ISSN 1342-2804

Research Reports on Mathematical and Computing Sciences

# A dual spectral projected gradient method for log-determinant semidefinite problems

Takashi Nakagaki, Mituhiro Fukuda,
Sunyoung Kim and Makoto Yamashita

December 2018, B–490

Department of
Mathematical and
Computing Sciences
Tokyo Institute of Technology

SERIES B: Operations Research

# A dual spectral projected gradient method for log-determinant semidefinite problems

Takashi Nakagaki* Mituhiro Fukuda† Sunyoung Kim‡ Makoto Yamashita§

December, 2018

**Abstract**

We extend the result on the spectral projected gradient method by Birgin *et al.* in 2000 to a log-determinant semidefinite problem (SDP) with linear constraints and propose a spectral projected gradient method for the dual problem. Our method is based on alternate projections on the intersection of two convex sets, which first projects onto the box constraints and then onto a set defined by a linear matrix inequality. By exploiting structures of the two projections, we show the same convergence properties can be obtained for the proposed method as Birgin's method where the exact orthogonal projection onto the intersection of two convex sets is performed. Using the convergence properties, we prove that the proposed algorithm attains the optimal value or terminates in a finite number of iterations. The efficiency of the proposed method is illustrated with the numerical results on randomly generated synthetic/deterministic data and gene expression data, in comparison with other methods including the inexact primal-dual path-following interior-point method, the adaptive spectral projected gradient method, and the adaptive Nesterov's smooth method. For the gene expression data, our results are compared with the quadratic approximation for sparse inverse covariance estimation method. We show that our method outperforms the other methods in obtaining a better optimal value fast.

**Key words.** Dual spectral projected gradient methods, log-determinant semidefinite programs with linear constraints, dual problem, theoretical convergence results, computational efficiency.

**AMS Classification.** 90C20, 90C22, 90C25, 90C26.

## 1 Introduction

We consider a convex semidefinite program with linear constraints of the form:

$$(\mathcal{P}) \qquad \begin{array}{rcl} \min & : & f(\boldsymbol{X}) := \mathrm{Tr}(\boldsymbol{C}\boldsymbol{X}) - \mu \log\det \boldsymbol{X} + \mathrm{Tr}(\boldsymbol{\rho}|\boldsymbol{X}|) \\ \text{s.t.} & : & \mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}, \boldsymbol{X} \succ \boldsymbol{O}, \end{array}$$

*Department of Mathematical and Computing Science, Tokyo Institute of Technology, 2-12-1-W8-41 Oh-Okayama, Meguro-ku, Tokyo 152-8552, Japan.

†Department of Mathematical and Computing Science, Tokyo Institute of Technology, 2-12-1-W8-41 Oh-Okayama, Meguro-ku, Tokyo 152-8552, Japan (mituhiro@is.titech.ac.jp). The research was partially supported by JSPS KAKENHI (Grant number: 26330024), and by the Research Institute for Mathematical Sciences, a Joint Usage/Research Center located in Kyoto University.

‡Department of Mathematics, Ewha W. University, 52 Ewhayeodae-gil, Sudaemoon-gu, Seoul 03760, Korea (skim@ewha.ac.kr). The research was supported by NRF 2017-R1A2B2005119.

§Department of Mathematical and Computing Science, Tokyo Institute of Technology, 2-12-1-W8-29 Oh-Okayama, Meguro-ku, Tokyo 152-8552, Japan (makoto.yamashita@is.titech.ac.jp). This research was partially supported by JSPS KAKENHI (Grant number: 18K11176).

where $\boldsymbol{C}$, $\boldsymbol{X}$ and $\boldsymbol{\rho}$ are $n \times n$ symmetric matrices $\mathbb{S}^n$, the elements of $\boldsymbol{\rho} \in \mathbb{S}^n$ are nonnegative, Tr denotes the trace of a matrix, $|\boldsymbol{X}| \in \mathbb{S}^n$ the matrix obtained by taking the absolute value of every element $X_{ij}$ $(1 \leq i, j \leq n)$ of $\boldsymbol{X}$, $\boldsymbol{X} \succ \boldsymbol{O}$ means that $\boldsymbol{X}$ is positive definite, and $\mathcal{A}$ a linear map of $\mathbb{S}^n \to \mathbb{R}^m$. In $(\mathcal{P})$, $\boldsymbol{C}, \boldsymbol{\rho} \in \mathbb{S}^n, \mu > 0, \boldsymbol{b} \in \mathbb{R}^m$, and the linear map $\mathcal{A}$ given by $\mathcal{A}(\boldsymbol{X}) = (\mathrm{Tr}(\boldsymbol{A}_1\boldsymbol{X}), \ldots, \mathrm{Tr}(\boldsymbol{A}_m\boldsymbol{X}))^T$, where $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m \in \mathbb{S}^n$, are input data.

Problem $(\mathcal{P})$ frequently appears in statistical models such as sparse covariance selection or Gaussian graphical models. In particular, the sparse covariance selection model [6] or its graphical interpretation known as Gaussian Graphical Model (GGM) [11] are special cases of $(\mathcal{P})$ for $\boldsymbol{\rho} = \boldsymbol{O}$ and linear constraints taking the form $X_{ij} = 0$ for $(i, j) \in \Omega \subseteq \{(i, j) \mid 1 \leq i < j \leq n\}$.

Many approximate solution methods for solving variants of $(\mathcal{P})$ have been proposed over the years. The methods mentioned below are mainly from recent computational developments. The adaptive spectral gradient (ASPG) method and the adaptive Nesterov's smooth (ANS) method proposed by Lu [14] are one of the earlier methods which can handle large-scale problems. Ueno and Tsuchiya [16] proposed a Newton method by localized approximation of the relevant data. Wang *et al.* [18] considered a primal proximal point algorithm which solves semismooth subproblems by the Newton-CG iterates. Employing the inexact primal-dual path-following interior-point method, Li and Toh in [12] demonstrated that the computational efficiency could be increased, despite the known inefficiency of interior-point methods for solving large-sized problems. Yuan [20] also proposed an improved Alternating Direction Method (ADM) to solve the sparse covariance problem by introducing an ADM-oriented reformulation. For a more general structured models/problems, Yang *et al.* [19] enhanced the method in [18] to handle block structured sparsity, employing an inexact generalized Newton method to solve the dual semismooth subproblem. They demonstrated that regularization using $\|\cdot\|_2$ or $\|\cdot\|_\infty$ norms instead of $\|\cdot\|_1$ in $(\mathcal{P})$ are more suitable for the structured models/problems. Wang [17] first generated an initial point using the proximal augmented Lagrangian method, then applied the Newton-CG augmented Lagrangian method to problems with an additional convex quadratic term in $(\mathcal{P})$. Li and Xiao [13] employed the symmetric Gauss-Seidel-type ADMM in the same framework of [18]. A more recent work by Zhang *et al.* [21] shows that $(\mathcal{P})$ with simple constraints as $X_{ij} = 0$ for $(i, j) \in \Omega$ can be converted into a more computationally tractable problem for large values of $\boldsymbol{\rho}$. Among the methods mentioned here, only the methods discussed in [18, 19, 17] can handle problems as general as $(\mathcal{P})$.

We propose a dual-type spectral projected gradient (SPG) method to obtain the optimal value of $(\mathcal{P})$. More precisely, an efficient algorithm is designed for the dual problem with $g : \mathbb{R}^m \times \mathbb{S}^n \to \mathbb{R}$:

$$(\mathcal{D}) \qquad \begin{array}{rcl} \max & : & g(\boldsymbol{y}, \boldsymbol{W}) := \boldsymbol{b}^T\boldsymbol{y} + \mu \log\det(\boldsymbol{C} + \boldsymbol{W} - \mathcal{A}^T(\boldsymbol{y})) + n\mu - n\mu\log\mu \\ \text{s.t.} & : & |\boldsymbol{W}| \leq \boldsymbol{\rho}, \boldsymbol{C} + \boldsymbol{W} - \mathcal{A}^T(\boldsymbol{y}) \succ \boldsymbol{O}, \end{array}$$

under the three assumptions: (i) $\mathcal{A}$ is surjective, that is, the set of $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m$ is linearly independent; (ii) The problem $(\mathcal{P})$ has an interior feasible point, *i.e.*, there exists $\boldsymbol{X} \succ \boldsymbol{O}$ such that $\mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}$; (iii) A feasible point for $(\mathcal{D})$ is given or can be easily computed. *i.e.*, there exists $\boldsymbol{y} \in \mathbb{R}^m$ and $\boldsymbol{W} \in \mathbb{S}^n$ such that $|\boldsymbol{W}| \leq \boldsymbol{\rho}$ and $\boldsymbol{C} + \boldsymbol{W} + \mathcal{A}^T(\boldsymbol{y}) \succ \boldsymbol{O}$. These assumptions are not strong as many applications satisfy these assumptions with slight modifications.

Our approach for solving $(\mathcal{D})$ by a projected gradient method is not the first one. A dual approach was examined in [7], however, their algorithm which employs the classical gradient projection method cannot handle linear constraints.

The spectral projection gradient (SPG) method by Birgin *et al.* [2], which is slightly modified in our method, minimizes a smooth objective function over a closed convex set. Each iteration of the SPG requires (a) projection(s) onto the feasible closed convex set and performs a non-monotone line search for the Barzilai-Borwein step size [1]. An important advantage of the SPG method is that it requires only the information of function values and first-order derivatives, therefore,

the computational cost of each iteration is much less than methods which employ second-order derivatives such as interior-point methods. The ASPG method [14] described above repeatedly applies the SPG method by decreasing $\boldsymbol{\rho}$ adaptively, but the ASPG method was designed for the only specific constraint $X_{ij} = 0$ for $(i, j) \in \Omega$. We extend these results to directly handle a more general linear constraint $\mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}$.

Our proposed algorithm called Dual SPG, which is a dual-type SPG, adapts the SPG methods of [2] to $(\mathcal{D})$. A crucial difference between our method and the original method is that the Dual SPG first performs an orthogonal projection onto the box constraints and subsequently onto the set defined by an LMI, while the original method computes the exact orthogonal projection of the search direction over the intersection of the two convex sets. The projection onto the intersection of the two sets requires some iterative methods, which frequently causes some numerical difficulties. Moreover, the projection by an iterative method is usually inexact, resulting in the search direction that may not be an ascent direction. We note that an ascent direction is necessary for the convergence analysis as shown in Lemma 3.2 in Section 3. On the other hand, the projections onto the box constraints and the LMI constraints can be exactly computed within numerical errors.

The convergence analysis for the Dual SPG (Algorithm 2.1) presented in Section 3 shows that such approximate orthogonal projections do not affect convergence, in fact, the convergence properties of the original SPG also hold for the Dual SPG. For instance, stopping criteria based on the fixed point of the projection (Lemma 3.8) and other properties described in the beginning of Section 3 can be proved for the Dual SPG. The properties are used to finally prove that the algorithm either terminates in a finite number of iterations or successfully attains the optimal value.

We should emphasize that the proof for the original SPG developed in [2] cannot be applied to the Dual SPG proposed here. As the Dual SPG utilizes the two different projections instead of the orthogonal projection onto the feasible region in the original SPG, a new proof is necessary, in particular, for Lemma 3.8 where the properties of the two projections are exploited. We also use the duality theorem to prove the convergence of a sub-sequence (Lemma 3.15) since the Dual SPG solves the dual problem. Lemma 3.15 cannot be obtained by simply applying the proof in [2].

The implementation of Algorithm 2.1, called DSPG in this paper, were run on three classes of problems: Randomly generated synthetic data (Section 4.1), deterministic synthetic data (Section 4.2), and gene expression data (Section 4.3; with no constraints) from the literature. Comparison of the DSPG against high-performance code such as ASPG [14], ANS [14], and IIPM [12] shows that our code can be superior or at least competitive with them in terms of computational time when high accuracy is required. In particular, against QUIC [9], the DSPG can be faster for denser instances.

This paper is organized as follows: We proposed our method DSPG in Section 2. Section 3 is mainly devoted to the convergence of the proposed method. Section 4 presents computational results of the proposed method in comparison with other methods. For the gene expression data, our results are compared with QUIC. We finally conclude in Section 5.

### 1.1 Notation

We use $||\boldsymbol{y}|| := \sqrt{\boldsymbol{y}^T\boldsymbol{y}}$ for $\boldsymbol{y} \in \mathbb{R}^m$ and $||\boldsymbol{W}|| := \sqrt{\boldsymbol{W} \bullet \boldsymbol{W}}$ for $\boldsymbol{W} \in \mathbb{S}^n$ where $\boldsymbol{W} \bullet \boldsymbol{V} = \mathrm{Tr}(\boldsymbol{W}\boldsymbol{V}) = \sum_{i=1}^n \sum_{j=1}^n W_{ij}V_{ij}$ for $\boldsymbol{V} \in \mathbb{S}^n$, as the norm of vectors and matrices, respectively. We extend the inner-product to the space of $\mathbb{R}^m \times \mathbb{S}^n$ by $(\boldsymbol{y}_1, \boldsymbol{W}_1) \bullet (\boldsymbol{y}_2, \boldsymbol{W}_2) := \boldsymbol{y}_1^T\boldsymbol{y}_2 + \boldsymbol{W}_1 \bullet \boldsymbol{W}_2$ for $(\boldsymbol{y}_1, \boldsymbol{W}_1), (\boldsymbol{y}_2, \boldsymbol{W}_2) \in \mathbb{R}^m \times \mathbb{S}^n$. The norm of linear maps is defined by $||\mathcal{A}|| := \max_{||\boldsymbol{X}||=1} ||\mathcal{A}(\boldsymbol{X})||$.

The superscript of $T$ indicates the transpose of vectors or matrices, or the adjoint of linear operators. For example, the adjoint of $\mathcal{A}$ is denoted by $\mathcal{A}^T : \mathbb{R}^m \to \mathbb{S}^n$. The notation $\boldsymbol{X} \succeq \boldsymbol{Y} (\boldsymbol{X} \succ \boldsymbol{Y})$ stands for $\boldsymbol{X} - \boldsymbol{Y}$ being a positive semidefinite matrix (a positive definite matrix, respectively).

We also use $\boldsymbol{X} \geq \boldsymbol{Y}$ to describe that $\boldsymbol{X} - \boldsymbol{Y}$ is a non-negative matrix, that is, $X_{ij} \geq Y_{ij}$ for all $i, j = 1, \ldots, n$.

The induced norm for $\mathbb{R}^m \times \mathbb{S}^n$ is given by $||(\boldsymbol{y}, \boldsymbol{W})|| := \sqrt{(\boldsymbol{y}, \boldsymbol{W}) \bullet (\boldsymbol{y}, \boldsymbol{W})}$. To evaluate the accuracy of the solution, we also use an element-wise infinity norm defined by

$$||(\boldsymbol{y}, \boldsymbol{W})||_\infty := \max\{\max_{i=1,\ldots,m} |y_i|, \max_{i,j=1,\ldots,n} |W_{ij}|\}.$$

For a matrix $\boldsymbol{W} \in \mathbb{S}^n$, $[\boldsymbol{W}]_{\leq \boldsymbol{\rho}}$ is the matrix whose $(i,j)$th element is $\min\{\max\{W_{ij}, -\rho_{ij}\}, \rho_{ij}\}$. The set of such matrices is denoted by $\mathcal{W} := \{[\boldsymbol{W}]_{\leq \boldsymbol{\rho}} : \boldsymbol{W} \in \mathbb{S}^n\}$. In addition, $\boldsymbol{P}_S$ denotes the projection onto a closed convex set $S$;

$$\boldsymbol{P}_S(\boldsymbol{x}) = \arg\min_{\boldsymbol{y} \in S} ||\boldsymbol{y} - \boldsymbol{x}||.$$

We denote an optimal solution of $(\mathcal{P})$ and $(\mathcal{D})$ by $\boldsymbol{X}^*$ and $(\boldsymbol{y}^*, \boldsymbol{W}^*)$, respectively. For simplicity, we use $\boldsymbol{X}(\boldsymbol{y}, \boldsymbol{W}) := \mu(\boldsymbol{C} + \boldsymbol{W} - \mathcal{A}^T(\boldsymbol{y}))^{-1}$. The gradient of $g$ is a map of $\mathbb{R}^m \times \mathbb{S}^n \to \mathbb{R}^m \times \mathbb{S}^n$ given by

$$\begin{aligned} \nabla g(\boldsymbol{y}, \boldsymbol{W}) &:= (\nabla_{\boldsymbol{y}} g(\boldsymbol{y}, \boldsymbol{W}), \nabla_{\boldsymbol{W}} g(\boldsymbol{y}, \boldsymbol{W})) \\ &= (\boldsymbol{b} - \mu\mathcal{A}((\boldsymbol{C} + \boldsymbol{W} - \mathcal{A}^T(\boldsymbol{y}))^{-1}), \mu(\boldsymbol{C} + \boldsymbol{W} - \mathcal{A}^T(\boldsymbol{y}))^{-1}) \\ &= (\boldsymbol{b} - \mathcal{A}(\boldsymbol{X}(\boldsymbol{y}, \boldsymbol{W})), \boldsymbol{X}(\boldsymbol{y}, \boldsymbol{W})) \end{aligned}$$

We use $\mathcal{F}$ and $\mathcal{F}^*$ to denote the feasible set and the set of optimal solutions of $(\mathcal{D})$, respectively;

$$\begin{aligned} \mathcal{F} &:= \{(\boldsymbol{y}, \boldsymbol{W}) \in \mathbb{R}^m \times \mathbb{S}^n : \boldsymbol{W} \in \mathcal{W}, \boldsymbol{C} + \boldsymbol{W} - \mathcal{A}^T(\boldsymbol{y}) \succ \boldsymbol{O}\} \\ \mathcal{F}^* &:= \{(\boldsymbol{y}^*, \boldsymbol{W}^*) \in \mathbb{R}^m \times \mathbb{S}^n : g(\boldsymbol{y}^*, \boldsymbol{W}^*) \geq g(\boldsymbol{y}, \boldsymbol{W}) \text{ for } (\boldsymbol{y}, \boldsymbol{W}) \in \mathcal{F}\}. \end{aligned}$$

Finally, $f^*$ and $g^*$ are used to denote the optimal values of $(\mathcal{P})$ and $(\mathcal{D})$, respectively.

