
Partial Correlation Network Estimation by Semismooth Newton Methods

Dongwon Kim*

Department of Statistics
Seoul National University
dongwonida@snu.ac.kr

Sungdong Lee*

Department of Medicine
National University of Singapore
sdlee23@nus.edu.sg

Joong-Ho Won

Department of Statistics
Seoul National University
won.j@snu.ac.kr

Abstract

We develop a scalable second-order algorithm for a recently proposed ℓ_1 -regularized pseudolikelihood-based partial correlation network estimation framework. While the latter method admits statistical guarantees and is inherently scalable compared to likelihood-based methods such as graphical lasso, the currently available implementations rely only on first-order information and require thousands of iterations to obtain reliable estimates even on high-performance supercomputers. In this paper, we further investigate the inherent scalability of the framework and propose locally and globally convergent semismooth Newton methods. Despite the nonsmoothness of the problem, these second-order algorithms converge at a locally quadratic rate, and require only a few tens of iterations in practice. Each iteration reduces to solving linear systems of small dimensions or linear complementary problems of smaller dimensions, making the computation also suitable for less powerful computing environments. Experiments on both simulated and real-world genomic datasets demonstrate the superior convergence behavior and computational efficiency of the proposed algorithm, which position our method as a promising tool for massive-scale network analysis sought for in, e.g., modern multi-omics research.

1 Introduction

A partial correlation network refers to a graph in which each node (vertex) corresponds to a variable and an edge between two nodes represents the partial correlation between the corresponding pair of variables. Partial correlation is defined as the correlation coefficient between the residuals resulting from the linear regression of each of the two variables with the rest of the variables, thereby measuring the degree of association between the two adjusting for the effect of the remaining variables. Learning partial correlation networks from data is particularly important in multi-omics studies of biological systems, where spurious edges due to shared regulatory factors should be avoided to detect direct biological interaction of causal implications [27, 24, 22].

Partial correlation network estimation is closely connected to covariance selection [6], which estimates nonzero entries of the inverse covariance or precision matrix of the variables. This is because a nonzero entry of the precision matrix implies a nonzero partial correlation between the corresponding variables. Moreover, if normality of the data distribution is assumed, then the zero entries of the precision matrix imply conditional independence. For this reason, when the number of variables is large and it is reasonable to assume that the number of edges is small, the graphical lasso [8, 34] that maximizes the Gaussian likelihood under ℓ_1 penalty on the entries of the precision matrix or solve the following optimization problem

$$\min_{\Theta \in \mathbb{S}^p} -\log \det \Theta + \text{tr}(S\Theta) + \lambda \|\Theta_{-D}\|_1, \lambda > 0, \quad (1)$$

*Equal contribution

is by far the most popular choice. Here, $\mathbb{S}^p$ denotes the space of p -dimensional symmetric matrices, and $\mathbf{S} = (1/n)\mathbf{X}^T\mathbf{X}$ is the sample covariance matrix of the data $\mathbf{X} = [x_1, \dots, x_n]^T \in \mathbb{R}^{n \times p}$. The Θ_{-D} denotes a matrix that takes the off-diagonal elements of Θ with zero diagonal elements. The ℓ_1 norm $\|\cdot\|_1$ sums up the absolute values of the elements of its input matrix. The solution $\hat{\Theta}$ to (1) estimates the target precision matrix Θ^* , such that $x_i \stackrel{iid}{\sim} N(0, (\Theta^*)^{-1})$.

However, many methods that directly solve (1) suffer from computational bottleneck when p reaches a few thousands, which is not desirable for analyzing modern massive-scaled data such as arising from multi-omics. The root cause is that the Karush-Kuhn-Tucker (KKT) optimality condition of (1)

$$-\Theta^{-1} + \mathbf{S} + \lambda \mathbf{Z} = \mathbf{0}, \quad \mathbf{Z} \in \partial\|\Theta_{-D}\|_1, \quad (2)$$

where $\partial\|\Theta_{-D}\|_1$ denotes the subdifferential of the ℓ_1 norm at Θ_{-D} , involves matrix inversion. Direct computation of the inverse Θ^{-1} requires $O(p^3)$ arithmetic operations. Since most existing algorithms for solving (1) (e.g., [5, 8, 20, 13]) can be understood as solving (2) in an iterative fashion, such a costly computation essentially occurs every iteration. Worse, matrix inversion is difficult to parallelize or distribute, limiting scalability.

A recently proposed framework called ACCORD [19] overcomes this limitation by employing a computation-friendly loss function in a one-to-one transformation of the precision matrix variable Θ . It solves a convex optimization problem

$$\min_{\Omega \in \mathbb{R}^{p \times p}} -\log \det \Omega_D + (1/2) \text{tr}(\Omega^T \Omega \mathbf{S}) + \lambda \|\Omega_{-D}\|_1, \quad (3)$$

where Ω_D denotes the diagonal matrix that takes the diagonal elements of Ω . Precision matrix Θ and the optimization variable Ω are related by $\Omega = \Theta_D^{-1/2} \Theta$ and $\Theta = \Omega_D \Omega$. Note Ω may not be symmetric. However, the sparsity patterns of Ω and Θ coincide with each other, so the ℓ_1 penalty on Ω_{-D} is legitimate. Under some regulatory conditions, the estimate $\hat{\Omega}$ that uniquely minimizes (3) is a consistent estimator of $\Omega^* := \Theta_D^{*-1/2} \Theta^*$ and its sparsity pattern. The ACCORD framework lies in line with ℓ_1 -penalized pseudolikelihood-based methods for partial correlation network estimation. These methods are computationally more scalable than the graphical lasso owing to the use of pseudolikelihood, which in turn is more robust to non-normality. SPACE [29] is the first of its kind, which directly targets partial correlations. While SPLICE [32] and SYMLASSO [9] follow this line, these methods are not convex, and convergence of the corresponding optimization algorithms are not established. CONCORD [17] convexifies SPACE by changing the target variable to precision matrix Θ . Nevertheless, consistency of the estimated precision matrix requires quite strong conditions. ACCORD relaxes those by asymmetrizing the target variable and establishes non-asymptotic bounds on estimation error and sample complexity [19].

The key computational attraction of ACCORD is its scalability. In (3), the log-determinant barrier is applied only to the diagonal elements of Ω , hence reduces to merely a sum of logarithms. This makes the KKT conditions for (3) much more tractable than (2). In [19], a proximal gradient algorithm is proposed to solve (3) at a linear rate. The major computational bottleneck of this algorithm is sparse-dense matrix-matrix multiplication, which easily allows for distributed computation [18]. This contrasts ACCORD with the graphical lasso. For massive-scale data from, e.g., multi-omic studies, [19] implements the algorithm on distributed-memory high-performance computing (HPC) systems equipped with the message passing interface (MPI). As a showcase, the liver hepatocellular carcinoma (LIHC) cohort from The Cancer Genome Atlas [1] with $p \approx 300,000$ molecular features, for which the most scalable graphical lasso algorithm [14] fails, are analyzed. Despite the success, this computation takes 200 node-hours on a powerful HPC system in a supercomputing center, accessibility to which is generally limited to public. For wider uses of ACCORD among a large number of scientists, a computational algorithm that can operate in a smaller but more accessible environment, such as a GPU workstation, and possesses a convergence rate faster than the linear rate of the proximal gradient algorithm (to compensate for the limited computing power) is on demand.

This paper addresses such a demand by proposing a semismooth Newton method for solving (3). Because of its superlinear convergence property, semismooth Newton methods [31] have been widely employed for learning problems with non-smooth objective functions [21, 23, 33, 35]. We develop the semismooth Newton method for ACCORD through two steps. First, we show that problem (3) is rowwise separable, which allows it to decompose into separate p -dimensional minimization problems for each row of Ω . Second, we show that the Karush-Kuhn-Tucker (KKT) condition for

each minimization problem can be formulated as finding a root of a nonlinear multivariate function F , which is nondifferentiable but strongly semismooth. This observation ensures that the iteration converges at a locally quadratic rate. The key to implementation is to find a suitable element of Clarke's generalized Jacobian (which is a set) of F . We achieve this goal by deriving the Bouligand derivative of F . The latter allows us to *globalize* the semismooth Newton algorithm, for which the initial point can be chosen with liberty. A connection to the primal-dual active set (PDAS) strategy [12] is also made. We then examine the performance of our algorithm in both synthetic and the LHC data, with concluding remarks provided thereafter. Proofs of the results are provided in the Appendix.

2 Locally convergent semismooth Newton method

2.1 KKT condition as root-finding

In this subsection, we show that the KKT condition for (3) can be formulated as finding a root of a multivariate nonlinear function. We begin with observing that the objective function in (3) is separable in rows. Let Ω_i be the i -th row of Ω . Then, (3) can be rewritten as finding a minimizer of

$$\sum_{i=1}^p \left(-\log \det \Omega_{i,i} + (1/2) \text{tr}(\Omega_i^T \Omega_i \mathbf{S}) + \lambda \|\Omega_{i,-i}\|_1 \right), \quad (4)$$

where $\Omega_{i,i}$ and $\Omega_{i,-i}$ denote the diagonal and the off-diagonal parts of Ω_i , respectively. Equation (4) clearly shows that the objective function is separable in Ω_i 's. Henceforth, we will derive the optimality condition for $i = 1$ without loss of generality. Let $\omega = (\omega_1, \omega_{-1}^\top)^\top$, where $\omega_1 = \Omega_{1,1}$ and $\omega_{-1} = \Omega_{1,-1}$. Then, the objective function to minimize is $f(\omega) = g(\omega) + h(\omega)$, where

$$g(\omega) = (1/2)\omega^T \mathbf{S} \omega, \quad h(\omega) = -\log \omega_1 + \lambda \|\omega_{-1}\|_1,$$

which are both convex functions. By introducing a dual variable $d = (d_1, d_{-1}^\top)^\top$, the optimality condition $\partial f(\omega^*) = 0$ can be expressed as

$$\mathbf{S}\omega^* + d^* = 0, \quad d^* \in \partial h(\omega^*) \quad (5)$$

where ∂f and ∂h denote the subdifferentials of f and h , respectively. Moreover, the condition $d^* \in \partial h(\omega^*)$ can be equivalently expressed using the proximal operator of h [15]

$$\omega^* = \text{prox}_h(\omega^* + d^*), \quad (6)$$

which can be explicitly written as

$$\text{prox}_h(\omega) := \arg \min_{v \in \mathbb{R}^p} \{h(v) + (1/2)\|v - \omega\|_2^2\} = \left(\frac{\omega_1 + \sqrt{\omega_1^2 + 4}}{2}, T_\lambda(\omega_{-1})^\top \right)^\top,$$

where $T_\lambda(\omega_{-1}) := (\text{sign}(\omega_j) \max(|\omega_j| - \lambda, 0))_{j \neq 1}$ is an element-wise soft-thresholding operator. Combining (5) and (6), the KKT condition for (3) can be equivalently written as

$$F(\omega^*, d^*) = 0, \quad (7)$$

where $F : \mathbb{R}^{2p} \rightarrow \mathbb{R}^{2p}$ is defined as

$$F(\omega, d) = \begin{pmatrix} \mathbf{S}\omega + d \\ \omega_1 - \text{prox}_{-\log(\cdot)}(\omega_1 + d_1) \\ \omega_{-1} - T_\lambda(\omega_{-1} + d_{-1}) \end{pmatrix}, \quad \text{prox}_{-\log(\cdot)}(x) = \frac{x + \sqrt{x^2 + 4}}{2}. \quad (8)$$

Therefore, solving (3) amounts to finding a root of nonlinear equation (7). Map F has a unique root almost surely, as long as the sample is drawn from a continuous distribution [19, Theorem 3.1].

2.2 Newton's method for semismooth functions

The formulation (7)–(8) of the KKT condition naturally suggests Newton's method for finding the root. However, because of the presence of the soft thresholding operator in (8), it is not differentiable everywhere, which means that the plain Newton iteration cannot be applied. In this section, we show that, despite the nondifferentiability, function F in (8) is strongly semismooth, hence a semismooth version of Newton's method can be employed. We begin by defining semismoothness [26, 31].

Definition 2.1 (Semismoothness). Function $F : \mathbb{R}^m \rightarrow \mathbb{R}^l$ is said to be semismooth at $x \in \mathbb{R}^m$ if it is locally Lipschitz, all directional derivatives $F'(x; v) := \lim_{t \rightarrow 0} \frac{F(x+tv) - F(x)}{t}$ exist, and for any $G \in \partial F(x + v)$, the following holds:

$$Gv - F'(x; v) = o(\|v\|).$$

Furthermore, F is strongly semismooth at x if it is semismooth at x and satisfies the condition for any $G \in \partial F(x + v)$,

$$Gv - F'(x; v) = O(\|v\|^2).$$

The F is said to be (strongly) semismooth on $U \in \mathbb{R}^m$ if it is (strongly) semismooth at any $x \in U$.

Theorem 2.2. *The map $(\omega, d) \mapsto F(\omega, d)$ as defined in (8) is strongly semismooth.*

The semismooth Newton method [31] relies on the notion of the generalized Jacobian [3].

Definition 2.3 (Generalized Jacobian). For a function $F : \mathbb{R}^m \rightarrow \mathbb{R}^l$ that is locally Lipschitz continuous around $x \in \mathbb{R}^m$, Clarke's generalized Jacobian at x is defined as

$$\partial F(x) = \text{conv} \left\{ \lim_k \nabla F(x^k) : x^k \rightarrow x, x^k \in D_F(x) \right\},$$

where $D_F(x) = \{y : F \text{ is differentiable at } y\}$, conv denotes the convex hull operation, and $\nabla F(y)$ represents the Jacobian of F at y . We call an element of $\partial F(x)$ a generalized Jacobian of F at x .

Thus for locally Lipschitz functions, a version of Newton's method for solving $F(x) = 0$ can be defined as

$$x^{k+1} = x^k - G(x^k)^{-1} F(x^k), \quad G(x^k) \in \partial F(x^k), \quad (9)$$

provided that $G(x^k)$ is nonsingular. If F semismoothness, then the Newton iteration (9) exhibits a superlinear convergence for suitable initialization [31]. If it is strongly semismooth, then it converges at a locally quadratic rate:

Theorem 2.4 ([31]). *Suppose that x^* is a root of $F : \mathbb{R}^m \rightarrow \mathbb{R}^l$, and that F is strongly semismooth in an open neighborhood U containing x^* , with a generalized Jacobian $G(x) \in \partial F(x)$. If $G(x)$ is nonsingular for all $x \in U$ and the set $\{\|G(x)^{-1}\| : x \in U\}$ is bounded, then the semismooth Newton method (9) converges quadratically to x^* , provided that $\|x^0 - x^*\|$ is sufficiently small.*

To verify the assumptions of Theorem 2.4 for the target problem (7), we provide a sufficient condition.

Proposition 2.5. *Suppose (ω^*, d^*) satisfies equation (7) for F defined in (8). Let $\bar{\mathcal{A}}^* = \{1\} \cup \{i \in \{2, \dots, p\} : |\omega_i^* + d_i^*| > \lambda\}$. If $\mathbf{X}_{\bar{\mathcal{A}}^*}$, the submatrix of $\mathbf{X}$ taking the columns indexed by $\bar{\mathcal{A}}^*$ has full column rank, then there exists a nonsingular $G(\omega, d) \in \partial F(\omega, d)$ for any $(\omega, d) \in \mathbb{R}^{2p}$ in a neighborhood of (ω^*, d^*) .*

In particular, we may choose (see Appendix D for the derivation)

$$G(\omega, d) = \begin{pmatrix} \mathbf{S} & \mathbf{I} \\ \mathbf{I} - \mathbf{J} & -\mathbf{J} \end{pmatrix}, \quad (10)$$

where $\mathbf{J} = \text{diag}(J_{ii})$ is a diagonal matrix with $J_{11} = \frac{1}{2} + \frac{\omega_1 + d_1}{2\sqrt{(\omega_1 + d_1)^2 + 4}}$ and $J_{ii} = 1$ if $|\omega_i + d_i| > \lambda$, $J_{ii} = 0$ if $|\omega_i + d_i| \leq \lambda$ for $i = 2, \dots, p$. In this case, semismooth Newton iteration (9) is equivalent to the primal-dual active set algorithm [11, 7, 2, 15].

