7 Appendix

7.1 Algorithm

We present the detailed algorithm description as Algorithm 2.

Algorithm 2: Quadratic Approximation method for Sparse Inverse Covariance Learning (*QUIC*)

Input : Empirical covariance matrix S, scalar λ, initial X_0, inner stopping tolerance ϵ, parameters $0 < \sigma < 0.5,\ 0 < \beta < 1$

Output: Sequence of X_t converging to $\arg\min_{X \succ 0} f(X)$, where $f(X) = -\log\det X + \text{tr}(SX) + \lambda\|X\|_1$.

Compute $W_0 = X_0^{-1}$.
for $t = 0, 1, \ldots$ **do**
  $D = 0, U = 0$
  while *not converged* **do**
    Partition the variables into fixed and free sets:
    $S_{fixed} := \{(i,j) \mid |\nabla_{ij} g(X_t)| < \lambda - \epsilon \text{ and } (X_t)_{ij} = 0\}, S_{free} := \mathcal{N} \setminus S_{fixed}$.
    for $(i,j) \in S_{free}$ **do**
      $a = w_{ij}^2 + w_{ii} w_{jj}$
      $b = s_{ij} - w_{ij} + \mathbf{w}_{:i}^T \mathbf{u}_{\cdot j}$
      $c = x_{ij} + d_{ij}$
      $\mu = -c + \mathcal{S}(c - b/a, \lambda/a)$
      $d_{ij} \leftarrow d_{ij} + \mu$
      $\mathbf{u}_{i\cdot} \leftarrow \mathbf{u}_{i\cdot} + \mu \mathbf{w}_{j\cdot}$
      $\mathbf{u}_{j\cdot} \leftarrow \mathbf{u}_{j\cdot} + \mu \mathbf{w}_{i\cdot}$
    end
  end
  for $\alpha = 1, \beta, \beta^2, \ldots$ **do**
    Compute the Cholesky factorization $LL^T = X_t + \alpha D$.
    if $X_t + \alpha D \not\succ 0$ **then**
      continue
    end
    Compute $f(X_t + \alpha D)$ from L and $X_t + \alpha D$
    if $f(X_t + \alpha D) \leq f(X_t) + \alpha\sigma\left[\text{tr}(\nabla g(X_t) D) + \lambda\|X_t + D\|_1 - \lambda\|X\|_1\right]$ **then**
      break
    end
  end
  $X_{t+1} = X_t + \alpha D$
  Compute $W_{t+1} = X_{t+1}^{-1}$ reusing the Cholesky factor.
end

7.2 Convergence guarantee (Proof of Theorem 1)

In this section, we prove that Algorithm 2 converges to the global optimum. Our proof is based on the proof in [17], which was developed for coordinate gradient descent methods. [17] considers composite objectives of the form

$$F(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x}), \tag{13}$$

where $g(\mathbf{x})$ is sufficiently smooth (continuously differentiable) and $h(\mathbf{x})$ is non-differentiable but separable. Recall, that in our case, $g(X) = -\log\det X + \text{tr}(SX)$ and $h(X) = \lambda\|X\|_1$. In [17] it is assumed that $g(X)$ is smooth over the domain $\mathbb{R}^n$. In our case $g(X)$ is smooth over the restricted domain of the positive definite cone S_n^{++}. We extend the analysis so that convergence still holds under our setting.

7.2.1 Notation

In the following arguments, capital letters such as $X, \bar{X}, A$ are $p \times p$ matrices, and I is the identity matrix. $f(X)$ is our objective function defined by (2). As is standard [13], the domain of the convex function $-\log\det$ is extended to S^p ($p \times p$ symmetric matrices) by

$$-\log\det X = \begin{cases} -\sum_{i=1}^{n} \log(\lambda_i(X)), & \text{if } X \succ 0 \\ \infty, & \text{otherwise} \end{cases}$$

where $\lambda_i(X)$ is the ith eigenvalue of X. We use $\|X\|_2$ to define the induced two norm of a matrix, and $\|D\|_F$ to denote the 2-norm of $\mathrm{vec}(D)$, which is equal to the Frobenius norm of the matrix D.