## 2 Spectral Projected Gradient Method for the Dual Problem

To propose a numerically efficient method, we focus on the fact that the feasible region of $(\mathcal{D})$ is the intersection of two convex sets: $\mathcal{F} = \widehat{\mathcal{W}} \cap \widehat{\mathcal{F}}$ where

$$\begin{aligned} \widehat{\mathcal{W}} &:= \mathbb{R}^m \times \mathcal{W} \\ \widehat{\mathcal{F}} &:= \{(\boldsymbol{y}, \boldsymbol{W}) \in \mathbb{R}^m \times \mathbb{S}^n : \boldsymbol{C} + \boldsymbol{W} - \mathcal{A}^T(\boldsymbol{y}) \succ \boldsymbol{O}\}. \end{aligned}$$

Although the projection onto this intersection requires elaborated computation, the projection onto the first set can be simply obtained by

$$\boldsymbol{P}_{\widehat{\mathcal{W}}}(\boldsymbol{y}, \boldsymbol{W}) = (\boldsymbol{y}, [\boldsymbol{W}]_{\leq \boldsymbol{\rho}}). \tag{1}$$

Next, we consider the second set $\widehat{\mathcal{F}}$. If the $k$th iterate $(\boldsymbol{y}^k, \boldsymbol{W}^k)$ satisfies $\boldsymbol{C} + \boldsymbol{W}^k - \mathcal{A}^T(\boldsymbol{y}^k) \succ \boldsymbol{O}$ and the direction toward the next iterate $(\boldsymbol{y}^{k+1}, \boldsymbol{W}^{k+1})$ is given by $(\Delta\boldsymbol{y}^k, \Delta\boldsymbol{W}^k)$, then the step length $\lambda$ can be computed such that $(\boldsymbol{y}^{k+1}, \boldsymbol{W}^{k+1}) := (\boldsymbol{y}^k, \boldsymbol{W}^k) + \lambda(\Delta\boldsymbol{y}^k, \Delta\boldsymbol{W}^k)$ satisfies $\boldsymbol{C} + \boldsymbol{W}^{k+1} - \mathcal{A}^T(\boldsymbol{y}^{k+1}) \succ \boldsymbol{O}$ using a similar procedure to interior-point methods. (See 4 below.) By the assumption (iii), we can start from some initial point $(\boldsymbol{y}^0, \boldsymbol{W}^0) \in \mathcal{F} = \widehat{\mathcal{W}} \cap \widehat{\mathcal{F}}$ and it is easy to keep all the iterations inside the intersection $\mathcal{F}$.

We now propose Algorithm 2.1 for solving the dual problem $(\mathcal{D})$. The notation $\boldsymbol{X}^k := \boldsymbol{X}(\boldsymbol{y}^k, \boldsymbol{W}^k) = \mu(\boldsymbol{C} + \boldsymbol{W}^k + \mathcal{A}^T(\boldsymbol{y}^k))^{-1}$ is used.

**Algorithm 2.1.** *(Dual Spectral Projected Gradient Method)*

Step 0: Set parameters $\epsilon \geq 0,\ \gamma \in (0,1),\ \tau \in (0,1),\ 0 < \sigma_1 < \sigma_2 < 1,\ 0 < \alpha_{\min} < \alpha_{\max} < \infty$ and an integer parameter $M \geq 1$. Take the initial point $(\boldsymbol{y}^0, \boldsymbol{W}^0) \in \mathcal{F}$ and an initial projection length $\alpha^0 \in [\alpha_{\min}, \alpha_{\max}]$. Set an iteration number $k := 0$.

Step 1: Compute a search direction (a projected gradient direction) for the stopping criterion

$$\begin{aligned}(\Delta \boldsymbol{y}^k_{(1)}, \Delta \boldsymbol{W}^k_{(1)}) &:= \boldsymbol{P}_{\widehat{\mathcal{W}}}((\boldsymbol{y}^k, \boldsymbol{W}^k) + \nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k)) - (\boldsymbol{y}^k, \boldsymbol{W}^k) \\ &= (\boldsymbol{b} - \mathcal{A}(\boldsymbol{X}^k), [\boldsymbol{W}^k + \boldsymbol{X}^k]_{\leq \boldsymbol{\rho}} - \boldsymbol{W}^k). \end{aligned} \tag{2}$$

If $||(\Delta \boldsymbol{y}^k_{(1)}, \Delta \boldsymbol{W}^k_{(1)})||_\infty \leq \epsilon$, stop and output $(\boldsymbol{y}^k, \boldsymbol{W}^k)$ as the approximate solution.

Step 2: Compute a search direction (a projected gradient direction)

$$\begin{aligned}(\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k) &:= \boldsymbol{P}_{\widehat{\mathcal{W}}}((\boldsymbol{y}^k, \boldsymbol{W}^k) + \alpha^k \nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k)) - (\boldsymbol{y}^k, \boldsymbol{W}^k) \\ &= (\alpha^k(\boldsymbol{b} - \mathcal{A}(\boldsymbol{X}^k)), [\boldsymbol{W}^k + \alpha^k \boldsymbol{X}^k]_{\leq \boldsymbol{\rho}} - \boldsymbol{W}^k). \end{aligned} \tag{3}$$

Step 3: Apply the Cholesky factorization to obtain a lower triangular matrix $\boldsymbol{L}$ such that $\boldsymbol{C} + \boldsymbol{W}^k - \mathcal{A}^T(\boldsymbol{y}^k) = \boldsymbol{L}\boldsymbol{L}^T$. Let $\theta$ be the minimum eigenvalue of $\boldsymbol{L}^{-1}(\Delta \boldsymbol{W}^k - \mathcal{A}^T(\Delta \boldsymbol{y}^k))\boldsymbol{L}^{-T}$. Then, compute

$$\overline{\lambda}^k := \begin{cases} 1 & (\theta \geq 0) \\ \min\left\{1, -\frac{1}{\theta} \times \tau\right\} & (\theta < 0) \end{cases} \tag{4}$$

and set $\lambda^k_1 := \overline{\lambda}^k$. Set an internal iteration number $j := 1$.

Step 3a: Set $(\boldsymbol{y}_+, \boldsymbol{W}_+) := (\boldsymbol{y}^k, \boldsymbol{W}^k) + \lambda^k_j(\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k)$.

Step 3b: If

$$g(\boldsymbol{y}_+, \boldsymbol{W}_+) \geq \min_{0 \leq h \leq \min\{k, M-1\}} g(\boldsymbol{y}^{k-h}, \boldsymbol{W}^{k-h}) + \gamma \lambda^k_j \nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k) \bullet (\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k) \tag{5}$$

is satisfied, then go to Step 4. Otherwise, choose $\lambda^k_{j+1} \in [\sigma_1 \lambda^k_j, \sigma_2 \lambda^k_j]$, and set $j := j + 1$, and return to Step 3a.

Step 4: Set $\lambda^k := \lambda^k_j$, $(\boldsymbol{y}^{k+1}, \boldsymbol{W}^{k+1}) := (\boldsymbol{y}^k, \boldsymbol{W}^k) + \lambda^k(\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k)$, $(\boldsymbol{s}_1, \boldsymbol{S}_1) := (\boldsymbol{y}^{k+1}, \boldsymbol{W}^{k+1}) - (\boldsymbol{y}^k, \boldsymbol{W}^k)$ and $(\boldsymbol{s}_2, \boldsymbol{S}_2) := \nabla g(\boldsymbol{y}^{k+1}, \boldsymbol{W}^{k+1}) - \nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k)$. Let $b^k := (\boldsymbol{s}_1, \boldsymbol{S}_1) \bullet (\boldsymbol{s}_2, \boldsymbol{S}_2)$. If $b^k \geq 0$, set $\alpha^{k+1} := \alpha_{\max}$. Otherwise, let $a^k := (\boldsymbol{s}_1, \boldsymbol{S}_1) \bullet (\boldsymbol{s}_1, \boldsymbol{S}_1)$ and set $\alpha^{k+1} := \min\{\alpha_{\max}, \max\{\alpha_{\min}, -a^k/b^k\}\}$.

Step 5: Increase the iteration counter $k := k + 1$ and return to Step 1.

The projection length $\alpha^{k+1} \in [\alpha_{\min}, \alpha_{\max}]$ in Step 4 is based on the Barzilai-Borwein step [1]. As investigated in [8, 15], this step has several advantages. For example, a linear convergence can be proven for unconstrained optimization problems without employing line search techniques on the conditions that its initial point is close to a local minimum and the Hessian matrix of the objective function is positive definite.

## 3 Convergence Analysis

We prove in Theorem 3.16, one of our main contributions, that Algorithm 2.1 with $\epsilon = 0$ generates a point of $\mathcal{F}^*$ in a finite number of iterations or it generates a sequence $\{(\boldsymbol{y}^k, \boldsymbol{W}^k)\} \subset \mathcal{F}$ that attains $\lim_{k\to\infty} g(\boldsymbol{y}^k, \boldsymbol{W}^k) = g^*$.

For the proof of Theorem 3.16, we present lemmas: Lemma 3.2 shows that the sequences $\{(\boldsymbol{y}^k, \boldsymbol{W}^k)\}$ by Algorithm 2.1 remain in a level set of $g$ for each $k$. Lemma 3.3 discusses on the boundedness of the level set, Lemma 3.7 on the uniqueness of the optimal solution in $(\mathcal{P})$, Lemma 3.8 on the validity of the stopping criteria in Algorithm 2.1, Lemma 3.10 on the bounds for the search direction $(\Delta\boldsymbol{y}^k, \Delta\boldsymbol{W}^k)$. Lemmas 3.12 and 3.15, which use Lemma 3.11 in their proofs, show that Algorithm 2.1 does not terminate before computing an approximate solution. Lemma 3.12 provides a lower bound for the step length $\lambda^k$ of Algorithm 2.1. Lemmas 3.13 and 3.15, which uses Lemma 3.14, discuss the termination of Algorithm 2.1 with $\epsilon = 0$ in a finite number of iterations attaining the optimal value $g^*$ or Algorithm 2.1 attains $\liminf_{k\to\infty} g(\boldsymbol{y}^k, \boldsymbol{W}^k) = g^*$.

In the proof of Theorem 3.16, the properties of projection will be repeatedly used. The representative properties are summarized in Proposition 2.1 of [8]. We list some of the properties related to this paper in the following and their proofs can also be found in [8] and the references therein.

**Proposition 3.1.** ([8]) For a convex set $S \subset \mathbb{R}^n$ and a function $f : \mathbb{R}^n \to \mathbb{R}$,

(P1) $(\boldsymbol{x} - \boldsymbol{P}_S(\boldsymbol{x}))^T(\boldsymbol{y} - \boldsymbol{P}_S(\boldsymbol{x})) \le 0 \quad \text{for} \quad \forall \boldsymbol{x} \in \mathbb{R}^n,\ \forall \boldsymbol{y} \in S.$

(P2) $(\boldsymbol{P}_S(\boldsymbol{x}) - \boldsymbol{P}_S(\boldsymbol{y}))^T(\boldsymbol{x} - \boldsymbol{y}) \ge ||(\boldsymbol{P}_S(\boldsymbol{x}) - \boldsymbol{P}_S(\boldsymbol{y})||^2 \quad \text{for} \quad \forall \boldsymbol{x}, \forall \boldsymbol{y} \in \mathbb{R}^n.$

(P3) $||\boldsymbol{P}_S(\boldsymbol{x}) - \boldsymbol{P}_S(\boldsymbol{y})|| \le ||\boldsymbol{x} - \boldsymbol{y}|| \quad \text{for} \quad \forall \boldsymbol{x}, \forall \boldsymbol{y} \in \mathbb{R}^n.$

(P4) $||\boldsymbol{P}_S(\boldsymbol{x} - \alpha\nabla f(\boldsymbol{x})) - \boldsymbol{x}||$ is non-decreasing in $\alpha > 0$ for $\forall \boldsymbol{x} \in S$.

(P5) $||\boldsymbol{P}_S(\boldsymbol{x} - \alpha\nabla f(\boldsymbol{x})) - \boldsymbol{x}||/\alpha$ is non-increasing in $\alpha > 0$ for $\forall \boldsymbol{x} \in S$.

To establish Theorem 3.16, we begin with a lemma that all the iterate points remain in a subset of $\mathcal{F}$.

**Lemma 3.2.** *Let $\mathcal{L}$ be the level set of $g$ determined by the initial value $g(\boldsymbol{y}^0, \boldsymbol{W}^0)$,*

$$\mathcal{L} := \{(\boldsymbol{y}, \boldsymbol{W}) \in \mathbb{R}^m \times \mathbb{S}^n : (\boldsymbol{y}, \boldsymbol{W}) \in \mathcal{F},\ g(\boldsymbol{y}, \boldsymbol{W}) \ge g(\boldsymbol{y}^0, \boldsymbol{W}^0)\}.$$

*Then, the sequence $\{(\boldsymbol{y}^k, \boldsymbol{W}^k)\}$ generated by Algorithm 2.1 satisfies $(\boldsymbol{y}^k, \boldsymbol{W}^k) \in \mathcal{L}$ for each $k$.*

*Proof.* First, we prove that $(\boldsymbol{y}^k, \boldsymbol{W}^k) \in \mathcal{F}$ for each $k$. By the assumption (iii), we have $(\boldsymbol{y}^0, \boldsymbol{W}^0) \in \mathcal{F}$. Assume that $(\boldsymbol{y}^k, \boldsymbol{W}^k) \in \mathcal{F}$ for some $k$. Since $0 \le \lambda^k \le 1$ in Step 4 and $\boldsymbol{W}^k \in \mathcal{W}$, the convexity of $\mathcal{W}$ indicates $\boldsymbol{W}^{k+1} = \boldsymbol{W}^k + \lambda^k\Delta\boldsymbol{W}^k = (1 - \lambda^k)\boldsymbol{W}^k + \lambda^k[\boldsymbol{W}^k + \alpha^k\boldsymbol{X}^k]_{\le\boldsymbol{\rho}} \in \mathcal{W}$. In addition, the value $\theta$ of Step 3 ensures $\boldsymbol{C} + (\boldsymbol{W}^k + \lambda\Delta\boldsymbol{W}^k) - \mathcal{A}^T(\boldsymbol{y}^k + \lambda\Delta\boldsymbol{y}^k) \succ \boldsymbol{O}$ for $\lambda \in [0, \overline{\lambda}^k]$. Hence, $(\boldsymbol{y}^{k+1}, \boldsymbol{W}^{k+1}) \in \mathcal{F}$.

Now, we verify that $g(\boldsymbol{y}^k, \boldsymbol{W}^k) \ge g(\boldsymbol{y}^0, \boldsymbol{W}^0)$ for each $k$. The case $k = 0$ is clear. The case

$k \geq 1$ depends on the fact $(\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k)$ is an ascent direction of $g$ at $(\boldsymbol{y}^k, \boldsymbol{W}^k)$;

$$\begin{aligned} & \nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k) \bullet (\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k) \\ = \; & (\nabla_{\boldsymbol{y}} g(\boldsymbol{y}^k, \boldsymbol{W}^k), \nabla_{\boldsymbol{W}} g(\boldsymbol{y}^k, \boldsymbol{W}^k)) \bullet (\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k) \\ = \; & \alpha^k ||\boldsymbol{b} - \mathcal{A}(\boldsymbol{X}^k)||^2 + \boldsymbol{X}^k \bullet ([\boldsymbol{W}^k + \alpha^k \boldsymbol{X}^k]_{\leq \boldsymbol{\rho}} - \boldsymbol{W}^k) \\ \geq \; & \alpha^k ||\boldsymbol{b} - \mathcal{A}(\boldsymbol{X}^k)||^2 + \frac{1}{\alpha^k} ||[\boldsymbol{W}^k + \alpha^k \boldsymbol{X}^k]_{\leq \boldsymbol{\rho}} - \boldsymbol{W}^k||^2 \\ = \; & \frac{1}{\alpha^k} ||(\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k)||^2 \qquad (6) \\ \geq \; & 0. \end{aligned}$$

The first inequality comes from (P2) by putting $\mathcal{W}$ as $S$, $\boldsymbol{W}^k + \alpha^k \boldsymbol{X}^k$ as $\boldsymbol{x}$ and $\boldsymbol{W}^k$ as $\boldsymbol{y}$, and using the relations $\boldsymbol{P}_{\mathcal{W}}(\boldsymbol{W}^k + \alpha^k \boldsymbol{X}^k) = [\boldsymbol{W}^k + \alpha^k \boldsymbol{X}^k]_{\leq \boldsymbol{\rho}}$ and $\boldsymbol{P}_{\mathcal{W}}(\boldsymbol{W}^k) = \boldsymbol{W}^k$ by $(\boldsymbol{y}^k, \boldsymbol{W}^k) \in \mathcal{F}$.

When the inner iteration terminates, we have

$$\begin{aligned} g(\boldsymbol{y}^{k+1}, \boldsymbol{W}^{k+1}) \; & \geq \; \min_{0 \leq h \leq \min\{k, M-1\}} g(\boldsymbol{y}^{k-h}, \boldsymbol{W}^{k-h}) + \gamma \lambda^k \nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k) \bullet (\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k) \\ & \geq \; \min_{0 \leq h \leq \min\{k, M-1\}} g(\boldsymbol{y}^{k-h}, \boldsymbol{W}^{k-h}). \end{aligned}$$

Therefore, if $\min_{0 \leq h \leq k} g(\boldsymbol{y}^h, \boldsymbol{W}^h) \geq g(\boldsymbol{y}^0, \boldsymbol{W}^0)$, we obtain $g(\boldsymbol{y}^{k+1}, \boldsymbol{W}^{k+1}) \geq g(\boldsymbol{y}^0, \boldsymbol{W}^0)$. By induction, we conclude $(\boldsymbol{y}^k, \boldsymbol{W}^k) \in \mathcal{L}$ for each $k$. □

The key to establishing Theorem 3.16 is the boundedness of the level set $\mathcal{L}$.

