Efficient QUIC-Based Damped Inexact Iterative Reweighting for Sparse Inverse Covariance Estimation with Nonconvex Partly Smooth Regularization

Xiangyu Yang[1,2], Jiarui Wei[2], and Makoto Yamashita[3]

[1]School of Mathematics and Statistics, Henan University, Kaifeng 475000, China
[2]Center for Applied Mathematics of Henan Province, Henan University, Zhengzhou, 450046, China
[3]Department of Mathematical and Computing Science, Institute of Science Tokyo, Tokyo 152-8552, Japan
[1]*yangxy@henu.edu.cn*
[2]*jrwei@henu.edu.cn*
[3]*Makoto.Yamashita@comp.isct.ac.jp*

Abstract

In this paper, we study sparse inverse covariance matrix estimation incorporating partly smooth nonconvex regularizers. To solve the resulting regularized log-determinant problem, we develop DIIR-QUIC—a novel Damped Inexact Iteratively Reweighted algorithm based on QUadratic approximate Inverse Covariance (QUIC) method. Our approach generalizes the classic iteratively reweighted ℓ_1 scheme through damped fixed-point updates. A key novelty of DIIR-QUIC is an inexact termination criterion for the subproblems that permits controlled inexactness in solutions to accelerate each iteration while still guaranteeing identification of the active manifold in finitely many

steps. We establish the global convergence of DIIR-QUIC and, under the Kurdyka-Łojasiewicz property, prove Q-linear convergence of the perturbed objective values and R-linear convergence of the iterates. Extensive numerical experiments on synthetic and real-world datasets demonstrate that DIIR-QUIC outperforms existing approaches in computational efficiency and estimation accuracy.

Keywords— Inverse covariance matrix estimation, Nonconvex regularization optimization, Smooth active manifold, Damped inexact iterative reweighting

1 Introduction

Estimating the inverse covariance matrix, also known as the precision matrix, is a fundamental problem in modern multivariate statistical analysis. A key motivation for the estimation is that its zero entries directly encode conditional-independence relationships among variables. This built-in sparsity makes precision estimation a versatile tool for high-dimensional problems: it underlies linear discriminant analysis in statistical learning [1], guides optimal asset allocation in portfolio optimization [2] and facilitates gene-network reconstruction in computational biology [3]. As data dimensions grow, developing effective methods that balance statistical reliability with computational scalability has become an increasingly active area of research.

Sparsity-promoting precision estimation is especially vital when the number of variables far exceeds the sample size [1]. In the setting of Gaussian Markov Random Fields (GMRF), we usually observe m independently drawn samples $\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in $\mathbb{R}^n$ and seek to recover the precision matrix $\boldsymbol{\Sigma}^{-1}$. Enforcing a sparse structure on $\boldsymbol{\Sigma}^{-1}$ generally enhances statistical inference and interpretability in the high-dimensional regime (i.e., $m \ll n$). A zero entry in the precision matrix corresponds to a conditional independence constraint, implying no direct dependency between the associated variables. As a result, estimating a sparse precision matrix is equivalent to learning a sparse undirected graph in a GMRF.

To obtain a sparse precision matrix, one widely studied approach in high-dimensional settings is the regularized log-likelihood formulation [4, 5]. This involves maximizing the log-likelihood function over the space of positive definite matrices while incorporating a sparsity-inducing regularizer. A natural choice for enforcing sparse pattern in the solution is the ℓ_0 (quasi-)norm, defined as $\|\boldsymbol{X}\|_0 \coloneqq \sum_{ij} \mathbb{I}(X_{ij} \neq 0)$, where $\mathbb{I}(\cdot)$ denotes the Boolean indicator function that returns 1 when the condition is true and 0 otherwise. The ℓ_0-norm is particularly appealing due to its ability to induce exact sparsity. However, its combinatorial nature renders the associated optimization problem computationally intractable, as

commented in [6]. To regain tractability, convex relaxations replace $\|\boldsymbol{X}\|_0$ with its tightest convex surrogate, the ℓ_1-norm. Regularizing the log-determinant function with an ℓ_1-penalty ensures strong convexity properties and, through Lagrangian duality, often yields computationally efficient solutions [7, 8]. Despite these advantages, the ℓ_1-penalty suffers from a well-known limitation: it tends to systematically over-penalize large coefficients, introducing bias and potentially misidentifying the true sparsity pattern. This bias generally arises due to the relaxation gap between ℓ_1- and ℓ_0-norm formulations [9].

To address this issue, nonconvex regularization approaches have gained considerable attention [10]. In particular, the nonconvex regularizers such as the ℓ_p^p-norm $(0 < p < 1)$ [11], the smoothly clipped absolute deviation (SCAD) and the minimax concave penalty (MCP) [12], as well as the piecewise exponential concave approximation function and the capped ℓ_1 regularizers [6], have demonstrated superior numerical performance compared to ℓ_1-norm regularization. These methods often yield sparser models with improved prediction accuracy. Such findings highlight the need for alternative regularization strategies that achieve a better balance between sparsity and estimation accuracy, thereby motivating further exploration of nonconvex sparsity-promoting penalties.

In this paper, we focus on the following matrix optimization problem with nonconvex sparsity-promoting regularizers:

$$\begin{array}{ll} \min & F(\boldsymbol{X}) = \{f(\boldsymbol{X}) \coloneqq \operatorname{tr}(\boldsymbol{S}\boldsymbol{X}) - \log\det \boldsymbol{X}\} + \rho \left\{ \Phi(\boldsymbol{X}) \coloneqq \sum_{ij} \phi(|X_{ij}|) \right\} \\ \text{s.t.} & \boldsymbol{X} \in \mathbb{S}^n_{++}, \end{array} \tag{$\mathcal{P}$}$$

where $\boldsymbol{S} \in \mathbb{S}^n_+$ is the empirical covariance matrix, and $\rho > 0$ is a regularization parameter. Throughout this paper, we write $\mathbb{S}^n_{++}$ (respectively, $\mathbb{S}^n_+$) for the set of n-by-n symmetric positive definite (respectively, positive semidefinite) matrices. The following assumption on the function ϕ is imposed throughout.

Assumption 1.1. *Let $\phi : [0, +\infty) \to [0, +\infty)$ satisfy:*

(i) $\phi(0) = 0$, and ϕ is concave on $[0, +\infty)$.

(ii) $\phi \in \mathcal{C}^1((0, +\infty))$, $\phi'(t) \geqslant 0$ for all $t > 0$, and ϕ' is nonincreasing on $(0, +\infty)$.

(iii) For any $\underline{\delta} > 0$, ϕ' is Lipschitz on $[\underline{\delta}, +\infty)$.

(iv) Let

$$\phi'(0^+) = \lim_{t \to 0^+} \phi'(t) \in (0, +\infty] \quad \text{and} \quad \phi'(+\infty) = \lim_{t \to +\infty} \phi'(t) \geqslant 0.$$

There exists $t^* \in (0, +\infty]$ *such that*

$$\phi'(t) > 0, \quad \forall\, t \in [0, t^*),\ \phi'(t) = 0, \quad \forall\, t \in [t^*, +\infty)\ (\textit{if } t^* < +\infty).$$

Then the inverse $(\phi')^{-1} : \big(\phi'(+\infty),\, \phi'(0^+)\big) \to (0, t^*)$ *defined by*

$$(\phi')^{-1}(s) = \inf\{t > 0 \mid \phi'(t) \leqslant s\}.$$

is well-defined and continuous on its effective domain.

Under Assumption 1.1, the function Φ covers a wide range of nonconvex surrogates for the ℓ_0 norm, which counts the number of nonzero entries in a matrix. Notable examples include ℓ_p (quasi-)norm [13], SCAD [14] and MCP [15]. Table 3 in Appendix summarizes several representative instances.

On the algorithmic front, addressing the nonconvex optimization problem $(\mathcal{P})$ remains a challenging task. To our knowledge, [11] was the first to address problem $(\mathcal{P})$ with $\Psi(\boldsymbol{X}) = \sum_{i \neq j} |X_{ij}|^p$ for $0 < p < 1$ in its maximization form. The authors proposed a two-stage alternating optimization algorithm. In the first stage, they reformulated the original matrix problem into an equivalent vector problem by exploiting permutations of the matrix variables and the empirical covariance matrix. In the second stage, they applied a cyclic descent method to solve the resulting ℓ_p^p-regularized least squares problem, updating each coordinate sequentially. At each iteration, their algorithm performs updates on a single row and column of the matrix, involving three components: one column vector, a scalar and a principal submatrix. However, the algorithm lacks theoretical guarantees for global convergence—precluding any convergence rate analysis—and its numerical evaluation has been confined to relatively small problem instances (typically on the order of 100 variables). Moreover, our reproduced experiments indicate that the algorithm incurs a comparatively high computational cost and generally fails to achieve satisfactorily low stationarity residuals. These limitations restrict its practical applicability to large-scale problems.

Another line of research, while not directly tackling the same problem, nonetheless offers useful insights. Notable examples include the works of Phan et al. [6] and Wei et al. [12]. Phan et al. [6] focused on a formulation involving a nonconvex loss function of the form $f(\boldsymbol{X}) = \log\det \boldsymbol{X} + \operatorname{tr}(\boldsymbol{S}\boldsymbol{X}^{-1})$, combined with nonconvex sparsity-promoting penalties, specifically the difference-of-convex (DC) representations of the piecewise exponential concave approximation and the capped ℓ_1 regularization. The authors proposed two tailored Difference-of-Convex Algorithm (DCA) variants and established global subsequential convergence. On the other hand, Wei et al. [12] addressed related covariance estimation problems involving a loss function of the form $f(\boldsymbol{X}) = \frac{1}{2}\|\boldsymbol{\Sigma} - \boldsymbol{S}\|_F^2 - \tau \log\det \boldsymbol{X}$ ($\tau > 0$ is used in their original texts), combined with the SCAD and MCP penalties. The authors applied

a standard iteratively reweighted ℓ_1 algorithm, with each subproblem handled by an inexact proximal gradient method and backtracking line-search. However, it is crucial to note that their statistical convergence—i.e., the consistency guarantees for the sequence of covariance estimators—and their iteration-complexity bounds for the inner subproblem solver are derived under the oracle assumption that the true support of the covariance matrix is known a priori. In addition, the global convergence of the generated iterates to a stationary point of the original nonconvex problem remains unestablished. Furthermore, the algorithms in both of these works cannot accommodate nonconvex regularizers such as the nonconvex ℓ_p^p norm, since its derivative is unbounded at the origin.

We propose to estimate large-scale sparse covariance matrices by solving the log-determinant program ($\mathcal{P}$) with more general nonconvex and nonsmooth regularizers. To this end, we develop an algorithmic framework that builds on efficient solvers for the weighted ℓ_1-regularized log-determinant subproblem. Our framework generalizes the iteratively reweighted ℓ_1 (IRℓ_1) schemes in [16]: at each iteration, we first apply a smoothing and reweighting step to define a weighted ℓ_1-regularized log-determinant subproblem and compute an intermediate solution. We then employ a damped update operator, as in [17, 18], to generate the next iterate.

While the underlying idea is simple and natural, two key challenges arise in practice. First, each weighted ℓ_1 subproblem can only be solved approximately, and it is unclear how much inexactness in the subproblem solution can be allowed to guarantee the global convergence of the method. Second, even if we assume the global convergence properties of the damped IRℓ_1 framework are guaranteed, it is also unclear whether the algorithm can identify the active manifold under inexact subproblem solutions. Smooth active manifolds are useful in nonsmooth optimization, as noted by the authors of [19], "*Once $\boldsymbol{\mathcal{M}}$ is identified, the nonsmoothness of the problem is largely irrelevant, since all future iterates lie on a smooth manifold along which f is smooth.*" Here, $\boldsymbol{\mathcal{M}}$ refers to a smooth active manifold.

In this paper, we address both issues in the damped IRℓ_1 framework by incorporating an inexact criterion for the subproblem solver and proving that the proposed algorithm is well-posed and converges globally. We further show that the algorithm correctly identifies the active manifold in a finite number of iterations under the proposed inexact criterion. Moreover, under the Kurdyka–Łojasiewicz (KŁ) property, we prove that the objective values converge at a Q-linear rate and the iterates uniquely converge at an R-linear rate. Numerical results confirm the effectiveness and efficiency of the proposed algorithm.

1.1 Notation and preliminaries

Throughout, let $\mathbb{N}$, $\mathbb{R}$, $\mathbb{R}_+ \coloneqq [0, +\infty)$ and $\mathbb{R}_{++} \coloneqq (0, +\infty)$ denote the sets of natural numbers, real numbers, nonnegative real numbers, and positive real numbers, respectively. Correspondingly, $\mathbb{R}^{m\times n}$ denotes the Euclidean space of m-by-n-dimensional real matrices, with inner product $\langle \boldsymbol{X}, \boldsymbol{Y}\rangle = \mathrm{tr}(\boldsymbol{X}^T\boldsymbol{Y})$. Let $\mathbb{S}^n_{>0} = \mathbb{R}^{n\times n}_{++} \cap \mathbb{S}^n$ denote the set of n-by-n-dimensional symmetric matrices with positive entries. Meanwhile, the notation $\boldsymbol{X} \succ \boldsymbol{0}$ indicates that $\boldsymbol{X}$ is positive definite, while $\boldsymbol{X} \succeq \boldsymbol{0}$ means that $\boldsymbol{X}$ is positive semidefinite. The Kronecker product of the matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ is denoted by $\boldsymbol{X}\otimes\boldsymbol{Y}$. We use $\mathcal{I}(\boldsymbol{X}) \coloneqq \{(i,j) \in [n]\times[n] \mid X_{ij} \neq 0\}$ and $\mathcal{Z}(\boldsymbol{X}) \coloneqq \{(i,j) \in [n] \times [n] \mid X_{ij} = 0\}$ to denote the index set of the nonzeros and zeros of $\boldsymbol{X}$, respectively. Given a matrix $\boldsymbol{X} \in \mathbb{S}^n$ and a positive weight matrix $\boldsymbol{W} \in \mathbb{R}^{n\times n}_+$, we define the weighted ℓ_1-norm of $\boldsymbol{X}$ as $\|\boldsymbol{X}\|_{1,\boldsymbol{W}} = \sum_{i,j=1}^n W_{ij}|X_{ij}|$. In addition, $\|\boldsymbol{X}\|_\infty \coloneqq \max\limits_{(i,j)\in[n]\times[n]} |X_{ij}|$. For any nonempty set $\Omega \subset \mathbb{R}^{n\times n}$, $\mathrm{rint}(\Omega)$ denotes its relative interior.

Throughout the paper we measure distances in $\mathbb{R}^{n\times n}$ with the Frobenius norm $\|\cdot\|_F$. In particular:

(i) Set-to-set distance. For any two nonempty sets $\bar{\Omega}, \Omega \subset \mathbb{R}^{n\times n}$, define

$$\mathrm{dist}(\bar{\Omega}, \Omega) \coloneqq \inf\{\|\boldsymbol{X} - \boldsymbol{Y}\|_F \mid \boldsymbol{X} \in \bar{\Omega}, \boldsymbol{Y} \in \Omega\}.$$

Note that here $\mathrm{dist}(\cdot,\cdot)$ is not a proper distance, since it may fail the triangle inequality.

(ii) Point-to-set distance. For any $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ and nonempty $\Omega \subset \mathbb{R}^{n\times n}$, we abbreviate $\mathrm{dist}(\boldsymbol{X}, \Omega) \coloneqq \mathrm{dist}(\{\boldsymbol{X}\}, \Omega)$.

(iii) Scalar case. When $n = 1$, this reduces to the usual absolute-value distance $\mathrm{dist}(x, y) = |x - y|,\ \forall x, y \in \mathbb{R}$.

The following proposition is useful in our analysis, and its proof can be found in [20, Appendix A.3].

Proposition 1.2 (Triangle inequality for set distances [20, Proposition 3])**.** *For any three sets $\Omega_1, \Omega_2, \Omega_3 \subseteq \mathbb{R}^{n\times n}$, it holds that*

$$\mathit{dist}(\Omega_1, \Omega_2) \leqslant \mathit{dist}_{\mathcal{H}}(\Omega_1 \mid \Omega_3) + \mathit{dist}(\Omega_3, \Omega_2), \tag{1}$$

where $\mathit{dist}_{\mathcal{H}}(\Omega_1 \mid \Omega_3) \coloneqq \sup\limits_{\boldsymbol{X}\in\Omega_3} \inf\limits_{\boldsymbol{Y}\in\Omega_1} \|\boldsymbol{Y} - \boldsymbol{X}\|_F$ refers to the Hausdorff distance between Ω_1 and Ω_3.

Before presenting the stationarity condition of $(\mathcal{P})$, we first characterize the subdifferentials of Φ in the following lemma.

Lemma 1.3 (Subgradients [21, Definition 8.3] and their relationships [21, Theorem 8.6]). *Consider* $(\mathcal{P})$. *Let* $\boldsymbol{X} \in \mathbb{S}^n_{++}$. *The following holds:*

(i) $\hat{\partial}\Phi(\boldsymbol{X}) = \partial\Phi(\boldsymbol{X}) = \partial\phi(|X_{11}|) \times \partial\phi(|X_{12}|) \times \cdots \times \partial\phi(|X_{nn}|)$ *with*

$$\partial\phi\big(|X_{ij}|\big) = \begin{cases} \{\phi'(|X_{ij}|)\}, & (i,j) \in \mathcal{I}(\boldsymbol{X}), \\ [-\phi'(0),\, \phi'(0)], & (i,j) \in \mathcal{Z}(\boldsymbol{X}) \text{ and } \lim\limits_{s\to 0^+} \phi'(s) < +\infty, \\ \mathbb{R}, & (i,j) \in \mathcal{Z}(\boldsymbol{X}) \text{ and } \lim\limits_{s\to 0^+} \phi'(s) = +\infty. \end{cases}$$

Here, $\hat{\partial}\Phi(\boldsymbol{X})$ *and* $\partial\Phi(\boldsymbol{X})$ *refer to the regular, limiting (or Mordukhovich) subdifferentials of* Φ *at* $\boldsymbol{X}$, *respectively. Both* $\partial\Phi(\boldsymbol{X})$ *and* $\hat{\partial}\Phi(\boldsymbol{X})$ *are closed sets, and* $\hat{\partial}\Phi(\boldsymbol{X})$ *is convex.*

(ii) $\hat{\partial}\Phi(\boldsymbol{X})^\infty = \partial^\infty\Phi(\boldsymbol{X}) = \partial^\infty\phi(|X_{11}|) \times \partial^\infty\phi(|X_{12}|) \times \cdots \times \partial^\infty\phi(|X_{nn}|)$ *with*

$$\partial^\infty\phi(|X_{ij}|) = \begin{cases} \{0\}, & (i,j) \in \mathcal{I}(\boldsymbol{X}), \\ \{0\}, & (i,j) \in \mathcal{Z}(\boldsymbol{X}) \text{ and } \lim\limits_{s\to 0^+} \phi'(s) < +\infty, \\ \mathbb{R}, & (i,j) \in \mathcal{Z}(\boldsymbol{X}) \text{ and } \lim\limits_{s\to 0^+} \phi'(s) = +\infty. \end{cases}$$

Here, $\partial^\infty\Phi(\boldsymbol{X})$ *refers to the horizon subdifferential of* Φ *at* $\boldsymbol{X}$, *and* $\hat{\partial}\Phi(\boldsymbol{X})^\infty$ *refers to the horizon cone of* $\hat{\partial}\Phi(\boldsymbol{X})$ *[21, Definition 3.3]. Furthermore,* $\partial^\infty\Phi(\boldsymbol{X})$ *and* $\hat{\partial}\Phi(\boldsymbol{X})^\infty$ *are closed cones, with* $\hat{\partial}\Phi(\boldsymbol{X})^\infty$ *convex.*

Consequently, by Assumption 1.1 and [21, Corollary 8.11], $\Phi(\boldsymbol{X})$ *is (subdifferentially) regular at* $\boldsymbol{X}$. *In addition, it follows from [21, Exercise 8.8] and* $f \in \mathcal{C}^1$ *that* F *is (subdifferentially) regular for all* $\boldsymbol{X} \in \mathbb{S}^n_{++}$.

In particular, given a positive weight matrix $\boldsymbol{W} \in \mathbb{R}^{n\times n}_+$, we have

$$\partial\|\boldsymbol{X}\|_{1,\boldsymbol{W}} = \left\{ \boldsymbol{G} \in \mathbb{S}^n \mid G_{ij} \begin{cases} = W_{ij}\mathrm{sgn}(X_{ij}), & \text{if } X_{ij} \neq 0, \\ \in [-W_{ij}, W_{ij}], & \text{if } X_{ij} = 0. \end{cases} \right\}. \tag{2}$$

We next provide the first-order necessary optimality condition for problem $(\mathcal{P})$.

Theorem 1.4 (Fermat's rule generalized [21, Theorem 10.1]). *Consider* $(\mathcal{P})$ *under Assumption 1.1. If* F *has a local minimum at* $\boldsymbol{X}^* \in \mathbb{S}^n_{++}$, *then*

$$-\nabla f(\boldsymbol{X}^*) \in \partial\Phi(\boldsymbol{X}^*).$$

Indeed, it holds that

$$(i,j) \in \mathcal{I}(\boldsymbol{X}^*): \quad \nabla_{ij} f(\boldsymbol{X}^*) + \rho \phi'(|X^*_{ij}|)\,\mathrm{sgn}(X^*_{ij}) = 0, \tag{3a}$$

$$(i,j) \in \mathcal{Z}(\boldsymbol{X}^*): \quad \begin{cases} |\nabla_{ij} f(\boldsymbol{X}^*)| \leqslant \rho \phi'(0), & \lim\limits_{s \to 0^+} \phi'(s) < +\infty, \\ -\nabla_{ij} f(\boldsymbol{X}^*) \in \mathbb{R}, & \lim\limits_{s \to 0^+} \phi'(s) = +\infty. \end{cases} \tag{3b}$$

Then we say that a matrix $\boldsymbol{X}^* \in \mathbb{S}^n_{++}$ is a stationary point of F if (3) *holds.*

We next recall the notion of partial smoothness, which captures the intrinsic smooth structure underlying a nonsmooth function. We begin with some elementary definitions.

Definition 1.5. Let $\Omega \subset \mathbb{R}^n$ be a nonempty convex set. The subspace parallel to the set Ω, denoted by $\mathrm{par}\,\Omega$, is defined as $\mathrm{par}\,\Omega = \mathrm{aff}\,\Omega - \boldsymbol{x},\ \forall \boldsymbol{x} \in \Omega$, where $\mathrm{aff}\,\Omega$ is the affine span of Ω.

Definition 1.6 (Partly smooth function [22, Definition 2.7])**.** Suppose that the set $\boldsymbol{\mathcal{M}} \subset \mathbb{R}^{m\times n}$ contains the point $\boldsymbol{X}$. A function $h : \mathbb{R}^{m\times n} \to \mathbb{R} \cup \{+\infty\}$ is said to be *partly smooth* at $\boldsymbol{X}$ relative to $\boldsymbol{\mathcal{M}}$ if $\boldsymbol{\mathcal{M}}$ is a manifold around $\boldsymbol{X}$ and the following four properties hold:

(i) (**Restricted smoothness**) the restriction $h|_{\boldsymbol{\mathcal{M}}}$ is smooth at $\boldsymbol{X}$;

(ii) (**Regularity**) at every point close to $\boldsymbol{X}$ in $\boldsymbol{\mathcal{M}}$, the function h is regular and has a subgradient;

(iii) (**Normals parallel to subdifferential**) $N_{\boldsymbol{\mathcal{M}}}(\boldsymbol{X}) \subset \mathrm{par}\,\partial h(\boldsymbol{X})$, where $N_{\boldsymbol{\mathcal{M}}}(\boldsymbol{X})$ denotes the normal space to $\boldsymbol{\mathcal{M}}$ at $\boldsymbol{X}$;

(iv) (**Subgradient continuity**) the subdifferential map ∂h is continuous at $\boldsymbol{X}$ relative to $\boldsymbol{\mathcal{M}}$.

We say that the function h is partly smooth relative to a set $\boldsymbol{\mathcal{M}}$ if $\boldsymbol{\mathcal{M}}$ is a manifold and h is partly smooth at each point in $\boldsymbol{\mathcal{M}}$ relative to $\boldsymbol{\mathcal{M}}$. In addition, $\boldsymbol{\mathcal{M}}$ is referred to the *active manifold* (of partial smoothness).

For a partly smooth function, the condition (iii) in Definition 1.6 reveals a "stable" property. We restate this result in the following proposition.

Proposition 1.7 (Local normal sharpness [22, Proposition 2.10])**.** *If the function $h : \mathbb{R}^{m\times n} \to \mathbb{R} \cup \{+\infty\}$ is partly smooth at the point $\boldsymbol{X}_0$ relative to the set $\boldsymbol{\mathcal{M}} \subset \mathbb{R}^{m\times n}$, then all points $\boldsymbol{X} \in \boldsymbol{\mathcal{M}}$ close to $\boldsymbol{X}_0$ satisfy the condition $N_{\boldsymbol{\mathcal{M}}}(\boldsymbol{X}) = \mathrm{par}\,\partial h(\boldsymbol{X})$.*

Definition 1.8 (Prox-regularity [23, Definition 2.1])**.** A function $h : \mathbb{R}^{m\times n} \to \mathbb{R} \cup \{+\infty\}$ is *prox-regular* at a point $\overline{\boldsymbol{X}}$ for a subgradient $\overline{\boldsymbol{V}} \in \partial h(\overline{\boldsymbol{X}})$ if h is finite at $\overline{\boldsymbol{X}}$, locally lower semi-continuous around $\overline{\boldsymbol{X}}$, and there exists $\rho > 0$ such that

$$h(\boldsymbol{X}') \geqslant h(\boldsymbol{X}) + \langle \boldsymbol{V}, \boldsymbol{X}' - \boldsymbol{X}\rangle - \frac{\rho}{2}\|\boldsymbol{X}' - \boldsymbol{X}\|_F^2$$

whenever $\boldsymbol{X}$ and $\boldsymbol{X}'$ are near $\overline{\boldsymbol{X}}$ with $h(\boldsymbol{X})$ near $h(\overline{\boldsymbol{X}})$ and $\boldsymbol{V} \in \partial h(\boldsymbol{X})$ is near $\overline{\boldsymbol{V}}$. Furthermore, h is *prox-regular* at $\overline{\boldsymbol{X}}$ if it is prox-regular at $\overline{\boldsymbol{X}}$ for every $\boldsymbol{V} \in \partial h(\overline{\boldsymbol{X}})$.

The following proposition states the basic conditions that guarantee the algorithm identifies the active manifold in a finite number of iterations.

Proposition 1.9 (Active manifold identification [23, Theorem 5.3])**.** *Let the function $h : \mathbb{R}^{m\times n} \to \mathbb{R} \cup \{+\infty\}$ be $\mathcal{C}^q$-partly smooth ($q \geqslant 2$) at the point $\overline{\boldsymbol{X}}$ relative to a smooth manifold $\boldsymbol{\mathcal{M}}$, and prox-regular there, with $\mathbf{0} \in$ rint $\partial h(\overline{\boldsymbol{X}})$. Suppose $\boldsymbol{X}^k \to \overline{\boldsymbol{X}}$ with $h(\boldsymbol{X}^k) \to h(\overline{\boldsymbol{X}})$. Then*

$$\boldsymbol{X}^k \in \boldsymbol{\mathcal{M}} \text{ for all sufficiently large } k$$

if and only if

$$dist(\mathbf{0}, \partial h(\boldsymbol{X}^k)) \to 0.$$

The KŁ property is needed in our convergence analysis. For completeness, we recall its essential components below.