We are only dealing with symmetric matrices, and therefore we restrict our attention to the upper triangular indices denoted by $\mathcal{N} \equiv \{(i,j) \mid 1 \le i \le j \le p\}$. The matrix function $g(X)$ can be viewed as an $\mathbb{R}^{|\mathcal{N}|} \to \mathbb{R}$ function operating on the vector containing the upper triangular elements of X. The gradient $\nabla g(X)$ accordingly becomes an $\mathbb{R}^{|\mathcal{N}|}$ vector, while the Hessian $\nabla^2 g(X) = X^{-1} \otimes X^{-1}$ can be represented by an $\mathbb{R}^{|\mathcal{N}| \times |\mathcal{N}|}$ matrix. We emphasize that we will treat any symmetric matrix as its vectorization of the upper diagonal elements, for example, we will denote $\mathrm{vec}(D)^T \nabla^2 g(X) \mathrm{vec}(D)$ by $D^T \nabla^2 g(X) D$.

For any $X \succ 0$, we define

$$D_J(X) \equiv \arg \min_{\substack{D: D_{ij}=0 \\ \forall (i,j) \notin J}} \nabla g(X)^T D + \frac{1}{2} D^T \nabla^2 g(X) D + \lambda \|X + D\|_1, \tag{14}$$

where $J \subseteq \mathcal{N}$ is any index set, and in particular $D_{\mathcal{N}}(X)$ takes the minimum over all variables.

We use $X_1, X_2, \ldots$ to denote the sequence of matrices generated by our algorithm, where each X_{t+1} is updated from X_t by

$$X_{t+1} = X_t + \alpha_t D_{J_t}(X_t),$$

where J_t is the index set selected at the kth iteration, and α_t is the step size which is the maximum value among $\{1, \beta, \beta^2, \ldots\}$ which satisfies

$$f(X_t + \alpha D_t) < f(X_t) + \alpha\sigma\Delta_t, \tag{15}$$

where $0.5 > \sigma > 0$ is a constant and

$$\Delta_t \equiv \Delta_{J_t}(X_t) \equiv \nabla g(X_t)^T D_t + \lambda \|X_t + D_t\|_1 - \lambda \|X_t\|_1.$$

We use $D_t \equiv D_{J_t}(X_t)$ for simplicity.

Following the setting in [17], the index sets $J_1, J_2, \ldots$ need to satisfy

$$\bigcup_{j=0,\ldots,T-1} J_{t+j} \supseteq \mathcal{N} \ \ \forall t = 1, 2, \ldots \tag{16}$$

for some fixed T. Our algorithm satisfies (16) as mentioned in Section 4.1: we set $J_1, J_3, \ldots$ to be the fixed sets, and $J_2, J_4, \ldots$ to be the free sets and $T = 3$ will suffice.

7.2.2 Lemmas

Our first lemma establishes that our iterates are in the set $mI \preceq X \preceq MI$ for some positive constants m and M.

Lemma 3. *The level set $U = \{X \mid f(X) < f(X_0)$ and $X \in S^p_{++}\}$ is contained in the set $\{X \mid mI \preceq X \preceq MI\}$ for positive constants $m, M > 0$.*

Proof. First, we prove that $X \preceq MI$ for all $X \in U$. The fact that $S \succeq 0$ and $X \succ 0$ implies $\mathrm{tr}(SX) \ge 0$ and $\|X\|_1 > 0$. Therefore we have

$$f(X_0) > f(X) \ge -\log\det X + \lambda \|X\|_1 \tag{17}$$

Since $\|X\|_2$ is the largest eigenvalue of X, we have $-\log\det X \ge -p\log(\|X\|_2)$. In addition, $\|X\|_1 \ge \mathrm{tr}(X) \ge \|X\|_2$. We combine these two facts and (17) to arrive at

$$f(X_0) > -p\log(\|X\|_2) + \lambda \|X\|_2.$$

Since $-p\log x + \lambda x$ is unbounded as x increases, there must exist an M that depends on X_0 such that $\|X\|_2 \leq M$.

Next, we prove that $mI \preceq X$ for all $X \in U$. We denote the smallest eigenvalue of X by a and use the upper bound on the other eigenvalues to get:

$$f(X_0) > f(X) > -\log\det X \geq -\log a - (p-1)\log M, \tag{18}$$

which shows that $m = e^{-f(X_0)}M^{-(p-1)}$ is a lower bound for a. □

Lemma 4. *There exists a unique minimizer X^* for* (2).