**Lemma 3.3.** *The level set $\mathcal{L}$ is bounded.*

*Proof.* If $(\boldsymbol{y}, \boldsymbol{W}) \in \mathcal{L}$, then $\boldsymbol{W} \in \mathcal{W}$. Thus, the boundedness of $\boldsymbol{W}$ is clear from $|W_{ij}| \leq \rho_{ij}$. We then fix $\widehat{\boldsymbol{W}} \in \mathcal{W}$ and show the boundedness of

$$\mathcal{L}_{\widehat{\boldsymbol{W}}} := \{\boldsymbol{y} \in \mathbb{R}^m : g(\boldsymbol{y}, \widehat{\boldsymbol{W}}) \geq g(\boldsymbol{y}^0, \boldsymbol{W}^0), \quad \boldsymbol{C} + \widehat{\boldsymbol{W}} - \mathcal{A}^T(\boldsymbol{y}) \succ \boldsymbol{O}\}.$$

Let $\boldsymbol{Z} := \boldsymbol{C} + \widehat{\boldsymbol{W}} - \mathcal{A}^T(\boldsymbol{y})$ for $\boldsymbol{y} \in \mathcal{L}_{\widehat{\boldsymbol{W}}}$. Since $\mathcal{A}$ is surjective, the map $\mathcal{A}\mathcal{A}^T : \mathbb{R}^m \to \mathbb{R}^m$ is nonsingular and

$$||\boldsymbol{y}|| = ||(\mathcal{A}\mathcal{A}^T)^{-1}\mathcal{A}(\boldsymbol{C} + \widehat{\boldsymbol{W}} - \boldsymbol{Z})|| \leq ||(\mathcal{A}\mathcal{A}^T)^{-1}|| \cdot ||\mathcal{A}|| \cdot (||\boldsymbol{C}|| + ||\widehat{\boldsymbol{W}}|| + ||\boldsymbol{Z}||).$$

Hence, if we can prove the boundedness of $\boldsymbol{Z}$, the desired result follows.

Since we assume that $(\mathcal{P})$ has at least one interior point, there exists $\widehat{\boldsymbol{X}}$ such that $\mathcal{A}(\widehat{\boldsymbol{X}}) = \boldsymbol{b}$ and $\widehat{\boldsymbol{X}} \succ \boldsymbol{O}$. We denote the eigenvalues of $\boldsymbol{Z}$ by $0 < \lambda_1(\boldsymbol{Z}) \leq \lambda_2(\boldsymbol{Z}) \leq \cdots \leq \lambda_n(\boldsymbol{Z})$. For simplicity, we use $\lambda_{\min}(\boldsymbol{Z}) := \lambda_1(\boldsymbol{Z})$ and $\lambda_{\max}(\boldsymbol{Z}) := \lambda_n(\boldsymbol{Z})$. Letting $\bar{c}_0 := g(\boldsymbol{y}^0, \boldsymbol{W}^0) - n\mu + n\mu \log \mu$, we can derive equivalent inequalities from $g(\boldsymbol{y}, \widehat{\boldsymbol{W}}) \geq g(\boldsymbol{y}^0, \boldsymbol{W}^0)$;

$$\begin{aligned} & g(\boldsymbol{y}, \widehat{\boldsymbol{W}}) \geq g(\boldsymbol{y}^0, \boldsymbol{W}^0) \\ \Leftrightarrow \; & \boldsymbol{b}^T \boldsymbol{y} + \mu \log \det(\boldsymbol{C} + \widehat{\boldsymbol{W}} - \mathcal{A}^T(\boldsymbol{y})) \geq \bar{c}_0 \\ \Leftrightarrow \; & \mathcal{A}(\widehat{\boldsymbol{X}})^T \boldsymbol{y} + \mu \log \det \boldsymbol{Z} \geq \bar{c}_0 \\ \Leftrightarrow \; & \widehat{\boldsymbol{X}} \bullet \mathcal{A}^T(\boldsymbol{y}) + \mu \log \det \boldsymbol{Z} \geq \bar{c}_0 \\ \Leftrightarrow \; & \widehat{\boldsymbol{X}} \bullet (\boldsymbol{C} + \widehat{\boldsymbol{W}} - \boldsymbol{Z}) + \mu \log \det \boldsymbol{Z} \geq \bar{c}_0 \\ \Leftrightarrow \; & \widehat{\boldsymbol{X}} \bullet \boldsymbol{Z} - \mu \log \det \boldsymbol{Z} \leq -\bar{c}_0 + \widehat{\boldsymbol{X}} \bullet (\boldsymbol{C} + \widehat{\boldsymbol{W}}) \end{aligned}$$

Since $\widehat{\boldsymbol{X}} \bullet \widehat{\boldsymbol{W}} = \sum_{i=1}^n \sum_{j=1}^n \widehat{X}_{ij}\widehat{W}_{ij} \leq \sum_{i=1}^n \sum_{j=1}^n |\widehat{X}_{ij}|\rho_{ij} = |\widehat{\boldsymbol{X}}| \bullet \boldsymbol{\rho}$, it holds that $\widehat{\boldsymbol{X}} \bullet \boldsymbol{Z} - \mu \log\det \boldsymbol{Z} \leq c$, where $c := -\bar{c}_0 + \widehat{\boldsymbol{X}} \bullet \boldsymbol{C} + |\widehat{\boldsymbol{X}}| \bullet \boldsymbol{\rho}$. From $\min_t\{at - \log t : t > 0\} = 1 + \log a$ for any $a > 0$, it follows that

$$\begin{aligned}
\widehat{\boldsymbol{X}} \bullet \boldsymbol{Z} - \mu \log\det \boldsymbol{Z} &\geq \sum_{i=1}^n [\lambda_{\min}(\widehat{\boldsymbol{X}})\lambda_i(\boldsymbol{Z}) - \mu \log \lambda_i(\boldsymbol{Z})] \\
&\geq (n-1)\mu \left(1 + \log \frac{\lambda_{\min}(\widehat{\boldsymbol{X}})}{\mu}\right) + \lambda_{\min}(\widehat{\boldsymbol{X}})\lambda_{\max}(\boldsymbol{Z}) - \mu \log \lambda_{\max}(\boldsymbol{Z}).
\end{aligned}$$

Hence,

$$\lambda_{\min}(\widehat{\boldsymbol{X}})\lambda_{\max}(\boldsymbol{Z}) - \mu \log \lambda_{\max}(\boldsymbol{Z}) \leq c - (n-1)\mu \left(1 + \log \frac{\lambda_{\min}(\widehat{\boldsymbol{X}})}{\mu}\right).$$

Note that the right-hand side is determined by only $\widehat{\boldsymbol{X}}$ and is independent from $\boldsymbol{Z}$, and that $\lambda_{\min}(\widehat{\boldsymbol{X}}) > 0$ from $\widehat{\boldsymbol{X}} \succ \boldsymbol{O}$. Hence, there exists $\beta_{\boldsymbol{Z}}^{\max} < \infty$ such that $\lambda_{\max}(\boldsymbol{Z}) \leq \beta_{\boldsymbol{Z}}^{\max}$ for all $(\boldsymbol{y}, \widehat{\boldsymbol{W}}) \in \mathcal{L}$.

In addition, from $\widehat{\boldsymbol{X}} \bullet \boldsymbol{Z} - \mu \log\det \boldsymbol{Z} \leq c$ and $\widehat{\boldsymbol{X}} \bullet \boldsymbol{Z} \geq 0$, we have

$$\begin{aligned}
\log\det \boldsymbol{Z} &\geq -\frac{c}{\mu} \\
\log \lambda_{\min}(\boldsymbol{Z}) &\geq -\frac{c}{\mu} - \sum_{i=2}^n \log \lambda_i(\boldsymbol{Z}) \geq -\frac{c}{\mu} - (n-1)\log \beta_{\boldsymbol{Z}}^{\max} \\
\lambda_{\min}(\boldsymbol{Z}) &\geq \beta_{\boldsymbol{Z}}^{\min} := \exp\left(-\frac{c}{\mu} - (n-1)\log \beta_{\boldsymbol{Z}}^{\max}\right) > 0.
\end{aligned}$$

Therefore, the minimum and maximum eigenvalues of $\boldsymbol{Z}$ are bounded for $(\boldsymbol{y}, \widehat{\boldsymbol{W}}) \in \mathcal{L}$. This completes the proof. □

**Remark 3.4.** From Lemmas 3.2 and 3.3, $||\boldsymbol{y}^k||$ and $||\boldsymbol{W}^k||$ are bounded; $||\boldsymbol{y}^k|| \leq \eta_{\boldsymbol{y}} := ||(\mathcal{A}\mathcal{A}^T)^{-1}|| \cdot ||\mathcal{A}|| \cdot (||\boldsymbol{C}|| + ||\boldsymbol{\rho}|| + \sqrt{n}\beta_{\boldsymbol{Z}}^{\max})$ and $||\boldsymbol{W}^k|| \leq \eta_{\boldsymbol{W}} := ||\boldsymbol{\rho}||$.

**Remark 3.5.** Lemma 3.3 implies that the set $\{\boldsymbol{X}(\boldsymbol{y}, \boldsymbol{W}) : (\boldsymbol{y}, \boldsymbol{W}) \in \mathcal{L}\}$ is also bounded. If we denote $\beta_{\boldsymbol{X}}^{\min} := \frac{\mu}{\beta_{\boldsymbol{Z}}^{\max}} > 0$ and $\beta_{\boldsymbol{X}}^{\max} := \frac{\mu}{\beta_{\boldsymbol{Z}}^{\min}} < \infty$, then we have $\beta_{\boldsymbol{X}}^{\min}\boldsymbol{I} \preceq \boldsymbol{X}(\boldsymbol{y}, \boldsymbol{W}) \preceq \beta_{\boldsymbol{X}}^{\max}\boldsymbol{I}$ for $(\boldsymbol{y}, \boldsymbol{W}) \in \mathcal{L}$. In particular, since $(\boldsymbol{y}^k, \boldsymbol{W}^k) \in \mathcal{L}$ from Lemma 3.2, $\boldsymbol{X}^k = \boldsymbol{X}(\boldsymbol{y}^k, \boldsymbol{W}^k) = \mu(\boldsymbol{C} + \boldsymbol{W} - \mathcal{A}^T(\boldsymbol{y}^k))^{-1}$ is also bounded; $\beta_{\boldsymbol{X}}^{\min}\boldsymbol{I} \preceq \boldsymbol{X}^k \preceq \beta_{\boldsymbol{X}}^{\max}\boldsymbol{I}$. Furthermore, for $(\boldsymbol{y}, \boldsymbol{W}) \in \mathcal{L}$, we obtain the bounds $||\boldsymbol{X}(\boldsymbol{y}, \boldsymbol{W})|| \leq \eta_{\boldsymbol{X}}$ and $||\boldsymbol{X}^{-1}(\boldsymbol{y}, \boldsymbol{W})|| \leq \eta_{\boldsymbol{X}^{-1}}$, where $\eta_{\boldsymbol{X}} := \sqrt{n}\beta_{\boldsymbol{X}}^{\max} > 0$ and $\eta_{\boldsymbol{X}^{-1}} := \frac{\sqrt{n}}{\beta_{\boldsymbol{X}}^{\min}} > 0$. Hence, it holds that $||\boldsymbol{X}^k|| \leq \eta_{\boldsymbol{X}}$ and $||(\boldsymbol{X}^k)^{-1}|| \leq \eta_{\boldsymbol{X}^{-1}}$ for each $k$.

**Remark 3.6.** It follows from Remark 3.5 that $||\Delta\boldsymbol{y}^k||$ and $||\Delta\boldsymbol{W}^k||$ are also bounded by $\eta_{\Delta\boldsymbol{y}} := \alpha_{\max}(||\boldsymbol{b}|| + ||\mathcal{A}||\eta_{\boldsymbol{X}})$ and $\eta_{\Delta\boldsymbol{W}} := \alpha_{\max}\eta_{\boldsymbol{X}}$, respectively. These bounds are found by

$$\begin{aligned}
||\Delta\boldsymbol{y}^k|| &= ||\alpha^k(\boldsymbol{b} - \mathcal{A}(\boldsymbol{X}^k))|| \leq \alpha^k(||\boldsymbol{b}|| + ||\mathcal{A}|| \cdot ||\boldsymbol{X}^k||) \leq \alpha_{\max}(||\boldsymbol{b}|| + ||\mathcal{A}||\eta_{\boldsymbol{X}}) \\
||\Delta\boldsymbol{W}^k|| &= |||[\boldsymbol{W}^k + \alpha^k\boldsymbol{X}^k]_{\leq\boldsymbol{\rho}} - \boldsymbol{W}^k|| \leq ||\alpha^k\boldsymbol{X}^k|| \leq \alpha_{\max}\eta_{\boldsymbol{X}}.
\end{aligned}$$

For $||\Delta\boldsymbol{W}^k||$, we substitute $S = \mathcal{W}$, $\boldsymbol{x} = \boldsymbol{W}^k + \alpha^k\boldsymbol{X}^k$ and $\boldsymbol{y} = \boldsymbol{W}^k = \boldsymbol{P}_{\mathcal{W}}(\boldsymbol{W}^k)$ to (P3).

From Lemma 3.3, the set of the optimal solutions $\mathcal{F}^*$ is a subset of $\{(\boldsymbol{y}, \boldsymbol{W}) \in \mathbb{R}^m \times \mathbb{S}^n : |\boldsymbol{W}| \leq \boldsymbol{\rho},\ \beta_{\boldsymbol{Z}}^{\min}\boldsymbol{I} \preceq \boldsymbol{C} + \boldsymbol{W} - \mathcal{A}^T(\boldsymbol{y}) \preceq \beta_{\boldsymbol{Z}}^{\max}\boldsymbol{I}\}$ and it is a closed convex set and bounded. From the continuity of the objective function $g$, the dual problem $(\mathcal{D})$ has an optimal solution. Furthermore, since both $(\mathcal{P})$ and $(\mathcal{D})$ has an interior feasible point, the duality theorem holds [3, 4], that is, the primal problem $(\mathcal{P})$ also has an optimal solution and there is no duality gap between $(\mathcal{P})$ and $(\mathcal{D})$, $f^* = g^*$. In the following Lemma 3.7, we show the uniqueness of the optimal solution in $(\mathcal{P})$ and a property of the optimal solutions in $(\mathcal{D})$.

**Lemma 3.7.** *The optimal solution of $(\mathcal{P})$ is unique. In addition, if both $(\boldsymbol{y}_1^*, \boldsymbol{W}_1^*)$ and $(\boldsymbol{y}_2^*, \boldsymbol{W}_2^*)$ are optimal solutions of $(\mathcal{D})$, then $\boldsymbol{X}(\boldsymbol{y}_1^*, \boldsymbol{W}_1^*) = \boldsymbol{X}(\boldsymbol{y}_2^*, \boldsymbol{W}_2^*)$ and $\boldsymbol{b}^T\boldsymbol{y}_1^* = \boldsymbol{b}^T\boldsymbol{y}_2^*$.*

*Proof.* Since the function $-\log\det\boldsymbol{X}$ is strictly convex [4], we have

$$-\log\det\left(\frac{\boldsymbol{X}_1+\boldsymbol{X}_2}{2}\right) < -\frac{1}{2}\log\det\boldsymbol{X}_1 - \frac{1}{2}\log\det\boldsymbol{X}_2 \quad \text{for} \quad \forall\boldsymbol{X}_1 \succ \boldsymbol{O}, \forall\boldsymbol{X}_2 \succ \boldsymbol{O}(\boldsymbol{X}_1 \neq \boldsymbol{X}_2). \tag{7}$$

Suppose that we have two different optimal solutions $(\boldsymbol{y}_1^*, \boldsymbol{W}_1^*)$ and $(\boldsymbol{y}_2^*, \boldsymbol{W}_2^*)$ for $(\mathcal{D})$ such that $\boldsymbol{C} + \boldsymbol{W}_1^* - \mathcal{A}^T(\boldsymbol{y}_1^*) \neq \boldsymbol{C} + \boldsymbol{W}_2^* - \mathcal{A}^T(\boldsymbol{y}_2^*)$. Since $(\boldsymbol{y}_1^*, \boldsymbol{W}_1^*)$ and $(\boldsymbol{y}_2^*, \boldsymbol{W}_2^*)$ attain the same objective value, it holds that $g^* = \boldsymbol{b}^T\boldsymbol{y}_1^* + \mu\log\det(\boldsymbol{C}+\boldsymbol{W}_1^*-\mathcal{A}^T(\boldsymbol{y}_1^*)) + n\mu - n\mu\log\mu = \boldsymbol{b}^T\boldsymbol{y}_2^* + \mu\log\det(\boldsymbol{C}+\boldsymbol{W}_2^*-\mathcal{A}^T(\boldsymbol{y}_2^*)) + n\mu - n\mu\log\mu$. Since the feasible set of $(\mathcal{D})$ is convex, $\left(\frac{\boldsymbol{y}_1^*+\boldsymbol{y}_2^*}{2}, \frac{\boldsymbol{W}_1^*+\boldsymbol{W}_2^*}{2}\right)$ is also feasible. However, the inequality (7) indicates

$$\begin{aligned}
& \boldsymbol{b}^T\left(\frac{\boldsymbol{y}_1^*+\boldsymbol{y}_2^*}{2}\right) + \mu\log\det\left(\boldsymbol{C} + \frac{\boldsymbol{W}_1^*+\boldsymbol{W}_2^*}{2} + \mathcal{A}^T\left(\frac{\boldsymbol{y}_1^*+\boldsymbol{y}_2^*}{2}\right)\right) + n\mu - n\mu\log\mu \\
> \ & \frac{1}{2}\left(\boldsymbol{b}^T\boldsymbol{y}_1^* + \mu\log\det(\boldsymbol{C}+\boldsymbol{W}_1^*-\mathcal{A}^T(\boldsymbol{y}_1^*)) + n\mu - n\mu\log\mu\right) \\
& + \frac{1}{2}\left(\boldsymbol{b}^T\boldsymbol{y}_2^* + \mu\log\det(\boldsymbol{C}+\boldsymbol{W}_2^*-\mathcal{A}^T(\boldsymbol{y}_2^*)) + n\mu - n\mu\log\mu\right) = \frac{g^*}{2} + \frac{g^*}{2} = g^*.
\end{aligned}$$

This is a contradiction to the optimality of $g^*$. Hence, we obtain $\boldsymbol{C} + \boldsymbol{W}_1^* - \mathcal{A}^T(\boldsymbol{y}_1^*) = \boldsymbol{C} + \boldsymbol{W}_2^* - \mathcal{A}^T(\boldsymbol{y}_2^*)$, which is equivalent to $\boldsymbol{X}(\boldsymbol{y}_1^*, \boldsymbol{W}_1^*) = \boldsymbol{X}(\boldsymbol{y}_2^*, \boldsymbol{W}_2^*)$. Since the objective values of both $(\boldsymbol{y}_1^*, \boldsymbol{W}_1^*)$ and $(\boldsymbol{y}_2^*, \boldsymbol{W}_2^*)$ are $g^*$, it is easy to show $\boldsymbol{b}^T\boldsymbol{y}_1^* = \boldsymbol{b}^T\boldsymbol{y}_2^*$ from $\boldsymbol{C} + \boldsymbol{W}_1^* - \mathcal{A}^T(\boldsymbol{y}_1^*) = \boldsymbol{C} + \boldsymbol{W}_2^* - \mathcal{A}^T(\boldsymbol{y}_2^*)$.

