On Sparse High-Dimensional Graph Estimation from Multi-Attribute Data

Jitendra K. Tugnait
Dept. of Electrical & Computer Eng.
Auburn University, Auburn, AL 36849, USA

2024 58th Asilomar Conference on Signals, Systems, and Computers | 979-8-3503-5405-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/IEEECONF60004.2024.10942781

***Abstract*—We consider the problem of inferring the conditional independence graph (CIG) of high-dimensional Gaussian vectors from multi-attribute data. Most existing methods for graph estimation are based on single-attribute models where one associates a scalar random variable with each node. In multi-attribute graphical models, each node represents a random vector. In this paper we provide a unified theoretical analysis of multi-attribute graph learning using a penalized log-likelihood objective function. We consider both convex (sparse-group lasso) and non-convex (log-sum and SCAD group penalties) penalty/regularization functions. We establish sufficient conditions in a high-dimensional setting for consistency (convergence of the precision matrix to true value in the Frobenius norm), local convexity when using non-convex penalties, and graph recovery. We do not impose any incoherence or irrepresentability condition for our convergence results.**

I. Introduction

Graphical models provide a powerful tool for analyzing multivariate data [1], [2]. In an undirected graphical model, the conditional dependency structure among p random variables $x_1, x_2, \cdots, x_p$, ($\boldsymbol{x} = [x_1\ x_2\ \cdots\ x_p]^\top$), is represented using an undirected graph $\mathcal{G} = (V, \mathcal{E})$, where $V = \{1, 2, \cdots, p\} = [p]$ is the set of p nodes corresponding to the p random variables x_i's, and $\mathcal{E} \subseteq [p] \times [p]$ is the set of undirected edges describing conditional dependencies among x_i's. The graph $\mathcal{G}$ then is a conditional independence graph (CIG) where there is no edge between nodes i and j iff x_i and x_j are conditionally independent given the remaining p-2 variables.

Gaussian graphical models (GGMs) are CIGs where $\boldsymbol{x}$ is multivariate Gaussian. Suppose $\boldsymbol{x}$ has positive-definite covariance matrix $\boldsymbol{\Sigma}$ with inverse covariance matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. Then Ω_{ij}, the (i, j)-th element of $\boldsymbol{\Omega}$, is zero iff x_i and x_j are conditionally independent. Given n samples of $\boldsymbol{x}$, in high-dimensional settings, one estimates $\boldsymbol{\Omega}$ under some sparsity constraints; see, e.g., [3]. In these graphs each node represents a scalar random variable. In many applications, there may be more than one random variable associated with a node. This class of graphical models has been called multi-attribute graphical models in [4], [5]. In [5], a sparse-group lasso [6], [7] based penalized log-likelihood approach for graph learning from multi-attribute data was presented whereas [4] consider only group lasso [8]. Both sparse-group lasso and group lasso are convex penalties. It is well-known that use of non-convex penalties such as Smoothly Clipped Absolute Deviation (SCAD) [9], [10] or log-sum [11], can yield more accurate results. Such penalties can produce sparse set of solution like lasso, and approximately unbiased coefficients for large coefficients, unlike lasso.

This work was supported by NSF Grant CCF-2308473. Author's email: tugnajk@auburn.edu

Contributions: In this paper we provide a unified theoretical analysis of multi-attribute graph learning using a penalized log-likelihood objective function. We consider both convex (sparse-group lasso) and non-convex (log-sum and SCAD group penalties) penalty/regularization functions. We establish sufficient conditions in a high-dimensional setting for consistency (convergence of the precision matrix to true value in the Frobenius norm), local convexity when using non-convex penalties, and graph recovery. We do not impose any incoherence or irrepresentability condition for our convergence results, unlike [4] where only group lasso is considered. In [4] the primal-dual witness technique of [12] is followed whereas we follow the proof technique of [13] (as in [5]). The SCAD penalty for multi-attribute graphs has been considered in [14] but it does not have counterparts to our Lemma 1 and Theorems 2 and 3. Moreover, the sparse-group SCAD penalty used in this paper is different than that in [14]. We provide only a theoretical analysis in this paper. Numerical optimization of the penalized log-likelihood can be done using an ADMM approach [15] as in [5], [14], where for non-convex penalties (SCAD or log-sum), one uses a local linear approximation of the penalties ([10], [14]) or reweighted ℓ_1 minimization [11], initialized via sparse-group lasso results of [5].

Notation: Given $\boldsymbol{A} \in \mathbb{R}^{p \times p}$, we use $\phi_{\min}(\boldsymbol{A})$, $\phi_{\max}(\boldsymbol{A})$, $|\mathbf{A}|$ and $\mathrm{tr}(\boldsymbol{A})$ to denote the minimum eigenvalue, maximum eigenvalue, determinant and trace of $\boldsymbol{A}$, respectively. For $\boldsymbol{B} \in \mathbb{R}^{p \times q}$, we have $\|\boldsymbol{B}\| = \sqrt{\phi_{\max}(\boldsymbol{B}^\top \boldsymbol{B})}$, $\|\boldsymbol{B}\|_F = \sqrt{\mathrm{tr}(\boldsymbol{B}^\top \boldsymbol{B})}$ and $\|\boldsymbol{B}\|_1 = \sum_{i,j} |B_{ij}|$ where B_{ij} is the (i, j)-th element of $\boldsymbol{B}$ (also denoted by $[\boldsymbol{B}]_{ij}$). Given $\boldsymbol{A} \in \mathbb{R}^{p \times p}$, $\boldsymbol{A}^+ = \mathrm{diag}(\boldsymbol{A})$ is a diagonal matrix with the same diagonal as $\boldsymbol{A}$, and $\boldsymbol{A}^- = \boldsymbol{A} - \boldsymbol{A}^+$ is $\boldsymbol{A}$ with all its diagonal elements set to zero. The notation $\boldsymbol{y}_n = \mathcal{O}_P(\boldsymbol{x}_n)$ for random vectors $\boldsymbol{y}_n, \boldsymbol{x}_n \in \mathbb{R}^p$ means that for any $\varepsilon > 0$, there exists $0 < M < \infty$ such that $P(\|\boldsymbol{y}_n\| \leq M \|\boldsymbol{x}_n\|) \geq 1 - \varepsilon\ \forall n \geq 1$.

