\mathcal{P}^\perp(\mathbf{\Sigma}^*)\|_*\\&\geq-\lambda\left\|\mathcal{P}\left(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\right)\right\|_*+\lambda\left\|\mathcal{P}^\perp\left(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\right)\right\|_*.\end{aligned} \tag{12}$$

Based on the inequalities in formulas (8), (11), and (12), when substituted into formula (6), we obtain

$$0\geq-\frac{3\lambda}{2}\left\|\mathcal{P}\left(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\right)\right\|_*-\frac{\lambda}{2}\left\|\mathcal{P}^\perp\left(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\right)\right\|_*-\lambda\left\|\mathcal{P}\left(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\right)\right\|_*+\lambda\left\|\mathcal{P}^\perp\left(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\right)\right\|_*.$$

From this inequality, we have

$$\left\|\mathcal{P}^\perp\left(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\right)\right\|_*\leq 5\left\|\mathcal{P}\left(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\right)\right\|_*,$$

which implies that $\widehat{\mathbf{\Delta}}=\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^*$ lies in the cone $\mathcal{B}$. □

A.2. Proof of Theorem 5

Proof. Applying Lemma 10 , we establish

$$\begin{aligned}\widetilde{f}\left(\widehat{\mathbf{\Sigma}}\right)&\geq\widetilde{f}\left(\mathbf{\Sigma}^*\right)+\left\langle\nabla\widetilde{f}\left(\mathbf{\Sigma}^*\right),\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\right\rangle+\frac{\rho^--\zeta^-}{2}\left\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\right\|_{\mathrm{F}}^2,\\\widetilde{f}\left(\mathbf{\Sigma}^*\right)&\geq\widetilde{f}\left(\widehat{\mathbf{\Sigma}}\right)+\left\langle\nabla\widetilde{f}\left(\widehat{\mathbf{\Sigma}}\right),\mathbf{\Sigma}^*-\widehat{\mathbf{\Sigma}}\right\rangle+\frac{\rho^--\zeta^-}{2}\left\|\mathbf{\Sigma}^*-\widehat{\mathbf{\Sigma}}\right\|_{\mathrm{F}}^2.\end{aligned}$$

By summing these two inequalities, we obtain

$$\left(\rho^--\zeta^-\right)\left\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\right\|_{\mathrm{F}}^2-\left\langle\nabla\widetilde{f}\left(\widehat{\mathbf{\Sigma}}\right)-\nabla\widetilde{f}\left(\mathbf{\Sigma}^*\right),\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\right\rangle\leq 0. \tag{13}$$

Considering the convexity of the nuclear norm $\|\cdot\|_*$, we deduce

$$\left\langle\lambda\widehat{\mathbf{W}}-\lambda\mathbf{W}^*,\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\right\rangle\geq 0, \tag{14}$$

where $\widehat{\mathbf{W}} \in \partial \left\| \widehat{\mathbf{\Sigma}} \right\|_*$ and $\mathbf{W}^* \in \partial \left\| \mathbf{\Sigma}^* \right\|_*$.

Since $\widehat{\mathbf{\Sigma}}$ is the optimal solution to (1), it satisfies the optimality condition:

$$\left\langle \nabla \widetilde{f}\left(\widehat{\mathbf{\Sigma}}\right) + \lambda \widehat{\mathbf{W}} + \widehat{\mathbf{\Xi}}, \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^* \right\rangle \leq 0, \tag{15}$$

where $\widehat{\mathbf{\Xi}} \in \partial \mathbb{I}_{\{\mathbf{\Sigma} \succeq \mathbf{0}\}}\left(\widehat{\mathbf{\Sigma}}\right)$.

Given the definition of the indicator function $\mathbb{I}_{\{\mathbf{\Sigma} \succeq \mathbf{0}\}}(\mathbf{\Sigma})$, it follows that:

$$\left\langle \widehat{\mathbf{\Xi}}, \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^* \right\rangle = 0. \tag{16}$$

By combining the inequalities (13), (14), (15) and (16), we obtain

$$\begin{aligned}
\left(\rho^- - \zeta^-\right) \left\| \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^* \right\|_{\mathrm{F}}^2 &\leq \left\langle \nabla \widetilde{f}\left(\widehat{\mathbf{\Sigma}}\right) - \nabla \widetilde{f}\left(\mathbf{\Sigma}^*\right), \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^* \right\rangle \\
&= \left\langle \nabla \widetilde{f}\left(\widehat{\mathbf{\Sigma}}\right) + \lambda \widehat{\mathbf{W}} + \widehat{\mathbf{\Xi}}, \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^* \right\rangle - \left\langle \widehat{\mathbf{\Xi}}, \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^* \right\rangle \\
&\quad - \left\langle \lambda \widehat{\mathbf{W}} - \lambda \mathbf{W}^*, \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^* \right\rangle - \left\langle \nabla \widetilde{f}\left(\mathbf{\Sigma}^*\right) + \lambda \mathbf{W}^*, \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^* \right\rangle \\
&\leq - \left\langle \nabla \widetilde{f}\left(\mathbf{\Sigma}^*\right) + \lambda \mathbf{W}^*, \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^* \right\rangle.
\end{aligned} \tag{17}$$

Consider the partitioning of the eigenvalues of $\mathbf{\Sigma}^*$ into three distinct sets:

- For indices $i \in \mathcal{S}_1$, the eigenvalues satisfy $\sigma_i^* \geq \alpha$.
- For indices $i \in \mathcal{S}_2$, we have $\alpha > \sigma_i^* > 0$.
- For indices $i \in \overline{\mathcal{S}}$, we have $\sigma_i^* = 0$, with $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$.

Based on this partition, we define two subspace of $\mathcal{T}$ associated with $\mathcal{S}_1$ and $\mathcal{S}_2$ as follows:

$$\mathcal{T}_{\mathcal{S}_1}\left(\mathbf{U}^*\right) \coloneqq \left\{ \mathbf{\Delta} \in \mathbb{R}^{d \times d} \mid \mathrm{row}\left(\mathbf{\Delta}\right) \subset \mathbf{U}_{\mathcal{S}_1}^* \text{ and } \mathrm{col}\left(\mathbf{\Delta}\right) \subset \mathbf{U}_{\mathcal{S}_1}^* \right\},$$

$$\mathcal{T}_{\mathcal{S}_2}\left(\mathbf{U}^*\right) \coloneqq \left\{ \mathbf{\Delta} \in \mathbb{R}^{d \times d} \mid \mathrm{row}\left(\mathbf{\Delta}\right) \subset \mathbf{U}_{\mathcal{S}_2}^* \text{ and } \mathrm{col}\left(\mathbf{\Delta}\right) \subset \mathbf{U}_{\mathcal{S}_2}^* \right\},$$

where $\mathbf{U}_{\mathcal{S}_1}^*$ is the sub-matrix of $\mathbf{U}^*$ formed by the rows indexed by $i \in \mathcal{S}_1$, and similarly, $\mathbf{U}_{\mathcal{S}_2}^*$ is the sub-matrix formed by the rows indexed by $i \in \mathcal{S}_2$. Corresponding to these subspaces, we define two projection operators $\mathcal{P}_{\mathcal{S}_1}(\cdot)$ and $\mathcal{P}_{\mathcal{S}_2}(\cdot)$, which project matrices onto the subspace $\mathcal{T}_{\mathcal{S}_1}$ and $\mathcal{T}_{\mathcal{S}_2}$, respectively. In the following, we decompose the RHS in (17) with respect to these three distinct components.