The uniqueness of optimal solution in $(\mathcal{P})$ can also be obtained by the same argument using (7). □

Next, we examine the validity of the stopping criteria in Algorithm 2.1.

**Lemma 3.8.** *$(\boldsymbol{y}^*, \boldsymbol{W}^*)$ is optimal for $(\mathcal{D})$ if and only if $(\boldsymbol{y}^*, \boldsymbol{W}^*) \in \mathcal{F}$ and*

$$\boldsymbol{P}_{\widehat{\mathcal{W}}}((\boldsymbol{y}^*, \boldsymbol{W}^*) + \alpha\nabla g(\boldsymbol{y}^*, W^*)) = (\boldsymbol{y}^*, \boldsymbol{W}^*) \tag{8}$$

*for some $\alpha > 0$.*

As proven in [8], for a general convex problem

$$\min f_1(\boldsymbol{x}) \quad \text{s.t.} \quad \boldsymbol{x} \in S_1$$

with a differentiable convex function $f_1 : \mathbb{R}^n \to \mathbb{R}$ and a closed convex set $S_1 \subset \mathbb{R}^n$, a point $\boldsymbol{x}^* \in S_1$ is optimal if and only if $\boldsymbol{P}_{S_1}(\boldsymbol{x}^* - \alpha\nabla f_1(\boldsymbol{x}^*)) = \boldsymbol{x}^*$ for some $\alpha > 0$. This condition is further extended to $\boldsymbol{P}_{S_1}(\boldsymbol{x}^* - \alpha\nabla f_1(\boldsymbol{x}^*)) = \boldsymbol{x}^*$ for *any* $\alpha > 0$. This results cannot be applied to $(\mathcal{D})$ since the projection onto the intersection $\mathcal{F} = \widehat{\mathcal{W}} \cap \widehat{\mathcal{F}}$ is not available at a low computation cost. The projection considered in the proposed method is onto $\widehat{\mathcal{W}}$, thus we prove Lemma 3.8 as follows.

*Proof.* It is easy to show that the condition (8) for some $\alpha > 0$ is equivalent to (8) for any $\alpha > 0$, following the proof for the condition (P6) of [8].

We now suppose that $(\boldsymbol{y}^*, \boldsymbol{W}^*) \in \mathcal{F}$ and $\boldsymbol{P}_{\widehat{\mathcal{W}}}((\boldsymbol{y}^*, \boldsymbol{W}^*) + \alpha \nabla g(\boldsymbol{y}^*, \boldsymbol{W}^*)) = (\boldsymbol{y}^*, \boldsymbol{W}^*)$ for any $\alpha > 0$. Let $\boldsymbol{X}^* := \boldsymbol{X}(\boldsymbol{y}^*, \boldsymbol{W}^*) = \mu(\boldsymbol{C} + \boldsymbol{W}^* - \mathcal{A}^T(\boldsymbol{y}^*))^{-1}$. By considering the definitions of $\boldsymbol{P}_{\widehat{\mathcal{W}}}$ and $\nabla g$ into (8), we have two equalities $\mathcal{A}(\boldsymbol{X}^*) = \boldsymbol{b}$ and $[\boldsymbol{W}^* + \alpha \boldsymbol{X}^*]_{\leq \boldsymbol{\rho}} = \boldsymbol{W}^*$. Since $\boldsymbol{X}^* \succ O$, $\boldsymbol{X}^*$ is a feasible point of $(\mathcal{P})$. The second equality $[\boldsymbol{W}^* + \alpha \boldsymbol{X}^*]_{\leq \boldsymbol{\rho}} = \boldsymbol{W}^*$ indicates the three cases:

Case 1 $(X^*_{ij} > 0)$ : There exists $\alpha > 0$ such that $W^*_{ij} + \alpha X^*_{ij} > \rho_{ij}$. From $[\boldsymbol{W}^* + \alpha \boldsymbol{X}^*]_{\leq \boldsymbol{\rho}} = \boldsymbol{W}^*$, we obtain $W^*_{ij} = \rho_{ij}$ .

Case 2 $(X^*_{ij} < 0)$ : In a similar way to Case 1, we obtain $W^*_{ij} = -\rho_{ij}$.

Case 3 $(X^*_{ij} = 0)$ : In this case, we know only $|W^*_{ij}| \leq \rho_{ij}$.

Using the relations $\boldsymbol{X}^* = \mu(\boldsymbol{C} + \boldsymbol{W}^* - \mathcal{A}^T(\boldsymbol{y}^*))^{-1}$ and $\mathcal{A}(\boldsymbol{X}^*) = \boldsymbol{b}$, we consider the difference of the primal and dual objective functions,

$$\begin{aligned} & f(\boldsymbol{X}^*) - g(\boldsymbol{y}^*, \boldsymbol{W}^*) \\ = \; & (\boldsymbol{C} \bullet \boldsymbol{X}^* - \mu \log\det \boldsymbol{X}^* + \boldsymbol{\rho} \bullet |\boldsymbol{X}^*|) \\ & - \big(\boldsymbol{b}^T \boldsymbol{y}^* + \mu \log\det(\boldsymbol{C} + \boldsymbol{W}^* - \mathcal{A}^T(\boldsymbol{y}^*)) + n\mu(1 - \log \mu)\big) \\ = \; & \boldsymbol{\rho} \bullet |\boldsymbol{X}^*| - \boldsymbol{W}^* \bullet \boldsymbol{X}^* \end{aligned} \tag{9}$$

The above three cases imply that this difference is 0. Note that $\boldsymbol{X}^*$ and $(\boldsymbol{y}^*, \boldsymbol{W}^*)$ are feasible for $(\mathcal{P})$ and $(\mathcal{D})$, respectively, and there is no duality gap, hence, $\boldsymbol{X}^*$ and $(\boldsymbol{y}^*, \boldsymbol{W}^*)$ are optimal for $(\mathcal{P})$ and $(\mathcal{D})$.

For the converse, we suppose that $(\boldsymbol{y}^*, \boldsymbol{W}^*)$ is an optimal solution of $(\mathcal{D})$. Again, let $\boldsymbol{X}^* = \mu(\boldsymbol{C} + \boldsymbol{W}^* - \mathcal{A}^*(\boldsymbol{y}^T))^{-1}$. Since $(\mathcal{D})$ is a concave maximization problem, $(\boldsymbol{y}^*, \boldsymbol{W}^*)$ satisfies

$$\nabla g(\boldsymbol{y}^*, \boldsymbol{W}^*) \bullet ((\boldsymbol{y}, \boldsymbol{W}) - (\boldsymbol{y}^*, \boldsymbol{W}^*)) \leq 0 \quad \text{for} \quad \forall (\boldsymbol{y}, \boldsymbol{W}) \in \mathcal{F},$$

or equivalently,

$$(\boldsymbol{b} - \mathcal{A}(\boldsymbol{X}^*))^T (\boldsymbol{y} - \boldsymbol{y}^*) + \boldsymbol{X}^* \bullet (\boldsymbol{W} - \boldsymbol{W}^*) \leq 0 \quad \text{for} \quad \forall (\boldsymbol{y}, \boldsymbol{W}) \in \mathcal{F}. \tag{10}$$

Since $\boldsymbol{C} + \boldsymbol{W}^* - \mathcal{A}^T(\boldsymbol{y}^*) \succ \boldsymbol{O}$ and $\mathcal{A}^T$ is a continuous map, there is a small $t > 0$ such that $\boldsymbol{C} + \boldsymbol{W}^* - \mathcal{A}^T(\boldsymbol{y}^* + t(\boldsymbol{b} - \mathcal{A}(\boldsymbol{X}^*))) \succ \boldsymbol{O}$. Therefore $(\boldsymbol{y}^* + t(\boldsymbol{b} - \mathcal{A}(\boldsymbol{X}^*)), \boldsymbol{W}^*)$ is feasible, and when we put $(\boldsymbol{y}^* + t(\boldsymbol{b} - \mathcal{A}(\boldsymbol{X}^*)), \boldsymbol{W}^*) \in \mathcal{F}$ into $(\boldsymbol{y}, \boldsymbol{W})$ of (10), we obtain $\mathcal{A}(\boldsymbol{X}^*) = \boldsymbol{b}$. Hence, we have $\boldsymbol{y}^* + \alpha(\boldsymbol{b} - \mathcal{A}(\boldsymbol{X}^*)) = \boldsymbol{y}^*$. Similarily, when we perturb $\boldsymbol{W}^*$ in element-wise, we obtain two indications; if $\boldsymbol{X}^*_{ij} > 0$ then $\boldsymbol{W}^*_{ij} = \rho_{ij}$ and if $\boldsymbol{X}^*_{ij} < 0$ then $\boldsymbol{W}^*_{ij} = -\rho_{ij}$. This leads to the results $[\boldsymbol{W}^* + \alpha \boldsymbol{X}^*]_{\leq \boldsymbol{\rho}} = \boldsymbol{W}^*$. Hence, we have shown that (8) holds for $\forall \alpha > 0$. □

From Lemma 3.8 and Lemma 3.7, we also find the relation of the optimal solutions of $(\mathcal{P})$ and $(\mathcal{D})$.

**Remark 3.9.** The matrix $\boldsymbol{X}^*$ computed by $\boldsymbol{X}^* := \boldsymbol{X}(\boldsymbol{y}^*, \boldsymbol{W}^*)$ for an optimal solution $(\boldsymbol{y}^*, \boldsymbol{W}^*)$ of $(\mathcal{D})$ is the unique optimal solution of $(\mathcal{P})$. Furthermore, from $(\boldsymbol{y}^*, \boldsymbol{W}^*) \in \mathcal{L}$ and Remark 3.5, the optimal solution $\boldsymbol{X}^*$ satisfies $\beta^{\min}_{\boldsymbol{X}} \boldsymbol{I} \preceq \boldsymbol{X}^* \preceq \beta^{\max}_{\boldsymbol{X}} \boldsymbol{I}$ and $||\boldsymbol{X}^*|| \leq \eta_{\boldsymbol{X}}$.

From the definition in (3), $(\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k)$ depends on $\alpha^k$. However, the stopping criteria shown in Lemma 3.8 is practically independent of $\alpha^k$. For the subsequent analysis, we introduce $(\Delta \boldsymbol{y}^k_{(1)}, \Delta \boldsymbol{W}^k_{(1)})$ by setting $\alpha^k = 1$;

$$\begin{aligned} (\Delta \boldsymbol{y}^k_{(1)}, \Delta \boldsymbol{W}^k_{(1)}) \; &:= \; \boldsymbol{P}_{\widehat{\mathcal{W}}}((\boldsymbol{y}^k, \boldsymbol{W}^k) + \nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k)) - (\boldsymbol{y}^k, \boldsymbol{W}^k) \\ &= \; (\boldsymbol{b} - \mathcal{A}(\boldsymbol{X}^k), [\boldsymbol{W}^k + \boldsymbol{X}^k]_{\leq \boldsymbol{\rho}} - \boldsymbol{W}^k). \end{aligned} \tag{11}$$

and we now investigate the relation between $(\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k)$ and $(\Delta \boldsymbol{y}^k_{(1)}, \Delta \boldsymbol{W}^k_{(1)})$.

**Lemma 3.10.** *The search direction $(\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k)$ is bounded by $(\Delta \boldsymbol{y}^k_{(1)}, \Delta \boldsymbol{W}^k_{(1)})$. More precisely,*

$$\min\{1, \alpha_{\min}\}||(\Delta \boldsymbol{y}^k_{(1)}, \Delta \boldsymbol{W}^k_{(1)})|| \leq ||(\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k)|| \leq \max\{1, \alpha_{\max}\}||(\Delta \boldsymbol{y}^k_{(1)}, \Delta \boldsymbol{W}^k_{(1)})||. \quad (12)$$

*Proof.* It holds that $\Delta \boldsymbol{y}^k = \alpha^k \Delta \boldsymbol{y}^k_{(1)}$ from the definitions. From (P4) of Proposition 3.1, we know that $||\boldsymbol{P}_{\mathcal{W}}(\boldsymbol{W}^k + \alpha \boldsymbol{X}^k) - \boldsymbol{W}^k||$ is non-decreasing for $\alpha > 0$, therefore, it holds for the case $\alpha^k > 1$ that $||\Delta \boldsymbol{W}^k|| = |||[\boldsymbol{W}^k + \alpha^k \boldsymbol{X}^k]_{\leq \boldsymbol{\rho}} - \boldsymbol{W}^k|| \geq |||[\boldsymbol{W}^k + \boldsymbol{X}^k]_{\leq \boldsymbol{\rho}} - \boldsymbol{W}^k|| = ||\Delta \boldsymbol{W}^k_{(1)}||$. In addition, (P5) of Proposition 3.1 indicates that $||\boldsymbol{P}_{\mathcal{W}}(\boldsymbol{W}^k + \alpha \boldsymbol{X}^k) - \boldsymbol{W}^k||/\alpha$ is non-increasing for $\alpha > 0$. Since we choose $\alpha^k$ from $[\alpha_{\min}, \alpha_{\max}]$, we have $||\Delta \boldsymbol{W}^k|| = |||[\boldsymbol{W}^k + \alpha^k \boldsymbol{X}^k]_{\leq \boldsymbol{\rho}} - \boldsymbol{W}^k|| \geq \alpha^k |||[\boldsymbol{W}^k + \boldsymbol{X}^k]_{\leq \boldsymbol{\rho}} - \boldsymbol{W}^k|| \geq \alpha_{\min} |||[\boldsymbol{W}^k + \boldsymbol{X}^k]_{\leq \boldsymbol{\rho}} - \boldsymbol{W}^k|| = \alpha_{\min} ||\Delta \boldsymbol{W}^k_{(1)}||$ for the case $\alpha^k \leq 1$. The combination of these two shows the left inequality of (12). The right inequality is also derived from (P4) and (P5) in a similar way.

□

The condition $||(\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k)|| > 0$ can be assumed without loss of generality, since $||(\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k)|| = 0$ indicates that $(\boldsymbol{y}^k, \boldsymbol{W}^k)$ is an optimal solution by Lemmas 3.8 and 3.10 and (11) and that Algorithm 2.1 stops at Step 2.

Algorithm 2.1 may terminate before computing an approximate solution with a required accuracy in the following two cases: (i) The step length $\lambda^k$ converges to 0 before $||(\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k)||$ reaches 0, and $(\boldsymbol{y}^k, \boldsymbol{W}^k)$ cannot proceed, (ii) The norm of the search direction $||(\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k)||$ converges to 0 before $g(\boldsymbol{y}^k, \boldsymbol{W}^k)$ reaches the optimal value $g^*$. Lemmas 3.12 and 3.15 show that the two cases will not happen. For the proofs of the two lemmas, we first discuss some inequalities related to matrix norms.

**Lemma 3.11.** *Suppose that $0 < \widehat{\beta}^{\min} < \widehat{\beta}^{\max} < \infty$. For $\forall \boldsymbol{X}, \forall \boldsymbol{Y} \in S_2 := \{\boldsymbol{X} \in \mathbb{S}^n : \widehat{\beta}^{\min} \boldsymbol{I} \preceq \boldsymbol{X} \preceq \widehat{\beta}^{\max} \boldsymbol{I}\}$, it holds*

*(i)* $(\boldsymbol{Y} - \boldsymbol{X}) \bullet (\boldsymbol{X}^{-1} - \boldsymbol{Y}^{-1}) \geq \frac{1}{(\widehat{\beta}^{\max})^2} ||\boldsymbol{Y} - \boldsymbol{X}||^2$,

*(ii)* $(\boldsymbol{Y} - \boldsymbol{X}) \bullet (\boldsymbol{X}^{-1} - \boldsymbol{Y}^{-1}) \geq (\widehat{\beta}^{\min})^2 ||\boldsymbol{Y}^{-1} - \boldsymbol{X}^{-1}||^2$,

*(iii)* $||\boldsymbol{Y} - \boldsymbol{X}|| \geq (\widehat{\beta}^{\min})^2 ||\boldsymbol{Y}^{-1} - \boldsymbol{X}^{-1}||$.