Definition 1.10 (Desingularizing function)**.** [24, Section 3.1.2] Let $\wp > 0$. We say that $\varphi : [0, \wp] \to \mathbb{R}_+$ is a desingularizing function if (i) $\varphi(0) = 0$; (ii) φ is continuous on $[0, \wp]$ and of class C^1 on $(0, \wp)$; (iii) $\varphi'(s) > 0$ for all $s \in (0, \wp)$.

Definition 1.11 (KŁ property)**.** [25, Definition 3] Let $F : \mathbb{R}^{m\times n} \to \mathbb{R} \cup \{+\infty\}$ be proper lower semicontinuous. We say that F satisfies the Kurdyka-Łojasiewicz property at $\bar{\boldsymbol{X}} \in \operatorname{dom}(\partial F) \coloneqq \{\boldsymbol{X} \in \mathbb{R}^{m\times n} \mid \partial F(\boldsymbol{X}) \neq \varnothing\}$ if there exist $\wp > 0$, a neighborhood $\mathbb{U}(\bar{\boldsymbol{X}}, \rho)$ of $\bar{\boldsymbol{X}}$, and a concave desingularizing function $\varphi : [0, \wp) \to \mathbb{R}_+$, such that the Kurdyka-Łojasiewicz inequality

$$\varphi'\left(F(\boldsymbol{X}) - F(\bar{\boldsymbol{X}})\right) \operatorname{dist}\left(0, \partial F(\bar{\boldsymbol{X}})\right) \geqslant 1 \tag{4}$$

holds, for all $\boldsymbol{X}$ in the strict local upper level set

$$\operatorname{Lev}(\bar{\boldsymbol{X}}, \rho) \coloneqq \{\boldsymbol{X} \in \mathbb{U}(\bar{\boldsymbol{X}}, \rho) \mid F(\bar{\boldsymbol{X}}) < F(\boldsymbol{X}) < F(\bar{\boldsymbol{X}}) + \wp\}.$$

Typical examples of desingularizing functions include those of the form $\varphi(s) = cs^{1-\theta}$, where $c > 0$ and $\theta \in [0, 1)$ is the KL exponent. If F satisfies the KŁ property with exponent θ at any $\boldsymbol{X} \in \operatorname{dom}(\partial F)$, then call F is said to be a KŁ function with exponent θ. According to [26, Lemma 2.1], any proper lower semicontinuous function has the KŁ property with exponent $1/2$ at all noncritical points.

2 Proposed Algorithm

In this section, we present Damped Inexact Iteratively Reweighted QUIC (DIIR-QUIC) to solve $(\mathcal{P})$. The proposed DIIR-QUIC alternates between applying the modified QUIC method to a weighted ℓ_1-regularized log-determinant subproblem and updating the iterate through a damped fixed-point step that blends the subproblem solution with the previous solution estimate. Therefore, DIIR-QUIC involves three main components: (i) a smoothing technique combined with the IRℓ_1 technique is employed to locally approximate the nonconvex regularizer to yield strongly-convex subproblems; (ii) an efficient subproblem solver for computational efficiency; (iii) a practical inexact termination criterion for the subproblem designed to trade off the inner and outer loop computations. An overall statement of DIIR-QUIC is formally presented in Algorithm 1.

Algorithm 1 Proposed DIIR-QUIC for solving $(\mathcal{P})$

1: Input: $\boldsymbol{S} \in \mathbb{S}^n_{++}$, $\rho > 0$, $\alpha \in (0,1)$ and $\mu \in (0,1)$
2: (Initialization) Choose $\boldsymbol{X}^0 \in \mathbb{S}^n_{++}$ and $\boldsymbol{\mathcal{E}}^0 \in \mathbb{S}^n_{>0}$. Set $k = 0$.
3: Set $W^k_{ij} = \phi'(|X^k_{ij}| + \mathcal{E}^k_{ij}), \forall (i,j) \in [n] \times [n]$
4: Solve $\boldsymbol{Y}^k \overset{(\approx)}{\leftarrow} \text{QUIC}(\rho, \boldsymbol{X}^k, \boldsymbol{W}^k, \boldsymbol{S})$ by invoking Algorithm 2
5: Update $\boldsymbol{X}^{k+1} = (1-\alpha)\boldsymbol{X}^k + \alpha \boldsymbol{Y}^k$
6: Update $\boldsymbol{\mathcal{E}}^{k+1} = (1-\alpha)\boldsymbol{\mathcal{E}}^k + \alpha(\mu \boldsymbol{\mathcal{E}}^k)$
7: Set $k \leftarrow k+1$ and go to step 3

To develop DIIR-QUIC, we first define a function for the objective function F in $(\mathcal{P})$ that is nonsmooth but locally Lipschitz continuous. This construction is motivated by the extensive literature on smoothing approximation techniques designed to address the nonsmoothness of the function Φ, which satisfies Assumption 1.1 (see, e.g., [27, 16]). To this end, we simply add perturbations to Φ to have the following optimization problem:

$$\begin{aligned} \min \quad & F(\boldsymbol{X}, \boldsymbol{\mathcal{E}}) = f(\boldsymbol{X}) + \rho \sum_{ij} \phi(|X_{ij}| + \mathcal{E}_{ij}) \\ \text{s.t.} \quad & \boldsymbol{X} \succ \boldsymbol{0}, \end{aligned} \tag{5}$$

where the perturbation parameters $\mathcal{E}_{ij} = \mathcal{E}_{ji} > 0, \forall (i,j) \in [n] \times [n]$. By construction and Assumption 1.1, for each entry X_{ij}, it holds that

(i) If $0 \leqslant \bar{\mathcal{E}}_{ij} \leqslant \mathcal{E}_{ij}$, then $\phi(|X_{ij}| + \bar{\mathcal{E}}_{ij}) \leqslant \phi(|X_{ij}| + \mathcal{E}_{ij})$.

(ii) For any $\mathcal{E}_{ij} \geqslant 0$, $0 \leqslant \phi(|X_{ij}| + \mathcal{E}_{ij}) - \phi(|X_{ij}|) \leqslant \phi(\mathcal{E}_{ij})$.

At the kth iterate $\boldsymbol{X}^k \in \mathbb{S}^n_{++}$, we compute an intermediate point $\boldsymbol{Y}^k$ by approximately solving the weighted ℓ_1-regularized log-determinant subproblem:

$$\boldsymbol{Y}^k \approx \operatorname{argmin}_{\boldsymbol{Y}\succ\mathbf{0}} Q_k(\boldsymbol{Y}) \coloneqq f(\boldsymbol{Y}) + \rho\|\boldsymbol{Y}\|_{1,\boldsymbol{W}^k}, \tag{$\mathcal{P}_{\mathrm{sub}}$}$$

where $W^k_{ij} = \boldsymbol{W}(X^k_{ij}, \mathcal{E}^k_{ij}) = \phi'(|X^k_{ij}| + \mathcal{E}^k_{ij}) > 0$, $\forall (i,j)$, and the concavity of ϕ underpins this construction. We accept $\boldsymbol{Y}^k$ once it satisfies the inexact termination criterion:

$$\operatorname{dist}\Big(\mathbf{0}, \partial Q_k(\boldsymbol{Y}^k)\Big) \leqslant \widetilde{\beta}_k \|\boldsymbol{X}^k - \boldsymbol{Y}^k\|_F, \tag{$\mathsf{C}^{(k)}_{\mathrm{inexact}}$}$$

where the scalar $\widetilde{\beta}_k \in (0,1]$ is tied to the kth subproblem's line-search step size (see Lemma 3.3(ii)). This inexact termination condition ($\mathsf{C}^{(k)}_{\mathrm{inexact}}$) provides several computational and theoretical benefits: (i) it can be efficiently evaluated within QUIC; (ii) thanks to the term $\|\boldsymbol{X}^k - \boldsymbol{Y}^k\|$, there is no need for $\widetilde{\beta}_k$ to shrink to zero. Indeed, $\tilde{\beta}_k$ is uniformly bounded away from zero, as shown in Lemma 3.3(ii) below; (iii) this condition allows DIIR-QUIC to identify the smooth active manifold within finitely many iterations.

With $\boldsymbol{Y}^k$ at hand, we compute the new iterate as $\boldsymbol{X}^{k+1} = (1-\alpha)\boldsymbol{X}^k + \alpha\boldsymbol{Y}^k$, and update the perturbation as $\boldsymbol{\mathcal{E}}^{k+1} = (1-\alpha)\boldsymbol{\mathcal{E}}^k + \alpha(\mu\boldsymbol{\mathcal{E}}^k)$, with a damping parameter $\alpha \in (0,1)$ and constant $\mu \in (0,1)$. Motivated by [17], we can describe the iterative scheme of DIIR-QUIC consistently with the following relaxed fixed-point iteration:

$$\begin{bmatrix} \boldsymbol{X}^{k+1} \\ \boldsymbol{\mathcal{E}}^{k+1} \end{bmatrix} = \mathcal{T}\left(\begin{bmatrix} \boldsymbol{X}^{k} \\ \boldsymbol{\mathcal{E}}^{k} \end{bmatrix}\right) = \begin{bmatrix} (1-\alpha)\boldsymbol{X}^k + \alpha\mathcal{S}_{\boldsymbol{X}}(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k) \\ (1-\alpha)\boldsymbol{\mathcal{E}}^k + \alpha\mathcal{S}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k) \end{bmatrix}, \tag{6}$$

where $\mathcal{T} : \mathbb{S}^n_{++} \times \mathbb{S}^n_{\succ 0} \to \mathbb{S}^n_{++} \times \mathbb{S}^n_{\succ 0}$ denotes a relaxed mapping, $\mathcal{S}_{\boldsymbol{X}}(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k) = \boldsymbol{Y}^k \overset{(\approx)}{\leftarrow} \operatorname{argmin}_{\boldsymbol{Y}\succ\mathbf{0}} Q_k(\boldsymbol{Y})$ and $\mathcal{S}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k) = \mu\boldsymbol{\mathcal{E}}^k$. In view of this, (6) yields a natural characterization of stationarity for problem ($\mathcal{P}$), as formalized in the proposition below.

Proposition 2.1. *A pair $(\boldsymbol{X}^*, \boldsymbol{\mathcal{E}}^*) \in \mathbb{S}^n_{++} \times \mathbb{S}^n_{\succ 0}$ is a stationary point of (5) if and only if $(\boldsymbol{X}^*, \boldsymbol{\mathcal{E}}^*) = \mathcal{T}(\boldsymbol{X}^*, \boldsymbol{\mathcal{E}}^*)$. In particular, a point $\boldsymbol{X}^* \in \mathbb{S}^n_{++}$ is a stationary point of F associated with ($\mathcal{P}$) if and only if $\boldsymbol{X}^* = \mathcal{S}_{\boldsymbol{X}}(\boldsymbol{X}^*, \mathbf{0})$.*

Proof. Note that $(\boldsymbol{X}^*, \boldsymbol{\mathcal{E}}^*) = \mathcal{T}(\boldsymbol{X}^*, \boldsymbol{\mathcal{E}}^*)$ if and only if $\boldsymbol{\mathcal{E}}^* = \mathbf{0}$. Thus, proving this equivalence reduces to establishing the second statement.

Suppose that $\boldsymbol{X}^*$ is stationary for ($\mathcal{P}$). Then $-\nabla f(\boldsymbol{X}^*) \in \partial\Phi(\boldsymbol{X}^*)$ by Theorem 1.4. Consider the problem

$$\min_{\boldsymbol{Y}\succ\mathbf{0}} \left\{ Q_*(\boldsymbol{Y}) = f(\boldsymbol{Y}) + \rho \sum_{i,j} \boldsymbol{W}^*_{ij} |\boldsymbol{Y}_{ij}| \right\},$$

where $\boldsymbol{W}_{ij}^* = \boldsymbol{W}(X_{ij}^*, 0) = \phi'(|X_{ij}^*|)$. Let $\boldsymbol{Y}^* = \mathcal{S}_{\boldsymbol{X}}(\boldsymbol{X}^*, \boldsymbol{0})$ be an inexact solution satisfying

$$\operatorname{dist}(\boldsymbol{0}, \partial Q_*(\boldsymbol{Y}^*)) \leqslant \widetilde{\beta}_* \|\boldsymbol{X}^* - \boldsymbol{Y}^*\|_F,$$

where $\widetilde{\beta}_* \in (0, 1]$. Since $\boldsymbol{0} \in \partial Q_*(\boldsymbol{X}^*)$ and Q_* is strictly convex (see Lemma 3.1), $\boldsymbol{X}^*$ is its unique minimizer. Initialized at $\boldsymbol{X}^*$, any solver satisfying the above inexactness bound should return $\boldsymbol{Y}^* = \boldsymbol{X}^*$, proving the forward direction.

Conversely, if $\boldsymbol{X}^* = \mathcal{S}_{\boldsymbol{X}}(\boldsymbol{X}^*, \boldsymbol{0})$, then by the inexact termination condition, we obtain $\operatorname{dist}(\boldsymbol{0}, \partial Q_*(\boldsymbol{X}^*)) = 0$. This directly implies $\boldsymbol{X}^*$ satisfies the optimality conditions (3), completing the proof. □

2.1 Modified QUIC for subproblem solution

As shown in Algorithm 1, the proposed DIIR-QUIC method consists of approximately solving a sequence of weighted ℓ_1-norm regularized log-determinant problems ($\mathcal{P}_{\text{sub}}$), a subject that has garnered substantial attention over the past decade. To efficiently solve each subproblem, we employ QUIC—a proximal Newton-type method known for its computational efficiency. To contextualize its use, we give a concise overview of QUIC.

The foundation of QUIC lies within the framework of inexact successive quadratic approximation methods [28], which sequentially solve subproblems constructed from a quadratic approximation of the smooth function f of Q_k. We use $\{\boldsymbol{Z}^t\}$ to denote the iterate sequence generated by QUIC for solving ($\mathcal{P}_{\text{sub}}$). At a given iterate $\boldsymbol{Z}^t$, QUIC first computes an approximate Newton direction $\boldsymbol{D}^t$ using a locally quadratic approximation of f at $\boldsymbol{Z}^t$, that is,

$$\boldsymbol{D}^t \approx \operatorname{argmin}_{\boldsymbol{D} \in \mathbb{S}^n} \; J(\boldsymbol{D}; \boldsymbol{Z}^t) = \bar{f}(\boldsymbol{D}; \boldsymbol{Z}^t) + g(\boldsymbol{D}; \boldsymbol{Z}^t), \tag{7}$$

where $\bar{f}(\boldsymbol{D}; \boldsymbol{Z}^t) \coloneqq f(\boldsymbol{Z}^t) + \langle \operatorname{vec}(\nabla f(\boldsymbol{Z}^t)), \operatorname{vec}(\boldsymbol{D}) \rangle + \frac{1}{2} \langle \operatorname{vec}(\boldsymbol{D}), \nabla^2 f(\boldsymbol{Z}^t) \operatorname{vec}(\boldsymbol{D}) \rangle$ and $g(\boldsymbol{D}; \boldsymbol{Z}^t) = \rho \|\boldsymbol{Z}^t + \boldsymbol{D}\|_{1, \boldsymbol{W}^k}$. A key insight of QUIC is the development of a closed-form solution for computing the Newton direction associated with (7), achieved through a coordinate descent update rule. Starting with an initial guess $\boldsymbol{\Delta}^{(0)}$, the algorithm updates $\boldsymbol{\Delta}^{(t+1)}$ via

$$\boldsymbol{\Delta}^{(t+1)} = \boldsymbol{\Delta}^{(t)} + \eta(\boldsymbol{e}_i \boldsymbol{e}_j^T + \boldsymbol{e}_j \boldsymbol{e}_i^T), \tag{8}$$

where $\boldsymbol{\Delta}^{(t)}$ represents a given completed updates, $\boldsymbol{e}_i, \boldsymbol{e}_j$ are unit vectors in $\mathbb{R}^n$, and $\eta \in \mathbb{R}$ is the parameter to be computed. In QUIC, this is done by solving a sequence of single-variable minimization problems:

$$\min_{\eta \in \mathbb{R}} \; J(\boldsymbol{\Delta}^{(t)} + \eta(\boldsymbol{e}_i \boldsymbol{e}_j^T + \boldsymbol{e}_j \boldsymbol{e}_i^T); \boldsymbol{Z}^t). \tag{9}$$

Iterating this process under appropriate termination criteria yields the desired solution $\boldsymbol{D}^t$.

Another key insight of QUIC is a dimension-reduction strategy that leverages the optimality condition of $(\mathcal{P}_{\text{sub}})$ to compute the Newton direction using only a subset of elements of $\boldsymbol{\Delta}^{(t)}$. The indices updated during this process are

$$\mathcal{I}_{\text{free}}(\boldsymbol{Z}^t) \coloneqq \{(i,j) \in [n] \times [n] \mid \boldsymbol{W}_{ij}^k < |\nabla_{ij} f(\boldsymbol{Z}^t)| \text{ or } \boldsymbol{Z}_{ij}^t \neq 0\}, \tag{10}$$

and the remaining indices are denoted as

$$\mathcal{I}_{\text{fixed}}(\boldsymbol{Z}^t) \coloneqq \{(i,j) \in [n] \times [n] \mid \boldsymbol{W}_{ij}^k \geqslant |\nabla_{ij} f(\boldsymbol{Z}^t)| \text{ and } \boldsymbol{Z}_{ij}^t = 0\}. \tag{11}$$

Hence, QUIC in fact solves the following optimization problem for the Newton direction:

$$\boldsymbol{D}^t \approx \operatorname{argmin}_{\boldsymbol{D}: D_{ij}=0, \forall (i,j) \in \mathcal{I}_{\text{fixed}}(\boldsymbol{Z}^t)} J(\boldsymbol{D}; \boldsymbol{Z}^t) = \bar{f}(\boldsymbol{D}; \boldsymbol{Z}^t) + g(\boldsymbol{D}; \boldsymbol{Z}^t). \tag{12}$$

To solve the reduced subproblem (12) by coordinate descent, we propose to use the following inexact termination criterion. We accept the current update $\boldsymbol{D}^t = \boldsymbol{\Delta}^{(t)}$ once

$$\|\boldsymbol{\Delta}^{(t)} - \boldsymbol{\Delta}^{(t-1)}\|_1 \leqslant \epsilon \|\boldsymbol{\Delta}^{(t)}\|_1, \tag{13}$$

where $0 < \epsilon \ll 1$ is a predetermined tolerance parameter.

Remark 2.2. In the original QUIC paper, the convergence analysis—both global and local—relies on the assumption that each Newton direction $\boldsymbol{D}^t$ is computed exactly by solving (12). However, as noted by the authors (see [5, §5.2.1]), its actual implementation employs an iterative coordinate descent method that only approximately solves the subproblem in practice. In contrast, our proposed inexact termination criterion (13) formalizes this practical inexactness and enables a rigorous convergence analysis. Specifically, we prove that the resulting algorithm retains both global and local convergence guarantees (see Lemma 3.4(i) and Corollary 3.10(ii) below), thereby strengthening the theoretical foundation of QUIC under inexact subproblem solves and aligning it more closely with practical implementations. Moreover, as shown in Lemma 3.4(ii) below, this inexact condition ensures that the obtained solution satisfies the inexactness criterion (18), which is crucial for establishing the active manifold property of QUIC when subproblems (7) are solved approximately.

After determining $\boldsymbol{D}^t$, a step size $\beta_t \in (0, 1]$ is chosen to ensure both feasibility (i.e., $\boldsymbol{Z}^t + \beta_t \boldsymbol{D}^t \succ \boldsymbol{0}$) and sufficient objective decrease. To this end, QUIC employs an Armijo-type line-search to identify β_t as the largest value in the candidate set $\{\pi^0, \pi^1, \ldots\}$, where $\pi \in (0, 1)$, satisfying:

$$Q_k(\boldsymbol{Z}^t + \beta_t \boldsymbol{D}^t) \leqslant Q_k(\boldsymbol{Z}^t) + \beta_t \sigma \triangle^t, \tag{14}$$

where $\sigma \in (0, 0.5 - \epsilon_\sigma)$ for some small $\epsilon_\sigma > 0$, which depends on the chosen inexactness parameter ϵ specified in (13), and

$$\triangle^t = \langle \nabla f(\boldsymbol{Z}^t), \boldsymbol{D}^t \rangle + \rho \|\boldsymbol{Z}^t + \boldsymbol{D}^t\|_{1,\boldsymbol{W}^k} - \rho \|\boldsymbol{Z}^t\|_{1,\boldsymbol{W}^k}. \tag{15}$$

The updated iterate is then computed as $\boldsymbol{Z}^{t+1} = \boldsymbol{Z}^t + \beta_t \boldsymbol{D}^t$. We summarize the computational steps of QUIC in Algorithm 2.

Algorithm 2 The modified QUIC for solving $(\mathcal{P}_{\text{sub}})$

Require: $\boldsymbol{S}$, $\boldsymbol{W}^k$, $\boldsymbol{Z}^0 \leftarrow \boldsymbol{X}^k$, $\beta_t \leftarrow 1$, $\pi \in (0,1)$ and $t \leftarrow 0$.
  while $\text{dist}\,(\boldsymbol{0}, \partial Q_k(\boldsymbol{Z}^t)) \leqslant \beta_t \|\boldsymbol{X}^k - \boldsymbol{Z}^t\|$ (refer to $(\mathsf{C}^{(k)}_{\text{inexact}})$) is not satisfied **do**
    while (13) is not satisfied **do**
      Partition the variables into free and fixed sets according to (10) and (11).
      for $(i,j) \in \mathcal{I}_{\text{free}}(\boldsymbol{Z}^t)$ **do**
        Solve (12) for $\boldsymbol{D}^t$ by coordinate descent update (8).
      end for
    end while
    while True **do**
      if $\boldsymbol{Z}^t + \beta_t \boldsymbol{D}^t \succ \boldsymbol{0}$ **then**
        if (14) holds **then**
          break
        end if
      end if
      $\beta_t \leftarrow \pi \beta_t$
    end while
    $\boldsymbol{Z}^{t+1} \leftarrow \boldsymbol{Z}^t + \beta_t \boldsymbol{D}^t$ and $t \leftarrow t + 1$.
  end while
  Return $\boldsymbol{Y}^k \leftarrow \boldsymbol{Z}^{t+1}$.

3 Convergence

In this section, we show the convergence properties of the proposed Algorithm 1.

3.1 Well-posedness and basic properties of subproblems

We start by demonstrating that each subproblem ($\mathcal{P}_{\rm sub}$) is well-defined, and also present some useful properties of the subproblems solved by Algorithm 2.

Lemma 3.1 (Well-posed subproblem)**.** *Let Assumption 1.1 hold. Each subproblem* ($\mathcal{P}_{\rm sub}$) *is well posed in the sense that it admits a unique optimal solution whose eigenvalues are uniformly bounded away from zero and infinity.*

Proof. It follows from $\mathcal{E}_{ij}^k > 0$ that $W_{ij}^k > 0,\ \forall k,\ \forall (i,j)$; hence the proof mainly follows [29, Theorem 1]. □

Proposition 3.2. *Consider the subproblem* ($\mathcal{P}_{\rm sub}$)*. Define the sublevel set of* Q_k *with respect to the initialization* $\boldsymbol{X}^k \in \mathbb{S}_{++}^n$ *as* $Lev_{Q_k}(\boldsymbol{X}^k) = \{\boldsymbol{Z} \succ \boldsymbol{0} \mid Q_k(\boldsymbol{Z}) \leqslant Q_k(\boldsymbol{X}^k)\}$*. Then the following statements hold.*

(i) For each k*, there exist constants* $\underline{c} > 0$ *and* $\bar{c} < +\infty$ *such that* $\underline{c} \leqslant l_k < u_k \leqslant \bar{c}$*, and the level set* $Lev_{Q_k}(\boldsymbol{X}^k) \subset \chi_k = \{\boldsymbol{Z} \succ \boldsymbol{0} \mid l_k \boldsymbol{I} \preceq \boldsymbol{Z} \preceq u_k \boldsymbol{I}\}$*.*

(ii) It holds that

$$\|\nabla f(\boldsymbol{X}) - \nabla f(\boldsymbol{Y})\| \leqslant l_k^{-2} \|\boldsymbol{X} - \boldsymbol{Y}\|,\ \forall \boldsymbol{X}, \boldsymbol{Y} \in \chi_k. \tag{16}$$

In addition, $\nabla^2 f(\boldsymbol{X}) = \boldsymbol{X}^{-1} \otimes \boldsymbol{X}^{-1}$ *and*

$$u_k^{-2} \boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{X}) \preceq l_k^{-2} \boldsymbol{I},\ \forall \boldsymbol{X} \in \chi_k, \tag{17}$$

where $0 < l_k < u_k < +\infty$ *are defined in statement (i).*

Proof. Statement (i) follows from the result in [5, Lemma 2]. Statement (ii) follows from statement (i) and [7, Lemma 6(iii)]. □

We next show that the termination condition (13) implies a theoretically useful inexactness criterion, and its proof can be found in Appendix A.1.

Lemma 3.3. *Consider the subproblem* ($\mathcal{P}_{\rm sub}$)*. Let* $\{\boldsymbol{Z}^t\}_{t \geqslant 0}$ *be the sequence generated by Algorithm 2, and let* $\{\boldsymbol{\Delta}^{(t)}\}$ *be the sequence generated by the update rule* (8) *for solving the problem associated with* (12)*. Furthermore, let* $\boldsymbol{D}^{(t)}$ *be an approximate solution to* (7) *satisfying the inexact condition* (13)*. Then the following statements hold.*

(i) Suppose that the initial point $\boldsymbol{\Delta}^{(0)} = \boldsymbol{0}$*. Then the finite quantity* $C_k \coloneqq \epsilon l_k^{-2}$ *satisfies the following implication for each iteration* t*:*

$$\|\boldsymbol{\Delta}^{(t)} - \boldsymbol{\Delta}^{(t-1)}\|_1 \leqslant \epsilon \|\boldsymbol{\Delta}^{(t)}\|_1 \implies \operatorname{dist}(\boldsymbol{0}, \partial J(\boldsymbol{\Delta}^{(t)}; \boldsymbol{Z}^{(t)})) \leqslant C_k \|\boldsymbol{\Delta}^{(t)}\|_1. \tag{18}$$

(ii) Under (13). Suppose that $u_k^{-2} - \epsilon l_k^{-2} > 0$. *Then there exists a constant* $\hat{\beta} > 0$ *such that for all* $\beta_t \geqslant \hat{\beta}$, $\boldsymbol{Z}^{t+1} = \boldsymbol{Z}^t + \beta_t \boldsymbol{D}^t$ *satisfies*

$$Q_k(\boldsymbol{Z}^t) - Q_k(\boldsymbol{Z}^{t+1}) \geqslant \left(\beta_t \left(u_k^{-2} - \epsilon l_k^{-2}\right) - (2l_k^2)^{-1}\beta_t^2\right) \|\boldsymbol{Z}^t - \boldsymbol{Z}^{t+1}\|_F^2. \quad (19)$$

Furthermore, the backtracking line-search procedure in Algorithm 2 terminates in a finite number of iterations and yields a step size β_t *that admits the following lower bound:*

$$\beta_t \geqslant \hat{\beta} \coloneqq \min\left\{1, 2\pi(1-\sigma)(u_k^{-2} - \epsilon l_k^{-2})l_k^2\right\}. \quad (20)$$

Consequently, it holds that

$$\lim_{t \to +\infty} \|\boldsymbol{Z}^{t+1} - \boldsymbol{Z}^t\|_F^2 = 0. \quad (21)$$

(iii) Suppose that Algorithm 2 is run with the condition (13). *Then there exist* $\bar{\epsilon}, \bar{\delta} > 0$ *such that the line-search condition* (14) *will be satisfied with unit stepsize (i.e,* $\beta_t = 1$*) whenever* $\|\boldsymbol{Z}^t - \boldsymbol{Z}^*\| \leqslant \bar{\delta}$ *and* $C_k \|\boldsymbol{\Delta}^{(t)}\|_1 \leqslant \bar{\epsilon}$.