Proof. According to Lemma 3, the level set is contained in the compact set $S = \{X \mid mI \preceq X \preceq MI\}$, where $\nabla^2 f(X) = X^{-1} \otimes X^{-1}$, $\nabla^2 f(X) \succeq M^{-2}I$. From Weierstrass' Theorem, any continuous function in a compact set attains its minimum. In addition, $f(X)$ is strongly convex in the compact set, so the minimizer X^* is unique. □

Lemma 5. *X^* is the optimal solution of* (2) *if and only if*

$$\operatorname{grad}^S_{ij} f(X^*) = 0 \quad \forall i, j,$$

where the minimum-norm sub-gradient $\operatorname{grad}^S_{ij} f(X)$ is defined by

$$\operatorname{grad}^S_{ij} f(X) = \begin{cases} \nabla_{ij} g(X) + \lambda & \text{if } X_{ij} > 0, \\ \nabla_{ij} g(X) - \lambda & \text{if } X_{ij} < 0, \\ \operatorname{sign}(\nabla_{ij} g(X)) \max(|\nabla_{ij} g(X)| - \lambda, 0) & \text{if } X_{ij} = 0. \end{cases}$$

Proof. The optimality condition for $f(X)$ is that for all $(i,j) \in \mathcal{N}$

$$\nabla_{ij} g(X) \begin{cases} = -\lambda & \text{if } X_{ij} > 0, \\ = \lambda & \text{if } X_{ij} < 0, \\ \in [-\lambda \;\; \lambda] & \text{if } X_{ij} = 0. \end{cases} \tag{19}$$

It is easy to prove that (19) holds if and only if $\operatorname{grad}^S_{ij} f(X) = 0$ for all i, j. Notice that in our case $\nabla g(X) = S - X^{-1}$ therefore

$$\operatorname{grad}^S_{ij} f(X) = \begin{cases} (S - X^{-1})_{ij} + \lambda & \text{if } X_{ij} > 0, \\ (S - X^{-1})_{ij} - \lambda & \text{if } X_{ij} < 0, \\ \operatorname{sign}((S - X^{-1})_{ij}) \max(|(S - X^{-1})_{ij}| - \lambda, 0) & \text{if } X_{ij} = 0. \end{cases}$$

□

Lemma 6. *For any index set $J \subseteq \mathcal{N}$, $D_J(X) = 0$ if and only if $\operatorname{grad}^S_{ij} f(X) = 0$ for all $(i,j) \in J$.*

Proof. $D_J(X) = 0$ if and only if $D = 0$ satisfy the optimality condition of (14). The condition can be written as (19) with $(i,j) \in J$. This is the same as (19) for a subset of indexes. Follow the same argument we can prove that this condition is equivalent to $\operatorname{grad}^S_{ij} f(X) = 0$ for all $(i,j) \in J$. □

Lemma 7. *$\Delta_J(X)$ in the line search condition* (15) *satisfies*

$$\Delta_J(X) = \nabla g(X)^T D_J(X) + \lambda\|X + D_J(X)\|_1 - \lambda\|X\|_1 \leq -D_J(X)^T \nabla^2 g(X) D_J(X)., \tag{20}$$

and consequently,

$$\Delta_J(X) \leq -m\|D_J(X)\|_F^2 \tag{21}$$

Proof. For simplicity, and since there can be no confusion, we drop index J. By definition of D in (14), $\forall \alpha \in [0,1]$:

$$\nabla g(X)^T D + \frac{1}{2} D^T \nabla^2 g(X) D + \lambda \|X + D\|_1 \leq \nabla g(X)^T (\alpha D) + \frac{1}{2} \alpha^2 D^T \nabla^2 g(X) D + \lambda \|X + \alpha D\|_1. \tag{22}$$

Since $\|\cdot\|_1$ is a norm, the following holds for all $\alpha \geq 0$:

$$\lambda \|X + \alpha D\|_1 = \lambda \|\alpha(X + D) + (1 - \alpha) X\|_1 \leq \lambda \alpha \|X + D\|_1 + \lambda (1 - \alpha) \|X\|_1. \tag{23}$$

Combining (22) and (23) yields:

$$\nabla g(X)^T D + \frac{1}{2} D^T \nabla^2 g(X) D + \lambda \|X + D\|_1 \leq \alpha \nabla g(X)^T D + \frac{1}{2} \alpha^2 D^T \nabla^2 f(X) D + \lambda \alpha \|X + D\|_1 + \lambda (1 - \alpha) \|X\|_1.$$