3 Globalizing semismooth Newton method

3.1 B-semismooth Newton method

In this section, we propose a globalization strategy for the semismooth Newton method (9). Although the quadratic convergence result (Theorem 2.4) is attractive, this result is only local and is silent on how to choose the initial point $x^0 = (\omega^0, d^0)$. An algorithm that converges from any initial point while maintaining locally quadratic convergence near the solution is more practical and desirable. To this end, it turns out that globalizing iteration (9) *per se* is difficult, while globalizing the semismooth Newton based on Bouligand derivatives or B-derivatives [28] is manageable.

Definition 3.1 (B-differentiability). A function $F : \mathbb{R}^m \rightarrow \mathbb{R}^l$ is said Bouligand differentiable (B-differentiable) at $x \in \mathbb{R}^m$ if it is locally Lipschitz, the directional derivative $F'(x; v) := \lim_{t \downarrow 0} \frac{F(x+tv) - F(x)}{t}$ exists for all v and fulfill the approximation property $\|F(x+v) - F(x) - F'(x; v)\| = o(\|v\|)$ as $v \rightarrow 0$. If F is B-differentiable for any $x \in \mathbb{R}^m$, then it is said B-differentiable. The directional derivative that satisfies these conditions is called the Bouligand derivative (B-derivative) of F at x in the direction v .

Using the directional derivative $F'(x; v)$, a Newton method for finding the root of a B-differentiable function F can be defined as

$$x^{k+1} = x^k + v^k, \quad F'(x^k; v^k) = -F(x^k). \quad (11)$$

Our target function is B-differentiable:

Proposition 3.2. *Function F as defined in (8) is B-differentiable. Moreover, its B-derivative at (ω, d) in the direction of $v = (v_\omega, v_d)$ is given by*

$$F'(\omega, d; v) = \begin{pmatrix} Sv_\omega + v_d \\ (z + 1/2)v_{\omega,1} + (z - 1/2)v_{d,1} \\ \begin{cases} -v_{d,i} & i \in \mathcal{A} \\ v_{\omega,i} & i \in \mathcal{I}^0 \\ \min\{v_{\omega,i}, -v_{d,i}\} & i \in \mathcal{I}^+ \\ \max\{v_{\omega,i}, -v_{d,i}\} & i \in \mathcal{I}^- \end{cases} \end{pmatrix}, \quad (12)$$

where $z = z(\omega_1, d_1) := -\frac{\omega_1 + d_1}{2\sqrt{(\omega_1 + d_1)^2 + 4}}$ and $\mathcal{A} = \mathcal{A}(\omega, d)$, $\mathcal{I}^0 = \mathcal{I}^0(\omega, d)$, $\mathcal{I}^+ = \mathcal{I}^+(\omega, d)$, and $\mathcal{I}^- = \mathcal{I}^-(\omega, d)$ such that

$$\begin{aligned} \mathcal{A} &= \{i \in \{2, \dots, p\} : |\omega_i + d_i| > \lambda\}, \quad \mathcal{I}^0 = \{i \in \{2, \dots, p\} : |\omega_i + d_i| < \lambda\}, \\ \mathcal{I}^+ &= \{i \in \{2, \dots, p\} : \omega_i + d_i = \lambda\}, \quad \mathcal{I}^- = \{i \in \{2, \dots, p\} : \omega_i + d_i = -\lambda\}. \end{aligned} \quad (13)$$

3.2 Solvability of the B-semismooth Newton method via linear complementarity problem

Applying (12) to (11), finding the Newton direction v such that $F'(\omega, d; v) = -F(\omega, d)$ reduces to

$$Sv_\omega + v_d = -S\omega - d \quad (14)$$

$$(z + 1/2)v_{\omega,1} + (z - 1/2)v_{d,1} = -\omega_1 + \text{prox}_{-\log(\cdot)}(\omega_1 + d_1) \quad (15)$$

$$\min\{v_{\omega,i}, -v_{d,i}\} = -\omega_i, \quad i \in \mathcal{I}^+, \quad \max\{v_{\omega,i}, -v_{d,i}\} = -\omega_i, \quad i \in \mathcal{I}^-, \quad (16)$$

where z is the same as in (12). Determining the existence of a solution to (14)–(16) and finding one is a quadratic programming problem in $v = (v_\omega, v_d) \in \mathbb{R}^{2p}$. However, under the condition similar to that of Proposition 3.2, we only need to solve a much smaller linear complementarity problem (LCP) [4]. Let $\bar{\mathcal{A}} = \bar{\mathcal{A}}(\omega, d) := \{1\} \cup \mathcal{A}$ and $\bar{\mathcal{I}} = \bar{\mathcal{I}}(\omega, d) := \bar{\mathcal{A}} \cup \mathcal{I}^+ \cup \mathcal{I}^-$ from (13) and $\mathbf{X}_{\bar{\mathcal{A}}}$, $\mathbf{X}_{\bar{\mathcal{I}}}$ be the submatrix of $\mathbf{X}$ obtained in a similar way to $\mathbf{X}_{\bar{\mathcal{A}}^*}$ in Proposition 2.5.

Lemma 3.3 (cf. [10]). *Assume $\mathbf{X}_{\bar{\mathcal{I}}}$ has full column rank. Given $(\omega, d) \in \mathbb{R}^{2p}$, $v = (v_\omega, v_d) \in \mathbb{R}^{2p}$ solves (14)–(16) if and only if*

$$\mathbf{x} = ((-v_{d,\mathcal{I}^+} + \omega_{\mathcal{I}^+})^\top, (v_{d,\mathcal{I}^-} - \omega_{\mathcal{I}^-})^\top)^\top, \quad \mathbf{y} = ((v_{\omega,\mathcal{I}^+} + \omega_{\mathcal{I}^+})^\top, (-v_{\omega,\mathcal{I}^-} - \omega_{\mathcal{I}^-})^\top)^\top$$

solves the LCP

$$\mathbf{x} \geq \mathbf{0}, \quad \mathbf{y} \geq \mathbf{0}, \quad \langle \mathbf{x}, \mathbf{y} \rangle = 0, \quad \text{and } \mathbf{y} = M\mathbf{x} + \mathbf{q} \quad (17)$$

where $M = M(\mathbf{S}, \omega, d)$ and $\mathbf{q} = \mathbf{q}(\mathbf{S}, \omega, d)$ are given in Appendix E.2. The M is symmetric positive definite and (17) admits a unique solution.

A LCP can be solved efficiently using standard techniques such as Lemke's or pivoting methods [4]. Let $S_{\mathcal{R}, \mathcal{C}}$ be the submatrix of $\mathbf{S}$ with rows indexed by $\mathcal{R}$ and columns indexed by $\mathcal{C}$. With the solution to (17) available, the Newton direction v^k in (11) for the F in (8) can be found in a closed form as stated in the following main theorem.

Theorem 3.4. In (13), let $\mathcal{A} = \mathcal{A}(\omega^k, d^k)$, $\bar{\mathcal{A}} = \bar{\mathcal{A}}(\omega^k, d^k)$, and $\mathcal{I}^\pm = \mathcal{I}^\pm(\omega^k, d^k)$. Assume $\mathbf{X}_{\bar{\mathcal{A}}}$ has full column rank. Let $\mathbf{x}, \mathbf{y}$ be the unique solution to (17) for an iterate (ω^k, d^k) . In addition to (13), let $\mathcal{A}^+ = \{i \in \{2, \dots, p\} : \omega_i^k + d_i^k > \lambda\}$ and $\mathcal{A}^- = \{i \in \{2, \dots, p\} : \omega_i^k + d_i^k < -\lambda\}$. Then, the B-semismooth Newton iteration (11) is well-defined with

$$\begin{aligned} v_{\omega, \mathcal{I}^0}^k &= -\omega_{\mathcal{I}^0}^k, \quad v_{d, \mathcal{A}}^k = \lambda(\mathbf{1}_{\mathcal{A}^+}^\top, -\mathbf{1}_{\mathcal{A}^-}^\top)^\top - d_{\mathcal{A}}^k, \quad \begin{pmatrix} v_{d, \mathcal{I}^+}^k \\ v_{d, \mathcal{I}^-}^k \end{pmatrix} = \begin{pmatrix} -\mathbf{x}_{\mathcal{I}^+} \\ \mathbf{x}_{\mathcal{I}^-} \end{pmatrix} + \begin{pmatrix} \omega_{\mathcal{I}^+}^k \\ \omega_{\mathcal{I}^-}^k \end{pmatrix}, \\ ((v_{\omega, 1}^k)^\top, (v_{\omega, \mathcal{A}}^k)^\top, (v_{\omega, \mathcal{I}^+}^k)^\top, -(v_{\omega, \mathcal{I}^-}^k)^\top, (v_{d, \mathcal{I}^0}^k)^\top)^\top &= N^{-1}(P^k - (0^\top, 0^\top, (v_{d, \mathcal{I}^+}^k)^\top, -(v_{d, \mathcal{I}^-}^k)^\top, 0^\top)^\top), \\ v_{d, 1} &= b^k v_{\omega, 1}^k - c^k, \quad N^{-1} = \begin{pmatrix} N_1^{-1} & 0 \\ -CN_1^{-1} & I \end{pmatrix} \end{aligned}$$

where $b^k, c^k \in \mathbb{R}$, $P^k \in \mathbb{R}^p$, $C \in \mathbb{R}^{(p-|\bar{\mathcal{A}}|) \times |\bar{\mathcal{A}}|}$, and $N_1 \in \mathbb{R}^{|\bar{\mathcal{A}}| \times |\bar{\mathcal{A}}|}$, given in Appendix E.2, are determined by ω^k, d^k and $\mathbf{S}$. In particular, N_1 is symmetric positive definite.

In particular, given that $\mathcal{I}^{+*} \cup \mathcal{I}^{-*} = \emptyset$ for $\mathcal{I}^{+*} = \{i \in \{2, \dots, p\} : \omega_i^* + d_i^* = \lambda\}$, $\mathcal{I}^{-*} = \{i \in \{2, \dots, p\} : \omega_i^* + d_i^* = -\lambda\}$, the convergence of (11) becomes locally quadratic:

Corollary 3.5 ([30, Corollary 3.3]). Let (ω^*, d^*) be the unique solution to $F(\omega, d) = 0$. If $\mathcal{I}^{+*} \cup \mathcal{I}^{-*} = \emptyset$ and $\mathbf{X}_{\bar{\mathcal{A}}^*}$ from Proposition 2.5 has full column rank, then the B-semismooth Newton iteration (11) is well-defined and converges quadratically to (ω^*, d^*) in its neighborhood.

Remark 3.6. Multiplication with matrix N^{-1} only requires a linear system solve involving the symmetric positive definite submatrix N_1 of size $|\bar{\mathcal{A}}| \times |\bar{\mathcal{A}}|$. In the sparse estimation context, the size of the active set $\bar{\mathcal{A}}$ is small, hence solving (11) is not a bottleneck for scalability.

Remark 3.7. Given that F in (8) is locally Lipschitz and directionally differentiable, there exists a generalized Jacobian G such that $F'(\omega, d; v) = G(\omega, d)v$ where v satisfies (11) [31]. More specifically, the B-semismooth Newton method can be regarded as the semismooth Newton method (9) that employs the generalized Jacobian G in (10) such that $J_{ii} = 1$ for $i \in \bar{\mathcal{A}} \cup \{i \in \mathcal{I}^+ \cup \mathcal{I}^- : \mathbf{x}_i = 0\}$ and $J_{ii} = 0$ for $i \in \mathcal{I}^0 \cup \{i \in \mathcal{I}^+ \cup \mathcal{I}^- : \mathbf{x}_i > 0\}$, where $\mathbf{x}$ is the solution to the LCP (17). Furthermore, since it is numerically unlikely that $|\omega_i + d_i|$ precisely equals to λ , in most scenarios $\mathcal{I}^+ \cup \mathcal{I}^- = \emptyset$ and the B-semismooth Newton method coincides with the PDAS algorithm.

3.3 Globalizing B-semismooth Newton with line search

Globalization of the B-semismooth Newton method (11) is based on the observation that, under the (unique) existence of the solution to (7), solving this nonlinear equation is equivalent to minimizing $\theta(\omega, d) = \frac{1}{2} \|F(\omega, d)\|_2^2$. Given the Newton direction v^k obtained by solving (14)–(16) for $(\omega, d) = (\omega^k, d^k)$, we can apply line search to construct a damped Newton algorithm that monotonically decreases the objective sequence $\{\theta(\omega^k, d^k)\}$. Following [25], we consider the line search algorithm that finds an integer m_k such that

$$\theta(x^k + \rho^{m_k} v^k) - \theta(x^k) \leq -2\sigma \rho^{m_k} \theta(x^k), \quad (18)$$

for algorithm parameters $\sigma, \rho \in (0, 1)$.

Also, Theorem 3.4 suggests that the full column rank of the submatrix $\mathbf{X}_{\bar{\mathcal{A}}_k}$ is essential in computing the direction v^k for $\bar{\mathcal{A}}_k = \bar{\mathcal{A}}(\omega^k, d^k)$. If this condition is not met, then we can resort to the proximal gradient algorithm of [19] which converges linearly to the solution. Note that this algorithm does not necessarily descent $\theta(\omega, d)$ in one iteration. Thus, we need repeat the step for $\theta(\omega, d)$ to decrease.

The resulting damped B-semismooth Newton algorithm is summarized in Algorithm 1. The main question is whether the iterate sequence $\{(\omega^k, d^k)\}$ will converge to the solution (ω^*, d^*) , and whether the Newton steps will eventually take place; these are all addressed in the following theorem.

Theorem 3.8. Suppose $\mathcal{I}^{+*} \cup \mathcal{I}^{-*} = \emptyset$ and $\mathbf{X}_{\bar{\mathcal{A}}^*}$ from Proposition 2.5 has full column rank. Let $\{(\omega^k, d^k)\}$ be the sequence of iterates generated by Algorithm 1. Then, the proximal gradient step occurs only a finite number of times, and if we let $\{\rho^{m_k}\}$ denote the corresponding step sizes,

1. if $\limsup_{k \rightarrow \infty} \rho^{m_k} > 0$, then $(\omega^k, d^k) \rightarrow (\omega^*, d^*)$, the unique solution to $F(\omega, d) = 0$;

Algorithm 1: Damped B-semismooth Newton method with line search

Input: Initial guess $x^0 = (\omega^0, d^0)$ for all rows, $\rho_0, \sigma \in (0, 1)$, and tolerance tol

```
1 foreach row in  $\Omega$  simultaneously do
2    $k \leftarrow 0$ ;
3   while  $\theta(x^k) > tol$  do
4     Compute the sets  $\mathcal{A}_k = \mathcal{A}(\omega^k, d^k)$ ,  $\mathcal{I}_k^0 = \mathcal{I}^0(\omega^k, d^k)$ , and  $\mathcal{I}_k^\pm = \mathcal{I}^+(\omega^k, d^k) \cup \mathcal{I}^-(\omega^k, d^k)$ ;
5     Set  $\bar{\mathcal{A}}_k = \{i\} \cup \mathcal{A}_k \cup \mathcal{I}_k^\pm$ ;
6     if  $X_{\bar{\mathcal{A}}_k}$  has full column rank then
7       Find the search direction  $v^k$  according to Theorem 3.4;
8        $m \leftarrow 0$ ;
9       repeat
10        |  $m \leftarrow m + 1$ ;
11        until  $\theta(x^k + \rho_k^m v^k) \leq (1 - 2\sigma\rho_k^m)\theta(x^k)$ ;
12         $x^{k+1} \leftarrow x^k + \rho_k^m v^k$ ;
13      else
14         $\tilde{x} \leftarrow x^k$ ;
15        repeat
16        | Update  $\tilde{x}$  with the proximal gradient algorithm of [19]
17        until  $\theta(\tilde{x}) \leq \theta(x^k)$ ;
18         $x^{k+1} \leftarrow \tilde{x}$ ;
19     $k \leftarrow k + 1$ ;
```

2. if $\limsup_{k \rightarrow \infty} \rho^{m_k} = 0$ and $(\omega^\dagger, d^\dagger)$ is an accumulation point of $\{(\omega^k, d^k)\}$, then $(\omega^\dagger, d^\dagger) = (\omega^*, d^*)$ and $(\omega^k, d^k) \rightarrow (\omega^*, d^*)$.