Next, we show the global subsequential convergence and the manifold identification property of Algorithm 2 under the inexact condition (13).

Lemma 3.4. *Consider the subproblem* $(\mathcal{P}_{\text{sub}})$. *Let* $\{\boldsymbol{Z}^t\}$ *be the sequence generated by Algorithm 2, converging to* $\boldsymbol{Z}^*$. *The following statements hold.*

*(i) (**Global subsequential convergence of QUIC under condition** (13)) Suppose that* $u_k^{-2} - \epsilon l_k^{-2} > 0$. *Then any cluster point of* $\{\boldsymbol{Z}^t\}$ *is a stationary point of* $(\mathcal{P}_{\text{sub}})$.

*(ii) (**Manifold identification of QUIC under condition** (13)) Given that the regularizer* $\|\cdot\|_{1,\boldsymbol{W}^*}$ *is partly smooth at* $\boldsymbol{Z}^*$ *relative to the manifold* $\boldsymbol{\mathcal{M}}_{QUIC}(\boldsymbol{Z}^*) = \{\boldsymbol{Z} \in \chi \mid \mathcal{I}(\boldsymbol{Z}) \subseteq \mathcal{I}(\boldsymbol{Z}^*)\}$, *suppose in addition that the nondegenerate condition*

$$\mathbf{0} \in \textit{rint}\ (\partial Q_k(\boldsymbol{Z}^*)) \quad (22)$$

holds at $\boldsymbol{Z}^*$, *and Algorithm 2 is executed under the inexactness criterion* (13). *Then*

(1.) $Q_k(\boldsymbol{Z}^t) \to Q_k(\boldsymbol{Z}^*)$.

(2.) $\textit{dist}(\mathbf{0}, \partial Q_k(\boldsymbol{Z}^t)) \to 0 \implies \boldsymbol{Z}^t \in \boldsymbol{\mathcal{M}}_{QUIC}(\boldsymbol{Z}^*)$ *for all sufficiently large* t.

Consequently, there exist constants $\bar{\epsilon}, \bar{\delta} > 0$ such that, for sufficiently large t, the following condition holds:

$$\|\boldsymbol{Z}^t - \boldsymbol{Z}^*\| \leqslant \bar{\delta},\ C_k\|\boldsymbol{\Delta}^{(t)}\|_1 \leqslant \bar{\epsilon},\ \beta_t = 1 \implies \boldsymbol{Z}^{t+1} \in \boldsymbol{\mathcal{M}}_{QUIC}(\boldsymbol{Z}^*).$$

Proof. The proof of statement (i) primarily follows from Proposition 3.2 and [28, Theorem 3.3] (applied within a monotone line-search framework (LS_1) while omitting the additional regularization function (i.e., setting $g = 0$ in their formulation)). As for statement (ii), it follows from Lemma 3.3 and [30, Lemmas 1-2 & Theorem 1]. □

3.2 Global convergence properties

This section analyzes the convergence properties of the proposed Algorithm 1. We first show the model reduction on $F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k)$ caused by $(\boldsymbol{X}^{k+1}, \boldsymbol{\mathcal{E}}^{k+1})$ with inexact subproblem solution. We now consider the outer loop of Algorithm 1. In the sequel, we define

$$l = \inf_{k\geqslant 0} l_k \text{ and } u = \sup_{k\geqslant 0} u_k, \tag{23}$$

where $0 < \underline{c} \leqslant l_k < u_k \leqslant \bar{c} < +\infty, \forall k \in \mathbb{N}$ is stated in Proposition 3.2. Correspondingly, we define $\chi = \{\boldsymbol{Z} \succ \boldsymbol{0} \mid l\boldsymbol{I} \preceq \boldsymbol{Z} \preceq u\boldsymbol{I}\}$.

We first prove some useful results in the following proposition, and its proof can be found in Appendix A.2.

Proposition 3.5 (Approximate sufficient descent property)**.** *Let Assumption 1.1 hold and let $\{(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k)\}$ be the sequence generated by Algorithm 1 with $\boldsymbol{X}^0 \in \chi$. The following assertions hold.*

(i) Suppose $l^2 \geqslant \left[\left(2\sqrt{\pi(1-\sigma)}\right)^{-1} + \epsilon\right] u^2$ under condition (13). The sequence $\{F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k)\}$ is monotonically decreasing. Indeed, there exists some constant $\widetilde{C} > 0$ such that

$$F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k) - F(\boldsymbol{X}^{k+1}, \boldsymbol{\mathcal{E}}^{k+1}) \geqslant \widetilde{C}\|\boldsymbol{X}^k - \boldsymbol{X}^{k+1}\|_F^2, \tag{24}$$

where $\widetilde{C} \coloneqq \left(\alpha^{-1}\tilde{\beta}(u^{-2} - \epsilon l^{-2}) - \frac{1}{2l^2\alpha}\right)$ with $\tilde{\beta} \coloneqq \min\left\{1, 2\pi(1-\sigma)(u^{-2} - \epsilon l^{-2})l^2\right\}$. Consequently, $\lim_{k\to+\infty} F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k)$ exists.

(ii) $\sum_{k=0}^{+\infty} \|\boldsymbol{X}^k - \boldsymbol{X}^{k+1}\|_F^2 < +\infty$, indicating $\lim_{k\to+\infty} \|\boldsymbol{X}^k - \boldsymbol{X}^{k+1}\|_F^2 = 0$ and $\lim_{k\to+\infty} \|\boldsymbol{X}^k - \boldsymbol{Y}^k\|_F^2 = 0$.

(iii) There exist constants $0 < l < u < +\infty$ such that the sequences $\{\boldsymbol{X}^k\}$ and $\{\boldsymbol{Y}^k\}$ are contained in a convex cone $\chi = \{\boldsymbol{X} \mid l\boldsymbol{I} \preceq \boldsymbol{X} \preceq u\boldsymbol{I}\}$. Additionally, the sequences $\{\boldsymbol{X}^k\}$ and $\{\boldsymbol{Y}^k\}$ are bounded. As a consequence, there exists a constant $C_{\max} > 0$ such that $\|\nabla f(\boldsymbol{X}^k)\|_\infty = \max_{ij} |\nabla_{ij} f(\boldsymbol{X}^k)| < C_{\max}$ for all k.

We next establish the global convergence of the proposed DIIR-QUIC.

Theorem 3.6 (Global subsequential convergence)**.** *Let Assumption 1.1 hold and let $\{(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k)\}$ be the sequence generated by Algorithm 1. Then the following statements holds.*

(i) The sequence $\{\boldsymbol{\mathcal{E}}^k\}$ converges to $\mathbf{0}$.

(ii) The set of cluster points Ω^∞ of $\{\boldsymbol{X}^k\}$ is nonempty, compact, connected, and dist$(\boldsymbol{X}^k, \Omega^\infty) \to 0$.

(iii) Any cluster point $\boldsymbol{X}^* \in \Omega^\infty$ is a stationary point of F.

(iv) The sequence $\{F(\boldsymbol{X}^k)\}$ is convergent. Moreover, the objective function F is constant on Ω^∞, and hence $\xi \coloneqq F(\boldsymbol{X}^*) = \lim_{k\to+\infty} F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k)$.

Proof. (i) This statement directly follows from the definition of Algorithm 1.

(ii) Proposition 3.5(iii) implies that Ω^∞ is nonempty. On the other hand, Proposition 3.5(ii)-(iii) leads to the desired results by combining [31, Corollary 2.7 (Ostrowski)]. This establishes statement (ii).

(iii) Let $\boldsymbol{X}^*$ be a cluster point of $\{\boldsymbol{X}^k\}$. Consider a subsequence $\mathscr{S}$ such that $\{\boldsymbol{X}^k\} \overset{\mathscr{S}}{\to} \boldsymbol{X}^*$. It follows from $\boldsymbol{Y}^k = \frac{1}{\alpha}(\boldsymbol{X}^{k+1} - \boldsymbol{X}^k) + \boldsymbol{X}^k$ that $\{\boldsymbol{Y}^k\} \overset{\mathscr{S}}{\to} \boldsymbol{X}^*$ by Proposition 3.5(ii), which further implies that $\operatorname{sgn}(Y_{ij}^k) = \operatorname{sgn}(X_{ij}^*), \forall (i,j) \in \mathcal{I}(\boldsymbol{X}^*)$. Let $\mathcal{R}^k \in \partial Q_k(\boldsymbol{Y}^k)$ be the minimum-norm subgradient [5, Definition 6] defined as follows:

$$\mathcal{R}_{ij}^k \;=\; \begin{cases} \nabla_{ij} f(\boldsymbol{Y}^k) \;+\; \rho W_{ij}^k, & \text{if } Y_{ij}^k > 0, \\ \nabla_{ij} f(\boldsymbol{Y}^k) \;-\; \rho W_{ij}^k, & \text{if } Y_{ij}^k < 0, \\ \operatorname{sgn}\big(\nabla_{ij} f(\boldsymbol{Y}^k)\big)\, \max\Big(\big|\nabla_{ij} f(\boldsymbol{Y}^k)\big| \;-\; \rho W_{ij}^k,\, 0\Big), & \text{if } Y_{ij}^k = 0. \end{cases} \tag{25}$$

We then know from the first-order optimality condition of $(\mathcal{P}_{\text{sub}})$ that

$$\mathcal{R}_{ij}^k \in \nabla_{ij} f(\boldsymbol{Y}^k) + \rho \boldsymbol{W}(X_{ij}^k, \mathcal{E}_{ij}^k) \partial |Y_{ij}^k|, \;\; \forall (i,j) \in [n] \times [n]. \tag{26}$$

Consider first that $(i,j) \in \mathcal{I}(\boldsymbol{X}^*)$. By [21, Proposition 8.7] and [32, Proposition 2.1.5], we know that, for sufficiently large $k \in \mathbb{N}$, there exists $\xi_{ij}^* \in \partial |X_{ij}^*|$ such that

$$0 = \nabla_{ij} f(\boldsymbol{X}^*) + \rho \boldsymbol{W}(X_{ij}^*, 0) \xi_{ij}^* \tag{27}$$

holds since $f \in \mathcal{C}^1$, $Y_{ij}^k \overset{\mathscr{S}}{\to} X_{ij}^*$, $\mathcal{E}_{ij}^k \overset{\mathscr{S}}{\to} 0$ and $\mathcal{R}_{ij}^k \overset{\mathscr{S}}{\to} 0$, $\forall (i,j) \in \mathcal{I}(\boldsymbol{X}^*)$. Thus, (3a) is satisfied at $X_{ij}^*, \forall (i,j) \in \mathcal{I}(\boldsymbol{X}^*)$. On the other hand, for $(i,j) \in \mathcal{Z}(\boldsymbol{X}^*)$ and $\lim_{t \to 0^+} \phi'(t) < +\infty$, we have from (26) that

$$0 \in \nabla_{ij} f(\boldsymbol{X}^*) + \rho \boldsymbol{W}(0,0) \partial |0| \iff |\nabla_{ij} f(\boldsymbol{X}^*)| \leqslant \rho \phi'(0). \tag{28}$$

Hence, $\boldsymbol{X}^* \in \Omega^\infty$ is stationary for F.

(iv) The convergence of $\{F(\boldsymbol{X}^k)\}$ is a direct consequence of Proposition 3.5(i)-(ii) and statement (i) in this theorem. On the other hand,

$$F(\boldsymbol{X}^*, \boldsymbol{\mathcal{E}}^*) = \lim_{k \to +\infty} f(\boldsymbol{X}^k) + \tilde{\Phi}(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k) \overset{(a)}{=} f(\boldsymbol{X}^*) + \Phi(\boldsymbol{X}^*) = F(\boldsymbol{X}^*), \tag{29}$$

where (a) holds by Assumption 1.1 and statement (iii) and $f \in \mathcal{C}^1$. The proof is complete. □

The following non-degenerate condition is assumed to hold throughout.

Assumption 3.7. *Under Assumption 1.1, let $\boldsymbol{X}^* \in \mathbb{S}_{++}^n$ be any stationary point of problem $(\mathcal{P})$. We assume in addition that*

(i) ***(Nondegenerate condition)****:*

$$\mathbf{0} \in \mathit{rint}\,(\partial F(\boldsymbol{X}^*)) = \nabla f(\boldsymbol{X}^*) + \mathit{rint}\,(\partial \Phi(\boldsymbol{X}^*))\,. \tag{30}$$

(ii) ***(No active-kink condition)****: Let $\mathcal{K} = \{t > 0 \mid \phi' \text{ is not } \mathcal{C}^1 \text{ at } t\}$. Then for each (i,j) with $X_{ij}^* \neq 0$, $|X_{ij}^*| \notin \mathcal{K}$.*

The following lemma establishes the conditions to guarantee manifold identification in our setting, and its proof can be found in Appendix A.3.

Lemma 3.8 (F in $(\mathcal{P})$ admits an active manifold)**.** *Let Assumptions 1.1-3.7 hold and let $\{(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k)\}$ be the sequence generated by Algorithm 1 converging to $(\boldsymbol{X}^*, \boldsymbol{\mathcal{E}}^*)$. The following statements hold.*

(i) Function Φ is partly smooth at the point $\boldsymbol{X}^*$ relative to the manifold $\mathcal{M}(\boldsymbol{X}^*) \coloneqq \{\boldsymbol{X} \in \mathbb{S}_{++}^n \mid sgn(\boldsymbol{X}_{\mathcal{I}(\boldsymbol{X}^*)}) = sgn(\boldsymbol{X}^*_{\mathcal{I}(\boldsymbol{X}^*)}), \boldsymbol{X}_{\mathcal{Z}(\boldsymbol{X}^*)} = \mathbf{0}\}$, and prox-regular there. Consequently, F is also partly smooth and prox-regular at $\boldsymbol{X}^*$ relative to $\boldsymbol{\mathcal{M}}(\boldsymbol{X}^*)$.

(ii) The following implication holds:

$$dist(\mathbf{0}, \partial F(\boldsymbol{X}^k)) \to 0 \iff \boldsymbol{X}^k \in \mathcal{M}(\boldsymbol{X}^*) \text{ for all sufficiently large } k. \tag{31}$$

Next, we claim that the proposed DIIR-QUIC identifies the active manifold in a finite number of iterations.

Theorem 3.9 (Finite identification of DIIR-QUIC)**.** *Let Assumptions 1.1-3.7 hold. Consider a point $\boldsymbol{X}^*$ that satisfies the non-degeneracy condition* (30)*. If Algorithm 1 is executed with the inexact stopping conditions* ($\mathsf{C}^{(k)}_{\text{inexact}}$)*, then there exist constants $\tilde{\tau}, \tilde{\kappa}, \tilde{\epsilon} > 0$ such that for any $\|\boldsymbol{X}^k - \boldsymbol{X}^*\| \leqslant \tilde{\tau}$, $\|\boldsymbol{X}^k - \boldsymbol{Y}^k\| \leqslant \tilde{\kappa}$, and $\|\boldsymbol{\mathcal{E}}^k\| \leqslant \tilde{\epsilon}$, it follows that $\boldsymbol{X}^{k+1} \in \boldsymbol{\mathcal{M}}(\boldsymbol{X}^*)$.*

Proof. We prove this by contradiction. Suppose the conclusion is false; then there exists a subsequence $\mathscr{S} \subset \mathbb{N}$ such that $\boldsymbol{X}^k \to \boldsymbol{X}^*$, $\boldsymbol{\mathcal{E}}^k \to \boldsymbol{0}$, and $\boldsymbol{X}^k - \boldsymbol{Y}^k \to \boldsymbol{0}$ as $k \in \mathscr{S} \to \infty$, yet $\boldsymbol{X}^{k+1} \notin \boldsymbol{\mathcal{M}}$ for all $k \in \mathscr{S}$.

By Proposition 3.5(ii) and standard variational analysis results [21, Proposition 8.7], [32, Proposition 2.1.5], we obtain from (25) and (26) that for sufficiently large $k \in \mathscr{S}$, it holds that

$$\left(\mathcal{R}^k_{ij} - \nabla_{ij} f(\boldsymbol{Y}^k)\right) \in \rho W(X^k_{ij}, \mathcal{E}^k_{ij}) \partial |X^k_{ij}|. \tag{32}$$

Then, we derive

$$\begin{aligned}
\operatorname{dist}(\boldsymbol{0}, \partial F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k)) &= \operatorname{dist}(-\nabla f(\boldsymbol{X}^k), \rho \partial \Phi(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k)) \\
&= \inf_{\mathcal{L}^k \in \rho \partial \Phi(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k)} \| - \nabla f(\boldsymbol{X}^k) - \mathcal{L}^k \|_F \\
&\overset{\text{Eq.(32)}}{\leqslant} \| - \nabla f(\boldsymbol{X}^k) - (\mathcal{R}^k - \nabla f(\boldsymbol{Y}^k)) \|_F \\
&\leqslant \|\nabla f(\boldsymbol{X}^k) - \nabla f(\boldsymbol{Y}^k)\|_F + \|\mathcal{R}^k\|_F \leqslant l^{-2} \|\boldsymbol{X}^k - \boldsymbol{Y}^k\|_F + \|\mathcal{R}^k\|_F.
\end{aligned} \tag{33}$$

On the other hand, we know that

$$\begin{aligned}
&\operatorname{dist}(\partial F(\boldsymbol{X}^k), \partial F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k)) = \operatorname{dist}(\rho \partial \Phi(\boldsymbol{X}^k), \rho \partial \Phi(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k)) \\
\leqslant\ & \rho \sum_{(i,j) \in \mathcal{I}(\boldsymbol{X}^k)} \operatorname{dist}(\phi'(|X^k_{ij}|), \phi'(|X^k_{ij}| + \mathcal{E}^k_{ij})) \\
&+ \rho \sum_{(i,j) \in \mathcal{Z}(\boldsymbol{X}^k)} \operatorname{dist}(\partial \phi(0), \phi'(\mathcal{E}^k_{ij})) \\
\leqslant\ & \rho \sum_{(i,j) \in \mathcal{I}^k} |\phi'(|X^k_{ij}|) - \phi'(|X^k_{ij}| + \mathcal{E}^k_{ij})| + \rho \sum_{(i,j) \in \mathcal{Z}(\boldsymbol{X}^k)} \inf_{\xi \in \partial \phi(0)} |\xi - \phi'(\mathcal{E}^k_{ij})| \\
\leqslant\ & \rho \sum_{(i,j) \in \mathcal{I}^k} L_{\phi'} \mathcal{E}^k_{ij} + \rho \sum_{(i,j) \in \mathcal{Z}(\boldsymbol{X}^k)} \inf_{\xi \in \partial \phi(0)} |\xi - \phi'(\mathcal{E}^k_{ij})| \\
=\ & \rho L_{\phi'} \|\boldsymbol{\mathcal{E}}^k_{\mathcal{I}^k}\|_1 + \rho \sum_{(i,j) \in \mathcal{Z}(\boldsymbol{X}^k)} \inf_{\xi \in \partial \phi(0)} |\xi - \phi'(\mathcal{E}^k_{ij})|.
\end{aligned} \tag{34}$$

Therefore, combining the above bounds, we obtain

$$
\begin{aligned}
&\operatorname{dist}(\mathbf{0}, \partial F(\boldsymbol{X}^k)) \\
\overset{\text{Eq.(1)}}{\leqslant}\ & \operatorname{dist}(\mathbf{0}, \partial F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k)) + \sup_{\mathcal{U}^k \in \partial F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k)} \inf_{\mathcal{V}^k \in \partial F(\boldsymbol{X}^k)} \|\mathcal{U}^k - \mathcal{V}^k\|_F \\
\overset{\text{Eq.(33)}}{\leqslant}\ & l^{-2}\|\boldsymbol{X}^k - \boldsymbol{Y}^k\|_F + \|\mathcal{R}^k\|_F + \sup_{\mathcal{U}^k \in \partial F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k)} \inf_{\mathcal{V}^k \in \partial F(\boldsymbol{X}^k)} \|\mathcal{U}^k - \mathcal{V}^k\|_F \\
\overset{\text{Eq.(34)}}{\leqslant}\ & l^{-2}\|\boldsymbol{X}^k - \boldsymbol{Y}^k\|_F + \|\mathcal{R}^k\|_F + \rho L_{\phi'} \|\boldsymbol{\mathcal{E}}^k_{\mathcal{I}^k}\|_1 + \rho \sum_{(i,j) \in \mathcal{Z}(\boldsymbol{X}^k)} \inf_{\xi \in \partial \phi(0)} |\xi - \phi'(\mathcal{E}^k_{ij})|.
\end{aligned}
$$

Since $\|\boldsymbol{X}^k - \boldsymbol{Y}^k\| \to 0$ and $\|\mathcal{R}^k\| \to 0$ by Proposition 3.5(ii), and since Theorem 3.6(i) ensures that $\|\boldsymbol{\mathcal{E}}^k_{\mathcal{I}(\boldsymbol{X}^k)}\|_1 \to 0$ and $\inf_{\xi \in \partial\phi(0)} |\xi - \phi'(\mathcal{E}^k_{ij})| \to 0$, we conclude that $\operatorname{dist}(\mathbf{0}, \partial F(\boldsymbol{X}^k)) \to 0$, which indicates that $\boldsymbol{X}^{k+1} \in \boldsymbol{\mathcal{M}}$ for all sufficiently large k by Lemma 3.8. This contradicts our assumption, and therefore, the conclusion holds. □

Corollary 3.10. *Let Assumptions 1.1-3.7 hold. Let $\{(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k)\}$ be the sequence generated by Algorithm 1. The following statements hold.*

*(i) (**Uniform lower bounds for nonzeros**) For any $(i,j) \in \mathcal{I}(\boldsymbol{X}^*)$, there exists a $\delta_k > 0$ and an index $K \in \mathbb{N}$ such that $X^k_{ij} \geqslant \delta_k > 0$ for all $k > K$. Indeed, it holds that*

$$
|X^k_{ij}| \geqslant \delta_k \coloneqq (\phi')^{-1}\left(\frac{C_{\max} + \mathcal{R}^k_{ij}}{\rho}\right) - \mathcal{E}^k_{ij} > 0, \forall k > K, \tag{35}
$$

where $\mathcal{R}^k_{ij}$ is defined in (25). Or equivalently, the corresponding weights for any $(i,j) \in \mathcal{I}(\boldsymbol{X}^*)$ are bounded, i.e.,

$$
0 < W^k_{ij} \leqslant \frac{C_{\max} + \mathcal{R}^k_{ij}}{\rho}, \forall k > K.
$$

As a consequence, it holds for any $(i,j) \in \mathcal{I}(\boldsymbol{X}^*)$ that

$$
|X^*_{ij}| \geqslant (\phi')^{-1}\left(\frac{C_{\max}}{\rho}\right) > 0. \tag{36}
$$

*(ii) (**Accelerating subproblem solution**) There exists an index $K \in \mathbb{N}$ such that for all $k > K$, Algorithm 2 solves the following optimization problem:*

$$
\min_{\boldsymbol{Y} > \mathbf{0}}\ f(\boldsymbol{Y}) + \rho \sum_{(i,j) \in \mathcal{I}(\boldsymbol{X}^k)} \boldsymbol{W}(X^k_{ij}, 0)\phi(\operatorname{sgn}(\boldsymbol{X}^k)Y_{ij}), \tag{37}
$$

which has the same optimum of $(\mathcal{P}_{\mathrm{sub}})$. *Consequently, QUIC is equivalent to the pure Newton method.*

Proof. (i) Let $(i,j) \in \mathcal{I}(\boldsymbol{X}^*)$ and $K \in \mathbb{N}$ be the index such that Theorem 3.9 holds. We first claim that $\boldsymbol{W}(X_{ij}^k, \mathcal{E}_{ij}^k) \leqslant \frac{C_{\max}+\mathcal{R}_{ij}^k}{\rho}$ for all $k > K$ with $(i,j) \in \mathcal{I}(\boldsymbol{X}^*)$. Seeking a contradiction, assume that $\boldsymbol{W}(X_{ij}^k, \mathcal{E}_{ij}^k) > \frac{C_{\max}+\mathcal{R}_{ij}^k}{\rho}$ for some $k > K$ with $(i,j) \in \mathcal{I}(\boldsymbol{X}^*)$. Now consider the first-order optimality condition of the kth subproblem $(\mathcal{P}_{\mathrm{sub}})$. By (26), it holds that

$$|\nabla_{ij} f(\boldsymbol{Y}^k)| = |\rho \boldsymbol{W}(X_{ij}^k, \mathcal{E}_{ij}^k) - \mathcal{R}_{ij}^k| > \left| \rho \frac{C_{\max} + \mathcal{R}_{ij}^k}{\rho} - \mathcal{R}_{ij}^k \right| = C_{\max}, \tag{38}$$

which contradicts Proposition 3.5(iii). Therefore, we know that for any $(i,j) \in \mathcal{I}(\boldsymbol{X}^*)$, $\boldsymbol{W}(X_{ij}^k, \mathcal{E}_{ij}^k) \leqslant \frac{C_{\max}+\mathcal{R}_{ij}^k}{\rho}$, which is equivalent to (35). On the other hand, since $\mathcal{R}^k \in \partial Q_k(\boldsymbol{Y}^k)$, it follows from $(\mathrm{C}_{\mathrm{inexact}}^{(k)})$ and Proposition 3.5(ii) that $\mathcal{R}_{ij} \to 0$. Moreover, Theorem 3.6(i)-(ii) and Theorem 3.9 guarantees that $\mathcal{E}_{ij}^k \to 0$. Hence, continuity of $(\phi')^{-1}$ indicates that (35) implies (36). Moreover, Theorem 3.9, Lemma 3.3(iii), Proposition 3.5(ii), together with the arguments in [5, Theorem 16], establish statement (ii). This completes the proof. □

3.3 Analysis Under the KŁ Property

In this subsection, assuming the KŁ property and drawing on techniques from [33], we demonstrate that the sequence of relaxed objective values $\{F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k)\}$ converges at a Q-linear rate. Moreover, we establish the global convergence of the iterates $\{\boldsymbol{X}^k\}$ and prove that they converge at an R-linear rate.