*Proof.* From the discussions of [5], the function $f_2(\boldsymbol{X}) = -\log\det(\boldsymbol{X})$ is strongly convex with the convexity parameter $\frac{1}{2(\widehat{\beta}^{\max})^2}$ on the set $S_2$. Therefore, it holds that

$$f_2(\boldsymbol{Y}) \geq f_2(\boldsymbol{X}) + \nabla f_2(\boldsymbol{X}) \bullet (\boldsymbol{Y} - \boldsymbol{X}) + \frac{1}{2(\widehat{\beta}^{\max})^2} ||\boldsymbol{Y} - \boldsymbol{X}||^2 \quad (13)$$

for $\forall \boldsymbol{X}, \forall \boldsymbol{Y} \in S_2$. By swapping $\boldsymbol{X}$ and $\boldsymbol{Y}$, we also have

$$f_2(\boldsymbol{X}) \geq f_2(\boldsymbol{Y}) + \nabla f_2(\boldsymbol{Y}) \bullet (\boldsymbol{X} - \boldsymbol{Y}) + \frac{1}{2(\widehat{\beta}^{\max})^2} ||\boldsymbol{X} - \boldsymbol{Y}||^2.$$

Since $\nabla f_2(\boldsymbol{X}) = -\boldsymbol{X}^{-1}$, adding these two inequalities generates (i). When we use $\boldsymbol{X}^{-1}, \boldsymbol{Y}^{-1} \in \{\boldsymbol{X} : \frac{1}{\widehat{\beta}^{\max}} \boldsymbol{I} \preceq \boldsymbol{X} \preceq \frac{1}{\widehat{\beta}^{\min}} \boldsymbol{I}\}$, we obtain (ii) in a similar way to (i). Finally, an application of the Cauchy-Schwartz inequality to (ii) lead to

$$(\widehat{\beta}^{\min})^2 ||\boldsymbol{Y}^{-1} - \boldsymbol{X}^{-1}||^2 \leq (\boldsymbol{Y} - \boldsymbol{X}) \bullet (\boldsymbol{X}^{-1} - \boldsymbol{Y}^{-1}) \leq ||\boldsymbol{Y} - \boldsymbol{X}|| \cdot ||\boldsymbol{X}^{-1} - \boldsymbol{Y}^{-1}||.$$

If $\boldsymbol{X} \neq \boldsymbol{Y}$, (iii) is obtained by dividing the both sides with $||\boldsymbol{X}^{-1} - \boldsymbol{Y}^{-1}||$, meanwhile if $\boldsymbol{X} = \boldsymbol{Y}$, (iii) is obvious. □

**Lemma 3.12.** *The step length $\lambda^k$ of Algorithm 2.1 has a lower bound,*

$$\lambda^k \geq \min\left\{\overline{\lambda}_{\min}, \frac{2\sigma_1(1-\gamma)}{L\alpha_{\max}}\right\}$$

*where* $\overline{\lambda}_{\min} := \min\left\{1, \frac{\beta_{\boldsymbol{Z}}^{\min}\tau}{\eta_{\Delta \boldsymbol{W}}+||\mathcal{A}^T||\eta_{\Delta \boldsymbol{y}}}\right\}$ *and* $L := \frac{\mu\sqrt{2(||\mathcal{A}||^2+1)}\max\{1,||\mathcal{A}||\}}{((1-\tau)\beta_{\boldsymbol{Z}}^{\min})^2}$.

*Proof.* We first show the lower bound of $\overline{\lambda}^k$ of Step 3. Since $\overline{\lambda}^k$ is determined by (4), we examine a bound of $\lambda$ such that $\boldsymbol{Z}(\lambda) := \boldsymbol{C}+(\boldsymbol{W}^k+\lambda\Delta\boldsymbol{W}^k)-\mathcal{A}^T(\boldsymbol{y}^k+\lambda\Delta\boldsymbol{y}^k) \succeq \boldsymbol{O}$. It follows from Remark 3.5 that $\mu(\boldsymbol{X}^k)^{-1} \succeq \beta_{\boldsymbol{Z}}^{\min}\boldsymbol{I}$. From Remark 3.6, we also have $||\Delta\boldsymbol{y}^k|| \leq \eta_{\Delta\boldsymbol{y}}$ and $||\Delta\boldsymbol{W}^k|| \leq \eta_{\Delta\boldsymbol{W}}$. Therefore, we obtain

$$\begin{aligned}\boldsymbol{Z}(\lambda) &= \mu(\boldsymbol{X}^k)^{-1} + \lambda(\Delta\boldsymbol{W}^k - \mathcal{A}^T(\Delta\boldsymbol{y}^k)) \\ &\succeq \beta_{\boldsymbol{Z}}^{\min}\boldsymbol{I} - \lambda(\eta_{\Delta\boldsymbol{W}} + ||\mathcal{A}^T||\eta_{\Delta\boldsymbol{y}})\boldsymbol{I}. \end{aligned} \tag{14}$$

Hence, for any $\lambda \in \left[0, \frac{\beta_{\boldsymbol{Z}}^{\min}}{\eta_{\Delta\boldsymbol{W}}+||\mathcal{A}^T||\eta_{\Delta\boldsymbol{y}}}\right]$, we have $\boldsymbol{Z}(\lambda) \succeq \boldsymbol{O}$, and consequently, we obtain $\overline{\lambda}^k \geq \overline{\lambda}_{\min}$.

If $\theta$ of (4) is non-negative, $\boldsymbol{Z}(\lambda) \succeq \boldsymbol{Z}(0) \succeq (1-\tau)\boldsymbol{Z}(0)$. In the case $\theta < 0$, we have $\overline{\lambda}^k \geq -\frac{1}{\theta}\times\tau$, and this leads to $\boldsymbol{Z}(\lambda) \succeq (1-\tau)\boldsymbol{Z}(0)$ for $\lambda \in [0, \overline{\lambda}^k]$. Therefore, $\boldsymbol{Z}(\lambda) \succeq (1-\tau)\boldsymbol{Z}(0) \succeq (1-\tau)\beta_{\boldsymbol{Z}}^{\min}\boldsymbol{I}$ for $\lambda \in [0, \overline{\lambda}^k]$. Hence, it follows from (iii) of Lemma 3.11 that

$$||\boldsymbol{Z}(\lambda)^{-1} - \boldsymbol{Z}(0)^{-1}|| \leq \frac{||\boldsymbol{Z}(\lambda) - \boldsymbol{Z}(0)||}{((1-\tau)\beta_{\boldsymbol{Z}}^{\min})^2} \quad \text{for} \quad \lambda \in [0, \overline{\lambda}^k].$$

Hence, we acquire some Lipschitz continuity on $\nabla g$ for the direction $(\Delta\boldsymbol{y}^k, \Delta\boldsymbol{W}^k)$. For $\lambda \in [0, \overline{\lambda}^k]$, we have

$$\begin{aligned} & ||\nabla g(\boldsymbol{y}^k + \lambda\Delta\boldsymbol{y}^k, \boldsymbol{W}^k + \lambda\Delta\boldsymbol{W}^k) - \nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k)|| \\ = & \left|\left|\left(\boldsymbol{b} - \mathcal{A}(\mu\boldsymbol{Z}(\lambda)^{-1}), \mu\boldsymbol{Z}(\lambda)^{-1}\right)\right) - \left(\boldsymbol{b} - \mathcal{A}(\mu\boldsymbol{Z}(0)^{-1}), \mu\boldsymbol{Z}(0)^{-1}\right)\right)\right| \\ = & \mu\left|\left|\left(-\mathcal{A}(\boldsymbol{Z}(\lambda)^{-1}) + \mathcal{A}(\boldsymbol{Z}(0)^{-1}), \boldsymbol{Z}(\lambda)^{-1} - \boldsymbol{Z}(0)^{-1}\right)\right|\right| \\ \leq & \mu\sqrt{||\mathcal{A}||^2+1}||\boldsymbol{Z}(\lambda)^{-1} - \boldsymbol{Z}(0)^{-1}|| \\ \leq & \frac{\mu\sqrt{||\mathcal{A}||^2+1}}{((1-\tau)\beta_{\boldsymbol{Z}}^{\min})^2}||\boldsymbol{Z}(\lambda) - \boldsymbol{Z}(0)|| \\ = & \frac{\lambda\mu\sqrt{||\mathcal{A}||^2+1}}{((1-\tau)\beta_{\boldsymbol{Z}}^{\min})^2}||\Delta\boldsymbol{W}^k - \mathcal{A}^T(\Delta\boldsymbol{y}^k)|| \\ \leq & \frac{\lambda\mu\sqrt{2(||\mathcal{A}||^2+1)}\max\{1,||\mathcal{A}||\}}{((1-\tau)\beta_{\boldsymbol{Z}}^{\min})^2}||(\Delta\boldsymbol{y}^k, \Delta\boldsymbol{W}^k)|| \\ = & \lambda L||(\Delta\boldsymbol{y}^k, \Delta\boldsymbol{W}^k)||, \end{aligned} \tag{15}$$

Here, we have used the inequalities $||\Delta\boldsymbol{W}^k - \mathcal{A}^T(\Delta\boldsymbol{y}^k)|| \leq ||\Delta\boldsymbol{W}^k|| + ||\mathcal{A}^T||\cdot||\Delta\boldsymbol{y}^k||$ and $||\Delta\boldsymbol{W}^k|| + ||\Delta\boldsymbol{y}^k|| \leq \sqrt{2}||(\Delta\boldsymbol{y}^k, \Delta\boldsymbol{W}^k)||$.

We examine how the inner loop, Step 3 of Algorithm 2.1, is executed. As in the Armijo rule, the inner loop terminates at a finite number of inner iterations. If (5) is satisfied at $j = 1$, then $\lambda^k = \overline{\lambda}^k \geq \overline{\lambda}_{\min}$. If (5) is satisfied at $j \geq 2$, then (5) is not satisfied at $j-1$. Thus, we have

$$\begin{aligned} & g(\boldsymbol{y}^k + \lambda_{j-1}^k\Delta\boldsymbol{y}^k, \boldsymbol{W}^k + \lambda_{j-1}^k\Delta\boldsymbol{W}^k) \\ < & \min_{0\leq h\leq\min\{k,M-1\}} g(\boldsymbol{y}^{k-h}, \boldsymbol{W}^{k-h}) + \gamma\lambda_{j-1}^k\nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k) \bullet (\Delta\boldsymbol{y}^k, \Delta\boldsymbol{W}^k) \\ \leq & g(\boldsymbol{y}^k, \boldsymbol{W}^k) + \gamma\lambda_{j-1}^k\nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k) \bullet (\Delta\boldsymbol{y}^k, \Delta\boldsymbol{W}^k). \end{aligned}$$

From Taylor's expansion and (15), it follows that

$$\begin{aligned} & g(\boldsymbol{y}^k + \lambda_{j-1}^k \Delta \boldsymbol{y}^k, \boldsymbol{W}^k + \lambda_{j-1}^k \Delta \boldsymbol{W}^k) - g(\boldsymbol{y}^k, \boldsymbol{W}^k) \\ = \; & \lambda_{j-1}^k \nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k) \bullet (\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k) \\ & + \int_0^{\lambda_{j-1}^k} \Big( \nabla g(\boldsymbol{y}^k + \lambda \Delta \boldsymbol{y}^k, \boldsymbol{W}^k + \lambda \Delta \boldsymbol{W}^k) - \nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k) \Big) \bullet (\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k) d\lambda \\ \geq \; & \lambda_{j-1}^k \nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k) \bullet (\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k) - \frac{(\lambda_{j-1}^k)^2 L}{2} ||(\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k)||^2, \end{aligned}$$

since $\lambda_{j-1}^k \leq \overline{\lambda}^k$. Combining these two inequalities, we obtain $\lambda_{j-1}^k \geq \frac{2(1-\gamma)}{L} \frac{\nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k) \bullet (\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k)}{||(\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k)||^2}$. It follows from (6) that

$$\frac{\nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k) \bullet (\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k)}{||(\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k)||^2} \geq \frac{1}{\alpha^k} \geq \frac{1}{\alpha_{\max}}. \tag{16}$$

Since $\lambda_j^k$ is chosen from $[\sigma_1 \lambda_{j-1}^k, \sigma_2 \lambda_{j-1}^k]$, we obtain $\lambda^k = \lambda_j^k \geq \frac{2\sigma_1(1-\gamma)}{L\alpha_{\max}}$. □

We now prove that the search direction generated by Algorithm 2.1 shrinks to zero in the infinite iterations.

**Lemma 3.13.** *Algorithm 2.1 with $\epsilon = 0$ stops in a finite number of iterations attaining the optimal value $g^*$, or the infimum of the norm of the search direction tends to zero as $k$ increases,*

$$\liminf_{k \to \infty} ||(\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k)|| = 0.$$

*Proof.* When Algorithm 2.1 stops in a finite number of iterations, the optimality is guaranteed by Lemma 3.8. From Lemma 3.10, it is sufficient to prove $\liminf_{k \to \infty} ||(\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k)|| = 0$, Suppose, to contrary, that there exist $\delta > 0$ and an integer $k_0$ such that $||(\Delta \boldsymbol{y}^k, \Delta \boldsymbol{W}^k)|| > \delta$ for $\forall k \geq k_0$. Let us denote $g_k := g(\boldsymbol{y}^k, \boldsymbol{W}^k)$ and $g_\ell^{\min} := \min\{g_{\ell M+1}, \ldots, g_{(\ell+1)M}\}$. It follows from Lemma 3.12, (5) and (16) that

$$g_{k+1} \geq \min\{g_k, \ldots, g_{k-M+1}\} + \widehat{\delta} \quad \text{for} \quad \forall k \geq \max\{k_0, M\},$$

where $\widehat{\delta} = \gamma \min\{\overline{\lambda}_{\min}, \frac{2\sigma_1(1-\gamma)}{L\alpha_{\max}}\} \frac{\delta^2}{\alpha_{\max}}$.

When $\ell$ is an integer such that $\ell > \frac{\max\{k_0, M\}}{M}$, we have

$$g_{(\ell+1)M+1} \geq \min\{g_{(\ell+1)M}, \ldots g_{(\ell+1)M-M+1}\} + \widehat{\delta} = g_\ell^{\min} + \widehat{\delta}.$$

By induction, for $j = 2, \ldots, M$,

$$g_{(\ell+1)M+j} \geq \min\{g_{(\ell+1)M+j-1}, \ldots g_{(\ell+1)M-M+j}\} + \widehat{\delta} \geq \min\{g_\ell^{\min} + \widehat{\delta}, g_\ell^{\min}\} + \widehat{\delta} = g_\ell^{\min} + \widehat{\delta}.$$

Therefore, we obtain

$$g_{\ell+1}^{\min} = \min\{g_{(\ell+1)M+1}, \ldots, g_{(\ell+1)M+M}\} \geq g_\ell^{\min} + \widehat{\delta}.$$

From Lemma 3.2, we know $g(\boldsymbol{y}^0, \boldsymbol{W}^0) \leq g_k \leq g^*$ for each $k$. Starting from an integer $\ell_0$ such that $\ell_0 > \frac{\max\{k_0, M\}}{M}$, it follows that

$$g^* \geq g_\ell^{\min} \geq g_{\ell_0}^{\min} + (\ell - \ell_0)\widehat{\delta} \geq g(\boldsymbol{y}^0, \boldsymbol{W}^0) + (\ell - \ell_0)\widehat{\delta} \quad \text{for} \quad \ell \geq \ell_0.$$

When we take large $\ell$ such that $\ell > \ell_0 + (g^* - g(\boldsymbol{y}^0, \boldsymbol{W}^0))/\widehat{\delta}$, we have a contradiction. This completes the proof.

□

For the proof of the main theorem, we further investigate the behavior of the objective function in Lemma 3.15, which requires Lemma 3.14. We use a matrix $\boldsymbol{U}^k \in \mathbb{S}^n$ defined by $U_{ij}^k := \rho_{ij}|X_{ij}^k| - W_{ij}^k X_{ij}^k$, and $\rho^{\max} := \max\{\rho_{ij} : i, j = 1, \ldots, n\}$. The notation $[\Delta \boldsymbol{W}_{(1)}^k]_{ij}$ denotes the $(i,j)$th element of $\Delta \boldsymbol{W}_{(1)}^k = [\boldsymbol{W}^k + \boldsymbol{X}^k]_{\leq \rho} - \boldsymbol{W}^k$.

**Lemma 3.14.** *It holds that*

$$|\boldsymbol{U}^k| \leq \max\{2\rho^{\max}, \eta_{\boldsymbol{X}}\}|\Delta \boldsymbol{W}_{(1)}^k|. \tag{17}$$

*Proof.* We investigate the inequality by dividing into three cases.

1. Case $X_{ij}^k = 0$: We have $U_{ij}^k = 0$, hence (17) holds.
2. Case $X_{ij}^k > 0$: We have $U_{ij}^k = (\rho_{ij} - W_{ij}^k)X_{ij}^k \geq 0$.
   (a) Case $W_{ij}^k = \rho_{ij}$: We have $U_{ij}^k = 0$, hence (17) holds.
   (b) Case $W_{ij}^k < \rho_{ij}$: If $W_{ij}^k + X_{ij}^k \leq \rho_{ij}$, then $[\Delta \boldsymbol{W}_{(1)}^k]_{ij} = W_{ij}^k + X_{ij}^k - W_{ij}^k = X_{ij}^k$. From $W_{ij}^k \geq -\rho_{ij}$, we have $0 \leq U_{ij}^k = (\rho_{ij} - W_{ij}^k)[\Delta \boldsymbol{W}_{(1)}^k]_{ij} \leq 2\rho_{ij}[\Delta \boldsymbol{W}_{(1)}^k]_{ij} \leq 2\rho^{\max}|[\Delta \boldsymbol{W}_{(1)}^k]_{ij}|$. Otherwise, if $W_{ij}^k + X_{ij}^k > \rho_{ij}$, then $[\Delta \boldsymbol{W}_{(1)}^k]_{ij} = \rho_{ij} - W_{ij}^k$, hence $U_{ij}^k = X_{ij}^k[\Delta \boldsymbol{W}_{(1)}^k]_{ij}$. From $|X_{ij}^k| \leq ||\boldsymbol{X}^k|| \leq \eta_{\boldsymbol{X}}$, we obtain $0 \leq U_{ij}^k \leq \eta_{\boldsymbol{X}}|[\Delta \boldsymbol{W}_{(1)}^k]_{ij}|$.
3. Case $X_{ij}^k < 0$: We compute simliarily to the case $X_{ij}^k > 0$.