Moreover, if $(\omega^k, d^k) \rightarrow (\omega^*, d^*)$, then the convergence is locally quadratic.

Comparison to the SSNAL framework There is a line of work that falls into the semismooth Newton-based augmented Lagrangian method (SSNAL) framework, developed for solving the lasso-related regression problems [21, 23, 35]. As the name suggests, this framework solves the problem via the inexact augmented Lagrangian method (ALM). ALM incurs an optimization subproblem, and SSNAL solves this subproblem via semismooth Newton. While the subproblem is solved at a locally superlinear (quadratic if it is strongly semismooth) rate, the convergence rate of the the outer ALM iteration is locally linear, even if the subproblem can be solved at a locally quadratic rate. Global convergence is also established.

In contrast, Algorithm 1 formulates the KKT condition as a strongly semismooth equation (8) and apply semismooth Newton. Globalization is achieved via damping instead of ALM. So our algorithmic design is fundamentally different from SSNAL: there is no subproblem to solve inexactly, and the convergence is locally quadratic. (If the KKT equation were merely semismooth, then the local convergence would be superlinear.)

Another distinction is that SSNAL typically constructs augmented Lagrangian via the *dual* of the original (primal) problem, while we focus on the primal. It is worth pointing out that [23] applies the SSNAL framework directly to the primal in addition to the dual, but its design and analysis focus on low-dimensional settings where the number of observations significantly exceeds the number of features (i.e., $n \gg p$). As such, it is not readily applicable to high-dimensional settings, which are the primary focus of our work.

4 Experiments

4.1 Convergence behavior of Damped B-semismooth Newton method

We present numerical experiments to empirically validate the convergence and descent property of the proposed damped B-semismooth Newton method. Specifically, we compare Algorithm 1 against the proximal gradient algorithm proposed in [19], by tracking the following three metrics over the

iteration:

$$\begin{aligned}
& \|F(\omega^k, d^k)\|_2^2, \quad (\text{objective function of Algorithm 1}) \\
& f(\omega^k) - f(\hat{\omega}), \quad (\text{objective function of ACCORD}) \\
& \|\omega^k - \omega^*\|_2^2. \quad (\text{squared distance to the optimum})
\end{aligned} \tag{19}$$

The $\hat{\omega}$ estimates the optimal solution ω^* by the output of Algorithm 1 at convergence. Convergence is declared if all rows satisfies the criterion $\|F(\omega^k, d^k)\|_2^2 < 10^{-10}$. The rank test for line 6 in Algorithm 1 is based on the sufficiency check $|\bar{\mathcal{A}}_k| \leq n$, while more stringent tests (e.g., condition number or singular value thresholds) are also possible. For the line search parameter (line 11), we use $\rho_k = \max\{0.7 - 0.003 \cdot k, 0.4\}$ and $\sigma = 0.001$. The algorithm is implemented in PyTorch 1.13.1. Matrix operations are parallelized on six NVIDIA RTX 6000 Ada generation GPUs using CUDA 11.7.

To assess the convergence behavior, we generated $n = 500$ observations from a $p = 1,000$ -dimensional zero-mean multivariate Gaussian distribution with precision matrix Θ^* that contained 3% non-zero entries, locations of which were sampled uniformly at random. The regularization parameter λ was set to 0.1, 0.15, and 0.2 to ensure that the estimated precision matrix exhibits a sparsity level comparable to Θ^* .

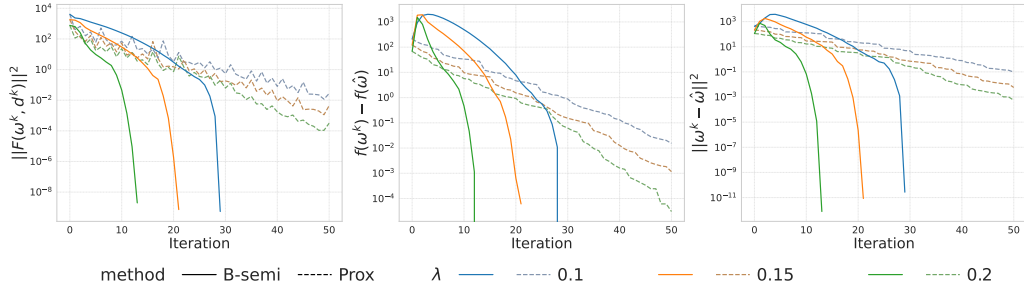


Figure 1: Convergence comparison: damped B-semismooth Newton vs. proximal gradient method

The results are summarized in Figure 1, where the vertical axis is plotted on a logarithmic scale. The proximal gradient method exhibits an approximately linear convergence rate with respect to both the ACCORD objective function f and the squared distance to the optimum, consistent with the theory [19]. In contrast, the damped B-semismooth Newton method demonstrates superlinear convergence for both metrics, achieving approximately an order of magnitude faster convergence. Furthermore, for $\|F(\omega^k, d^k)\|_2^2$ in which Algorithm 1 has a descent property, the convergence appears locally quadratic (as predicted by theory) and occurs within a few tens of iterations.

As Algorithm 1 includes two components in addition to the semismooth Newton steps, namely the proximal gradient steps and the LCP solve step, we also conducted an ablation study to assess the contribution of each component. See Appendix G.1 for details.

4.2 Graph density-convergence trade-off

In Algorithm 1, the B-semismooth Newton update cannot be applied to the row in which $\text{rank}(X_{\bar{\mathcal{A}}_k}) > \min(n, p)$. Since the size of the active set is inherently linked to the sparsity of the true precision matrix Θ^* (hence that of Ω^*), we vary the regularization parameter λ to study the robustness of the algorithm (in terms of the occurrence of the Newton step) to the density of the underlying partial correlation graph.

To conduct this investigation, we define the graph density as $\text{Density} = |E|/\binom{p}{2}$ where $|E|$ denotes the number of undirected edges and p is the number of nodes in the graph implied by Θ^* . Based on this definition, we constructed a graph structure sampled uniformly at random with densities of 0.03, 0.05, and 0.10. Note that the graph density in Section 4.1 is roughly 0.03. The partial correlation matrix corresponding to the graph was constructed to have the minimum eigenvalues of at least 0.2, with all non-zero entries having absolute values no less than 0.1. Based on this matrix, we generate $n = 100$ observations from a $p = 1,000$ -dimensional zero-mean multivariate Gaussian distribution with precision matrix Θ^* .

The hyperparameter λ was chosen so that the estimated solution exhibits a sparsity level comparable to that of the true precision matrix Θ^* , while the line search parameters were set to $\rho = 0.5, \sigma = 0.001$. We then assesses whether convergence is achieved within 300 iterations for each case. This limit of iterations was chosen because from Section 4.1 we see B -semismooth Newton method typically converges within a few tens of iterations. The results are presented in Table 1.

Density \ λ	0.2	0.15	0.1	0.09	0.08	0.07
0.03	Y	Y	—	—	—	—
0.05	—	Y	Y	—	—	—
0.10	—	—	Y	Y	Y	N

Table 1: Convergence within 300 iterations for varying graph density and λ (Yes / No)

Density \ λ	0.07	0.06	0.05
0.10	9.1%	97.7%	100%

Table 2: Proportion of rows updated by proximal method after 300 iterations

With one exception, Algorithm 1 converged within 300 iterations. The exception occurred when the underlying graph was the densest and the regularization parameter was the smallest. Usually the denser the underlying graph becomes, the smaller regularization parameter is required to recover its structure. To further investigate this phenomenon, we summarize in Table 2 the proportion of rows updated by the proximal gradient method after 300 iterations for smaller values of λ .

This result shows that there is a trade-off between the density of the underlying graph to estimate and the convergence speed of Algorithm 1. Nevertheless, for a consistent estimation of the support of the underlying graph, a sparsity assumption is essential, which is tied to the identifiability condition of Proposition 3.2 [7]. Under this condition, the Newton step occurs eventually. Adaptive selection of the regularization parameter λ under the sparsity regime, e.g., continuation strategy [7, 15], is an interesting future direction of study.

Further experiments on other types of random graphs were also conducted, the results of which can be found in Appendix G.3.

4.3 Application to large-scale genomic data

In this subsection, we assess the performance of Algorithm 1 in estimating the partial correlation graph from a large-scale multi-omic dataset. Specifically, we apply Algorithm 1 and proximal gradient method (ACCORD-FBS) to the LIHC dataset from The Cancer Genome Atlas [1], consisting of $p = 305,471$ features including RNA transcription levels and the DNA methylation status of human genomes from $n = 365$ samples. Estimating partial correlations in such high-dimensional biological data can facilitate the identification and validation of epigenomic and transcriptomic regulatory mechanisms in human cells. This dataset was analyzed in [19], in which an HPC version of ACCORD-FBS was run on two national supercomputing centers for several hours. We emphasize that we employed our own implementation of ACCORD-FBS for the GPU workstation of the previous subsection, utilizing the row-wise separability.

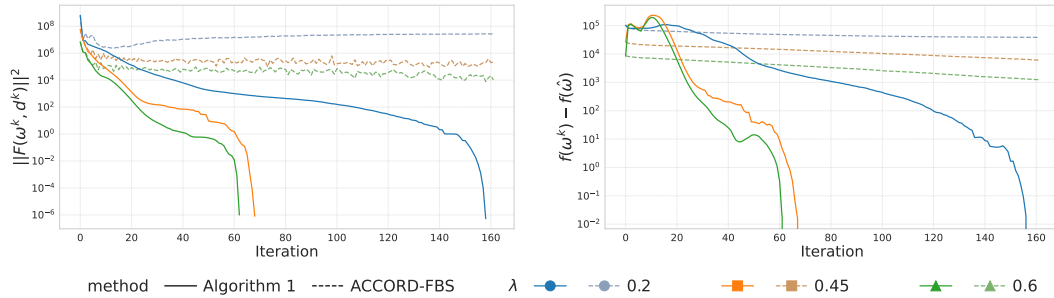


Figure 2: Convergence comparison on a large-scale multi-omic cancer dataset

The regularization parameter was set to $\lambda = 0.45$, selected via the extended pseudo-Bayesian information criterion [19]; a detailed sensitivity analysis of estimation performance around this value of λ is provided in Appendix G.2. To test robustness of Algorithm 1, we further examined the optimization performance for $\lambda \in \{0.20, 0.60\}$, corresponding to relatively dense and sparse estimates. Figure 2

illustrates the convergence behavior of Algorithm 1 and ACCORD-FBS, evaluated using the first two metrics defined in (19). While ACCORD-FBS typically requires over a thousand iterations to converge [19], our algorithm reached the solution within only a few dozen iterations. Despite the reduced number of iterations, the computational cost per iteration remains nearly unchanged, as the dominant operation in each step involves only the inversion of a relatively small submatrix of dimension $|\bar{\bar{A}}_k| \times |\bar{\bar{A}}_k|$. Due to the rapid convergence of the B-semismooth Newton method, each iteration involves significantly fewer row updates compared to ACCORD-FBS.

Method \ λ	0.2		0.45		0.6	
	Time (s)	$\ F\ $	Time (s)	$\ F\ $	Time (s)	$\ F\ $
Algorithm 1	5341	1.260×10^1	3352	0	2083	0
ACCORD-FBS	6778	4.709×10^3	5150	4.035×10^2	3452	2.671×10^2

Table 3: Wall-clock time (in seconds) and $\|F\|$ until convergence or 100 iterations

As shown in Table 3, Algorithm 1 consistently achieved shorter wall-clock time and smaller $\|F\|$ than ACCORD-FBS for all tested values of λ within 100 iterations or at convergence, whichever came first.

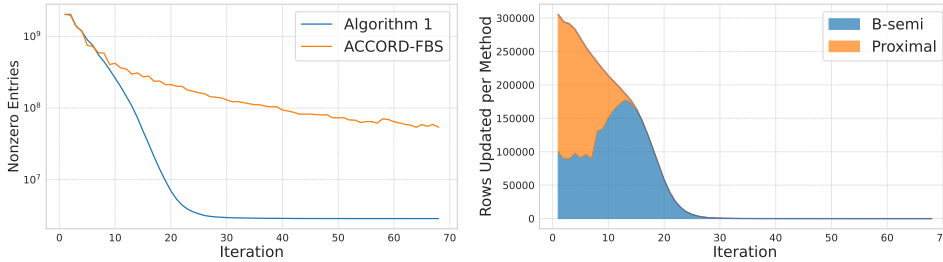


Figure 3: Interaction between sparsity and update strategy in Algorithm 1

As expected, more iterations were required to converge as the regularization parameter λ decreases, and more likely the proximal gradient step occurs (rather than the Newton step) in Algorithm 1. As a means of dissecting this phenomenon, we fix λ at 0.45 and monitor the number of nonzero elements in the iterate sequence Ω^k and the number of rows updated by either of the two methods; see Figure 3. In the early stage, updates were performed primarily by the proximal gradient method. Over time, the number of nonzero entries in Ω^k decreased and the proportion of rows updated by the Newton step increased. This transition aligns with intuition and enables Algorithm 1 to take advantage of the fast convergence behavior of the B-semismooth Newton method in later iterations.

5 Conclusion

We explored the applicability of semismooth Newton methods in partial correlation network estimation to derive a scalable algorithm with superlinear convergence. These methods exhibit locally quadratic convergence to the solution, where the global convergence is also achievable by combining with the proposed damping strategy. They also enjoy parallelism by leveraging the rowwise separability of the ACCORD objective function. Simulations and genomic data applications on GPUs validate the superior computational performance of the semismooth methods compared to the proximal method. In conclusion, semismooth Newton methods suggest a promising outlook for fast and efficient partial correlation analysis of large-scale data with modest computing resources.

6 Acknowledgements

S. Lee was partially supported by the Sejong Science Fellowship through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (MSIT, No. RS-2024-00347617), Singapore Ministry of Education (MOE-000244-00, MOE-000617-00), and National Medical Research Council of Singapore (MOH-000986-00). D. Kim and J.-H. Won were supported

by the AI-Bio Research Grant through Seoul National University (No. 0413-20230050), Artificial Intelligence Graduate School Program (Seoul National University), and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT, No. RS-2024-00337691).