Therefore

$$(1 - \alpha) \nabla g(X)^T D + (1 - \alpha) \lambda \|X + D\|_1 - (1 - \alpha) \lambda \|X\|_1 + \frac{1}{2} (1 - \alpha^2) D^T \nabla^2 g(X) D \leq 0.$$

Divide both sides by $1 - \alpha$ to get:

$$\nabla g(X)^T D + \lambda \|X + D\|_1 - \lambda \|X\|_1 + \frac{1}{2} (1 + \alpha) D^T \nabla^2 g(X) D \leq 0.$$

By setting $\alpha \uparrow 1$, we have

$$\nabla g(X)^T D + \lambda \|X + D\|_1 - \lambda \|X\|_1 \leq -D^T \nabla^2 g(X) D,$$

which proves (20). Combine with Lemma 3 to get (21). □

Lemma 8. *For any convergent subsequence $X_{s_t} \to \bar{X}$,*

$$D_{s_t} \equiv D_{J_{s_t}}(X_{s_t}) \to 0.$$

Proof. The objective value is monotonically decreasing and bounded below, therefore $f(X_{s_t})$ cannot go to negative infinity, so $f(X_{s_t}) - f(X_{s_{t+1}}) \to 0$. From (15), we have $\alpha_{s_t} \Delta_{s_t} \to 0$.

We proceed to prove by contradiction. If D_{s_t} does not converge to 0, then there exist an infinite index set $\mathcal{T} \subseteq \{s_1, s_2, \ldots\}$ and $\delta > 0$ such that $\|D_t\|_F > \delta$ for all $t \in \mathcal{T}$. We will work in this index set $\mathcal{T}$ in what follows.

Let α_t denote the line search step size which satisfies (15), by our line search procedure $\frac{\alpha_t}{\beta}$ will not satisfy (15), so we have:

$$f(X_t + (\frac{\alpha_t}{\beta}) D_t) - f(X_t) \geq \sigma (\frac{\alpha_t}{\beta}) \Delta_t. \tag{24}$$

If $X_t + \frac{\alpha_t}{\beta} D_t$ is not positive definite, then we define $f(X_t + \frac{\alpha_t}{\beta} D_t)$ to be ∞, so (24) still holds. We have

$$\begin{aligned}
\sigma \Delta_t &\leq \frac{g(X_t + (\frac{\alpha_t}{\beta}) D_t) - g(X_t) + \lambda \|X_t + \frac{\alpha_t}{\beta} D_t\|_1 - \lambda \|X_t\|_1}{\frac{\alpha_t}{\beta}} \\
&\leq \frac{g(X_t + (\frac{\alpha_t}{\beta}) D_t) - g(X_t) + (\frac{\alpha_t}{\beta}) \lambda \|X_t + D_t\|_1 + (1 - \frac{\alpha_t}{\beta}) \lambda \|X_t\|_1 - \lambda \|X_t\|_1}{\frac{\alpha_t}{\beta}} \quad \text{(by (23))} \\
&= \frac{g(X_t + (\frac{\alpha_t}{\beta}) D_t) - g(X_t)}{\frac{\alpha_t}{\beta}} + \lambda \|X_t + D_t\|_1 - \lambda \|X_t\|_1, \forall t \in \mathcal{T}.
\end{aligned}$$

By the definition of Δ_t we can replace the last two terms and get

$$\begin{aligned}
\sigma \Delta_t &\leq \frac{g(X_t + (\frac{\alpha_t}{\beta}) D_t) - g(X_t)}{\frac{\alpha_t}{\beta}} + \Delta_t - \nabla g(X_t)^T D_t, \\
(1 - \sigma)(-\Delta_t) &\leq \frac{g(X_t + (\frac{\alpha_t}{\beta}) D_t) - g(X_t)}{\frac{\alpha_t}{\beta}} - \nabla g(X_t)^T D_t
\end{aligned}$$

By (21) in Lemma 7,

$$(1-\sigma)m\|D_t\|_F^2 \leq \frac{g(X_t+(\frac{\alpha_t}{\beta})D_t)-g(X_t)}{\frac{\alpha_t}{\beta}} - \nabla g(X_t)^T D_t$$
$$(1-\sigma)m\|D_t\|_F \leq \frac{g(X_t+(\frac{\alpha_t}{\beta})\|D_t\|_F\frac{D_t}{\|D_t\|_F})-g(X_t)}{\|D_t\|_F\frac{\alpha_t}{\beta}} - \nabla g(X_t)^T\frac{D_t}{\|D_t\|_F}.$$

