A First Order Free Lunch for SQRT-Lasso∗

Xingguo Li, Jarvis Haupt, Raman Arora, Han Liu, Mingyi Hong, and Tuo Zhao

Abstract

Many statistical machine learning techniques sacrifice convenient computational structures to gain estimation robustness and modeling flexibility. In this paper, we study this fundamental tradeoff through a SQRT-Lasso problem for sparse linear regression and sparse precision matrix estimation in high dimensions. We explain how novel optimization techniques help address these computational challenges. Particularly, we propose a pathwise iterative smoothing shrinkage thresholding algorithm for solving the SQRT-Lasso optimization problem. We further provide a novel model-based perspective for analyzing the smoothing optimization framework, which allows us to establish a nearly linear convergence (R-linear convergence) guarantee for our proposed algorithm. This implies that solving the SQRT-Lasso optimization is almost as easy as solving the Lasso optimization. Moreover, we show that our proposed algorithm can also be applied to sparse precision matrix estimation, and enjoys good computational properties. Numerical experiments are provided to support our theory.

1 Introduction

Given a design matrix \( X \in \mathbb{R}^{n \times d} \) and a response vector \( y \in \mathbb{R}^n \), we consider a linear model

\[
    y = X \theta^* + \epsilon,
\]

where \( \theta^* \in \mathbb{R}^d \) is an unknown coefficient vector, and \( \epsilon \in \mathbb{R}^n \) is a random noise vector with i.i.d. sub-Gaussian entries, \( \mathbb{E}[\epsilon_i] = 0 \) and \( \mathbb{E}[\epsilon_i^2] = \sigma^2 \) for all \( i = 1, \ldots, n \). We are interested in estimating \( \theta^* \) in high dimensions where \( n/d \to 0 \). A popular assumption in high dimensions is that only a small subset of variables are relevant in modeling, i.e., many entries of \( \theta^* \) are zero. To get such a sparse estimator, Tibshirani (1996) proposed Lasso, which solves

\[
    \hat{\theta} = \arg\min_{\theta} \frac{1}{n} \| y - X \theta \|_2^2 + \lambda_{\text{Lasso}} \| \theta \|_1,
\]

where \( \lambda_{\text{Lasso}} \) is the regularization parameter and encourages the solution sparsity. The statistical properties of Lasso have been established in Zhang and Huang (2008); Zhang (2009); Bickel et al.

∗Xingguo Li and Jarvis Haupt are affiliated with Department of Electrical and Computer Engineering at University of Minnesota, Minneapolis, MN, 55455, USA; Raman Arora and Tuo Zhao is affiliated with Department of Computer Science at Johns Hopkins University Baltimore, MD, 21210, USA; Han Liu is affiliated with Department of Operations Research and Financial Engineering at Princeton University, Princeton, NJ 08544, USA; Mingyi Hong is affiliated with Department of Industrial and Manufacturing Systems Engineering at Iowa State University. Emails: lixx1661@umn.edu, jdhaupt@umn.edu, arora@cs.jhu.edu, hanliu@princeton.edu, mingyi@iastate.edu, tour@cs.jhu.edu
In particular, given \( \lambda_{\text{Lasso}} \approx \sigma \sqrt{\log d/n} \), the Lasso estimator in (1.1) attains the minimax optimal rates of convergence in parameter estimation, \[ \| \hat{\theta} - \theta^* \|_2 = O_P \left( \sigma \sqrt{s^* \log d/n} \right), \] (1.2)
where \( s^* \) denotes the number of nonzero entries in \( \theta^* \) (Ye and Zhang, 2010; Raskutti et al., 2011).

Despite these favorable properties, the Lasso approach has a significant drawback: The selected regularization parameter \( \lambda_{\text{Lasso}} \) linearly scales with the unknown quantity \( \sigma \). Therefore, we need to carefully tune \( \lambda_{\text{Lasso}} \) over a wide range of potential values in order to get a good finite-sample performance. To overcome this drawback, Belloni et al. (2011) proposed SQRT-Lasso, which solves
\[
\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{\sqrt{n}} \| y - X\theta \|_2 + \lambda_{\text{SQRT}} \| \theta \|_1. \tag{1.3}
\]
They further show that SQRT-Lasso require no prior knowledge of \( \sigma \). Choosing \( \lambda_{\text{SQRT}} \approx \sqrt{\log d/n} \), the SQRT-Lasso estimator in (1.3) attains the same optimal statistical rate of convergence in parameter estimation as (1.2). This means that the regularization selection for SQRT-Lasso does not scale with \( \sigma \). We can easily specify a smaller range of potential values for tuning \( \lambda_{\text{SQRT}} \) than Lasso.

Besides estimating \( \theta^* \), SQRT-Lasso can also estimate \( \sigma \), which further makes it applicable to sparse precision matrix estimation; this is not the case with Lasso. Specifically, given \( n \) observations i.i.d. sampled from a \( d \)-variate normal distribution with mean \( 0 \) and a sparse precision matrix \( \Theta^* \), Liu and Wang (2012) proposed an estimator based on SQRT-Lasso (See more details in §4), and showed that it attains the minimax optimal statistical rate of convergence in parameter estimation
\[
\| \hat{\Theta} - \Theta^* \|_2 = O_P \left( \| \Theta^* \|_2 \cdot s^* \sqrt{\log d/n} \right),
\] where \( \| \Theta^* \|_2 \) denotes the spectral norm of \( \Theta^* \) (i.e., the largest singular value of \( \Theta^* \)), and \( s^* \) denotes the maximum number of nonzero entries in each column of \( \Theta^* \) (i.e. \( \max_{\ell} 1(\Theta_{j\ell} \neq 0) \leq s^* \)).

Though SQRT-Lasso simplifies its tuning efforts and achieves the optimal statistical properties for both sparse linear regression and sparse precision matrix estimation in high dimensions, the optimization problem in (1.3) is computationally more challenging than (1.1) for Lasso, because the \( \ell_2 \) loss in SQRT-Lasso does not have the same nice computational structures as the least square loss in Lasso. For example, the \( \ell_2 \) loss can be nondifferentiable, and does not have a Lipschitz continuous gradient. Belloni et al. (2011) converted (1.3) to a second order cone optimization problem, and further solved it by an interior point method; Li et al. (2015) then solved (1.3) by an ADMM algorithm. Neither of them, however, can scale to large problems. In contrast, Xiao and Zhang (2013) proposed an efficient pathwise iterative shrinkage thresholding algorithm (PISTA) for solving (1.1), which attains a linear convergence to the unique sparse global optimum with high probability.