Theorem 3.11. *Suppose that Assumptions 1.1-3.7 hold and that* $F(\boldsymbol{X}, \boldsymbol{\mathcal{E}})$ *is a KŁ function with exponent* $1/2$. *Let* $\{(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k)\}$ *be the sequence generated by Algorithm 1 converging to some limit point* $(\boldsymbol{X}^*, \boldsymbol{0}) \in \Gamma^\infty$. *Then* $\{F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k)\}$ *converges Q-linearly to* $F(\boldsymbol{X}^*, \boldsymbol{0})$

Proof. We first note that if there exists an index $k_0 \in \mathbb{N}$ such that $F(\boldsymbol{X}^{k_0}, \boldsymbol{\mathcal{E}}^{k_0}) = F(\boldsymbol{X}^{k_0+1}, \boldsymbol{\mathcal{E}}^{k_0+1})$, then Proposition 3.5(i) implies $\boldsymbol{X}^{k_0} = \boldsymbol{X}^{k_0+1}$, and hence $\boldsymbol{\mathcal{E}}^{k_0} = \boldsymbol{\mathcal{E}}^{k_0+1}$. In view of Proposition 2.1, the sequence $\{\boldsymbol{X}^k\}$ then converges to a stationary point in a finite number of iterations. Hence, we assume that $F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k) > F(\boldsymbol{X}^{k+1}, \boldsymbol{\mathcal{E}}^{k+1})$ for all $k \in \mathbb{N}$ in the subsequent proof.

Suppose that $F(\boldsymbol{X}, \boldsymbol{\mathcal{E}})$ satisfies the KŁ property with exponent $1/2$. Then, by [25, Lemma 6] (in conjunction with Theorem 3.6(ii) and (iv)), there exist constants

$\varepsilon > 0$ and $\vartheta > 0$ such that the following holds: for every cluster point $\hat{\boldsymbol{X}} \in \Omega^\infty$ and for every $(\boldsymbol{X}, \boldsymbol{\mathcal{E}}) \in \mathbb{S}^n_{++} \times \mathbb{S}^n_{\succ 0}$ satisfying

$$\operatorname{dist}((\boldsymbol{X}, \boldsymbol{\mathcal{E}}), \Gamma^\infty) < \varepsilon \text{ and } F(\hat{\boldsymbol{X}}, \boldsymbol{0}) < F(\boldsymbol{X}, \boldsymbol{\mathcal{E}}) < F(\hat{\boldsymbol{X}}, \boldsymbol{0}) + \vartheta,$$

the inequality

$$\varphi'(F(\boldsymbol{X}, \boldsymbol{\mathcal{E}}) - F(\hat{\boldsymbol{X}}, \boldsymbol{0})) \cdot \operatorname{dist}(\boldsymbol{0}, \partial F(\boldsymbol{X}, \boldsymbol{\mathcal{E}})) \geqslant 1 \tag{39}$$

holds. Here, the desingularizing function is defined by $\varphi(t) = ct^{1/2}$ with $c > 0$, so that its derivative is given by $\varphi'(t) = \frac{c}{2} t^{-1/2}$.

Let $(\boldsymbol{X}^*, \boldsymbol{0})$ be a cluster point of the sequence $\{(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k)\}$. By Theorem 3.6(i)-(ii) and (iv), there exists an index $K \in \mathbb{N}$ such that for all $k \geqslant K$ we have $(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k) \in \{(\boldsymbol{X}, \boldsymbol{\mathcal{E}}) \in \mathbb{S}^n_{++} \times \mathbb{S}^n_{\succ 0} \mid \operatorname{dist}((\boldsymbol{X}, \boldsymbol{\mathcal{E}}), \Gamma^\infty) < \epsilon\} \cap \{(\boldsymbol{X}, \boldsymbol{\mathcal{E}}) \mid F(\boldsymbol{X}^*, \boldsymbol{0}) < F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k) < F(\boldsymbol{X}^*, \boldsymbol{0}) + \vartheta\}$ with $\vartheta > 0$ and . For any such k, the KŁ inequality yields

$$\frac{c}{2}(F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k) - F(\boldsymbol{X}^*, \boldsymbol{0}))^{-1/2} \cdot \operatorname{dist}(\boldsymbol{0}, \partial F(\boldsymbol{X}^k; \boldsymbol{\mathcal{E}}^k)) \geqslant 1. \tag{40}$$

Introducing the notion $\tilde{\Delta}_k = F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k) - F(\boldsymbol{X}^*, \boldsymbol{0})$, (40) can be rewritten as

$$\operatorname{dist}(\boldsymbol{0}, \partial F(\boldsymbol{X}^k; \boldsymbol{\mathcal{E}}^k)) \geqslant \frac{2}{c} \tilde{\Delta}_k^{1/2}.$$

Next, we bound the left-hand side from above in terms of the iterate differences. By virtue of (33) and using the inexactness condition $(\mathsf{C}^{(k)}_{\text{inexact}})$, one obtains

$$\begin{aligned} \operatorname{dist}(\boldsymbol{0}, \partial F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k)) &\leqslant l^{-2} \|\boldsymbol{X}^k - \boldsymbol{Y}^k\|_F + \|\mathcal{R}^k\|_F \\ &\leqslant (l^{-2} + \tilde{\beta}_k) \|\boldsymbol{X}^k - \boldsymbol{Y}^k\|_F. \end{aligned} \tag{41}$$

Then, we deduce that

$$\frac{2}{c} \tilde{\Delta}_k^{1/2} \leqslant \frac{l^{-2} + \tilde{\beta}_k}{\alpha} \|\boldsymbol{X}^k - \boldsymbol{X}^{k+1}\|_F.$$

On the other hand, the sufficient decrease property (see Eq. (24)) guarantees that

$$\|\boldsymbol{X}^k - \boldsymbol{X}^{k+1}\|_F^2 \leqslant \widetilde{C}^{-1}(F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k) - F(\boldsymbol{X}^{k+1}, \boldsymbol{\mathcal{E}}^{k+1})) = \widetilde{C}^{-1}(\tilde{\Delta}_k - \tilde{\Delta}_{k+1}).$$

Substituting the above bound into the previous squared inequality, we obtain

$$\frac{4}{c^2} \tilde{\Delta}_k \leqslant \left(\frac{l^{-2} + \tilde{\beta}_k}{\alpha} \right)^2 \widetilde{C}^{-1}(\tilde{\Delta}_k - \tilde{\Delta}_{k+1}).$$

Rearranging the terms, we deduce that there exists a constant $\tilde{c} \in (0, 1)$ such that

$$\tilde{\Delta}_{k+1} \leqslant \tilde{c} \tilde{\Delta}_k \quad \text{for all sufficiently large } k.$$

This recursive inequality implies that

$$\tilde{\Delta}_k \leqslant \big(\tilde{c}\big)^{k-K}\tilde{\Delta}_K, \tag{42}$$

i.e., the sequence $\{\tilde{\Delta}_k\}$ converges to zero Q-linearly, which is equivalent to saying that $\{F(\boldsymbol{X}^k,\boldsymbol{\mathcal{E}}^k)\}$ converges Q-linearly. This completes the proof. □

We next prove the convergence properties of the sequence $\{\boldsymbol{X}^k\}$.

Theorem 3.12. *Suppose that Assumptions 1.1-3.7 hold, and let $\{(\boldsymbol{X}^k,\boldsymbol{\mathcal{E}}^k)\}$ be the sequence generated by Algorithm 1. The following statements hold.*

(i) If $F(\boldsymbol{X},\boldsymbol{\mathcal{E}})$ is a KŁ function, then $\sum_{k=1}^{+\infty}\|\boldsymbol{X}^{k+1}-\boldsymbol{X}^k\|_F < +\infty$. Consequently, the sequence $\{\boldsymbol{X}^k\}$ converges to a stationary point of $(\mathcal{P})$.

(ii) If, in addition, $F(\boldsymbol{X},\boldsymbol{\mathcal{E}})$ has the KŁ property with exponent $1/2$ *at $\boldsymbol{X}^*$, then $\{\boldsymbol{X}^k\}$ converges* at least R-linearly *to $\boldsymbol{X}^*$.*

Proof. (i) As stated in the proof of Theorem 3.11, it suffices to consider the case in which $F(\boldsymbol{X}^k,\boldsymbol{\mathcal{E}}^k) > F(\boldsymbol{X}^{k+1},\boldsymbol{\mathcal{E}}^{k+1})$ for all sufficiently large k. Let $(\boldsymbol{X}^*,\boldsymbol{0})$ be a cluster point of the sequence $\{(\boldsymbol{X}^k,\boldsymbol{\mathcal{E}}^k)\}$.

From Theorem 3.6(i)–(ii) and (iv), there exist $\epsilon > 0$, $\vartheta > 0$, and an index $K \in \mathbb{N}$ such that for all $k \geqslant K$, $(\boldsymbol{X}^k,\boldsymbol{\mathcal{E}}^k) \in \big\{(\boldsymbol{X},\boldsymbol{\mathcal{E}}) \mid \mathrm{dist}\big((\boldsymbol{X},\boldsymbol{\mathcal{E}}),\Gamma^\infty\big) < \epsilon\big\} \cap \big\{(\boldsymbol{X},\boldsymbol{\mathcal{E}}) \mid F(\boldsymbol{X}^*,\boldsymbol{0}) < F(\boldsymbol{X}^k,\boldsymbol{\mathcal{E}}^k) < F(\boldsymbol{X}^*,\boldsymbol{0})+\vartheta\big\}$. Because F is a KŁ function on this neighborhood, there is a *desingularizing function* φ such that for each $k \geqslant K$,

$$\varphi'\big(F(\boldsymbol{X}^k,\boldsymbol{\mathcal{E}}^k)-F(\boldsymbol{X}^*,\boldsymbol{0})\big)\cdot\mathrm{dist}\big(\boldsymbol{0},\partial F(\boldsymbol{X}^k,\boldsymbol{\mathcal{E}}^k)\big) \geqslant 1. \tag{43}$$

Moreover, since φ is concave on $[0,\vartheta)$, for any such k, one obtains the standard KŁ -type inequality

$$\begin{aligned}&\varphi'(F(\boldsymbol{X}^k,\boldsymbol{\mathcal{E}}^k)-F(\boldsymbol{X}^*,\boldsymbol{0}))(F(\boldsymbol{X}^k,\boldsymbol{\mathcal{E}}^k)-F(\boldsymbol{X}^{k+1},\boldsymbol{\mathcal{E}}^{k+1}))\\ \leqslant\ &\varphi(F(\boldsymbol{X}^k)-F(\boldsymbol{X}^*))-\varphi(F(\boldsymbol{X}^{k+1})-F(\boldsymbol{X}^*)).\end{aligned} \tag{44}$$

Set $\bar{\Delta}_k \coloneqq \varphi(F(\boldsymbol{X}^k)-F(\boldsymbol{X}^*))$ for each k. It follows from (43) and (44) that for each $k \geqslant K$,

$$\begin{aligned}\bar{\Delta}_k-\bar{\Delta}_{k+1} &\geqslant \varphi'(F(\boldsymbol{X}^k,\boldsymbol{\mathcal{E}}^k)-F(\boldsymbol{X}^*,\boldsymbol{0}))(F(\boldsymbol{X}^k,\boldsymbol{\mathcal{E}}^k)-F(\boldsymbol{X}^{k+1},\boldsymbol{\mathcal{E}}^{k+1}))\\ &\geqslant \frac{F(\boldsymbol{X}^k,\boldsymbol{\mathcal{E}}^k)-F(\boldsymbol{X}^{k+1},\boldsymbol{\mathcal{E}}^{k+1})}{\mathrm{dist}(\boldsymbol{0},\partial F(\boldsymbol{X}^k,\boldsymbol{\mathcal{E}}^k))}\\ &\overset{(a)}{\geqslant} \frac{\widetilde{C}\|\boldsymbol{X}^k-\boldsymbol{X}^{k+1}\|_F^2}{\frac{l^{-2}+\tilde{\beta}_k}{\alpha}\|\boldsymbol{X}^k-\boldsymbol{X}^{k+1}\|_F}\\ &= \frac{\alpha\widetilde{C}}{l^{-2}+\tilde{\beta}_k}\|\boldsymbol{X}^k-\boldsymbol{X}^{k+1}\|_F \overset{(b)}{\geqslant} \frac{\alpha\widetilde{C}}{l^{-2}+1}\|\boldsymbol{X}^k-\boldsymbol{X}^{k+1}\|_F,\end{aligned} \tag{45}$$

where (a) follows from (24), (41) and $\boldsymbol{Y}^k = \frac{1}{\alpha}(\boldsymbol{X}^{k+1} - \boldsymbol{X}^k) + \boldsymbol{X}^k$, and (b) holds since $\tilde{\beta}_k \leqslant 1$.

Denote $\omega \coloneqq \frac{\alpha\tilde{C}}{l^{-2}+1} > 0$. Summing up (45) from K to any $\bar{k} > K$ yields

$$\sum_{k=K}^{\bar{k}} \|\boldsymbol{X}^k - \boldsymbol{X}^{k+1}\|_F \leqslant \frac{1}{\omega} \sum_{k=K}^{\bar{k}} \left(\bar{\Delta}_k - \bar{\Delta}_{k+1}\right) = \frac{1}{\omega}\left(\bar{\Delta}_K - \bar{\Delta}_{\bar{k}+1}\right) \leqslant \frac{1}{\omega}\bar{\Delta}_K.$$

Letting $\bar{k} \to +\infty$, we have $\sum_{k=K}^{+\infty} \|\boldsymbol{X}^k - \boldsymbol{X}^{k+1}\|_F < +\infty$, and hence $\{\boldsymbol{X}^k\}$ is a Cauchy sequence. By Theorem 3.6(iii), its limit $\boldsymbol{X}^*$ lies in Ω^∞. One then verifies that $\boldsymbol{X}^*$ is a stationary point of $(\mathcal{P})$, completing the proof of statement (i).

(ii). By Theorem 3.11 (see (42)), we know that there exists an index $K \in \mathbb{N}$ such that for all $k \geqslant K$,

$$F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k) - F(\boldsymbol{X}^*, \boldsymbol{0}) \leqslant \left(\tilde{c}\right)^{k-K} \tilde{\Delta}_K \quad \text{for some } \tilde{c} \in (0,1).$$

Using statement (i), we know $\{\boldsymbol{X}^k\}$ converges. For any $k \geqslant K$, we have

$$\|\boldsymbol{X}^k - \boldsymbol{X}^*\|_F = \left\| \lim_{t\to\infty} \sum_{\ell=k}^{t} \left(\boldsymbol{X}^\ell - \boldsymbol{X}^{\ell+1}\right) \right\|_F \leqslant \sum_{\ell=k}^{\infty} \|\boldsymbol{X}^\ell - \boldsymbol{X}^{\ell+1}\|_F.$$

By (24), we can bound each $\|\boldsymbol{X}^\ell - \boldsymbol{X}^{\ell+1}\|_F$ by a constant multiple of $\left(F(\boldsymbol{X}^\ell, \boldsymbol{\mathcal{E}}^\ell) - F(\boldsymbol{X}^{\ell+1}, \boldsymbol{\mathcal{E}}^{\ell+1})\right)^{1/2}$. Hence

$$\|\boldsymbol{X}^k - \boldsymbol{X}^*\|_F \;\leqslant\; \sum_{\ell=k}^{\infty} \left(\tilde{C}^{-1}\left(F(\boldsymbol{X}^\ell, \boldsymbol{\mathcal{E}}^\ell) - F(\boldsymbol{X}^{\ell+1}, \boldsymbol{\mathcal{E}}^{\ell+1})\right)\right)^{\frac{1}{2}} \;\leqslant\; \left(\frac{\tilde{\Delta}_K}{\tilde{C}\,(\tilde{c})^K}\right)^{\frac{1}{2}} \sum_{\ell=k}^{\infty} (\tilde{c})^{\frac{\ell}{2}}.$$

Since $\sum_{\ell=k}^{\infty} (\tilde{c})^{\ell/2} = \dfrac{(\tilde{c})^{k/2}}{1-\sqrt{\tilde{c}}}$, it follows that

$$\|\boldsymbol{X}^k - \boldsymbol{X}^*\|_F \;\leqslant\; \left(\frac{\tilde{\Delta}_K}{\tilde{C}\,(\tilde{c})^K}\right)^{\frac{1}{2}} \frac{(\tilde{c})^{k/2}}{1-\sqrt{\tilde{c}}} \;=\; \mathcal{O}\big((\sqrt{\tilde{c}}\,)^k\big).$$

Because $0 < \sqrt{\tilde{c}} < 1$, this shows $\{\boldsymbol{X}^k\}$ converges *R-linearly* to $\boldsymbol{X}^*$. The proof is complete. □

4 Numerical Experiments

This section evaluates the numerical performance of DIIR-QUIC in solving $(\mathcal{P})$ on both synthetic and real-world datasets and investigates its convergence behavior

numerically. All experiments are conducted on a PC equipped with an AMD Ryzen 5 4600H processor (3.00 GHz base frequency), 16 GB of RAM, running 64-bit Ubuntu 22 LTS. At each outer iteration, the weighted ℓ_1-regularized log-determinant subproblem ($\mathcal{P}_{\text{sub}}$) is solved using a QUIC implementation [1] modified to enforce the inexact termination condition ($\mathsf{C}^{(k)}_{\text{inexact}}$). We have publicly released the C++ implementation of DIIR-QUIC at `https://github.com/Optimizater/DIIR-QUIC`.

4.1 Nonconvex ℓ_p-regularized log-determinant problems

The nonconvex ℓ_p (quasi-)norm ($0 < p < 1$) is used to promote sparsity in inverse covariance estimation [11]. As a representative instance of our broader class of sparsity-promoting formulations, we take $\Phi(\boldsymbol{X}) = \|\boldsymbol{X}\|_p^p = \sum_{ij} |X_{ij}|^p$ and consider the following problem:

$$\begin{aligned} \min \;\; & F(\boldsymbol{X}) = \{f(\boldsymbol{X}) \coloneqq \operatorname{tr}(\boldsymbol{S}\boldsymbol{X}) - \log\det \boldsymbol{X})\} + \rho \sum_{ij} |X_{ij}|^p \\ \text{s.t.} \;\; & \boldsymbol{X} \in \mathbb{S}^n_{++}, \end{aligned} \tag{46}$$

where $p \in (0, 1)$. To our knowledge, [11] is the only existing algorithm capable of directly solving the special case of problem (46) in which the regularizer is $\rho \sum_{i \neq j} |X_{ij}|^p$. Accordingly, we use it as the benchmark algorithm in our numerical experiments and, following their terminology, refer to it as ℓ_pCOV.

For DIIR-QUIC, we set $\alpha = 0.98$ and $\mu = 0.1$. In Algorithm 2, an iterative coordinate descent method is used to compute an approximate Newton direction $\boldsymbol{D}^t$ in (12), under the inexactness condition (13) with $\epsilon = 0.05$. We terminate the outer loop when either

(i) the iteration count reaches `MaxIter` = 3000, or

(ii) the stationarity residual (see Theorem 1.4)

$$\max_{(i,j) \in \mathcal{I}(\boldsymbol{X}^k)} |S_{ij} - [(\boldsymbol{X}^k)^{-1}]_{ij} + \rho p |X^k_{ij}|^{p-1} \operatorname{sgn}(X^k_{ij})| \cdot n < \text{tol}, \tag{47}$$

is satisfied. Here, tol = 10^{-5}.

In our implementation, we treat any entry of $\boldsymbol{X}^k$ whose magnitude exceeds 10^{-8} as active, and these indices define $\mathcal{I}(\boldsymbol{X}^k)$. Following [34], we initialize

$$[\boldsymbol{X}^0]_{ij} = \begin{cases} 0, & \text{if } i \neq j \\ \frac{1}{S_{ii} + \rho}, & \text{if } i = j. \end{cases} \tag{48}$$

[1] A publicly available implementation of the original QUIC algorithm—written in C++ with a Python wrapper—can be found at `https://github.com/osdf/pyquic`.

and initialize the perturbation matrix $\boldsymbol{\mathcal{E}}^0$ by drawing an $n \times n$ matrix $\boldsymbol{W}$ with i.i.d. $\mathcal{N}(0,1)$, scaling each entry by 0.5, symmetrizing via $\widehat{\boldsymbol{W}} = (\boldsymbol{W} + \boldsymbol{W}^T)/2$, and then setting $\mathcal{E}^0_{ij} = |\widehat{W}_{ij}|$ for all (i,j).

For the ℓ_pCOV, the authors have not publicly released their MATLAB implementation. Consequently, we reimplemented the algorithm primarily based on the descriptions provided in the original paper. Since our Algorithm 2 is implemented in C++, we also developed the core computational components of ℓ_pCOV in C++ to ensure a relatively fair comparison. As described in [11, §IV SIMULATIONS], the authors employed a warm-start (WS) strategy with respect to the model parameter p to potentially improve numerical performance with a good initialization. This strategy has also been incorporated into our reimplementation. More precisely, this warm-start procedure generates a short sequence of exponents $1 = p_{(0)} > p_{(1)} > \cdots > p_{(K)} = p > 0$ that interpolates linearly between 1 and the target $p \in (0,1)$. The number of intermediate steps K is chosen as

$$K = \begin{cases} 2, & 1 > p \geqslant 0.9, \\ 3, & 0.7 \leqslant p < 0.9, \\ 4, & 0.4 \leqslant p < 0.7, \\ 5, & 0.2 \leqslant p < 0.4, \\ 6, & 0 < p < 0.2. \end{cases}$$

Then it proceeds as follows:

(i) **Initialization:** Set $p_{(0)} = 1$ and adopt the GLASSO algorithm [4] to solve the resulting convex weighted ℓ_1-regularized log-determinant program with initialization (48).

(ii) **Warm-start loop:** For $k = 1, \ldots, K$, run the ℓ_pCOV algorithm at exponent $p_{(k)}$, using the solution obtained at step $k-1$ as the initial estimation.

Since ℓ_pCOV tackles (46) with the off-diagonal regularizer $\rho \sum_{i \neq j} |X_{ij}|^p$, we terminate its iterations when either

(i) the iteration count reaches `MaxIter` $= 3000$, or

(ii) the stationarity residual (47) is satisfied. Here, since the regularizer excludes diagonal entries, we set $\rho p |X^k_{ij}|^{p-1} \mathrm{sgn}(X^k_{ij}) = 0$ for $i = j$ and $(i,j) \in \mathcal{I}(\boldsymbol{X}^k)$.

4.2 Synthetic data

We generate two synthetic test cases of n-variate Gaussian data following [35], each defined by a known precision matrix:

(i) **Tridiagonal precision:** A strongly diagonally dominant precision matrix

$$\boldsymbol{\Sigma}_{ii} = 1.25,\ \boldsymbol{\Sigma}_{i,i+1} = \boldsymbol{\Sigma}_{i+1,i} = -0.5,\ \boldsymbol{\Sigma}_{ij} = 0 \text{ otherwise.}$$

This structure induces simple chain-like dependencies among the variables and ensures positive definiteness by virtue of diagonal dominance [36].

(ii) **Clustered precision:** A *random* structured precision matrix with $n/100$ clusters of size 100. Each variable is conditionally dependent on approximately 10 others, with 90% of edges located within the same cluster and only 10% connecting different clusters. This yields dense intra-cluster and sparse inter-cluster connectivity.

As noted in [35], the tridiagonal case provides a simple baseline with one-dimensional dependence, while the clustered case models more realistic community-structured dependencies commonly observed in practice [37]. In particular, we follow the procedure presented in [38, Example 1] to generate the clustered precision matrix. Using the true sparse inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ generated above, we then draw $m = n/2$ i.i.d. samples from the corresponding GMRF distribution. When $\boldsymbol{S}$ is singular, we regularize it by setting $\boldsymbol{S} \leftarrow \boldsymbol{S} + \hat{\varepsilon}\boldsymbol{I}_{n\times n}$, where $\hat{\varepsilon} \in \{10^{-8}, 10^{-7}, \ldots, 10^{-4}\}$, and repeat this update until $\boldsymbol{S}$ becomes positive definite.

The regularization parameter ρ on synthetic datasets is chosen by five-fold cross-validation, as in [6]. We first construct a logarithmic grid of 10 candidate values: $\{\rho_j = 10^{a_j}\}_{j=1}^{10}$, where the exponents a_j are equally spaced over $[-1, 0]$ for tridiagonal matrices and over $[-2, 0]$ for clustered matrices. For each ρ_j and each fold k, we estimate the precision matrix on the training subset and evaluate its negative log-likelihood (NLL_k) on the test subset:

$$\mathrm{NLL}_k(\rho_j) = \frac{1}{2}\left(\mathrm{tr}(\boldsymbol{S}_{\text{test}}^{(k)}\boldsymbol{\Sigma}_{\text{train}}^{\dagger,(k)}) - \log\det\boldsymbol{\Sigma}_{\text{train}}^{\dagger,(k)}\right), \tag{49}$$

where $\boldsymbol{S}_{\text{test}}^{(k)}$ is the empirical covariance on fold k and $\boldsymbol{\Sigma}_{\text{train}}^{\dagger,(k)}$ is the estimated precision matrix from the training set. We then average over folds:

$$\mathrm{NLL}_{\mathrm{CV}} = \frac{1}{5}\sum_{k=1}^{5}\mathrm{NLL}_k(\rho_j), \tag{50}$$

and choose the ρ_j that minimizes $\mathrm{NLL}_{\mathrm{CV}}$ as the optimal ρ^*.

4.2.1 Empirical convergence behavior on synthetic data

We now investigate the convergence behaviors of DIIR-QUIC. All experiments run DIIR-QUIC and ℓ_pCOV on the same dataset. Each curve in our plots corresponds to a single run of the respective algorithm on that shared dataset. Specifically,

(i) We plot the objective values $F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k)$ and stationarity residuals versus elapsed time for both DIIR-QUIC and ℓ_pCOV. In ℓ_pCOV, we consider the penalty $\rho \sum_{i,j} |X_{ij}|^p$ for objective comparison. When warm-starting over a sequence of exponents $1 = p_{(0)} > p_{(1)} > \cdots > p_{(K)} = p > 0$, we record performance metrics only at the final exponent $p_{(K)} = p$. During the initial warm-up stages $(p_{(i)}, \forall 1 \leqslant i < K)$, we terminate each run when either the iteration count reaches `MaxIter` $= 3000$, or the change in successive objective values falls below 10^{-4}, and we do not log performance metrics for these intermediate stages. Fig 1 confirms that DIIR-QUIC not only drives down the objective rapidly but also achieves high-quality first-order stationarity—effectively solving the nonconvex ℓ_p-regularized log-determinant problem in far fewer iterations than ℓ_pCOV algorithm, across varying matrix dimensions and exponents p.

(ii) To illustrate the Q-linear convergence of DIIR-QUIC, we plot the error $F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k) - F(\boldsymbol{X}^\dagger, \boldsymbol{\mathcal{E}}^\dagger)$ over the last few iterations, where $\boldsymbol{X}^\dagger$ denotes the returned estimate of DIIR-QUIC. To avoid plotting numerical noise, we only display those $F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k) - F(\boldsymbol{X}^\dagger, \boldsymbol{\mathcal{E}}^\dagger)$ values exceeding 10^{-8}. The plots are shown in Fig 2.