II. System Model

We will call $\mathcal{G}$ considered earlier a *single-attribute graphical model* for $\boldsymbol{x}$. Now consider p jointly Gaussian random vectors $\boldsymbol{z}_i \in \mathbb{R}^m$, $i = 1, 2, \cdots, p$. We associate $\boldsymbol{z}_i$ with the ith node of an undirected graph $\mathcal{G} = (V, \mathcal{E})$ where $V = [p]$ and edges in $\mathcal{E}$ describe the conditional dependencies among vectors

$\{\boldsymbol{z}_i,\ i \in V\}$. As in the scalar case ($m = 1$), there is no edge between node i and node j in $\mathcal{G}$ iff random vectors $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ are conditionally independent given all the remaining random vectors [4]. This is the *multi-attribute Gaussian graphical model* of interest in this paper.

Define the mp-vector

$$\boldsymbol{x} = [\boldsymbol{z}_1^\top \ \boldsymbol{z}_2^\top \ \cdots \ \boldsymbol{z}_p^\top]^\top \in \mathbb{R}^{mp}\,. \tag{1}$$

Suppose we have n i.i.d. observations $\boldsymbol{x}(t), t = 0, 1, \cdots, n-1$, of zero-mean $\boldsymbol{x}$. Our objective is to estimate the inverse covariance matrix $(\mathbb{E}\{\mathbf{x}\mathbf{x}^\top\})^{-1}$ and to determine if edge $\{i, j\} \in \mathcal{E}$, given data $\{\boldsymbol{x}(t)\}_{t=0}^{n-1}$. Let us associate $\boldsymbol{x}$ with an "enlarged" graph $\bar{\mathcal{G}} = (\bar{V}, \bar{\mathcal{E}})$, where $\bar{V} = [mp]$ and $\bar{\mathcal{E}} \subseteq \bar{V} \times \bar{V}$. Now $[\boldsymbol{z}_j]_\ell$, the ℓth component of $\boldsymbol{z}_j$ associated with node j of $\mathcal{G} = (V, \mathcal{E})$, is the random variable $x_q = [\boldsymbol{x}]_q$, where $q = (j-1)m + \ell$, $j = 1, 2, \cdots, p$ and $\ell = 1, 2, \cdots, m$. The random variable x_q is associated with node q of $\bar{\mathcal{G}} = (\bar{V}, \bar{\mathcal{E}})$. Corresponding to the edge $\{j, k\} \in \mathcal{E}$ in the multi-attribute $\mathcal{G} = (V, \mathcal{E})$, there are m^2 edges $\{q, r\} \in \bar{\mathcal{E}}$ specified by $q = (j-1)m + s$ and $r = (k-1)m + t$, where $s = 1, 2, \cdots, m$ and $t = 1, 2, \cdots, m$. The graph $\bar{\mathcal{G}} = (\bar{V}, \bar{\mathcal{E}})$ is a single-attribute graph. In order for $\bar{\mathcal{G}}$ to reflect the conditional independencies encoded in $\mathcal{G}$, we must have the equivalence $\{j, k\} \notin \mathcal{E} \Leftrightarrow \bar{\mathcal{E}}^{(jk)} \cap \bar{\mathcal{E}} = \emptyset$, where $\bar{\mathcal{E}}^{(jk)} = \big\{\{q, r\} : q = (j-1)m + s,\, r = (k-1)m + t,\, s, t = 1, 2, \cdots, m\big\}$. Let $\boldsymbol{R}_{xx} = \mathbb{E}\{\boldsymbol{x}\boldsymbol{x}^\top\} \succ \mathbf{0}$ and $\boldsymbol{\Omega} = \boldsymbol{R}_{xx}^{-1}$. Define the (j, k)th $m \times m$ subblock $\boldsymbol{\Omega}^{(jk)}$ of $\boldsymbol{\Omega}$ as

$$[\boldsymbol{\Omega}^{(jk)}]_{st} = [\boldsymbol{\Omega}]_{(j-1)m+s,(k-1)m+t}\,,\ s, t = 1, 2, \cdots, m\,. \tag{2}$$

It is established in [4, Sec. 2.1] that $\boldsymbol{\Omega}^{(jk)} = \mathbf{0} \Leftrightarrow \{j, k\} \notin \mathcal{E}$. Since $\boldsymbol{\Omega}^{(jk)} = \mathbf{0}$ is equivalent to $[\boldsymbol{\Omega}]_{qr} = 0$ for every $\{q, r\} \in \bar{\mathcal{E}}^{(jk)}$, and since, by [2, Proposition 5.2], $[\boldsymbol{\Omega}]_{qr} = 0$ iff x_q and x_r are conditionally independent, hence, iff $\{q, r\} \notin \bar{\mathcal{E}}$, it follows that the aforementioned equivalence holds true.

III. Penalized Negative Log-Likelihood

Consider a finite set of data comprised of n i.i.d. zero-mean observations $\boldsymbol{x}(t), t = 0, 1, 2, \cdots, n-1$. Parameterizing in terms of the precision (inverse covariance) matrix $\boldsymbol{\Omega}$, the negative log-likelihood, up to some irrelevant constants, is given by

$$\mathcal{L}(\boldsymbol{\Omega}) := \ln(|\boldsymbol{\Omega}|) - \operatorname{tr}\left(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega}\right) \tag{3}$$

$$\text{where } \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{t=0}^{n-1} \boldsymbol{x}(t)\boldsymbol{x}^\top(t)\,. \tag{4}$$