To address this computational challenge, we propose a pathwise iterative smoothing shrinkage thresholding algorithm (PIS\(^2\)TA) to solve (1.3). Specifically, we first apply the conjugate dual

---

\(^1\)The notation \( O_P(\cdot) \) is defined in Line 84 on Page 3.
smoothing approach to the nonsmooth $\ell_2$ loss (Nesterov, 2005; Beck and Teboulle, 2012), and obtain
a smooth surrogate denoted by $\|y - X\theta\|_\mu$, where $\mu > 0$ is a smoothing parameter (See more details
in §2). We then apply PISTA to solve the partially smoothed optimization problem as follows:

$$
\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{\sqrt{n}} \|y - X\theta\|_\mu + \lambda_{\text{SQRT}} \|\theta\|_1.
$$

(1.4)

Existing computational theory guarantees that our proposed PIS$^2$TA algorithm attains a sublinear convergence to the global optimum in term of the objective value (Nesterov, 2005). However, our numerical experiments show that PIS$^2$TA achieves far better experimental computational performance (better than sublinear convergence) for solving SQRT-Lasso, and is significantly more efficient than other competing algorithms, and nearly as efficient as Xiao and Zhang (2013) for solving Lasso.

This is because the existing computational analyses of the conjugate dual smoothing approach do not take certain specific modeling structures into consideration. For example: (I) The $\ell_2$ loss is only nonsmooth when all residuals are equal to zero (significantly overfitted). But this is very unlikely to happen because we are solving (1.3) with a sufficiently large regularization; (II) Although the smoothed $\ell_2$ loss is not strongly convex, if we restrict the solution to a sparse domain, the smoothed $\ell_2$ loss can behave like strongly convex functions over a neighborhood of $\theta^*$. Motivated by these observations, we establish a new computational theory for PIS$^2$TA, which exploits the above modeling structures. Particularly, we show that PIS$^2$TA achieves a nearly linear convergence (R-linear convergence) to the unique sparse global optimum for solving (1.3) with high probability, and also gives us a well fitted model. There are two implications: (I) We can solve the SQRT-Lasso optimization as nearly efficiently as solving the Lasso optimization; (II) We pay almost no price in optimization accuracy when using the smoothing approach for solving the SQRT-Lasso optimization, because (1.4) and (1.3) share the same unique sparse global optimum with high probability.

As an extension of our theory for the SQRT-Lasso optimization, we further analyze the computational properties of our proposed PIS$^2$TA algorithm for sparse precision matrix estimation in high dimensions. We show that PIS$^2$TA also achieves an R-linear convergence to the unique sparse global optimum with high probability. We provide numerical experiments on simulated and real data to support our theory. All proofs of our analysis are deferred to the supplementary material.

**Notations:** Given a vector $v = (v_1, \ldots, v_d)^\top \in \mathbb{R}^d$, we define vector norms: $\|v\|_1 = \sum_j |v_j|$, $\|v\|_2^2 = \sum_j v_j^2$, and $\|v\|_\infty = \max_j |v_j|$. We denote the number of nonzero entries in $v$ as $\|v\|_0 = \sum_j 1(v_j \neq 0)$. We denote $v_{\setminus j} = (v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_d)^\top \in \mathbb{R}^{d-1}$ as the subvector of $v$ with the $j$-th entry removed. Let $A \subseteq \{1, \ldots, d\}$ be an index set. We use $\overline{A}$ to denote the complementary set to $A$, i.e., $\overline{A} = \{j \mid j \in \{1, \ldots, d\}, j \notin A\}$. We use $v_A$ to denote a subvector of $v$ by extracting all entries of $v$ with indices in $A$. Given a matrix $A \in \mathbb{R}^{d \times d}$, we use $A_{\setminus j} = (A_{1j}, \ldots, A_{dj})^\top$ to denote the $j$-th column of $A$, and $A_{\setminus k} = (A_{k1}, \ldots, A_{kd})^\top$ to denote the $k$-th row of $A$. Let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ be the largest and smallest eigenvalues of $A$. We define $\|A\|_F^2 = \sum_j \|A_{\setminus j}\|_2^2$ and $\|A\|_2 = \sqrt{\lambda_{\max}(A^\top A)}$. We denote $A_{\setminus i,j}$ as the submatrix of $A$ with the $i$-th row and the $j$-th column removed. We denote $A_{\setminus i,j}$ as the $i$-th row of $A$ with its $j$-th entry removed. Let $A \subseteq \{1, \ldots, d\}$
be an index set. We use $A_{AA}$ to denote a submatrix of $A$ by extracting all entries of $A$ with both row and column indices in $A$. We denote $A > 0$ if $A$ is a positive-definite matrix. Given two real sequences $\{A_n\}, \{a_n\}$, $A_n = O(a_n)$ (or $A_n = \Omega(a_n)$) if and only if $\exists M \in \mathbb{R}^+$ and $N \in \mathbb{N}$ such that $|A_n| \leq M|a_n|$ (or $|A_n| \geq M|a_n|$) for all $n \geq N$. $A_n \times a_n$ if $A_n = O(a_n)$ and $A_n = \Omega(a_n)$ simultaneously. $A_n = O_P(a_n)$ if $\forall \delta \in (0, 1), \exists M \in \mathbb{R}^+$ and $N_\delta \in \mathbb{N}$ such that $\mathbb{P}[|A_n| > M|a_n|] < \delta$ for all $n \geq N_\delta$. $A_n = o(a_n)$ if $\forall \delta > 0, \exists N_\delta \in \mathbb{N}$ such that $|A_n| \leq \delta|a_n|$ for all $n \geq N_\delta$, i.e., $\lim_{n \to \infty} A_n/a_n = 0$. Given a vector $x \in \mathbb{R}^d$ and a real value $\lambda > 0$, we denote the soft thresholding operator $S_\lambda(x) = [\text{sign}(x_j) \max\{|x_j| - \lambda, 0\}]_{j=1}^d$.