(iii) We empirically validate Proposition 2.1 by checking that, once the approximate fixed-point conditions

$$\|\boldsymbol{\mathcal{E}}^k\|_F < 10^{-8} \quad \text{and} \quad \|\boldsymbol{X}^k - \boldsymbol{Y}^k\|_F < 10^{-5},$$

are satisfied, the stationarity residual condition (47) is generally satisfied. To illustrate this, Fig 3 plots, over the elapsed time, the scaled infinity norms

$$n \cdot \|\boldsymbol{X}^k - \boldsymbol{Y}^k\|_\infty \quad \text{and} \quad n \cdot \|\boldsymbol{\mathcal{E}}^k\|_\infty,$$

where we only display those $n \cdot \|\boldsymbol{\mathcal{E}}^k\|_\infty$ values exceeding 10^{-10}. The plots confirm that as soon as the two fixed-point thresholds are reached, the stationarity residual falls below its prescribed tolerance.

(iv) To empirically confirm Theorem 3.9, we monitor the cardinality of the index set $\mathcal{I}(\boldsymbol{X}^k)$ over successive iterations. In all our tests, each of these cardinalities stabilizes after a finite number of iterations, indicating that the algorithm has correctly identified the smooth active manifold. The plots are shown in Fig 4.

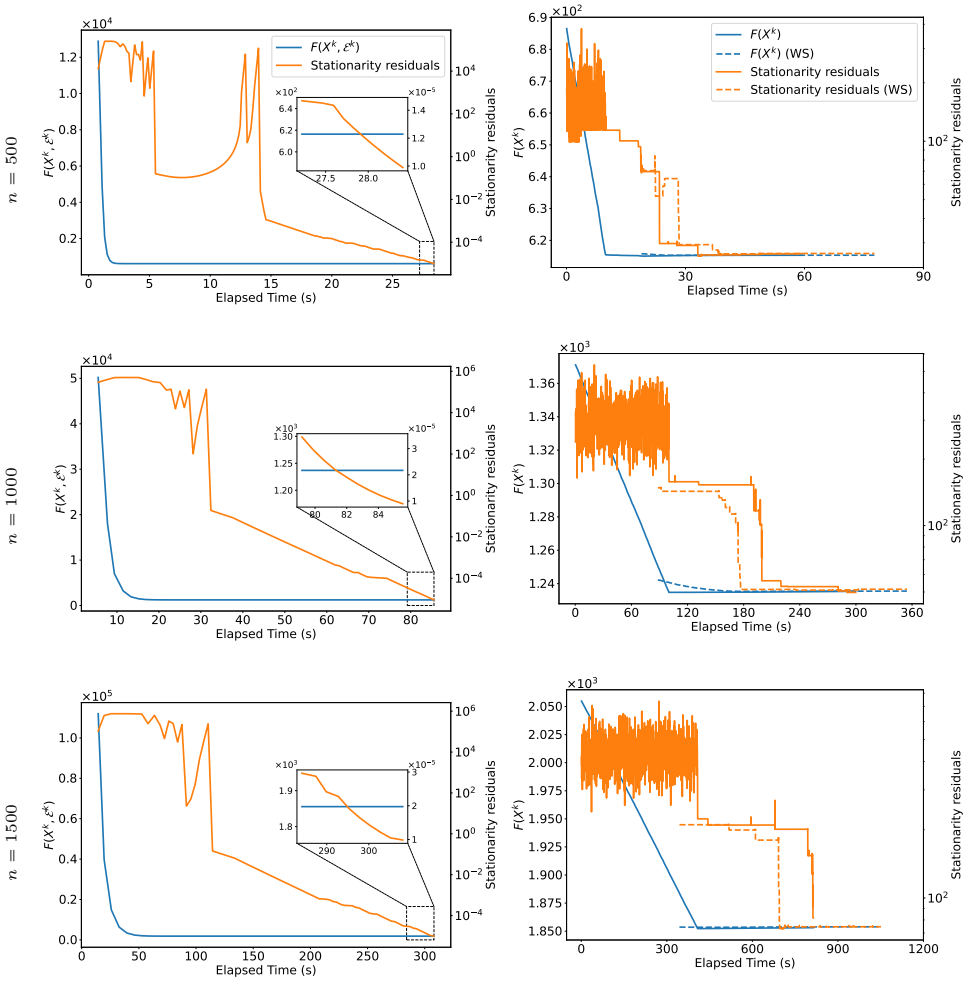

Figure 1: Convergence plots of objective value and stationarity residual (plotted on a logarithmic scale) versus elapsed time (seconds) for DIIR-QUIC (left) and ℓ_pCOV (right) on a tridiagonal precision matrix with $p = 0.5$. DIIR-QUIC not only converges faster—consistently attaining the prescribed stationarity residual when ℓ_pCOV often does not—but also reduces total runtime by approximately 70% across all tested dimensions.

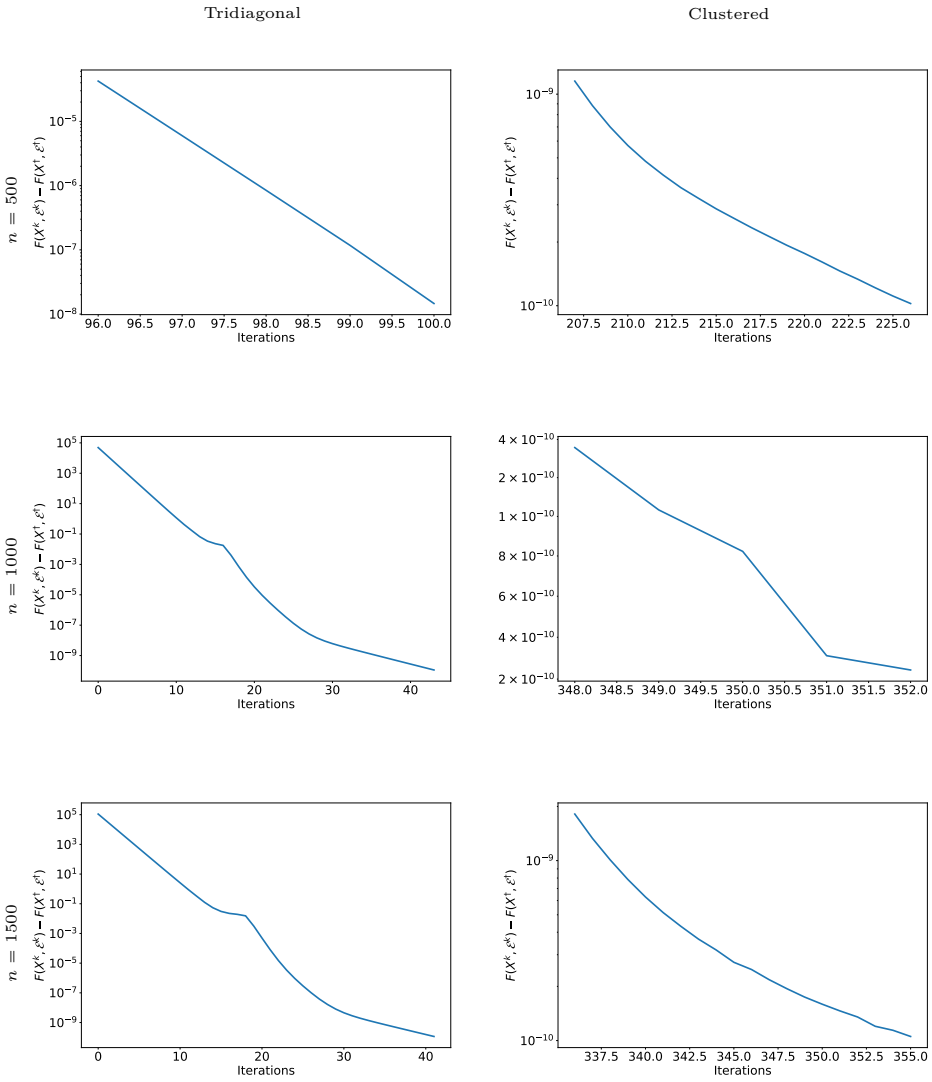

Figure 2: Q-linear convergence of the perturbed objective error $F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k) - F(\boldsymbol{X}^\dagger, \boldsymbol{\mathcal{E}}^\dagger)$ plotted versus iteration for DIIR-QUIC on tridiagonal precision matrices and clustered matrices with $p = 0.5$, across varying matrix dimensions. The error is displayed on a logarithmic scale, and only values above 10^{-8} are shown.

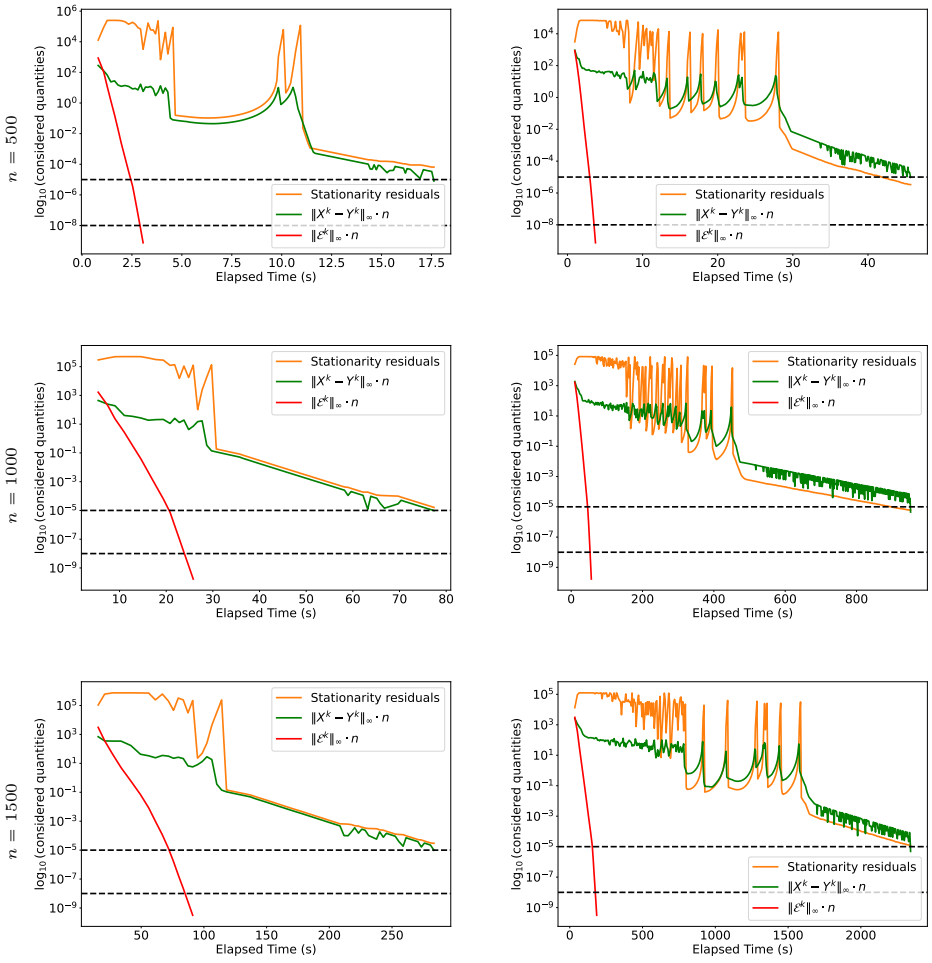

Figure 3: Fixed-point conditions and stationarity residual condition versus elapsed time (seconds) for DIIR-QUIC and ℓ_pCOV on tridiagonal precision matrices and cluster precision matrices with $p = 0.5$

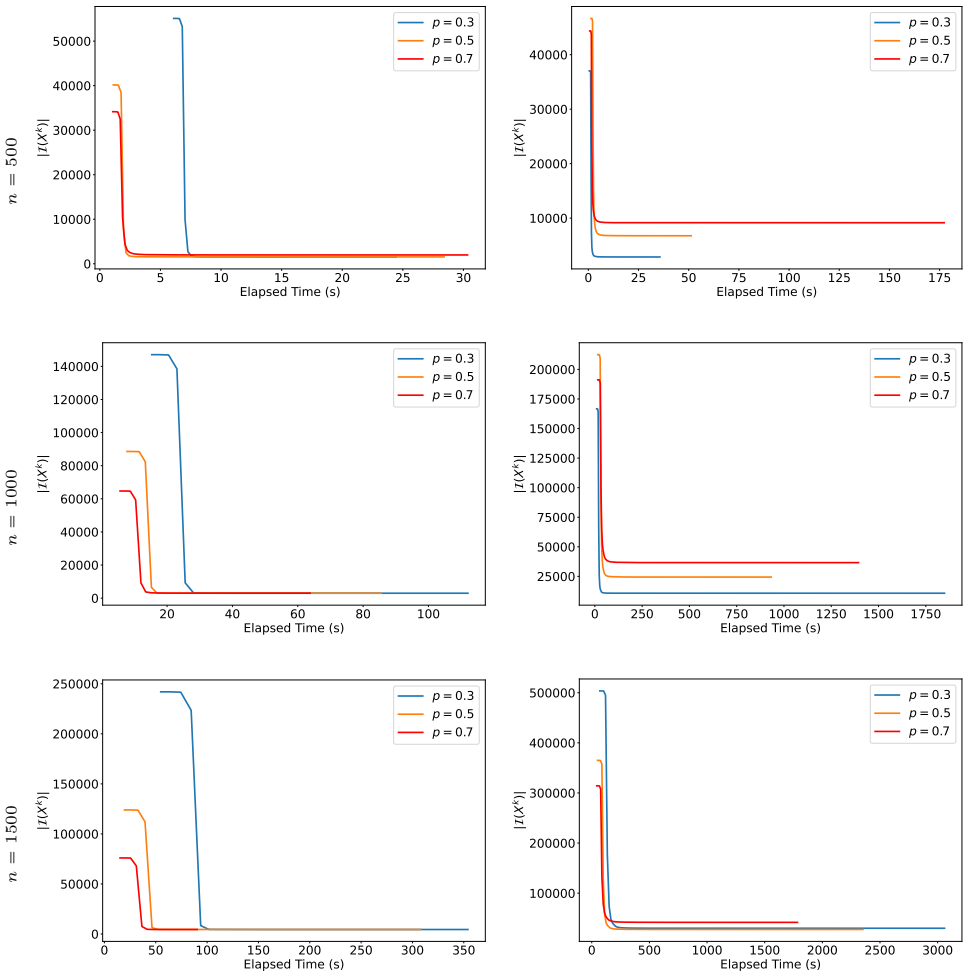

Figure 4: Plots of the cardinality of $\mathcal{I}(\boldsymbol{X}^k)$ versus elapsed time for DIIR-QUIC on tridiagonal precision matrices and clustered matrices.

4.2.2 Statistical properties of the estimator on synthetic data

To measure how well an algorithm recovers the true precision matrix $\mathbf{\Sigma}^{-1}$, we follow [35, 11, 7] and compute the following metrics for the returned estimate $\boldsymbol{X}^{\dagger}$:

(i) Normalized Kullback-Leibler (KL) loss [39]

$$\mathrm{KL}(\boldsymbol{X}^{\dagger}) \coloneqq \frac{1}{n}\left[\mathrm{tr}(\mathbf{\Sigma}\boldsymbol{X}^{\dagger}) - \log\det(\mathbf{\Sigma}\boldsymbol{X}^{\dagger}) - n\right]$$

and quadratic loss (Loss_Q)

$$\mathrm{Loss}_Q(\boldsymbol{X}^{\dagger}) \coloneqq \frac{1}{n}\|\mathbf{\Sigma}\boldsymbol{X}^{\dagger} - \boldsymbol{I}\|_F,$$

measure entry-wise fit between $\mathbf{\Sigma}\boldsymbol{X}^{\dagger}$ and the identity. For these two measures, smaller value indicates a preferred estimator.

(ii) Matthews correlation coefficient (MCC) [40]

$$\mathrm{MCC}(\boldsymbol{X}^{\dagger}) \coloneqq \frac{\mathrm{TP}\times\mathrm{TN} - \mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}.$$

Here, TP, TN, FP, and FN count true positives, true negatives, false positives and false negatives in the support of $\boldsymbol{X}^{\dagger}$. MCC ranges from -1 to 1, with 1 denoting perfect support recovery.

(iii) Sensitivity and specificity [7]

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \quad \text{and} \quad \mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}.$$

Sensitivity measures the fraction of true nonzeros recovered, while specificity measures the fraction of true zeros recovered; ideal values are close to 1.

(iv) F_1 score [35]:

$$\mathrm{F}_1\text{-score} \coloneqq [1 + 0.5(\mathrm{FP}+\mathrm{FN})/(\|\mathbf{\Sigma}^{-1}\|_0 - \mathrm{FN})]^{-1}.$$

Here $\|\mathbf{\Sigma}^{-1}\|_0$ counts the number of nonzeros of the true precision matrix $\mathbf{\Sigma}^{-1}$. The F_1-score balances precision and recall on the nonzero support. A value of 1 indicates perfect support recovery (see, e.g., [41] for details).

To comprehensively evaluate the performance of DIIR-QUIC and ℓ_pCOV, we conducted 10 independent trials on synthetic datasets featuring two distinct covariance structures (tridiagonal and clustered patterns) across varying dimensions and exponents p. As shown in Table 1, the results demonstrate that DIIR-QUIC consistently outperformed competing methods in recovering the underlying sparse structure.

Table 1: Comparison of DIIR-QUIC and ℓ_pCOV on synthetic covariance estimation. For each method, the table reports the mean ($\pm$ one standard deviation) over 10 independent trials under two covariance structures—tridiagonal (Tri) and clustered (Clu)—with $p = 0.5$ at dimensions $n = 500, 1000$, and 1500. Boldface entries indicate the better performance.

Data	n	Algorithm	$\mathbf{F}_1$	Time (s)	nnz(Σ^{-1})	nnz($X^\dagger$)	Loss Q	KL Loss	Sensitivity	Specificity	MCC
tri	500	DIIR-QUIC	0.972 (0.004)	**14.615 (12.238)**	1498.000 (0.000)	1525.600 (10.947)	0.011 (0.000)	0.028 (0.001)	0.981 (0.002)	**1.000 (0.000)**	0.972 (0.004)
		ℓ_pCOV	0.999 (0.001)	59.706 (0.804)	1498.000 (0.000)	1498.400 (0.800)	**0.009 (0.000)**	**0.015 (0.001)**	0.999 (0.001)	**1.000 (0.000)**	0.999 (0.001)
		ℓ_pCOV (WS)	**1.000 (0.000)**	78.437 (1.282)	1498.000 (0.000)	1498.400 (0.800)	**0.009 (0.000)**	**0.015 (0.001)**	**1.000 (0.000)**	**1.000 (0.000)**	**1.000 (0.000)**
	1000	DIIR-QUIC	0.984 (0.002)	**78.944 (27.952)**	2998.000 (0.000)	2994.400 (4.630)	0.007 (0.000)	0.023 (0.000)	0.984 (0.001)	**1.000 (0.000)**	0.984 (0.002)
		ℓ_pCOV	**1.000 (0.000)**	250.391 (25.703)	2998.000 (0.000)	2998.000 (0.000)	**0.005 (0.000)**	**0.011 (0.000)**	**1.000 (0.000)**	**1.000 (0.000)**	**1.000 (0.000)**
		ℓ_pCOV (WS)	**1.000 (0.000)**	308.795 (45.818)	2998.000 (0.000)	2998.000 (0.000)	**0.005 (0.000)**	**0.011 (0.000)**	**1.000 (0.000)**	**1.000 (0.000)**	**1.000 (0.000)**
	1500	DIIR-QUIC	0.988 (0.001)	**346.117 (137.118)**	4498.000 (0.000)	4490.200 (5.400)	0.005 (0.000)	0.021 (0.000)	0.987 (0.001)	**1.000 (0.000)**	0.988 (0.001)
		ℓ_pCOV	**1.000 (0.000)**	579.475 (103.963)	4498.000 (0.000)	4498.000 (0.000)	**0.004 (0.000)**	0.010 (0.000)	**1.000 (0.000)**	**1.000 (0.000)**	**1.000 (0.000)**
		ℓ_pCOV (WS)	**1.000 (0.000)**	840.307 (114.144)	4498.000 (0.000)	4498.000 (0.000)	**0.004 (0.000)**	**0.009 (0.000)**	**1.000 (0.000)**	**1.000 (0.000)**	**1.000 (0.000)**
clu	500	DIIR-QUIC	**0.423 (0.005)**	**83.094 (38.935)**	11439.800 (136.802)	6761.000 (95.734)	**0.038 (0.000)**	**0.252 (0.003)**	**0.337 (0.005)**	0.988 (0.000)	0.418 (0.005)
		ℓ_pCOV	0.379 (0.007)	142.713 (38.860)	11439.800 (136.802)	3711.600 (77.386)	0.040 (0.000)	0.263 (0.004)	0.251 (0.005)	**0.996 (0.000)**	0.427 (0.007)
		ℓ_pCOV (WS)	0.403 (0.006)	142.709 (38.859)	11439.800 (136.802)	3975.400 (83.852)	0.039 (0.000)	0.256 (0.004)	0.272 (0.005)	**0.996 (0.000)**	**0.448 (0.006)**
	1000	DIIR-QUIC	0.568 (0.004)	**934.045 (311.597)**	23021.000 (187.743)	24441.800 (264.264)	0.023 (0.000)	0.174 (0.002)	**0.586 (0.004)**	0.989 (0.000)	0.558 (0.004)
		ℓ_pCOV	0.610 (0.004)	993.646 (311.517)	23021.000 (187.743)	14357.400 (150.439)	0.023 (0.000)	0.173 (0.002)	0.495 (0.005)	**0.997 (0.000)**	0.621 (0.004)
		ℓ_pCOV (WS)	**0.652 (0.005)**	993.647 (311.516)	23021.000 (187.743)	15598.600 (174.613)	**0.022 (0.000)**	**0.160 (0.002)**	0.547 (0.006)	**0.997 (0.000)**	**0.658 (0.004)**
	1500	DIIR-QUIC	0.675 (0.003)	**3178.596 (1392.902)**	34577.400 (187.755)	27870.600 (229.703)	**0.016 (0.000)**	0.146 (0.001)	**0.610 (0.004)**	0.997 (0.000)	0.675 (0.003)
		ℓ_pCOV	0.654 (0.003)	3238.136 (1393.000)	34577.400 (187.755)	18417.000 (170.050)	0.017 (0.000)	0.155 (0.001)	0.501 (0.003)	**1.000 (0.000)**	0.684 (0.002)
		ℓ_pCOV (WS)	**0.700 (0.003)**	3238.125 (1393.003)	34577.400 (187.755)	20222.200 (174.566)	**0.016 (0.000)**	**0.141 (0.001)**	0.555 (0.003)	**1.000 (0.000)**	**0.723 (0.002)**

4.3 Real-world datasets

We conduct performance comparisons between DIIR-QUIC and ℓ_pCOV on two classes of real-world datasets: gene expression datasets [38] and stock datasets [6]. In our first experiments, we consider three benchmark datasets from [38, Examples 3, 5 and 6]: the Lymph node status data ($n = 587$), the Arabidopsis thaliana data ($n = 834$) and the Leukemia data ($n = 1255$). Following [11], we select the regularization parameter ρ via the extended Bayesian information criterion (EBIC). For a candidate precision-matrix estimate $\boldsymbol{X}^\dagger$ and sample size m, the EBIC for a Gaussian graphical model is commonly defined as

$$\mathrm{EBIC}(\boldsymbol{X}^\dagger) = -m(\log\det \boldsymbol{X}^\dagger - \mathrm{tr}(S\boldsymbol{X}^\dagger)) + \frac{\log m + 4\tilde{\gamma}\log n}{2}|\mathcal{I}(\boldsymbol{X}^\dagger)|_{\mathrm{off}},$$

where $|\mathcal{I}(\boldsymbol{X}^\dagger)|_{\mathrm{off}}$ denotes the cardinality of the index set of nonzero off-diagonal entries of $\boldsymbol{X}^\dagger$ and $\tilde{\gamma} \in [0, 1]$ is a user-defined constant. We then evaluate EBIC over a logarithmic grid

$$\rho_j = 10^{a_j},\ a_j = -1 + \frac{j-1}{N-1}(0-(-1)) = -1 + \frac{j-1}{N-1},\ j \in [N],$$

where $N = 10$, so that ρ_j ranges smoothly from 10^{-1} to 1. Denoting the resulting EBIC values by $\mathrm{EBIC}(\rho_j)$, we choose $\rho_{\mathrm{EBIC}} = \mathrm{argmin}_{j\in[N]}\ \mathrm{EBIC}(\rho_j)$.

To evaluate both the statistical accuracy and computational efficiency of DIIR-QUIC at the chosen regularization level ρ_{EBIC}, we compute the five-fold cross-validated test-set negative log-likelihood, as defined in (50). The metric $\mathrm{NLL}_{\mathrm{CV}}$ provides a cross-validated assessment of predictive performance, effectively balancing the trade-off between under- and over-fitting. The values of per-fold held-out negative log-likelihoods are summarized via box plots to visualize variability and robustness. The right column of Figure 5 plots the stationarity residuals versus elapsed time for both algorithms at ρ_{EBIC}. Under an identical wall-clock time limit (rather than an iteration cap) imposed on ℓ_pCOV, DIIR-QUIC converges more rapidly, demonstrating its computational efficiency in reaching a stationary point.