For $i \in \overline{\mathcal{S}}$, we proceed under (a) in Assumption 1, which ensures that $q_\lambda'(0) = 0$. This property enables us to express the gradient as follows:

$$\nabla Q_\lambda\left(\mathbf{\Sigma}^*\right) = \mathbf{U}^* q_\lambda'\left(\mathbf{\Gamma}^*\right) \mathbf{U}^{*\top},$$

where $\mathbf{\Gamma}^* \in \mathbb{R}^{r \times r}$ is the diagonal matrix with $\mathrm{diag}\left(\mathbf{\Gamma}^*\right) = \boldsymbol{\sigma}^*$. By projecting $\nabla Q_\lambda(\mathbf{\Sigma}^*)$ onto $\mathcal{T}^\perp$, we obtain

$$\begin{aligned}
\mathcal{P}^\perp\left(\nabla Q_\lambda\left(\mathbf{\Sigma}^*\right)\right) &= \left(\mathbf{I} - \mathbf{U}^* \mathbf{U}^{*\top}\right) \mathbf{U}^* q_\lambda'\left(\mathbf{\Gamma}^*\right) \mathbf{U}^{*\top} \left(\mathbf{I} - \mathbf{U}^* \mathbf{U}^{*\top}\right) \\
&= \left(\mathbf{U}^* - \mathbf{U}^*\right) q_\lambda'\left(\mathbf{\Gamma}^*\right) \left(\mathbf{U}^{*\top} - \mathbf{U}^{*\top}\right) \\
&= \mathbf{0}.
\end{aligned} \tag{18}$$

Subsequently, we decompose $\mathbf{W}^* \in \partial \left\| \mathbf{\Sigma}^* \right\|_*$ as:

$$\mathbf{W}^* = \mathcal{P}(\mathbf{W}^*) + \mathcal{P}^\perp(\mathbf{W}^*) = \mathbf{U}^* \mathbf{U}^{*\top} - \lambda^{-1} \mathcal{P}^\perp\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right), \tag{19}$$

which follows from the norm condition:

$$\left\|\lambda^{-1}\mathcal{P}^{\perp}\left(\nabla f\left(\mathbf{\Sigma}^{*}\right)\right)\right\|_{2} \leq \lambda^{-1}\left\|\nabla f\left(\mathbf{\Sigma}^{*}\right)\right\|_{2} = \lambda^{-1}m^{-1}\left\|\mathcal{A}^{*}\left(\boldsymbol{y}-\mathcal{A}\left(\mathbf{\Sigma}\right)\right)\right\|_{2} \leq 1.$$

Combining the results from (18) and (19), we obtain

$$\begin{aligned}\left\langle \mathcal{P}^{\perp}\left(\nabla\widetilde{f}\left(\mathbf{\Sigma}^{*}\right)+\lambda\mathbf{W}^{*}\right),\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^{*}\right\rangle &= \left\langle \mathcal{P}^{\perp}\left(\nabla f\left(\mathbf{\Sigma}^{*}\right)\right)+\lambda\mathcal{P}^{\perp}(\mathbf{W}^{*}),\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^{*}\right\rangle \\ &= \left\langle \mathbf{0},\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^{*}\right\rangle \\ &= 0. \end{aligned} \tag{20}$$

For $i \in \mathcal{S}_1$, under (d) in Assumption 1, since $\sigma_i^* \geq \alpha$, we have $q'_\lambda\left(\sigma_i^*\right) = -\lambda$. Consequently, we can project $\nabla Q_\lambda(\mathbf{\Sigma}^*)$ onto $\mathcal{T}_{\mathcal{S}_1}$ as follows:

$$\mathcal{P}_{\mathcal{S}_1}\left(\nabla Q_\lambda\left(\mathbf{\Sigma}^{*}\right)\right) = \mathbf{U}^*_{\mathcal{S}_1} q'_\lambda\left(\mathbf{\Gamma}^*_{\mathcal{S}_1}\right) {\mathbf{U}^*_{\mathcal{S}_1}}^{\top} = -\lambda \mathbf{U}^*_{\mathcal{S}_1}{\mathbf{U}^*_{\mathcal{S}_1}}^{\top},$$

where $\mathbf{\Gamma}^*_{\mathcal{S}_1} \in \mathbb{R}^{r\times r}$. Then, we have

$$\begin{aligned}&\left\langle \mathcal{P}_{\mathcal{S}_1}\left(\nabla\widetilde{f}\left(\mathbf{\Sigma}^{*}\right)+\lambda\mathbf{W}^{*}\right),\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^{*}\right\rangle \\ =&\left\langle \mathcal{P}_{\mathcal{S}_1}\left(\nabla f\left(\mathbf{\Sigma}^{*}\right)\right)+\mathcal{P}_{\mathcal{S}_1}\left(\lambda\mathbf{W}^{*}\right)+\mathcal{P}_{\mathcal{S}_1}\left(\nabla Q_\lambda\left(\mathbf{\Sigma}^{*}\right)\right),\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^{*}\right\rangle \\ =&\left\langle \mathcal{P}_{\mathcal{S}_1}\left(\nabla f\left(\mathbf{\Sigma}^{*}\right)\right)+\mathcal{P}_{\mathcal{S}_1}\left(\lambda\mathbf{U}^{*}\mathbf{U}^{*\top}\right)-\lambda\mathbf{U}^*_{\mathcal{S}_1}{\mathbf{U}^*_{\mathcal{S}_1}}^{\top},\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^{*}\right\rangle \\ =&\left\langle \mathcal{P}_{\mathcal{S}_1}\left(\nabla f\left(\mathbf{\Sigma}^{*}\right)\right),\mathcal{P}_{\mathcal{S}_1}\left(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^{*}\right)\right\rangle. \end{aligned} \tag{21}$$

By applying Hölder's inequality and define $\vartheta = \|\mathcal{P}_{S_1}(\nabla f(\mathbf{\Sigma}^*))\|_2$, we update (21) to

$$\left\langle \mathcal{P}_{\mathcal{S}_1}\left(\nabla\widetilde{f}\left(\mathbf{\Sigma}^{*}\right)+\lambda\mathbf{W}^{*}\right),\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^{*}\right\rangle \geq -\vartheta\left\|\mathcal{P}_{\mathcal{S}_1}\left(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^{*}\right)\right\|_{*}. \tag{22}$$

Since $\operatorname{rank}\left(\mathcal{P}_{\mathcal{S}_1}\left(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^{*}\right)\right) \leq |\mathcal{S}_1|$, we further refine (22) as

$$\left\langle \mathcal{P}_{\mathcal{S}_1}\left(\nabla\widetilde{f}\left(\mathbf{\Sigma}^{*}\right)+\lambda\mathbf{W}^{*}\right),\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^{*}\right\rangle \geq -\vartheta\sqrt{|\mathcal{S}_1|}\left\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^{*}\right\|_{\mathrm{F}}.$$

For $i \in \mathcal{S}_2$, applying Hölder's inequality, we derive

$$\begin{aligned}&\left\langle \mathcal{P}_{\mathcal{S}_2}\left(\nabla\widetilde{f}\left(\mathbf{\Sigma}^{*}\right)+\lambda\mathbf{W}^{*}\right),\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^{*}\right\rangle \\ =&\left\langle \mathcal{P}_{\mathcal{S}_2}\left(\nabla f\left(\mathbf{\Sigma}^{*}\right)+\nabla Q_\lambda\left(\mathbf{\Sigma}\right)+\lambda\mathbf{W}^{*}\right),\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^{*}\right\rangle \\ =&\left\langle \mathcal{P}_{\mathcal{S}_2}\left(\nabla f\left(\mathbf{\Sigma}^{*}\right)\right),\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^{*}\right\rangle+\left\langle \mathcal{P}_{\mathcal{S}_2}\left(\nabla Q_\lambda\left(\mathbf{\Sigma}\right)\right),\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^{*}\right\rangle+\left\langle \mathcal{P}_{\mathcal{S}_2}\left(\lambda\mathbf{W}^{*}\right),\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^{*}\right\rangle \\ \geq&-\left[\underbrace{\left\|\mathcal{P}_{\mathcal{S}_2}\left(\nabla f\left(\mathbf{\Sigma}^{*}\right)\right)\right\|_{2}}_{\mathrm{I}}+\underbrace{\left\|\mathcal{P}_{\mathcal{S}_2}\left(\nabla Q_\lambda\left(\mathbf{\Sigma}\right)\right)\right\|_{2}}_{\mathrm{II}}+\underbrace{\left\|\mathcal{P}_{\mathcal{S}_2}\left(\lambda\mathbf{W}^{*}\right)\right\|_{2}}_{\mathrm{III}}\right]\cdot\left\|\mathcal{P}_{\mathcal{S}_2}\left(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^{*}\right)\right\|_{2}. \end{aligned} \tag{23}$$

It remains to analysis the terms I, II and III in (23) individually.