In the high-dimensional case ($n < p$ or n comparable to p), to enforce sparsity and to make the problem well-conditioned, we propose to minimize a penalized version $\bar{\mathcal{L}}(\boldsymbol{\Omega})$ of $\mathcal{L}(\boldsymbol{\Omega})$ where we penalize (regularize) both element-wise and group-wise. We have

$$\bar{\mathcal{L}}(\boldsymbol{\Omega}) = \mathcal{L}(\boldsymbol{\Omega}) + \alpha P_e(\boldsymbol{\Omega}) + (1-\alpha) P_g(\boldsymbol{\Omega}), \tag{5}$$

$$P_e(\boldsymbol{\Omega}) = \sum_{i \neq j}^{mp} \rho_\lambda\left(\left|[\boldsymbol{\Omega}]_{ij}\right|\right), \tag{6}$$

$$P_g(\boldsymbol{\Omega}) = m \sum_{q \neq \ell}^{p} \rho_\lambda\left(\|\boldsymbol{\Omega}^{(q\ell)}\|_F\right) \tag{7}$$

where $\boldsymbol{\Omega}^{(q\ell)} \in \mathbb{R}^{m \times m}$ is defined as in (2), $\lambda > 0$, $\alpha \in [0, 1]$, m in (7) reflects the number of group variables [8], and for $u \in \mathbb{R}$, $\rho_\lambda(u)$ is a penalty function that is function of $|u|$. In (6), the penalty term is applied to each off-diagonal element of $\boldsymbol{\Omega}$ and in (7), the penalty term is applied to the off-block-diagonal group of m^2 terms via $\boldsymbol{\Omega}^{(q\ell)}$. The parameter $\alpha \in [0, 1]$ "balances" element-wise and group-wise penalties [5]–[7].

The following penalty functions are considered:

- *Lasso*. For some $\lambda > 0$, $\rho_\lambda(u) = \lambda|u|, \quad u \in \mathbb{R}$.
- *Log-sum*. For some $\lambda > 0$ and $1 \gg \epsilon > 0$, $\rho_\lambda(u) = \lambda\epsilon \ln\left(1 + \frac{|u|}{\epsilon}\right)$.
- *Smoothly Clipped Absolute Deviation (SCAD)*. For some $\lambda > 0$ and $a > 2$, $\rho_\lambda(u) = \lambda|u|$ for $|u| \leq \lambda$, $= (2a\lambda|u| - |u|^2 - \lambda^2)/(2(a-1))$ for $\lambda < |u| < a\lambda$ and $= \lambda^2(a+1)/2$ for $|u| \geq a\lambda$.

In the terminology of [16], all of the above three penalties are "μ-amenable" for some $\mu \geq 0$. As defined in [16, Sec. 2.2], $\rho_\lambda(u)$ is μ-amenable for some $\mu \geq 0$ if

(i) The function $\rho_\lambda(u)$ is symmetric around zero, i.e., $\rho_\lambda(u) = \rho_\lambda(-u)$ and $\rho_\lambda(0) = 0$.
(ii) The function $\rho_\lambda(u)$ is nondecreasing on $\mathbb{R}_+$.
(iii) The function $\rho_\lambda(u)/u$ is nonincreasing on $\mathbb{R}_+$.
(iv) The function $\rho_\lambda(u)$ is differentiable for $u \neq 0$.
(v) The function $\rho_\lambda(u) + \frac{\mu}{2}u^2$ is convex, for some $\mu \geq 0$.
(vi) $\lim_{u \to 0^+} \frac{d\rho_\lambda(u)}{du} = \lambda$.

It is shown in [16, Appendix A.1], that all of the above three penalties are μ-amenable with $\mu = 0$ for Lasso and $\mu = 1/(a-1)$ for SCAD. In [16] the log-sum penalty is defined as $\rho_\lambda(u) = \ln(1 + \lambda|u|)$ whereas in [11], it is defined as $\rho_\lambda(u) = \lambda \ln\left(1 + \frac{|u|}{\epsilon}\right)$. We follow [11] but modify it so that property (vi) in the definition of μ-amenable penalties holds. In our case $\mu = \frac{\lambda}{\epsilon}$ for the log-sum penalty since $\frac{d^2\rho_\lambda(u)}{du^2} = -\lambda\epsilon/(\epsilon + |u|)^2$ for $u \neq 0$.

The following properties also hold for the three penalty functions:

(vii) For some $C_\lambda > 0$ and $\delta_\lambda > 0$, we have

$$\rho_\lambda(u) \geq C_\lambda|u| \text{ for } |u| \leq \delta_\lambda\,. \tag{8}$$

(viii) $\frac{d\rho_\lambda(u)}{d|u|} \leq \lambda$ for $u \neq 0$.

Property (viii) is straightforward to verify. For Lasso, $C_\lambda = \lambda$ and $\delta_\lambda = \infty$. For SCAD, $C_\lambda = \lambda$ and $\delta_\lambda = \lambda$. Since $\ln(1+x) \geq x/(1+x)$ for $x > -1$, we have $\ln(1+x) \geq x/C_1$ for $0 \leq x \leq C_1 - 1$, $C_1 > 1$. Take $C_1 = 2$. Then log-sum $\rho_\lambda(u) \geq \frac{\lambda}{2}|u|$ for any $|u| \leq \epsilon$, leading to $C_\lambda = \frac{\lambda}{2}$ and $\delta_\lambda = \epsilon$. We may and will take $C_\lambda = \frac{\lambda}{2}$ for lasso and SCAD penalties as well.

IV. Theoretical Analysis

We now allow p and λ to be functions of sample size n, denoted as p_n and λ_n, respectively. Recall that we have the original multi-attribute graph $\mathcal{G} = (V, \mathcal{E})$ with $|V| = p_n$ and the corresponding enlarged graph $\bar{\mathcal{G}} = (\bar{V}, \bar{\mathcal{E}})$ with $|\bar{V}| = mp_n$. We assume the following regarding $\mathcal{G}$.

(A1) Denote the true edge set of the graph by $\mathcal{E}_0$, implying that $\mathcal{E}_0 = \{\{j,k\} \ : \ \|\boldsymbol{\Omega}_0^{(jk)}\|_F > 0, \ j \neq k\}$ where $\boldsymbol{\Omega}_0$ denotes the true precision matrix of $\boldsymbol{x}(t)$. Assume that $\text{card}(\mathcal{E}_0) = |(\mathcal{E}_0)| \leq s_{n0}$.

(A2) The minimum and maximum eigenvalues of $(mp_n) \times (mp_n)$ true covariance $\boldsymbol{\Sigma_0} \succ \boldsymbol{0}$ satisfy

$$0 < \beta_{\min} \leq \phi_{\min}(\boldsymbol{\Sigma_0}) \leq \phi_{\max}(\boldsymbol{\Sigma_0}) \leq \beta_{\max} < \infty .$$

Here $\beta_{\min}$ and $\beta_{\max}$ are not functions of n (or p_n).