2 Algorithm

Our proposed algorithm consists of three components: (I) Conjugate Dual Smoothing, (II) Iterative Shrinkage Thresholding Algorithm (ISTA), and (III) Pathwise Optimization.

(I) The **Conjugate Dual Smoothing** approach is adopted to obtain a smooth surrogate of $\ell_2$ loss (Nesterov, 2005; Beck and Teboulle, 2012). We denote the smoothed $\ell_2$ loss function as

$$
\|y - X\theta\|_\mu = \max_{\|z\|_2 \leq 1} z^T(y - X\theta) - \frac{\mu}{2}\|z\|_2^2.
$$

(2.1)

The optimization problem in (2.1) admits a closed form solution:

$$
\mathcal{L}_\mu(\theta) = \frac{1}{\sqrt{n}}\|y - X\theta\|_\mu = \begin{cases} 
\frac{1}{2\mu\sqrt{n}}\|y - X\theta\|_2^2, & \text{if } \|y - X\theta\|_2 < \mu \\
\frac{1}{\sqrt{n}}\|y - X\theta\|_2 - \frac{\mu}{2}, & \text{o.w.}
\end{cases}
$$

We present several two-dimensional examples of the smoothed $\ell_2$ norm using different $\mu$’s in Figure 1. A larger $\mu$ introduces a larger approximation error, but makes the approximation smoother. We then consider the following partially smoothed optimization problem,

$$
\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \mathcal{F}_{\mu, \lambda}(\theta), \quad \text{where } \mathcal{F}_{\mu, \lambda}(\theta) = \mathcal{L}_\mu(\theta) + \lambda\|\theta\|_1.
$$

(2.2)

![Figure 1: Examples of $\|x\|_2$ (a) $\mu = 0$ and $\|x\|_\mu$ with $\mu = 0.1$, 0.5, and 1 respectively for $x \in \mathbb{R}^2$.](image)

(II) The **ISTA** Algorithm is applied to solve (2.2) (Nesterov, 2013). Particularly, given $\theta^{(t)}$ at $t$-th iteration, we consider the quadratic approximation of $\mathcal{F}_{\mu, \lambda}(\theta)$ at $\theta = \theta^{(t)}$,

$$
\mathcal{Q}_{\mu, \lambda}(\theta, \theta^{(t)}) = \mathcal{L}_\mu(\theta^{(t)}) + \nabla \mathcal{L}_\mu(\theta^{(t)})^T(\theta - \theta^{(t)}) + \frac{L^{(t)}}{2}\|\theta - \theta^{(t)}\|_2^2 + \lambda\|\theta\|_1,
$$

(2.3)
where \( L^{(t)} \) is a step size parameter determined by the backtracking line search. We then take
\[
\theta^{(t+1)} = \arg \min_{\theta} Q_{\mu,\lambda}(\theta, \theta^{(t)}) = S_{\lambda/L^{(t)}}(\theta^{(t)}) - \nabla L_{\mu}(\theta^{(t)})/L^{(t)},
\]
For simplicity, we denote \( \theta^{(t+1)} = T_{L^{(t+1)}}(\theta^{(t)}) \). Given a pre-specified precision \( \varepsilon \), we terminate
the iterations when the approximate KKT condition holds:
\[
\omega_{\lambda}(\theta^{(t)}) = \min_{g \in \partial \theta^{(t)}} \| \nabla L_{\mu}(\theta^{(t)}) + \lambda g \|_{\infty} \leq \varepsilon.
\]

(III) The Pathwise Optimization is essentially a multistage optimization scheme for boosting computational performance. We solve (2.2) using a geometrically decreasing sequence of regularization parameters \( \lambda_1 > \ldots > \lambda_N \), where \( \lambda_N = \lambda_{\text{SQRT}} \). This yields a sequence of output solutions \( \hat{\theta}_{[1]}, \ldots, \hat{\theta}_{[N]} \) from sparse to dense.

Particularly, at the \( K \)-th optimization stage, we choose \( \hat{\theta}_{[K-1]} \) (the output solution of the \( (K-1) \)-th stage) as the initial solution, i.e., \( \hat{\theta}_{[K]}^{(0)} = \hat{\theta}_{[K-1]} \), and solve (2.2) with \( \lambda = \lambda_K \) using the ISTA algorithm. This is also referred as the warm start initialization in existing literature. We summarize our approach in Algorithm 1.

3 Computational and Statistical Analysis

We first define the locally restricted strong convexity and smoothness.

**Definition 3.1.** Given a constant \( r \in \mathbb{R}^+ \), let \( B_r = \{ \theta \in \mathbb{R}^d : \| \theta - \theta^* \|_2^2 \leq r \} \). For any \( v, w \in B_r \), which satisfies \( \| v - w \|_0 \leq s \), \( L_{\mu} \) is locally restricted strongly convex (LRSC) and smooth (LRSS) on \( B_r \) at sparsity level \( s \) if there exist universal constants \( \rho_s^-, \rho_s^+ \in (0, \infty) \) such that
\[
\frac{\rho_s^-}{2} \| v - w \|_2^2 \leq L_{\mu}(v) - L_{\mu}(w) - \nabla L_{\mu}(w)^{\top}(v - w) \leq \frac{\rho_s^+}{2} \| v - w \|_2^2. \tag{3.1}
\]
We define the locally restricted condition number at sparsity level \( s \) as \( \kappa_s = \rho_s^+/\rho_s^- \).

The LRSC and LRSS properties are locally constrained variants of restricted strong convexity and smoothness (Agarwal et al., 2010; Xiao and Zhang, 2013) with respect to a neighborhood of the true model parameter \( \theta^* \), which are keys to establishing the strong convergence guarantees of our proposed algorithm in high dimensions.

Next, we introduce two key assumptions for establishing our computational theory.

**Assumption 3.2.** The sequence of the regularization parameters satisfies \( \lambda_N \geq 6 \| \nabla L_{\mu}(\theta^*) \|_{\infty} \).

Assumption 3.2 requires that \( \lambda_N \) is sufficiently large such that the irrelevant variables can be eliminated (Bickel et al., 2009; Negahban et al., 2012).