Combining these cases results in (17). □

**Lemma 3.15.** *Algorithm 2.1 with $\epsilon = 0$ stops in a finite number of iterations attaining the optimal value $g^*$, or the infimum of the difference of the objective functions between $(\boldsymbol{y}^k, \boldsymbol{W}^k)$ and $(\boldsymbol{y}^*, \boldsymbol{W}^*) \in \mathcal{F}^*$ tends to zero as $k$ increases, i.e.,*

$$\liminf_{k \to \infty} |g(\boldsymbol{y}^k, \boldsymbol{W}^k) - g^*| = 0. \tag{18}$$

*Proof.* If Algorithm 2.1 stops at the $k$th iteration, $(\boldsymbol{y}^k, \boldsymbol{W}^k)$ is an optimal solution, therefore, $g^* = g(\boldsymbol{y}^k, \boldsymbol{W}^k)$. The proof for (18) is based on an inequality

$$|g(\boldsymbol{y}^k, \boldsymbol{W}^k) - g(\boldsymbol{y}^*, \boldsymbol{W}^*)| \leq |g(\boldsymbol{y}^k, \boldsymbol{W}^k) - f(\boldsymbol{X}^k)| + |f(\boldsymbol{X}^k) - f(\boldsymbol{X}^*)| + |f(\boldsymbol{X}^*) - g(\boldsymbol{y}^*, \boldsymbol{W}^*)|. \tag{19}$$

We know that $f(\boldsymbol{X}^*) = g(\boldsymbol{y}^*, \boldsymbol{W}^*)$ from the duality theorem, hence, we evaluate the first and second terms.

From the definition of $f$ and $g$, the first term will be bounded by

$$\begin{aligned} & |f(\boldsymbol{X}^k) - g(\boldsymbol{y}^k, \boldsymbol{W}^k)| \\ = & \left| \boldsymbol{\rho} \bullet |\boldsymbol{X}^k| - \boldsymbol{W}^k \bullet \boldsymbol{X}^k + (\mathcal{A}(\boldsymbol{X}^k) - \boldsymbol{b})^T \boldsymbol{y}^k \right| \\ \leq & \left| \boldsymbol{\rho} \bullet |\boldsymbol{X}^k| - \boldsymbol{W}^k \bullet \boldsymbol{X}^k \right| + \eta_{\boldsymbol{y}} ||\mathcal{A}|| \cdot ||\boldsymbol{X}^k - \boldsymbol{X}^*||. \end{aligned} \tag{20}$$

Using Lemma 3.14, we further have

$$\begin{aligned} \left| \boldsymbol{\rho} \bullet |\boldsymbol{X}^k| - \boldsymbol{W}^k \bullet \boldsymbol{X}^k \right| = \left| \sum_{i=1}^n \sum_{j=1}^n U_{ij}^k \right| = \sum_{i=1}^n \sum_{j=1}^n |U_{ij}^k| \leq \max\{2\rho^{\max}, \eta_{\boldsymbol{X}}\} \sum_{i=1}^n \sum_{j=1}^n |[\Delta \boldsymbol{W}_{(1)}^k]_{ij}| \\ \leq \max\{2\rho^{\max}, \eta_{\boldsymbol{X}}\} n ||\Delta \boldsymbol{W}_{(1)}^k|| \leq \max\{2\rho^{\max}, \eta_{\boldsymbol{X}}\} n ||(\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k)||. \end{aligned} \tag{21}$$

For the second inequality, we have used the relation between the two norms $\sum_{i=1}^n \sum_{j=1}^n |V_{ij}| \leq n||\boldsymbol{V}||$ that holds for any $\boldsymbol{V} \in \mathbb{S}^n$.

Next, we evaluate the second term of (19). Since $f_2(\boldsymbol{X}) = -\log\det(\boldsymbol{X})$ is a convex function,

$$f_2(\boldsymbol{X}^k) \geq f_2(\boldsymbol{X}^*) + \nabla f_2(\boldsymbol{X}^*) \bullet (\boldsymbol{X}^k - \boldsymbol{X}^*)$$

and

$$f_2(\boldsymbol{X}^*) \geq f_2(\boldsymbol{X}^k) + \nabla f_2(\boldsymbol{X}^k) \bullet (\boldsymbol{X}^* - \boldsymbol{X}^k).$$

These two inequalities indicate

$$|f_2(\boldsymbol{X}^k) - f_2(\boldsymbol{X}^*)| \leq \max\{||\nabla f_2(\boldsymbol{X}^k)||, ||\nabla f_2(\boldsymbol{X}^*)||\}||\boldsymbol{X}^k - \boldsymbol{X}^*|| \leq \eta_{\boldsymbol{X}^{-1}}||\boldsymbol{X}^k - \boldsymbol{X}^*||.$$

For the last inequality, we have used $\nabla f_2(\boldsymbol{X}) = -\boldsymbol{X}^{-1}$ for any $\boldsymbol{X} \succ \boldsymbol{O}$ and Remark 3.5. In addition, we have

$$\begin{aligned} |\boldsymbol{\rho} \bullet (|\boldsymbol{X}^k| - |\boldsymbol{X}^*|)| &\leq \sum_{i=1}^n \sum_{j=1}^n \rho_{ij} \left| |X_{ij}^k| - |X_{ij}^*| \right| \\ &\leq \sum_{i=1}^n \sum_{j=1}^n \rho_{ij} |X_{ij}^k - X_{ij}^*| \leq ||\boldsymbol{\rho}|| \cdot ||\boldsymbol{X}^k - \boldsymbol{X}^*||. \end{aligned}$$

Hence, the second term of (19) is bounded by

$$\begin{aligned} & |f(\boldsymbol{X}^k) - f(\boldsymbol{X}^*)| \\ \leq\ & |\boldsymbol{C} \bullet (\boldsymbol{X}^k - \boldsymbol{X}^*)| + \mu |f_2(\boldsymbol{X}^k) - f_2(\boldsymbol{X}^*)| + |\boldsymbol{\rho} \bullet (|\boldsymbol{X}^k| - |\boldsymbol{X}^*|)| \\ \leq\ & (||\boldsymbol{C}|| + \mu\eta_{\boldsymbol{X}^{-1}} + ||\boldsymbol{\rho}||)||\boldsymbol{X}^k - \boldsymbol{X}^*||. \end{aligned} \tag{22}$$

We now evaluate the norm $||\boldsymbol{X}^k - \boldsymbol{X}^*||$. It follows from (P1) of Proposition 3.1 and $(\boldsymbol{y}^*, \boldsymbol{W}^*) \in \widehat{\mathcal{W}}$ that

$$\begin{aligned} \Big( ((\boldsymbol{y}^k, \boldsymbol{W}^k) + \nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k)) - P_{\widehat{\mathcal{W}}}((\boldsymbol{y}^k, \boldsymbol{W}^k) + \nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k)) \Big) \\ \bullet \Big( (\boldsymbol{y}^*, \boldsymbol{W}^*) - P_{\widehat{\mathcal{W}}}((\boldsymbol{y}^k, \boldsymbol{W}^k) + \nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k)) \Big) \leq 0. \end{aligned}$$

Therefore, we obtain

$$\Big( \nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k)) - (\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k) \Big) \bullet \Big( (\boldsymbol{y}^*, \boldsymbol{W}^*) - (\boldsymbol{y}^k, \boldsymbol{W}^k) - (\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k) \Big) \leq 0,$$

and this is equivalent to

$$\begin{aligned} & \Big( \Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k \Big) \bullet \Big( (\boldsymbol{y}^*, \boldsymbol{W}^*) - (\boldsymbol{y}^k, \boldsymbol{W}^k) \Big) + \nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k) \bullet \Big( \Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k \Big) - ||(\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k)||^2 \\ \geq\ & \nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k) \bullet \Big( (\boldsymbol{y}^*, \boldsymbol{W}^*) - (\boldsymbol{y}^k, \boldsymbol{W}^k) \Big). \end{aligned} \tag{23}$$

On the other hand, it follows from (i) of Lemma 3.11 that

$$\begin{aligned} & \Big( \nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k) - \nabla g(\boldsymbol{y}^*, \boldsymbol{W}^*) \Big) \bullet \Big( (\boldsymbol{y}^*, \boldsymbol{W}^*) - (\boldsymbol{y}^k, \boldsymbol{W}^k) \Big) \\ =\ & (-\mathcal{A}(\boldsymbol{X}^k) + \mathcal{A}(\boldsymbol{X}^*), \boldsymbol{X}^k - \boldsymbol{X}^*) \bullet (\boldsymbol{y}^* - \boldsymbol{y}^k, \boldsymbol{W}^* - \boldsymbol{W}^k) \\ =\ & (\boldsymbol{X}^k - \boldsymbol{X}^*) \bullet (-\mathcal{A}^T(\boldsymbol{y}^* - \boldsymbol{y}^k)) + (\boldsymbol{X}^k - \boldsymbol{X}^*) \bullet (\boldsymbol{W}^* - \boldsymbol{W}^k) \\ =\ & (\boldsymbol{X}^k - \boldsymbol{X}^*) \bullet ((\boldsymbol{C} + \boldsymbol{W}^* - \mathcal{A}^T(\boldsymbol{y}^*)) - (\boldsymbol{C} + \boldsymbol{W}^k - \mathcal{A}^T(\boldsymbol{y}^k))) \\ =\ & (\boldsymbol{X}^k - \boldsymbol{X}^*) \bullet (\mu(\boldsymbol{X}^*)^{-1} - \mu(\boldsymbol{X}^k)^{-1}) \\ \geq\ & \frac{\mu}{(\beta_{\boldsymbol{X}}^{\max})^2} ||\boldsymbol{X}^k - \boldsymbol{X}^*||^2, \end{aligned}$$

and this is equivalent to

$$\nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k) \bullet \Big((\boldsymbol{y}^*, \boldsymbol{W}^*) - (\boldsymbol{y}^k, \boldsymbol{W}^k)\Big) \;\; \geq \;\; \frac{\mu}{(\beta_{\boldsymbol{X}}^{\max})^2}||\boldsymbol{X}^k - \boldsymbol{X}^*||^2 + \nabla g(\boldsymbol{y}^*, \boldsymbol{W}^*) \bullet \Big((\boldsymbol{y}^*, \boldsymbol{W}^*) - (\boldsymbol{y}^k, \boldsymbol{W}^k)\Big)$$

By connecting this inequality and (23), we obtain

$$\begin{aligned} & \Big(\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k\Big) \bullet \Big((\boldsymbol{y}^*, \boldsymbol{W}^*) - (\boldsymbol{y}^k, \boldsymbol{W}^k)\Big) + \nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k) \bullet \Big(\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k\Big) - ||(\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k)||^2 \\ \geq \;\; & \frac{\mu}{(\beta_{\boldsymbol{X}}^{\max})^2}||\boldsymbol{X}^k - \boldsymbol{X}^*||^2 + \nabla g(\boldsymbol{y}^*, \boldsymbol{W}^*) \bullet \Big((\boldsymbol{y}^*, \boldsymbol{W}^*) - (\boldsymbol{y}^k, \boldsymbol{W}^k)\Big), \end{aligned}$$

and this is equivalent to

$$\begin{aligned} & \frac{\mu}{(\beta_{\boldsymbol{X}}^{\max})^2}||\boldsymbol{X}^k - \boldsymbol{X}^*||^2 - \Big(\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k\Big) \bullet \Big((\boldsymbol{y}^*, \boldsymbol{W}^*) - (\boldsymbol{y}^k, \boldsymbol{W}^k)\Big) + ||(\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k)||^2 \\ \leq \;\; & \nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k) \bullet \Big(\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k\Big) - \nabla g(\boldsymbol{y}^*, \boldsymbol{W}^*) \bullet \Big((\boldsymbol{y}^*, \boldsymbol{W}^*) - (\boldsymbol{y}^k, \boldsymbol{W}^k)\Big). \end{aligned} \tag{24}$$

Since (9) and there is no duality gap, we know that $\boldsymbol{X}^* \bullet \boldsymbol{W}^* = \boldsymbol{\rho} \bullet |\boldsymbol{X}^*|$. Therefore,

$$\begin{aligned} & \nabla g(\boldsymbol{y}^*, \boldsymbol{W}^*) \bullet \Big((\boldsymbol{y}^*, \boldsymbol{W}^*) - (\boldsymbol{y}^k, \boldsymbol{W}^k) - (\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k)\Big) \\ = \;\; & (\boldsymbol{b} - \boldsymbol{A}(\boldsymbol{X}^*), \boldsymbol{X}^*) \bullet \Big((\boldsymbol{y}^*, \boldsymbol{W}^*) - (\boldsymbol{y}^k, \boldsymbol{W}^k) - (\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k)\Big) \\ = \;\; & (\boldsymbol{0}, \boldsymbol{X}^*) \bullet \Big((\boldsymbol{y}^* - \boldsymbol{y}^k - \Delta \boldsymbol{y}_{(1)}^k, \boldsymbol{W}^* - \boldsymbol{W}^k - \Delta \boldsymbol{W}_{(1)}^k)\Big) \\ = \;\; & \boldsymbol{X}^* \bullet \boldsymbol{W}^* - \boldsymbol{X}^* \bullet [\boldsymbol{W}^k + \boldsymbol{X}^k]_{\leq \boldsymbol{\rho}} \\ = \;\; & |\boldsymbol{X}^*| \bullet \boldsymbol{\rho} - \boldsymbol{X}^* \bullet [\boldsymbol{W}^k + \boldsymbol{X}^k]_{\leq \boldsymbol{\rho}} \\ \geq \;\; & 0. \end{aligned}$$

Hence, it follows that

$$\begin{aligned} & \nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k) \bullet (\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k) - \nabla g(\boldsymbol{y}^*, \boldsymbol{W}^*) \bullet \Big((\boldsymbol{y}^*, \boldsymbol{W}^*) - (\boldsymbol{y}^k, \boldsymbol{W}^k)\Big) \\ = \;\; & \Big(\nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k) - \nabla g(\boldsymbol{y}^*, \boldsymbol{W}^*)\Big) \bullet (\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k) \\ & \quad -\nabla g(\boldsymbol{y}^*, \boldsymbol{W}^*) \bullet \Big((\boldsymbol{y}^*, \boldsymbol{W}^*) - (\boldsymbol{y}^k, \boldsymbol{W}^k) - (\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k)\Big) \\ \leq \;\; & \Big(\nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k) - \nabla g(\boldsymbol{y}^*, \boldsymbol{W}^*)\Big) \bullet (\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k) \\ \leq \;\; & ||\nabla g(\boldsymbol{y}^*, \boldsymbol{W}^*) - \nabla g(\boldsymbol{y}^k, \boldsymbol{W}^k)|| \cdot ||(\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k)|| \\ = \;\; & ||(-\mathcal{A}(\boldsymbol{X}^*) + \mathcal{A}(\boldsymbol{X}^k), \boldsymbol{X}^* - \boldsymbol{X}^k)|| \cdot ||(\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k)|| \\ \leq \;\; & (1 + ||\mathcal{A}||)||\boldsymbol{X}^k - \boldsymbol{X}^*|| \cdot ||(\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k)||. \end{aligned} \tag{25}$$

From (24) and (25), we obtain

$$\begin{aligned} & \frac{\mu}{(\beta_{\boldsymbol{X}}^{\max})^2}||\boldsymbol{X}^k - \boldsymbol{X}^*||^2 - (\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k) \bullet \Big((\boldsymbol{y}^*, \boldsymbol{W}^*) - (\boldsymbol{y}^k, \boldsymbol{W}^k)\Big) + ||(\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k)||^2 \\ \leq \;\; & (1 + ||\mathcal{A}||)||\boldsymbol{X}^k - \boldsymbol{X}^*|| \cdot ||(\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k)||. \end{aligned}$$

Using $||(\boldsymbol{y}^*, \boldsymbol{W}^*) - (\boldsymbol{y}^k, \boldsymbol{W}^k)|| = \sqrt{||\boldsymbol{y}^* - \boldsymbol{y}^k||^2 + ||\boldsymbol{W}^* - \boldsymbol{W}^k||^2} \leq ||\boldsymbol{y}^* - \boldsymbol{y}^k|| + ||\boldsymbol{W}^* - \boldsymbol{W}^k|| \leq ||\boldsymbol{y}^*|| + ||\boldsymbol{y}^k|| + ||\boldsymbol{W}^*|| + ||\boldsymbol{W}^k|| \leq 2(\eta_{\boldsymbol{y}} + \eta_{\boldsymbol{W}})$ and $||(\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k)||^2 \geq 0$, it holds that

$$\frac{\mu}{(\beta_{\boldsymbol{X}}^{\max})^2}||\boldsymbol{X}^k - \boldsymbol{X}^*||^2 - 2(\eta_{\boldsymbol{y}} + \eta_{\boldsymbol{W}})||(\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k)|| \leq (1 + ||\mathcal{A}||)||\boldsymbol{X}^k - \boldsymbol{X}^*|| \cdot ||(\Delta \boldsymbol{y}_{(1)}^k, \Delta \boldsymbol{W}_{(1)}^k)||.$$

This is a quadratic inequality with respect to $||\boldsymbol{X}^k - \boldsymbol{X}^*||$, and solving this quadratic inequality leads us to

$$||\boldsymbol{X}^k - \boldsymbol{X}^*|| \leq u_1(||(\Delta \boldsymbol{y}^k_{(1)}, \Delta \boldsymbol{W}^k_{(1)})||), \tag{26}$$

where $u_1(t) := \frac{1}{2\mu}(1 + ||\mathcal{A}||)(\beta^{\max}_{\boldsymbol{X}})^2 t + \frac{\beta^{\max}_{\boldsymbol{X}}}{2\mu}\sqrt{\left((1 + ||\mathcal{A}||)(\beta^{\max}_{\boldsymbol{X}})\right)^2 t^2 + 8\mu(\eta_{\boldsymbol{y}} + \eta_{\boldsymbol{W}})t}$.