Let $\hat{\boldsymbol{\Omega}}_\lambda = \arg\min_{\boldsymbol{\Omega}\succ\boldsymbol{0}} \bar{\mathcal{L}}(\boldsymbol{\Omega})$. Theorem 1 establishes local consistency of $\hat{\boldsymbol{\Omega}}_\lambda$ (a sketch of the proof is in the Appendix).

Theorem 1 (Local Consistency). For $\tau > 2$, let

$$C_0 = 40 \max_k([\boldsymbol{\Sigma}_0]_{kk}) \sqrt{N_1/\ln(mp_n)}\,, \tag{9}$$

$$R = 8(1+m)C_0/\beta_{\min}^2\,, \tag{10}$$

$$r_n = \sqrt{(mp_n + m^2 s_{n0})\ln(mp_n)/n} = o(1)\,, \tag{11}$$

$$N_1 = 2\ln(4(mp_n)^\tau)\,, \tag{12}$$

$$N_2 = \arg\min\left\{n \ : \ r_n \leq 0.1/(R\beta_{\min})\right\}, \tag{13}$$

$$N_3 = \arg\min\left\{n \ : \ r_n \leq \frac{\epsilon}{R}\right\}, \tag{14}$$

$$N_4 = \arg\min\left\{n \ : \ \lambda_n \leq \frac{\min_{(i,j):\,[\boldsymbol{\Omega}_0]_{ij}\neq 0} |[\boldsymbol{\Omega}_0]_{ij}|}{a+1}\right\}, \tag{15}$$

$$\lambda_{n\ell} = 2C_0\sqrt{\ln(mp_n)/n}\,, \tag{16}$$

$$\lambda_{nu1} = C_0(m+1)r_n/(m\sqrt{s_{n0}})\,, \tag{17}$$

$$\lambda_{nu2} = \min\left(Rr_n, \lambda_{nu1}\right) . \tag{18}$$

Under assumptions (A1)-(A2), there exists a local minimizer $\hat{\boldsymbol{\Omega}}_\lambda$ of $\bar{\mathcal{L}}(\boldsymbol{\Omega})$ satisfying

$$\|\hat{\boldsymbol{\Omega}}_\lambda - \boldsymbol{\Omega}_0\|_F \leq Rr_n \tag{19}$$

with probability greater than $1 - 1/(mp_n)^{\tau-2}$ if

(i) for the lasso penalty $n > \max\{N_1, N_2\}$ and λ_n satisfies $\lambda_{n\ell} \leq \lambda_n \leq \lambda_{nu1}$,

(ii) for the SCAD penalty $n > \max\{N_1, N_2, N_4\}$ and λ_n satisfies $\lambda_n = \lambda_{nu2}$,

(iii) for the log-sum penalty $n > \max\{N_1, N_2, N_3\}$ and λ_n satisfies $\lambda_{n\ell} \leq \lambda_n \leq \lambda_{nu1}$.

For the lasso penalty, $\hat{\boldsymbol{\Omega}}_\lambda$ is a global minimizer whereas for the other two penalties, it is a local minimizer. •

We follow the proof technique of [16, Lemma 6] in establishing Lemma 1 (the proof is in the Appendix).

Lemma 1 (Local Convexity). The optimization problem

$$\hat{\boldsymbol{\Omega}}_\lambda = \arg\min_{\boldsymbol{\Omega}\in\mathcal{B}} \bar{\mathcal{L}}(\boldsymbol{\Omega})\,, \tag{20}$$

$$\mathcal{B} = \{\boldsymbol{\Omega} \ : \ \boldsymbol{\Omega} \succ \boldsymbol{0}, \ \|\boldsymbol{\Omega}\| \leq 0.99\,\bar{\mu}\}\,, \tag{21}$$

$$\bar{\mu} = \begin{cases} \infty & : \ \text{lasso} \\ \sqrt{(a-1)/m} & : \ \text{SCAD} \\ \sqrt{\epsilon/(m\lambda_n)} & : \ \text{log-sum,} \end{cases} \tag{22}$$

consists of a strictly convex objective function over a convex constraint set, for all three penalties, where λ_n is as in Theorem 1. •

Lemma 1 and Theorem 1 lead to Theorem 2.

Theorem 2. Assume the conditions of Theorem 1. Then $\hat{\boldsymbol{\Omega}}_\lambda$ as defined in Lemma 1 is unique, satisfying $\|\hat{\boldsymbol{\Omega}}_\lambda - \boldsymbol{\Omega}_0\|_F \leq Rr_n$ with probability greater than $1 - 1/(mp_n)^{\tau-2}$ if $Rr_n + 1/\beta_{\min} \leq 0.99\,\bar{\mu}$.

Sketch of Proof. If $1/\beta_{\min} \leq 0.99\bar{\mu}$, then $\boldsymbol{\Omega}_0 \in \mathcal{B}$ since $\|\boldsymbol{\Omega}_0\| \leq 1/\beta_{\min}$, and also $\hat{\boldsymbol{\Omega}} \in \mathcal{B}$ since $\|\hat{\boldsymbol{\Omega}}\| \leq Rr_n + 1/\beta_{\min}$. Thus, both $\hat{\boldsymbol{\Omega}}_\lambda$ and $\boldsymbol{\Omega}_0$ are feasible. ■

Remark 1. We see from Theorem 1 that as $n \to \infty$, $\lambda_n \to 0$ (since $r_n = o(1)$), therefore, we eventually have "global" convexity for log-sum penalty by (22) for any $\boldsymbol{\Omega}_0$. But such is not the case for SCAD where one may need a to become large in which case it would behave more like lasso. □

We now turn to graph recovery. Define

$$\hat{\mathcal{E}} = \left\{\{q,\ell\} \ : \ \|\hat{\boldsymbol{\Omega}}^{(q\ell)}\|_F > \gamma_n > 0, q \neq \ell\right\}, \tag{23}$$

$$\mathcal{E}_0 = \left\{\{q,\ell\} \ : \ \|\boldsymbol{\Omega}_0^{(q\ell)}\|_F > 0, q \neq \ell\right\}, \tag{24}$$

$$\bar{\sigma}_n = Rr_n\,, \tag{25}$$

$$\nu = \min_{\{q,\ell\}\in\mathcal{E}_0} \|\boldsymbol{\Omega}_0^{(q\ell)}\|_F\,, \tag{26}$$

$$N_4 = \arg\min\left\{n \ : \ \bar{\sigma}_n \leq 0.4\nu\right\}, \tag{27}$$

where R and r_n are as in (10) and (11), respectively.