**Assumption 3.3.** \( L_{\mu} \) satisfies LRSC and LRSS properties on \( B_r \), where \( r \geq s^* \left( 8\lambda_{N_1}/\rho_{s^{*}+\tilde{s}}^- \right)^2 \) for some \( N_1 < N \), \( N_1 \in \mathbb{Z}^+ \), and \( \lambda_{N_1} > \lambda_N \). Specifically, (3.1) holds with \( \rho_{s^{*}+\tilde{s}}^-, \rho_{s^{*}+\tilde{s}}^+ \in (0, \infty) \), where \( \tilde{s} = C_1 s^* > (196\kappa_{s^{*}+\tilde{s}}^2 + 144\kappa_{s^{*}+\tilde{s}}^2 s^*)^{-1} \), \( C_1 \in \mathbb{R}^+ \) is a constant and \( \kappa_{s^{*}+\tilde{s}} = \rho_{s^{*}+\tilde{s}}^+ / \rho_{s^{*}+\tilde{s}}^- \).

Assumption 3.3 guarantees that \( L_{\mu} \) satisfies LRSC and LRSS properties as long as the estimation error satisfies \( \| \theta - \theta^* \|_2^2 \leq r \) and the number of irrelevant coordinates of solutions is bounded by \( \tilde{s} \).
Algorithm 1: Pathwise Iterative Smoothing Shrinkage Thresholding Algorithm (PIS\textsuperscript{2}TA) for solving the SQRT-Lasso optimization (1.4). \( \hat{\theta}_{[K]} \) denotes the output solution corresponding to \( \lambda_K \); \( \theta_{[K]}^{(t)} \) denotes the solution at the \( t \)-th iteration of the \( K \)-th optimization stage; \( \varepsilon_K \) is a pre-specified precision for the \( K \)-th optimization stage. The line search procedure is described in Algorithm 2.

Input: \( y, X, N, \lambda_N, \varepsilon_N, L_{\text{max}} > 0 \)

Initialize:\( \hat{\theta}_{[0]} = 0, \lambda_0 = \|\nabla L_\mu(0)\|_\infty, \eta = (\lambda_N/\lambda_0)^{1/N} \)

For:\( K = 1, \ldots, N \)

\( t \leftarrow 0, \lambda_K \leftarrow \eta\lambda_{K-1}, \theta_{[K]}^{(0)} \leftarrow \hat{\theta}_{[K-1]}, L_{[K]}^{(0)} \leftarrow L_{\text{max}} \)

Repeat:

\( t \leftarrow t + 1 \)

\( L_{[K]}^{(t)} \leftarrow \min\{2\tilde{L}_{[K]}^{(t)}, L_{\text{max}}\} \), where \( \tilde{L}_{[K]}^{(t)} \leftarrow \text{LineSearch}\left(\lambda_K, \theta_{[K]}^{(t-1)}, L_{[K]}^{(t-1)}\right) \)

\( \theta_{[K]}^{(t)} \leftarrow \mathcal{T}_{\mu,\lambda_K}^{\tilde{L}_{[K]}^{(t)},\lambda_K}(\theta_{[K]}^{(t-1)}) \)

Until:\( \omega_{\lambda_K}(\theta_{[K]}^{(t)}) \leq \varepsilon_K \)

\( \hat{\theta}_{[K]} \leftarrow \theta_{[K]}^{(t)} \)

End For

Return:\( \hat{\theta}_{[N]} \)

Algorithm 2: Line search of PIS\textsuperscript{2}TA for SQRT-Lasso.

Input:\( \lambda_K, \theta_{[K]}^{(t-1)}, L_{[K]}^{(t-1)} \)

Initialize:\( \tilde{L}_{[K]}^{(t)} = L_{[K]}^{(t-1)} \)

Repeat:

\( \theta_{[K]}^{(t)} = \mathcal{T}_{\tilde{L}_{[K]}^{(t)},\lambda_K}^{L_{[K]}^{(t)},\lambda_K}(\theta_{[K]}^{(t-1)}) \)

If\( \mathcal{F}_{\mu,\lambda_K}(\theta_{[K]}^{(t)}) < Q_{\mu,\lambda_K}(\theta_{[K]}^{(t)}, \theta_{[K]}^{(t-1)}) \)

\( \tilde{L}_{[K]}^{(t)} = L_{[K]}^{(t-1)}/2 \)

End If

Until:\( \mathcal{F}_{\mu,\lambda_K}(\theta_{[K]}^{(t)}) \geq Q_{\mu,\lambda_K}(\theta_{[K]}^{(t)}, \theta_{[K]}^{(t-1)}) \)

Return:\( \tilde{L}_{[K]}^{(t)} \)

3.1 Computational Theory

Our analysis consists of two phases depending on the estimation error and sparsity of the solution \( \theta \) along the path. Specifically, we denote \( \mathcal{B}_{s^*+\tilde{s}} = \mathcal{B}_r \cap \{ \theta \in \mathbb{R}^d : \|\theta - \theta^*\|_0 \leq s^* + \tilde{s} \} \). Let \( N_1 \in \{1, \ldots, N\} \) be a cut-off between Phase I and Phase II. We can show that Phase I corresponds to the first \( N_1 \) stages of pathwise optimization, in which we cannot guarantee \( \theta \notin \mathcal{B}_{r}^{s^*+\tilde{s}} \). Thus we only establish a sublinear convergence for Phase I. But Phase I is still computationally efficient,
since we can choose reasonably large $\varepsilon_K$ for $K = 1, \ldots, N_1$ to facilitate early stopping; Phase II corresponds to the consequent $(N - N_1)$ stages, in which we guarantee $\theta \in B^*_{r^* + \tilde{s}}$. Thus LRSC and LRSS hold, and a linear convergence can be established accordingly.