In the second experiment, following [6], we evaluate portfolio performance within the Markowitz mean-variance framework. The optimal portfolio weights $\boldsymbol{w} \in \mathbb{R}^p$ are obtained by solving:

$$\min_{\boldsymbol{w}}\ \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w} \quad \text{s.t.} \quad \boldsymbol{w}^T\boldsymbol{e} = 1, \tag{51}$$

where $\boldsymbol{\Sigma}$ is the estimated covariance matrix of asset returns and $\boldsymbol{e}$ is the all-ones vector. For a given weight vector $\boldsymbol{w}$, we compute on the test set $\boldsymbol{X}_{\mathrm{test}}$

$$R(\boldsymbol{w}) = \sum_{\boldsymbol{x}\in\boldsymbol{X}_{\mathrm{test}}} \boldsymbol{w}^T\boldsymbol{x}, \quad \sigma(\boldsymbol{w}) = \sqrt{\boldsymbol{w}^T\boldsymbol{S}_{\mathrm{test}}\boldsymbol{w}},$$

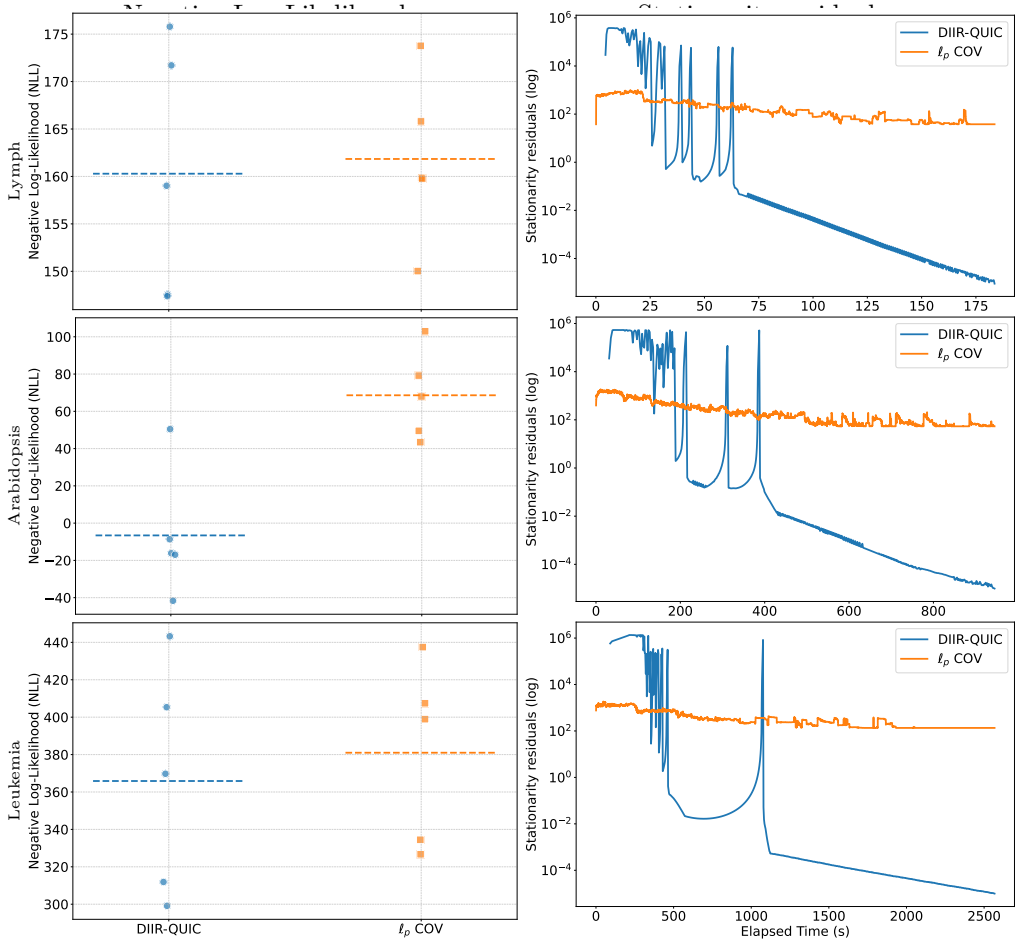

Figure 5: Comparison of DIIR-QUIC and ℓ_pCOV algorithms. **Left:** Five-fold cross-validated negative log-likelihood (lower is better). **Right:** Stationarity residuals versus elapsed time.

where $\boldsymbol{S}_{\text{test}}$ is the sample covariance matrix of X_{test}. Portfolio performance is then measured by the Sharpe ratio

$$S(\boldsymbol{w}) = \frac{R(\boldsymbol{w})}{\sigma(\boldsymbol{w})}. \tag{52}$$

We analyze weekly returns of 30 Dow Jones Industrial Index components from January 6, 2020, to June 26, 2023, with data obtained from East Money Information.[2] Following [6], we partition the series into training (first 50 weeks), validation (next 50 weeks) and test sets. Regularization parameters are tuned to maximize the Sharpe ratio on the validation set. We compare three methods—DIIR-QUIC, ℓ_pCOV, and its warm-start ℓ_pCOV (WS)—to evaluate their impact on out-of-sample performance. Table 2 reports the results.

Table 2: Out-of-sample performance comparison of DIIR-QUIC and ℓ_pCOV methods on the Dow Jones weekly returns: realized return $R(\boldsymbol{w})$, realized risk $\sigma(\boldsymbol{w})$, and Sharpe ratio $S(\boldsymbol{w})$.

p	Algorithm	$R(\boldsymbol{w})$	$\sigma(\boldsymbol{w})$	$S(\boldsymbol{w})$	Time (s)	Stationarity residuals
	DIIR-QUIC	0.0205	0.0263	**0.7815**	0.0037	9.65e-06
0.3	ℓ_p COV	0.0191	0.0251	0.7621	0.0006	1.04e-16
	ℓ_p COV (WS)	0.0191	0.0251	0.7621	0.1708	1.04e-16
	DIIR-QUIC	0.0231	0.0267	**0.8635**	0.0025	8.40e-06
0.5	ℓ_p COV	0.0191	0.0251	0.7621	0.0007	1.04e-16
	ℓ_p COV (WS)	0.0191	0.0251	0.7621	0.1138	1.04e-16
	DIIR-QUIC	0.0233	0.0267	**0.8732**	0.0014	3.63e-06
0.7	ℓ_p COV	0.0191	0.0251	0.7621	0.0007	1.04e-16
	ℓ_p COV (WS)	0.0191	0.0251	0.7621	0.0577	1.04e-16

5 Conclusion

In this paper, we have proposed DIIR-QUIC, an inexact QUIC-based iteratively reweighting algorithm tailored for solving log-determinant optimization problems involving nonconvex partly smooth regularizers. We established that, under mild conditions, the inexactly solved subproblems are sufficient to identify the smooth

[2]Please refer to `https://www.eastmoney.com`

active manifold in finitely many iterations. Moreover, we proved the global convergence of the generated iterates, along with convergence rates for both the perturbed objective value sequence and the iterates under Kurdyka-Łojasiewicz property. Finally, extensive numerical experiments on synthetic and real inverse-covariance estimation tasks confirmed that DIIR-QUIC consistently delivered superior computational efficiency and estimation accuracy compared with existing methods, demonstrating its practical value for large-scale nonconvex log-determinant optimization.

A Auxiliary proofs

For the sake of clarity, some technical proofs are deferred to this appendix.

A.1 Proof of Lemma 3.3

Proof. As for statement (i), recall that $\bar{f}$ is differentiable and

$$\partial J(\boldsymbol{\Delta}^{(t)}; \boldsymbol{Z}^t) = \nabla \bar{f}(\boldsymbol{\Delta}^{(t)}; \boldsymbol{Z}^t) + \partial g(\boldsymbol{\Delta}^{(t)}; \boldsymbol{Z}^t).$$

It follows that

$$\begin{aligned}
&\operatorname{dist}(\mathbf{0}, \partial J(\boldsymbol{\Delta}^{(t)}; \boldsymbol{Z}^t)) = \inf_{\boldsymbol{G}^{(t)} \in \partial g(\boldsymbol{\Delta}^{(t)}; \boldsymbol{Z}^t)} \|\nabla \bar{f}(\boldsymbol{\Delta}^{(t)}; \boldsymbol{Z}^t) + \boldsymbol{G}^{(t)}\|_F \\
\leqslant &\inf_{\boldsymbol{G}^{(t)} \in \partial g(\boldsymbol{\Delta}^{(t)}; \boldsymbol{Z}^t)} \|\nabla \bar{f}(\boldsymbol{\Delta}^{(t)}; \boldsymbol{Z}^t) + \boldsymbol{G}^{(t)}\|_1 \\
= &\inf_{\boldsymbol{G}^{(t)} \in \partial g(\boldsymbol{\Delta}^{(t)}; \boldsymbol{Z}^t)} \left\{ \sum_{(i,j) \in \mathcal{I}_{\text{free}}(\boldsymbol{Z}^t)} |\nabla_{ij} \bar{f}(\boldsymbol{\Delta}^{(t)}; \boldsymbol{Z}^t) + G_{ij}^{(t)}| + \right. \\
&\qquad \left. \sum_{(i,j) \in \mathcal{I}_{\text{fixed}}(\boldsymbol{Z}^t)} |\nabla_{ij} \bar{f}(\boldsymbol{\Delta}^{(t)}; \boldsymbol{Z}^t) + G_{ij}^{(t)}| \right\}.
\end{aligned} \tag{53}$$

Define $\varsigma_{ij}(\eta) = J(\boldsymbol{\Delta}^{(t-1)} + \eta(\boldsymbol{e}_i \boldsymbol{e}_j^T + \boldsymbol{e}_j \boldsymbol{e}_i^T); \boldsymbol{Z}^t), \forall (i,j) \in \mathcal{I}_{\text{free}}(\boldsymbol{Z}^t)$ and let $\eta^* \geqslant 0$ denote the minimizer of $\varsigma_{ij}(\eta)$. Then it follows from [21, Theorem 10.1] that $0 \in \partial \varsigma_{ij}(\eta^*), \forall (i,j) \in \mathcal{I}_{\text{free}}(\boldsymbol{Z}^t)$. By the chain rule for subdifferentials (see, e.g., [21, Theorem 10.49]), there exists $S_{ij}^* \in \partial \psi_{ij}(\eta^*)$ with $\psi_{ij}(\eta) = g(\boldsymbol{\Delta}^{(t-1)} + \eta(\boldsymbol{e}_i \boldsymbol{e}_j^T + \boldsymbol{e}_j \boldsymbol{e}_i^T); \boldsymbol{Z}^t)$ such that

$$\langle \nabla \bar{f}(\boldsymbol{\Delta}^{(t-1)} + \eta^*(\boldsymbol{e}_i \boldsymbol{e}_j^T + \boldsymbol{e}_j \boldsymbol{e}_i^T); \boldsymbol{Z}^t), \boldsymbol{e}_i \boldsymbol{e}_j^T + \boldsymbol{e}_j \boldsymbol{e}_i^T \rangle + S_{ij}^* = 0, \forall (i,j) \in \mathcal{I}_{\text{free}}(\boldsymbol{Z}^t).$$

On the other hand, we know the overall update is given by $\boldsymbol{\Delta}^{(t)} = \boldsymbol{\Delta}^{(t-1)} + \eta^*(\boldsymbol{e}_i\boldsymbol{e}_j^T + \boldsymbol{e}_j\boldsymbol{e}_i^T), \forall (i,j) \in \mathcal{I}_{\text{free}}(\boldsymbol{Z}^t)$. Then the coordinate-wise optimality conditions imply that for every updated index $(i,j) \in \mathcal{I}_{\text{free}}(\boldsymbol{Z}^t)$, it holds that

$$|\nabla_{ij}\bar{f}(\boldsymbol{\Delta}^{(t)}; \boldsymbol{Z}^t) + G_{ij}^{(t)}| = 0.$$

Additionally, for any $(i,j) \in \mathcal{I}_{\text{fixed}}(\boldsymbol{Z}^t)$, we have $G_{ij}^{(t-1)} = G_{ij}^{(t)}$. Moreover, by the definition of $\mathcal{I}_{\text{fixed}}(\boldsymbol{Z}^t)$, it holds that $Z_{ij}^t = 0$ and $W_{ij}^k \geqslant |\nabla_{ij} f(\boldsymbol{Z}^t)|$. These conditions, together with the expression for the subdifferential in (2) and $\Delta_{ij}^{(t-1)} = 0$ imply that

$$\inf_{\boldsymbol{G}^{(t-1)} \in \partial g(\boldsymbol{\Delta}^{(t-1)}; \boldsymbol{Z}^t)} \sum_{(i,j) \in \mathcal{I}_{\text{fixed}}(\boldsymbol{Z}^t)} |\nabla_{ij}\bar{f}(\boldsymbol{\Delta}^{(t-1)}; \boldsymbol{Z}^t) + G_{ij}^{(t-1)}| = 0.$$

Combining this condition with (53) yields

$$\begin{aligned}
&\operatorname{dist}(\mathbf{0}, \partial J(\boldsymbol{\Delta}^{(t)}; \boldsymbol{Z}^t)) \\
&= \inf_{\boldsymbol{G}^{(t)} \in \partial g(\boldsymbol{\Delta}^{(t)}; \boldsymbol{Z}^t)} \sum_{(i,j) \in \mathcal{I}_{\text{fixed}}(\boldsymbol{Z}^t)} |\nabla_{ij}\bar{f}(\boldsymbol{\Delta}^{(t)}; \boldsymbol{Z}^t) + G_{ij}^{(t)}| \\
&\overset{(a)}{\leqslant} \inf_{\boldsymbol{G}^{(t-1)} \in \partial g(\boldsymbol{\Delta}^{(t-1)}; \boldsymbol{Z}^t)} \sum_{(i,j) \in \mathcal{I}_{\text{fixed}}(\boldsymbol{Z}^t)} \Big\{ |\nabla_{ij}\bar{f}(\boldsymbol{\Delta}^{(t)}; \boldsymbol{Z}^t) - \nabla_{ij}\bar{f}(\boldsymbol{\Delta}^{(t-1)}; \boldsymbol{Z}^t)| + \\
&\qquad\qquad |\nabla_{ij}\bar{f}(\boldsymbol{\Delta}^{(t-1)}; \boldsymbol{Z}^t) + G_{ij}^{(t-1)}| \Big\} \\
&= \sum_{(i,j) \in \mathcal{I}_{\text{fixed}}(\boldsymbol{Z}^t)} |\nabla_{ij}\bar{f}(\boldsymbol{\Delta}^{(t)}; \boldsymbol{Z}^t) - \nabla_{ij}\bar{f}(\boldsymbol{\Delta}^{(t-1)}; \boldsymbol{Z}^t)| \\
&\quad + \inf_{\boldsymbol{G}^{(t-1)} \in \partial g(\boldsymbol{\Delta}^{(t-1)}; \boldsymbol{Z}^t)} \sum_{(i,j) \in \mathcal{I}_{\text{fixed}}(\boldsymbol{Z}^t)} |\nabla_{ij}\bar{f}(\boldsymbol{\Delta}^{(t-1)}; \boldsymbol{Z}^t) + G_{ij}^{(t-1)}| \\
&= \sum_{(i,j) \in \mathcal{I}_{\text{fixed}}(\boldsymbol{Z}^t)} |\nabla_{ij}\bar{f}(\boldsymbol{\Delta}^{(t)}; \boldsymbol{Z}^t) - \nabla_{ij}\bar{f}(\boldsymbol{\Delta}^{(t-1)}; \boldsymbol{Z}^t)| \\
&\leqslant \|\nabla\bar{f}(\boldsymbol{\Delta}^{(t)}; \boldsymbol{Z}^t) - \nabla\bar{f}(\boldsymbol{\Delta}^{(t-1)}; \boldsymbol{Z}^t)\|_1 \\
&\overset{(b)}{\leqslant} l_k^{-2}\|\boldsymbol{\Delta}^{(t)} - \boldsymbol{\Delta}^{(t-1)}\|_1 \leqslant \epsilon l_k^{-2}\|\boldsymbol{\Delta}^{(t)}\|_1,
\end{aligned} \tag{54}$$

where the inequality (a) follows from the triangle inequality, and inequality (b) holds because Proposition 3.2(ii) applies to $\bar{f}$ as well. Consequently, setting $C_k \coloneqq \epsilon l_k^{-2}$ yields the desired result.

As for statement (ii), by the standard convention in convex analysis, it holds that $Q_k(\boldsymbol{Z}) = +\infty,\ \forall \boldsymbol{Z} \preceq \mathbf{0}$. Note that for any $\boldsymbol{Z}$, $\boldsymbol{D}$, scalar $a \in [0,1]$ and

$\boldsymbol{W} \in \mathbb{R}_+^{n\times n}$, it follows from the convexity and positive homogeneity of the $\|\cdot\|_{1,\boldsymbol{W}^k}$ that

$$\|\boldsymbol{Z}+a\boldsymbol{D}\|_{1,\boldsymbol{W}} = \|a(\boldsymbol{Z}+\boldsymbol{D})+(1-a)\boldsymbol{Z}\|_{1,\boldsymbol{W}} \leqslant a\|\boldsymbol{Z}+\boldsymbol{D}\|_{1,\boldsymbol{W}}+(1-a)\|\boldsymbol{Z}\|_{1,\boldsymbol{W}}. \quad (55)$$

Then, we have for all $\beta_t \in (0,1]$ that

$$\begin{aligned} & Q_k(\boldsymbol{Z}^t+\beta_t\boldsymbol{D}^t) - Q_k(\boldsymbol{Z}^t) \\ &= f(\boldsymbol{Z}^t+\beta_t\boldsymbol{D}^t) - f(\boldsymbol{Z}^t) + \rho\|\boldsymbol{Z}^t+\beta_t\boldsymbol{D}^t\|_{1,\boldsymbol{W}^k} - \rho\|\boldsymbol{Z}^t\|_{1,\boldsymbol{W}^k} \\ &\overset{(a)}{\leqslant} \beta_t\langle\nabla f(\boldsymbol{Z}^t),\boldsymbol{D}^t\rangle - \beta_t\rho(\|\boldsymbol{Z}^t\|_{1,\boldsymbol{W}^k} - \|\boldsymbol{Z}^t+\boldsymbol{D}^t\|_{1,\boldsymbol{W}^k}) \\ &\quad + \beta_t\int_0^1\langle\nabla f(\boldsymbol{Z}^t+\theta\beta_t\boldsymbol{D}^t)-\nabla f(\boldsymbol{Z}^t),\boldsymbol{D}^t\rangle d\theta \overset{(b)}{\leqslant} \beta_t\triangle^t + \frac{\beta_t^2}{2l_k^2}\|\boldsymbol{D}^t\|_F^2, \end{aligned} \quad (56)$$

where inequality (a) follows from the mean value theorem and (55), and inequality (b) holds by (15) together with the the Lipschitz continuity of ∇f as stated in (16). Then, by the strong convexity of $J(\boldsymbol{D};\boldsymbol{Z}^t)$, it follows from (54) that there exists $\bar{\boldsymbol{S}}^t \in \partial J(\boldsymbol{D}^t;\boldsymbol{Z}^t)$ such that

$$\begin{aligned} J(\boldsymbol{D}^t;\boldsymbol{Z}^t) - J(\boldsymbol{0};\boldsymbol{Z}^t) &\leqslant \langle\bar{\boldsymbol{S}}^t,\boldsymbol{D}^t\rangle - \frac{u_k^{-2}}{2}\|\boldsymbol{D}^t\|_F^2 \\ &\leqslant \|\bar{\boldsymbol{S}}^t\|_F\|\boldsymbol{D}^t\|_F^2 - \frac{u_k^{-2}}{2}\|\boldsymbol{D}^t\|_F^2 \leqslant \left(\frac{\epsilon}{l_k^2}-\frac{1}{2u_k^2}\right)\|\boldsymbol{D}^t\|_F^2. \end{aligned}$$

Rearranging gives

$$\begin{aligned} & \left(\frac{\epsilon}{l_k^2}-\frac{1}{2u_k^2}\right)\|\boldsymbol{D}^t\|_F^2 \\ &\geqslant J(\boldsymbol{D}^t;\boldsymbol{Z}^t) - J(\boldsymbol{0};\boldsymbol{Z}^t) \\ &= \langle\nabla f(\boldsymbol{Z}^t),\boldsymbol{D}^t\rangle + \frac{1}{2}\langle\mathrm{vec}(\boldsymbol{D}^t),\nabla^2 f(\boldsymbol{Z}^t)\mathrm{vec}(\boldsymbol{D}^t)\rangle + \rho\|\boldsymbol{Z}^t+\boldsymbol{D}^t\|_{1,\boldsymbol{W}^k} - \rho\|\boldsymbol{Z}^t\|_{1,\boldsymbol{W}^k} \\ &= \triangle^t + \frac{1}{2}\langle\mathrm{vec}(\boldsymbol{D}^t),\nabla^2 f(\boldsymbol{Z}^t)\mathrm{vec}(\boldsymbol{D}^t)\rangle \overset{(a)}{\geqslant} \triangle^t + \frac{u_k^{-2}}{2}\|\boldsymbol{D}^t\|_F^2, \end{aligned}$$

where inequality (a) holds by (17). Then, we have

$$\triangle^t \leqslant \left(\epsilon l_k^{-2} - u_k^{-2}\right)\|\boldsymbol{D}^t\|_F^2. \quad (57)$$

This, together with (56), leads to the desired (19).

On the other hand, combining (57) with the fact that $\triangle^t < 0$, we deduce from (56) that

$$\begin{aligned} Q_k(\boldsymbol{Z}^t+\beta_t\boldsymbol{D}^t) - Q_k(\boldsymbol{Z}^t) &\leqslant \beta_t\triangle^t + \frac{\beta_t^2}{2l_k^2}\|\boldsymbol{D}^t\|_F^2 \\ &\leqslant \left(\beta_t - \frac{l_k^{-2}\beta_t^2}{2(u_k^{-2}-\epsilon l_k^{-2})}\right)\triangle^t. \end{aligned} \quad (58)$$

Therefore, (14) is satisfied whenever

$$\beta_t \triangle^t - \frac{l_k^{-2}(\beta_t)^2}{2(u_k^{-2} - \epsilon l_k^{-2})}\triangle^t \leqslant \beta_t \sigma \triangle^t,$$

which indicates that (14) holds whenever

$$\beta_t \leqslant 2(1-\sigma)(u_k^{-2} - \epsilon l_k^{-2}) l_k^2.$$

Hence, we deduce that (20) holds, where $\pi \in (0,1)$ is introduced to mitigate potential undershooting in the backtracking procedure. In addition, rearranging (19) and letting $t \to +\infty$, it follows from (20) that $\|\boldsymbol{Z}^{t+1} - \boldsymbol{Z}^t\|_F^2 \to 0$, as desired.

(iii). For notational ease, define the function

$$\iota(s) \coloneqq J(s\boldsymbol{D}^t; \boldsymbol{Z}^t),\ \forall s \in [0,1].$$

We first prove that

$$\iota(a) \geqslant \iota(1) - (1-a)C_k\|\boldsymbol{D}^t\|_1^2,\ \forall a \in (0,1]. \tag{59}$$

Since $g(s\boldsymbol{D}^t)$ is convex with respect to $s \in [0,1]$, we denote its right–directional derivative at $s=1$ by

$$\mathrm{Dir}^+ g(\boldsymbol{Z}^t + \boldsymbol{D}^t; \boldsymbol{D}^t) \coloneqq \lim_{h \downarrow 0} \frac{g(\boldsymbol{Z}^t + (1+h)\boldsymbol{D}^t; \boldsymbol{Z}^t) - g(\boldsymbol{Z}^t + \boldsymbol{D}^t; \boldsymbol{Z}^t)}{h}.$$

Thus, we obtain $\iota'_+(1) = \frac{d}{ds}\bar{f}(s\mathbf{D}^{\mathbf{t}}; \boldsymbol{Z}^t)\big|_{s=1} + \mathrm{Dir}^+ g(\boldsymbol{Z}^t + \boldsymbol{D}^t; \boldsymbol{D}^t)$. Due to the convexity of g, there exists a subgradient $\boldsymbol{G}^t \in \partial g(\boldsymbol{Z}^t + \boldsymbol{D}^t; \boldsymbol{Z}^t)$ such that $\mathrm{Dir}^+ g(\boldsymbol{Z}^t + \boldsymbol{D}^t; \boldsymbol{D}^t) = \langle \boldsymbol{G}^t, \boldsymbol{D}^t \rangle$. Therefore, $\iota'_+(1) = \langle \nabla f(\boldsymbol{Z}^t) + \nabla^2 f(\boldsymbol{Z}^t)\boldsymbol{D}^t + \boldsymbol{G}^t, \boldsymbol{D}^t \rangle$.

Let $\boldsymbol{J}^{(t)} \in \partial J(\boldsymbol{D}^t; \boldsymbol{Z}^t)$ such that $\iota'_+(1) = \langle \boldsymbol{J}^{(t)}, \boldsymbol{D}^t \rangle$. Then, by applying the inexactness condition (18) together with the matrix Hölder inequality, we deduce that

$$\iota'_+(1) \leqslant C_k\|\boldsymbol{D}^t\|_1^2.$$

Exploiting the convexity of ι, for any $a \in (0,1)$ we have

$$\iota(1) \leqslant \iota(a) + (1-a)\iota'_+(1) \leqslant \iota(a) + (1-a)C_k\|\boldsymbol{D}^t\|_1^2,$$

Rearranging the above inequality yields the desired (59). Moreover, by (55) and (59), we have

$$\begin{aligned}
&\langle \mathrm{vec}(\nabla f(\boldsymbol{Z}^t)), \mathrm{vec}(\boldsymbol{D}^t) \rangle + \frac{1}{2}\langle \mathrm{vec}(\boldsymbol{D}^t), \nabla^2 f(\boldsymbol{Z}^t)\mathrm{vec}(\boldsymbol{D}^t) \rangle + \rho\|\boldsymbol{Z}^t + \boldsymbol{D}^t\|_{1,\boldsymbol{W}^k} \\
\leqslant\ & a\langle \mathrm{vec}(\nabla f(\boldsymbol{Z}^t)), \mathrm{vec}(\boldsymbol{D}^t) \rangle + \frac{a^2}{2}\langle \mathrm{vec}(\boldsymbol{D}^t), \nabla^2 f(\boldsymbol{Z}^t)\mathrm{vec}(\boldsymbol{D}^t) \rangle + \rho\, a\|\boldsymbol{Z}^t + \boldsymbol{D}^t\|_{1,\boldsymbol{W}^k} \\
&+ (1-a)\rho\|\boldsymbol{Z}^t\|_{1,\boldsymbol{W}^k} + (1-a)C_k\|\boldsymbol{D}^t\|_1^2.
\end{aligned}$$

Rearranging gives

$$0 \geqslant (1-a)\left[\langle \mathrm{vec}(\nabla f(\boldsymbol{Z}^t)), \mathrm{vec}(\boldsymbol{D}^t)\rangle + \rho\|\boldsymbol{Z}^t + \boldsymbol{D}^t\|_{1,\boldsymbol{W}^k} - \rho\|\boldsymbol{Z}^t\|_{1,\boldsymbol{W}^k} - C_k\|\boldsymbol{D}^t\|_1^2\right] + \frac{1}{2}(1-a^2)\langle \mathrm{vec}(\boldsymbol{D}^t), \nabla^2 f(\boldsymbol{Z}^t)\mathrm{vec}(\boldsymbol{D}^t)\rangle.$$

Dividing both sides of the above inequality by $(1-a)$ and then taking $a \uparrow 1$, we have

$$\triangle^t \leqslant -\langle \mathrm{vec}(\boldsymbol{D}^t), \nabla^2 f(\boldsymbol{Z}^t)\mathrm{vec}(\boldsymbol{D}^t)\rangle + C_k\|\boldsymbol{D}^t\|_1^2. \tag{60}$$

Define $\tilde{f}(\varrho) = f(\boldsymbol{Z}^t + \varrho\boldsymbol{D}^t)$. We know that

$$\begin{aligned} |\tilde{f}''(\varrho) - \tilde{f}''(0)| &= |\mathrm{vec}(\boldsymbol{D}^t)^T(\nabla^2 f(\boldsymbol{Z}^t + t\boldsymbol{D}^t) - \nabla^2 f(\boldsymbol{Z}^t))\mathrm{vec}(\boldsymbol{D}^t)| \\ &\overset{(a)}{\leqslant} \|\nabla^2 f(\boldsymbol{Z}^t + t\boldsymbol{D}^t) - \nabla^2 f(\boldsymbol{Z}^t)\|_F\|\boldsymbol{D}^t\|_F^2 \overset{(b)}{\leqslant} \varrho l_k^{-2}\|\boldsymbol{D}^t\|_F^3, \end{aligned} \tag{61}$$

where (a) holds by the Cauchy-Schwartz inequality and (b) holds by the mean value theorem and Proposition 3.2(ii). Then we have that

$$\tilde{f}''(\varrho) \leqslant \tilde{f}''(0) + \varrho l_k^{-2}\|\boldsymbol{D}^t\|_F^3 = \mathrm{vec}(\boldsymbol{D}^t)^T\nabla^2 f(\boldsymbol{Z}^t)\mathrm{vec}(\boldsymbol{D}^t) + \varrho l_k^{-2}\|\boldsymbol{D}^t\|_F^3.$$

