Dual Spectral Projected Gradient Method for Generalized Log-det Semidefinite Programming

Charles Namchaisiri[*] Makoto Yamashita[†]

September 27, 2024

Abstract

Log-det semidefinite programming (SDP) problems are optimization problems that often arise from Gaussian graphic models. A log-det SDP problem with an ℓ_1-norm term has been examined in many methods, and the dual spectral projected gradient (DSPG) method by Nakagaki et al. in 2020 is designed to efficiently solve the dual problem of the log-det SDP by combining a non-monotone line-search projected gradient method with the step adjustment for positive definiteness. In this paper, we extend the DSPG method for solving a generalized log-det SDP problem involving additional terms to cover more structures in Gaussian graphical models in a unified style. We establish the convergence of the proposed method to the optimal value. We conduct numerical experiments to illustrate the efficiency of the proposed method.

1 Introduction

In this paper, we address the following log-determinant semidefinite programming (SDP) optimization problem:

$$\begin{aligned} \min_{\boldsymbol{X} \in \mathbb{S}^n} \quad & f(\boldsymbol{X}) := \boldsymbol{C} \bullet \boldsymbol{X} - \mu \log\det \boldsymbol{X} + \sum_{h=1}^{H} \lambda_h \|\mathcal{Q}_h(\boldsymbol{X})\|_{p_h} \\ \text{s.t.} \quad & \mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}, \boldsymbol{X} \succ \boldsymbol{O}. \end{aligned} \tag{$\mathcal{P}$}$$

We use $\mathbb{R}^n$ and $\mathbb{S}^n$ to denote the sets of n-dimensional vectors and $n \times n$ symmetric matrices, respectively. The inner product between $\boldsymbol{C} \in \mathbb{S}^n$ and $\boldsymbol{X} \in \mathbb{S}^n$ is defined by $\boldsymbol{C} \bullet \boldsymbol{X} := \sum_{i=1}^n \sum_{j=1}^n C_{ij} X_{ij}$. We use nonnegative μ and $\lambda_1, \ldots, \lambda_H$ as weight parameters in the objective function, $\mathcal{Q}_h : \mathbb{S}^n \to \mathbb{R}^{n_h}$ for each $h = 1, \ldots, H$ is a linear map, $\|\boldsymbol{y}\|_{p_h} := (\sum_{i=1}^{n_h} y_i^{p_h})^{\frac{1}{p_h}}$ is the ℓ_{p_h}-norm of $\boldsymbol{y} \in \mathbb{R}^{n_h}$ with $p_h \geq 1$. In the constraints, $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$ is a linear map defined by $\mathcal{A}(\boldsymbol{X}) = (\boldsymbol{A}_1 \bullet \boldsymbol{X}, \boldsymbol{A}_2 \bullet \boldsymbol{X}, \ldots, \boldsymbol{A}_m \bullet \boldsymbol{X})^\top$ with $\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_m \in \mathbb{S}^n$, and the vector $\boldsymbol{b} \in \mathbb{R}^m$ in $(\mathcal{P})$ is a given vector. The symbol $\boldsymbol{X} \succ \boldsymbol{O}$ ($\boldsymbol{X} \succeq \boldsymbol{O}$) for a matrix $\boldsymbol{X} \in \mathbb{S}^n$ denotes that $\boldsymbol{X}$ is a positive definite (positive semidefinite, respectively) matrix.

[*] School of Computing, Tokyo Institute of Technology, Japan. (namchaisiri.c.aa@m.titech.ac.jp)

[†] School of Computing, Tokyo Institute of Technology, Japan. (Makoto.Yamashita@c.titech.ac.jp)

Many log-determinant SDP problems can be rewritten as the form of $(\mathcal{P})$. A model $(\mathcal{P})$ with $H = 1$ and $\mathcal{Q}_1(\boldsymbol{X})$ being the vector of elements of $\boldsymbol{X}$ is equivalent to the following problem:

$$\begin{aligned} \min_{\boldsymbol{X} \in \mathbb{S}^n} \quad & f(\boldsymbol{X}) := \boldsymbol{C} \bullet \boldsymbol{X} - \mu \log\det \boldsymbol{X} + \lambda \sum_{i=1}^{n} \sum_{j=1}^{n} |X_{ij}| \\ \text{s.t.} \quad & \mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}, \boldsymbol{X} \succ \boldsymbol{O}. \end{aligned} \tag{1.1}$$

Furthermore, when $\lambda = 0$ and the linear constraint is $X_{ij} = 0$ for $(i,j) \in \Omega \subseteq \{(i,j) \mid 1 \le i < j \le n\}$, (1.1) turns into the Gaussian graphical models [10], which corresponds to the graphical interpretation of sparse covariance selection model [5].

The model $(\mathcal{P})$ also covers the following problem in Lin et al. [12] that estimates sparse Gaussian graphical models with hidden clustering structures. The fourth term in the objective function induces a clustering structure of the concentration matrix.

$$\begin{aligned} \min_{\boldsymbol{X} \in \mathbb{S}^n} \quad & f(\boldsymbol{X}) := \boldsymbol{C} \bullet \boldsymbol{X} - \mu \log\det \boldsymbol{X} + \rho \sum_{i<j} |X_{ij}| + \lambda \sum_{i<j} \sum_{s<t} |X_{ij} - X_{st}| \\ \text{s.t.} \quad & \mathcal{A}(\boldsymbol{X}) = \boldsymbol{b},\ \boldsymbol{X} \succ \boldsymbol{O}, \end{aligned}$$

Another important model in $(\mathcal{P})$ is the block ℓ_∞-regularized log-likelihood minimization problem in Duchi et al. [6] to estimate sparsity between entire blocks of variables:

$$\min_{\boldsymbol{X} \in \mathbb{S}^n} \quad \boldsymbol{f}(\boldsymbol{X}) := \boldsymbol{C} \bullet \boldsymbol{X} - \log\det \boldsymbol{X} + \sum_{k=1}^{K} \lambda_k \max\{|X_{ij}| \big| (i,j) \in G_k\}, \tag{1.2}$$

where entries in $\boldsymbol{X}$ are divided into disjoint subsets $G_1, G_2, \ldots, G_K (K < n^2)$. The last term in (1.2) is the group ℓ_∞-regularized covariance selection used to enforce the sparsity between blocks. Other than this regularization, there is also the group ℓ_1 and ℓ_2-regularized regression, see *,e.g.,* [1, 18]. Extensions of model (1.2) have also been examined. Honorio et al. [8] proposed the multi-task structure learning problem for Gaussian graphical models that consider promoting a consistent sparseness pattern across arbitrary tasks by using the regularizer to penalize corresponding edges across the task. Yang et al. [17] discussed a model that replaces the last term in the objective function in (1.2) with ℓ_p-norm for $p \in \{1, 2, \infty\}$.

When the regularizer is restricted to the ℓ_1-norm (1.1), many methods have been proposed. Wang et al. [16] proposed a Newton-CG primal proximal point algorithm, and Li et al. [11] modified an inexact primal-dual path-following interior-point algorithm to solve the log-det SDP with a large number of linear constraints. Hsie et al. [9] proposed a quadratic approximation for sparse inverse covariance estimation (QUIC) based on the Newton method and a quadratic approximation. Wang et al. [15] generated an initial point of the algorithm by using a proximal augmented Lagrangian method and then computed the accurate solution by applying a Newton-CG augmented Lagrangian method.

Nakagaki et al. [13] proposed a dual spectral projected gradient method (DSPG) for solving the dual problem of (1.1). The method is an iterative method that uses an inexact projection to avoid the difficulty of computing the orthogonal projection to the intersection of two convex sets, while still having the advantages of the spectral projected gradient (SPG) method [4]. In particular, an important advantage here is that it requires only the function values and the first-order derivatives, making it faster than other second-order derivatives methods. However, their convergence analysis heavily depended on the ℓ_1 norm in the objective function of (1.1). In $(\mathcal{P})$ that we address in this paper, we do not assume a specific structure of the linear map $\mathcal{Q}_h$ and also the number of the terms H that corresponds to the number of variables in the dual problem, therefore, we cannot simply apply [13] to the generalized problem $(\mathcal{P})$.

In this paper, we extend the DSPG method to deal with the general form of log-determinant optimization problems $(\mathcal{P})$. To enhance the efficiency, we apply a similar reformulation technique as in [14]. We embed the ℓ_p-norm structure in the objective function in $(\mathcal{P})$ into constraints so that the objective is differentiable, and we combine the projection onto the constraints related to the ℓ_p-norm.

In this paper, our main contributions are as follows.

- We propose the generalized model $(\mathcal{P})$, which covers many log-det models.
- We develop a numerical method (Algorithm 1) for solving $(\mathcal{P})$, and present the convergence analysis of the algorithm.
- We show the efficiency of the proposed method with numerical experiments (Section 4). For $(\mathcal{P})$ with the matrix dimension $n = 2000$, the method in Duchi et al. [6] demanded 163.48 seconds while the proposed method consumed only 65.88 seconds to attain the same solution accuracy.

The remainder of this paper is organized as follows. We describe the structure of the dual problem and the proposed DSPG-based method (Algorithm 1) in Section 2 and discuss the convergence analysis in Section 3 focusing on the generalized part of $(\mathcal{P})$. In Section 4, we present the result of numerical experiments on the log-likelihood minimization problem, block constraint problem, and multi-task structure. Finally, we conclude in Section 5.

1.1 Notation and symbols

Let $\|\boldsymbol{X}\| := \sqrt{\boldsymbol{X} \bullet \boldsymbol{X}}$ denote the Frobenius norm for a matrix $\boldsymbol{X} \in \mathbb{S}^n$. We also use $\|\boldsymbol{y}\|$ to represent the Euclidean norm of the vector $\boldsymbol{y} \in \mathbb{R}^n$, that is, $\|\boldsymbol{y}\| := \|\boldsymbol{y}\|_2$.

Given a linear map $\mathcal{A}$. We denote the adjoint operator of $\mathcal{A}$ as $\mathcal{A}^\top$. We define the operator norm of $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$ and $\mathcal{A}^\top : \mathbb{R}^m \to \mathbb{S}^n$ as $\|\mathcal{A}\| := \sup_{\boldsymbol{X} \neq \boldsymbol{O}} \left\{ \frac{\|\mathcal{A}(\boldsymbol{X})\|}{\|\boldsymbol{X}\|} \right\}$ and $\|\mathcal{A}^\top\| := \sup_{\boldsymbol{y} \neq \boldsymbol{0}} \left\{ \frac{\|\mathcal{A}^\top(\boldsymbol{y})\|}{\|\boldsymbol{y}\|} \right\}$, respectively.