**Theorem 3.4.** Suppose Assumptions 3.2 and 3.3 hold. Let $\hat{\theta}_{[K]}$ be the output solution that satisfies $\omega_{\lambda_K}(\hat{\theta}_{[K]}) \leq \varepsilon_K$ of the $K$-th stage respectively for all $K = 1, \ldots, N$. We denote $S^* = \{j \mid \theta_j^* \neq 0\}$, $S^r = \{j \mid \theta_j^* = 0\}$ and $s^* = |S^*|$. Recall that $\eta$ is the decaying ratio of the geometrically decreasing regularization sequence. Given $\mu \leq \frac{\sqrt{n}}{4}$, $\lambda_0 = \|\nabla L_\mu(0)\|_\infty$, and $\eta = (\lambda_N/\lambda_0)^{1/N} \in (5/6, 1)$, there exists $N_1 \in \{1, 2, \ldots, N\}$ such that the following results hold:

**Phase I:** Let $R = \max_{K \leq N_1} \{(\sup_{\theta} \|\theta - \hat{\theta}_{[K]}\|_2 : F_{\mu, \lambda_K}(\theta) \leq F_{\mu, \lambda_K}(\hat{\theta}_{[K]})\}$. At the $K$-th stage, where $K = 1, \ldots, N_1$, we need at most $T_K = \mathcal{O}\left(\frac{|X|^2 R^2}{\varepsilon_K \mu \sqrt{n}}\right)$ iterations to guarantee

$$\omega_{\lambda_K}(\theta_{[K]}^t) \leq \varepsilon_K \quad \text{and} \quad F_{\mu, \lambda_K}(\hat{\theta}_{[K]}) - F_{\mu, \lambda_K}(\theta_{[K]}^t) = \mathcal{O}\left(\frac{|X|^2 R^2}{T_K \mu \sqrt{n}}\right),$$

where $\theta_{[K]}$ is a global optimum to (1.4). Moreover, we have $\theta_{[N_1]} \in B_r$ and $\|\theta_{[N_1]}\|_{\infty} \leq s^* + \tilde{s}$.

**Phase II:** Let $\alpha = 1 - \frac{1}{8s^* + 2\tilde{s}}$. At the $K$-th stage, where $K = N_1 + 1, \ldots, N$, we have sparse solutions throughout all iterations, i.e., $\|\theta_{[K]}^{(i)}\|_{S^*} \leq s^* + \tilde{s}$. Moreover, we need at most $T_K = \mathcal{O}\left(\kappa_{s^* + 2\tilde{s}} \log \left(\frac{\kappa_{s^* + 2\tilde{s}} s^* \lambda_K}{\varepsilon_K}\right)\right)$ iterations to guarantee $\omega_{\lambda_K}(\theta_{[K]}^t) \leq \varepsilon_K$,

$$F_{\mu, \lambda_K}(\hat{\theta}_{[K]}) - F_{\mu, \lambda_K}(\theta_{[K]}^t) = \mathcal{O}\left(\alpha^T \varepsilon_K \lambda_K s^*\right), \quad \text{and} \quad \|\hat{\theta}_{[K]} - \theta_{[K]}^t\|_2 = \mathcal{O}\left(\alpha^T \varepsilon_K \lambda_K s^*\right),$$

where $\theta_{[K]}$ is the unique sparse global optimum to (1.4) with $\lambda_K$ satisfying $\|\theta_{[K]}\|_{S^*} \leq s^* + \tilde{s}$.

Theorem 3.4 guarantees that PIS²TA achieves an R-linear convergence to the unique sparse global optimum to (1.4) in terms of both objective value and solution parameter, which is as nearly efficient as Xiao and Zhang (2013) for solving Lasso with much less turning effort (since $\lambda_N$ is independent of $\sigma$). This further explains why PIS²TA is much more efficient than other competing algorithm for solving SQRT-Lasso such as ADMM and SOCP + interior point method.

A geometric interpretation of Theorem 3.4 is provided in Figure 2. The first $N - 1$ stages serve as intermediate processes to facilitate fast convergence to $\theta_{[N]}$, which do not require high precision solutions. Thus we choose $\varepsilon_K = \lambda_K/4 \gg \varepsilon_N$ for $K = 1, \ldots, N - 1$ such that Phase I is efficient, as shown in Figure 3, and only high precision for the last stage, e.g. $\varepsilon_N = 10^{-5}$ (because only the last regularization parameter is of our interest). The total number of iterations for computing the entire solution path is at most

$$\mathcal{O}\left(\frac{N_1 |X|^2 R^2}{\lambda_{N_1} \mu \sqrt{n}} + \kappa_{s^* + 2\tilde{s}} (N - N_1) \log(\kappa_{s^* + 2\tilde{s}} s^* \lambda_N / \varepsilon_N)\right).$$

Moreover, given properly chosen $\mu$ and $\lambda_N$, we guarantee that none of the linear convergence region, true model parameter $\theta^*$ and output solution $\theta_{[N]}$ fall into the smooth region. This implies that (1.3) and (1.4) share the same global optimum in Phase II. A formal claim is presented in §3.2.
Figure 2: A geometric interpretation of two phases of convergence. Phase I (yellow region): Sublinear convergence, and Phase II (orange region): Linear convergence. The linear convergence region, the true parameter $\theta^*$ and output solution $\hat{\theta}_{[N]}$ do not fall into the smoothed region.

Figure 3: Plots of the objective gaps for all iterations $t$ of each path following stage (only a few stages are demonstrated for clarity).

3.2 Statistical Theory

To analyze the statistical properties of our estimator obtained via PIS$^2$TA, we assume that the design matrix $X$ satisfies the restricted eigenvalue condition as follows.

**Assumption 3.5.** The design matrix $X$ satisfies the Restricted Eigenvalue (RE) condition, i.e., there exist constants $\psi_{\min}, \psi_{\max}, \varphi_{\min}, \varphi_{\max} \in (0, \infty)$, which do not scale with $(s^*, n, d)$, such that

$$\psi_{\min} \|v\|^2_2 - \psi_{\min} \frac{\log d}{n} \|v\|^2_2 \leq \|Xv\|^2_2 \leq \psi_{\max} \|v\|^2_2 + \varphi_{\max} \frac{\log d}{n} \|v\|^2_1,$$

(3.2)

A wide family of examples satisfy the RE condition, such as the correlated sub-Gaussian random design (Rudelson and Zhou, 2013), which has been extensively studied for sparse recovery (Candes and Tao, 2005; Bickel et al., 2009; Raskutti et al., 2010).