References

- [1] Adrian Ally, Miruna Balasundaram, Rebecca Carlsen, Eric Chuah, Amanda Clarke, Noreen Dhalla, Robert A Holt, Steven JM Jones, Darlene Lee, Yussanne Ma, et al. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell*, 169(7):1327–1341, 2017.
- [2] Dejun Chu, Changshui Zhang, Shiliang Sun, and Qing Tao. Semismooth Newton algorithm for efficient projections onto $\ell_{1,\infty}$ -norm ball. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 1974–1983. PMLR, 2020.
- [3] Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.
- [4] Richard W Cottle, Jong-Shi Pang, and Richard E Stone. *The linear complementarity problem*. SIAM, 2009.
- [5] Alexandre d’Aspremont, Onureena Banerjee, and Laurent El Ghaoui. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30(1):56–66, 2008.
- [6] Arthur P Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.
- [7] Qibin Fan, Yuling Jiao, and Xiliang Lu. A primal dual active set algorithm with continuation for compressed sensing. *IEEE Transactions on Signal Processing*, 62(23):6276–6285, 2014.
- [8] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9(3):432–441, 2008.
- [9] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Applications of the Lasso and grouped Lasso to the estimation of sparse graphical models. Technical report, Department of Statistics, Stanford University, 2010.
- [10] Esther Hans and Thorsten Raasch. Global convergence of damped semismooth Newton methods for ℓ_1 tikhonov regularization. *Inverse Problems*, 31(2):025005, 2015.
- [11] Michael Hintermüller. Semismooth Newton methods and applications. Technical report, Department of Mathematics, Humboldt-University of Berlin, 2010.
- [12] Michael Hintermüller, Kazufumi Ito, and Karl Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM Journal on Optimization*, 13(3):865–888, 2002.
- [13] Cho-Jui Hsieh, Mátyás A Sustik, Inderjit S Dhillon, and Pradeep Ravikumar. QUIC: quadratic approximation for sparse inverse covariance estimation. *Journal of Machine Learning Research*, 15(1):2911–2947, 2014.
- [14] Cho-Jui Hsieh, Mátyás A Sustik, Inderjit S Dhillon, Pradeep K Ravikumar, and Russell Poldrack. BIG & QUIC: Sparse inverse covariance estimation for a million variables. *Advances in Neural Information Processing Systems*, 26:3165–3173, 2013.
- [15] Jian Huang, Yuling Jiao, Bangti Jin, Jin Liu, Xiliang Lu, and Can Yang. A unified primal dual active set algorithm for nonconvex sparse recovery. *Statistical Science*, 36(2):215–238, 2021.
- [16] Kaifeng Jiang, Defeng Sun, and Kim-Chuan Toh. A partial proximal point algorithm for nuclear norm regularized matrix least squares problems. *Mathematical Programming Computation*, 6(3):281–325, 2014.
- [17] Kshitij Khare, Sang-Yun Oh, and Bala Rajaratnam. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(4):803–825, 2015.

- [18] Penporn Koanantakool, Alnur Ali, Ariful Azad, Aydin Buluc, Dmitriy Morozov, Leonid Oliker, Katherine Yelick, and Sang-Yun Oh. Communication-avoiding optimization methods for distributed massive-scale sparse inverse covariance estimation. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84, pages 1376–1386. PMLR, 2018.
- [19] Sungdong Lee, Joshua Bang, Youngra Kim, Hyungwon Choi, Sang-Yun Oh, and Joong-Ho Won. Learning massive-scale partial correlation networks in clinical multi-omics studies with HP-ACCORD. *The Annals of Applied Statistics*, 2025. to appear.
- [20] Lu Li and Kim-Chuan Toh. An inexact interior point method for L_1 -regularized sparse covariance selection. *Mathematical Programming Computation*, 2(3):291–315, 2010.
- [21] Xudong Li, Defeng Sun, and Kim-Chuan Toh. A highly efficient semismooth Newton augmented Lagrangian method for solving Lasso problems. *SIAM Journal on Optimization*, 28(1):433–458, 2018.
- [22] Faming Liang and Bochao Jia. *Sparse graphical modeling for high dimensional data: a paradigm of conditional independence tests*. Chapman and Hall/CRC, 2023.
- [23] Meixia Lin, Yong-Jin Liu, Defeng Sun, and Kim-Chuan Toh. Efficient sparse semismooth Newton methods for the clustered Lasso problem. *SIAM Journal on Optimization*, 29(3):2026–2052, 2019.
- [24] Chenjin Ma, Mengyun Wu, and Shuangge Ma. Analysis of cancer omics data: a selective review of statistical techniques. *Briefings in Bioinformatics*, 23(2):bbab585, 2022.
- [25] José Mario Martínez and Liqun Qi. Inexact Newton methods for solving nonsmooth equations. *Journal of Computational and Applied Mathematics*, 60(1-2):127–145, 1995.
- [26] Robert Mifflin. Semismooth and semiconvex functions in constrained optimization. *SIAM Journal on Control and Optimization*, 15(6):959–972, 1977.
- [27] Felicia SL Ng, David Ruau, Lorenz Wernisch, and Berthold Göttgens. A graphical model approach visualizes regulatory relationships between genome-wide transcription factor binding profiles. *Briefings in Bioinformatics*, 19(1):162–173, 2018.
- [28] Jong-Shi Pang. Newton’s method for B-differentiable equations. *Mathematics of Operations Research*, 15(2):311–341, 1990.
- [29] Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.
- [30] Liqun Qi. Convergence analysis of some algorithms for solving nonsmooth equations. *Mathematics of Operations Research*, 18(1):227–244, 1993.
- [31] Liqun Qi and Jie Sun. A nonsmooth version of Newton’s method. *Mathematical Programming*, 58(1):353–367, 1993.
- [32] Guilherme V Rocha, Peng Zhao, and Bin Yu. A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (SPLICE). *arXiv preprint arXiv:0807.3734*, 2008.
- [33] Joong-Ho Won. Proximity operator of the matrix perspective function and its applications. *Advances in Neural Information Processing Systems*, 33:6305–6314, 2020.
- [34] Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.
- [35] Yangjing Zhang, Ning Zhang, Defeng Sun, and Kim-Chuan Toh. An efficient Hessian based algorithm for solving large-scale sparse group Lasso problems. *Mathematical Programming*, 179(1):223–263, 2020.

Supplementary Material for “Partial Correlation Network Estimation by Semismooth Newton Methods”

A Proof of Theorem 2.2

Recall that F is given in (8),

$$F(\omega, d) = \begin{bmatrix} S\omega + d \\ \omega_1 - \text{prox}_{-\log(\cdot)}(\omega_1 + d_1) \\ \omega_{-1} - T_\lambda(\omega_{-1} + d_{-1}) \end{bmatrix},$$

where S is the sample covariance matrix, T_λ is the soft-thresholding operator, and

$$\text{prox}_{-\log(\cdot)}(x) = \frac{x + \sqrt{x^2 + 4}}{2}.$$

A.1 F is locally Lipschitz

To prove the local Lipschitz continuity of F , it suffices to show that each component of F is Lipschitz continuous. Note that each row of $S\omega + d$ is differentiable, with gradient $(S_i, 1)^T$, which norm is bounded by $\sqrt{\|S\|^2 + 1}$, and $\|S\|$ is a largest eigenvalue of S . $\omega_1 - \text{prox}_{-\log(\cdot)}(\omega_1 + d_1)$ is also differentiable with gradient $(\frac{1}{2} - \frac{\omega_1 + d_1}{2\sqrt{(\omega_1 + d_1)^2 + 4}}, -\frac{1}{2} - \frac{\omega_1 + d_1}{2\sqrt{(\omega_1 + d_1)^2 + 4}})$, where the absolute value of each component are within $[0, 1]$. Finally, the soft-thresholding operator $T_\lambda(\omega_i + d_i)$ is Lipschitz continuous with a constant of 1, so we can conclude that each component of F is Lipschitz continuous.

A.2 Directional differentiability of F

In here, we not only establish the directional differentiability of F , but also derive its directional derivative explicitly. A vector-valued function $F : \mathbb{R}^m \rightarrow \mathbb{R}^l$ is directionally differentiable at $x \in \mathbb{R}^m$ in the direction v if and only if each of its components F_i (for $i = 1, \dots, l$) is directionally differentiable at x in the direction v . Note that it is not necessary to consider the differentiable components, as the directional derivative of a differentiable function F at a point x in the direction v is given by $F'(x; v) = \nabla F(x) \cdot v$. Accordingly, the directional derivative of the linear term $S\omega + d$ is simply $Sv_\omega + v_d$. Similarly, the directional derivative of the composite term $\omega_1 - \text{prox}_{-\log(\cdot)}(\omega_1 + d_1)$ is $(z + 1/2)v_{\omega,1} + (z - 1/2)v_{d,1}$ where $z = -\frac{\omega_1 + d_1}{2\sqrt{(\omega_1 + d_1)^2 + 4}}$.

Now, we compute directional derivative of the non-differentiable component. We use the definition of active sets and inactive sets in (13). Then, the direction derivatives falls into one of the following cases:

i) $i \in \mathcal{I}^+(\omega, d)$ and $v_{\omega,i} + v_{d,i} > 0$. Then $i \in \mathcal{A}^+(\omega + tv_\omega, d + tv_d)$. In this case, $F(\omega_i + tv_{\omega,i}, d_i + tv_{d,i}) - F(\omega_i, d_i) = -tv_{d,i} + \lambda - \omega_i - d_i = -tv_{d,i}$. The same argument works for $i \in \mathcal{I}^-(\omega, d)$ and $v_{\omega,i} + v_{d,i} < 0$.

ii) $i \in \mathcal{I}^+(\omega, d)$ and $v_{\omega,i} + v_{d,i} < 0$. Then, $i \in \mathcal{I}^0(\omega + tv_\omega, d + tv_d)$. In this case, $F(\omega_i + tv_{\omega,i}, d_i + tv_{d,i}) - F(\omega_i, d_i) = \omega_i + tv_{\omega,i} - \omega_i = tv_{\omega,i}$. The same argument works for $i \in \mathcal{I}^-(\omega, d)$ and $v_{\omega,i} + v_{d,i} > 0$.

iii) $i \in \mathcal{A}(\omega, d)$. Then, we can find $t_0 > 0$ such that $\mathcal{A}(\omega, d) \subset \mathcal{A}(\omega + tv_\omega, d + tv_d)$ for all $0 < t \leq t_0$ with

$$t_0 \leq \min_{j \in \mathcal{A}(\omega, d)} \inf \{t > 0 : |\omega_j + tv_{\omega,j} + d_j + tv_{d,j}| > \lambda\}.$$

For $0 < t \leq t_0$ and $i \in \mathcal{A}(\omega, d) \subset \mathcal{A}(\omega + tv_\omega, d + tv_d)$,

$$F(\omega_i + tv_{\omega,i}, d_i + tv_{d,i}) - F(\omega_i, d_i) = -tv_{d,i}.$$

iv) $i \in \mathcal{I}^0(\omega, d)$. Then, we can find $t_1 > 0$ such that $\mathcal{I}^0(\omega, d) \subset \mathcal{I}^0(\omega + tv_\omega, d + tv_d)$ for all $0 < t \leq t_1$ with

$$t_1 \leq \min_{j \in \mathcal{I}^0(\omega, d)} \inf \{t > 0 : |\omega_j + tv_{\omega,j} + d_j + tv_{d,j}| < \lambda\}.$$

For $0 < t \leq t_1$ and $i \in \mathcal{I}^0(\omega, d) \subset \mathcal{I}^0(\omega + tv_\omega, d + tv_d)$,

$$F(\omega_i + tv_{\omega,i}, d_i + tv_{d,i}) - F(\omega_i, d_i) = tv_{\omega,i}.$$

Note that t_0 and t_1 are independent of i . Hence, directional derivative of F is (12).

A.3 Strong Semismoothness of F

Now, we prove that F is strongly semismooth. From [31, Corollary 4.4], it is sufficient to check the component-wise strong semismoothness of F . First, since $S\omega + d$ is an affine function, strong semismoothness holds trivially.

Next, we verify the strong semismoothness of $((\omega_1 - d_1) - \sqrt{(\omega_1 + d_1)^2 + 4})/2$. It suffices to consider only the non-linear term $H(\omega_1, d_1) := \sqrt{(\omega_1 + d_1)^2 + 4}$. The derivatives of $H(\omega_1, d_1)$ with respect to ω_1 and d_1 are $\frac{\omega_1 + d_1}{\sqrt{(\omega_1 + d_1)^2 + 4}}$. The directional derivative of $H(\omega_1, d_1)$ in the direction of a vector $v = (v_\omega, v_d)^T$ is given by

$$H'(\omega_1, d_1; v_\omega, v_d) = \frac{\partial H}{\partial \omega_1} v_\omega + \frac{\partial H}{\partial d_1} v_d = \frac{\omega_1 + d_1}{\sqrt{(\omega_1 + d_1)^2 + 4}} (v_\omega + v_d).$$

and the generalized Jacobian of $H(\omega_1 + v_\omega, d_1 + v_d)$ is a matrix given by the partial derivatives with respect to ω_1 and d_1 :

$$V = \partial H(\omega_1 + h_\omega, d_1 + h_d) = \left(\frac{\omega_1 + h_\omega + d_1 + h_d}{\sqrt{(\omega_1 + h_\omega + d_1 + h_d)^2 + 4}} \quad \frac{\omega_1 + h_\omega + d_1 + h_d}{\sqrt{(\omega_1 + h_\omega + d_1 + h_d)^2 + 4}} \right).$$

Then,

$$Vh - H'(\omega_1, d_1; h_\omega, h_d) = \left(\frac{\omega_1 + h_\omega + d_1 + h_d}{\sqrt{(\omega_1 + h_\omega + d_1 + h_d)^2 + 4}} - \frac{\omega_1 + d_1}{\sqrt{(\omega_1 + d_1)^2 + 4}} \right) (h_\omega + h_d).$$

Using a Taylor expansion, we obtain $\frac{\omega_1 + h_\omega + d_1 + h_d}{\sqrt{(\omega_1 + h_\omega + d_1 + h_d)^2 + 4}} - \frac{\omega_1 + d_1}{\sqrt{(\omega_1 + d_1)^2 + 4}} = H(\omega_1 + h_\omega, d_1 + h_d) - H(\omega_1, d_1) = \mathcal{O}(\|h\|)$. Thus, $Vh - H'(\omega_1, d_1; h_\omega, h_d) = \mathcal{O}(\|h\|^2)$.

Lastly, we have the facts that the soft thresholding operator is strongly semismooth everywhere [16] and $\omega \rightarrow \omega + d$ does not introduce any discontinuity or nonlinear behavior in $T_\lambda(\omega)$. Therefore, $\omega_{-1} - T_\lambda(\omega_{-1} + d_{-1})$ is strongly semismooth.

Since all components of F are strongly semismooth, it follows that F is strongly semismooth.

B Proof of Theorem 2.4

Before proving Theorem 2.4, we present a well-known condition that are equivalent to semismoothness from [11, Theorem 2.9, Theorem 2.12].

Proposition B.1. *Let $F : U \rightarrow \mathbb{R}^m$ be defined on the open set $U \subset \mathbb{R}^n$. Then, for $x \in U$, the following statements are equivalent:*

- (a) F is semismooth at x .
- (b) F is locally Lipschitz continuous at x , all directional derivatives $F'(x; v)$ exist, and for any $G \in \partial F(x + v)$,

$$\|F(x + v) - F(x) - Gv\| = o(\|v\|) \quad \text{as } v \rightarrow 0.$$

Furthermore, the following statements are equivalent:

- (c) F is strongly semismooth at x .
- (d) F is locally Lipschitz continuous at x , all directional derivatives $F'(x; v)$ exist, and for any $G \in \partial F(x + v)$,

$$\|F(x + v) - F(x) - Gv\| = \mathcal{O}(\|v\|^2) \quad \text{as } v \rightarrow 0.$$

Note that the Newton iterates satisfy

$$\|x^{k+1} - x^*\| \leq \|G(x^k)^{-1}\| \|F(x^k) - F(x^*) - G(x^k)(x^k - x^*)\|, \quad (20)$$

provided that $x^k \in U$. Let $B(x^*, r)$ denote a ball of radius r centered at x^* contained in U , and let m be such that $\|G(x)^{-1}\| \leq m$ for all $x \in B(x^*, r)$. Assume that F is semismooth at x^* and let $\eta \in (0, 1]$ be arbitrary. Then, there exists $\rho \in (0, r)$ such that

$$\|F(x^* + v) - F(x^*) - G(x^* + v)v\| \leq \frac{\eta}{m} \|v\| \leq \frac{1}{m} \|v\| \quad (21)$$

for all $\|v\| < \rho$. Consequently, if we choose x^0 such that $\|x^0 - x^*\| < \rho$, then by induction from (20) and (21), for $v = x^k - x^*$, we have $\|x^{k+1} - x^*\| < \rho$ and $x^{k+1} \in B(x^*, \rho)$. It follows that the iterates are well-defined.