Let $\mathcal{U}$ be the direct product space $\mathbb{R}^m \times \mathbb{S}^n \times \cdots \times \mathbb{S}^n$. We define the inner product of $\boldsymbol{U}_1 = (\boldsymbol{y}_1, \boldsymbol{S}_1^1, \dots, \boldsymbol{S}_H^1) \in \mathcal{U}$ and $\boldsymbol{U}_2 = (\boldsymbol{y}_2, \boldsymbol{S}_1^2, \dots, \boldsymbol{S}_H^2) \in \mathcal{U}$ by $\langle \boldsymbol{U}_1, \boldsymbol{U}_2 \rangle := \boldsymbol{y}_1^\top \boldsymbol{y}_2 + \boldsymbol{S}_1^1 \bullet \boldsymbol{S}_1^2 + \cdots + \boldsymbol{S}_H^1 \bullet \boldsymbol{S}_H^2$. We also define the norm of $\boldsymbol{U} \in \mathcal{U}$ as $||\boldsymbol{U}|| := \sqrt{\langle \boldsymbol{U}, \boldsymbol{U} \rangle}$.

For $p \geq 1$ and $\lambda > 0$, we define the ℓ_p-ball with radius λ as

$$\mathcal{B}_p^\lambda := \{\boldsymbol{x} \mid ||\boldsymbol{x}||_p \leq \lambda\}.$$

We use $P_{\mathcal{S}}(\cdot)$ to denote the projection onto the convex set $\mathcal{S}$, i.e.,

$$P_{\mathcal{S}}(\cdot) := \arg \min_{\boldsymbol{X} \in \Omega} \|\boldsymbol{X} - \cdot\|.$$

2 The proposed method

Let $\mathcal{Q}_h^\top : \mathbb{R}^{n_h} \to \mathbb{R}^m$ be the adjoint operators of $\mathcal{Q}_h$. To extend the DSPG method developed in [13] for the dual problem of (1.1), we need the dual problem of our problem $(\mathcal{P})$:

$$\begin{aligned} \max_{\boldsymbol{y}\in\mathbb{R}^m,\,\boldsymbol{z}_h\in\mathbb{R}^{n_h}(h=1,\dots,H)} \quad & \boldsymbol{b}^\top\boldsymbol{y} + \mu\log\det\left(\boldsymbol{C} - \mathcal{A}^\top(\boldsymbol{y}) + \sum_{h=1}^{H}\mathcal{Q}_h^\top(\boldsymbol{z}_h)\right) + n\mu - n\mu\log\mu \\ \text{s.t.} \quad & \|\boldsymbol{z}_h\|_{p_h^*} \le \lambda_h \ (h=1,\dots,H),\ \boldsymbol{C} - \mathcal{A}^\top(\boldsymbol{y}) + \sum_{h=1}^{H}\mathcal{Q}_h^\top(\boldsymbol{z}_h) \succ \boldsymbol{O}, \end{aligned} \tag{2.1}$$

where $\|\cdot\|_{p_h^*}$ is the dual norm of $\|\cdot\|_{p_h}$ such that $\frac{1}{p_h} + \frac{1}{p_h^*} = 1$. More precisely, p_h^* for $p_h \in [1,\infty]$ is given by

$$p_h^* = \begin{cases} \infty & \text{if } p_h = 1 \\ p_h/(p_h-1) & \text{if } 1 < p_h < \infty \\ 1 & \text{if } p_h = \infty. \end{cases}$$

To increase the computational efficiency, we employ a similar approach as in Namchaisiri et al. [14]. We introduce sets $\mathcal{S}_h = \{\boldsymbol{S} \ \in \mathbb{S}^n |\ \boldsymbol{S} = \mathcal{Q}_h^\top(\boldsymbol{z}), \boldsymbol{z} \in \mathcal{B}_{p_h^*}^{\lambda_h}\}$ for $h = 1,2,\dots,H$. Denoting the variables $(\boldsymbol{y}, \boldsymbol{S}_1, \dots, \boldsymbol{S}_H) \in \mathcal{U}$ as one composite variable $\boldsymbol{U}$, we can rewrite (2.1) as follows:

$$\begin{aligned} \max_{\boldsymbol{U}\in\mathcal{U}} \quad & g(\boldsymbol{U}) := \boldsymbol{b}^T\boldsymbol{y} + \mu\log\det\left(\boldsymbol{C} - \mathcal{A}^\top(\boldsymbol{y}) + \sum_{h=1}^{H}\boldsymbol{S}_h\right) + n\mu - n\mu\log\mu \\ \text{s.t.} \quad & \boldsymbol{S}_h \in \mathcal{S}_h(h=1,\dots,H),\ \boldsymbol{C} - \mathcal{A}^\top(\boldsymbol{y}) + \sum_{h=1}^{H}\boldsymbol{S}_h \succ \boldsymbol{O}. \end{aligned} \tag{$\mathcal{D}$}$$

Note that the difficulty due to the ℓ_{p_h}-norm in the problem is embedded into the set $\mathcal{S}_h$.

For the convergence analysis in the following section, we employ the same assumption as [13, 14]:

Assumption 1. *We assume the following statements hold for problem $(\mathcal{P})$ and its corresponding dual $(\mathcal{D})$:*

(i) *The set of matrices $\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{A}_m$ is linearly independent. In the other words, $\mathcal{A}$ is surjective;*

(ii) *There exists a strictly feasible solution $\widehat{\boldsymbol{X}} \succ \boldsymbol{O}$ of primal problem $(\mathcal{P})$ that satisfies $\mathcal{A}(\widehat{\boldsymbol{X}}) = \boldsymbol{b}$;*

(iii) *There exists a composite variable $\widehat{\boldsymbol{U}}$ that satisfies the constraint of dual problem $(\mathcal{D})$.*

Let $\mathcal{F}$ be the feasible region of $(\mathcal{D})$. We express this feasible region as the intersection $\mathcal{F} = \mathcal{M} \cap \mathcal{N}$ of $\mathcal{M} := \{\boldsymbol{U} \in \mathcal{U} \mid \boldsymbol{y} \in \mathbb{R}^m, \boldsymbol{S}_h \in \mathcal{S}_h(h=1,\dots,H)\}$ and $\mathcal{N} := \{\boldsymbol{U} \in \mathcal{U} \mid \boldsymbol{C} - \mathcal{A}^\top(\boldsymbol{y}) + \sum_{h=1}^{H}\boldsymbol{S}_h \succ \boldsymbol{O}\}$. Let $\boldsymbol{X}(\boldsymbol{U}) := \mu\left(\boldsymbol{C} - \mathcal{A}^\top(\boldsymbol{y}) + \sum_{h=1}^{H}\boldsymbol{S}_h\right)^{-1}$. Thus, the gradient of $g(\boldsymbol{U})$ can be expressed as

$$\begin{aligned} \nabla g(\boldsymbol{U}) &= \left(\nabla_{\boldsymbol{y}} g(\boldsymbol{U}), \nabla_{\boldsymbol{S}_1} g(\boldsymbol{U}), \dots, \nabla_{\boldsymbol{S}_H} g(\boldsymbol{U})\right) \\ &= \left(\boldsymbol{b} - \mathcal{A}(\boldsymbol{X}(\boldsymbol{U})), \boldsymbol{X}(\boldsymbol{U}), \dots, \boldsymbol{X}(\boldsymbol{U})\right). \end{aligned}$$

We introduce a map

$$\mathcal{B}(\boldsymbol{U}) := -\mathcal{A}^\top(\boldsymbol{y}) + \sum_{h=1}^{H} \boldsymbol{S}_h, \quad \text{where} \;\; \boldsymbol{U} = (\boldsymbol{y}, \boldsymbol{S}_1, \dots, \boldsymbol{S}_H),$$

and define $\boldsymbol{U}^k := \left(\boldsymbol{y}^k, \boldsymbol{S}_1^k, \dots, \boldsymbol{S}_H^k\right)$. For solving the dual problem $(\mathcal{D})$, we propose Algorithm 1 below by extending the DSPG method in [14]. The step length of Step 3 is the Barzilai-Borwein step [3]. The advantages of this step length are mentioned in [7, 13], in particular, the linear convergence can be obtained under a mild condition.

Algorithm 1 A DSPG algorithm for generalized log-det semidefinite programming

Step 0. Choose parameters $\varepsilon > 0, \tau \in (0,1), \gamma \in (0,1), 0 < \beta < 1, 0 < \alpha_{\min} < \alpha_{\max} < \infty$ and integer $M > 0$. Take $\boldsymbol{U}^0 \in \mathcal{F}$ and $\alpha_0 \in [\alpha_{\min}, \alpha_{\max}]$. Set the iteration number $k = 0$.

Step 1. Let $\Delta \boldsymbol{U}_{(1)}^k := \left([\Delta \boldsymbol{y}^k]_{(1)}, [\Delta \boldsymbol{S}_1^k]_{(1)}, \dots, [\Delta \boldsymbol{S}_H^k]_{(1)}\right) := P_{\mathcal{M}}\left(\boldsymbol{U}^k + \nabla g(\boldsymbol{U}^k)\right) - \boldsymbol{U}^k$. If $\|\Delta \boldsymbol{U}_{(1)}^k\| \leq \varepsilon$, terminate; otherwise, go to **Step 2**.

Step 2. Let $\boldsymbol{D}^k := \left(\Delta \boldsymbol{y}^k, \Delta \boldsymbol{S}_1^k, \dots, \Delta \boldsymbol{S}_H^k\right) := P_{\mathcal{M}}\left(\boldsymbol{U}^k + \alpha_k \nabla g(\boldsymbol{U}^k)\right) - \boldsymbol{U}^k$. Let $\boldsymbol{L}_k$ be the Cholesky factorization of $\boldsymbol{C} + \mathcal{B}(\boldsymbol{U}^k)$, that is $\boldsymbol{L}_k \boldsymbol{L}_k^\top = \boldsymbol{C} + \mathcal{B}(\boldsymbol{U}^k)$, and θ be the minimum eigenvalue of $\boldsymbol{L}_k^{-1} \mathcal{B}(\boldsymbol{D}^k) \left(\boldsymbol{L}_k^\top\right)^{-1}$. Set

$$\nu_k := \begin{cases} 1 & \text{if } \theta \geq 0, \\ \min\{1, \, -\tau/\theta\} & \text{otherwise.} \end{cases}$$

Apply a line search to find the largest element $\sigma_k \in \{1, \beta, \beta^2, \dots\}$ such that

$$g(\boldsymbol{U}^k + \sigma_k \nu_k \boldsymbol{D}^k) \geq \min_{[k-M+1]_+ \leq l \leq k} g(\boldsymbol{U}^l) + \gamma \sigma_k \nu_k \langle \nabla g(\boldsymbol{U}^k), \, \boldsymbol{D}^k \rangle.$$

Step 3. Let $\boldsymbol{U}^{k+1} = \boldsymbol{U}^k + \sigma_k \nu_k \boldsymbol{D}^k$. Let $p_k := \langle \boldsymbol{U}^{k+1} - \boldsymbol{U}^k, \, \nabla g(\boldsymbol{U}^{k+1}) - \nabla g(\boldsymbol{U}^k) \rangle$. Set

$$\alpha_{k+1} := \begin{cases} \alpha_{\max} & \text{if } p_k \geq 0, \\ \min\left\{\alpha_{\max}, \, \max\left\{\alpha_{\min}, \, -\|\boldsymbol{U}^{k+1} - \boldsymbol{U}^k\|^2 / p_k\right\}\right\} & \text{otherwise.} \end{cases}$$

Set $k \leftarrow k + 1$. Return to **Step 1**.