We next integrate both sides of the above inequality with respect to $\varrho \in [0,1]$ to obtain an upper bound on $\tilde{f}'(\varrho)$. Note that

$$\int_0^\varrho \tilde{f}''(\varrho)\, d\varrho \leqslant \int_0^\varrho (\tilde{f}''(0) + \varrho l_k^{-2}\|\boldsymbol{D}^t\|_F^3)\, d\varrho,$$

then we have

$$\tilde{f}'(\varrho) \leqslant \tilde{f}'(0) + \varrho\mathrm{vec}(\boldsymbol{D}^t)^T\nabla^2 f(\boldsymbol{Z}^t)\mathrm{vec}(\boldsymbol{D}^t) + \frac{1}{2}\varrho^2 l_k^{-2}\|\boldsymbol{D}^t\|_F^3. \tag{62}$$

We again integrate both sides of (62) with respect to $\varrho \in [0,1]$ to obtain an upper bound on $\tilde{f}(\varrho)$, and we hence have

$$\tilde{f}(\varrho) \leqslant \tilde{f}(0) + \varrho\langle \nabla f(\boldsymbol{Z}^t), \boldsymbol{D}^t\rangle + \frac{1}{2}\varrho^2\mathrm{vec}(\boldsymbol{D}^t)^T\nabla^2 f(\boldsymbol{Z}^t)\mathrm{vec}(\boldsymbol{D}^t) + \frac{1}{6}\varrho^3 l_k^{-2}\|\boldsymbol{D}^t\|_F^3. \tag{63}$$

On the other hand,

$$
\begin{aligned}
&Q_k(\boldsymbol{Z}^t+\boldsymbol{D}^t) = f(\boldsymbol{Z}^t+\boldsymbol{D}^t)+\rho\|\boldsymbol{Z}^t+\boldsymbol{D}^t\|_{1,\boldsymbol{W}^k}\\
&\overset{(a)}{\leqslant} f(\boldsymbol{Z}^t)+\rho\|\boldsymbol{Z}^t\|_{1,\boldsymbol{W}^k}+(\langle\nabla f(\boldsymbol{Z}^t),\boldsymbol{D}^t\rangle+\rho\|\boldsymbol{Z}^t+\boldsymbol{D}^t\|_{1,\boldsymbol{W}^k}-\rho\|\boldsymbol{Z}^t\|_{1,\boldsymbol{W}^k})\\
&\quad+\frac{1}{2}\mathrm{vec}(\boldsymbol{D}^t)^T\nabla^2 f(\boldsymbol{Z}^t)\mathrm{vec}(\boldsymbol{D}^t)+\frac{1}{6}l_k^{-2}\|\boldsymbol{D}^t\|_F^3\\
&\leqslant Q_k(\boldsymbol{Z}^t)+\triangle^t+\frac{1}{2}\mathrm{vec}(\boldsymbol{D}^t)^T\nabla^2 f(\boldsymbol{Z}^t)\mathrm{vec}(\boldsymbol{D}^t)+\frac{1}{6}l_k^{-2}\|\boldsymbol{D}^t\|_F^3\\
&\overset{(b)}{\leqslant} Q_k(\boldsymbol{Z}^t)+\frac{\triangle^t}{2}+\frac{n\,\epsilon l_k^{-2}}{2}\|\boldsymbol{D}^t\|_F^2+\frac{1}{6}l_k^{-2}\|\boldsymbol{D}^t\|_F^3,\\
&\overset{(c)}{\leqslant} Q_k(\boldsymbol{Z}^t)+\left(\left(\frac{1}{2}-\frac{n\epsilon l_k^{-2}}{2(u_k^{-2}-\epsilon l_k^{-2})}\right)-\frac{l_k^{-2}}{6(u_k^{-2}-\epsilon l_k^{-2})}\|\boldsymbol{D}^t\|_F\right)\triangle^t\\
&\overset{(d)}{\leqslant} Q_k(\boldsymbol{Z}^t)+\sigma\triangle^t,
\end{aligned}
$$

where (a) follows from (63) with $\varrho=1$, (b) follows from (60) and the fact that $\|\boldsymbol{D}\|_1\leqslant n\|\boldsymbol{D}\|_F, \forall \boldsymbol{D}\in\mathbb{S}^n$, (c) follows from (57) and (d) holds provided

$$
\left(\frac{1}{2}-\frac{n\epsilon l_k^{-2}}{2(u_k^{-2}-\epsilon l_k^{-2})}\right)-\frac{l_k^{-2}}{6(u_k^{-2}-\epsilon l_k^{-2})}\|\boldsymbol{D}^t\|_F>\sigma,
$$

which is guaranteed by $\sigma\in(0,0.5-\epsilon_\sigma)$, by (21), by the boundedness of β_t in statement (ii) and for $0<\epsilon\ll 1$. Therefore, the line-search condition (14) holds with $\beta_t=1$. This completes the proof. □

A.2 Proof of Proposition 3.5

Proof. (i). By the concavity of $\phi(\cdot)$ over $\mathbb{R}_{++}$, we have for each $(i,j)\in[n]\times[n]$ that

$$
\phi(|X_{ij}^{k+1}|+\mathcal{E}_{ij}^k)\leqslant\phi(|X_{ij}^k|+\mathcal{E}_{ij}^k)+\phi'(|X_{ij}^k|+\mathcal{E}_{ij}^k)(|X_{ij}^{k+1}|-|X_{ij}^k|). \tag{64}
$$

Then

$$
\begin{aligned}
&F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k) - F(\boldsymbol{X}^{k+1}, \boldsymbol{\mathcal{E}}^k) \\
&= f(\boldsymbol{X}^k) - f(\boldsymbol{X}^{k+1}) + \rho \sum_{ij} \phi(|X_{ij}^k| + \mathcal{E}_{ij}^k) - \rho \sum_{ij} \phi(|X_{ij}^{k+1}| + \mathcal{E}_{ij}^k) \\
&\overset{\text{Eq.(64)}}{\geqslant} f(\boldsymbol{X}^k) - f(\boldsymbol{X}^{k+1}) + \rho \sum_{ij} W_{ij}^k (|X_{ij}^k| - |X_{ij}^{k+1}|) \\
&\overset{(a)}{=} f(\boldsymbol{X}^k) + \rho \sum_{ij} W_{ij}^k |X_{ij}^k| - \left(f((1-\alpha)\boldsymbol{X}^k + \alpha \boldsymbol{Y}^k) + \rho \sum_{ij} W_{ij}^k |\alpha Y_{ij}^k + (1-\alpha) X_{ij}^k| \right) \\
&\overset{(b)}{\geqslant} \alpha \left(f(\boldsymbol{X}^k) - f(\boldsymbol{Y}^k) + \rho \sum_{ij} W_{ij}^k (|X_{ij}^k| - |Y_{ij}^k|) \right) = \alpha (Q_k(\boldsymbol{X}^k) - Q_k(\boldsymbol{Y}^k)),
\end{aligned} \tag{65}
$$

where inequality (a) uses the identity $\boldsymbol{X}^{k+1} = (1-\alpha)\boldsymbol{X}^k + \alpha \boldsymbol{Y}^k$ and inequality (b) holds due to the convexity of f and $|\cdot|$. Let $K_k \in \mathbb{N}$ denote the number of iterations required to obtain $\boldsymbol{Y}^k$ for the kth subproblem ($\mathcal{P}_{\text{sub}}$). Starting from the initialization $\boldsymbol{Z}^0 = \boldsymbol{X}^k$, it follows that

$$
\begin{aligned}
Q_k(\boldsymbol{X}^k) - Q_k(\boldsymbol{Y}^k) &= \sum_{t=0}^{K_k - 1} Q_k(\boldsymbol{Z}^t) - Q_k(\boldsymbol{Z}^{t+1}) \\
&\overset{\text{Eq.(19)}}{\geqslant} \sum_{t=0}^{K_k - 1} \left(\beta_t (u_k^{-2} - \epsilon l_k^{-2}) - \frac{\beta_t^2}{2 l_k^2} \right) \|\boldsymbol{Z}^t - \boldsymbol{Z}^{t+1}\|_F^2 \\
&\overset{(a)}{\geqslant} \sum_{t=0}^{K_k - 1} \left(\tilde{\beta} (u^{-2} - \epsilon l^{-2}) - \frac{1}{2 l^2} \right) \|\boldsymbol{Z}^t - \boldsymbol{Z}^{t+1}\|_F^2 \\
&\overset{(b)}{\geqslant} \left(\tilde{\beta} (u^{-2} - \epsilon l^{-2}) - \frac{1}{2 l^2} \right) \|\boldsymbol{X}^k - \boldsymbol{Y}^k\|_F^2 \\
&\overset{(c)}{=} \left(\frac{\tilde{\beta} (u^{-2} - \epsilon l^{-2})}{\alpha^2} - \frac{1}{2 l^2 \alpha^2} \right) \|\boldsymbol{X}^k - \boldsymbol{X}^{k+1}\|_F^2,
\end{aligned}
$$

where (a) follows from $\beta_t \leqslant 1$, (23) and $\tilde{\beta} \coloneqq \min\left\{1, 2\pi(1-\sigma)(u^{-2} - \epsilon l^{-2}) l^2\right\}$, (b) holds by the triangle inequality, and (c) results from the identity $\boldsymbol{Y}^k = \frac{1}{\alpha}(\boldsymbol{X}^{k+1} - \boldsymbol{X}^k) + \boldsymbol{X}^k$. This, together with (65), gives

$$
\begin{aligned}
F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k) - F(\boldsymbol{X}^{k+1}, \boldsymbol{\mathcal{E}}^{k+1}) &\geqslant F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k) - F(\boldsymbol{X}^{k+1}, \boldsymbol{\mathcal{E}}^k) \\
&\geqslant \left(\frac{\tilde{\beta} (u^{-2} - \epsilon l^{-2})}{\alpha} - \frac{1}{2 l^2 \alpha} \right) \|\boldsymbol{X}^k - \boldsymbol{X}^{k+1}\|_F^2.
\end{aligned} \tag{66}
$$

(ii) Rearranging and summing up both sides of inequality (66) from $k = 0$ to

$\bar{k}-1$ gives

$$\begin{aligned}
&\left(\frac{\tilde{\beta}(u^{-2}-\epsilon l^{-2})}{\alpha}-\frac{1}{2l^2\alpha}\right)\sum_{k=0}^{\bar{k}-1}\|\boldsymbol{X}^k-\boldsymbol{X}^{k+1}\|_F^2\\
\leqslant&\sum_{k=0}^{\bar{k}-1}F(\boldsymbol{X}^k,\boldsymbol{\mathcal{E}}^k)-F(\boldsymbol{X}^{k+1},\boldsymbol{\mathcal{E}}^{k+1})\\
\leqslant& F(\boldsymbol{X}^0,\boldsymbol{\mathcal{E}}^0)-F(\boldsymbol{X}^{\bar{k}},\boldsymbol{\mathcal{E}}^{\bar{k}})\overset{(a)}{\leqslant}F(\boldsymbol{X}^0,\boldsymbol{\mathcal{E}}^0)-\inf_{\boldsymbol{X}\in\mathrm{Lev}_{F(\boldsymbol{X},\boldsymbol{\mathcal{E}})}(\boldsymbol{X}^0,\boldsymbol{\mathcal{E}}^0)}F(\boldsymbol{X},\boldsymbol{\mathcal{E}})<+\infty,
\end{aligned}$$

where inequality (a) holds since the value $\inf_{\boldsymbol{X}\in\mathrm{Lev}_{F(\boldsymbol{X},\boldsymbol{\mathcal{E}})}(\boldsymbol{X}^0,\boldsymbol{\mathcal{E}}^0)}F(\boldsymbol{X},\boldsymbol{\mathcal{E}})$ is finite by [21, Theorem 1.9]. Let $\bar{k}\to+\infty$, and it leads us to the desired results.

(iii) It follows from $\boldsymbol{X}^0\in\chi$ and $\{\boldsymbol{Y}^k\}\subset\chi,\ \forall k$ (by Proposition 3.2(i)) that $\{\boldsymbol{X}^k\}\subset\chi$. This completes the proof. □

A.3 Proof of Lemma 3.8

Proof. (i) It is immediate that $\boldsymbol{\mathcal{M}}(\boldsymbol{X}^*)$ forms a smooth manifold in a neighborhood of $\boldsymbol{X}^*$. We next claim that Φ is partly smooth at $\boldsymbol{X}^*$ relative to $\boldsymbol{\mathcal{M}}(\boldsymbol{X}^*)$. To see this, observe first that for all $\boldsymbol{X}\in\boldsymbol{\mathcal{M}}(\boldsymbol{X}^*)$ we have

$$\begin{aligned}
\Phi|_{\boldsymbol{\mathcal{M}}(\boldsymbol{X}^*)}(\boldsymbol{X})&=\sum_{(i,j)\in\mathcal{I}(\boldsymbol{X}^*)}\phi(|X_{ij}|)+\sum_{(i,j)\in\mathcal{Z}(\boldsymbol{X}^*)}\phi(|X_{ij}|)\\
&=\sum_{(i,j)\in\mathcal{I}(\boldsymbol{X}^*)}\phi(\mathrm{sgn}(\boldsymbol{X}^*_{ij})X_{ij}),
\end{aligned}\tag{67}$$

so $\Phi|_{\boldsymbol{\mathcal{M}}(\boldsymbol{X}^*)}$ is smooth. Second, since Φ is continuous by Assumption 1.1 and $\Phi|_{\boldsymbol{\mathcal{M}}(\boldsymbol{X}^*)}$ is smooth, we then know that Φ is regular [21, Example 7.28] at each $\boldsymbol{X}\in\boldsymbol{\mathcal{M}}(\boldsymbol{X}^*)$ with $\frac{\partial\Phi}{\partial X_{ij}}=\left\{\frac{d\phi(\mathrm{sgn}(\boldsymbol{X}^*_{ij})X_{ij})}{X_{ij}}\right\}$, for any $(i,j)\in\mathcal{I}(\boldsymbol{X}^*)$ by [21, Exercise 8.8]. Next, we check the local normal sharpness condition. Routine calculation (e.g., using [21, Example 6.8]) shows that at any point $\boldsymbol{X}\in\boldsymbol{\mathcal{M}}(\boldsymbol{X}^*)$, we have

$$\begin{aligned}
N_{\boldsymbol{\mathcal{M}}(\boldsymbol{X}^*)}(\boldsymbol{X})&=\{\boldsymbol{H}\in\mathbb{S}^n\mid H_{ij}=0,\forall(i,j)\in\mathcal{I}(\boldsymbol{X}^*)\},\\
\mathrm{par}\,\partial\Phi(\boldsymbol{X})&=\{\boldsymbol{H}\in\mathbb{S}^n\mid H_{ij}=0,\forall(i,j)\in\mathcal{I}(\boldsymbol{X}^*)\}.
\end{aligned}\tag{68}$$

Hence the normal space is parallel to the subdifferential. Finally, the subdifferential map $\partial\Phi$ is continuous relative to $\boldsymbol{\mathcal{M}}(\boldsymbol{X}^*)$ since $\boldsymbol{\mathcal{M}}(\boldsymbol{X}^*)$ contains Φ to a smooth subspace by (67). On the other hand, by [21, Proposition 13.34], we also know that Φ is prox-regular at $\boldsymbol{X}^*$ relative to $\boldsymbol{\mathcal{M}}(\boldsymbol{X}^*)$. Consequently, it follows from

[22, Corollary 4.7] and $f \in \mathcal{C}^1$ that F is partly smooth at $\boldsymbol{X}^*$ relative to $\boldsymbol{\mathcal{M}}(\boldsymbol{X}^*)$, and prox-regular there.

(ii) Theorem 3.6(iii), together with $f \in \mathcal{C}^1$ and Assumption 1.1, gives $F(\boldsymbol{X}^k) \to F(\boldsymbol{X}^*)$. Thus, the premises in Proposition 1.9 on F are satisfied. Note that

$$\begin{aligned} \operatorname{dist}(\mathbf{0}, \partial F(\boldsymbol{X}^k)) \to 0 &\iff \operatorname{dist}(-\nabla f(\boldsymbol{X}^k), \partial \Phi(\boldsymbol{X}^k)) \to 0 \\ &\overset{(a)}{\iff} \operatorname{dist}(-\nabla f(\boldsymbol{X}^*), \partial \Phi(\boldsymbol{X}^*)) = 0, \end{aligned} \tag{69}$$

where (a) follows from $\nabla f(\boldsymbol{X}^k) \to \nabla f(\boldsymbol{X}^*)$ and the non-degenerate condition (30). By Proposition 1.9, we know that $\boldsymbol{X}^k \in \boldsymbol{\mathcal{M}}(\boldsymbol{X}^*)$ for all sufficiently large $k \in \mathbb{N}$. This completes the proof. □

B Tests with other nonconvex regularizers

For completeness, we provide additional experimental results with $p = 0.3$ (Figure 6, Figure 8 and Table 4) and $p = 0.7$ (Figure 7, Figure 9 and Table 5). Moreover, we validate the performance of DIIR-QUIC by incorporating nonconvex SCAD and MCP penalties. Synthetic precision-matrix estimation problems of size $n \in \{500, 1000, 2000\}$ were solved under both penalty types. Figure 10 summarizes the results.

References

[1] D. Bertsimas, J. Lamperski, and J. Pauphilet, "Certifiably optimal sparse inverse covariance estimation," *Mathematical Programming*, vol. 184, no. 1, pp. 491–530, 2020.

[2] M. Gulliksson and S. Mazur, "An iterative approach to ill-conditioned optimal portfolio selection," *Computational Economics*, vol. 56, no. 4, pp. 773–794, 2020.

[3] T. T. Cai, H. Li, W. Liu, and J. Xie, "Covariate-adjusted precision matrix estimation with an application in genetical genomics," *Biometrika*, vol. 100, no. 1, pp. 139–156, 2013.

[4] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

Table 3: Examples of the function Φ with tuning parameters $p \in (0,1)$, $\varpi > 0$, and SCAD/MCP parameters (λ, a) or (λ, γ). Here $\phi'(|X_{ij}|)$ is the derivative at $|X_{ij}|$, and $(\phi')^{-1}(s)$ is the inverse mapping defined on the strictly decreasing portion $s \in (0, \phi'(0^+))$.

Regularizer	$\Phi(\mathbf{X})$	$\phi'(\lvert X_{ij}\rvert)$	$\phi'(0^+)$	$\phi'(\infty)$	$(\phi')^{-1}(s)$	Flat-tail t^*
ℓ_p (quasi-norm) [13]	$\sum_{ij} \lvert X_{ij}\rvert^p$	$p\,\lvert X_{ij}\rvert^{p-1}$	$+\infty$	0	$\left(\frac{s}{p}\right)^{1/(p-1)}$	∞
Log-sum [42]	$\sum_{ij} \log\!\left(1 + \frac{\lvert X_{ij}\rvert}{\varpi}\right)$	$\frac{1}{\lvert X_{ij}\rvert + \varpi}$	$\frac{1}{\varpi}$	0	$\frac{1}{s} - \varpi$	∞
Geman [43]	$\sum_{ij} \frac{\lvert X_{ij}\rvert}{\lvert X_{ij}\rvert + \varpi}$	$\frac{\varpi}{(\lvert X_{ij}\rvert + \varpi)^2}$	$\frac{1}{\varpi}$	0	$\sqrt{\frac{\varpi}{s}} - \varpi$	∞
Arctan [44]	$\sum_{ij} \arctan\!\left(\frac{\lvert X_{ij}\rvert}{\varpi}\right)$	$\frac{\varpi}{\varpi^2 + \lvert X_{ij}\rvert^2}$	$\frac{1}{\varpi}$	0	$\sqrt{\frac{\varpi}{s} - \varpi^2}$	∞
Exp [45]	$\sum_{ij} \left(1 - e^{-\lvert X_{ij}\rvert/\varpi}\right)$	$\frac{1}{\varpi} e^{-\lvert X_{ij}\rvert/\varpi}$	$\frac{1}{\varpi}$	0	$-\varpi \ln(\varpi\, s)$	∞
SCAD [14]	$\sum_{ij} \phi_{\mathrm{SCAD}}(\lvert X_{ij}\rvert)$	$\begin{cases} \lambda, & \lvert X_{ij}\rvert \leqslant \lambda, \\ \frac{a\lambda - \lvert X_{ij}\rvert}{a-1}, & \lambda < \lvert X_{ij}\rvert \leqslant a\lambda, \\ 0, & \lvert X_{ij}\rvert > a\lambda, \end{cases}$	λ	0	$a\lambda - (a-1)\, s$	$a\lambda$
MCP [15]	$\sum_{ij} \phi_{\mathrm{MCP}}(\lvert X_{ij}\rvert)$	$\begin{cases} \lambda - \frac{\lvert X_{ij}\rvert}{\gamma}, & \lvert X_{ij}\rvert \leqslant \gamma\lambda, \\ 0, & \lvert X_{ij}\rvert > \gamma\lambda, \end{cases}$	λ	0	$\gamma\,(\lambda - s)$	$\gamma\lambda$

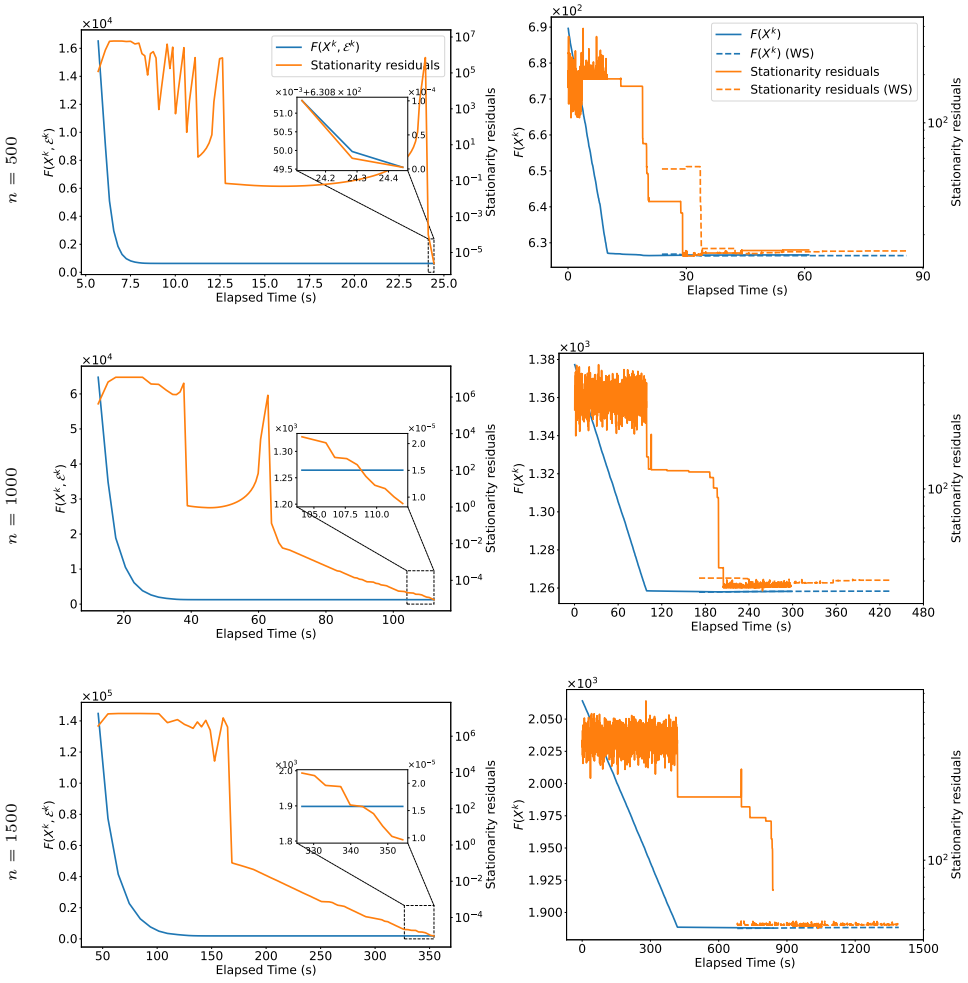

Figure 6: Convergence plots of objective value and stationarity residual (plotted on a logarithmic scale) versus elapsed time (seconds) for DIIR-QUIC and ℓ_pCOV on a tridiagonal precision matrix with $p = 0.3$.

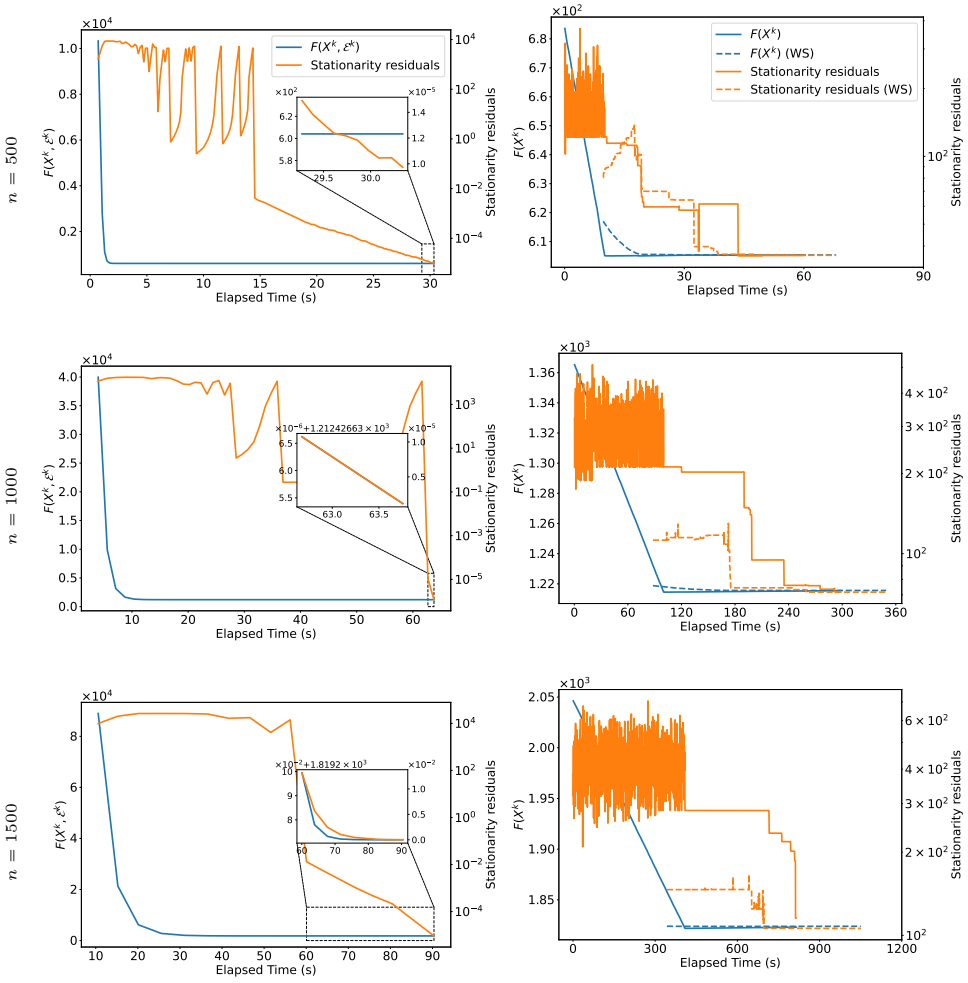

Figure 7: Convergence plots of objective value and stationarity residual (plotted on a logarithmic scale) versus elapsed time (seconds) for DIIR-QUIC and ℓ_pCOV on a tridiagonal precision matrix with $p = 0.7$.