Theorem 3. For $\gamma_n = 0.5\nu$ and $n \geq N_4$, $\hat{\mathcal{E}} = \mathcal{E}_0$ with probability$> 1 - 1/(mp_n)^{\tau-2}$ under the conditions of Theorem 1. •

V. Conclusions

A unified theoretical analysis of multi-attribute graph learning using a penalized log-likelihood objective function was presented where both convex (sparse-group lasso) and non-convex (log-sum and SCAD group penalties) regularization functions were considered. Sufficient conditions in a high-dimensional setting for consistency (convergence of the precision matrix to true value in the Frobenius norm), local convexity when using non-convex penalties, and graph recovery were analyzed. We did not impose any incoherence or irrepresentability condition for our convergence results.

References

[1] J. Whittaker, *Graphical Models in Applied Multivariate Statistics*. New York: Wiley, 1990.

[2] S.L. Lauritzen, *Graphical models*. Oxford, UK: Oxford Univ. Press, 1996.
[3] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional data*. Berlin: Springer, 2011.
[4] M. Kolar, H. Liu and E.P. Xing, "Graph estimation from multi-attribute data," *J. Machine Learning Research*, vol. 15, pp. 1713-1750, 2014.
[5] J.K. Tugnait, "Sparse-group lasso for graph learning from multi-attribute data," *IEEE Trans. Signal Process.*, vol. 69, pp. 1771-1786, 2021. (Corrections, vol. 69, p. 4758, 2021.)
[6] J. Friedman, T. Hastie and R. Tibshirani, "A note on the group lasso and a sparse group lasso," *arXiv:1001.0736v1 [math.ST]*, 5 Jan 2010.
[7] N. Simon, J. Friedman, T. Hastie and R. Tibshirani, "A sparse-group lasso," *J. Computational Graphical Statistics*, vol. 22, pp. 231-245, 2013.
[8] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal Statistical Society: Statistical Methodology, Series B*, vol. 68, no. 1, pp. 49-67, 2006.
[9] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. American Statistical Assoc.*, vol. 96, pp. 1348-1360, Dec. 2001.
[10] C. Lam and J. Fan, "Sparsistency and rates of convergence in large covariance matrix estimation," *Ann. Statist.*, vol. 37, no. 6B, pp. 4254-4278, 2009.
[11] E.J. Candès, M.B. Wakin and S.P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, pp. 877-905, 2008.
[12] P. Ravikumar, M.J. Wainwright, G. Raskutti and B. Yu, "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Electronic J. Statistics*, vol. 5, pp. 935-980, 2011.
[13] A.J. Rothman, P.J. Bickel, E. Levina and J. Zhu, "Sparse permutation invariant covariance estimation," *Electronic J. Statistics*, vol. 2, pp. 494-515, 2008.
[14] J.K. Tugnait, "Sparse-group non-convex penalized multi-attribute graphical model selection," in *Proc. 29th European Signal Processing Conference (EUSIPCO 2021)*, pp. 1850-1854, Dublin, Ireland, Aug. 23-27, 2021.
[15] S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1-122, 2010.
[16] P.-L. Loh and M.J. Wainwright, "Support recovery without incoherence: A case for nonconvex regularization," *Annals of Statistics*, vol. 45, pp. 2455-2482, 2017.
[17] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge, UK: Cambridge Univ. Press, 2004.

Appendix

Sketch of Proof of Theorem 1: Let $\mathbf{\Omega} = \mathbf{\Omega}_0 + \mathbf{\Delta}$ with both $\mathbf{\Omega}$, $\mathbf{\Omega}_0 \succ \mathbf{0}$, and

$$Q(\mathbf{\Omega}) := \bar{\mathcal{L}}(\mathbf{\Omega}) - \bar{\mathcal{L}}(\mathbf{\Omega}_0)\,. \tag{28}$$

The estimate $\hat{\mathbf{\Omega}}_\lambda$, denoted by $\hat{\mathbf{\Omega}}$ hereafter suppressing dependence upon λ, minimizes $Q(\mathbf{\Omega})$, or equivalently, $\hat{\mathbf{\Delta}} = \hat{\mathbf{\Omega}} - \mathbf{\Omega}_0$ minimizes $G(\mathbf{\Delta}) := Q(\mathbf{\Omega}_0 + \mathbf{\Delta})$. We will follow the method of proof of [5, Theorem 1], which, in turn, for the most part, follows the method of proof of [13, Theorem 1] pertaining to lasso penalty. Consider the set

$$\Theta_n(R) := \left\{ \mathbf{\Delta} \,:\, \mathbf{\Delta} = \mathbf{\Delta}^\top,\ \|\mathbf{\Delta}\|_F = R r_n \right\} \tag{29}$$

where R and r_n are as in (10) and (11), respectively. Since $G(\hat{\mathbf{\Delta}}) \le G(\mathbf{0}) = 0$, if we can show that $\inf_{\mathbf{\Delta}}\{G(\mathbf{\Delta}) \,:\, \mathbf{\Delta} \in \Theta_n(R)\} > 0$, then the minimizer $\hat{\mathbf{\Delta}}$ must be inside $\Theta_n(R)$, and hence $\|\hat{\mathbf{\Delta}}\|_F \le R r_n$. It is shown in [13, (9)] that

$$\ln(|\mathbf{\Omega_0} + \mathbf{\Delta}|) - \ln(|\mathbf{\Omega}_0|) = \mathrm{tr}(\mathbf{\Sigma_0 \Delta}) - A_1 \tag{30}$$

where, with $\boldsymbol{H}(\mathbf{\Omega}_0, \mathbf{\Delta}, v) = (\mathbf{\Omega_0} + v\mathbf{\Delta})^{-1} \otimes (\mathbf{\Omega_0} + v\mathbf{\Delta})^{-1}$ and v denoting a scalar,

$$A_1 = \mathrm{vec}(\mathbf{\Delta})^\top \left(\int_0^1 (1-v)\boldsymbol{H}(\mathbf{\Omega}_0, \mathbf{\Delta}, v)\, dv \right) \mathrm{vec}(\mathbf{\Delta})\,. \tag{31}$$