From the viewpoint of the convergence analysis, Algorithm 1 converges for any linear map $\mathcal{Q}_h$ as shown in Section 3. This is not proved in the previous DSPG papers [13, 14]. On the other hand, the efficiency of Algorithm 1 depends on the computation of the projection $P_{\mathcal{M}}(\cdot)$. The objective functions of the numerical experiments in Section 4 were chosen so that the projection onto each $\mathcal{S}_h$ can be computed within appropriate computation costs.

3 Convergence analysis

We show the convergence of Algorithm 1 to the optimal value by extending the analysis in [14]. In particular, we focus on the proof of the part $\sum_{h=1}^{H} \boldsymbol{S}_h$ that corresponds to $\sum_{h=1}^{H} \lambda_h \|\mathcal{Q}_h(\boldsymbol{X})\|_{p_h}$ in the primal generalized log-det SDPs $(\mathcal{P})$.

For the convergence proof below, we will show the validity of the stopping criterion in Lemma 6, the lower bound of step length that prevents the premature termination in Lemma 8, and the convergence of the output sequence of Algorithm 1 to the optimal solution in Theorem 11. Let $\{\boldsymbol{U}^k\}$ be the sequence generated by Algorithm 1. We use the notation $\boldsymbol{X}^k := \boldsymbol{X}(\boldsymbol{U}^k)$ in the proof.

We start with Lemma 2 that shows the feasibility and the boundedness of the sequence generated from Algorithm 1. We introduce a level set

$$\mathcal{L} := \left\{\boldsymbol{U} \in \mathcal{F} : \ g(\boldsymbol{U}) \geq g(\boldsymbol{U}^0)\right\}.$$

Lemma 2. [14, Lemma 3.1]*$\{\boldsymbol{U}^k\} \subseteq \mathcal{L}$ and $\{\boldsymbol{U}^k\}$ is bounded.*

In particular, the surjectivity of $\mathcal{A}$ and the existence of the primal interior feasible point $\widehat{\boldsymbol{X}}$ in Assumption 1 play an essential part in the proof of [14].

With the boundedness of $\mathcal{S}_1, \ldots, \mathcal{S}_H$, we also obtain the following lemma, which will be used in Lemma 10.

Lemma 3. [13, Lemma 3.3]*The level set $\mathcal{L}$ is bounded.*

The proof of this lemma also utilizes the surjectivity of $\mathcal{A}$ to derive the boundedness of the component of $\boldsymbol{y}$ in $\boldsymbol{U}$.

Using Lemma 2 and strict concavity of log det term in the objective function of $(\mathcal{D})$, we can show that the eigenvalues of $\boldsymbol{C} + \mathcal{B}(\boldsymbol{U}^k)$ are sandwiched with a positive lower bound and a finite upper bound. Since $\boldsymbol{X}^k = \mu(\boldsymbol{C} + \mathcal{B}(\boldsymbol{U}^k))^{-1}$, we obtain the boundedness of the sequence $\{\boldsymbol{X}^k\}$.

Corollary 4. [13, Remark 3.5] *There exist bounds $\beta_{\boldsymbol{X}}^{\min}$ and $\beta_{\boldsymbol{X}}^{\max}$ such that $\boldsymbol{O} \preceq \beta_{\boldsymbol{X}}^{\min}\boldsymbol{I} \preceq \boldsymbol{X}^k \preceq \beta_{\boldsymbol{X}}^{\max}\boldsymbol{I}$ for all k. Thus $\eta_{\boldsymbol{X}} := \sqrt{n}\beta_{\boldsymbol{X}}^{\max}$ is also a bound for $\{\boldsymbol{X}^k\}$ such that $\|\boldsymbol{X}^k\| \leq \eta_{\boldsymbol{X}}$ for all k.*

Furthermore, we can also show the boundedness of other components related to $\boldsymbol{U}^k$.

Lemma 5. *There exist bounds $\eta_{\boldsymbol{X}^{-1}}$ and $\eta_{\Delta \boldsymbol{y}} > 0$ such that $\|(\boldsymbol{X}^k)^{-1}\| \leq \eta_{\boldsymbol{X}^{-1}}$ and $\|\Delta \boldsymbol{y}^k\| \leq \eta_{\Delta \boldsymbol{y}}$ for all k. There also exists $\eta_{\Delta \boldsymbol{S}} > 0$ that satisfies $\|\Delta \boldsymbol{S}_h^k\| \leq \eta_{\Delta \boldsymbol{S}}$ for all $h = 1, 2, \ldots, H$ and all k.*

Proof. We can obtain $\eta_{\boldsymbol{X}^{-1}}$ and $\eta_{\Delta \boldsymbol{y}}$ by applying Lemma 3.4 in [14]. For $h = 1, 2, \ldots, H$, since $\boldsymbol{U}^k \in \mathcal{M} \subset \mathcal{F}$ from Lemma 2, we have $\boldsymbol{S}_h^k \in \mathcal{S}_h$ for any $k \geq 1$. Therefore, due to the property of the projection, it holds that $\|\Delta \boldsymbol{S}_h^k\| = \left\|P_{\mathcal{S}_h}\left(\boldsymbol{S}_h^k + \alpha_k \boldsymbol{X}^k\right) - \boldsymbol{S}_h^k\right\| \leq \|\alpha_k \boldsymbol{X}^k\| \leq \alpha_{\max}\eta_{\boldsymbol{X}} =: \eta_{\Delta \boldsymbol{S}}$. □

The following lemma derives an optimality condition for the dual problem $(\mathcal{D})$ from the viewpoint of the projection, and this guarantees the validity of the stopping criterion in Step 1 of Algorithm 1. The proofs in [14] cannot directly give this lemma, since the structure of set $\mathcal{M}$ now includes multiple sets $\mathcal{S}_1, \ldots, \mathcal{S}_H$.

Lemma 6. *$\boldsymbol{U}^* \in \mathcal{F}$ is an optimal solution of the dual problem $(\mathcal{D})$ if and only if there exists $\alpha > 0$ such that*

$$P_{\mathcal{M}}(\boldsymbol{U}^* + \alpha \nabla g(\boldsymbol{U}^*)) = \boldsymbol{U}^*.$$

Proof. We can see that the objective equation is equivalent to

$$\boldsymbol{U}^* = \arg\min \left\{\frac{1}{2}\|\boldsymbol{U} - (\boldsymbol{U}^* + \alpha \nabla g(\boldsymbol{U}^*))\|^2 + \delta_{\mathcal{M}}(\boldsymbol{U})\right\}. \tag{3.1}$$

where $\delta_{\mathcal{M}}$ is the indicator function of $\mathcal{M}$.

Let $\boldsymbol{U}^*$ be decomposed into $\boldsymbol{U}^* = (\boldsymbol{y}^*, \boldsymbol{S}_1^*, \boldsymbol{S}_2^*, \ldots, \boldsymbol{S}_H^*)$. From the definition of the projection onto $\boldsymbol{S}_h$ for each $h = 1, \ldots, H$, the equality $P_{\boldsymbol{\mathcal{S}}_h}(\boldsymbol{S}_h^* + \alpha \nabla_{\boldsymbol{S}_h} g(\boldsymbol{U}^*)) = \boldsymbol{S}_h^*$ holds if and only if

$$\begin{aligned}\boldsymbol{S}_h^* &= \arg\min_{\boldsymbol{S}_h \in \mathcal{S}_h} \frac{1}{2}\|\boldsymbol{S}_h - (\boldsymbol{S}_h^* + \alpha \nabla_{\boldsymbol{S}_h} g(\boldsymbol{U}^*))\|^2 \\ &= \arg\min \left\{\frac{1}{2}\|\boldsymbol{S}_h - (\boldsymbol{S}_h^* + \alpha \nabla_{\boldsymbol{S}_h} g(\boldsymbol{U}^*))\|^2 + \delta_{\mathcal{S}_h}(\boldsymbol{S}_h)\right\}.\end{aligned}$$

From the definition $\mathcal{M} := \{\boldsymbol{U} \in \mathcal{U} \mid \boldsymbol{y} \in \mathbb{R}^m, \boldsymbol{S}_h \in \mathcal{S}_h (h = 1, \ldots, H)\}$, we can see that projection computation for each component of $\boldsymbol{U}$ to $\mathcal{M}$ is independent from the other components.

Since the subgradient of the indicator function is a normal cone, (3.1) is equivalent to $0 \in \alpha \nabla g(\boldsymbol{U}^*) + N_{\mathcal{M}}(\boldsymbol{U}^*)$, where $N_{\mathcal{M}}(\boldsymbol{U}^*)$ denotes the normal cone of $\mathcal{M}$ at $\boldsymbol{U}^*$. Since $\mathcal{N} = \left\{\boldsymbol{U} \in \mathcal{U} \mid \boldsymbol{C} - \mathcal{A}^\top(\boldsymbol{y}) + \sum_{h=1}^H \boldsymbol{S}_h \succ \boldsymbol{O}\right\}$, we can see that $\mathcal{N}$ is an open set, which means all of the elements in $\mathcal{N}$ are interior points of $\mathcal{N}$. According to Theorem 3.30 in [2], we obtain that $N_{\mathcal{M}}(\boldsymbol{U}^*) = N_{\mathcal{M} \cap \mathcal{N}}(\boldsymbol{U}^*) = N_{\mathcal{F}}(\boldsymbol{U}^*)$. This implies $0 \in \alpha \nabla g(\boldsymbol{U}^*) + N_{\mathcal{F}}(\boldsymbol{U}^*)$, and this condition is equivalent to the optimality of $\boldsymbol{U}^*$ in the dual problem $(\mathcal{D})$. □

We will show in Lemma 7 that the difference between $\lambda_h \|\mathcal{Q}_h(\boldsymbol{X}^k)\|_{p_h}$ in the primal objective function and an inner product $\boldsymbol{S}_h^k \bullet \boldsymbol{X}^k$ can be estimated with $\|[\Delta \boldsymbol{S}_h^k]_{(1)}\|$. This difference is a part of the duality gap between $(\mathcal{P})$ and $(\mathcal{D})$ in the kth iteration. Therefore, we employ the limit $\liminf_{k\to\infty} \|[\Delta \boldsymbol{S}_h^k]_{(1)}\| \to 0$ (which will be shown in Lemma 9) to show the limit $\liminf_{k\to\infty} |g(\boldsymbol{U}^k) - g^*|$ in Lemma 10 which leads to the convergence to the objective value.

Lemma 7. *For $h = 1, 2, \ldots, H$, $|\lambda_h \|\mathcal{Q}_h(\boldsymbol{X}^k)\|_{p_h} - \boldsymbol{S}_h^k \bullet \boldsymbol{X}^k|$ is bounded by $\|[\Delta \boldsymbol{S}_h^k]_{(1)}\|$. More precisely,*

$$|\lambda_h \|\mathcal{Q}_h(\boldsymbol{X}^k)\|_{p_h} - \boldsymbol{S}_h^k \bullet \boldsymbol{X}^k| \le c_h \|[\Delta \boldsymbol{S}_h^k]_{(1)}\|$$

holds for all k with $c_h := \eta_{\boldsymbol{X}} + \lambda_h \sqrt{n_h}(||\mathcal{Q}_h|| + ||\mathcal{Q}_h^\top||)$.