We next verify Assumption 3.2 and 3.3 based on the RE condition in the following lemma.
Lemma 3.6. Suppose Assumption 3.5 holds. Given $\mu \leq \frac{\sqrt{n}\sigma}{4}$ and $\lambda_N = 24\sqrt{\frac{\log d}{n}}$, $\lambda_N \geq 6\|\nabla L_\mu(\theta^*)\|_{\infty}$ with high probability. Moreover, for large enough $n$, $L_\mu(\theta)$ satisfies LRSC and LRSS properties on $\mathcal{B}_r$ with high probability, where $r = \frac{\sigma^2_{\max}}{8\psi_{\min}}$. Specifically, (3.1) holds with $\rho_{s^*+2}\delta \leq \frac{8\psi_{\max}}{\sigma}$ and $\rho_{s^*+2}\delta \geq \frac{\psi_{\min}}{8\sigma}$, where $\delta = C_2s^* > (196\kappa_{s^*+2}\delta + 144\kappa_{s^*+2}\delta)s^*$, $C_2 \in \mathbb{R}^+$ is a generic constant and $\kappa_{s^*+2}\delta \leq 64\psi_{\max}/\psi_{\min}$.

Lemma 3.6 guarantees that Assumption 3.2 holds given properly chosen $\mu$ and $\lambda_N$, and Assumption 3.3 holds given the design $X$ satisfying RE condition, both with high probability. Therefore, by Theorem 3.4, PIS$^2$TA achieves an R-linear convergence to the unique sparse global optimum. In the next theorem, we characterize the statistical rate of convergence of PIS$^2$TA.

Theorem 3.7. Suppose Assumption 3.5 holds. Let the output solution $\hat{\theta}_{[N]}$ satisfy $\omega_{\lambda_N}(\hat{\theta}_{[N]}) \leq \varepsilon_N$ for a small enough $\varepsilon_N$. Given $\mu \leq \frac{\sqrt{n}\sigma}{4}$, $\lambda_N = 24\sqrt{\frac{\log d}{n}}$ and a large enough $n$, we have:

$$||\hat{\theta}_{[N]} - \theta^*||_2 = O_P\left(\sigma\sqrt{s^*\log d/n}\right) \quad \text{and} \quad ||\hat{\theta}_{[N]} - \theta^*||_1 = O_P\left(\sigma s^*\sqrt{\log d/n}\right).$$

Moreover, let $\hat{\sigma} = \frac{||y-X\hat{\theta}_{[N]}||_2}{\sqrt{n}}$ be the estimation of $\sigma$. Then we have $|\hat{\sigma} - \sigma| = O_P(\sigma s^*\log d/n)$.

Theorem 3.7 guarantees that the output solution $\hat{\theta}_{[N]}$ obtained by PIS$^2$TA achieves the minimax optimal rate of convergence in parameter estimation (Ye and Zhang, 2010; Raskutti et al., 2011). The next proposition shows that (1.3) and (1.4) share the same global optimum, which corresponds to a well fitted model. This implies that neither the unique sparse global optimum $\bar{\theta}_{[N]}$ nor the linear convergence region falls into the smooth region, as shown in Figure 2.

Proposition 3.8. Under the same assumptions as Theorem 3.7, for all $\lambda_K$’s, where $K = N_1 + 1, \ldots, N$, (1.3) and (1.4) share the unique sparse global optimum with high probability.

4 Extension to Sparse Precision Matrix Estimation

We consider the TIGER approach proposed in Liu and Wang (2012) for estimating sparse precision matrix. To be clear, we emphasize that we use $v_1, v_2, \ldots$ (without [] for subscript) to index vectors. Let $X = [x_1^\top, \ldots, x_n^\top]^\top \in \mathbb{R}^{n \times d}$ be $n$ observed data points from a $d$-variate Gaussian distribution $\mathcal{N}_d(0, \Sigma)$. Our goal is to estimate the sparse precision matrix $\Theta = \Sigma^{-1}$. Let $Z = X\hat{\Gamma}^{-1/2} = [z_1^\top, \ldots, z_n^\top]^\top \in \mathbb{R}^{n \times d}$ be the standardized data matrix, where $\hat{\Gamma} = \text{diag}(\hat{\Sigma}_{11}, \ldots, \hat{\Sigma}_{dd})$ is a diagonal matrix and $\hat{\Sigma} = \frac{1}{n}X^\top X$. Then, we can write $z_i = Z_{*,i\setminus i}^\top \theta_i^* + \hat{\Gamma}^{-1/2}_{ii} \epsilon_i$ for all $i = 1, \ldots, d$, where $\theta_i^* = \hat{\Gamma}^{-1/2}_{ii} \hat{\Gamma}^{-1/2}_{i\setminus i}(\hat{\Sigma}_{i\setminus i})^{-1} \Sigma_{i,i}$ and $\epsilon_i \sim \mathcal{N}_n(0, \sigma_i^2\mathbb{I}_n)$ with $\sigma_i^2 = \Sigma_{ii} - \Sigma_{i\setminus i}^{-1}(\Sigma_{i\setminus i})^{-1} \Sigma_{i,i}$. We denote $\tau_i^2 = \sigma_i^2\hat{\Gamma}^{-1}_{ii}$, and solve:

$$\bar{\theta}_i = \arg\min_{\theta_i \in \mathbb{R}^{d-1}} L_\mu,i(\theta_i) + \lambda||\theta_i||_1 \quad \text{and} \quad \bar{\tau}_i = ||z_i - Z_{*,i\setminus i}\bar{\theta}_i||_\mu,$$

(4.1)
for all $i = 1, \ldots, d$, where $L_{\mu,i}(\theta_i) = \frac{1}{\sqrt{n}} \| z_i - Z_{i}^{\star}, \theta_{i} \|_{\mu}$. Then the $i$-th column of precision matrix $\Theta$ is estimated by: $\hat{\Theta}_{ii} = \tau_i^{-2} \tilde{\Gamma}_{ii}^{-1}$, and $\hat{\Theta}_{\setminus i,i} = -\tau_i^{-2} \tilde{\Gamma}_{\setminus i,i}^{-1/2} \tilde{\Theta}_{\setminus i,i}^{1/2}$.