Set $\hat{\alpha}_t = \frac{\alpha_t}{\beta}\|D_t\|_F$, and since $\|D_t\|_F > \delta$ for all $t \in \mathcal{T}$ we have

$$(1-\sigma)m\delta \leq \frac{g(X_t+\hat{\alpha}_t\frac{D_t}{\|D_t\|_F})-g(X_t)}{\hat{\alpha}_t} - \frac{\nabla g(X_t)^T D_t}{\|D_t\|_F}. \tag{25}$$

By (21),

$$-\alpha_t\Delta_t \geq \alpha_t m\|D_t\|_F^2 \geq m\alpha_t\|D_t\|_F\delta,$$

and $\{\alpha_t\Delta_t\}_t \to 0$, so $\{\alpha_t\|D_t\|_F\}_t \to 0$, so $\{\hat{\alpha}^k\}_t \to 0$. Since $\frac{D_t}{\|D_t\|_F}$ is in the compact 1-norm ball, there exists a subset $\bar{\mathcal{T}} \subset \mathcal{T}$ such that $\{\frac{D_t}{\|D_t\|_F}\}_{\bar{\mathcal{T}}} \to \bar{D}$, so

$$(1-\sigma)m\delta \leq \frac{g(X_t+\hat{\alpha}_t\bar{D})-g(X_t)}{\hat{\alpha}_t} - \nabla g(X_t)^T\bar{D}. \tag{26}$$

Our algorithm guarantees that X_t is positive definite. Also $X_t + \hat{\alpha}_t\bar{D}$ is positive definite when $\hat{\alpha}_t \to 0$. So taking limit of (26) as $t \in \bar{\mathcal{T}}$ and $k \to \infty$ on (25), we have

$$(1-\sigma)m\delta \leq \nabla g(\bar{X})^T\bar{D} - \nabla g(\bar{X})^T\bar{D} = 0,$$

a contradiction, finishing the proof. □

Lemma 9. *For any $X \succ 0$ and symmetric D, there exists an $\bar{\alpha} > 0$ such that for all $\alpha < \bar{\alpha}$, (1) $X + \alpha D \succ 0$ and (2) $X + \alpha D$ satisfies the line search condition* (15).

Proof. First, when $\alpha < \sigma_n(X)/\|D\|_2$ ($\sigma_n(X)$ stands for the smallest eigen-value of X), $\|\alpha D\|_2 < \sigma_n(X)$, so $X + \alpha D \succ 0$.

Second,

$$\begin{aligned} f(X+\alpha D) - f(X) &= g(X+\alpha D) - g(X) + \lambda\|X+\alpha D\|_1 - \lambda\|X\|_1 \\ &\leq g(X+\alpha D) - g(X) + \alpha\lambda(\|X+D\|_1 - \|X\|_1) \text{ by (23)} \\ &= \alpha\Delta + o(\alpha). \end{aligned}$$

It follows that for a fixed $0 < \sigma < 1$, when α is sufficiently small, the line search condition must hold. □

7.2.3 Proof of Lemma 1

Since the *fixed* set S_{fixed} is defined by

$$S_{fixed} := \{(i,j) \mid |\nabla_{ij}g(X_t)| < \lambda - \epsilon \text{ and } (X_t)_{ij} = 0\},$$

so $\text{grad}_{ij}^S f(X_t) = 0$ for all $(i,j) \in S_{fixed}$. From Lemma 6, this implies $D_{S_{fixed}} = 0$, therefore the solution of the following optimization problem is $\Delta = 0$:

$$\arg\min_{\Delta} f(X_t+\Delta) \text{ such that } \Delta_{ij} = 0 \ \ \forall (i,j) \in S_{free}.$$

7.2.4 Main proof

Theorem 3. *Our algorithm* QUIC *converges to a unique global optimum*.

Proof. Assume a subsequence $\{X_t\}_{\mathcal{T}}$ converges to $\bar{X}$. Since the choice of the index set J_t selected at each step is finite, we can further assume that $J_t = \bar{J}_0$ for all $t \in \mathcal{T}$. From Lemma 8, $D_{\bar{J}_0}(X_t) \to 0$. By the continuity of $\nabla f(X)$ and $\nabla^2 f(X)$, it is easy to show $D_{\bar{J}_0}(X_t) \to D_{\bar{J}_0}(\bar{X})$. Therefore $D_{\bar{J}_0}(\bar{X}) = 0$.