Assume further that F is strongly semismooth at x^* , meaning there exists a constant $c > 0$ such that

$$\|F(x^* + v) - F(x^*) - G(x^* + v)v\| \leq c\|v\|^2 = \frac{c \cdot m}{m} \|v\|^2, \quad (22)$$

for all $\|v\| < \rho$.

Then, if we choose x^0 such that $\|x^0 - x^*\| < \rho$, then by induction from (20) and (22), for $v = x^k - x^*$, we have

$$\|x^{k+1} - x^*\| \leq m\|F(x^k) - F(x^*) - G(x^k)(x^k - x^*)\| \leq c \cdot m \|x^k - x^*\|^2,$$

where $\|x^{k+1} - x^*\| < \rho$. Therefore, the iterates remain well-defined and satisfy

$$\|x^{k+1} - x^*\| / \|x^k - x^*\|^2 = O(1),$$

indicating quadratic convergence of the sequence $\{x^k\}$ to x^* .

Thus, under the assumption of strong semismoothness of F , the Newton iterates converge quadratically to x^* . $\square$

C Proof of Proposition 2.5

Let

$$\begin{aligned} \mathcal{A} &= \mathcal{A}(\omega, d) := \{i \in \{2, \dots, p\} : |\omega_i + d_i| > \lambda\}, \\ \mathcal{I} &= \mathcal{I}(\omega, d) := \{i \in \{2, \dots, p\} : |\omega_i + d_i| \leq \lambda\}. \end{aligned}$$

Since F is continuous, there exists a neighborhood $\mathcal{N}$ of (ω^*, d^*) such that for all $(\omega, d) \in \mathcal{N}$, we have $\bar{\mathcal{A}} = \{1\} \cup \mathcal{A} = \{1\} \cup \mathcal{A}^* = \bar{\mathcal{A}}^*$. Since $\mathbf{X}_{\bar{\mathcal{A}}^*}$ has full column rank, $\mathbf{X}_{\bar{\mathcal{A}}}$ also has full column rank for all (ω, d) sufficiently close to (ω^*, d^*) .

One of the generalized Jacobian $G(\omega, d) \in \partial F(\omega, d)$ can be expressed as

$$G(\omega, d) = \begin{pmatrix} X_1^T X_1 & X_1^T X_{\mathcal{A} \cup \mathcal{I}} & 1 & 0 \\ X_{\mathcal{A} \cup \mathcal{I}}^T X_1 & X_{\mathcal{A} \cup \mathcal{I}}^T X_{\mathcal{A} \cup \mathcal{I}} & 0 & \mathbf{I} \\ z + \frac{1}{2} & 0 & z - \frac{1}{2} & 0 \\ 0 & I_{\mathcal{I}} & 0 & -I_{\mathcal{A}} \end{pmatrix}$$

where $z = -\frac{\omega_1 + d_1}{2\sqrt{(\omega_1 + d_1)^2 + 4}}$. Note that we can permute the rows and columns of G as

$$G(\omega, d) = \begin{pmatrix} \frac{1}{n} X_{\bar{\mathcal{A}}}^T X_{\bar{\mathcal{A}}} & \frac{1}{n} X_{\bar{\mathcal{A}}}^T X_{\mathcal{I}} & I_{\bar{\mathcal{A}}} & 0 \\ \frac{1}{n} X_{\mathcal{I}}^T X_{\bar{\mathcal{A}}} & \frac{1}{n} X_{\mathcal{I}}^T X_{\mathcal{I}} & 0 & I_{\mathcal{I}} \\ ((z + \frac{1}{2})_1 + 0_{\mathcal{A}}) & 0 & ((z - \frac{1}{2})_1 - I_{\mathcal{A}}) & 0 \\ 0 & I_{\mathcal{I}} & 0 & 0_{\mathcal{I}} \end{pmatrix}.$$

where $(x_1 + X_{\mathcal{A}}) = \begin{pmatrix} x_1 & 0 \\ 0 & X_{\mathcal{A}} \end{pmatrix}$.

Let $G := \begin{pmatrix} M_1 & M_2 \\ M_3 & M_4 \end{pmatrix}$, where $M_1 = (\frac{1}{n} X_{\bar{\mathcal{A}}}^T X_{\bar{\mathcal{A}}})$, $M_2 = (\frac{1}{n} X_{\mathcal{I}}^T X_{\bar{\mathcal{A}}} \quad I_{\bar{\mathcal{A}}} \quad 0)$,

$$M_3 = \begin{pmatrix} \frac{1}{n} X_{\mathcal{A}}^T X_{\mathcal{I}} \\ ((z + \frac{1}{2})_1 + 0_{\mathcal{A}}) \\ 0 \end{pmatrix} \text{ and } M_4 = \begin{pmatrix} \frac{1}{n} X_{\mathcal{I}}^T X_{\mathcal{I}} & 0 & I_{\mathcal{I}} \\ 0 & ((z - \frac{1}{2})_1 - I_{\mathcal{A}}) & 0 \\ I_{\mathcal{I}} & 0 & 0_{\mathcal{I}} \end{pmatrix}.$$

Note that $((z - \frac{1}{2})_1 - I_{\mathcal{A}})^{-1}$ is $((1/(z - \frac{1}{2}))_1 - I_{\mathcal{A}})$ and

$$\begin{pmatrix} \frac{1}{n} X_{\mathcal{I}}^T X_{\mathcal{I}} & 0 & I_{\mathcal{I}} \\ 0 & ((z - \frac{1}{2})_1 - I_{\mathcal{A}}) & 0 \\ I_{\mathcal{I}} & 0 & 0_{\mathcal{I}} \end{pmatrix} \begin{pmatrix} 0_{\mathcal{I}} & 0 & I_{\mathcal{I}} \\ 0 & ((1/(z - \frac{1}{2}))_1 - I_{\mathcal{A}}) & 0 \\ I_{\mathcal{I}} & 0 & -\frac{1}{n} X_{\mathcal{I}}^T X_{\mathcal{I}} \end{pmatrix} = I.$$

We utilize Schur complement of G to formulate the inverse of G . Note that

$$\begin{aligned} G/M_4 &= M_1 - M_2 M_4^{-1} M_3 \\ &= \frac{1}{n} X_{\mathcal{A}}^T X_{\mathcal{A}} - \begin{pmatrix} \frac{1}{n} X_{\mathcal{I}}^T X_{\mathcal{A}} & I_{\mathcal{A}} & 0 \end{pmatrix} \\ &\quad \cdot \begin{pmatrix} 0_{\mathcal{I}} & 0 & I_{\mathcal{I}} \\ 0 & ((1/(z - \frac{1}{2}))_1 - I_{\mathcal{A}}) & 0 \\ I_{\mathcal{I}} & 0 & -\frac{1}{n} X_{\mathcal{I}}^T X_{\mathcal{I}} \end{pmatrix} \begin{pmatrix} \frac{1}{n} X_{\mathcal{A}}^T X_{\mathcal{I}} \\ ((z + \frac{1}{2})_1 + 0_{\mathcal{A}}) \\ 0 \end{pmatrix} \\ &= \frac{1}{n} X_{\mathcal{A}}^T X_{\mathcal{A}} - \left(\left(\frac{z + \frac{1}{2}}{z - \frac{1}{2}} \right)_1 + 0_{\mathcal{A}} \right). \end{aligned}$$

Given that $(X_{\mathcal{A}}^T X_{\mathcal{A}})$ is nonsingular, $(X_{\mathcal{A}}^T X_{\mathcal{A}}) - n \left(\left(\frac{z + \frac{1}{2}}{z - \frac{1}{2}} \right)_1 + 0_{\mathcal{A}} \right)$ is also nonsingular since $|z| < \frac{1}{2}$.

Thus, we can construct the inverse of G as

$$G^{-1}(x) = \begin{pmatrix} (G/M_4)^{-1} & -(G/M_4)^{-1} M_2 M_4^{-1} \\ -M_4^{-1} M_3 (G/M_4)^{-1} & M_4^{-1} + M_4^{-1} M_3 (G/M_4)^{-1} M_2 M_4^{-1} \end{pmatrix}.$$

□

D Derivation of a Generalized Jacobian for F

Let $F : \mathbb{R}^{2p} \rightarrow \mathbb{R}^{2p}$ be

$$F(\omega, d) = \begin{pmatrix} \mathbf{S}\omega + d \\ \omega_1 - \text{prox}_{-\log(\cdot)}(\omega_1 + d_1) \\ \omega_{-1} - T_{\lambda}(\omega_{-1} + d_{-1}) \end{pmatrix}, \quad \text{prox}_{-\log(\cdot)}(x) = \frac{x + \sqrt{x^2 + 4}}{2}$$

where $T_{\lambda}(\cdot)$ is the componentwise soft-thresholding operator.

$$\text{Let } F_1(\omega, d) = \mathbf{S}\omega + d \text{ and } F_2(\omega, d) = \begin{pmatrix} \text{prox}_{-\log(\cdot)}(\omega_1 + d_1) \\ T_{\lambda}(\omega_{-1} + d_{-1}) \end{pmatrix}.$$

Then, it is straightforward to see that $\frac{\partial F_1}{\partial \omega} = \mathbf{S}$, $\frac{\partial F_1}{\partial d} = \mathbf{I}$ and

$$\frac{\partial \text{prox}_{-\log(\cdot)}(\omega_1 + d_1)}{\partial \omega_1} = \frac{\partial \text{prox}_{-\log(\cdot)}(\omega_1 + d_1)}{\partial d_1} = \frac{1}{2} + \frac{\omega_1 + d_1}{2\sqrt{(\omega_1 + d_1)^2 + 4}}.$$

Note that for $i = 2, \dots, p$,

$$\frac{\partial T_{\lambda}(\omega_i + d_i)}{\partial \omega_i} = \frac{\partial T_{\lambda}(\omega_i + d_i)}{\partial d_i} = \begin{cases} 0 & |\omega_i + d_i| > \lambda \\ [0, 1] & |\omega_i + d_i| = \lambda \\ 1 & |\omega_i + d_i| < \lambda \end{cases}.$$

By adopting the convention $\frac{\partial T_{\lambda}(\omega_i + d_i)}{\partial \omega_i} = \frac{\partial T_{\lambda}(\omega_i + d_i)}{\partial d_i} = 0$ at the kink $\{|\omega_i + d_i| = \lambda\}$, the Jacobians simplify to $\frac{\partial F_2}{\partial \omega} = \mathbf{I} - \mathbf{J}$ and $\frac{\partial F_2}{\partial d} = -\mathbf{J}$ where $\mathbf{J} = \text{diag}(J_{ii})$ with

$$J_{11} = \frac{1}{2} + \frac{\omega_1 + d_1}{2\sqrt{(\omega_1 + d_1)^2 + 4}}, \quad J_{ii} = \begin{cases} 1 & |\omega_i + d_i| > \lambda \\ 0 & |\omega_i + d_i| \leq \lambda, \end{cases} \quad i = 2, \dots, p.$$

Thus,

$$\partial F(\omega, d) = \begin{pmatrix} \frac{\partial F_1}{\partial \omega} & \frac{\partial F_1}{\partial d} \\ \frac{\partial F_2}{\partial \omega} & \frac{\partial F_2}{\partial d} \end{pmatrix} = \begin{pmatrix} \mathbf{S} & \mathbf{I} \\ \mathbf{I} - \mathbf{J} & -\mathbf{J} \end{pmatrix},$$

which is equivalent to the example of the generalize Jacobian (10).

E Proofs of Theorem 3.4

E.1 Proof of Proposition 3.2

In Appendix A.2, we established more than the existence of the directional derivative; we demonstrated a stronger property that $F(x + tv) - F(x) - tF'(x; v)$ is exactly zero for sufficiently small $t > 0$. Specifically, the main result in A.2 can be summarized as follows.

Let $(\omega, d) \in \mathbb{R}^{2p}$. There exists $s > 0$ such that, for all $0 < t \leq s$,

$$\frac{F(\omega + tv_\omega, d + tv_d) - F(\omega, d)}{t} = \begin{pmatrix} Sv_\omega + v_d \\ (z + 1/2)v_{\omega,1} + (z - 1/2)v_{d,1} \\ \begin{cases} -v_{d,i} & i \in \mathcal{A} \\ v_{\omega,i} & i \in \mathcal{I}^0 \\ \min\{v_{\omega,i}, -v_{d,i}\} & i \in \mathcal{I}^+ \\ \max\{v_{\omega,i}, -v_{d,i}\} & i \in \mathcal{I}^- \end{cases} \end{pmatrix}.$$

The right-hand side defines the component-wise directional derivative $F'(\omega, d; v)$ and $F'(\omega, d; \lambda v) = \lambda F'(\omega, d; v)$ for $\lambda \geq 0$. Consequently, F satisfies the approximation property $\|F(x + v) - F(x) - F'(x; v)\| = o(\|v\|)$ as $v \rightarrow 0$. Since F is locally Lipschitz (see Appendix A.1) and the directional derivative exists at every v , we conclude that F is B-differentiable. $\square$

E.2 Proofs of Lemma 3.3 and Theorem 3.4

Proof of Lemma 3.3: For the “only if” part, suppose $v = (v_\omega, v_d) \in \mathbb{R}^{2p}$ satisfies (14)–(16). We then observe that (16) is equivalent to

$$\begin{aligned} & (v_{\omega,i} + \omega_i > 0 \cap -v_{d,i} + \omega_i = 0) \cup (v_{\omega,i} + \omega_i = 0 \cap -v_{d,i} + \omega_i \geq 0), \\ & (v_{\omega,i} + \omega_i < 0 \cap -v_{d,i} + \omega_i = 0) \cup (v_{\omega,i} + \omega_i = 0 \cap -v_{d,i} + \omega_i \leq 0), \end{aligned}$$

which in turn is equivalent to

$$\begin{aligned} v_{\omega, \mathcal{I}^+} + \omega_{\mathcal{I}^+} &\geq 0, \quad -v_{d, \mathcal{I}^+} + \omega_{\mathcal{I}^+} \geq 0, \quad \langle v_{\omega, \mathcal{I}^+} + \omega_{\mathcal{I}^+}, -v_{d, \mathcal{I}^+} + \omega_{\mathcal{I}^+} \rangle = 0, \\ v_{\omega, \mathcal{I}^-} + \omega_{\mathcal{I}^-} &\leq 0, \quad -v_{d, \mathcal{I}^-} + \omega_{\mathcal{I}^-} \leq 0, \quad \langle v_{\omega, \mathcal{I}^-} + \omega_{\mathcal{I}^-}, -v_{d, \mathcal{I}^-} + \omega_{\mathcal{I}^-} \rangle = 0. \end{aligned} \quad (23)$$

Rewrite (14)–(15) as

$$\begin{aligned} Sv_\omega + v_d &= -S\omega - d, \\ (z + 1/2)v_{\omega,1} + (z - 1/2)v_{d,1} &= -\frac{1}{2} \left((\omega_1 - d_1) - \sqrt{(\omega_1 + d_1)^2 + 4} \right), \end{aligned}$$

where $z = z(\omega_1, d_1) = -\frac{\omega_1 + d_1}{2\sqrt{(\omega_1 + d_1)^2 + 4}}$.