Noting that $\mathbf{\Omega}^{-1} = \mathbf{\Sigma}$, we can rewrite $G(\mathbf{\Delta})$ as

$$G(\mathbf{\Delta}) = A_1 + A_2 + A_3 + A_4\,, \tag{32}$$

$$\text{where } A_2 = \mathrm{tr}\left((\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0)\mathbf{\Delta} \right)\,, \tag{33}$$

$$A_3 = \alpha \sum_{i,j=1; i\neq j}^{mp_n} \left(\rho_\lambda(|\Omega_{0ij} + \Delta_{ij}|) - \rho_\lambda(|\Omega_{0ij}|) \right)\,, \tag{34}$$

$$\begin{aligned} A_4 =& (1-\alpha) m \sum_{q,\ell=1; q\neq \ell}^{p_n} \big(\rho_\lambda(\|\mathbf{\Omega}_0^{(q\ell)} + \mathbf{\Delta}^{(q\ell)}\|_F) \\ & - \rho_\lambda(\|\mathbf{\Omega}_0^{(q\ell)}\|_F) \big)\,. \end{aligned} \tag{35}$$

Following [13, p. 502], we have

$$A_1 \ge \frac{\|\mathbf{\Delta}\|_F^2}{2(\|\mathbf{\Omega}_0\| + \|\mathbf{\Delta}\|)^2} \ge \frac{\|\mathbf{\Delta}\|_F^2}{2\left(\beta_{\min}^{-1} + R r_n\right)^2} \tag{36}$$

where we have used the fact that $\|\mathbf{\Omega_0}\| = \|\mathbf{\Sigma}_0^{-1}\| = \phi_{\max}(\mathbf{\Sigma}_0^{-1}) = (\phi_{\min}(\mathbf{\Sigma}_0))^{-1} \le \beta_{\min}^{-1}$ and $\|\mathbf{\Delta}\| \le \|\mathbf{\Delta}\|_F = R r_n$. We now consider A_2 in (33). We have

$$A_2 = \underbrace{\sum_{i,j=1; i\neq j}^{mp_n} [\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0]_{ij} \Delta_{ji}}_{L_1} + \underbrace{\sum_{i=1}^{mp_n} [\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0]_{ii} \Delta_{ii}}_{L_2} \tag{37}$$

Recall that by [5, Lemma 2], the sample covariance $\hat{\mathbf{\Sigma}}$ satisfies the tail bound

$$P\left(\max_{k,\ell} \left| [\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0]_{kl} \right| > C_0 \sqrt{\frac{\ln(mp_n)}{n}} \right) \le \frac{1}{(mp_n)^{\tau-2}} \tag{38}$$

for $\tau > 2$, if the sample size $n > N_1$ (N_1 is defined in (12)). To bound L_1, using [5, Lemma 2], we can show that with probability $> 1 - 1/(mp_n)^{\tau-2}$,

$$|L_1| \le \|\mathbf{\Delta}^-\|_1\, C_0 \sqrt{\frac{\ln(mp_n)}{n}}\,, \tag{39}$$

and

$$|L_2| \le \|\mathbf{\Delta}^+\|_F C_0 r_n\,. \tag{40}$$

Therefore, with probability $> 1 - 1/(mp_n)^{\tau-2}$,

$$|A_2| \le \|\mathbf{\Delta}^-\|_1\, C_0 \sqrt{\frac{\ln(mp_n)}{n}} + \|\mathbf{\Delta}^+\|_F C_0 r_n\,. \tag{41}$$

We now derive a different bound on A_2. Define $\tilde{\mathbf{\Delta}} \in \mathbb{R}^{p_n \times p_n}$ with (i,j)-th element $\tilde{\Delta}_{ij} = \|\mathbf{\Delta}^{(ij)}\|_F$, where $\mathbf{\Delta}^{(ij)}$ is defined from $\mathbf{\Delta}$ similar to (2). Using the Cauchy-Schwarz inequality an alternative bound can be derived as

$$|A_2| \le m\|\tilde{\mathbf{\Delta}}^-\|_1\, C_0 \sqrt{\frac{\ln(mp_n)}{n}} + \sqrt{m}\,\|\tilde{\mathbf{\Delta}}^+\|_F C_0 r_n\,. \tag{42}$$

For Lasso and Log-Sum Penalties: We now bound A_3 in (34). Let $\bar{\mathcal{E}}_0$ denote the true enlarged edge-set corresponding to $\mathcal{E}_0$ when one interprets multi-attribute model as a single-attribute model. Let

$\bar{\mathcal{E}}_0^c$ denote its complement. Using the mean-value theorem, we have ($\rho'_\lambda(u) = \frac{d\rho_\lambda(u)}{du}$)

$$\begin{aligned}\rho_\lambda(|\Omega_{0ij} + \Delta_{ij}|) =& \rho_\lambda(|\Omega_{0ij}|) \\ &+ \rho'_\lambda(|\tilde{\Omega}_{ij}|)(|\Omega_{0ij} + \Delta_{ij}| - |\Omega_{0ij}|) \end{aligned} \tag{43}$$

where $|\tilde{\Omega}_{ij}| = |\Omega_{0ij}| + \gamma(|\Omega_{0ij} + \Delta_{ij}| - |\Omega_{0ij}|)$ for some $\gamma \in [0,1]$. Using (43) we can show that

$$\begin{aligned} A_3 \geq -\alpha \sum_{(i,j)\in\bar{\mathcal{E}}_0} \rho'_\lambda(|\tilde{\Omega}_{ij}|)|\Delta_{ij}| + \alpha \sum_{(i,j)\in\bar{\mathcal{E}}_0^c} C_\lambda |\Delta_{ij}| \\ \text{for } \ |\Delta_{ij}| \leq \delta_\lambda . \end{aligned} \tag{44}$$

Now use property (viii) of the penalty functions and $C_\lambda = \lambda/2$ to conclude that

$$A_3 \geq -\alpha\lambda_n \sum_{(i,j)\in\bar{\mathcal{E}}_0} |\Delta_{ij}| + \alpha(\lambda_n/2) \sum_{(i,j)\in\bar{\mathcal{E}}_0^c} |\Delta_{ij}| . \tag{45}$$