For term I, we have

$$\left\|\mathcal{P}_{\mathcal{S}_2}\left(\nabla f\left(\mathbf{\Sigma}^{*}\right)\right)\right\|_{2} \leq \left\|\nabla f\left(\mathbf{\Sigma}^{*}\right)\right\|_{2} \leq \lambda. \tag{24}$$

For term II, we have

$$\left\|\mathcal{P}_{\mathcal{S}_2}\left(\nabla Q_\lambda\left(\mathbf{\Sigma}^{*}\right)\right)\right\|_{2} = \left\|\mathbf{U}^*_{\mathcal{S}_1} q'_\lambda\left(\mathbf{\Gamma}^*_{\mathcal{S}_1}\right){\mathbf{U}^*_{\mathcal{S}_2}}^{\top}\right\|_{2}.$$

Additionally, under (a) in Assumption 1, for any $i \in \mathcal{S}_2$, σ_i^* satisfies $\alpha > \sigma_i^* > 0$, we have $|q'_\lambda\left(\sigma_i^*\right)| \leq \lambda$. This leads to

$$\left\|\mathcal{P}_{\mathcal{S}_2}\left(\nabla Q_\lambda\left(\mathbf{\Sigma}\right)\right)\right\|_{2} \leq \lambda. \tag{25}$$

For term III, we have

$$\left\|\mathcal{P}_{\mathcal{S}_2}\left(\lambda \mathbf{W}^*\right)\right\|_2 \leq \left\|\mathcal{P}_{\mathcal{S}_2}\left(\lambda \mathbf{U}^* \mathbf{U}^{*\top}\right)\right\|_2 = \lambda. \tag{26}$$

By substituting (24), (25) and (26) into (23), we have

$$\left\langle \mathcal{P}_{\mathcal{S}_2}\left(\nabla \widetilde{f}\left(\boldsymbol{\Sigma}^*\right) + \lambda \mathbf{W}^*\right), \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^* \right\rangle \geq -3\lambda \left\|\mathcal{P}_{\mathcal{S}_2}\left(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\right)\right\|_2.$$

Since $\operatorname{rank}\left(\mathcal{P}_{\mathcal{S}_2}\left(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\right)\right) \leq |\mathcal{S}_2|$, we have

$$\left\langle \mathcal{P}_{\mathcal{S}_2}\left(\nabla \widetilde{f}\left(\boldsymbol{\Sigma}^*\right) + \lambda \mathbf{W}^*\right), \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^* \right\rangle \geq -3\lambda \sqrt{|\mathcal{S}_2|} \left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\right\|_F. \tag{27}$$

By substituting (20), (22) and (27) into (17), we have

$$\begin{aligned}
&\left(\rho^- - \zeta^-\right) \left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\right\|_{\mathrm{F}}^2 \\
\leq& - \left\langle \mathcal{P}_{\mathcal{S}_1}\left(\nabla \widetilde{f}\left(\boldsymbol{\Sigma}^*\right) + \lambda \mathbf{W}^*\right), \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^* \right\rangle - \left\langle \mathcal{P}_{\mathcal{S}_2}\left(\nabla \widetilde{f}\left(\boldsymbol{\Sigma}^*\right) + \lambda \mathbf{W}^*\right), \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^* \right\rangle \\
&- \left\langle \mathcal{P}^{\perp}\left(\nabla \widetilde{f}\left(\boldsymbol{\Sigma}^*\right) + \lambda \mathbf{W}^*\right), \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^* \right\rangle \\
\leq& \vartheta \sqrt{|\mathcal{S}_1|} \left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\right\|_{\mathrm{F}} + 3\lambda \sqrt{|\mathcal{S}_2|} \left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\right\|_F.
\end{aligned}$$

Consequently, we get

$$\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\right\|_{\mathrm{F}} \leq \frac{\vartheta \sqrt{|\mathcal{S}_1|}}{\rho^- - \zeta^-} + \frac{3\lambda \sqrt{|\mathcal{S}_2|}}{\rho^- - \zeta^-}.$$

□

A.3. Proof of Corollary 7

Proof. According to the definitions of $f(\cdot)$, we have

$$\begin{aligned}
f\left(\widehat{\boldsymbol{\Sigma}}^O\right) - f\left(\boldsymbol{\Sigma}^*\right) &= \frac{1}{2m}\left\|\boldsymbol{y} - \mathcal{A}\left(\widehat{\boldsymbol{\Sigma}}^O\right)\right\|_2^2 - \frac{1}{2m}\left\|\boldsymbol{y} - \mathcal{A}\left(\boldsymbol{\Sigma}^*\right)\right\|_2^2 \\
&= \frac{1}{2m}\left\|\boldsymbol{\eta} - \mathcal{A}\left(\widehat{\boldsymbol{\Sigma}}^O - \boldsymbol{\Sigma}^*\right)\right\|_2^2 - \frac{1}{2m}\left\|\boldsymbol{\eta}\right\|_2^2 \\
&= \frac{1}{2m}\left\|\mathcal{A}\left(\widehat{\boldsymbol{\Sigma}}^O - \boldsymbol{\Sigma}^*\right)\right\|_2^2 - \frac{1}{m}\left\langle \mathcal{A}^*\left(\boldsymbol{y} - \boldsymbol{\mathcal{A}}\left(\boldsymbol{\Sigma}\right)\right), \widehat{\boldsymbol{\Sigma}}^O - \boldsymbol{\Sigma}^* \right\rangle \\
&= \frac{1}{2m}\left\|\mathcal{A}\left(\widehat{\boldsymbol{\Sigma}}^O - \boldsymbol{\Sigma}^*\right)\right\|_2^2 + \left\langle \nabla f\left(\boldsymbol{\Sigma}^*\right), \widehat{\boldsymbol{\Sigma}}^O - \boldsymbol{\Sigma}^* \right\rangle \leq 0,
\end{aligned} \tag{28}$$

where the last inequality applies due to $\widehat{\boldsymbol{\Sigma}}^O$ is the oracle estimator.

Next, applying the assumption 3, we have

$$f\left(\widehat{\boldsymbol{\Sigma}}^O\right) \geq f\left(\boldsymbol{\Sigma}^*\right) + \left\langle \nabla f\left(\boldsymbol{\Sigma}^*\right), \widehat{\boldsymbol{\Sigma}}^O - \boldsymbol{\Sigma}^* \right\rangle + \frac{\rho^-}{2}\left\|\widehat{\boldsymbol{\Sigma}}^O - \boldsymbol{\Sigma}^*\right\|_{\mathrm{F}}^2. \tag{29}$$

By combining inequalities (28) and (29), we obtain

$$\begin{aligned}
\frac{\rho^-}{2}\left\|\widehat{\boldsymbol{\Sigma}}^O - \boldsymbol{\Sigma}^*\right\|_{\mathrm{F}}^2 &\leq f\left(\widehat{\boldsymbol{\Sigma}}^O\right) - f\left(\boldsymbol{\Sigma}^*\right) - \left\langle \nabla f\left(\boldsymbol{\Sigma}^*\right), \widehat{\boldsymbol{\Sigma}}^O - \boldsymbol{\Sigma}^* \right\rangle \\
&= \frac{1}{2m}\left\|\mathcal{A}\left(\widehat{\boldsymbol{\Sigma}}^O - \boldsymbol{\Sigma}^*\right)\right\|_2^2 \\
&\leq - \left\langle \nabla f\left(\boldsymbol{\Sigma}^*\right), \widehat{\boldsymbol{\Sigma}}^O - \boldsymbol{\Sigma}^* \right\rangle.
\end{aligned}$$