Furthermore, $\{D_{\bar{J}_0}(X_t)\}_t \to 0$ and $\|X_t - X_{t+1}\|_F \leq \|D_{\bar{J}_0}(X_t)\|_F$, so $\{X_{t+1}\}_t$ also converges to $\bar{X}$. By further subsetting of $\mathcal{T}$ we can assume that $J_{t+1} = \bar{J}_1$ for all $t \in \mathcal{T}$. By the same argument we can prove $\{D_{J_{t+1}}(X_t)\}_t \to 0$, so $D_{\bar{J}_1}(\bar{X}) = 0$. Similarly, we can show that $D_{\bar{J}_i}(\bar{X}) = 0 \; \forall i = 0, \ldots, T-1$ can be assumed for an appropriate subset of $\mathcal{T}$. According to Lemma 6 and assumption (16), $\bar{X}$ is a stationary point:

$$\mathrm{grad}^S_{ij} f(\bar{X}) = 0 \; \forall i, j.$$

Moreover, by Lemma 4, there exists a unique optimal point, so the sequence $\{X_t\}$ generated by our algorithm must converge to the global optimum. □

7.3 Quadratic Convergence Rate

7.3.1 Existing results for Newton method on Bounded constrain

The convergence rate of Newton method on bounded constrained minimization has been studied in [10] and [6]. Here we briefly mention their results.

Assume we want to solve a constrained minimization problem

$$\min_{x \in \Omega} F(x),$$

where Ω is a nonempty subset of R^n and $F : R^n \to R$ has a second derivative $\nabla^2 F(x)$. Then beginning from x^0, a natural extension of Newton method is to compute x^{k+1} by

$$x^{k+1} = \arg\min_{x \in \Omega} \nabla F(x^k)^T (x - x^k) + \frac{1}{2}(x - x^k)^T \nabla^2 F(x^k)(x - x^k). \tag{27}$$

For simplicity, we assume F is strictly convex and has a unique minimizer x^* in Ω. Then the following theorem holds

Theorem 4. *Assume F is strictly convex, has a unique minimizer x_* in Ω, and $\nabla^2 F(x)$ is Lipschitz continuous, then for all x_0 sufficiently close to x_*, the sequence $\{x_k\}$ generated by* (27) *converges quadratically to x_*.*

This theorem is proved in [6].

7.3.2 Proof for the quadratic convergence of QUIC

Again we consider the composite objectives as (13), and $g(X)$ has Lipschitz continuous second order derivatives. Assume X^* is the optimal solution, then we can divide the indexes into

$$P = \{(i,j) \mid \nabla_{ij} g(X^*) = -\lambda\}, \;\; N = \{(i,j) \mid \nabla_{ij} g(X^*) = \lambda\}, \;\; Z = \{(i,j) \mid -\lambda < \nabla_{ij} g(X^*) < \lambda\}. \tag{28}$$

Notice that $X^*_{ij} \geq 0$ for all $(i,j) \in P$, $X^*_{ij} \leq 0$ for all $(i,j) \in N$ and $X^*_{ij} = 0$ for all $(i,j) \in Z$.

Lemma 10. *If the second order derivative of $g(\cdot)$ is Lipschitz continuous, then when X_t is close enough to X^*, the line search condition* (15) *will be satisfied with step size $\alpha = 1$.*

Proof. To simplify the notation, here we denote X_t by X, D_t by D, and Δ_t by Δ. We bound the decrease in objective function value by the following argument. First, define

$$\tilde{g}(t) = g(X + tD),$$

so $\tilde{g}''(t) = D^T \nabla^2 g(X + tD) D$. From the Lipschitz continuity of $\nabla^2 g(\cdot)$, we have

$$\|\nabla^2 g(X + tD) - \nabla^2 g(X)\| \leq tL\|D\|,$$

where L is the Lipschitz constant. By definition

$$|\tilde{g}''(t) - \tilde{g}''(0)| = |D^T(\nabla^2 g(X + tD) - \nabla^2 g(X))D| \leq tL\|D\|^3.$$

Therefore we can upper bound $\tilde{g}''(t)$ by

$$\tilde{g}''(t) \leq \tilde{g}''(0) + tL\|D\|^3 = D^T\nabla^2 g(X)D + tL\|D\|^3.$$