Proof. Since $\boldsymbol{U}^k \in \mathcal{M}$, we know $\boldsymbol{S}_h^k \in \mathcal{S}_h$, thus there exists $\boldsymbol{z}_h^k \in \mathbb{R}^{n_h}$ such that $\|\boldsymbol{z}_h^k\|_{p_h^*} \le \lambda_h$ and $\boldsymbol{S}_h^k = \mathcal{Q}_h^\top(\boldsymbol{z}_h^k)$. Therefore, we have

$$\boldsymbol{S}_h^k \bullet \boldsymbol{X}^k = \mathcal{Q}_h^\top(\boldsymbol{z}_h^k) \bullet \boldsymbol{X}^k = \mathcal{Q}_h(\boldsymbol{X}^k)^\top \boldsymbol{z}_h^k.$$

The Holder's inequality $|\boldsymbol{a}^T \boldsymbol{b}| \le \|\boldsymbol{a}\|_{p^*} \|\boldsymbol{b}\|_p$ holds for any vectors $\boldsymbol{a}, \boldsymbol{b}$ with the same length and $p \ge 1$. Thus, $|\boldsymbol{S}_h^k \bullet \boldsymbol{X}^k| = |\mathcal{Q}_h(\boldsymbol{X}^k)^\top \boldsymbol{z}_h^k| \le ||\boldsymbol{z}_h^k||_{p_h^*} ||\mathcal{Q}_h(\boldsymbol{X}^k)||_{p_h} \le \lambda_h ||\mathcal{Q}_h(\boldsymbol{X}^k)||_{p_h}$, and this leads to $\lambda_h ||\mathcal{Q}_h(\boldsymbol{X}^k)||_p - \boldsymbol{S}_h^k \bullet \boldsymbol{X}^k \ge 0$. Furthermore, we know that $||\mathcal{Q}_h(\boldsymbol{X})||_{p_h} \le ||\mathcal{Q}_h(\boldsymbol{X})||_1$ for $p_h \ge 1$, thus it holds that

$$|\lambda_h \|\mathcal{Q}_h(\boldsymbol{X}^k)\|_{p_h} - \boldsymbol{S}_h^k \bullet \boldsymbol{X}^k| \le |\lambda_h \|\mathcal{Q}_h(\boldsymbol{X}^k)\|_1 - \boldsymbol{S}_h^k \bullet \boldsymbol{X}^k|. \tag{3.2}$$

Let $\hat{\boldsymbol{S}}_h^k = P_{\mathcal{S}_h}(\boldsymbol{S}_h^k + \boldsymbol{X}^k)$. Then there exists $\hat{\boldsymbol{z}}_h^k \in \mathbb{R}^{n_h}$ such that $\|\hat{\boldsymbol{z}}_h^k\|_{p^*} \le \lambda_h$ and $\hat{\boldsymbol{S}}_h^k = \mathcal{Q}_h^\top(\hat{\boldsymbol{z}}_h^k)$. Let $\tilde{\boldsymbol{e}}_h^k \in \mathbb{R}^{n_h}$ be a vector whose elements are the signs of $\mathcal{Q}_h(\boldsymbol{X}^k)$. We have

$$\begin{aligned}|\lambda_h \|\mathcal{Q}_h(\boldsymbol{X}^k)\|_1 - \boldsymbol{S}_h^k \bullet \boldsymbol{X}^k| = |\lambda_h (\tilde{\boldsymbol{e}}_h^k)^\top \mathcal{Q}_h(\boldsymbol{X}^k) - \boldsymbol{S}_h^k \bullet \boldsymbol{X}^k| &= |\lambda_h \mathcal{Q}_h^\top(\tilde{\boldsymbol{e}}_h^k) \bullet \boldsymbol{X}^k - \boldsymbol{S}_h^k \bullet \boldsymbol{X}^k| \\ &\le |\lambda_h \mathcal{Q}_h^\top(\tilde{\boldsymbol{e}}_h^k) \bullet \boldsymbol{X}^k - \hat{\boldsymbol{S}}_h^k \bullet \boldsymbol{X}^k| + |(\hat{\boldsymbol{S}}_h^k - \boldsymbol{S}_h^k) \bullet \boldsymbol{X}^k|\end{aligned}$$

$$\leq |(\lambda_h \mathcal{Q}_h^\top(\tilde{\boldsymbol{e}}_h^k) - \hat{\boldsymbol{S}}_h^k) \bullet \boldsymbol{X}^k| + \|[\Delta \boldsymbol{S}_h^k]_{(1)}\| \cdot \|\boldsymbol{X}^k\|. \tag{3.3}$$

We can see that the second term of (3.3) is bounded by $\|[\Delta \boldsymbol{S}_h^k]_{(1)}\|$ due to $\|\boldsymbol{X}^k\| \leq \eta_{\boldsymbol{X}}$ in Lemma 5. Therefore, our focus here is the first term. Due to properties P1 in [7, Proposition 2.1] as a property of the projection $\hat{\boldsymbol{S}}_h^k = P_{\mathcal{S}_h}(\boldsymbol{S}_h^k + \boldsymbol{X}^k)$, we can derive an inequality

$$\left(\boldsymbol{X}^k + [\Delta \boldsymbol{S}_h^k]_{(1)}\right) \bullet \left(\lambda_h \mathcal{Q}_h^\top(\tilde{\boldsymbol{e}}_h^k) - \hat{\boldsymbol{S}}_h^k\right) = \left(\boldsymbol{S}_h^k + \boldsymbol{X}^k - \hat{\boldsymbol{S}}_h^k\right) \bullet \left(\lambda_h \mathcal{Q}_h^\top(\tilde{\boldsymbol{e}}_h^k) - \hat{\boldsymbol{S}}_h^k\right) \leq 0.$$

This indicates

$$\boldsymbol{X}^k \bullet \left(\lambda_h \mathcal{Q}_h^\top(\tilde{\boldsymbol{e}}_h^k) - \hat{\boldsymbol{S}}_h^k\right) \leq [\Delta \boldsymbol{S}_h^k]_{(1)} \bullet \left(\lambda_h \mathcal{Q}_h^\top(\tilde{\boldsymbol{e}}_h^k) - \hat{\boldsymbol{S}}_h^k\right). \tag{3.4}$$

We show the nonnegativity of $\boldsymbol{X}^k \bullet \left(\lambda_h \mathcal{Q}_h^\top(\tilde{\boldsymbol{e}}_h^k) - \hat{\boldsymbol{S}}_h^k\right) = \boldsymbol{X}^k \bullet \left(\lambda_h \mathcal{Q}_h^\top(\tilde{\boldsymbol{e}}_h^k) - \mathcal{Q}_h^\top(\hat{\boldsymbol{z}}_h^k)\right) = \mathcal{Q}(\boldsymbol{X}^k) \bullet (\lambda_h \tilde{\boldsymbol{e}}_h^k - \hat{\boldsymbol{z}}_h^k) = \sum_{j=1}^{n_h} [\mathcal{Q}_h(\boldsymbol{X}^k)]_j [\lambda_h \tilde{\boldsymbol{e}}_h^k - \hat{\boldsymbol{z}}_h^k]_j$. For each j, we can see that the sign of $[\lambda_h \tilde{\boldsymbol{e}}_h^k - \hat{\boldsymbol{z}}_h^k]_j$ is the same as $\tilde{\boldsymbol{e}}_h^k$ because $|[\hat{\boldsymbol{z}}_h^k]_j| \leq \|\hat{\boldsymbol{z}}^k\|_{p_h^*} \leq \lambda_h$. Since $\tilde{\boldsymbol{e}}_h^k$ is the sign of $[\mathcal{Q}_h(\boldsymbol{X}^k)]_j$, we obtain $[\mathcal{Q}_h(\boldsymbol{X}^k)]_j [\lambda_h \tilde{\boldsymbol{e}}_h^k - \hat{\boldsymbol{z}}_h^k]_j \geq 0$, hence,

$$\boldsymbol{X}^k \bullet \left(\lambda_h \mathcal{Q}_h^\top(\tilde{\boldsymbol{e}}_h^k) - \hat{\boldsymbol{S}}_h^k\right) = \sum_{j=1}^{n_h} [\mathcal{Q}_h(\boldsymbol{X}^k)]_j [\lambda_h \tilde{\boldsymbol{e}}_h^k - \hat{\boldsymbol{z}}_h^k]_j \geq 0.$$

Applying this result into (3.4) and using Lemma 5, we obtain

$$|\boldsymbol{X}^k \bullet \left(\lambda_h \mathcal{Q}_h^\top(\tilde{\boldsymbol{e}}_h^k) - \hat{\boldsymbol{S}}_h^k\right)| \leq |[\Delta \boldsymbol{S}_h^k]_{(1)} \bullet \left(\lambda_h \mathcal{Q}_h^\top(\tilde{\boldsymbol{e}}_h^k) - \hat{\boldsymbol{S}}_h^k\right)|. \tag{3.5}$$

Since $\hat{\boldsymbol{S}}_h^k \in \mathcal{S}_h$, we obtain the bound

$$||\hat{\boldsymbol{S}}_h^k|| \leq ||\mathcal{Q}_h|| ||\hat{\boldsymbol{z}}_h^k|| \leq \sqrt{n_h} ||\mathcal{Q}_h|| ||\hat{\boldsymbol{z}}_h^k||_\infty \leq \sqrt{n_h} ||\mathcal{Q}_h|| ||\hat{\boldsymbol{z}}_h^k||_{p_h^*} \leq \sqrt{n_h} \lambda_h ||\mathcal{Q}_h||. \tag{3.6}$$

Using (3.6) in (3.5), we have

$$|\boldsymbol{X}^k \bullet \left(\lambda_h \mathcal{Q}_h^\top(\tilde{\boldsymbol{e}}_h^k) - \hat{\boldsymbol{S}}_h^k\right)| \leq (\lambda_h \sqrt{n_h} ||\mathcal{Q}_h^\top|| + \lambda_h \sqrt{n_h} ||\mathcal{Q}_h||) \|[\Delta \boldsymbol{S}_h^k]_{(1)}\|. \tag{3.7}$$

This indicates that $|\boldsymbol{X}^k \bullet \left(\lambda_h \mathcal{Q}_h^\top(\tilde{\boldsymbol{e}}_h^k) - \hat{\boldsymbol{S}}_h^k\right)|$ is bounded by $\|[\Delta \boldsymbol{S}_h^k]_{(1)}\|$. Combining (3.2), (3.3) and (3.7), we can conclude this lemma with the value of $c_h = \eta_{\boldsymbol{X}} + \lambda_h \sqrt{n_h}(||\mathcal{Q}_h|| + ||\mathcal{Q}_h^\top||)$. □

Lemma 8 shows the lower bound of the step length, which prevents the Algorithm 1 from terminating before the stopping criterion is satisfied.