This inequality can be further refined using Hölder's inequality as follows

$$\left\|\widehat{\mathbf{\Sigma}}^O - \mathbf{\Sigma}^*\right\|_{\mathrm{F}}^2 \leq \frac{2\left\langle \mathcal{P}\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right), \widehat{\mathbf{\Sigma}}^O - \mathbf{\Sigma}^*\right\rangle}{\rho^-} \leq \frac{2\left\|\mathcal{P}\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)\right\|_2 \cdot \left\|\widehat{\mathbf{\Sigma}}^O - \mathbf{\Sigma}^*\right\|_*}{\rho^-}. \tag{30}$$

Given that $\operatorname{rank}\left(\widehat{\mathbf{\Sigma}}^O - \mathbf{\Sigma}^*\right) = r$, we have $\left\|\widehat{\mathbf{\Sigma}}^O - \mathbf{\Sigma}^*\right\|_* \leq \sqrt{r}\left\|\widehat{\mathbf{\Sigma}}^O - \mathbf{\Sigma}^*\right\|_{\mathrm{F}}$. Using this in (30) gives our desired result

$$\left\|\widehat{\mathbf{\Sigma}}^O - \mathbf{\Sigma}^*\right\|_{\mathrm{F}}^2 \leq \frac{2\sqrt{r}\left\|\mathcal{P}\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)\right\|_2 \cdot \left\|\widehat{\mathbf{\Sigma}}^O - \mathbf{\Sigma}^*\right\|_{\mathrm{F}}}{\rho^-}.$$

Thus, we have

$$\left\|\widehat{\mathbf{\Sigma}}^O - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} \leq \frac{2\sqrt{r}\left\|\mathcal{P}\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)\right\|_2}{\rho^-}.$$

Next, we will prove that $\widehat{\mathbf{\Sigma}}^O = \widehat{\mathbf{\Sigma}}$ under our assumptions. From Lemma 10 , the following inequalities holds

$$\widetilde{f}\left(\widehat{\mathbf{\Sigma}}\right) \geq \widetilde{f}\left(\widehat{\mathbf{\Sigma}}^O\right) + \left\langle \nabla\widetilde{f}\left(\widehat{\mathbf{\Sigma}}^O\right), \widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Sigma}}^O\right\rangle + \frac{\rho^- - \zeta^-}{2}\left\|\widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Sigma}}^O\right\|_{\mathrm{F}}^2,$$
$$\widetilde{f}\left(\widehat{\mathbf{\Sigma}}^O\right) \geq \widetilde{f}\left(\widehat{\mathbf{\Sigma}}\right) + \left\langle \nabla\widetilde{f}\left(\widehat{\mathbf{\Sigma}}\right), \widehat{\mathbf{\Sigma}}^O - \widehat{\mathbf{\Sigma}}\right\rangle + \frac{\rho^- - \zeta^-}{2}\left\|\widehat{\mathbf{\Sigma}}^O - \widehat{\mathbf{\Sigma}}\right\|_{\mathrm{F}}^2.$$

By adding the above two inequality, we obtain

$$\begin{aligned}
&(\rho^- - \zeta^-)\left\|\widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Sigma}}^O\right\|_{\mathrm{F}}^2 \\
\leq& \left\langle \nabla\widetilde{f}\left(\widehat{\mathbf{\Sigma}}\right) - \nabla\widetilde{f}\left(\widehat{\mathbf{\Sigma}}^O\right), \widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Sigma}}^O\right\rangle \\
=& \left\langle \nabla\widetilde{f}\left(\widehat{\mathbf{\Sigma}}\right) + \lambda\widehat{\mathbf{W}}, \widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Sigma}}^O\right\rangle - \left\langle \nabla\widetilde{f}\left(\widehat{\mathbf{\Sigma}}^O\right) + \lambda\widehat{\mathbf{W}}^O, \widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Sigma}}^O\right\rangle - \left\langle \lambda\widehat{\mathbf{W}} - \lambda\widehat{\mathbf{W}}^O, \widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Sigma}}^O\right\rangle \\
=& \underbrace{\left\langle \nabla\widetilde{f}\left(\widehat{\mathbf{\Sigma}}\right) + \lambda\widehat{\mathbf{W}}, \widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Sigma}}^O\right\rangle}_{\mathrm{I}} - \underbrace{\left\langle \nabla\widetilde{f}\left(\widehat{\mathbf{\Sigma}}^O\right) + \lambda\widehat{\mathbf{W}}^O, \mathcal{P}\left(\widehat{\mathbf{\Sigma}}^O - \widehat{\mathbf{\Sigma}}\right)\right\rangle}_{\mathrm{III}} \\
&- \underbrace{\left\langle \nabla\widetilde{f}\left(\widehat{\mathbf{\Sigma}}^O\right) + \lambda\widehat{\mathbf{W}}^O, \mathcal{P}^\perp\left(\widehat{\mathbf{\Sigma}}^O - \widehat{\mathbf{\Sigma}}\right)\right\rangle}_{\mathrm{III}} - \underbrace{\left\langle \lambda\widehat{\mathbf{W}} - \lambda\widehat{\mathbf{W}}^O, \widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Sigma}}^O\right\rangle}_{\mathrm{IV}}.
\end{aligned}$$

Then we analysis the terms I, II, III and IV, respectively.

Since $\widehat{\mathbf{\Sigma}}$ is the optimal solution of (1), it satisfies the optimality condition. Therefore, we have

$$\left\langle \nabla\widetilde{f}\left(\widehat{\mathbf{\Sigma}}\right) + \lambda\widehat{\mathbf{W}} + \widehat{\mathbf{\Xi}}, \widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Sigma}}^O\right\rangle \leq 0, \tag{31}$$

where $\widehat{\mathbf{\Xi}} \in \partial\mathbb{I}_{\{\mathbf{\Sigma}\succeq\mathbf{0}\}}\left(\widehat{\mathbf{\Sigma}}\right)$. According to the definition of $\mathbb{I}_{\{\mathbf{\Sigma}\succeq\mathbf{0}\}}\left(\mathbf{\Sigma}\right)$, we have

$$\left\langle \widehat{\mathbf{\Xi}}, \widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Sigma}}^O\right\rangle = 0. \tag{32}$$

By subtracting (32) from (31), we obtain

$$\left\langle \nabla\widetilde{f}\left(\widehat{\mathbf{\Sigma}}\right) + \lambda\widehat{\mathbf{W}}, \widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Sigma}}^O\right\rangle \leq 0. \tag{33}$$

Based on the definition of the oracle estimator, the eigen-decompositions of $\widehat{\mathbf{\Sigma}}^O$ is $\widehat{\mathbf{\Sigma}}^O = \mathbf{U}^*\widehat{\mathbf{\Gamma}}^O\mathbf{U}^{*\top}$.
Next, we decompose $\widehat{\mathbf{W}}^O \in \partial\left\|\widehat{\mathbf{\Sigma}}^O\right\|_*$ as

$$\widehat{\mathbf{W}}^O = \mathcal{P}(\widehat{\mathbf{W}}^O) + \mathcal{P}^\perp(\widehat{\mathbf{W}}^O) = \mathbf{U}^*\mathbf{U}^{*\top} + \widehat{\mathbf{Z}}^O,$$

where $\widehat{\mathbf{Z}}^O \in \mathcal{T}^\perp$ and $\left\|\widehat{\mathbf{Z}}^O\right\|_2 \leq 1$. Since $\sum_{i=1}^d p_\lambda\left(\sigma_i\left(\mathbf{\Sigma}\right)\right) = \lambda\left\|\mathbf{\Sigma}\right\|_* + Q_\lambda\left(\mathbf{\Sigma}\right)$, we have