The second equation can be equivalently written as

$$v_{d,1} = bv_{\omega,1} + c, \quad (24)$$

where

$$b = -\left(z - \frac{1}{2}\right)^{-1} \left(z + \frac{1}{2}\right) > 0, \quad c = -\frac{1}{2} \left(z - \frac{1}{2}\right)^{-1} \left((\omega_1 - d_1) - \sqrt{(\omega_1 + d_1)^2 + 4} \right). \quad (25)$$

Let $E = -S\omega - d$ and $S_{\mathcal{R}, \mathcal{C}}$ denote the submatrix of S with rows indexed by $\mathcal{R}$ and columns indexed by $\mathcal{C}$. Then, for the first equation,

$$\begin{aligned} Sv_\omega + v_d = E &\iff \begin{pmatrix} S_{1,1}v_{\omega,1} + S_{1,2}v_{\omega,2} + \cdots + S_{1,p}v_{\omega,p} + v_{d,1} \\ S_{2,1}v_{\omega,1} + S_{2,2}v_{\omega,2} + \cdots + S_{1,p}v_{\omega,p} + v_{d,2} \\ \vdots \\ S_{p,1}v_{\omega,1} + S_{p,2}v_{\omega,2} + \cdots + S_{p,p}v_{\omega,p} + v_{d,p} \end{pmatrix} = \begin{pmatrix} E_1 \\ E_2 \\ \vdots \\ E_p \end{pmatrix} \\ &\iff \begin{pmatrix} (S_{1,1} + b)v_{\omega,1} + S_{1,2}v_{\omega,2} + \cdots + S_{1,p}v_{\omega,p} \\ S_{2,1}v_{\omega,1} + S_{2,2}v_{\omega,2} + \cdots + S_{1,p}v_{\omega,p} + v_{d,2} \\ \vdots \\ S_{p,1}v_{\omega,1} + S_{p,2}v_{\omega,2} + \cdots + S_{p,p}v_{\omega,p} + v_{d,p} \end{pmatrix} = \begin{pmatrix} E_1 - c \\ E_2 \\ \vdots \\ E_p \end{pmatrix}. \end{aligned}$$

Invoking the definition of $\mathcal{I}$ in (12), we can rearrange this equation as

$$\begin{pmatrix} (S_{1,1} + b)v_{\omega,1} + S_{1,-1}^T v_{\omega,-1} \\ S_{\mathcal{A},1} v_{\omega,1} + S_{\mathcal{A},-1}^T v_{\omega,-1} + v_{d,\mathcal{A}} \\ S_{\mathcal{I}^+,1} v_{\omega,1} + S_{\mathcal{I}^+,-1}^T v_{\omega,-1} + v_{d,\mathcal{I}^+} \\ S_{\mathcal{I}^-,1} v_{\omega,1} + S_{\mathcal{I}^-,-1}^T v_{\omega,-1} + v_{d,\mathcal{I}^-} \\ S_{\mathcal{I}^0,1} v_{\omega,1} + S_{\mathcal{I}^0,-1}^T v_{\omega,-1} + v_{d,\mathcal{I}^0} \end{pmatrix} = \begin{pmatrix} E_1 - c \\ E_{\mathcal{A}} \\ E_{\mathcal{I}^+} \\ E_{\mathcal{I}^-} \\ E_{\mathcal{I}^0} \end{pmatrix}.$$

In the above system, transpose the known $v_{\omega,\mathcal{I}^0}$ and $v_{d,\mathcal{A}}$ (from (8) and (12)) to the right-hand side of the equation to yield

$$\begin{pmatrix} (S_{1,1} + b)v_{\omega,1} + S_{1,\mathcal{A}\cup\mathcal{I}^+\cup\mathcal{I}^-}^T v_{\omega,\mathcal{A}\cup\mathcal{I}^+\cup\mathcal{I}^-} \\ S_{\mathcal{A},1} v_{\omega,1} + S_{\mathcal{A},\mathcal{A}\cup\mathcal{I}^+\cup\mathcal{I}^-}^T v_{\omega,\mathcal{A}\cup\mathcal{I}^+\cup\mathcal{I}^-} \\ S_{\mathcal{I}^+,1} v_{\omega,1} + S_{\mathcal{I}^+,\mathcal{A}\cup\mathcal{I}^+\cup\mathcal{I}^-}^T v_{\omega,\mathcal{A}\cup\mathcal{I}^+\cup\mathcal{I}^-} + v_{d,\mathcal{I}^+} \\ S_{\mathcal{I}^-,1} v_{\omega,1} + S_{\mathcal{I}^-,\mathcal{A}\cup\mathcal{I}^+\cup\mathcal{I}^-}^T v_{\omega,\mathcal{A}\cup\mathcal{I}^+\cup\mathcal{I}^-} + v_{d,\mathcal{I}^-} \\ S_{\mathcal{I}^0,1} v_{\omega,1} + S_{\mathcal{I}^0,\mathcal{A}\cup\mathcal{I}^+\cup\mathcal{I}^-}^T v_{\omega,\mathcal{A}\cup\mathcal{I}^+\cup\mathcal{I}^-} + v_{d,\mathcal{I}^0} \end{pmatrix} = \begin{pmatrix} E_1 - S_{1,\mathcal{I}^0}^T v_{\omega,\mathcal{I}^0} - c \\ E_{\mathcal{A}} - S_{\mathcal{A},\mathcal{I}^0}^T v_{\omega,\mathcal{I}^0} - v_{d,\mathcal{A}} \\ E_{\mathcal{I}^+} - S_{\mathcal{I}^+,\mathcal{I}^0}^T v_{\omega,\mathcal{I}^0} \\ E_{\mathcal{I}^-} - S_{\mathcal{I}^-,\mathcal{I}^0}^T v_{\omega,\mathcal{I}^0} \\ E_{\mathcal{I}^0} - S_{\mathcal{I}^0,\mathcal{I}^0}^T v_{\omega,\mathcal{I}^0} \end{pmatrix}.$$

Now, rearranging the above equation gives the following.

$$\underbrace{\begin{pmatrix} S_{1,1} + b & S_{1,\mathcal{A}} & S_{1,\mathcal{I}^+} & -S_{1,\mathcal{I}^-} & 0_{1,\mathcal{I}^0} \\ S_{\mathcal{A},1} & S_{\mathcal{A},\mathcal{A}} & S_{\mathcal{A},\mathcal{I}^+} & -S_{\mathcal{A},\mathcal{I}^-} & 0_{\mathcal{A},\mathcal{I}^0} \\ S_{\mathcal{I}^+,1} & S_{\mathcal{I}^+,\mathcal{A}} & S_{\mathcal{I}^+,\mathcal{I}^+} & -S_{\mathcal{I}^+,\mathcal{I}^-} & 0_{\mathcal{I}^+,\mathcal{I}^0} \\ -S_{\mathcal{I}^-,1} & -S_{\mathcal{I}^-,\mathcal{A}} & -S_{\mathcal{I}^-,\mathcal{I}^+} & S_{\mathcal{I}^-,\mathcal{I}^-} & 0_{\mathcal{I}^-,\mathcal{I}^0} \\ S_{\mathcal{I}^0,1} & S_{\mathcal{I}^0,\mathcal{A}} & S_{\mathcal{I}^0,\mathcal{I}^+} & -S_{\mathcal{I}^0,\mathcal{I}^-} & I_{\mathcal{I}^0,\mathcal{I}^0} \end{pmatrix}}_{=N} \begin{pmatrix} v_{\omega,1} \\ v_{\omega,\mathcal{A}} \\ v_{\omega,\mathcal{I}^+} \\ -v_{\omega,\mathcal{I}^-} \\ v_{d,\mathcal{I}^0} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ v_{d,\mathcal{I}^+} \\ -v_{d,\mathcal{I}^-} \\ 0 \end{pmatrix} = \underbrace{\begin{pmatrix} E_1 - S_{1,\mathcal{I}^0}^T v_{\omega,\mathcal{I}^0} - c \\ E_{\mathcal{A}} - S_{\mathcal{A},\mathcal{I}^0}^T v_{\omega,\mathcal{I}^0} - v_{d,\mathcal{A}} \\ E_{\mathcal{I}^+} - S_{\mathcal{I}^+,\mathcal{I}^0}^T v_{\omega,\mathcal{I}^0} \\ -E_{\mathcal{I}^-} + S_{\mathcal{I}^-,\mathcal{I}^0}^T v_{\omega,\mathcal{I}^0} \\ E_{\mathcal{I}^0} - S_{\mathcal{I}^0,\mathcal{I}^0}^T v_{\omega,\mathcal{I}^0} \end{pmatrix}}_{=P}. \quad (26)$$

Since $\mathbf{X}_{\bar{\mathcal{A}}}$ is full-rank,

$$\begin{pmatrix} S_{1,1} & S_{1,\mathcal{A}} & S_{1,\mathcal{I}^+} & S_{1,\mathcal{I}^-} \\ S_{\mathcal{A},1} & S_{\mathcal{A},\mathcal{A}} & S_{\mathcal{A},\mathcal{I}^+} & S_{\mathcal{A},\mathcal{I}^-} \\ S_{\mathcal{I}^+,1} & S_{\mathcal{I}^+,\mathcal{A}} & S_{\mathcal{I}^+,\mathcal{I}^+} & S_{\mathcal{I}^+,\mathcal{I}^-} \\ S_{\mathcal{I}^-,1} & S_{\mathcal{I}^-,\mathcal{A}} & S_{\mathcal{I}^-,\mathcal{I}^+} & S_{\mathcal{I}^-,\mathcal{I}^-} \end{pmatrix} = (1/n) (X_{1\cup\mathcal{A}\cup\mathcal{I}^+\cup\mathcal{I}^-}^T X_{1\cup\mathcal{A}\cup\mathcal{I}^+\cup\mathcal{I}^-}) \quad (27)$$

is nonsingular (in fact, positive definite). Denote the 4×4 principal block submatrix of N by

$$N_1 = \begin{pmatrix} S_{1,1} + b & S_{1,\mathcal{A}} & S_{1,\mathcal{I}^+} & -S_{1,\mathcal{I}^-} \\ S_{\mathcal{A},1} & S_{\mathcal{A},\mathcal{A}} & S_{\mathcal{A},\mathcal{I}^+} & -S_{\mathcal{A},\mathcal{I}^-} \\ S_{\mathcal{I}^+,1} & S_{\mathcal{I}^+,\mathcal{A}} & S_{\mathcal{I}^+,\mathcal{I}^+} & -S_{\mathcal{I}^+,\mathcal{I}^-} \\ -S_{\mathcal{I}^-,1} & -S_{\mathcal{I}^-,\mathcal{A}} & -S_{\mathcal{I}^-,\mathcal{I}^+} & S_{\mathcal{I}^-,\mathcal{I}^-} \end{pmatrix} \in \mathbb{R}^{|\bar{\mathcal{A}}| \times |\bar{\mathcal{A}}|} \quad (28)$$

so that $N = \begin{pmatrix} N_1 & 0 \\ C & I_{\mathcal{I}^0,\mathcal{I}^0} \end{pmatrix}$ with

$$C = (S_{\mathcal{I}^0,1}, S_{\mathcal{I}^0,\mathcal{A}}, S_{\mathcal{I}^0,\mathcal{I}^+}, -S_{\mathcal{I}^0,\mathcal{I}^-}). \quad (29)$$

Then N_1 is also nonsingular since $b > 0$, and thus

$$N^{-1} = \begin{pmatrix} N_1^{-1} & 0 \\ -CN_1^{-1} & I_{\mathcal{I}^0,\mathcal{I}^0} \end{pmatrix}.$$

Therefore,

$$Sv_{\omega} + v_d = E \iff \begin{pmatrix} v_{\omega,1} \\ v_{\omega,\mathcal{A}} \\ v_{\omega,\mathcal{I}^+} \\ -v_{\omega,\mathcal{I}^-} \\ v_{d,\mathcal{I}^0} \end{pmatrix} = N^{-1} \left[P - \begin{pmatrix} 0 \\ 0 \\ v_{d,\mathcal{I}^+} \\ -v_{d,\mathcal{I}^-} \\ 0 \end{pmatrix} \right]. \quad (30)$$

If we let the 2×2 principal diagonal block submatrix of N_1 be $N_2 = \begin{pmatrix} S_{1,1} + b & S_{1,\mathcal{A}} \\ S_{\mathcal{A},1} & S_{\mathcal{A},\mathcal{A}} \end{pmatrix}$, then the bottom right 2×2 block of N_1^{-1} is the Schur complement of N_2 in N_1 :

$$M := \left[\begin{pmatrix} S_{\mathcal{I}^+, \mathcal{I}^+} & -S_{\mathcal{I}^+, \mathcal{I}^-} \\ -S_{\mathcal{I}^-, \mathcal{I}^+} & S_{\mathcal{I}^-, \mathcal{I}^-} \end{pmatrix} - \begin{pmatrix} S_{\mathcal{I}^+, 1} & S_{\mathcal{I}^+, \mathcal{A}} \\ -S_{\mathcal{I}^-, 1} & -S_{\mathcal{I}^-, \mathcal{A}} \end{pmatrix} N_2^{-1} \begin{pmatrix} S_{1, \mathcal{I}^+} & -S_{1, \mathcal{I}^-} \\ S_{\mathcal{A}, \mathcal{I}^+} & -S_{\mathcal{A}, \mathcal{I}^-} \end{pmatrix} \right]^{-1} \quad (31)$$

Thus,

$$Sv_\omega + v_d = E \implies \begin{pmatrix} v_{\omega, \mathcal{I}^+} \\ -v_{\omega, \mathcal{I}^-} \end{pmatrix} = M \begin{pmatrix} -v_{d, \mathcal{I}^+} \\ v_{d, \mathcal{I}^-} \end{pmatrix} + [N_1^{-1}]_{3:4} P, \quad (32)$$

where $[N_1^{-1}]_{3:4}$ refers to the submatrix of N_1^{-1} spanning the third and fourth row blocks.

From the property of the Schur complement, M is positive definite if and only if N_1 is positive definite. To see this, write

$$N_1 = I_0 \begin{pmatrix} S_{1,1} + b & S_{1,\mathcal{A}} & S_{1,\mathcal{I}^+} & S_{1,\mathcal{I}^-} \\ S_{\mathcal{A},1} & S_{\mathcal{A},\mathcal{A}} & S_{\mathcal{A},\mathcal{I}^+} & S_{\mathcal{A},\mathcal{I}^-} \\ S_{\mathcal{I}^+,1} & S_{\mathcal{I}^+,\mathcal{A}} & S_{\mathcal{I}^+,\mathcal{I}^+} & S_{\mathcal{I}^+,\mathcal{I}^-} \\ S_{\mathcal{I}^-,1} & S_{\mathcal{I}^-,\mathcal{A}} & S_{\mathcal{I}^-,\mathcal{I}^+} & S_{\mathcal{I}^-,\mathcal{I}^-} \end{pmatrix} I_0,$$

where

$$I_0 = \begin{pmatrix} I_{1,1} & & & \\ & I_{\mathcal{A},\mathcal{A}} & & \\ & & I_{\mathcal{I}^+,\mathcal{I}^+} & \\ & & & -I_{\mathcal{I}^-, \mathcal{I}^-} \end{pmatrix}$$

is symmetric and orthogonal. The middle matrix is (27) that is positive definite.

If we let

$$\mathbf{x} = \begin{pmatrix} -v_{d, \mathcal{I}^+} + \omega_{\mathcal{I}^+} \\ v_{d, \mathcal{I}^-} - \omega_{\mathcal{I}^-} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} v_{\omega, \mathcal{I}^+} + \omega_{\mathcal{I}^+} \\ -v_{\omega, \mathcal{I}^-} - \omega_{\mathcal{I}^-} \end{pmatrix}, \quad (33)$$

then we see

$$\langle \mathbf{x}, \mathbf{y} \rangle = 0 \quad \text{and} \quad \mathbf{y} = M\mathbf{x} + q,$$

where

$$q = -M \begin{pmatrix} \omega_{\mathcal{I}^+} \\ -\omega_{\mathcal{I}^-} \end{pmatrix} + [N_1^{-1}]_{3:4} P + \begin{pmatrix} \omega_{\mathcal{I}^+} \\ -\omega_{\mathcal{I}^-} \end{pmatrix}. \quad (34)$$

from (23) and (32). In other words, the $\mathbf{x}$ and $\mathbf{y}$ in (33) jointly solve the LCP (17) with the M and q given in (31) and (34). In fact, (17) has a unique solution, since a finite-dimensional LCP of the form (17) has a unique solution if M is a P -matrix, meaning all of its principal minors are positive. A symmetric positive definite matrix is a P -matrix.