Integrate both sides to get

$$\tilde{g}'(t) \leq \tilde{g}'(0) + tD^T\nabla^2 g(X)D + \frac{1}{2}t^2L\|D\|^3 = \nabla g(X)^T D + tD^T\nabla^2 g(X)D + \frac{1}{2}t^2L\|D\|^3.$$

Integrating both sides again, we have

$$\tilde{g}(t) \leq \tilde{g}(0) + t\nabla g(X)^T D + \frac{1}{2}t^2 D^T\nabla^2 g(X)D + \frac{1}{6}t^3L\|D\|^3.$$

Taking $t = 1$ the inequality becomes

$$g(X+D) = \tilde{g}(1) \leq g(X) + \nabla g(X)^T D + \frac{1}{2}D^T\nabla^2 g(X)D + \frac{1}{6}L\|D\|^3$$

$$g(X+D) + \lambda\|X+D\|_1 \leq g(X) + \lambda\|X\|_1 + (\nabla g(X)^T D + \lambda\|X+D\|_1 - \lambda\|X\|_1) + \frac{1}{2}D^T\nabla^2 g(X)D + \frac{1}{6}L\|D\|^3,$$

so

$$\begin{aligned} f(X+D) &\leq f(X) + \Delta + \frac{1}{2}D^T\nabla^2 g(X)D + \frac{1}{6}L\|D\|^3 \\ &\leq f(X) + \frac{1}{2}\Delta - \frac{1}{6}\frac{L}{m}\|D\|\Delta \ \text{ (by (20) and (21) in Lemma 7)} \\ &= f(X) + (\frac{1}{2} - \frac{1}{6}\frac{L}{m}\|D\|)\Delta. \end{aligned}$$

And from Lemma 8 we have $D^k \to 0$, therefore when k is large enough, $(\frac{1}{2} - \frac{1}{6}\frac{L}{m}\|D^k\|)$ will be larger than σ $(0 < \sigma < 0.5)$, so the line search condition holds with step size 1. □

Lemma 11. *Assume that the sequence $\{X_t\}$ converges to the global optimum X^*. There exists a $\bar{t} > 0$ such that*

$$(X_t)_{ij} \begin{cases} \geq 0 & \text{if } (i,j) \in P \\ \leq 0 & \text{if } (i,j) \in N \\ = 0 & \text{if } (i,j) \in Z \end{cases} \tag{29}$$

for all $t > \bar{t}$.

Proof. We prove the case for $(i,j) \in P$ by contradiction, the other two cases can be handled similarly. Assume that there exists an infinite subsequence $\{X_{s_t}\}$ such that $(X_{s_t})_{ij} < 0$. We consider the update from X_{s_t-1} to X_{s_t}. From Lemma 10, we can assume that s_t is large enough so that the step size equals 1, therefore $X_{s_t} = X_{s_t-1} + d_{s_t}$. Note that D_{s_t} is the optimal solution of

$$\min_D \nabla g(X_{s_t-1})^T D + \frac{1}{2}D^T\nabla^2 g(X_{s_t-1})D + \|X+D\|_1 - \|X\|_1. \tag{30}$$

Since $(X_{s_t})_{ij} = (X_{s_t-1})_{ij} + (D_{s_t})_{ij} < 0$, from the optimality condition of (30) we have

$$(\nabla g(X_{s_t-1}) + \nabla^2 g(X_{s_t-1})(D_{s_t}))_{ij} = \lambda. \tag{31}$$

Since D_{s_t} converges to 0, (31) implies that $\{\nabla_{ij} g(X_{s_t-1})\}$ will converge to λ. However, by the definition of P, $\nabla_{ij} g(X^*) = -\lambda$, and by the continuity of ∇g we get that $\{\nabla_{ij} g(X_t)\}$ converges to $\nabla_{ij} g(X^*) = -\lambda$, a contradiction finishing the proof for the case with $(i,j) \in P$ in (29). □