Lemma 8. *There exists a bound $(\sigma\nu)_{\min} > 0$ such that step length $\sigma_k \nu_k > (\sigma\nu)_{\min}$ for all k.*

Proof. Let $\mathcal{B}(\boldsymbol{D}^k) := \mathcal{A}^\top(\Delta \boldsymbol{y}) + \sum_{h=1}^H \Delta \boldsymbol{S}_h$. Using Lemma 5, we have its bound

$$\begin{aligned} ||\mathcal{B}(\boldsymbol{D}^k)|| \leq\ & ||\mathcal{A}^\top|| ||\Delta \boldsymbol{y}|| + \sum_{h=1}^H ||\Delta \boldsymbol{S}_h|| \leq ||\mathcal{A}^\top|| \eta_{\Delta \boldsymbol{y}} + H \eta_{\Delta \boldsymbol{S}} \\ \leq\ & \sqrt{H+1} \max\{||\mathcal{A}^\top||, 1\} ||\boldsymbol{D}^k||. \end{aligned} \tag{3.8}$$

We divide the proof into two sections. Firstly, we show that ν_k has a lower bound, and then we will show that $\sigma_k\nu_k$ has a lower bound.

From the definition of ν_k in Step 2 of Algorithm 1, we consider only the case that $\theta < 0$. The definition of θ that is the minimum eigenvalue of $\boldsymbol{L}_k^{-1}\mathcal{B}(\boldsymbol{D}^k)(\boldsymbol{L}_k^\top)^{-1}$ implies that θ is the maximum value that satisfies $\boldsymbol{L}_k^{-1}\mathcal{B}(\boldsymbol{D}^k)(\boldsymbol{L}_k^\top)^{-1} \succeq \theta\boldsymbol{I}$. Using the property $\boldsymbol{L}_k\boldsymbol{L}_k^\top = \boldsymbol{C} + \mathcal{B}(\boldsymbol{U}^k)$ and $\theta < 0$, $\nu' = -1/\theta$ is the maximum value that satisfies

$$\boldsymbol{C} + \mathcal{B}(\boldsymbol{U}^k) + \nu'\mathcal{B}(\boldsymbol{D}^k) \succeq \boldsymbol{O}.$$

On the other hand, using the bound from Corollary 4, we have

$$\begin{aligned}\boldsymbol{C} + \mathcal{B}(\boldsymbol{U}^k) + \nu'\mathcal{B}(\boldsymbol{D}^k) &= \mu(\boldsymbol{X}^k)^{-1} + \nu'\mathcal{B}(\boldsymbol{D}^k) \\ \succeq \frac{\mu}{\beta_{\boldsymbol{X}}^{\max}}\boldsymbol{I} - \nu'\|\mathcal{B}(\boldsymbol{D}^k)\|\boldsymbol{I} &\succeq \left(\frac{\mu}{\beta_{\boldsymbol{X}}^{\max}} - \nu'\left(\|\mathcal{A}^\top\|\eta_{\Delta\boldsymbol{y}} + H\eta_{\Delta\boldsymbol{S}}\right)\right)\boldsymbol{I}.\end{aligned}$$

Therefore, if we consider the interval $0 \le \nu' \le \mu/\left(\beta_{\boldsymbol{X}}^{\max}(\|\mathcal{A}^\top\|\eta_{\Delta\boldsymbol{y}} + H\eta_{\Delta\boldsymbol{S}})\right) =: \nu_{\min}$, we can guarantee the positive semidefiniteness $\boldsymbol{C} + \mathcal{B}(\boldsymbol{U}^k) + \nu'\mathcal{B}(\boldsymbol{D}^k) \succeq 0$. This implies a positive lower bound $\nu_k \ge \min\{1, \tau\nu_{\min}\}$.

From the discussion of [14, Lemma 3.5], we have the bound

$$\left\|\left(\boldsymbol{X}(\boldsymbol{U}^k + \nu\boldsymbol{D}^k) - \boldsymbol{X}(\boldsymbol{U}^k)\right)\right\| \le \frac{\nu}{\mu\left(\frac{1-\tau}{\beta_{\boldsymbol{X}}^{\max}}\right)^2}\|\mathcal{B}(\boldsymbol{D}^k)\|.$$

Therefore, it holds for $\lambda \ge 0$ that

$$\begin{aligned}&\|\nabla g(\boldsymbol{U}^k + \lambda\boldsymbol{D}^k) - \nabla g(\boldsymbol{U}^k)\| \\ &= \left\|\left(-\mathcal{A}(\boldsymbol{X}(\boldsymbol{U}^k + \lambda\boldsymbol{D}^k) - \boldsymbol{X}(\boldsymbol{U}^k)), \boldsymbol{X}(\boldsymbol{U}^k + \lambda\boldsymbol{D}^k) - \boldsymbol{X}(\boldsymbol{U}^k), \dots, \boldsymbol{X}(\boldsymbol{U}^k + \lambda\boldsymbol{D}^k) - \boldsymbol{X}(\boldsymbol{U}^k)\right)\right\| \\ &\le \sqrt{\|\mathcal{A}\|^2 + H}\ \|\boldsymbol{X}(\boldsymbol{U}^k + \lambda\boldsymbol{D}^k) - \boldsymbol{X}(\boldsymbol{U}^k)\| \\ &\le \frac{\sqrt{\|\mathcal{A}\|^2 + H}}{\mu\left(\frac{1-\tau}{\beta_{\boldsymbol{X}}^{\max}}\right)^2}\lambda\|\mathcal{B}(\boldsymbol{D}^k)\| \le \left(\frac{\sqrt{\|\mathcal{A}\|^2 + H}}{\mu\left(\frac{1-\tau}{\beta_{\boldsymbol{X}}^{\max}}\right)^2}\sqrt{H+1}\max\{\|\mathcal{A}^\top\|, 1\}\right)\lambda\|\boldsymbol{D}^k\| = \lambda L\|\boldsymbol{D}^k\|, \qquad (3.9)\end{aligned}$$

where the last inequality was due to (3.8) and we introduce $L := \frac{\sqrt{\|\mathcal{A}\|^2 + H}}{\mu(\frac{1-\tau}{\beta_{\boldsymbol{X}}^{\max}})^2}\sqrt{H+1}\max\{\|\mathcal{A}^\top\|, 1\}$.

We focus on the termination condition of the non-monotone Armijo rule in the line search of Step 2 in Algorithm 1. If it terminates at $\sigma_k = 1$, then we obtain a lower bound as $\sigma_k\nu_k \ge \min\{1, \tau\nu_{\min}\}$. If it terminates at $\sigma_k = \beta^j$ for some $j \ge 1$, this indicates that the termination condition is not satisfied at $\sigma_k = \beta^{j-1}$. This further implies

$$\begin{aligned}g(\boldsymbol{U}^k + \beta^{j-1}\nu_k\boldsymbol{D}^k) &< \min_{[k-M+1]_+ \le l \le k} g(\boldsymbol{U}^l) + \gamma\beta^{j-1}\nu_k\langle\nabla g(\boldsymbol{U}^k),\, \boldsymbol{D}^k\rangle \\ &\le g(\boldsymbol{U}^k) + \gamma\beta^{j-1}\nu_k\langle\nabla g(\boldsymbol{U}^k),\, \boldsymbol{D}^k\rangle.\end{aligned}$$

Therefore, we have

$$\gamma\beta^{j-1}\nu_k\langle\nabla g(\boldsymbol{U}^k),\, \boldsymbol{D}^k\rangle \ge g(\boldsymbol{U}^k + \beta^{j-1}\nu_k\boldsymbol{D}^k) - g(\boldsymbol{U}^k) \qquad (3.10)$$

$$
\begin{aligned}
&= \beta^{j-1}\nu_k \langle \nabla g(\boldsymbol{U}^k), \boldsymbol{D}^k \rangle + \int_0^{\beta^{j-1}\nu_k} \langle \nabla g(\boldsymbol{U}^k + \lambda \boldsymbol{D}^k) - \nabla g(\boldsymbol{U}^k), \boldsymbol{D}^k \rangle d\lambda. \\
&\geq \beta^{j-1}\nu_k \langle \nabla g(\boldsymbol{U}^k), \boldsymbol{D}^k \rangle - \frac{L(\beta^{j-1}\nu_k)^2}{2} \|\boldsymbol{D}^k\|^2,
\end{aligned}
$$

where the equality is due to Taylor's expansion and the last inequality holds from (3.9). If $||\boldsymbol{D}^k|| = 0$, Algorithm 1 should be terminated in Step 1, so we can here assume $||\boldsymbol{D}^k|| > 0$. From $\beta > 0$ and $\nu_k \geq \min\{1, \tau\nu_{\min}\}$, we know that $\beta^{j-1}\nu_k > 0$. Therefore, (3.10) is equivalent to

$$
\beta^{j-1}\nu_k \geq \frac{2(1-\gamma)}{L} \frac{\langle \nabla g(\boldsymbol{U}^k), \boldsymbol{D}^k \rangle}{||\boldsymbol{D}^k||^2}. \tag{3.11}
$$

Using a property of the projection (see [7, Proposition 2.1, P1]), it holds that

$$
\langle (\boldsymbol{U}^k + \alpha_k \nabla g(\boldsymbol{U}^k)) - P_{\mathcal{M}}(\boldsymbol{U}^k + \alpha_k \nabla g(\boldsymbol{U}^k)),\, \boldsymbol{U}^k - P_{\mathcal{M}}(\boldsymbol{U}^k + \alpha_k \nabla g(\boldsymbol{U}^k)) \rangle \leq 0.
$$

The left-hand side is

$$
\langle -\boldsymbol{D}^k + \alpha_k \nabla g(\boldsymbol{U}^k),\, -\boldsymbol{D}^k \rangle = -\alpha_k \langle \nabla g(\boldsymbol{U}^k),\, \boldsymbol{D}^k \rangle + \|\boldsymbol{D}^k\|^2,
$$

thus we obtain

$$
\frac{\langle \nabla g(\boldsymbol{U}^k),\, \boldsymbol{D}^k \rangle}{||\boldsymbol{D}^k||^2} \geq \frac{1}{\alpha_k} \geq \frac{1}{\alpha_{\max}}. \tag{3.12}
$$

Applying (3.12) to (3.11), we obtain the positive lower bound $\beta^{j-1}\nu_k \geq \frac{2(1-\gamma)}{\alpha_{\max}L}$, which leads to the positive lower bound $\sigma_k\nu_k := \beta^j\nu_k \geq \frac{2\beta(1-\gamma)}{\alpha_{\max}L}$. Combining the case $\sigma_k = 1$, we obtain $\sigma_k\nu_k \geq \min\{1, \tau\nu_{\min}, \frac{2\beta(1-\gamma)}{\alpha_{\max}L}\}$ for all k. This completes the proof. □

Lemma 9 indicates the limit of the search direction of Algorithm 1, which will be used in the proof of Lemma 10.

Lemma 9. *Algorithm 1 with the stopping criterion parameter $\varepsilon = 0$ terminates after reaching the optimal value g^*, or it generates the sequence $\{\boldsymbol{U}^k\}$ that satisfies*

$$
\liminf_{k\to\infty} \|\Delta \boldsymbol{U}^k_{(1)}\| = 0.
$$

Proof. As discussed in [14, Lemma 3.6], the dual objective value increases at least in every M iteration, where M is used in the line search of Step 2 and it has an upper bound since the level set $\mathcal{L}$ is bounded from Lemma 3. Using the existence of the lower bound of $\sigma_k\nu_k$ from Lemma 8, we can show $\liminf_{k\to\infty} \|\boldsymbol{D}^k\| = 0$. Furthermore, from the properties P4 and P5 in [7, Proposition 2.1], we can show that the inequality $\|\Delta \boldsymbol{U}^k_{(1)}\| \leq \max\{1, \alpha_{\max}\}\|\boldsymbol{D}^k\|$ holds. This implies the statement of this lemma. □

We now discuss the convergence of the objective value to the optimal value of $(\mathcal{D})$, denoted with g^*.