Next we bound A_4 in (35). Considering the true edge-set $\mathcal{E}_0$ for the multi-attribute graph, let $\mathcal{E}_0^c$ denote its complement. If the edge $\{i,j\} \in \mathcal{E}_0^c$, then $\mathbf{\Omega}_0^{(ij)} = \mathbf{0}$, therefore, $\|\mathbf{\Omega}_0^{(ij)} + \mathbf{\Delta}^{(ij)}\|_F - \|\mathbf{\Omega}_0^{(ij)}\|_F = \|\mathbf{\Delta}^{(ij)}\|_F$. For $\{i,j\} \in \mathcal{E}_0$, by the triangle inequality, $\|\mathbf{\Omega}_0^{(ij)} + \mathbf{\Delta}^{(ij)}\|_F - \|\mathbf{\Omega}_0^{(ij)}\|_F \geq -\|\mathbf{\Delta}^{(ij)}\|_F$. Thus, similar to bounding A_3, we have

$$\begin{aligned} A_4 \geq& -(1-\alpha)m\lambda_n \sum_{(i,j)\in\mathcal{E}_0} \|\mathbf{\Delta}^{(ij)}\|_F \\ &+ (1-\alpha)m(\lambda_n/2) \sum_{(i,j)\in\mathcal{E}_0^c} \|\mathbf{\Delta}^{(ij)}\|_F . \end{aligned} \tag{46}$$

Bounding $\alpha A_2 + A_3$ and $(1-\alpha)A_2 + A_4$ separately, we can show that

$$\begin{aligned} A_2 + A_3 + A_4 &\geq -\|\mathbf{\Delta}\|_F \left(\lambda_n m \sqrt{s_{n0}} + (1+m)C_0 r_n\right) \\ &\geq -2(1+m)C_0 r_n \|\mathbf{\Delta}\|_F \end{aligned} \tag{47}$$

where we used the fact that since $\lambda_n \leq \lambda_{nu1}$, $\lambda_n m \sqrt{s_{n0}} \leq C_0(1+m)r_n$. Using (32), the bound (36) on A_1, bound (47) on $A_2 + A_3 + A_4$, and $\|\mathbf{\Delta}\|_F = Rr_n$, we have with probability $> 1 - 1/(mp_n)^{\tau-2}$,

$$G(\mathbf{\Delta}) \geq \|\mathbf{\Delta}\|_F^2 \left[\frac{1}{2(\beta_{\min}^{-1} + Rr_n)^2} - \frac{2C_0(1+m)}{R}\right] . \tag{48}$$

For the given choice of N_2, $Rr_n \leq Rr_{N_2} \leq 0.1/\beta_{\min}$ for $n \geq N_2$. Also, $2C_0(1+m)/R = \beta_{\min}^2/4$ by (10). Then for $n \geq N_2$,

$$\frac{1}{2(\beta_{\min}^{-1} + Rr_n)^2} - \frac{2C_0(1+m)}{R} \geq \beta_{\min}^2 \left(\frac{1}{2.42} - \frac{1}{4}\right) > 0 ,$$

implying $G(\mathbf{\Delta}) > 0$. This proves (19). The choice of N_3 for log-sum penalty ensures that $|\Delta_{ij}| \leq \delta_\lambda = \epsilon$ needed in (44) is satisfied w.h.p.: if $Rr_n \leq \epsilon$, then $|\Delta_{ij}| \leq \|\mathbf{\Delta}\|_F \leq Rr_n \leq \epsilon$.

For SCAD Penalty: Here we address (43) differently. Using triangle inequality, we have

$$\begin{aligned} |\tilde{\Omega}_{ij}| &\geq |\Omega_{0ij}| + \gamma\big(|\Omega_{0ij}| - |\Delta_{ij}| - |\Omega_{0ij}|\big) \\ &\geq |\Omega_{0ij}| - |\Delta_{ij}| . \end{aligned} \tag{49}$$

Since $|\Delta_{ij}| \leq \|\mathbf{\Delta}\|_F \leq Rr_n$, the choice $\lambda_n = \lambda_{nu2}$ implies that $\lambda_n \geq Rr_n$, satisfying $|\Delta_{ij}| \leq \lambda_n$. Therefore, $|\tilde{\Omega}_{ij}| \geq |\Omega_{0ij}| - \lambda_n$. For $\bar{n} \geq N_4$, $\rho'_\lambda(|\tilde{\Omega}_{ij}|) = 0$ (see (15)) if $\{i,j\} \in \bar{\mathcal{E}}_0$, i.e, $|\Omega_{0ij}| \neq 0$, since in this case $|\tilde{\Omega}_{ij}| \geq (a+1)\lambda_n - \lambda_n = a\lambda_n$. Therefore, $A_3 = \alpha \sum_{(i,j)\in\bar{\mathcal{E}}_0^c} \rho_\lambda(|\Delta_{ij}|) \geq \alpha \sum_{(i,j)\in\bar{\mathcal{E}}_0^c} C_\lambda |\Delta_{ij}|$ for $|\Delta_{ij}| \leq \delta_\lambda$, leading to

$$A_3 \geq \alpha(\lambda_n/2) \sum_{(i,j)\in\bar{\mathcal{E}}_0^c} |\Delta_{ij}| . \tag{50}$$

Mimicking the steps for bounding A_3 above and under same conditions, we have

$$A_4 \geq (1-\alpha)m(\lambda_n/2) \sum_{(i,j)\in\mathcal{E}_0^c} \|\mathbf{\Delta}^{(ij)}\|_F . \tag{51}$$