$$\begin{aligned}
\mathcal{P}\left(\nabla\left(\sum_{i=1}^d p_\lambda\left(\sigma_i\left(\widehat{\mathbf{\Sigma}}^O\right)\right)\right)\right) &= \mathcal{P}\left(\lambda\widehat{\mathbf{W}}^O + \nabla Q_\lambda\left(\widehat{\mathbf{\Sigma}}^O\right)\right) \\
&= \mathcal{P}\left(\lambda\left(\mathbf{U}^*\mathbf{U}^{*\top} + \widehat{\mathbf{Z}}^O\right) + \nabla Q_\lambda\left(\widehat{\mathbf{\Sigma}}^O\right)\right) \\
&= \mathcal{P}\left(\lambda\mathbf{U}^*\mathbf{U}^{*\top} + \nabla Q_\lambda\left(\widehat{\mathbf{\Sigma}}^O\right)\right) \\
&= \lambda\mathbf{U}^*\mathbf{U}^{*\top} + \mathbf{U}^* q'_\lambda\left(\left(\widehat{\mathbf{\Gamma}}^O\right)_{\mathcal{S}}\right)\mathbf{U}^{*\top} \\
&= \mathbf{U}^*\left(\lambda\mathbf{I} + q'_\lambda\left(\left(\widehat{\mathbf{\Gamma}}^O\right)_{\mathcal{S}}\right)\right)\mathbf{U}^{*\top}, \qquad (34)
\end{aligned}$$

and the diagonal matrix $\left(\lambda\mathbf{I} + q'_\lambda\left(\left(\widehat{\mathbf{\Gamma}}^O\right)_{\mathcal{S}}\right)\right)$ satisfies

$$\left[\lambda\mathbf{I} + q'_\lambda\left(\left(\widehat{\mathbf{\Gamma}}^O\right)_{\mathcal{S}}\right)\right]_{ii} = \lambda + q'_\lambda\left(\sigma_i\left(\widehat{\mathbf{\Sigma}}^O\right)\right) = p'_\lambda\left(\sigma_i\left(\widehat{\mathbf{\Sigma}}^O\right)\right), \qquad (35)$$

where $i \in \mathcal{S}$. According to the perturbation bounds for eigenvalues, we have

$$\begin{aligned}
\min_{i\in\mathcal{S}}\left|\sigma_i\left(\widehat{\mathbf{\Sigma}}^O\right)\right| &= \min_{i\in\mathcal{S}}\left|\sigma_i\left(\widehat{\mathbf{\Sigma}}^O\right) - \sigma_i^* + \sigma_i^*\right| \\
&\geq -\max_{i\in\mathcal{S}}\left|\sigma_i\left(\widehat{\mathbf{\Sigma}}^O\right) - \sigma_i^*\right| + \min_{i\in\mathcal{S}}\left|\sigma_i^*\right| \\
&\geq -\left\|\widehat{\mathbf{\Sigma}}^O - \mathbf{\Sigma}^*\right\|_2 + \min_{i\in\mathcal{S}}\left|\sigma_i^*\right| \\
&\geq -\left\|\widehat{\mathbf{\Sigma}}^O - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} + \min_{i\in\mathcal{S}}\left|\sigma_i^*\right| \\
&\geq -\frac{2\sqrt{r}\left\|\mathcal{P}\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)\right\|_2}{\rho^-} + \min_{i\in\mathcal{S}}\left|\sigma_i^*\right|. \qquad (36)
\end{aligned}$$

Since $\boldsymbol{\sigma}^*$ satisfies $\min_{i\in\mathcal{S}}\left|\sigma_i^*\right| \geq \alpha + \frac{2\sqrt{r}\|\mathcal{A}^*(\boldsymbol{y}-\boldsymbol{\mathcal{A}}(\mathbf{\Sigma}))\|_2}{m\rho^-} = \alpha + \frac{2\sqrt{r}\|\mathcal{P}(\nabla f(\mathbf{\Sigma}^*))\|_2}{\rho^-}$, (36) can be updated to

$$\min_{i\in\mathcal{S}}\left|\sigma_i\left(\widehat{\mathbf{\Sigma}}^O\right)\right| \geq \alpha.$$

Given $\widehat{\mathbf{\Sigma}}^O \in \mathcal{T}$, we have $\operatorname{rank}\left(\widehat{\mathbf{\Sigma}}^O\right) = r$, then $\sigma_i\left(\widehat{\mathbf{\Sigma}}^O\right)$ satisfies

$$\sigma_i\left(\widehat{\mathbf{\Sigma}}^O\right)\begin{cases}\geq \alpha & i \in \mathcal{S}, \\ = 0 & i \notin \mathcal{S}.\end{cases}$$

Under (d) in Assumption 1, for $i \in \mathcal{S}$, we have

$$p'_\lambda\left(\sigma_i\left(\widehat{\mathbf{\Sigma}}^O\right)\right) = 0. \qquad (37)$$

Then combining (34), (35) and (37), we get

$$\mathcal{P}\left(\nabla\left(\sum_{i=1}^d p_\lambda\left(\sigma_i\left(\widehat{\mathbf{\Sigma}}^O\right)\right)\right)\right) = \mathbf{0}.$$

Since $\widehat{\mathbf{\Sigma}}^O$ is the oracle estimator, it satisfies the optimality condition. Therefore, we have the following inequality

$$\left\langle\nabla f\left(\widehat{\mathbf{\Sigma}}^O\right) + \widehat{\mathbf{\Xi}}^O, \widehat{\mathbf{\Sigma}}^O - \widehat{\mathbf{\Sigma}}\right\rangle \leq 0,$$

where $\widehat{\mathbf{\Xi}}^O \in \partial\mathbb{I}_{\{\mathbf{\Sigma}\succeq\mathbf{0}\}}\left(\widehat{\mathbf{\Sigma}}^O\right)$.

According to the definition of $\mathbb{I}_{\{\mathbf{\Sigma}\succeq\mathbf{0}\}}(\mathbf{\Sigma})$, we have

$$\left\langle \widehat{\mathbf{\Xi}}^O, \widehat{\mathbf{\Sigma}}^O - \widehat{\mathbf{\Sigma}} \right\rangle = 0.$$

Then we have that

$$\left\langle \nabla f\left(\widehat{\mathbf{\Sigma}}^O\right), \mathcal{P}\left(\widehat{\mathbf{\Sigma}}^O - \widehat{\mathbf{\Sigma}}\right) \right\rangle \leq \left\langle \nabla f\left(\widehat{\mathbf{\Sigma}}^O\right), \widehat{\mathbf{\Sigma}}^O - \widehat{\mathbf{\Sigma}} \right\rangle \leq 0.$$

Therefore, we have

$$\begin{aligned}
&\left\langle \nabla \widetilde{f}\left(\widehat{\mathbf{\Sigma}}^O\right) + \lambda \widehat{\mathbf{W}}^O, \mathcal{P}\left(\widehat{\mathbf{\Sigma}}^O - \widehat{\mathbf{\Sigma}}\right) \right\rangle \\
=&\left\langle \nabla f\left(\widehat{\mathbf{\Sigma}}^O\right), \mathcal{P}\left(\widehat{\mathbf{\Sigma}}^O - \widehat{\mathbf{\Sigma}}\right) \right\rangle + \left\langle \mathcal{P}\left(\nabla\left(\sum_{i=1}^{d} p_\lambda\left(\sigma_i\left(\widehat{\mathbf{\Sigma}}^O\right)\right)\right)\right), \mathcal{P}\left(\widehat{\mathbf{\Sigma}}^O - \widehat{\mathbf{\Sigma}}\right) \right\rangle \\
\leq& 0. \qquad (38)
\end{aligned}$$