For the “if” part, suppose $(\mathbf{x}, \mathbf{y})$ solves the LCP (17) with the M and q given in (31) and (34). Then $(\mathbf{x}, \mathbf{y})$ is unique. Let

$$\begin{pmatrix} v_{d, \mathcal{I}^+} \\ v_{d, \mathcal{I}^-} \end{pmatrix} = \begin{pmatrix} -\mathbf{x}_{\mathcal{I}^+} \\ \mathbf{x}_{\mathcal{I}^-} \end{pmatrix} + \begin{pmatrix} \omega_{\mathcal{I}^+} \\ \omega_{\mathcal{I}^-} \end{pmatrix} \quad \text{for} \quad \mathbf{x} = \begin{pmatrix} \mathbf{x}_{\mathcal{I}^+} \\ \mathbf{x}_{\mathcal{I}^-} \end{pmatrix} \quad (35)$$

and

$$\begin{pmatrix} v_{\omega, \mathcal{I}^+} \\ -v_{\omega, \mathcal{I}^-} \end{pmatrix} = \left[N^{-1} \left(P - \begin{pmatrix} 0 \\ 0 \\ v_{d, \mathcal{I}^+} \\ -v_{d, \mathcal{I}^-} \\ 0 \end{pmatrix} \right) \right]_{3:4}.$$

Then

$$\begin{aligned} \begin{pmatrix} v_{\omega, \mathcal{I}^+} \\ -v_{\omega, \mathcal{I}^-} \end{pmatrix} &= \left[N^{-1} \left(P - \begin{pmatrix} 0 \\ 0 \\ -\mathbf{x}_{\mathcal{I}^+} + \omega_{\mathcal{I}^+} \\ -\mathbf{x}_{\mathcal{I}^-} - \omega_{\mathcal{I}^-} \\ 0 \end{pmatrix} \right) \right]_{3:4} \\ &= [N_1^{-1}]_{3:4} P - M \begin{pmatrix} \omega_{\mathcal{I}^+} \\ -\omega_{\mathcal{I}^-} \end{pmatrix} + M\mathbf{x} = M\mathbf{x} + q - \begin{pmatrix} \omega_{\mathcal{I}^+} \\ -\omega_{\mathcal{I}^-} \end{pmatrix} = \mathbf{y} - \begin{pmatrix} \omega_{\mathcal{I}^+} \\ -\omega_{\mathcal{I}^-} \end{pmatrix}, \end{aligned}$$

which implies (14)–(16). $\square$

Proof of Theorem 3.4: From the uniqueness of the solution to LCP (17), the v^k in (11) is well-defined: We know

$$v_{\omega, \mathcal{I}^0}^k = -\omega_{\mathcal{I}^0}^k, \quad v_{d, \mathcal{A}}^k = \lambda(\mathbf{1}_{\mathcal{A}^+}^\top, -\mathbf{1}_{\mathcal{A}^-}^\top)^\top - d_{\mathcal{A}}^k$$

from (8) and (12). That

$$\begin{pmatrix} v_{d, \mathcal{I}^+}^k \\ v_{d, \mathcal{I}^-}^k \end{pmatrix} = \begin{pmatrix} -\mathbf{x}_{\mathcal{I}^+} \\ \mathbf{x}_{\mathcal{I}^-} \end{pmatrix} + \begin{pmatrix} \omega_{\mathcal{I}^+}^k \\ \omega_{\mathcal{I}^-}^k \end{pmatrix}$$

follows from (35); the M and q for solving $(\mathbf{x}, \mathbf{y})$ in LCP (17) are given in (31) and (34). Also, from (24),

$$v_{d, 1}^k = b^k v_{\omega, 1}^k - c^k,$$

where b^k and c^k can be deduced from (25). Finally, with the above quantities computed, we have from (30) and (26),

$$\begin{pmatrix} v_{\omega, 1}^k \\ v_{\omega, \mathcal{A}}^k \\ v_{\omega, \mathcal{I}^+}^k \\ -v_{\omega, \mathcal{I}^-}^k \\ v_{d, \mathcal{I}^0}^k \end{pmatrix} = N^{-1} \left[P^k - \begin{pmatrix} 0 \\ 0 \\ v_{d, \mathcal{I}^+}^k \\ -v_{d, \mathcal{I}^-}^k \\ 0 \end{pmatrix} \right]$$

with

$$N^{-1} = \begin{pmatrix} N_1^{-1} & 0 \\ -CN_1^{-1} & I \end{pmatrix} \quad \text{and} \quad P^k = \begin{pmatrix} E_1^k - S_{1, \mathcal{I}^0}^T v_{\omega, \mathcal{I}^0}^k - c^k \\ E_{\mathcal{A}}^k - S_{\mathcal{A}, \mathcal{I}^0}^T v_{\omega, \mathcal{I}^0}^k - v_{d, \mathcal{A}}^k \\ E_{\mathcal{I}^+}^k - S_{\mathcal{I}^+, \mathcal{I}^0}^T v_{\omega, \mathcal{I}^0}^k \\ -E_{\mathcal{I}^-}^k + S_{\mathcal{I}^-, \mathcal{I}^0}^T v_{\omega, \mathcal{I}^0}^k \\ E_{\mathcal{I}^0}^k - S_{\mathcal{I}^0, \mathcal{I}^0}^T v_{\omega, \mathcal{I}^0}^k \end{pmatrix},$$

where N_1 and C are given in (28) and (29) and $E^k = -S\omega^k - d^k$. $\square$

F Proofs of Theorem 3.8

F.1 Proof of Theorem 3.8

To verify the global convergence of Algorithm 1, we utilize some regularity conditions of F as in [28, 25]. Consider following lemmas:

Lemma F.1. *The proximal gradient step in Algorithm 1 occurs only a finite number of times.*

Lemma F.2. *There exists a constant $L > 0$ and a neighborhood $\mathcal{N}$ of x^* such that for all $x \in \mathcal{N}$,*

$$\|F'(x; v)\| \geq L\|v\| \quad (36)$$

holds for all vectors $v \in \mathbb{R}^{2p}$.

Lemma F.3. *For every sequence $\{x^k\}$ converging to $x^\dagger$, $\{v^k\}$ converging to $v^\dagger$, and positive scalar sequence $\{t_k\}$ converging to zero,*

$$\limsup_{k \rightarrow \infty} \frac{\theta(x^k + t_k v^k) - \theta(x^k)}{t_k} \leq \lim_{k \rightarrow \infty} F(x^k)^T F'(x^k; v^k)$$

Proof of Theorem 3.8: We exclude the trivial case where x^k exactly hits the optimal solution during the iteration, in other words, $\theta(x^k) > 0$. First, we show that m_k in (18) always exists for all Newton iterations. To see this assume that

$$\theta(x^k + \rho^m v^k) - \theta(x^k) > -2\sigma \rho^m \theta(x^k), \quad (37)$$

holds for all $m \in \{m_j\}$ where m_j is an infinite increasing integer sequence. Then, dividing both sides by ρ^{m_j} and taking the limit as $j \rightarrow \infty$, we have

$$-2\sigma \theta(x^k) \leq \lim_{j \rightarrow \infty} \frac{\theta(x^k + \rho^{m_j} v^k) - \theta(x^k)}{\rho^{m_j}} = F'(x^k; v^k)^T F(x^k) = -2\theta(x^k).$$

as we assumed that $\sigma \in (0, 1)$ and $\theta(x^k) > 0$, we have a contradiction. So, there always exists m_k such that (18) holds. Since $\{\theta(x^k)\}$ is a strictly decreasing sequence bounded above zero, its limit exists and therefore,

$$0 = \lim_{k \rightarrow \infty} \{\theta(x^k) - \theta(x^{k+1})\} \geq 2\sigma \limsup_{k \rightarrow \infty} (\rho^{m_k} \theta(x^k)) = 0.$$

If $\limsup_{k \rightarrow \infty} \rho^{m_k} > 0$, then $\lim_{k \rightarrow \infty} \theta(x^k) = 0$. Since we assume that $\mathcal{I}^{+*} \cup \mathcal{I}^{-*} = \emptyset$ and $X_{\bar{\mathcal{A}}^*}$ has full column rank, every (in particular, the unique) generalized jacobian $\partial F(x^*)$ is nonsingular. Then, by Clarke's inverse function theorem [3], F admits a locally Lipschitz continuous inverse in a neighborhood of x^* . Consequently, for all sufficiently large k , we have $x^k = F^{-1}(F(x^k)) \rightarrow F^{-1}(0) = x^*$.

If $\limsup_{k \rightarrow \infty} \rho^{m_k} = 0$, then we have $\rho^{m_k} \theta(x^k) \rightarrow 0$ and

$$\theta(x^k + \rho^{m_k-1} v^k) - \theta(x^k) > -2\sigma \rho^{m_k-1} \theta(x^k).$$

Consider a subsequence of $\{x^k : k \in K\}$ that converges to $x^\dagger$ such that $\bar{\mathcal{A}}^\dagger = \bar{\mathcal{A}}^*$. For all $k \in K$ large enough, $x^k \in \mathcal{N}$, where $\mathcal{N}$ is the neighborhood of x^* specified in Lemma F.2. By Lemma F.2, it follows that $\|v^k\| \leq \frac{1}{L} \|F(x^k)\|$, which implies that the corresponding subsequence $\{v^k\}_{k \in K}$ is bounded. Thus, we can find a subsequence k_l, k_2 such that x^{k_l} and v^{k_l} converges into some $x^\dagger$ and $v^\dagger$, respectively. Then, by Lemma F.3, we have

$$\lim_{l \rightarrow \infty} F(x^{k_l})^T F'(x^{k_l}; v^{k_l}) \geq \lim_{l \rightarrow \infty} \frac{\theta(x^{k_l} + \rho^{m_{k_l}-1} v^{k_l}) - \theta(x^{k_l})}{\rho^{m_{k_l}-1}} \geq -2\sigma \theta(x^\dagger)$$

On the other hand, since v^{k_l} is a Newton direction for (11),

$$\lim_{l \rightarrow \infty} F(x^{k_l})^T F'(x^{k_l}; v^{k_l}) = \lim_{l \rightarrow \infty} -2\theta(x^{k_l}) = -2\theta(x^\dagger).$$

Therefore, $\theta(x^\dagger) = 0$ and since the solution is unique, the accumulation point $x^\dagger$ coincides with x^* .

Finally, we establish the convergence of the full sequence $\{x^k\}$. Since x^* is the unique solution to $F(x) = 0$, there exists $\delta > 0$ such that x^* is the unique minimizer of $\|F(x)\|$ for $\|x - x^*\| \leq \delta$. Recall that for all k large enough, $\|v^k\| \leq \frac{1}{L} \|F(x^k)\|$ from Lemma F.2. Since $\|F(x^k)\| \rightarrow 0$ as $k \rightarrow \infty$, it follows that $\|v^k\| \rightarrow 0$. Therefore, there exists a constant $\delta' \in (0, \delta)$ such that, we have

$$\|x^{k+1} - x^k\| \leq \|v^k\| \leq \delta - \delta'.$$

whenever $\|x^k - x^*\| \leq \delta'$. Let $\mu > 0$ denote the minimum value of $\|F(x)\|$ over the region $\delta' \leq \|x - x^*\| \leq \delta$. Suppose that $\|x^{k_0} - x^*\| \leq \delta'$ and $\|F(x^{k_0})\| < \mu$. Then, since the sequence $\|F(x^k)\|$ is strictly decreasing, it follows that $\|F(x^{k_0+1})\| < \mu$. But, since $\|x^{k_0+1} - x^{k_0}\| \leq \delta - \delta'$, we also have that

$$\|x^{k_0+1} - x^*\| \leq \|x^{k_0+1} - x^{k_0}\| + \|x^{k_0} - x^*\| \leq \delta.$$

Hence, by the definition of μ , we have $\|x^{k_0+1} - x^*\| \leq \delta'$. By induction, it follows that $\|x^k - x^*\| \leq \delta'$ and $\|F(x^k)\| < \mu$ for all $k \geq k_0$. Since x^* is the unique accumulation point of $\{x^k\}$ within this neighborhood, we conclude that the entire sequence $\{x^k\}$ converges to x^* .

We now characterize the local convergence behavior of the iterates. By [30, Theorem 4.3], since $x^k \rightarrow x^*$, the unique solution to $F(x) = 0$, and $F'(\cdot; \cdot)$ in (12) is semicontinuous of degree 2 [30, Lemma 2.3], the sequence $\{x^k\}$ converges quadratically to x^* in its neighborhood and ρ^m eventually becomes 1. $\square$

F.2 Proof of Lemma F.1

The proximal gradient algorithm of [19] generates an iterate sequence to converge linearly to the unique solution x^* . We distinguish between two contexts in which the proximal gradient step is invoked: the inner loop (lines 14–16) and the outer loop (lines 12–17).

First consider the inner loop. Let $\{x_{\text{prox}}^{k_j}\}$ denote the sequence of iterates computed by the proximal gradient update. This update is applied iteratively from the initial point $x_{\text{prox}}^{k_0} = x^k$. Assume $\theta(x_{\text{prox}}^{k_0}) > 0$. (Otherwise we are done.) Since $x_{\text{prox}}^{k_j} \rightarrow x^*$ as $j \rightarrow \infty$, $\theta(x_{\text{prox}}^{k_j}) \rightarrow 0$ as $j \rightarrow \infty$, and

by linear convergence, there exists a finite index J such that $\theta(x_{\text{prox}}^{k_J}) \leq \theta(x_{\text{prox}}^{k_0})$. Thus, the inner loop terminates in finitely many iterations.

Next, consider the proximal gradient phase of the outer loop. This phase is potentially invoked until the iterate enters a neighborhood $\mathcal{N}$ of $x^* = (\omega^*, d^*)$ where $\bar{\mathcal{A}} = \bar{\mathcal{A}}^*$ holds (such a neighborhood exists due to the assumption of Theorem 3.8). If the iterate does not enter $\mathcal{N}$ after a finite number of invocations, then $\theta(x_{\text{prox}}^{k_j}) \not\rightarrow 0$, a contradiction.