Using (20), (21), (22) and (26), the inequality (19) is now evaluated as

$$|g(\boldsymbol{y}^k, \boldsymbol{W}^k) - g(\boldsymbol{y}^*, \boldsymbol{W}^*)| \leq u_2(||(\Delta \boldsymbol{y}^k_{(1)}, \Delta \boldsymbol{W}^k_{(1)})||)$$

where

$$u_2(t) := \max\{2\rho^{\max}, \eta_{\boldsymbol{X}}\}nt + (\eta_{\boldsymbol{y}}||\mathcal{A}|| + ||\boldsymbol{C}|| + \mu\eta_{\boldsymbol{X}^{-1}} + ||\rho||)u_1(t). \tag{27}$$

Since all the coefficients are positive, the function $u_2(t)$ is continuous for $t \geq 0$, and $u_2(t) > 0$ for $t > 0$. Hence, it follows Lemma 3.13 that

$$\liminf_{k\to\infty} |g(\boldsymbol{y}^k, \boldsymbol{W}^k) - g(\boldsymbol{y}^*, \boldsymbol{W}^*)| = 0.$$

□

Finally, we are ready to show the main result, the convergence of the sequence generated by Algorithm 2.1 to the optimal value.

**Theorem 3.16.** *Algorithm 2.1 with $\epsilon = 0$ stops in a finite number of iterations attaining the optimal value $g^*$, or generate a sequence $\{(\boldsymbol{y}^k, \boldsymbol{W}^k)\}$ such that*

$$\lim_{k\to\infty} |g(\boldsymbol{y}^k, \boldsymbol{W}^k) - g^*| = 0.$$

*Proof.* Suppose, to contrary, that there exists $\bar{\epsilon} > 0$ such that we have an infinite sequence $\{k_1, k_2, \ldots, k_j, \ldots\}$ that satisfies $g_{k_j} < g^* - \bar{\epsilon}$.

We should remark that it holds $k_{j+1} - k_j \leq M$. If $k_{j+1} - k_j > M$, since we can assume that $g_i + \bar{\epsilon} \geq g^*$ for each $i \in [k_j + 1, \ldots, k_{j+1} - 1]$, the inequality (5) indicates $g_{k_{j+1}} \geq \min\{g_{k_{j+1}-1}, \ldots, g_{k_{j+1}-M}\} \geq g^* - \bar{\epsilon}$. Hence, we know $k_{j+1} - k_j \leq M$ and the sequence $\{k_1, k_2, \ldots, k_j, \ldots\}$ should be actually infinite.

Since $u_2(t)$ in (27) is continuous for $t \geq 0$ and $u_2(t) > 0$ for $t > 0$, there exists $\bar{\delta}$ such that $||(\Delta \boldsymbol{y}^{k_j}, \Delta \boldsymbol{W}^{k_j})|| > \bar{\delta}$ for each $j$. We apply the same discussion as Lemma 3.13 to the infinite sequence $\{g_{k_1}, g_{k_2}, \ldots, g_{k_j}, \ldots\}$. If $j$ becomes sufficiently large, we have a contradiction to the upper bound $g_{k_j} \leq g^*$.

□

# 4 Numerical Experiments

We present numerical results obtained from implementing Algorithm 2.1 on the randomly generated synthetic data, deterministic synthetic data and gene expression data in [12] which includes one of most efficient computational results. Our numerical experiments were conducted on larger instances than the test problems in [12] whenever it was possible.

We compare our code DSPG, Algorithm 2.1, with the inexact primal-dual path-following interior-point method (IIPM) [12], the Adaptive Spectral Projected Gradient method (ASPG) [14], and the Adaptive Nesterov's Smooth method (ANS) [14]. For the gene expression data, our results are also compared with the QUadratic approximation for sparse Inverse Covariance

estimation method (QUIC) [9] in Section 4.3. A comparison with the results on the Newton-CG primal proximal-point algorithm (PPA) [18] is not included since its performance was reported to be inferior to the IIPM [12] and it failed to solve some instances.

We note that different stopping criteria are used in each of the aforementioned codes. They obviously affect the number of iterations and consequently the overall computational time. For a fair comparison, we set the threshold values for the IIPM, ASPG, ANS, and QUIC comparable to that of DSPG. More precisely, the stopping criteria of the DSPG was set to

$$||(\Delta \boldsymbol{y}^k_{(1)}, \Delta \boldsymbol{W}^k_{(1)})||_\infty \le \epsilon,$$

where $\epsilon = 10^{-5}$. For the IIPM, we employed

$$\max\left\{\frac{gap}{1+|f(\boldsymbol{X}^k)|+|g(\boldsymbol{y}^k,\boldsymbol{W}^k)|}, pinf, dinf\right\} \le \text{gaptol} := 10^{-6},$$

where *gap*, *pinf*, *dinf* were specified in [12], and for the ASPG and ANS, we used two thresholds $\epsilon_0 := 10^{-3}$ and $\epsilon_c := 10^{-5}$ such that $f(\boldsymbol{X}) \ge f(\boldsymbol{X}^*) - \epsilon_0$ and $\max_{(i,j)\in\Omega} |X_{ij}| \le \epsilon_c$ [12]. The QUIC stops when $\|\partial f(\boldsymbol{X}^k)\|/\text{Tr}(\boldsymbol{\rho}|\boldsymbol{X}^k|) < 10^{-6}$.

The DSPG was experimented with the following parameters: $\gamma = 10^{-4}$, $\tau = 0.5$, $0.1 = \sigma_1 < \sigma_2 = 0.9$, $\alpha_{\min} = 10^{-15} = 1/\alpha_{\max}$, $\alpha_0 = 1$, and $M = 50$. In the DSPG, the mexeig routine of the IIPM was used to reduce the computational time. All numerical experiments were performed on a computer with Intel Xeon X5365 (3.0 GHz) with 48 GB memory using MATLAB.

We set the initial solution as $(\boldsymbol{y}^0, \boldsymbol{W}^0) = (\boldsymbol{0}, \boldsymbol{O})$, which satisfies the assumption (iii) for the instances tested in Sections 4.1 and 4.2. Let $(\boldsymbol{y}^k, \boldsymbol{W}^k)$ be the output of Algorithm 2.1. The recovered primal solution $\boldsymbol{X}^k := \mu(\boldsymbol{C} + \boldsymbol{W}^k - \mathcal{A}^T(\boldsymbol{y}^k))^{-1}$ may not satisfy the equalities $X_{ij} = 0$ for $(i,j) \in \Omega$ in $(\mathcal{P})$ due to numerical errors. In this case, we replace the value of $X_{ij}$ with 0 for $(i,j) \in \Omega$. For the tested instances, this replacement did not affect the semidefiniteness of $\boldsymbol{X}$, since the primal optimal solution was unique (Lemma 3.7) and the nonzero values of $X_{ij}$ were very small.

In the tables in Sections 4.1 and 4.2, the entry corresponding to the DSPG under the column "primal obj." indicates the minimized function value $(\mathcal{P})$ for $\boldsymbol{X}$ after replacing nonzero values of $X_{ij}$ with 0 for $(i,j) \in \Omega$, while "gap" means the maximized function value $(\mathcal{D})$ for $(\boldsymbol{y}, \boldsymbol{W})$ minus the primal one. Therefore, it should have a minus sign. The entries for the IIPM, ASPG, and ANS under "primal obj." column show the difference between the corresponding function values and the primal objective function values of the DSPG. Thus, if this value is positive, it means that the DSPG obtained a lower value for the minimization problem. The tables also show the minimum eigenvalues for the primal variable, number of (outer) iterations, and computational time.

In order to measure the effectiveness of recovering the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$, we adopt the strategy in [12]. The normalized entropy loss ($\text{loss}_E$) and the quadratic loss ($\text{loss}_Q$) are computed as

$$\text{loss}_E := \frac{1}{n}(\text{tr}(\boldsymbol{\Sigma}\boldsymbol{X}) \log\det(\boldsymbol{\Sigma}\boldsymbol{X}) - n), \qquad \text{loss}_Q := \frac{1}{n}\|\boldsymbol{\Sigma}\boldsymbol{X} - \boldsymbol{I}\|,$$

respectively. Notice that the two values should ideally be zero if the regularity term $\boldsymbol{\rho} \bullet |\boldsymbol{X}|$ is disregarded in $(\mathcal{P})$. Also, the sensitivity and the specificity defined as

$$\text{the sensitivity} := \frac{\text{TP}}{\text{TP}+\text{FN}}, \qquad \text{the specificity} := \frac{\text{TN}}{\text{TN}+\text{FP}},$$

are computed, where TP, TN, FP, and FN are the true positives, true negatives, false positive, and false negative, respectively. In our case, the true positives are correct nonzero entries in $\boldsymbol{\Sigma}^{-1}$

and the true negatives are correct zero entries in the same matrix. Therefore, the sensitivity and specificity measure the correct rates of nonzero and of zero entries of $\mathbf{\Sigma}^{-1}$, respectively. The values close to one for both sensitivity and specificity would be desirable. Thus, we set values of $\rho > 0$ such that $\boldsymbol{\rho} = \rho \boldsymbol{E}$ where $\boldsymbol{E}$ is the matrix of all ones in $(\mathcal{P})$ for which the sensitivity and specificity become close to each other, and also $\mu$ equals to one.

### 4.1 Randomly generated synthetic data

As in [12, Section 4.1], we generated the test data by first generating a sparse positive definite matrix $\mathbf{\Sigma}^{-1} \in \mathbb{S}^n$ for a density parameter $\delta > 0$, and then computing a sample covariance matrix $\boldsymbol{C} \in \mathbb{S}^n$ from $2n$ i.i.d. random vectors selected from the $n$-dimensional Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$.

Our experiments were carried out on different sizes $n$ of matrix $\mathbf{\Sigma}^{-1}$, two choices of density parameters $\delta = 0.1$ and 0.9, and problem $(\mathcal{P})$ without the linear constraints $\mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}$ and with linear constraints $X_{ij} = 0$ for $(i, j) \in \Omega$, where $\Omega$ specifies the zero elements of $\mathbf{\Sigma}^{-1}$.

Table 1: Comparative numerical results for the DSPG, IIPM, ASPG and ANS on unconstrained randomly generated synthetic data. $n$ =1000, 3000, and 5000, density $\delta = 0.1$ and 0.9.

| $n$ | $\rho$ | method | primal obj. | iter. | time (s) | $\delta = 0.1$ | |
|---|---|---|---|---|---|---|---|
| | | DSPG | −648.85805752 | 89 | 42.7 | $\lambda_{\min}(\boldsymbol{X})$ | 7.64e−02 |
| | | (gap) | −0.00006098 | | | $\text{loss}_E$ | 1.8e−01 |
| 1000 | 5/1000 | IIPM | +0.00000385 | 15 | 78.0 | $\text{loss}_Q$ | 2.2e−02 |
| | = 0.005 | ASPG | +0.00046235 | 77 | 49.5 | sensitivity | 0.90 |
| | | ANS | +0.00093895 | 310 | 172.4 | specificity | 0.88 |
| | | DSPG | −4440.85648991 | 62 | 657.2 | $\lambda_{\min}(\boldsymbol{X})$ | 2.42e−01 |
| | | (gap) | −0.00009711 | | | $\text{loss}_E$ | 1.3e−02 |
| 3000 | 4/3000 | IIPM | +0.00015710 | 15 | 1219.9 | $\text{loss}_Q$ | 2.0e−01 |
| | = 0.001333 | ASPG | +0.00082640 | 49 | 801.9 | sensitivity | 0.82 |
| | | ANS | +0.00089732 | 269 | 3255.9 | specificity | 0.85 |
| | | DSPG | −9576.24150224 | 57 | 3015.4 | $\lambda_{\min}(\boldsymbol{X})$ | 1.0e−02 |
| | | (gap) | −0.00015026 | | | $\text{loss}_E$ | 1.9e−01 |
| 5000 | 3/5000 | IIPM | +0.00039297 | 15 | 4730.0 | $\text{loss}_Q$ | 1.0e−02 |
| | = 0.0006 | ASPG | +0.00012477 | 52 | 4137.0 | sensitivity | 0.82 |
| | | ANS | +0.00084603 | 248 | 14929.4 | specificity | 0.81 |
| $n$ | $\rho$ | method | primal obj. | iter. | time (s) | $\delta = 0.9$ | |
| | | DSPG | −3584.93243464 | 33 | 16.5 | $\lambda_{\min}(\boldsymbol{X})$ | 3.30e+01 |
| | | (gap) | −0.00000122 | | | $\text{loss}_E$ | 9.4e−02 |
| 1000 | 0.15/1000 | IIPM | +0.00031897 | 15 | 56.1 | $\text{loss}_Q$ | 1.5e−02 |
| | = 0.00015 | ASPG | +0.00070753 | 21 | 18.0 | sensitivity | 0.50 |
| | | ANS | +0.00094435 | 78 | 49.2 | specificity | 0.53 |
| | | DSPG | −13012.61749049 | 26 | 278.9 | $\lambda_{\min}(\boldsymbol{X})$ | 7.56e+01 |
| | | (gap) | −0.00000818 | | | $\text{loss}_E$ | 8.3e−02 |
| 3000 | 0.125/3000 | IIPM | +0.00125846 | 18 | 1133.4 | $\text{loss}_Q$ | 8.1e−03 |
| | = 0.0000417 | ASPG | +0.00049848 | 21 | 474.8 | sensitivity | 0.49 |
| | | ANS | +0.00097430 | 81 | 1135.8 | specificity | 0.54 |
| | | DSPG | −23487.45518427 | 26 | 1381.3 | $\lambda_{\min}(\boldsymbol{X})$ | 1.07e+02 |
| | | (gap) | −0.00000534 | | | $\text{loss}_E$ | 9.0e−02 |
| 5000 | 0.1/5000 | IIPM | +0.00068521 | 23 | 5928.7 | $\text{loss}_Q$ | 6.5e−03 |
| | = 0.00002 | ASPG | +0.00044082 | 21 | 2405.7 | sensitivity | 0.53 |
| | | ANS | +0.00097990 | 90 | 6150.1 | specificity | 0.49 |

Table 1 shows the results for problems without any linear constraints in $(\mathcal{P})$. Clearly, the

DSPG requires less time to compute a lower objective value than the other codes. The advantage of the DSPG is greater for the denser problems ($\delta = 0.9$, which is the case not considered in [12]) or larger problems ($n = 5000$). Moreover, the dense problems tend to be easier to compute in terms of computational time, although their recovery can be slightly worse than the problems with $\delta = 0.1$, as indicated by the values of the sensitivity and specificity. For denser instances, $\text{loss}_E$ and $\text{loss}_Q$ are improved.

Table 2: Comparative numerical results for the DSPG, IIPM, ASPG and ANS on constrained randomly generated synthetic data. $n$ =1000, 3000, and 5000, density $\delta = 0.1$ and 0.9.

| $n$ | $\rho$/# constraints | method | primal obj. | iter. | time (s) | $\delta = 0.1$ | |
|---|---|---|---|---|---|---|---|
| | 5/1000 | DSPG | −631.25522377 | 144 | 76.1 | $\lambda_{\min}(\boldsymbol{X})$ | 7.70$e$−02 |
| | =0.005 | (gap) | −0.00013566 | | | $\text{loss}_E$ | 1.7$e$−01 |
| 1000 | | IIPM | −0.00013004 | 16 | 103.1 | $\text{loss}_Q$ | 2.1$e$−02 |
| | 221,990 | ASPG | +0.00074651 | 1025 | 635.8 | sensitivity | 0.93 |
| | | ANS | +0.00076506 | 5464 | 3027.2 | specificity | 0.92 |
| | 3/3000 | DSPG | −4582.28297352 | 126 | 1383.8 | $\lambda_{\min}(\boldsymbol{X})$ | 2.41$e$−01 |
| | =0.001 | (gap) | −0.00006496 | | | $\text{loss}_E$ | 1.6$e$−01 |
| 3000 | | IIPM | −0.00004689 | 17 | 1692.4 | $\text{loss}_Q$ | 1.2$e$−02 |
| | 1,898,796 | ASPG | +0.00062951 | 755 | 9658.4 | sensitivity | 0.92 |
| | | ANS | +0.00083835 | 5863 | 67170.0 | specificity | 0.88 |
| | 3/5000 | DSPG | −9489.67203718 | 96 | 5180.8 | $\lambda_{\min}(\boldsymbol{X})$ | 4.85$e$−01 |
| | = 0.0006 | (gap) | −0.00005274 | | | $\text{loss}_E$ | 1.8$e$−01 |
| 5000 | | IIPM | +0.00001554 | 16 | 6359.0 | $\text{loss}_Q$ | 9.7$e$−03 |
| | 5,105,915 | ASPG | +0.00074531 | 704 | 43955.2 | sensitivity | 0.85 |
| | | ANS | +0.00085980 | 5056 | 286746.6 | specificity | 0.89 |
| $n$ | $\rho$/# constraints | method | primal obj. | iter. | time (s) | $\delta = 0.9$ | |
| | 0.1/1000 | DSPG | −3625.96768067 | 42 | 20.7 | $\lambda_{\min}(\boldsymbol{X})$ | 3.08$e$+01 |
| | = 0.0001 | (gap) | −0.00000072 | | | $\text{loss}_E$ | 1.3$e$−01 |
| 1000 | | IIPM | +0.00014852 | 17 | 65.0 | $\text{loss}_Q$ | 1.9$e$−02 |
| | 32,565 | ASPG | +0.00079319 | 376 | 547.9 | sensitivity | 0.64 |
| | | ANS | +0.00098958 | 1938 | 1102.3 | specificity | 0.69 |
| | 0.07/3000 | DSPG | −13178.75746518 | 35 | 372.6 | $\lambda_{\min}(\boldsymbol{X})$ | 6.84$e$+01 |
| | = 0.0000233 | (gap) | −0.00000049 | | | $\text{loss}_E$ | 1.4$e$−01 |
| 3000 | | IIPM | +0.00089508 | 24 | 1528.1 | $\text{loss}_Q$ | 1.1$e$−02 |
| | 238,977 | ASPG | +0.00030434 | 451 | 15295.4 | sensitivity | 0.67 |
| | | ANS | +0.00099513 | 3309 | 38990.8 | specificity | 0.67 |
| | 0.07/5000 | DSPG | −23644.31706813 | 29 | 1543.3 | $\lambda_{\min}(\boldsymbol{X})$ | 1.01$e$+02 |
| | = 0.000014 | (gap) | −0.00000833 | | | $\text{loss}_E$ | 1.2$e$−01 |
| 5000 | | IIPM | +0.00101943 | 28 | 7247.3 | $\text{loss}_Q$ | 7.9$e$−03 |
| | 604,592 | ASPG | +0.00034229 | 344 | 30880.9 | sensitivity | 0.64-0.65 |
| | | ANS | +0.00098642 | 3272 | 188957.0 | specificity | 0.68-0.69 |

For the problems tested in Table 2, the sparsity of $\mathbf{\Sigma}^{-1} \in \mathbb{S}^n$ is imposed as linear constraints in $(\mathcal{P})$ as $X_{ij} = 0$ for $(i,j) \in \Omega$, where $|\Omega| \equiv$"# constraints" in the table. From the results in Table 2, we observe that the ASPG and ANS require much more computational time than in the unconstrained case. The IIPM is the only code which violates the linear constraints $X_{ij} = 0$ for $(i,j) \in \Omega$, resulting in values less than $6.01 \times 10^{-9}$ for $\max_{i,j=1,\ldots,n} |X_{ij}|$ at the final iteration. We also see that $\text{loss}_E$ and $\text{loss}_Q$ do not change when density $\delta$ is changed.