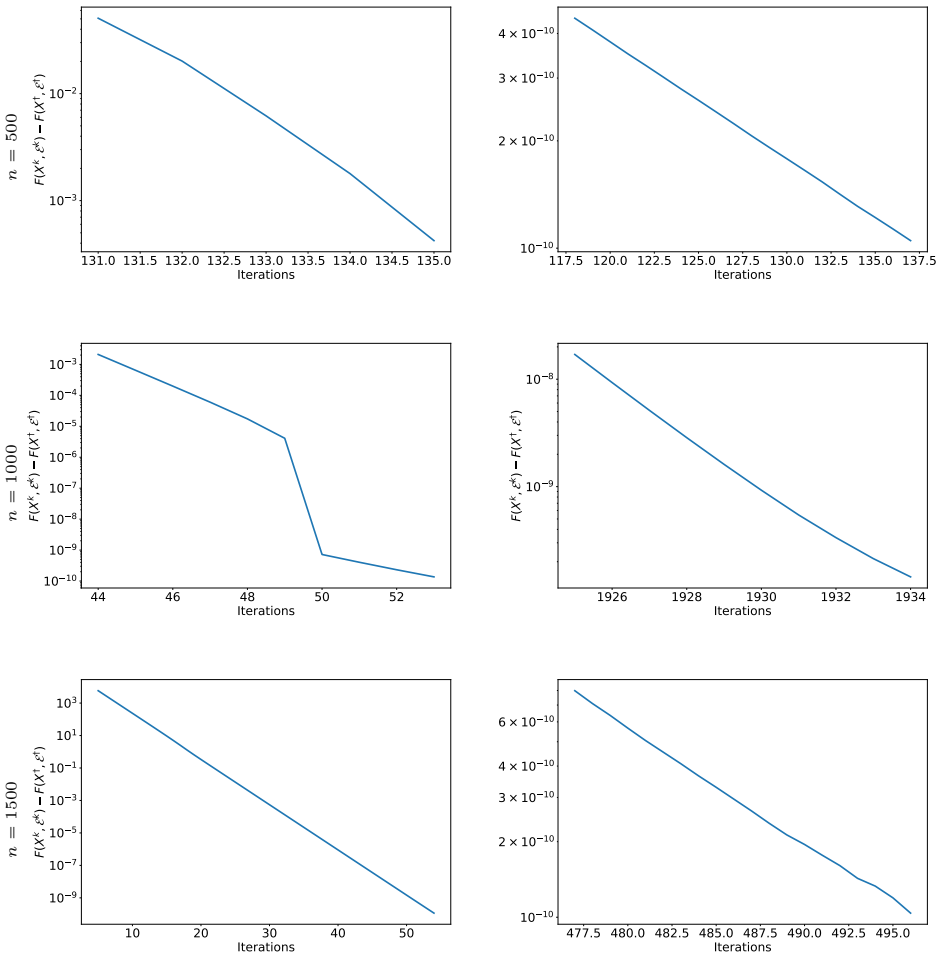

Figure 8: Q-linear convergence of the perturbed objective error $F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k) - F(\boldsymbol{X}^\dagger, \boldsymbol{\mathcal{E}}^\dagger)$ plotted versus iteration for DIIR-QUIC on tridiagonal precision matrices and clustered matrices with $p = 0.3$, across varying matrix dimensions. The error is displayed on a logarithmic scale, and only values above 10^{-8} are shown.

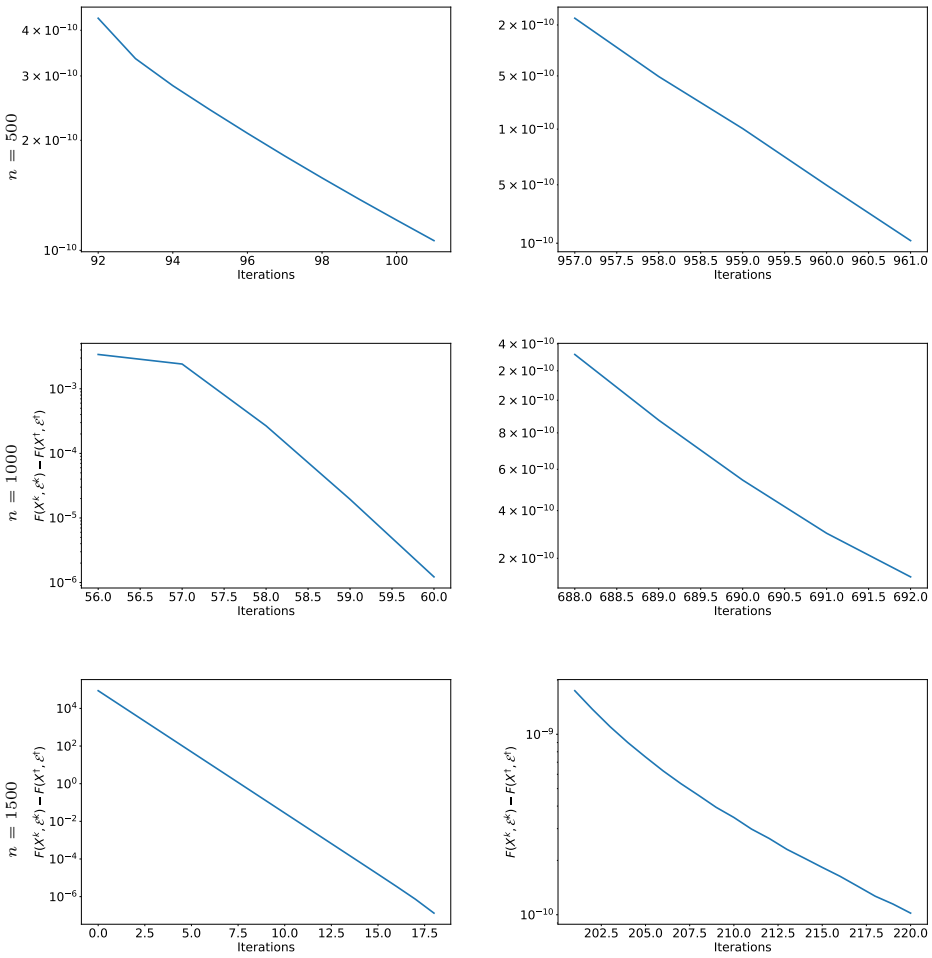

Figure 9: Q-linear convergence of the perturbed objective error $F(\boldsymbol{X}^k, \boldsymbol{\mathcal{E}}^k) - F(\boldsymbol{X}^\dagger, \boldsymbol{\mathcal{E}}^\dagger)$ plotted versus iteration for DIIR-QUIC on tridiagonal precision matrices and clustered matrices with $p = 0.7$, across varying matrix dimensions. The error is displayed on a logarithmic scale, and only values above 10^{-8} are shown.

Table 4: Comparison of DIIR-QUIC and ℓ_pCOV on synthetic covariance estimation. For each method, the table reports the mean ($\pm$ one standard deviation) over 10 independent trials under two covariance structures—tridiagonal (Tri) and clustered (Clu)—with $p = 0.3$ at dimensions $n = 500, 1000$, and 1500. Boldface entries indicate the better performance.

Data	n	Algorithm	$\mathbf{F}_1$	Time (s)	nnz(Σ^{-1})	nnz($X^\dagger$)	Loss Q	KL Loss	Sensitivity	Specificity	MCC
tri	500	DIIR-QUIC	0.959 (0.002)	**20.296 (6.954)**	1498.000 (0.000)	1513.200 (7.111)	0.011 (0.000)	0.027 (0.001)	0.964 (0.001)	**1.000 (0.000)**	0.959 (0.002)
		ℓ_pCOV	0.995 (0.001)	59.292 (1.507)	1498.000 (0.000)	1492.800 (2.857)	0.009 (0.000)	0.013 (0.001)	0.994 (0.002)	**1.000 (0.000)**	0.995 (0.001)
		ℓ_pCOV (WS)	**0.999 (0.001)**	78.470 (2.660)	1498.000 (0.000)	1495.800 (1.887)	**0.008 (0.000)**	**0.012 (0.001)**	**0.998 (0.001)**	**1.000 (0.000)**	**0.999 (0.001)**
	1000	DIIR-QUIC	0.971 (0.001)	**112.508 (35.052)**	2998.000 (0.000)	2976.600 (5.219)	0.007 (0.000)	0.022 (0.001)	0.968 (0.001)	**1.000 (0.000)**	0.971 (0.001)
		ℓ_pCOV	**1.000 (0.000)**	248.449 (24.727)	2998.000 (0.000)	2997.600 (0.800)	**0.004 (0.000)**	**0.007 (0.000)**	**1.000 (0.000)**	**1.000 (0.000)**	**1.000 (0.000)**
		ℓ_pCOV (WS)	**1.000 (0.000)**	381.302 (53.667)	2998.000 (0.000)	2997.800 (0.600)	**0.004 (0.000)**	**0.007 (0.000)**	**1.000 (0.000)**	**1.000 (0.000)**	**1.000 (0.000)**
	1500	DIIR-QUIC	0.973 (0.001)	**397.025 (63.827)**	4498.000 (0.000)	4462.200 (9.775)	0.005 (0.000)	0.020 (0.000)	0.970 (0.001)	**1.000 (0.000)**	0.973 (0.001)
		ℓ_pCOV	**1.000 (0.000)**	592.189 (119.818)	4498.000 (0.000)	4498.000 (0.000)	**0.003 (0.000)**	0.007 (0.000)	**1.000 (0.000)**	**1.000 (0.000)**	**1.000 (0.000)**
		ℓ_pCOV (WS)	**1.000 (0.000)**	1113.915 (172.059)	4498.000 (0.000)	4498.000 (0.000)	**0.003 (0.000)**	**0.006 (0.000)**	**1.000 (0.000)**	**1.000 (0.000)**	**1.000 (0.000)**
clu	500	DIIR-QUIC	**0.333 (0.009)**	**49.281 (22.177)**	11432.800 (158.622)	2974.200 (71.181)	**0.039 (0.000)**	**0.279 (0.004)**	**0.210 (0.007)**	0.998 (0.000)	**0.400 (0.009)**
		ℓ_pCOV	0.146 (0.003)	108.808 (22.124)	11432.800 (158.622)	905.200 (21.264)	0.041 (0.000)	0.327 (0.005)	0.079 (0.002)	**1.000 (0.000)**	0.273 (0.004)
		ℓ_pCOV (WS)	0.155 (0.005)	108.811 (22.124)	11432.800 (158.622)	970.000 (26.803)	0.041 (0.000)	0.324 (0.005)	0.084 (0.003)	**1.000 (0.000)**	0.282 (0.005)
	1000	DIIR-QUIC	**0.539 (0.006)**	**613.051 (274.081)**	23093.400 (193.779)	10933.800 (113.900)	**0.024 (0.000)**	**0.208 (0.003)**	**0.397 (0.006)**	0.998 (0.000)	**0.571 (0.005)**
		ℓ_pCOV	0.287 (0.003)	672.529 (274.001)	23093.400 (193.779)	3915.000 (44.553)	0.027 (0.000)	0.275 (0.003)	0.168 (0.002)	**1.000 (0.000)**	0.403 (0.002)
		ℓ_pCOV (WS)	0.324 (0.006)	672.537 (273.988)	23093.400 (193.779)	4513.600 (83.258)	0.026 (0.000)	0.262 (0.004)	0.194 (0.004)	**1.000 (0.000)**	0.434 (0.005)
	1500	DIIR-QUIC	**0.682 (0.003)**	**4185.551 (1598.464)**	34701.000 (233.142)	30211.200 (277.186)	**0.017 (0.000)**	**0.140 (0.001)**	**0.638 (0.004)**	0.996 (0.000)	0.679 (0.003)
		ℓ_pCOV	0.559 (0.003)	4245.262 (1598.553)	34701.000 (233.142)	13793.800 (149.019)	0.019 (0.000)	0.182 (0.002)	0.391 (0.003)	**1.000 (0.000)**	0.617 (0.002)
		ℓ_pCOV (WS)	0.643 (0.002)	4245.306 (1598.517)	34701.000 (233.142)	16760.200 (152.054)	**0.017 (0.000)**	0.153 (0.002)	0.477 (0.003)	**1.000 (0.000)**	**0.683 (0.002)**

Table 5: Comparison of DIIR-QUIC and ℓ_pCOV on synthetic covariance estimation. For each method, the table reports the mean ($\pm$ one standard deviation) over 10 independent trials under two covariance structures—tridiagonal (Tri) and clustered (Clu)—with $p = 0.7$ at dimensions $n = 500, 1000$, and 1500. Boldface entries indicate the better performance.

Data	n	Algorithm	$\mathbf{F}_1$	Time (s)	nnz(Σ^{-1})	nnz($X^\dagger$)	Loss Q	KL Loss	Sensitivity	Specificity	MCC
tri	500	DIIR-QUIC	0.874 (0.005)	**43.538 (27.415)**	1498.000 (0.000)	1925.000 (18.852)	0.012 (0.000)	0.032 (0.001)	0.998 (0.001)	0.998 (0.000)	0.880 (0.004)
		ℓ_pCOV	**0.944 (0.003)**	58.648 (1.675)	1498.000 (0.000)	1674.400 (10.575)	**0.010 (0.000)**	**0.020 (0.001)**	**1.000 (0.000)**	**0.999 (0.000)**	**0.946 (0.003)**
		ℓ_pCOV (WS)	0.943 (0.003)	66.769 (3.191)	1498.000 (0.000)	1677.600 (11.586)	**0.010 (0.000)**	**0.020 (0.001)**	**1.000 (0.000)**	**0.999 (0.000)**	0.945 (0.003)
	1000	DIIR-QUIC	0.995 (0.001)	**82.146 (32.319)**	2998.000 (0.000)	3018.800 (5.075)	0.007 (0.000)	0.025 (0.000)	0.998 (0.000)	**1.000 (0.000)**	0.995 (0.001)
		ℓ_pCOV	**1.000 (0.000)**	252.932 (27.301)	2998.000 (0.000)	3000.800 (1.833)	**0.005 (0.000)**	**0.013 (0.000)**	**1.000 (0.000)**	**1.000 (0.000)**	**1.000 (0.000)**
		ℓ_pCOV (WS)	0.999 (0.000)	303.844 (50.265)	2998.000 (0.000)	3001.200 (2.400)	**0.005 (0.000)**	**0.013 (0.000)**	**1.000 (0.000)**	**1.000 (0.000)**	0.999 (0.000)
	1500	DIIR-QUIC	0.998 (0.000)	**128.492 (28.726)**	4498.000 (0.000)	4502.600 (1.562)	0.005 (0.000)	0.024 (0.000)	0.999 (0.000)	**1.000 (0.000)**	0.998 (0.000)
		ℓ_pCOV	**1.000 (0.000)**	588.834 (93.311)	4498.000 (0.000)	4498.000 (0.000)	**0.004 (0.000)**	**0.012 (0.000)**	**1.000 (0.000)**	**1.000 (0.000)**	**1.000 (0.000)**
		ℓ_pCOV (WS)	**1.000 (0.000)**	856.249 (128.172)	4498.000 (0.000)	4498.000 (0.000)	**0.004 (0.000)**	**0.012 (0.000)**	**1.000 (0.000)**	**1.000 (0.000)**	**1.000 (0.000)**
clu	500	DIIR-QUIC	0.428 (0.005)	**75.779 (26.243)**	11522.200 (111.905)	9139.600 (128.473)	**0.036 (0.000)**	**0.247 (0.001)**	**0.384 (0.006)**	0.980 (0.000)	0.407 (0.005)
		ℓ_pCOV	0.430 (0.007)	135.312 (26.093)	11522.200 (111.905)	7171.600 (106.714)	0.039 (0.000)	0.251 (0.001)	0.349 (0.007)	**0.987 (0.000)**	0.422 (0.006)
		ℓ_pCOV (WS)	**0.440 (0.006)**	135.311 (26.093)	11522.200 (111.905)	7362.600 (105.864)	0.039 (0.000)	0.248 (0.001)	0.361 (0.007)	**0.987 (0.000)**	**0.431 (0.006)**
	1000	DIIR-QUIC	0.503 (0.005)	**1222.015 (210.478)**	23026.600 (212.091)	36873.600 (438.393)	**0.022 (0.000)**	0.173 (0.002)	**0.654 (0.007)**	0.978 (0.000)	0.503 (0.005)
		ℓ_pCOV	0.558 (0.005)	1281.443 (210.630)	23026.600 (212.091)	28733.200 (342.888)	0.023 (0.000)	0.169 (0.002)	0.627 (0.007)	**0.985 (0.000)**	0.550 (0.005)
		ℓ_pCOV (WS)	**0.569 (0.005)**	1281.438 (210.626)	23026.600 (212.091)	29595.600 (353.259)	0.023 (0.000)	**0.166 (0.002)**	0.650 (0.006)	**0.985 (0.000)**	**0.562 (0.005)**
	1500	DIIR-QUIC	0.611 (0.006)	**5759.561 (3675.491)**	34713.600 (256.051)	42335.800 (313.913)	**0.015 (0.000)**	0.145 (0.002)	**0.678 (0.008)**	0.992 (0.000)	0.607 (0.007)
		ℓ_pCOV	0.663 (0.006)	5818.995 (3675.481)	34713.600 (256.051)	32665.200 (258.192)	0.016 (0.000)	0.140 (0.002)	0.644 (0.007)	**0.995 (0.000)**	0.659 (0.007)
		ℓ_pCOV (WS)	**0.678 (0.007)**	5818.996 (3675.463)	34713.600 (256.051)	33525.200 (260.962)	**0.015 (0.000)**	**0.136 (0.002)**	0.667 (0.008)	**0.995 (0.000)**	**0.674 (0.007)**

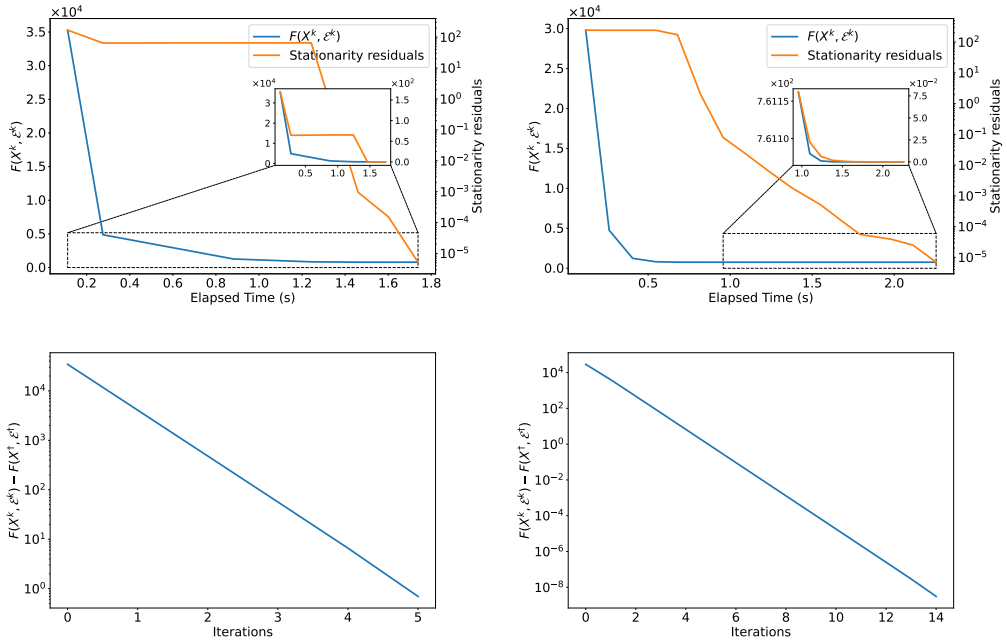

Figure 10: Convergence behavior of DIIR-QUIC on a tridiagonal synthetic precision matrix ($n = 500$) with SCAD and MCP penalties. **Top row:** Evaluation of the penalized objective value and stationarity residuals versus elapsed time. **Bottom row:** Q-linear convergence of the objective error in log scale.

[5] C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, P. Ravikumar, *et al.*, "QUIC: quadratic approximation for sparse inverse covariance estimation.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2911–2947, 2014.

[6] D. N. Phan, H. A. Le Thi, and T. P. Dinh, "Sparse covariance matrix estimation by DCA-based algorithms," *Neural Computation*, vol. 29, no. 11, pp. 3040–3077, 2017.

[7] T. Nakagaki, M. Fukuda, S. Kim, and M. Yamashita, "A dual spectral projected gradient method for log-determinant semidefinite problems," *Computational Optimization and Applications*, vol. 76, pp. 33–68, 2020.

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

[9] D. Bertsimas, A. King, and R. Mazumder, "Best subset selection via a modern optimization lens," *The Annals of Statistics*, vol. 44, no. 2, pp. 813–852, 2016.

[10] J. Fan, H. Liu, Q. Sun, and T. Zhang, "I-lamm for sparse learning: Simultaneous control of algorithmic complexity and statistical error," *Annals of Statistics*, vol. 46, no. 2, p. 814, 2018.

[11] G. Marjanovic and V. Solo, "On ℓ_q optimization and sparse inverse covariance selection," *IEEE Transactions on Signal Processing*, vol. 62, no. 7, pp. 1644–1654, 2014.

[12] Q. Wei and Z. Zhao, "Large covariance matrix estimation with oracle statistical rate via majorization-minimization," *IEEE Transactions on Signal Processing*, vol. 71, pp. 3328–3342, 2023.

[13] L. E. Frank and J. H. Friedman, "A statistical view of some chemometrics regression tools," *Technometrics*, vol. 35, no. 2, pp. 109–135, 1993.

[14] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[15] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 2010.

[16] Z. Lu, "Iterative reweighted minimization methods for ℓ_p regularized unconstrained nonlinear programming," *Mathematical Programming*, vol. 147, no. 1, pp. 277–307, 2014.

[17] L. Bai, Y. Hu, H. Wang, and X. Yang, "Avoiding strict saddle points of nonconvex regularized problems," *arXiv preprint arXiv:2401.09274*, 2024.

[18] D. Davis and D. Drusvyatskiy, "Proximal methods avoid active strict saddles of weakly convex functions," *Foundations of Computational Mathematics*, vol. 22, no. 2, pp. 561–606, 2022.

[19] D. Davis, D. Drusvyatskiy, and L. Jiang, "Active manifolds, stratifications, and convergence to local minima in nonsmooth optimization," *Foundations of Computational Mathematics*, pp. 1–83, 2025.

[20] A. Flinth, F. de Gournay, and P. Weiss, "Grid is good. adaptive refinement algorithms for off-the-grid total variation minimization," *Open Journal of Mathematical Optimization*, vol. 6, pp. 1–27, 2025.

[21] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, vol. 317. Heidelberg, Germany: Springer Berlin, 2009.

[22] A. S. Lewis, "Active sets, nonsmoothness, and sensitivity," *SIAM Journal on Optimization*, vol. 13, no. 3, pp. 702–725, 2002.

[23] W. L. Hare and A. S. Lewis, "Identifying active constraints via partial smoothness and prox-regularity," *Journal of Convex Analysis*, vol. 11, no. 2, pp. 251–266, 2004.

[24] G. Garrigos, *Descent dynamical systems and algorithms for tame optimization and multi-objective problems.* PhD thesis, Université de Montpellier; Universidad Tecnica Federico Santa Maria, 2015.

[25] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, vol. 146, no. 1, pp. 459–494, 2014.

[26] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, "Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-łojasiewicz inequality," *Mathematics of Operations Research*, vol. 35, no. 2, pp. 438–457, 2010.

[27] X. Chen, L. Niu, and Y. Yuan, "Optimality conditions and a smoothing trust region newton method for nonlipschitz optimization," *SIAM Journal on Optimization*, vol. 23, no. 3, pp. 1528–1552, 2013.

[28] T. Liu and A. Takeda, "An inexact successive quadratic approximation method for a class of difference-of-convex optimization problems," *Computational Optimization and Applications*, vol. 82, no. 1, pp. 141–173, 2022.

[29] O. Banerjee, L. El Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data," *The Journal of Machine Learning Research*, vol. 9, pp. 485–516, 2008.

[30] C.-P. Lee, "Accelerating inexact successive quadratic approximation for regularized optimization through manifold identification," *Mathematical Programming*, vol. 201, no. 1, pp. 599–633, 2023.

[31] H. H. Bauschke, M. N. Dao, and W. M. Moursi, "On Fejér monotone sequences and nonexpansive mappings," *arXiv preprint arXiv:1507.05585*, 2015.

[32] F. H. Clarke, *Optimization and nonsmooth analysis*. Philadelphia, PA: SIAM, 1990.

[33] Y. Wu, S. Pan, and X. Yang, "A regularized newton method for ℓ_q-norm composite optimization problems," *SIAM Journal on Optimization*, vol. 33, no. 3, pp. 1676–1706, 2023.

[34] S. Wu, "Stochastic optimization methods for structure learning in Gaussian graphical models and the general log-determinant optimization," *Doctoral Dissertation*, 2020.

[35] A. Eftekhari, L. Gaedke-Merzhäuser, D. Pasadakis, M. Bollhöfer, S. Scheidegger, and O. Schenk, "Algorithm 1042: Sparse precision matrix estimation with squic," *ACM Transactions on Mathematical Software*, vol. 50, no. 2, pp. 1–18, 2024.

[36] M. Bollhöfer, A. Eftekhari, S. Scheidegger, and O. Schenk, "Large-scale sparse inverse covariance matrix estimation," *SIAM Journal on Scientific Computing*, vol. 41, no. 1, pp. A380–A401, 2019.

[37] D. Pasadakis, M. Bollhöfer, and O. Schenk, "Sparse quadratic approximation for graph learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 11256–11269, 2023.

[38] L. Li and K.-C. Toh, "An inexact interior point method for ℓ_1-regularized sparse covariance selection," *Mathematical Programming Computation*, vol. 2, pp. 291–315, 2010.

[39] M. Yuan and Y. Lin, "Model selection and estimation in the Gaussian graphical model," *Biometrika*, vol. 94, no. 1, pp. 19–35, 2007.

[40] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.

[41] H. Dalianis and H. Dalianis, "Evaluation metrics and evaluation," *Clinical Text Mining: Secondary Use of Electronic Patient Records*, pp. 45–53, 2018.

[42] M. Fazel, H. Hindi, and S. P. Boyd, "Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices," in *Proceedings of the 2003 IEEE American Control Conference, 2003*, vol. 3, pp. 2156–2162, 2003.

[43] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Transactions on Image Processing*, vol. 4, no. 7, pp. 932–946, 1995.

[44] Y. Wang and L. Zhu, "Variable selection and parameter estimation with the atan regularization method," *Journal of Probability and Statistics*, vol. 2016, no. 1, p. 6495417, 2016.

[45] C. Gao, N. Wang, Q. Yu, and Z. Zhang, "A feasible nonconvex relaxation approach to feature selection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 25, pp. 356–361, 2011.