Here we solve (4.1) by PIS$^2$TA. We then introduce a few mild technical assumptions:

**Assumption 4.1.** Suppose the true covariance matrix $\Sigma^{\star}$ and precision matrix $\Theta^{\star}$ satisfy: (A1) $\Theta^{\star} \in M(\kappa_{\Theta}, s^{\star}) = \{ \Theta \in \mathbb{R}^{d \times d} : \Theta > 0, \ \Lambda_{\max}(\Theta)/\Lambda_{\min}(\Theta) \leq \kappa_{\Theta}, \ \max_{i} \sum_{j}(\Theta_{ij} \neq 0) \leq s^{\star} \}$; (A2) $(s^{\star})^{2} \log d = o(n)$; (A3) $\lim \sup_{n \to \infty} \max_{i}(\Sigma_{ii}^{\star})^{2} \log d/n < 1/4$.

We first verify the assumptions required by our computational theory by the following lemma.

**Lemma 4.2.** Suppose Assumption 4.1 holds. Given $\mu \leq \frac{1}{5} \sqrt{\frac{n}{\kappa_{\Theta}}}$ and $\lambda_{N} = 6 \sqrt{\frac{5 \log d}{n}}$, we have $\lambda_{N} \geq 6 \max_{i \in \{1, \ldots, d\}} \| \nabla L_{\mu,i}(\Theta_{i}^{\star}) \|_{\infty}$ with high probability. Moreover, $L_{\mu,i}(\Theta_{i})$ satisfies LRSC and LRSS properties on $B_{r_{i}}$ for all $i = 1, \ldots, d$ with high probability, where $r_{i} = \frac{s^{2}}{12 \kappa_{\Theta}}$. Specifically, for all $i = 1, \ldots, d$, (3.1) holds with $\rho_{s^{\star} + 2s}^{\star} \leq 12 \kappa_{\Theta} s^{\star}$ and $\rho_{s^{\star} + 2s}^{\star} \geq \frac{1}{12 \kappa_{\Theta} s^{\star}}$, where $s = C_{3}s^{\star} > (196 \kappa_{s^{\star} + 2s}^{\star} + 144 \kappa_{s^{\star} + 2s}^{\star})s^{\star}$ for a generic constant $C_{3}$, and $\kappa_{s^{\star} + 2s}^{\star} \leq 144 \kappa_{\Theta}^{2}$.

Lemma 4.2 guarantees that Assumption 3.2 and 3.3 holds with high probability given properly chosen $\mu$ and $\lambda_{N}$. Thus, by Theorem 3.4, PIS$^2$TA achieves an R-linear convergence to the unique sparse global optimum of (4.1) with high probability for all columns of $\Theta^{\star}$. The next theorem characterizes the statistical rate of convergence of the obtained precision matrix estimator using PIS$^2$TA.

**Theorem 4.3.** Suppose Assumption 4.1 holds. Let $\hat{\Theta}_{[N]}$ be the output solution for the regularization parameter $\lambda_{N}$. Given $\mu \leq \frac{1}{5} \sqrt{\frac{n}{\kappa_{\Theta}}}$ and $\lambda_{N} = 6 \sqrt{\frac{5 \log d}{n}}$, we have

$$
\| \hat{\Theta}_{[N]} - \Theta^{\star} \|_{2} = O_{P}\left( s^{\star} \| \Theta^{\star} \|_{2} \sqrt{\log d/n} \right)
$$

Theorem 4.3 implies that our obtained precision matrix estimator attains the minimax optimal rate of convergence in parameter estimation. Moreover, we guarantee that neither the linear convergence region nor output solution $\hat{\Theta}_{[N]}$ falls into the smooth region with high probability.

## 5 Numerical Experiments

We investigate the computational performance of the proposed PIS$^2$TA algorithm through numerical experiments over both simulated and real data example. All simulations are implemented in C with double precision using a PC with an Intel 3.3GHz Core i5 CPU and 16GB memory.

For simulated data, we generate a training dataset of 200 samples, where each row of the design matrix $X_{i}^{\star}$, $i = 1, \ldots, 200$, independently from a 2000-dimensional normal distribution $N(0, \Sigma)$ where $\Sigma_{jj} = 1$ and $\Sigma_{jk} = 0.5$ for all $k \neq j$. We set $s^{\star} = 3$ with $\theta_{1}^{\star} = 3$, $\theta_{2}^{\star} = -2$, and $\theta_{4}^{\star} = 1.5$, and $\theta_{j}^{\star} = 0$ for all $j \neq 1, 2, 4$. A validation set of 200 samples for the regularization parameter selection and a testing set of 10,000 samples are also generated to evaluate the prediction accuracy.
We set $\sigma = 0.5, 1, 2, 4$ respectively to illustrate the tuning insensitivity. The regularization parameter of both Lasso and SQRT-Lasso is chosen over a geometrically decreasing sequence $\{\lambda_K\}_{t=0}^{50}$ with $
olinebreak \lambda_{50} = \sigma \sqrt{\log d / n / 2}$ for Lasso and $\lambda_{50} = \sqrt{\log d / n / 2}$ for SQRT-Lasso. The optimal regularization parameter is determined by $\lambda_{opt} = \lambda_{\hat{N}}$ as $\hat{N} = \arg\min_{K \in \{0, \ldots, 50\}} \|\hat{y} - \hat{X}\hat{\theta}_{[K]}\|_2^2$, where $\hat{\theta}_{[K]}$ denotes the obtained estimate using the regularization parameter $\lambda_K$, and $\hat{y}$ and $\hat{X}$ denote the response vector and design matrix of the validation set. For both Lasso and SQRT-Lasso, we set the stopping precision $\varepsilon_K = 10^{-5}$ for all $K = 1, \ldots, 30$ and $\varepsilon_K = 10^{-4}$ for all $K = 31, \ldots, 50$. For SQRT-Lasso, we set the smoothing parameter $\mu = 10^{-4}$.

First of all, we compare PIS^2TA with ADMM proposed in Li et al. (2015). The backtracking line search described in Algorithm 2 is adopted to accelerate both algorithms. We conduct 500 simulations for all $\sigma$’s. The results are presented in Table 1. The PIS^2TA and ADMM algorithms attain similar objective values, but PIS^2TA is about 20 times faster than ADMM. Both algorithms also achieve similar estimation errors. Throughout all 500 simulations, we have $\|y - X\hat{\theta}_{[\hat{N}]\|_2} > \sqrt{n}\mu \approx 0.0015$. This implies that all obtained optimal estimators are outside the smoothed region of the optimization problem, i.e., the smoothing approach does not hurt the solution accuracy.