Lemma 10. *Algorithm 1 with the stopping criterion parameter $\varepsilon = 0$ terminates after reaching the optimal value g^*, or it generates the sequence $\{\boldsymbol{U}^k\}$ that satisfies*

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

Proof. Let $\boldsymbol{X}^*$ be the optimal solution of $(\mathcal{P})$. Due to the strict convexity of the logarithm function, $\boldsymbol{X}^*$ is the unique solution. We decompose the difference of $|g(\boldsymbol{U}^k) - g^*|$ into the summation of three parts by the following inequality:

$$|g(\boldsymbol{U}^k) - g^*| \leq |g(\boldsymbol{U}^k) - f(\boldsymbol{X}^k)| + |f(\boldsymbol{X}^k) - f(\boldsymbol{X}^*)| + |f(\boldsymbol{X}^*) - g^*|. \tag{3.13}$$

The first part of (3.13) can be evaluated as

$$\begin{aligned}|g(\boldsymbol{U}^k) - f(\boldsymbol{X}^k)| &= \Bigg| \boldsymbol{b}^T \boldsymbol{y}^k + \mu \log\det\left(\boldsymbol{C} - \mathcal{A}^\top(\boldsymbol{y}^k) + \sum_{h=1}^{H} \boldsymbol{S}_h^k\right) + n\mu - n\mu\log\mu \\ &\qquad - \boldsymbol{C} \bullet \boldsymbol{X}^k + \mu \log\det \boldsymbol{X}^k - \sum_{h=1}^{H} \|\mathcal{Q}_h(\boldsymbol{X}^k)\|_{p_h} \Bigg| \\ &= \left| (\boldsymbol{y}^k)^\top (\boldsymbol{b} - \mathcal{A}(\boldsymbol{X}^k)) - \sum_{h=1}^{H} (\lambda_h \|\mathcal{Q}(\boldsymbol{X}^k)\|_p - \boldsymbol{S}_h^k \bullet \boldsymbol{X}^k) \right| \\ &\leq \|\boldsymbol{y}^k\| \|\mathcal{A}\| \|\boldsymbol{X}^* - \boldsymbol{X}^k\| + \sum_{h=1}^{H} \left| \lambda_h \|\mathcal{Q}(\boldsymbol{X}^k)\|_p - \boldsymbol{S}_h^k \bullet \boldsymbol{X}^k \right|.\end{aligned}$$

The first term is bounded by $\|\boldsymbol{X}^k - \boldsymbol{X}^*\|$ due to the boundedness of $\boldsymbol{y}^k$ from Lemma 5, and the second summation is bounded by $\|[\Delta \boldsymbol{S}_h^k]_{(1)}\|$ as shown in Lemma 7. From [13, Lemma 3.15], $\|\boldsymbol{X}^k - \boldsymbol{X}^*\|$ can be bounded by $||\Delta \boldsymbol{U}_{(1)}^k||$. Therefore, $|g(\boldsymbol{U}^k) - f(\boldsymbol{X}^k)|$ is bounded by $||\Delta \boldsymbol{U}_{(1)}^k||$. We can use the same discussion as [14, Lemma 3.8] to show that the second term of (3.13) is also bounded by $\|\Delta \boldsymbol{U}_{(1)}^k\|$. Due to Assumption 1, the duality theorem between $(\mathcal{P})$ and $(\mathcal{D})$ holds, and makes the third term in (3.13) zero. Combining three terms, we can show that $|g(\boldsymbol{U}^k) - g^*|$ is bounded by $||\Delta \boldsymbol{U}_{(1)}^k||$. Finally, from Lemma 9, we have $\liminf_{k\to\infty} ||\Delta \boldsymbol{U}_{(1)}^k|| = 0$. This completes the proof. □

Combining these results, we can show the main convergence of Algorithm 1.

Theorem 11. *Algorithm 1 with the stopping criterion parameter $\varepsilon = 0$ terminates after reaching the optimal value g^*, or it generates the sequence $\{\boldsymbol{U}^k\}$ that satisfies*

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

Proof. We use the contradiction to prove this theorem. Suppose that there exists $\epsilon > 0$ and an infinite increasing sequence $\{k_1, k_2, \dots\}$ such that $g(\boldsymbol{U}^{k_i}) \leq g^* - \epsilon$ for all i and $g(\boldsymbol{U}^l) > g^* - \epsilon$ for all $l \notin \{k_1, k_2, \dots\}$.

Firstly, we will show that $k_{i+1} - k_i \leq M$ for all i. Suppose that there exists an i such that $k_{i+1} - k_i > M$, which means all elements in $\{k_{i+1} - M, k_{i+1} - M + 1, \dots k_{i+1} - 1\}$ is not contained in the sequence of k_i. Therefore, $g(\boldsymbol{U}^l) > g^* - \epsilon$ for all $k_{i+1} - M \leq l \leq k_{i+1} - 1$, and this is equivalent to $g(\boldsymbol{U}^l) > g^* - \epsilon$.

From (3.12), we obtain

$$\alpha_k \langle \nabla g(\boldsymbol{U}^k),\, \boldsymbol{D}^k \rangle \geq \|\boldsymbol{D}^k\|^2 \geq 0.$$

Applying this to the inequality in Step 2 of Algorithm 1, we obtain that

$$g(\boldsymbol{U}^{k_{i+1}}) \geq \min_{k_{i+1} - M \leq l \leq k_{i+1} - 1} g(\boldsymbol{U}^l) \geq g^* - \epsilon.$$

However, this contradicts to $g(\boldsymbol{U}^{k_i}) \leq g^* - \epsilon$ for all i. Therefore, we obtain $k_{i+1} - k_i \leq M$ for all i.

From the proof in Lemma 10, we know that $|g(\boldsymbol{U}^k) - g^*|$ is bounded by $||\Delta \boldsymbol{U}^k_{(1)}||$. This means the sequence $||\Delta \boldsymbol{U}^{k_i}_{(1)}||$ has a lower bound $\bar{\epsilon}$. In addition, the proof of Lemma 9 employed $\|\Delta \boldsymbol{U}^k_{(1)}\| \leq \max\{1, \alpha_{\max}\}\|\boldsymbol{D}^k\|$. This leads to the existence of the lower bound of $\|\boldsymbol{D}^{k_i}\| = \bar{\epsilon}/\max\{1, \alpha_{\max}\} =: \delta$. From (3.12), we know $\left\langle \nabla g(\boldsymbol{U}^k), \boldsymbol{D}^k \right\rangle \geq \frac{\left\|\boldsymbol{D}^k\right\|^2}{\alpha_k} \geq \frac{\delta^2}{\alpha_{\max}}$. Therefore, we derive

$$g(\boldsymbol{U}^{k_i}) \geq \min_{k_i - M \leq l \leq k_i - 1} g(\boldsymbol{U}^l) + \bar{\delta},$$

where $\bar{\delta} := \gamma(\sigma\nu)_{\min} \frac{\delta^2}{\alpha_{\max}}$. Let $l(k)$ be an integer such that $k - M \leq l(k) \leq k - 1$ and $g(\boldsymbol{U}^{l(k)}) = \min_{k-M \leq l \leq k-1} g(\boldsymbol{U}^l)$. Since $k_i - k_{i-1} \leq M$, we have $g(\boldsymbol{U}^{l(k_i)}) \leq g(\boldsymbol{U}^{k_{i-1}}) \leq g^* - \epsilon$. This means $l(k_i)$ is in the sequence $\{k_1, k_2, \dots\}$. Therefore, for each $k_i > M$, there exists k_j such that $k_i - k_j \leq M$ and

$$g(\boldsymbol{U}^{k_i}) \geq g(\boldsymbol{U}^{k_j}) + \bar{\delta}.$$

Hence, if $\{k_1, k_2, \dots\}$ is an infinite sequence, we know $g(\boldsymbol{U}^{k_i}) \to \infty$ when we take $i \to \infty$, and this contradicts the existence of the optimal solution g^*. This completes the proof. □

4 Numerical experiments

In this section, we present numerical results of Algorithm 1 on a problem of the form of $(\mathcal{P})$ with synthetic data. In the second experiment that performs on a problem with block regularization, we compare Algorithm 1 with the projected gradient (PG) method proposed in Duchi et al. [6, Algorithm 3]. All experiments in this section were conducted in Matlab R2022b on a 64-bit PC with Intel Core i7-7700K CPU (4.20 GHz, 4 cores) and 16 GB RAM.

For Algorithm 1, we set the parameters $\tau = 0.5, \gamma = 10^{-3}, \beta = 0.5, \alpha_{\min} = 10^{-8}, \alpha_{\max} = 10^8$ and $M = 5$. For PG, we set the parameters $\alpha = 0.5$, $\beta = 0.5$. We take an initial point of Algorithm 1 and PG as $\boldsymbol{U}^0 = (\boldsymbol{y}^0, \boldsymbol{S}^0_1, \dots, \boldsymbol{S}^k_H) = (\boldsymbol{0}, \boldsymbol{O}, \dots, \boldsymbol{O})$ and $W = \boldsymbol{O}$, respectively. We set the stopping criterion of Algorithm 1 as $\|\Delta \boldsymbol{U}^k_{(1)}\| \leq \varepsilon$ with $\varepsilon = 10^{-12}$. The limit of the iterations is 5000 iterations, and the computation time limit is 7200 seconds.

For the projection onto the $\mathcal{B}^{\lambda_h}_{p^*_h}$, we use the direct computation when $p^*_h \in \{1, 2, \infty\}$:

$$[P_{\mathcal{B}^{\lambda}_{p^*_h}}(\boldsymbol{z})]_i = \begin{cases} z_i - (\operatorname{sign}(z_i)\min\{|z_i|, s\}) & \text{for} \quad p^*_h = 1 \\ \frac{\lambda z_i}{\max\{||\boldsymbol{z}||_2, \lambda\}} & \text{for} \quad p^*_h = 2 \\ \max\{-\lambda, \min\{\lambda, z_i\}\} & \text{for} \quad p^*_h = \infty. \end{cases}$$

where s for the case $p^*_h = 1$ is a real number that satisfies $\sum_{i=1}^N \max\{0, |x_i| - s\} = \lambda$. For $p^*_h \notin \{1, 2, \infty\}$, we use the Newton method implemented in `bpdq_proj_lpball`[1].

To evaluate the performance of the algorithm, we use the relative gap defined in [14] as

$$\text{Gap} = \frac{|P - D|}{\max\{1, (|P| + |D|)/2\}},$$

where P and D are the output values of primal and dual objective functions, respectively.