Similar to the lasso and log-sum case, we then have

$$\begin{aligned} A_2 + A_3 + A_4 &\geq -\|\mathbf{\Delta}\|_F \left(C_0 d_1 m \sqrt{s_{n0}} + mC_0 r_n\right) \\ &\geq -(1+m)C_0 r_n \|\mathbf{\Delta}\|_F \end{aligned} \tag{52}$$

where we used the fact that $C_0 d_1 m \sqrt{s_{n0}} \leq C_0 r_n$. Mimicking (48), we have with probability $> 1 - 1/(mp_n)^{\tau-2}$, we have

$$\begin{aligned} G(\mathbf{\Delta}) &\geq \|\mathbf{\Delta}\|_F^2 \left[\frac{1}{2(\beta_{\min}^{-1} + Rr_n)^2} - \frac{(1+m)C_0}{R}\right] \\ &\geq \beta_{\min}^2 \left(\frac{1}{2.42} - \frac{1}{8}\right) > 0 , \end{aligned} \tag{53}$$

implying $G(\mathbf{\Delta}) > 0$. This proves (19). For the SCAD penalty, we need $|\Delta_{ij}| \leq \delta_\lambda = \lambda_n$ in (50). Since $|\Delta_{ij}| \leq \|\mathbf{\Delta}\|_F \leq Rr_n$, the choice $\lambda_n = \lambda_{nu2}$ implies that $\lambda_n \geq Rr_n$, satisfying $|\Delta_{ij}| \leq \lambda_n$. This completes the proof. ■

Proof of Lemma 1: Consider $h(\mathbf{\Omega}) = \mathcal{L}(\mathbf{\Omega}) - \frac{\mu}{2}\|\mathbf{\Omega}\|_F^2$ for some $\mu \geq 0$. The Hessian of $\mathcal{L}(\mathbf{\Omega})$ w.r.t. vec$(\mathbf{\Omega})$ is $\nabla^2\mathcal{L}(\mathbf{\Omega}) = \mathbf{\Omega}^{-1} \otimes \mathbf{\Omega}^{-1}$ with

$$\phi_{\min}(\nabla^2\mathcal{L}(\mathbf{\Omega})) = \phi_{\min}^2(\mathbf{\Omega}^{-1}) = 1/\phi_{\max}^2(\mathbf{\Omega}) = 1/\|\mathbf{\Omega}\|^2 . \tag{54}$$

Since $\nabla^2 h(\mathbf{\Omega}) = \mathbf{\Omega}^{-1} \otimes \mathbf{\Omega}^{-1} - \mu \boldsymbol{I}_{(mp)^2}$, it follows that $h(\mathbf{\Omega})$ is positive semi-definite, hence convex, if

$$\|\mathbf{\Omega}\| \leq 1/\sqrt{\mu} .$$

By property (v) of the penalty functions, $g(u) := \rho_\lambda(u) + \frac{\mu}{2}u^2$ is convex, for some $\mu \geq 0$, and by property (ii), it is non-decreasing on $\mathbb{R}_+$. Therefore, by the composition rules [17, Sec. 3.2.4], $g(|[\mathbf{\Omega}]_{ij}|)$ and $g(\|\mathbf{\Omega}^{(q\ell)}\|_F)$ are convex. Hence,

$$P_e(\mathbf{\Omega}) + \frac{\mu_e}{2}\|\mathbf{\Omega}\|_F^2 = \sum_{i\neq j}^{mp_n} \left(\rho_\lambda\big(|[\mathbf{\Omega}]_{ij}|\big) + \frac{\mu_e}{2}\big|[\mathbf{\Omega}]_{ij}\big|^2\right)$$

is convex for $\mu_e = \mu \geq 0$, and similarly,

$$P_g(\mathbf{\Omega}) + \frac{\mu_g}{2}\|\mathbf{\Omega}\|_F^2 = m \sum_{q\neq\ell}^{p_n} \left(\rho_\lambda(\|\mathbf{\Omega}^{(q\ell)}\|_F) + \frac{\mu_g}{2m}\|\mathbf{\Omega}^{(q\ell)}\|_F^2\right)$$

is convex for $\mu_g = m\,\mu$, where μ is the value that renders $\rho_\lambda(u) + \frac{\mu}{2}u^2$ convex. Now express $\bar{\mathcal{L}}(\mathbf{\Omega})$ as

$$\bar{\mathcal{L}}(\mathbf{\Omega}) = \alpha\bar{\mathcal{L}}_e(\mathbf{\Omega}) + (1-\alpha)\bar{\mathcal{L}}_g(\mathbf{\Omega}) , \tag{55}$$

$$\bar{\mathcal{L}}_e(\mathbf{\Omega}) = \mathcal{L}(\mathbf{\Omega}) - \frac{\mu}{2}\|\mathbf{\Omega}\|_F^2 + P_e(\mathbf{\Omega}) + \frac{\mu}{2}\|\mathbf{\Omega}\|_F^2 , \tag{56}$$

$$\bar{\mathcal{L}}_g(\mathbf{\Omega}) = \mathcal{L}(\mathbf{\Omega}) - \frac{\mu}{2}\|\mathbf{\Omega}\|_F^2 + P_g(\mathbf{\Omega}) + \frac{\mu}{2}\|\mathbf{\Omega}\|_F^2 . \tag{57}$$

Now $\bar{\mathcal{L}}_e(\mathbf{\Omega})$ is convex function of $\mathbf{\Omega}$ if $\|\mathbf{\Omega}\| \leq 1/\sqrt{\mu}$, and $\bar{\mathcal{L}}_g(\mathbf{\Omega})$ is convex in $\mathbf{\Omega}$ if $\|\mathbf{\Omega}\| \leq 1/\sqrt{\mu_g} = 1/\sqrt{m\mu} =: \bar{\mu}$. Thus, for $\bar{\mathcal{L}}(\mathbf{\Omega})$ to be strictly convex, using the (minimum) values of μ to make $\rho_\lambda(u) + \frac{\mu}{2}u^2$ convex, we require

$$\begin{aligned} \|\mathbf{\Omega}\| \leq& \ \bar{\mu} \\ =& \begin{cases} \infty & : \ \text{lasso} \\ \sqrt{(a-1)/m} & : \ \text{SCAD} \\ \sqrt{\epsilon/(m\lambda_n)} & : \ \text{log-sum,} \end{cases} \end{aligned} \tag{58}$$

The choice $\|\mathbf{\Omega}\| < \bar{\mu}$ makes $\mathcal{L}(\mathbf{\Omega}) - \frac{\mu}{2}\|\mathbf{\Omega}\|_F^2$ positive definite, hence strictly convex. We take $\|\mathbf{\Omega}\| = 0.99\,\bar{\mu}$, completing the proof. ■