Once the (outer) sequence $\{x^k\}$ enters $\mathcal{N}$, within this neighborhood the rank condition for local B-semismooth Newton convergence is satisfied, so the algorithm remains in the Newton phase. Consequently, the proximal gradient step is invoked only a finite number of times throughout the entire procedure. $\square$

F.3 Proof of Lemma F.2

Let

$$\begin{aligned}\mathcal{A} &= \{i \in \{2, \dots, p\} : |\omega_i + d_i| > \lambda\}, \\ \mathcal{I} &= \{i \in \{2, \dots, p\} : |\omega_i + d_i| \leq \lambda\}.\end{aligned}$$

Recall that the Bouligand derivative $F'(\omega, d; v)$ can be expressed with the generalized Jacobian $G_{\mathcal{A}}(\omega, d)$ as

$$F'(\omega, d; v) = G_{\mathcal{A}}(\omega, d)v = \begin{pmatrix} X_1^T X_1 & X_1^T X_{\mathcal{A} \cup \mathcal{I}} & 1 & 0 \\ X_{\mathcal{A} \cup \mathcal{I}}^T X_1 & X_{\mathcal{A} \cup \mathcal{I}}^T X_{\mathcal{A} \cup \mathcal{I}} & 0 & \mathbf{I} \\ z + \frac{1}{2} & 0 & z - \frac{1}{2} & 0 \\ 0 & I_{\mathcal{I}} & 0 & -I_{\mathcal{A}} \end{pmatrix} v$$

for the differentiable points ($|\omega_i + d_i| \neq \lambda$) where $z = -\frac{\omega_1 + d_1}{2\sqrt{(\omega_1 + d_1)^2 + 4}}$. Since we know that the value of the $(p+i)$ -th element of $F'(\omega, d; v)$ for $i \in \mathcal{I}^\pm$ is $-v_{d,i}$ or $v_{\omega,i}$, we can construct the following inequality:

$$\|F'(\omega, d; v)\|^2 \geq \inf_{B \subseteq \mathcal{I}^\pm} \|G_{\mathcal{A} \cup B}(\omega, d)v\|$$

Thus, it suffices to show that the generalized Jacobian $G_{\mathcal{A}}(\omega, d)$ is nonsingular. Then, it can be deduced that

$$\|G_{\mathcal{A}}^{-1}(\omega, d)\| \|G_{\mathcal{A}}(\omega, d)v\| \geq \|v\|,$$

and thus, L in (36) holds for the minimum value of $\|G_{\mathcal{A}}^{-1}(\omega, d)\|$. Since F is continuous and $\mathcal{I}^{+*} \cup \mathcal{I}^{-*} = \emptyset$, there exists a neighborhood $\mathcal{N}$ of (ω^*, d^*) such that for all $(\omega, d) \in \mathcal{N}$, we have $\bar{\mathcal{A}} = \bar{\mathcal{A}}^*$. Since $\mathbf{X}_{\bar{\mathcal{A}}^*}$ has full column rank, $\mathbf{X}_{\bar{\mathcal{A}}}$ also has full column rank for all (ω, d) sufficiently close to (ω^*, d^*) . We already proved that such $G_{\mathcal{A}}$ has an inverse in Appendix C. $\square$

F.4 Proof of Lemma F.3

Note that

$$\lim_{k \rightarrow \infty} \frac{F_i^2(x^k + t_k v^k) - F_i^2(x^k)}{2t_k} \leq \lim_{k \rightarrow \infty} F(x^k)^T F'(x^k; v^k) \quad (38)$$

holds as the equality within its domain of x for $i = 1, \dots, p+1$, and the equality also holds for $i = p+1, \dots, 2p$ where $|\omega_j + d_j| \neq \lambda$ ($j = 2, \dots, p$), as F_i 's are differentiable and F_i' 's are continuous at these points. Hence, it suffices to prove that (38) also holds for non-differential points $|\omega + d| = \lambda$. Suppose that (ω^k, d^k) converges to $(\omega^\dagger, d^\dagger)$ and (v_ω^k, v_d^k) converges to $(v_\omega^\dagger, v_d^\dagger)$. Without loss of generality, let $\omega^\dagger + d^\dagger = \lambda$. Let $k_1, k_2, \dots$ be any increasing sequence of a positive integer. Then, for any $i \in \{p+1, \dots, 2p\}$, we can find some subsequence k_{m_l} where

$$\lim_{l \rightarrow \infty} \frac{F_i((\omega^{k_{m_l}} + t_{k_{m_l}} v_\omega^{k_{m_l}}, d^{k_{m_l}} + t_{k_{m_l}} v_d^{k_{m_l}})) - F_i((\omega^{k_{m_l}}, d^{k_{m_l}}))}{t_{k_{m_l}}} \quad (39)$$

exists and falls into at least one of the following cases:

1. $\omega^{k_{m_l}} + d^{k_{m_l}} > \lambda$ and $\omega^{k_{m_l}} + d^{k_{m_l}} + t_{k_{m_l}}(v_\omega^{k_{m_l}} + v_d^{k_{m_l}}) \geq \lambda$. Then,

$$(39) = \lim_{l \rightarrow \infty} \frac{t_{k_{m_l}} v_\omega^{k_{m_l}} - t_{k_{m_l}}(v_\omega^{k_{m_l}} + v_d^{k_{m_l}})}{t_{k_{m_l}}} = -v_d^\dagger = \lim_{l \rightarrow \infty} F'_i(x^{k_{m_l}}; v^{k_{m_l}}).$$

2. $\omega^{k_{m_l}} + d^{k_{m_l}} < \lambda$ and $\omega^{k_{m_l}} + d^{k_{m_l}} + t_{k_{m_l}}(v_\omega^{k_{m_l}} + v_d^{k_{m_l}}) \geq \lambda$. Then,

$$(39) = \lim_{l \rightarrow \infty} \frac{t_{k_{m_l}} v_\omega^{k_{m_l}} - (\omega^{k_{m_l}} + d^{k_{m_l}} + t_{k_{m_l}}(v_\omega^{k_{m_l}} + v_d^{k_{m_l}}) - \lambda)}{t_{k_{m_l}}} \\ \leq \lim_{l \rightarrow \infty} (t_{k_{m_l}} v_\omega^{k_{m_l}})/t_{k_{m_l}} = v_\omega^\dagger = \lim_{l \rightarrow \infty} F'_i(x^{k_{m_l}}; v^{k_{m_l}}).$$

3. $\omega^{k_{m_l}} + d^{k_{m_l}} > \lambda$ and $\omega^{k_{m_l}} + d^{k_{m_l}} + t_{k_{m_l}}(v_\omega^{k_{m_l}} + v_d^{k_{m_l}}) \leq \lambda$. Then,

$$(39) = \lim_{l \rightarrow \infty} \frac{t_{k_{m_l}} v_\omega^{k_{m_l}} + (\omega^{k_{m_l}} + d^{k_{m_l}} - \lambda)}{t_{k_{m_l}}} \\ \leq \lim_{l \rightarrow \infty} \frac{t_{k_{m_l}} v_\omega^{k_{m_l}} - t_{k_{m_l}}(v_\omega^{k_{m_l}} + v_d^{k_{m_l}})}{t_{k_{m_l}}} = -v_d^\dagger = \lim_{l \rightarrow \infty} F'_i(x^{k_{m_l}}; v^{k_{m_l}}).$$

4. $\omega^{k_{m_l}} + d^{k_{m_l}} < \lambda$ and $\omega^{k_{m_l}} + d^{k_{m_l}} + t_{k_{m_l}}(v_\omega^{k_{m_l}} + v_d^{k_{m_l}}) \leq \lambda$. Then,

$$(39) = \lim_{l \rightarrow \infty} (t_{k_{m_l}} v_\omega^{k_{m_l}})/t_{k_{m_l}} = v_\omega^\dagger = \lim_{l \rightarrow \infty} F'_i(x^{k_{m_l}}; v^{k_{m_l}}).$$

5. $\omega^{k_{m_l}} + d^{k_{m_l}} = \lambda$. Then,

$$(39) = \lim_{l \rightarrow \infty} (t_{k_{m_l}} v_\omega^{k_{m_l}} - t_{k_{m_l}} \max\{v_\omega^{k_{m_l}} + v_d^{k_{m_l}}, 0\})/t_{k_{m_l}} \\ = \min\{v_\omega^\dagger, -v_d^\dagger\} \leq \lim_{l \rightarrow \infty} F'_i(x^{k_{m_l}}; v^{k_{m_l}}).$$

Hence, we can conclude that

$$\limsup_{k \rightarrow \infty} \frac{F_i(x^k + t_k v^k) - F_i(x^k)}{t_k} \leq \lim_{k \rightarrow \infty} F'_i(x^k; v^k)$$

holds for all i . Then, using the continuity of F it follows that

$$\limsup_{k \rightarrow \infty} \frac{\theta(x^k + t_k v^k) - \theta(x^k)}{t_k} \leq \lim_{k \rightarrow \infty} F(x^k)^T F'(x^k; v^k)$$

with some simple arithmetic. $\square$

G Additional Experiments

G.1 Ablation study: roles of the proximal update and LCP solving

Algorithm 1 includes two components other than the semismooth Newton steps, that are the proximal update steps and the LCP solve step. To disentangle the roles of these two components, we replicated the setup of Section 4.1 and ran each configuration for 50 iterations under three ablations: (i) disabling only the proximal update steps, (ii) disabling only the LCP solve step, and (iii) disabling both. The LCP solve step is invoked only when, for a given row, $\mathcal{I}^+ \cup \mathcal{I}^- \neq \emptyset$. In this ablation study, “disabling the LCP solve step” means reclassifying all indices in $\mathcal{I}^+ \cup \mathcal{I}^-$ as inactive by assigning them to $\mathcal{I}^0$, thereby bypassing the LCP solve subroutine. Recall that with both components enabled, Algorithm 1 terminated within 30 iterations.

At $\lambda = 0.10$, disabling the proximal steps prevented the algorithm from converging within 50 iterations, regardless of whether the LCP solve component was enabled. By contrast, convergence was achieved whenever the proximal gradient component was retained, even without the LCP solve

comeponent. In the original run at this value of λ , the proximal update was invoked in 229 out of 1000 rows in the first iteration, indicating its importance in the denser regime. For $\lambda = 0.15, 0.2$, the algorithm converged under all ablation settings. In the original runs, at most one proximal update (out of 1000 rows) occurred in the first iteration, after which the semismooth Newton updates predominated. Thus, while the proximal gradient steps can be dispensable in these sparser regimes, it remains a practical safeguard unless the sparsity level is known a priori.

The LCP solve step was activated only on the kink set $\{i : |\omega_i + d_i| = \lambda\}$. Since exact equality is numerically rare, such activations almost never occurred. Consistent with this observation, Table 4 shows that Algorithm 1 is largely insensitive to this event, and Table 5 reports that the LCP solve occurs for a negligible fraction of rows. Even with 10,000 Newton row updates per iteration, the LCP solve occurred in about one or fewer row on average, as reflected by the small average and maximum per-iteration fractions. Thus, the practical impact of the LCP module is limited, although it remains theoretically indispensable for correctness and robustness, especially on unseen data.

λ	Full method	No Proximal	No LCP	No Proximal + No LCP
0.10	Converged (30)	Diverged	Converged (30)	Diverged
0.15	Converged (22)	Converged (26)	Converged (22)	Converged (26)
0.20	Converged (14)	Converged (14)	Converged (14)	Converged (14)

Table 4: Convergence within 50 iterations under ablations of the proximal and LCP modules.

λ	Avg. frac. LCP / iter.	Max frac. LCP / iter.	Last iter. LCP occurred	Total iters.
0.20	0.00005	0.0011	62	159
0.45	0.00004	0.0012	29	69
0.60	0.00012	0.0045	31	63

Table 5: Incidence of LCP updates in the original runs. “Frac.” denotes the fraction of rows updated by LCP in an iteration.

G.2 Sensitivity to the choice of regularization parameter λ

We assessed the convergence behavior of Algorithm 1 across a range of regularization parameters under the same experimental setup as described in Section 4.3, focusing on real-world datasets. The baseline choice $\lambda = 0.450$ was selected by the extended Bayesian information criterion in [19]. To examine sensitivity, we varied λ from 0.325 to 0.575 with an increment of 0.025 and recorded the iteration count, the number of nonzeros in the estimated precision matrix, and the resulting fraction of nonzero elements (density). The result is summarized in Table 6.

The change of the density remained modest within the range $\lambda \in [0.4, 0.5]$, indicating that the sparsity pattern is relatively insensitive around the “optimal” $\lambda = 0.450$. Across the tested λ values, Algorithm 1 converged within a few tens of iterations; while the iteration count tended to increase mildly as λ decreases (i.e., as the solution becomes denser), the overall convergence behavior remained stable.

G.3 Convergence across densities and alternate random graph topologies

In this section, we further investigate the trade-off between graph density and convergence using alternative random graph structures. In Section 4.2, the ground truth precision matrix Θ^* was constructed based on an Erdős–Rényi graph, where edges are sampled uniformly at random. To capture a broader variety of graph structures, we additionally consider two alternative random graph topologies: hub networks and scale-free graphs.

A hub network was generated using the following procedure: (1) for each cluster, $(300 \times \text{density})$ hub nodes were selected, each of which was connected to 66 randomly chosen nodes; (2) the remaining $(100 - 300 \times \text{density})$ nodes within each cluster were connected via an Erdős–Rényi graph, using the remaining intra-cluster edges.

λ	Iterations	Nonzeros	Density (%)
0.325	103	5,494,644	5.89×10^{-3}
0.350	69	4,806,952	5.15×10^{-3}
0.375	78	4,214,966	4.52×10^{-3}
0.400	67	3,699,259	3.96×10^{-3}
0.425	94	3,248,215	3.48×10^{-3}
0.450	64	2,852,968	3.06×10^{-3}
0.475	67	2,505,706	2.68×10^{-3}
0.500	71	2,200,810	2.36×10^{-3}
0.525	56	1,930,194	2.07×10^{-3}
0.550	66	1,692,240	1.81×10^{-3}
0.575	65	1,481,495	1.59×10^{-3}

Table 6: Iteration count, number of nonzero elements, and the fraction of nonzero elements in the estimated precision matrix for the LHC dataset with varying regularization parameter λ .

A scale-free graph was generated so that the degree k of the nodes has a probability $P(k) \sim k^{-2.3}$.

Each graph comprised 10 clusters with 100 nodes per cluster. The total number of edges was determined by the graph density defined in Section 4.2, with 80% of the edges allocated to intra-cluster connections and the remaining 20% to inter-cluster connections. The resulting graphs are summarized in Table 7.

Density	Hub Nodes	Intra-cluster Edges per Cluster
0.03	9	1198
0.05	15	1998
0.10	30	3996

Table 7: The number of hub nodes and intra-cluster edges per cluster according to graph density

Following the experimental setup described in Section 4.2, the regularization parameter λ was chosen so that the estimated precision matrix exhibits a sparsity level similar to that of the true matrix Θ^* . We then assessed whether Algorithm 1 converges within 300 iterations across various graph types and density levels (0.03, 0.05, and 0.1). The results are summarized in Tables 8–11.

We observe that the convergence behavior in different graph density levels is primarily determined by the overall density of the true precision matrix (i.e., graph structure), rather than by the specific graph structures. Intuitively, although the graph structures considered in this experiment may exhibit diverse local structural patterns, their overall density was controlled to remain constant across all settings. Since the applicability of the B-semismooth Newton method depends more on overall graph density than on local structural variations, the observed results are consistent with this interpretation.

Hub Network						
Density \ λ	0.2	0.15	0.1	0.09	0.08	0.07
0.03	Y	Y	-	-	-	-
0.05	-	Y	Y	-	-	-
0.1	-	-	Y	Y	Y	N

Table 8: Convergence within 300 iterations for varying graph density and λ

Density \ λ	0.07	0.06	0.05
0.1	3%	83.3%	100%

Table 9: Proportion of rows updated by proximal method after 300 iterations

Scale-free graph						
Density \ λ	0.2	0.15	0.1	0.09	0.08	0.07
0.03	Y	Y	-	-	-	-
0.05	-	Y	Y	-	-	-
0.1	-	-	Y	Y	Y	N

Table 10: Convergence within 300 iterations for varying graph density and λ

Density \ λ	0.07	0.06	0.05
0.1	1.6%	85.5%	99.8%

Table 11: Proportion of rows updated by proximal method after 300 iterations

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [\[Yes\]](#)

Justification: All claims made in the abstract and introduction are justified with theorems and proofs.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[No\]](#)

Justification: We do not provide explicit discussions on the limitations of our method as they align with the original method in [19].

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: The theoretical results of the paper are provided with assumptions and proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: The code for the main experiments is provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

- (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: **[Yes]**

Justification: Sufficient access to the data and the code is provided.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: **[Yes]**

Justification: The experiment details are specified.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: **[No]**

Justification: Reporting error bars on the experiments is not relevant with our research.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Computer resources for the experiments are specified.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We have reviewed the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This research discusses an optimization algorithm with no societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The release of data or models does not have such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The data used in the experiment is properly credited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This research does not involve crowdsourcing nor human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This research does not involve human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method of this research does not involve LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.