By projecting $\nabla Q_\lambda\left(\widehat{\mathbf{\Sigma}}^O\right)$ into $\mathcal{T}^\perp$, we have

$$\begin{aligned}
\mathcal{P}^\perp\left(\nabla Q_\lambda\left(\widehat{\mathbf{\Sigma}}^O\right)\right) &= \left(\mathbf{I} - \mathbf{U}^*\mathbf{U}^{*\top}\right)\mathbf{U}^* q'_\lambda\left(\widehat{\mathbf{\Gamma}}^O\right)\mathbf{U}^{*\top}\left(\mathbf{I} - \mathbf{U}^*\mathbf{U}^{*\top}\right) \\
&= \left(\mathbf{U}^* - \mathbf{U}^*\right) q'_\lambda\left(\left(\widehat{\mathbf{\Gamma}}^O\right)_{\overline{\mathcal{S}}}\right)\left(\mathbf{U}^* - \mathbf{U}^*\right) \\
&= \mathbf{0}.
\end{aligned}$$

Under Assumption 4, it holds that

$$f\left(\mathbf{\Sigma}^*\right) + \left\langle \nabla f\left(\mathbf{\Sigma}^*\right), \widehat{\mathbf{\Sigma}}^O - \mathbf{\Sigma}^* \right\rangle + \frac{\rho^+}{2}\left\|\widehat{\mathbf{\Sigma}}^O - \mathbf{\Sigma}^*\right\|_{\mathrm{F}}^2 \geq f\left(\widehat{\mathbf{\Sigma}}^O\right).$$

$$f\left(\widehat{\mathbf{\Sigma}}^O\right) + \left\langle \nabla f\left(\widehat{\mathbf{\Sigma}}^O\right), \mathbf{\Sigma}^* - \widehat{\mathbf{\Sigma}}^O \right\rangle + \frac{\rho^+}{2}\left\|\mathbf{\Sigma}^* - \widehat{\mathbf{\Sigma}}^O\right\|_{\mathrm{F}}^2 \geq f\left(\mathbf{\Sigma}^*\right).$$

Adding the above two inequalities up, we get

$$\begin{aligned}
\rho^+\left\|\widehat{\mathbf{\Sigma}}^O - \mathbf{\Sigma}^*\right\|_{\mathrm{F}}^2 &\geq \left\langle \nabla f\left(\widehat{\mathbf{\Sigma}}^O\right) - \nabla f\left(\mathbf{\Sigma}^*\right), \widehat{\mathbf{\Sigma}}^O - \mathbf{\Sigma}^* \right\rangle \\
&\geq \left\|\nabla f\left(\widehat{\mathbf{\Sigma}}^O\right) - \nabla f\left(\mathbf{\Sigma}^*\right)\right\|_{\mathrm{F}} \cdot \left\|\widehat{\mathbf{\Sigma}}^O - \mathbf{\Sigma}^*\right\|_{\mathrm{F}},
\end{aligned}$$

which can be simplified as

$$\left\|\nabla f\left(\widehat{\mathbf{\Sigma}}^O\right) - \nabla f\left(\mathbf{\Sigma}^*\right)\right\|_{\mathrm{F}} \leq \rho^+\left\|\widehat{\mathbf{\Sigma}}^O - \mathbf{\Sigma}^*\right\|_{\mathrm{F}}.$$

Thus, we have

$$\begin{aligned}
\left\|\mathcal{P}^\perp\left(\nabla f\left(\widehat{\mathbf{\Sigma}}^O\right)\right)\right\|_2 &\leq \left\|\nabla f\left(\widehat{\mathbf{\Sigma}}^O\right)\right\|_2 \\
&\leq \left\|\nabla f\left(\widehat{\mathbf{\Sigma}}^O\right) - \nabla f\left(\mathbf{\Sigma}^*\right)\right\|_2 + \left\|\nabla f\left(\mathbf{\Sigma}^*\right)\right\|_2 \\
&\leq \left\|\nabla f\left(\widehat{\mathbf{\Sigma}}^O\right) - \nabla f\left(\mathbf{\Sigma}^*\right)\right\|_{\mathrm{F}} + \left\|\nabla f\left(\mathbf{\Sigma}^*\right)\right\|_2 \\
&\leq \rho^+\left\|\widehat{\mathbf{\Sigma}}^O - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} + \left\|\nabla f\left(\mathbf{\Sigma}^*\right)\right\|_2 \\
&\leq \frac{2\rho^+\sqrt{r}\left\|\mathcal{P}\left(\nabla f\left(\mathbf{\Sigma}^*\right)\right)\right\|_2}{\rho^-} + \left\|\nabla f\left(\mathbf{\Sigma}^*\right)\right\|_2 \\
&\leq \frac{2\rho^+\sqrt{r}\left\|\nabla f\left(\mathbf{\Sigma}^*\right)\right\|_2}{\rho^-} + \left\|\nabla f\left(\mathbf{\Sigma}^*\right)\right\|_2 \\
&= \frac{2\rho^+\sqrt{r}\left\|\mathcal{A}^*\left(\boldsymbol{y} - \boldsymbol{\mathcal{A}}\left(\mathbf{\Sigma}\right)\right)\right\|_2}{m\rho^-} + \frac{\left\|\mathcal{A}^*\left(\boldsymbol{y} - \boldsymbol{\mathcal{A}}\left(\mathbf{\Sigma}\right)\right)\right\|_2}{m}.
\end{aligned}$$

Under the condition that $\lambda \geq 2m^{-1}\left\|\mathcal{A}^*\left(\boldsymbol{y}-\boldsymbol{\mathcal{A}}\left(\boldsymbol{\Sigma}\right)\right)\right\|_2 + 2m^{-1}\sqrt{r}\rho^+\left\|\mathcal{A}^*\left(\boldsymbol{y}-\boldsymbol{\mathcal{A}}\left(\boldsymbol{\Sigma}\right)\right)\right\|_2/\rho^-$, the upper bound of $\left\|\mathcal{P}^{\perp}\left(\nabla f\left(\widehat{\boldsymbol{\Sigma}}^{O}\right)\right)\right\|_2$ satisfies

$$\left\|\mathcal{P}^{\perp}\left(\nabla f\left(\widehat{\boldsymbol{\Sigma}}^{O}\right)\right)\right\|_2 \leq \lambda.$$

So we can decompose $\widehat{\mathbf{W}}^{O} \in \partial\left\|\widehat{\boldsymbol{\Sigma}}^{O}\right\|_*$ as

$$\widehat{\mathbf{W}}^{O} = \mathcal{P}(\widehat{\mathbf{W}}^{O}) + \mathcal{P}^{\perp}(\widehat{\mathbf{W}}^{O}) = \mathbf{U}^*\mathbf{U}^{*\top} - \lambda^{-1}\mathcal{P}^{\perp}\left(\nabla f\left(\widehat{\boldsymbol{\Sigma}}^{O}\right)\right).$$

Then we get

$$\begin{aligned}
&\left\langle \nabla\widetilde{f}\left(\widehat{\boldsymbol{\Sigma}}^{O}\right) + \lambda\widehat{\mathbf{W}}^{O}, \mathcal{P}^{\perp}\left(\widehat{\boldsymbol{\Sigma}}^{O}-\widehat{\boldsymbol{\Sigma}}\right)\right\rangle \\
=&\left\langle \mathcal{P}^{\perp}\left(\nabla f\left(\widehat{\boldsymbol{\Sigma}}^{O}\right) + \lambda\widehat{\mathbf{W}}^{O}\right), \mathcal{P}^{\perp}\left(\widehat{\boldsymbol{\Sigma}}^{O}-\widehat{\boldsymbol{\Sigma}}\right)\right\rangle + \left\langle \mathcal{P}^{\perp}\left(\nabla Q_{\lambda}\left(\widehat{\boldsymbol{\Sigma}}^{O}\right)\right), \mathcal{P}^{\perp}\left(\widehat{\boldsymbol{\Sigma}}^{O}-\widehat{\boldsymbol{\Sigma}}\right)\right\rangle \\
=&\left\langle \mathbf{0}, \mathcal{P}^{\perp}\left(\widehat{\boldsymbol{\Sigma}}^{O}-\widehat{\boldsymbol{\Sigma}}\right)\right\rangle + \left\langle \mathbf{0}, \mathcal{P}^{\perp}\left(\widehat{\boldsymbol{\Sigma}}^{O}-\widehat{\boldsymbol{\Sigma}}\right)\right\rangle \\
=&0.
\end{aligned} \tag{39}$$