Lemma 12. *Assume $X_t \to X^*$. There exists a $\bar{t} > 0$ such that variables in P or N will not be selected as fixed set (denoted by S_{fixed}) after $t > \bar{t}$. That is,*

$$S_{fixed} \subset Z = \mathcal{N} \setminus (P \cup N).$$

Proof. Since X_t converges to X^* and $\nabla g(\cdot)$ is continuous, $\nabla g(X_t)$ will converge to $\nabla g(X^*)$. Therefore, $\nabla_{ij} g(X_t)$ converges to $-\lambda$ if $(i,j) \in P$ and to λ if $(i,j) \in N$. Since we select fixed set by testing whether $(X_t)_{ij} = 0$ and

$$-\lambda + \epsilon < \nabla_{ij} g(X_t) < \lambda - \epsilon,$$

when k is large enough $|\nabla_{ij} g(X_t) - \nabla_{ij} g(X^*)|$ will be smaller than ϵ, then all variables in P or N will not be selected in the fixed set. □

Theorem 5. *$\{X_t\}$ generated by our algorithm* QUIC *converges asymptotic quadratically to X^* when t is large enough.*

Proof. First, if we the index sets P, N and Z (related to the optimal solution) are given, solving (2) is the same as solving the following constrained minimization problem.

$$\begin{aligned} \min_X \quad & -\log\det(X) + \mathrm{tr}(SX) + \sum_{(i,j)\in P} \lambda X_{ij} - \sum_{(i,j)\in N} \lambda X_{ij} \\ \text{s.t.} \quad & X_{ij} \geq 0 \ \ \forall (i,j) \in P, \\ & X_{ij} \leq 0 \ \ \forall (i,j) \in N, \\ & X_{ij} = 0 \ \ \forall (i,j) \in Z. \end{aligned} \tag{32}$$

Next we claim that when k is large enough, our algorithm is equivalent to applying the Newton method in Section 7.3.1 to minimize (32). Since the objective function values of (32) and (2) are the same if we restrict variables to follow the sign patterns in (32), to prove the equivalence it suffices to show:

1. The sign of the optimal solution for the original sub-problem (5) will always be the same as (32) after a finite number of iterations. This is the result of Lemma 11.
2. The fixed set selection does not affect the Newton sub-problem. This can be proved by Lemma 12 because at each iteration the fixed set $S_{fixed} \subset Z$, and Z is the set which always satisfies $(D_t)_Z = 0$ after t large enough. So we will never fix the wrong variables (choose variables in P or N in the fixed set) after t is large enough.

Moreover, Lemma 10 shows the step size will always be 1 when t large enough. Therefore our algorithm is equivalent to the Newton method in Section 7.3.1, which converges quadratically to the optimal solution of (32). Since the revised problem (32) and our original problem (2) has the same minimum, our algorithm converges quadratically to the optimum of (2) when the iteration t is large enough. □

7.4 Size of free sets in experiments

In Figure 2, we plot the size of the free set versus iterations for Hereditarybc dataset. Starting from a total of $1869^2 = 3,493,161$ variables, the size of the free set progressively drops, in fact to less than $120,000$ in the very first iteration. We can see the super-linear convergence of QUIC even more clearly when we plot it against the number of iterations.

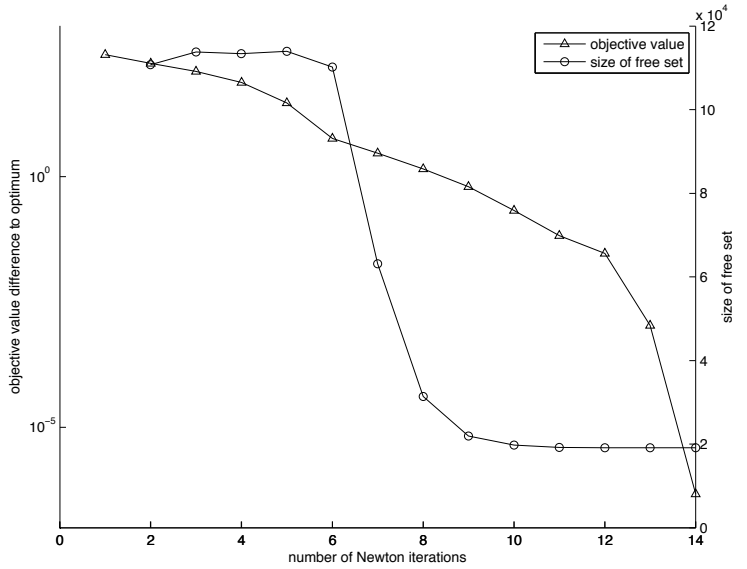

Figure 2: Size of free sets and objective value versus iterations (Hereditarybc dataset). There are total $3,493,161$ variables, but the size of free set reduce to less than $120,000$ in one iteration, and become about $20,000$ at the end.