[1] https://wiki.epfl.ch/bpdq/documents/help/bpdq_toolbox/common/bpdq_proj_lpball.html

4.1 Log-likelihood minimization problem with ℓ_p-norm extension

In this experiment, we evaluate the efficiency of the proposed method by solving the following synthetic problem:

$$\begin{aligned} \min_{\boldsymbol{X}\in\mathbb{S}^n} \quad & \boldsymbol{C}\bullet\boldsymbol{X} - \mu\log\det\boldsymbol{X} + \sum_{h=1}^{H}\lambda_h\left(\sum_{1\le i<j\le n}|X_{ij}|^{p_h}\right)^{\frac{1}{p_h}} \\ \text{s.t.} \quad & X_{ij}=0\ \forall(i,j)\in\Omega, \boldsymbol{X}\succ\boldsymbol{O}. \end{aligned}$$

Here, $\Omega\subset\{(i,j):1\le i\le j\le n\}$ is a set that defines the linear constraints. We introduce a linear map $vect:\mathbb{S}^n\to\mathbb{R}^{\frac{n(n-1)}{2}}$ that reshapes a $n\times n$ symmetric matrix into a $\frac{n(n-1)}{2}$-dimensional vector by stacking the column vectors in the upper triangular part of the matrix. We can rewrite the problem into the form of $(\mathcal{P})$ as follows:

$$\begin{aligned} \min_{\boldsymbol{X}\in\mathbb{S}^n} \quad & f(\boldsymbol{X}) := \boldsymbol{C}\bullet\boldsymbol{X} - \mu\log\det\boldsymbol{X} + \sum_{h=1}^{H}\lambda_h||vect(\boldsymbol{X})||_{p_h} \\ \text{s.t.} \quad & X_{ij}=0\ \forall(i,j)\in\Omega, \boldsymbol{X}\succ\boldsymbol{O}. \end{aligned}$$

To generate the input matrix $\boldsymbol{C}$ and the set Ω, we used the same procedure in [13]. Firstly, we randomly generated a $n\times n$ sparse positive matrix Σ^{-1} with a density parameter $\sigma=0.1$, and constructed the covariance matrix $\boldsymbol{C}\in\mathbb{S}^n$ from $\max\{2n,2000\}$ samples of the multivariate Gaussian distribution $\mathcal{N}(0,\Sigma)$. We made a set $\Omega^{'}:=\{(i,j)\mid\Sigma^{-1}_{ij}=0,|i-j|>5,1\le i<j\le n\}$, and randomly selected a half of entries in $\Omega^{'}$ to be Ω.

Firstly, we conducted experiments on $H=1$, which corresponds to the following problem:

$$\begin{aligned} \min_{\boldsymbol{X}\in\mathbb{S}^n} \quad & \boldsymbol{C}\bullet\boldsymbol{X} - \mu\log\det\boldsymbol{X} + \lambda_1||vect(\boldsymbol{X})||_{p_1} \\ \text{s.t.} \quad & X_{ij}=0\ \forall(i,j)\in\Omega, \boldsymbol{X}\succ\boldsymbol{O}. \end{aligned} \tag{4.1}$$

Here, we set $\lambda_1=0.001\cdot n^{1-\frac{1}{p_1}}$.

Table 1 shows the numerical results of problem (4.1). The first column is the size n and the number of constraints $|\Omega|$. The second, third, and fourth columns are the number of iterations, the computation time in seconds, and the relative gap for Algorithm 1. The other six columns are for the different values of p_1.

Table 1: Numerical results on ℓ_{p_1}-norm log-likelihood minimization problem ($H=1$).

	$p_1=1$			$p_1=2$			$p_1=\infty$		
$(n,\|\Omega\|)$	Iterations	Time	Gap	Iterations	Time	Gap	Iterations	Time	Gap
$(500,56086)$	207	15.03	3.70e-7	241	17.58	7.23e-7	193	15.62	1.66e-7
$(1000,220647)$	169	57.75	2.36e-7	187	63.95	2.30e-7	153	61.40	1.46e-7
$(2000,859795)$	127	292.69	7.05e-7	202	468.78	5.75e-7	155	442.88	2.55e-7
$(4000,3311218)$	123	2104.99	5.66e-7	217	3662.50	6.02e-6	196	3300.02	3.09e-7

We can see that Algorithm 1 can solve the problem (4.1) in different values of p_1. The original DSPG method in Nakagaki et al. [13] can solve only the case of $p_1=1$, while Table 1 indicates Algorithm 1 can solve other $p_1>1$ with enough accuracy.

Table 2 reports a more detailed computation time of the projection of problem (4.1). The second column is the computation time of projection of all the iterations, the third column is the average time of projection per iteration, and the fourth column is the average of the entire computation per iteration. The other six columns are for $p_1 = 2$ and $p_1 = \infty$. We can observe from Table 2 that even the projection time for $p_1 = \infty$ is higher than others due to the complexity of the projection onto the $\ell_{p_1^*} = \ell_1$-ball, the average times per iteration of the three experiments are not much different. This is because the computation cost of each projection is much lower than the cost of Cholesky factorization in Step 2 which demands $O(n^3)$ operations.

Table 2: Projection time on ℓ_{p_1}-norm log-likelihood minimization problem ($H = 1$).

	$p_1 = 1$			$p_1 = 2$			$p_1 = \infty$		
$(n, \|\Omega\|)$	Time.Proj	Avg.Proj.	Avg.Comp.	Time.Proj	Avg.Proj.	Avg.Comp.	Time.Proj	Avg.Proj.	Avg.Comp.
$(500, 56086)$	0.27	1.31e-3	7.26e-2	0.34	1.41e-3	0.07	2.79	1.44e-2	8.09e-2
$(1000, 220647)$	0.87	5.19e-3	0.34	0.97	5.20e-3	0.34	6.94	4.54e-2	0.40
$(2000, 859795)$	2.31	1.81e-2	2.30	4.31	2.13e-2	2.32	31.1	0.20	2.57
$(4000, 3311218)$	9.45	7.69e-2	17.11	18.48	8.52e-2	18.88	165.22	0.84	16.84

Next, we did the experiments with $H = 2$:

$$\begin{aligned} \min_{\boldsymbol{X} \in \mathbb{S}^n} \quad & \boldsymbol{C} \bullet \boldsymbol{X} - \mu \log\det \boldsymbol{X} + \lambda_{p_1} ||vect(\boldsymbol{X})||_1 + \lambda_2 ||vect(\boldsymbol{X})||_{p_2} \\ \text{s.t.} \quad & X_{ij} = 0 \ \forall (i,j) \in \Omega, \boldsymbol{X} \succ \boldsymbol{O}. \end{aligned} \tag{4.2}$$

Similar to the previous problem, we set $\lambda_1 = 0.001 \cdot n^{1-\frac{1}{p_1}}$ and $\lambda_2 = 0.001 \cdot n^{1-\frac{1}{p_2}}$. This problem has different norms in the objective function so we can evaluate that Algorithm 1 can handle the problem by summating the extension structure. We mention that the existing DSPG methods cannot apply to this problem.

Table 3 reports the numerical results on problem (4.2). Similarly to the previous experiment, the computation time is almost proportional to the number of iterations, thus the average time for each iteration does not vary so much. For example, the numerical experiments at $n = 4000$ takes 20.03 seconds per iteration for the case $(p_1, p_2) = (1, 2)$, 18.06 seconds per iteration for the case $(p_1, p_2) = (1, \infty)$, and 17.54 seconds per iteration for the case $(p_1, p_2) = (2, \infty)$.

Table 3: Numerical results on ℓ_p-norm log-likelihood minimization problem ($H = 2$).

	$(p_1, p_2) = (1, 2)$			$(p_1, p_2) = (1, \infty)$			$(p_1, p_2) = (2, \infty)$		
$(n, \|\Omega\|)$	Iterations	Time	Gap	Iterations	Time	Gap	Iterations	Time	Gap
$(500, 56086)$	283	30.86	6.67e-7	202	19.36	9.25e-8	210	18.70	2.36e-7
$(1000, 220647)$	191	97.29	5.50e-7	164	78.58	2.69e-7	150	60.26	2.13e-8
$(2000, 859795)$	120	351.24	6.35e-7	124	348.99	2.45e-7	166	412.15	5.45e-7
$(4000, 3311218)$	105	2103.68	1.04e-6	76	1372.87	1.19e-6	212	3718.64	2.63e-6

We further conducted an experiment with the value p_h that requires high costs for the projection to $\ell_{p_h^*}$-ball. We examined cases with $H = 1$, $p_1 = 3/2$ and $4/3$, which need the projection on ℓ_3-ball and ℓ_4-ball respectively.

Table 4 shows the numerical results of problem (4.1) with $p_1 = 3/2$ and $4/3$ respectively. Now we observe that the computation time for the projection is very high compared to the previous

Table 4: Numerical results on $\ell_{3/2}$-norm and $\ell_{4/3}$-norm log-likelihood minimization problems.

$(n, \lvert\Omega\rvert)$	Iterations	Time	Time.Proj	Avg.Comp	Avg.Proj	Gap
			$p = \frac{3}{2}$			
$(500, 56086)$	337	95.42	72.92	0.28	0.22	8.32e-7
$(1000, 220647)$	229	292.18	207.29	1.28	0.91	8.98e-7
$(2000, 859795)$	166	953.22	563.64	5.74	3.40	1.53e-6
$(4000, 3311218)$			OOT			
			$p = \frac{4}{3}$			
$(500, 56086)$	344	147.16	124.02	0.42	0.36	2.19e-6
$(1000, 220647)$	343	606.59	493.98	1.44	1.77	1.84e-6
$(2000, 859795)$	387	3103.23	2218.91	8.01	5.73	2.66e-6
$(4000, 3311218)$			OOT			

experiments with $p_1 = 1, 2, \infty$. In all pairs of $(n, |\Omega|)$, the computation time of the projection occupies more than half of the entire computation time. This makes Algorithm 1 cannot solve the experiments with $(n, |\Omega|) = (4000, 3311218)$ in the computation time limit of 7200 seconds.

4.2 Block ℓ_∞-regularized log-likelihood minimization problem

In this experiment, we present the numerical results on the following block l_∞-regularized log-likelihood minimization problem (1.2) from Duchi et al. [6]:

$$\min_{\boldsymbol{X}\in\mathbb{S}^n} \; \boldsymbol{C} \bullet \boldsymbol{X} - \log\det \boldsymbol{X} + \sum_{h_1,h_2=1}^{H} \lambda_{h_1h_2} \max\{|X_{ij}| \big| (i,j) \in G_{h_1h_2}\}. \tag{4.3}$$

We constructed the covariance matrix $\boldsymbol{C} \in \mathbb{S}^n$ as in Section 4.1. The sets $G_{11}, \dots, G_{H_1H_2}$ are the disjoint subsets of $\{1, \dots, n\} \times \{1, \dots, n\}$ that partition the inverse covariance matrix $\boldsymbol{\Sigma}$. More precisely, to inherit the symmetry of $\boldsymbol{\Sigma} \in \mathbb{S}^n$, we first divide the set of row/column indexes $\{1, \dots, n\}$ into $\bar{G}_1, \bar{G}_2, \dots, \bar{G}_H$ and define $G_{h_1h_2} = (\bar{G}_{h_1} \times \bar{G}_{h_2}) \cup (\bar{G}_{h_2} \times \bar{G}_{h_1})$ for $1 \le h_1 \le h_2 \le H$.