Additionally, since $\|\cdot\|_*$ is convex, $\widehat{\mathbf{W}} \in \partial\left\|\widehat{\boldsymbol{\Sigma}}\right\|_*$ and $\widehat{\mathbf{W}}^{O} \in \partial\left\|\boldsymbol{\Sigma}^*\right\|_*$, we have

$$\left\langle \lambda\widehat{\mathbf{W}} - \lambda\widehat{\mathbf{W}}^{O}, \widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}^{O}\right\rangle \geq 0. \tag{40}$$

Therefore, by adding up (33), (38), (39) and (40), we obtain

$$\left(\rho^- - \zeta^-\right)\left\|\widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}^{O}\right\|_{\mathrm{F}}^2 \leq 0.$$

Since $\rho^- > \zeta^-$, it follows that $\widehat{\boldsymbol{\Sigma}}^{O} = \widehat{\boldsymbol{\Sigma}}$. Thus, we have

$$\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\right\|_{\mathrm{F}} = \left\|\widehat{\boldsymbol{\Sigma}}^{O} - \boldsymbol{\Sigma}^*\right\|_{\mathrm{F}} \leq \frac{2\sqrt{r}\left\|\mathcal{P}\left(\nabla f\left(\boldsymbol{\Sigma}^*\right)\right)\right\|_2}{\rho^-}$$

□

A.4. Proof of Corollary 9

First, we give three essential Lemmas.

Lemma 12 (Proposition 1 in [29]). *Consider the sampling operator of $\boldsymbol{\Theta}$-ensemble, it holds with probability at least $1-2\exp\left(-m/32\right)$ that*

$$\frac{\left\|\mathcal{A}\left(\boldsymbol{\Delta}\right)\right\|_2}{\sqrt{m}} \geq \frac{1}{4}\left\|\sqrt{\boldsymbol{\Theta}}\mathrm{vec}\left(\boldsymbol{\Delta}\right)\right\|_2 - 24\pi\left(\boldsymbol{\Theta}\right)\sqrt{d/m}\left\|\boldsymbol{\Delta}\right\|_*.$$

Lemma 13 (Lemma 6 in [29]). *There're universal constants C_0, C_1 and C_2 such that*

$$\mathbb{P}\left[\frac{\left\|\mathcal{A}^*\left(\boldsymbol{y}-\mathcal{A}\left(\boldsymbol{\Sigma}\right)\right)\right\|_2}{m} \geq 2C_0\delta\varpi\sqrt{d/m}\right] \leq C_1\exp\left(-2dC_2\right).$$

Lemma 14 (Lemma D.1 in [49]). *Let $\mathbf{x}$ be a sub-Gaussian random vector with zero mean and covariance $\boldsymbol{\Sigma}^*$ and $\{\mathbf{x}_i\}_{i=1}^T$ be a collection of i.i.d. samples from $\mathbf{x}$. There exists some constants c_1, c_2, and z_0 such that for all z with $0 < z < z_0$, the sample covariance matrix $\boldsymbol{\Sigma}_T$ satisfies the following tail bound*

$$\mathbb{P}\left(\left|\Sigma^*_{ij} - \left(\Sigma_T\right)_{ij}\right| > z\right) \leq c_1\exp(-c_2Tz^2).$$

Then, we give the proof of Corollary 9.

Proof. By projecting $\left\|\widehat{\boldsymbol{\Delta}}\right\|_*$ into $\mathcal{T}$ and $\mathcal{T}^\perp$, we have

$$\left\|\widehat{\boldsymbol{\Delta}}\right\|_* = \left\|\mathcal{P}\left(\widehat{\boldsymbol{\Delta}}\right)\right\|_* + \left\|\mathcal{P}^\perp\left(\widehat{\boldsymbol{\Delta}}\right)\right\|_* \leq 6\left\|\mathcal{P}\left(\widehat{\boldsymbol{\Delta}}\right)\right\|_* = 6\sqrt{r}\left\|\mathcal{P}\left(\widehat{\boldsymbol{\Delta}}\right)\right\|_{\mathrm{F}} \leq 6\sqrt{r}\left\|\widehat{\boldsymbol{\Delta}}\right\|_{\mathrm{F}}.$$

Then according to Lemma 12, we have

$$\begin{aligned}\frac{\left\|\mathcal{A}\left(\widehat{\boldsymbol{\Delta}}\right)\right\|_2}{\sqrt{m}} &\geq \frac{1}{4}\left\|\sqrt{\boldsymbol{\Theta}}\mathrm{vec}\left(\widehat{\boldsymbol{\Delta}}\right)\right\|_2 - 24\pi\left(\boldsymbol{\Theta}\right)\sqrt{d/m}\left\|\widehat{\boldsymbol{\Delta}}\right\|_* \\ &\geq \frac{\sqrt{\lambda_{\min}\left(\boldsymbol{\Theta}\right)}}{4}\left\|\widehat{\boldsymbol{\Delta}}\right\|_{\mathrm{F}} - 144\sqrt{r}\pi\left(\boldsymbol{\Theta}\right)\sqrt{d/m}\left\|\widehat{\boldsymbol{\Delta}}\right\|_{\mathrm{F}}.\end{aligned}$$

Then for $m \geq \frac{C_1 r\pi^2(\boldsymbol{\Theta})d^2}{\lambda_{min}(\boldsymbol{\Theta})}$ where C_1 is sufficiently large such that

$$144\sqrt{r}\pi\left(\boldsymbol{\Theta}\right)\sqrt{d/m} \leq \frac{\sqrt{\lambda_{\min}\left(\boldsymbol{\Theta}\right)}}{8},$$

we have

$$\frac{\left\|\mathcal{A}\left(\widehat{\boldsymbol{\Delta}}\right)\right\|_2}{\sqrt{m}} \geq \frac{\sqrt{\lambda_{\min}\left(\boldsymbol{\Theta}\right)}}{8}\left\|\widehat{\boldsymbol{\Delta}}\right\|_{\mathrm{F}},$$

which implies that

$$\frac{\left\|\mathcal{A}\left(\widehat{\boldsymbol{\Delta}}\right)\right\|_2^2}{m} \geq \frac{\lambda_{\min}\left(\boldsymbol{\Theta}\right)}{64}\left\|\widehat{\boldsymbol{\Delta}}\right\|_{\mathrm{F}}^2,$$

for $\rho^- = C_2\lambda_{\min}\left(\boldsymbol{\Theta}\right)$, the above inequality can be updated to

$$\frac{\left\|\mathcal{A}\left(\widehat{\boldsymbol{\Delta}}\right)\right\|_2^2}{m} \geq \frac{\rho^-}{64C_2}\left\|\widehat{\boldsymbol{\Delta}}\right\|_{\mathrm{F}}^2,$$

when we set $C_2 = \frac{1}{32}$, the Restricted Strong Convexity of $\mathcal{A}(\cdot)$ is established. Additionally, we have

$$\left\|\mathcal{P}_{\mathcal{S}_1}\left(\nabla f\left(\boldsymbol{\Sigma}\right)\right)\right\|_2 = \left\|\mathbf{U}_{\mathcal{S}_1}^{*\top}\nabla f\left(\boldsymbol{\Sigma}\right)\mathbf{U}_{\mathcal{S}_1}^*\right\|_2.$$