### 4.2 Deterministic synthetic data

The numerical results on eight problems where $\boldsymbol{A} \in \mathbb{S}^n$ has a special structure such as diagonal band, fully dense, or arrow-shaped [12] are shown in Tables 3 and 4. For each $\boldsymbol{A}$, a sample covariance matrix $\boldsymbol{C} \in \mathbb{S}^n$ is computed from $2n$ i.i.d. random vectors selected from the $n$-dimensional Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^{-1})$. Finally, we randomly select 50% of the zero entries for each $\boldsymbol{A}$ to be the linear constraints in $(\mathcal{P})$, excepting for the Full problem in Table 3.

Similar observation to Section 4.1 can be made for the results presented in Tables 3 and 4. The DSPG took less computational time than the other methods in most cases and obtained slightly worse objective function values.

Table 3: Comparative numerical results for the DSPG, IIPM, ASPG and ANS on unconstrained deterministic synthetic data. $n$ =2000.

| problem | $\rho$ | method | primal obj. | iter. | time (s) | | |
|---|---|---|---|---|---|---|---|
| Full | 0.1 | DSPG | 2189.07471338 | 20 | 57.8 | $\lambda_{\min}(\boldsymbol{X})$ | $8.42e{-}01$ |
| | | (gap) | $-0.33302912$ | | | $\text{loss}_E$ | $7.9e{-}03$ |
| | | IIPM | $-0.33297893$ | 11 | 185.9 | $\text{loss}_Q$ | $2.1e{-}03$ |
| | | ASPG | $-0.33297903$ | 54 | 244.1 | | |
| | | ANS | $-0.33283013$ | 40 | 150.5 | | |

### 4.3 Gene expression data

Five problems from the gene expression data [12] were tested for performance comparison. Since it was assumed that the conditional independence of their gene expressions is not known, linear constraints were not imposed in $(\mathcal{P})$. In this experiment, we additionally compared the performance of the DSPG with QUIC [9] which is known to be fast for sparse problems.

Figures 1-3 show the computational time (left axis) for each problem when $\rho$ is changed. As $\rho$ grows larger, the final solution $\boldsymbol{X}^k$ (of the DSPG) becomes sparser, as shown in the right axis for the number of nonzero elements of $\boldsymbol{X}^k$.

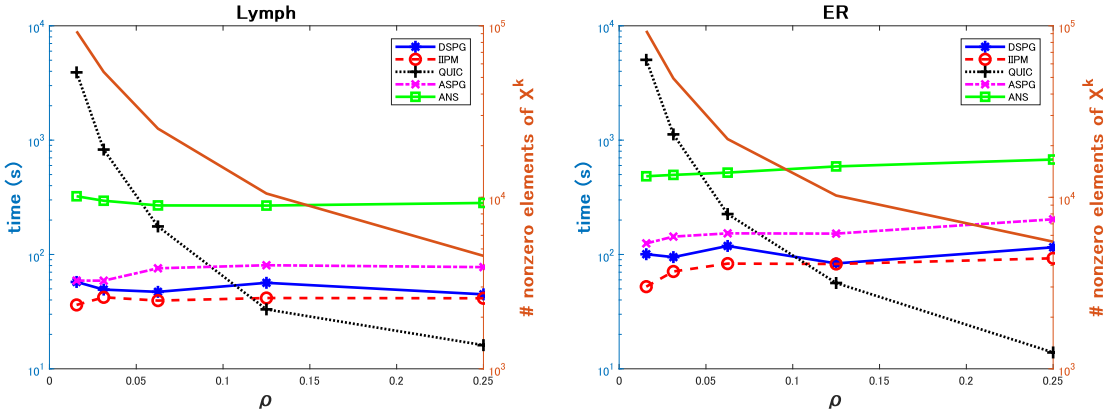


Figure 1: Computational time (the left axis) for the DSPG, IIPM, QUIC, ASPG, ANS on the problems "Lymph" ($n = 587$) and "ER" ($n = 692$) when $\rho$ is changed; the number of nonzero elements of $\boldsymbol{X}^k$ for the final iterate of the DSPG (the right axis).

Table 4: Comparative numerical results for the DSPG, IIPM, ASPG and ANS on constrained deterministic synthetic data. $n$ =2000.

| problem | $\rho$/# constraints | method | primal obj. | iter. | time (s) | | |
|---|---|---|---|---|---|---|---|
| ar1 | 0.1 | DSPG | 3707.57716442 | 2001 | 6060.5 | $\lambda_{\min}(\boldsymbol{X})$ | 1.00-1.25$e$−06 |
| | | (gap) | −0.32561268 | | | $\text{loss}_E$ | 3.1$e$−02 |
| | | IIPM | −0.32526710 | 38 | 3577.3 | $\text{loss}_Q$ | 2.3$e$−01 |
| | 998,501 | ASPG | −0.32474270 | 19034 | 69534.6 | sensitivity | 1.00 |
| | | ANS | −0.32448637 | 29347 | 88733.8 | specificity | 1.00 |
| ar2 | 0.1 | DSPG | 3029.94934978 | 55 | 167.6 | $\lambda_{\min}(\boldsymbol{X})$ | 2.73$e$−01 |
| | | (gap) | −0.00329417 | | | $\text{loss}_E$ | 4.4$e$−02 |
| | | IIPM | −0.00291044 | 11 | 290.1 | $\text{loss}_Q$ | 5.8$e$−03 |
| | 997,502 | ASPG | −0.00309541 | 196 | 821.2 | sensitivity | 1.00 |
| | | ANS | −0.00241116 | 1241 | 4230.7 | specificity | 1.00 |
| ar3 | 0.03 | DSPG | 2552.71613399 | 78 | 236.8 | $\lambda_{\min}(\boldsymbol{X})$ | 1.70$e$−01 |
| | | (gap) | −0.00553547 | | | $\text{loss}_E$ | 1.8$e$−02 |
| | | IIPM | −0.00545466 | 14 | 433.2 | $\text{loss}_Q$ | 4.4$e$−03 |
| | 996,503 | ASPG | −0.00480321 | 353 | 1242.4 | sensitivity | 1.00 |
| | | ANS | −0.00468946 | 2712 | 8592.6 | specificity | 1.00 |
| ar4 | 0.01 | DSPG | 2340.10866746 | 73 | 222.7 | $\lambda_{\min}(\boldsymbol{X})$ | 2.31$e$−01 |
| | | (gap) | −0.00050381 | | | $\text{loss}_E$ | 5.6$e$−02 |
| | | IIPM | −0.00048223 | 14 | 403.3 | $\text{loss}_Q$ | 8.4$e$−03 |
| | 995,505 | ASPG | +0.00030934 | 1095 | 3975.8 | sensitivity | 1.00 |
| | | ANS | +0.00044155 | 5996 | 19379.6 | specificity | 1.00 |
| Decay | 0.1 | DSPG | 2253.67375651 | 14 | 44.4 | $\lambda_{\min}(\boldsymbol{X})$ | 7.70$e$−01 |
| | | (gap) | −0.00114736 | | | $\text{loss}_E$ | 1.5$e$−02 |
| | | IIPM | −0.00094913 | 10 | 170.5 | $\text{loss}_Q$ | 3.6$e$−03 |
| | 981,586 | ASPG | −0.00106549 | 12 | 69.9 | sensitivity | 0.00 |
| | | ANS | −0.00089883 | 32 | 126.6 | specificity | 1.00 |
| Star | 0.1 | DSPG | 2204.50539735 | 82 | 248.7 | $\lambda_{\min}(\boldsymbol{X})$ | 2.50-2.51$e$−07 |
| | | (gap) | −0.00018704 | | | $\text{loss}_E$ | 4.8$e$−03 |
| | | IIPM | −0.00001083 | 11 | 179.6 | $\text{loss}_Q$ | 4.5$e$−01 |
| | 997,501 | ASPG | −0.00002462 | 31 | 159.0 | sensitivity | 0.33 |
| | | ANS | −0.00017677 | 92 | 311.4 | specificity | 1.00 |
| Circle | 0.05 | DSPG | 3519.14112855 | 1094 | 3307.0 | $\lambda_{\min}(\boldsymbol{X})$ | 1.24-1.61$e$−06 |
| | | (gap) | −0.07034481 | | | $\text{loss}_E$ | 2.9$e$−02 |
| | | IIPM | −0.07032168 | 28 | 1976.8 | $\text{loss}_Q$ | 2.6$e$−01 |
| | 998,500 | ASPG | −0.06948986 | 11557 | 42437.1 | sensitivity | 1.00 |
| | | ANS | −0.06946870 | 19714 | 59672.3 | specificity | 1.00 |

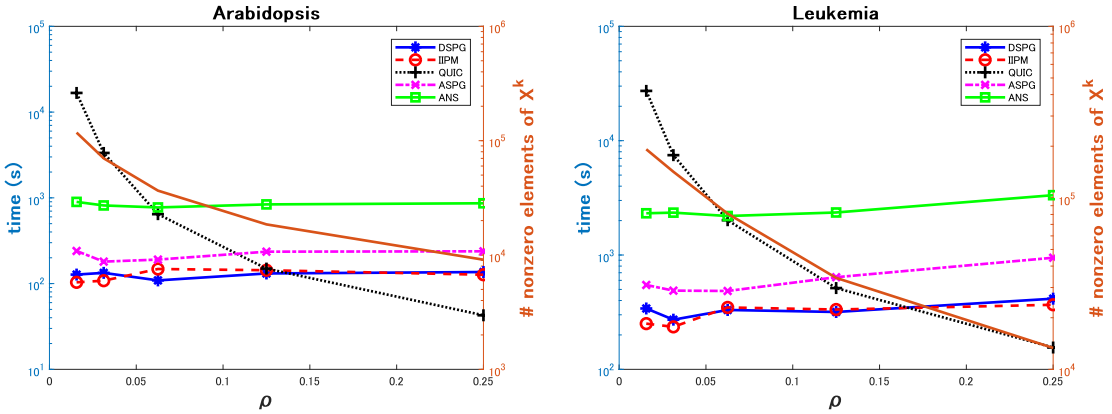


Figure 2: Computational time (the left axis) for the DSPG, IIPM, QUIC, ASPG, ANS on problems "Arabidopsis" ($n = 834$) and "Leukemia" ($n = 1255$) when $\rho$ is changed; the number of nonzero elements of $\boldsymbol{X}^k$ for the final iteration of the DSPG (the right axis).

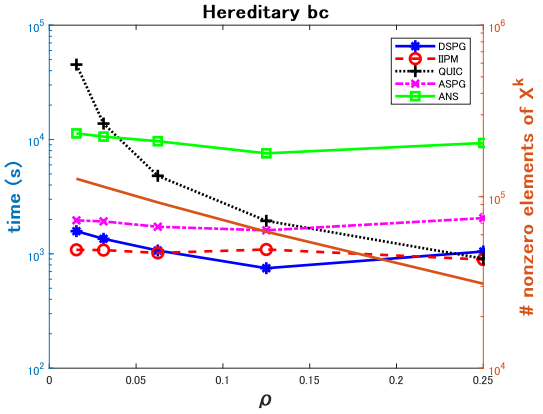


Figure 3: Computational time (the left axis) for the DSPG, IIPM, QUIC, ASPG, ANS on the problem "Hereditary bc" ($n = 1869$) when $\rho$ is changed; the number of nonzero elements of $\boldsymbol{X}^k$ for the final iteration of the DSPG (the right axis).

We see that the DSPG (solid blue line) is as competitive with the IIPM (dashed red line) and even faster than the QUIC (dotted black line), which is known for their fast convergence, when $\rho$ is small. The performance of the QUIC is closely related to the sparsity of the final iterate of $\boldsymbol{X}^k$ for the DSPG (solid brown line) as expected. Here we used the threshold $|\boldsymbol{X}^k|_{ij} \geq 0.05$ to determine nonzero elements.

# 5 Conclusion

We have proposed a dual-type spectral projected gradient method for $(\mathcal{P})$ to efficiently handle large-scale problems. Based on the theoretical convergence results of the proposed method, the Dual SPG algorithm has been implemented and the numerical results on randomly generated synthetic data, deterministic synthetic data and gene expression data are reported. We have demonstrated the efficiency in computational time to obtain a better optimal value for $(\mathcal{P})$. In particular, when $\rho$ is small, we have observed that the performance of the proposed method increases.

To further improve the performance of the Dual SPG method, our future research includes reducing the computational time by employing an approach similar to Dahl *et al.* [21] and/or exploiting the structured sparsity as discussed in [10].

# References


[1] J. Barzilai and J. M. Borwein. Two-point step size gradient methods. *IMA J. Numer. Anal.*, 8(1):141–148, 1988.

[2] E. G. Birgin, J. M. Martinez, and M. Raydan. Nonmonotone spectral projected gradient methods on convex sets. *SIAM J. Optim.*, 10(4):1196–1211, 2000.

[3] J. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples (2nd Edition)*. Springer, New York, 2006.

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[5] A. d'Aspremont, O. Banerjee, and L. E. Ghaoui. First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. and Appl.*, 30(1):56–66, 2008.

[6] A. P. Dempster. Covariance selection. *Biometrics*, 28(157-175), 1972.

[7] J. C. Duchi, S. Gould, and D. Koller. Projected subgradient methods for learning sparse gaussians. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, 2008.

[8] W. Hager and H. Zhang. A new active set algorithm for box constrained optimization. *SIAM J. Optim.*, 17(2):526–557, 2006.

[9] C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, and P. Ravikumar. Quic: Quadratic approximation for sparse inverse covariance estimation. *J. Mach. Learn. Res.*, 15:2911–2947, 2014.

[10] S. Kim, M. Kojima, M. Mevissen, and M. Yamashita. Exploiting sparsity in linear and nonlinear matrix inequalities via positive semidefinite matrix completion. *Math. Program. Series B*, 129(1):33–68, 2011.

[11] S. L. Lauritzen. *Graphical Models*. The Clarendon Press/Oxford University Press, Oxford, 1996.

[12] L. Li and K.-C. Toh. An inexact interior point method for $l_1$-regularized sparse covariance selection. *Math. Prog. Comp.*, 2(3–4):291–315, 2010.

[13] P. Li and Y. Xiao. An efficient algorithm for sparse inverse covariance matrix estimation based on dual formulation. *Comput. Stat. Data Anal.*, 128:292–307, 2018.

[14] Z. Lu. Adaptive first-order methods for general sparse inverse covariance selection. *SIAM J. Matrix Anal. Appl.*, 31(4):2000–2016, 2010.

[15] R. Tavakoli and H. Zhang. A nonmonotone spectral projected gradient method for large-scale topology optimization problems. *Numer. Algebr. Control Optim.*, 2(2):395–412, 2012.

[16] G. Ueno and T. Tsuchiya. Covariance regularization in inverse space. *Q. J. R. Meteorol. Soc.*, 135:1133–1156, 2009.

[17] C. Wang. On how to solve large-scale log-determinant optimization problems. *Comput. Optim. Appl.*, 64:489–511, 2016.

[18] C. Wang, D. Sun, and K.-C. Toh. Solving log-determinant optimization problems by a Newton-CG primal proximal point algorithm. *SIAM J. Optim.*, 20(6):2994–3013, 2010.

[19] J. Yang, D. Sun, and K.-C. Toh. A proximal point algorithm for log-determinant optimization with group lasso regularization. *SIAM J. Optim.*, 23(2):857–893, 2013.

[20] X. Yuan. Alternating direction method for sparse covariance models. *J. Sci. Comput.*, 51:261–273, 2012.

[21] R. Y. Zhang, S. Fattahi, and S. Sojoudi. Large-scale sparse inverse covariance estimation via thresholding and max-det matrix completion. *Proc. Mach. Learn. Res*, 80:5766–5775, 2018.