We can rewrite (4.3) into the form of $(\mathcal{P})$ by defining a linear map $\mathcal{Q}_{h_1h_2}(\boldsymbol{X})$ that stacks $\{X_{ij} \big| (i,j) \in G_{h_1h_2}\}$ as a vector and taking its infinity norm (that is, $p_{h_1h_2} = \infty$). We set the penalty $\lambda_{h_1h_2} := \rho|G_{h_1h_2}|$ where ρ is a positive parameter and $|G_{h_1h_2}|$ is the cardinality of $G_{h_1h_2}$ so that $\lambda_{h_1h_2}$ is proportional to $|G_{h_1h_2}|$. In the following result, we used $\rho = 0.001$.

We compare the performance of Algorithm 1 with that of the PG method. PG uses a stopping criterion based on the KKT condition in [11]

$$\max\left\{\frac{|P - D|}{1 + |P| + |D|}, pinf, dinf\right\} \le \text{gaptol},$$

where gaptol $= 10^{-6}$. The values of $pinf$ and $dinf$ are the residuals of the constraints in the primal and dual problems, respectively, as in [11]. For the comparison, we include Algorithm 1 that also employs the same stopping criterion denoted as Algorithm 1 (KKT). We note that $pinf = 0$ holds in this experiment since the problem is unconstrained, and $dinf = 0$ holds in Algorithm 1 because the generated sequence $\{\boldsymbol{U}^k\}$ is always dual feasible. All algorithms can obtain the result with the relative gap around 10^{-6} with these settings.

Table 5: Numerical results on block ℓ_∞-regularized log-likelihood minimization problem (OOT is out of time)

	Algorithm 1			Algorithm 1 (KKT)			Projected Gradient		
(n,k)	Iterations	Time	Gap	Iterations	Time	Gap	Iterations	Time	Gap
(500,10)	35	1.35	1.37e-6	35	1.72	1.37e-6	59	5.40	1.72e-6
(1000,20)	36	6.39	3.54e-7	33	8.74	1.78e-6	50	23.21	1.76e-6
(2000,50)	42	43.26	3.08e-6	44	65.88	1.64e-6	74	180.57	1.83e-6
(4000,50)	62	325.59	6.16e-6	66	511.20	8.85e-7	161	1764.43	1.85e-6
(6000,50)	104	1602.37	3.18e-5	131	3001.14	1.58e-6		OOT	

Table 5 shows the performance comparison between the two algorithms on (4.3). Algorithm 1 works well in the large instance. We can also see that the number of iterations that Algorithm 1 takes is less than PG. The convergence efficiency of Algorithm 1 will be clearer for larger instances. We can see from the experiment with $(n, k) = (6000, 50)$ that Algorithm 1 can give the estimated solution of (4.3) in 1602 seconds, but PG cannot solve it in the computation time limit of 7200 seconds.

It was shown in [17] that the different norm constraint of (4.3) (considering max function as the ℓ_∞-norm) can be better in some cases, thus we further investigate the following synthetic problem:

$$\min_{\boldsymbol{X}\in\mathbb{S}^n} \ \boldsymbol{f}(\boldsymbol{X}) := \boldsymbol{C}\bullet\boldsymbol{X} - \log\det\boldsymbol{X} + \sum_{1\le h_1\le h_2\le H} \lambda'_{h_1h_2}\left(\sum_{(i,j)\in G_{h_1h_2}} X_{ij}^2\right)^{\frac{1}{2}}, \tag{4.4}$$

when $\lambda'_{h_1h_2} := \rho|G_{h_1h_2}|^{\frac{1}{2}}$. This problem is a variant of (4.3) which changes the *max* function to the Frobenious norm.

Table 6 shows the performance comparison between the two algorithms on (4.4). Focusing on the computation time, Algorithm 1 again performs well in this problem. In the experiment with $(n, k) = (4000, 50)$, Algorithm 1 executed in 163 seconds and reached a relative gap of 4.98×10^{-15}, while PG executed in 1038 seconds and gives the solution with a relative gap 2.65×10^{-7}. We can see that Algorithm 1 takes a shorter time and outputs a highly accurate solution than PG.

Table 6: Numerical results on (4.5)

	Algorithm 1			Algorithm 1 (KKT)			Projected Gradient		
(n,k)	Iterations	Time	Gap	Iterations	Time	Gap	Iterations	Time	Gap
(500,10)	15	0.78	2.48e-15	6	0.73	8.73e-7	6	2.01	7.65e-7
(1000,20)	16	3.11	3.07e-14	7	4.05	1.20e-6	8	11.24	5.01e-7
(2000,50)	22	20.20	3.35e-14	13	37.80	1.21e-6	22	108.16	7.86e-7
(4000,50)	31	163.00	4.98e-15	17	336.43	6.05e-7	27	1037.94	2.65e-7

4.3 Multi-task structure learning problem

In this experiment, we present the numerical results on the following multi-task structure learning problem [8, Equation (3)]:

$$\min_{\boldsymbol{X}^1,\dots,\boldsymbol{X}^K \in \mathbb{S}^n} \quad \sum_{k=1}^{K} \left(\boldsymbol{C}^k \bullet \boldsymbol{X}^k - \log\det \boldsymbol{X}^k \right) + \lambda \sum_{i,j=1}^{n} ||(X^1_{ij}, \dots, \boldsymbol{X}^k_{ij})||_\infty, \tag{4.5}$$

where $\boldsymbol{C}^1, \dots, \boldsymbol{C}^K \in \mathbb{S}^n$. Following the transformation in [17, Section 1.2], we let $\boldsymbol{C} := \mathrm{diag}(\boldsymbol{C}_1, \dots, \boldsymbol{C}_K) \in \mathbb{S}^{nK}$ and $\boldsymbol{X} = \mathrm{diag}(\boldsymbol{X}_1, \dots, \boldsymbol{X}_K) \in \mathbb{S}^{nK}$. By adding the linear constraint to the non-block diagonal elements $\boldsymbol{X}$, we can modify (4.5) into

$$\begin{aligned} \min_{\boldsymbol{X}^1,\dots,\boldsymbol{X}^K \in \mathbb{S}^n} \quad & \boldsymbol{C} \bullet \boldsymbol{X} - \log\det \boldsymbol{X} + \lambda \sum_{i,j=1}^{n} ||(X^1_{ij}, \dots, X^K_{ij})||_\infty \\ \text{s.t.} \quad & X_{ij} = 0 \ \forall (i,j) \in \Omega, \boldsymbol{X} \succ \boldsymbol{O}, \end{aligned} \tag{4.6}$$

where $\Omega := \{(i,j) \mid 1 \le i,j \le nK, |\lceil i/n \rceil - \lceil j/n \rceil| \ge 1\}$ with $\lceil x \rceil$ being the ceiling function that takes the largest integer which does not exceed x. We set the penalty $\lambda = 0.005$.

Table 7 shows the performance of Algorithm 1 on (4.6) with $K = 5, 10$, and 15. We can observe that Algorithm 1 generates accurate solutions even if we increase the value of K. This shows the ability of the Algorithm 1 to solve the problem with a complicated structure by adapting the objective function.

Table 7: Numerical results on Multi-task Structure Learning Problem

	$K = 5$			$K = 10$			$K = 15$		
n	Iterations	Time	Gap	Iterations	Time	Gap	Iterations	Time	Gap
100	39	7.27	2.44e-8	39	14.73	9.79e-9	32	29.57	1.18e-8
200	42	34.57	1.75e-8	35	79.87	5.47e-9	32	213.43	1.34e-8
300	56	114.12	4.93e-6	36	253.93	2.90e-8	34	672.73	2.99e-8
400	55	221.69	1.64e-5	40	609.38	6.29e-9	34	1552.03	1.68e-8

5 Conclusion

In this paper, we addressed the generalized log-det SDP ($\mathcal{P}$) that covers many existing optimization models and proposed Algorithm 1 based on DSPG. We show the convergence of Algorithm 1 to the optimal value under the mild assumptions (Assumption 1). We also provide the results of numerical experiments on the synthetic problem, the number of components in the extension structure (block-constraint), and their combination (multi-task structure). Algorithm 1 can obtain accurate solutions on the large instances within the acceptable computation time.

One of the future directions is to incorporate squared terms of the ℓ_p norm in the objective function like $\|\mathcal{Q}_h(\boldsymbol{X})\|^2_{p_h}$. The combination of the ℓ_1 norm and the squared ℓ_2 norm appears in statistics. Another important factor is to discuss the types of projection. In the numerical experiments, the main bottleneck was the Cholesky factorization. We still have some flexibility in choosing the projection if the cost in each iteration is at most $O(n^3)$.

Data Availability

The test instances in Section 4 were generated randomly following the steps described in these sections.

Conflict of Interest

All authors have no conflicts of interest.

Acknowledgments

The research of M. Y. was partially supported by JSPS KAKENHI (Grant Number: 21K11767).

References

[1] F. R. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9(6), 2008.

[2] A. Bagirov, N. Karmitsa, and M. M. Makela. *Introduction to nonsmooth optimization: Theory, practice and software*. Springer, 2014.

[3] J. Barzilai and J. M. Borwein. Two-point step size gradient methods. *IMA journal of numerical analysis*, 8(1):141–148, 1988.

[4] E. G. Birgin, J. M. Mart´*ı*nez, and M. Raydan. Nonmonotone spectral projected gradient methods on convex sets. *SIAM Journal on Optimization*, 10(4):1196–1211, 2000.

[5] A. P. Dempster. Covariance selection. *Biometrics*, pages 157–175, 1972.

[6] J. 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*, pages 153–160, 2008.

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

[8] J. Honorio and D. Samaras. Multi-task learning of gaussian graphical models. In *ICML*, pages 447–454, 2010.

[9] C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, and P. Ravikumar. Quic: Quadratic approximation for sparse inverse covariance estimation. *Journal of Machine Learning Research*, 15(83):2911–2947, 2014.

[10] S. L. Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[11] L. Li and K.-C. Toh. An inexact interior point method for l 1-regularized sparse covariance selection. *Mathematical Programming Computation*, 2:291–315, 2010.

[12] M. Lin, D. Sun, K.-C. Toh, and C. Wang. Estimation of sparse gaussian graphical models with hidden clustering structure. *arXiv preprint arXiv:2004.08115*, 2020.

[13] T. Nakagaki, M. Fukuda, S. Kim, and M. Yamashita. A dual spectral projected gradient method for log-determinant semidefinite problems. *Computational Optimization and Applications*, 76(1):33–68, 2020.

[14] C. Namchaisiri, T. Liu, and M. Yamashita. A new dual spectral projected gradient method for log-determinant semidefinite programming with hidden clustering structures. *arXiv preprint arXiv:2403.18284*, 2024.

[15] C. Wang. On how to solve large-scale log-determinant optimization problems. *Computational Optimization and Applications*, 64:489–511, 2016.

[16] C. Wang, D. Sun, and K.-C. Toh. Solving log-determinant optimization problems by a newton-cg primal proximal point algorithm. *SIAM Journal on Optimization*, 20(6):2994–3013, 2010.

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

[18] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006.