According to Lemma 13, since $\mathbf{U}_{\mathcal{S}_1}^{*\top}\nabla f\left(\boldsymbol{\Sigma}\right)\mathbf{U}_{\mathcal{S}_1}^* \in \mathbb{R}^{d\times d}$, combining with Lemma 14, we have

$$\begin{aligned}\left\|\mathbf{U}^{*\top}\nabla f\left(\boldsymbol{\Sigma}\right)\mathbf{U}^*\right\|_2 &\leq 2C_0\delta\varpi\sqrt{d/(mT)}, \\ \left\|\mathbf{U}_{\mathcal{S}_1}^{*\top}\nabla f\left(\boldsymbol{\Sigma}\right)\mathbf{U}_{\mathcal{S}_1}^*\right\|_2 &\leq 2C_0\delta\varpi\sqrt{\left|\mathcal{S}_1\right|/(mT)},\end{aligned}$$

which holds with probability at least $1 - C_1\exp\left(-2\left|\mathcal{S}_1\right|C_2\right)$. Based on Theorem 5, since $\vartheta = \left\|\mathcal{P}_{S_1}(\nabla\widetilde{f}(\boldsymbol{\Sigma}^*))\right\|_2$, $\lambda = \left\|\nabla f(\boldsymbol{\Sigma}^*)\right\|_2$ and $\rho^- = C_2\lambda_{\min}\left(\boldsymbol{\Theta}\right)$

$$\begin{aligned}\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\right\|_{\mathrm{F}} &\leq \frac{2\vartheta\sqrt{|\mathcal{S}_1|}}{\rho^-} + \frac{6\lambda\sqrt{|\mathcal{S}_2|}}{\rho^-} \\ &\leq \frac{1}{C_2\lambda_{\min}\left(\boldsymbol{\Theta}\right)}\left(4C_0\delta\varpi\sqrt{\left|\mathcal{S}_1\right|/(mT)}\cdot\sqrt{|\mathcal{S}_1|} + 12C_0\delta\varpi\sqrt{d/(mT)}\cdot\sqrt{|\mathcal{S}_2|}\right) \\ &\leq \frac{\delta\varpi}{\lambda_{\min}\left(\boldsymbol{\Theta}\right)\sqrt{mT}}\left(4C_0\left|\mathcal{S}_1\right|/C_2 + 12C_0\sqrt{d}/C_2\right),\end{aligned}$$

which holds with probability at least $1 - C_1\exp\left(-2\left|\mathcal{S}_1\right|C_2\right)$. Therefore, by choosing appropriate value of C_0, C_1, C_2, C_5 and C_6, we get

$$\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\right\|_{\mathrm{F}} \leq \frac{\delta\varpi}{\sqrt{mT}\lambda_{\min}\left(\boldsymbol{\Theta}\right)}\left[C_1\left|\mathcal{S}_1\right| + C_2\sqrt{\left|\mathcal{S}_2\right|d}\right],$$

which holds with probability at least $1 - C_5 \exp(-2dC_6)$. Since $\rho^- = C_1 \lambda_{\min}(\mathbf{\Theta})$ and Lemma 13, according to Corollary 7, we have

$$\begin{aligned}\min_{i \in \mathcal{S}} |\sigma_i^*| &\geq \alpha + \frac{2\sqrt{r} \left\| \sum_{i=1}^m \eta_i \mathbf{A}_i \right\|_2}{m\rho^-} \\ &\geq \alpha + \frac{C_2 \delta \varpi \sqrt{rd/m}}{\lambda_{\min}(\mathbf{\Theta})} \\ &\geq \alpha + \frac{C_2 m^{-1} \left\| \sum_{i=1}^m \eta_i \mathbf{A}_i \right\|_2 \sqrt{r}}{2\rho^-},\end{aligned}$$

which holds with probability at least $1 - C_1 \exp(-2dC_2)$. With suitable value of C_2, the above inequality is equal to

$$\min_{i \in \mathcal{S}} |\sigma_i^*| \geq \alpha + \frac{2\sqrt{r} \left\| \sum_{i=1}^m \eta_i \mathbf{A}_i \right\|_2}{m\rho^-},$$

which holds with probability at least $1 - C_1 \exp(-2dC_2)$ and is the requirement of $\boldsymbol{\sigma}^*$ in Corollary 7. According to the definition of $f(\cdot)$, we have

$$\mathbf{H}_m = \frac{\sum_{i=1}^m \mathrm{vec}(\mathbf{A}_i) \mathrm{vec}(\mathbf{A}_i)^\top}{m},$$

where $\mathbf{H}_m$ is the Hessian matrix of $f(\cdot)$. Then we have $\mathbb{E}(\mathbf{H}_m) = \mathbf{\Theta}$.

By concentration, when m is sufficiently large, with high probability, $\lambda_{\max}(\mathbf{H}_m) \leq 2\lambda_{\max}(\mathbf{\Theta})$. Based on 4, we could obtain that $\rho^+ = \lambda_{\max}(\mathbf{H}_m)$. Then we have that when m is sufficiently large, with high probability, $\rho^+ \leq 2\lambda_{\max}(\mathbf{\Theta})$. Additionally, given $\rho^- = C_1 \lambda_{\min}(\mathbf{\Theta})$ and Lemma 13, we have

$$\begin{aligned}\lambda &\geq C_3 \delta \varpi \sqrt{d/m} \left(1 + \frac{\sqrt{r} \lambda_{\max}(\mathbf{\Theta})}{\lambda_{\min}(\mathbf{\Theta})}\right) \\ &\geq C_3 \delta \varpi \sqrt{d/m} \left(1 + \frac{\sqrt{r} \rho^+ C_1}{2\rho^-}\right) \\ &\geq \frac{C_3}{2C_1} m^{-1} \left\| \sum_{i=1}^m \eta_i \mathbf{A}_i \right\|_2 \left(1 + \frac{\sqrt{r} \rho^+ C_1}{2\rho^-}\right),\end{aligned}$$

which holds with probability at least $1 - C_1 \exp(-2dC_2)$. With suitable value of C_1 and C_3, the above inequality is equal to

$$\lambda \geq 2m^{-1} \left\| \sum_{i=1}^m \eta_i \mathbf{A}_i \right\|_2 + 2m^{-1} \sqrt{r} \rho^+ \left\| \sum_{i=1}^m \eta_i \mathbf{A}_i \right\|_2 / \rho^-,$$

which holds with probability at least $1 - C_1 \exp(-2dC_2)$ and is the requirement of λ in Corollary 7. Since the requirements of $\boldsymbol{\sigma}^*$ and λ are reached with probability at least $1 - C_1 \exp(-2dC_2)$, according to Corollary 7, we have that $\widehat{\mathbf{\Sigma}}$ is the optimal solution of the estimator, $\mathrm{rank}\left(\widehat{\mathbf{\Sigma}}\right) = \mathrm{rank}(\mathbf{\Sigma}^*) = r$, with probablity at least $1 - C_4 \exp(-C_5 d)$ by setting the proper values of C_4 and C_5.

In addition, we have

$$\begin{aligned}\left\| \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^* \right\|_{\mathrm{F}} &\leq \frac{2\sqrt{r} \left\| \mathcal{P}_{S_1}(\nabla \widetilde{f}(\mathbf{\Sigma}^*)) \right\|_2}{\rho^-} \\ &\leq \frac{2\sqrt{r} \left\| \mathcal{P}_{S_1}(\nabla \widetilde{f}(\mathbf{\Sigma}^*)) \right\|_2}{C_1 \lambda_{\min}(\mathbf{\Theta})} \\ &\leq \frac{4\sqrt{r} C_0 \delta \varpi \sqrt{r/(mT)}}{C_1 \lambda_{\min}(\mathbf{\Theta})} \\ &\leq \frac{4 C_0 r \delta \varpi}{C_1 \sqrt{mT} \lambda_{\min}(\mathbf{\Theta})} \\ &\leq \frac{C_6 r \delta \varpi}{\sqrt{mT} \lambda_{\min}(\mathbf{\Theta})},\end{aligned}$$

which holds with probability at least $1 - C_4 \exp(-C_5 d)$ with suitable value of C_1, C_2, C_4, C_5 and C_6. $\square$