Next, we compare the computational and statistical performance between Lasso (solved by PISTA Xiao and Zhang (2013)) and SQRT-Lasso (solved by PIS^2TA). The results averaged over 500 simulations are summarized in Tables 1. In terms of statistical performance, Lasso and SQRT-Lasso attains similar estimation and prediction error. In terms of computational performance, the PIS^2TA algorithm for solving SQRT-Lasso is as efficient as PISTA for Lasso, which matches our computational analysis.

Moreover, we also examine the optimal regularization parameters for Lasso and SQRT-Lasso. We visualize the distribution of all 500 selected $\lambda_{opt}$’s using the kernel density estimator. Particularly, we adopt the Gaussian kernel, and the kernel bandwidth is selected based on the 10-fold cross validation. Figure 4 illustrates the estimated density functions. The horizontal axis corresponds to the rescaled regularization parameter $\lambda_{opt} / \sqrt{\log d / n}$. We see that the optimal regularization parameters of Lasso significantly vary with different $\sigma$. In contrast, the optimal regularization parameters of SQRT-Lasso are more concentrated. This is consistent with the claimed tuning insensitivity.

Finally, we compare PIS^2TA with ADMM over real data sets for precision matrix estimation. Particularly, we use four real world biology data sets preprocessed by Li and Toh (2010): Estrogen ($d = 692$), Arabidopsis ($d = 834$), Leukemia ($d = 1, 225$), Hereditary ($d = 1, 869$). We set three different values for $\lambda_N$ such that the obtained estimators achieve different levels of sparse recovery. We set $N = 50$, and $\varepsilon_K = 10^{-4}$ for all $K$’s. The timing performance is summarized in Table 2. As can be seen, PIS^2TA is 5 to 20 times faster than ADMM on all four data sets.

---

2We do not have any results for the algorithm proposed in Belloni et al. (2011), because it failed to finish 500 simulations in 12 hours. The implementation was based on SDPT3.

3We do not have any results for the algorithm proposed in Belloni et al. (2011), because it failed to finish the experiments on all four data sets in 12 hours. The implementation was also based on SDPT3.
Table 1: Quantitative comparison between Lasso (PISTA) and SQRT-Lasso (PIS²TA and ADMM) over 500 simulations. The estimation error is defined as $\|\hat{\theta} - \theta^*\|_2$. The prediction error is defined as $\|\bar{y} - \bar{X}\hat{\theta}[\bar{X}]\|_2/\sqrt{n}$. The residual is defined as $\|y - X\hat{\theta}[\hat{N}]\|_2$ for PIS²TA only. PIS²TA attains nearly the same estimation and prediction errors as ADMM, but is significantly faster than ADMM over all different settings. Besides, we also find that all obtained estimators are outside the smoothed region throughout all 500 simulations.

| Variance of Noise | Est. Err. | Pred. Err. | Time (Second) | Residual |
|-------------------|-----------|------------|---------------|----------|
|                   | Lasso     | PIS²TA    | Lasso         | PIS²TA   | PIS²TA | ADMM | PIS²TA |
| $\sigma = 0.5$    | 0.2761    | 0.2760    | 0.5403        | 0.5399   | 0.8817 | 0.9526 | 17.260 | 5.2505 |
|                   | (0.0651)  | (0.0537)  | (0.0172)      | (0.0143) | (0.2824) | (0.2646) | (3.1723) | (0.9591) |
| $\sigma = 1$     | 0.5271    | 0.5319    | 1.0722        | 1.0757   | 0.9146 | 1.0170 | 19.762 | 10.5209 |
|                   | (0.1744)  | (0.1065)  | (0.0303)      | (0.0280) | (0.3185) | (0.2959) | (4.4387) | (1.7760) |
| $\sigma = 2$     | 1.0621    | 1.1065    | 2.1551        | 2.1492   | 1.1772 | 1.1263 | 24.464 | 20.886 |
|                   | (0.2521)  | (0.2141)  | (0.0595)      | (0.0589) | (0.4582) | (0.4128) | (4.2786) | (3.6506) |
| $\sigma = 4$     | 2.1275    | 2.1356    | 4.2928        | 4.2963   | 1.2913 | 1.2544 | 26.101 | 45.7623 |
|                   | (0.4247)  | (0.4033)  | (0.1112)      | (0.1079) | (0.4855) | (0.5074) | (5.7725) | (5.6333) |

Table 2: Timing comparison between PIS²TA and ADMM on biology data under different levels of sparsity recovery. PIS²TA is significantly faster than ADMM over all settings and data sets.

|                | Estrogen | Arabidopsis | Leukemia | Hereditary |
|----------------|----------|-------------|----------|------------|
|                | PIS²TA   | ADMM        | PIS²TA   | ADMM       |
| Sparsity 1%    | 16.562   | 175.98      | 18.404   | 373.83     |
| Sparsity 3%    | 70.622   | 338.96      | 81.557   | 707.52     |
| Sparsity 10%   | 188.03   | 703.24      | 226.97   | 1378.1     |

Figure 4: Estimated distributions of $\lambda_{opt}/\sqrt{\log d/n}$ over different values of $\sigma$ for Lasso and PIS²TA.
References

Agarwal, A., Negahban, S. and Wainwright, M. J. (2010). Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In Advances in Neural Information Processing Systems.

Beck, A. and Teboulle, M. (2012). Smoothing and first order methods: A unified framework. SIAM Journal on Optimization 22 557–580.

Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. Biometrika 98 791–806.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. The Annals of Statistics 37 1705–1732.

Bühlmann, P. and van de Geer, S. (2011). Statistics for high-dimensional data: methods, theory and applications. Springer.

Candes, E. J. and Tao, T. (2005). Decoding by linear programming. IEEE Transactions on Information Theory 51 4203–4215.

Johnstone, I. M. (2001). Chi-square oracle inequalities. Lecture Notes-Monograph Series 399–418.

Li, L. and Toh, K.-C. (2010). An inexact interior point method for l1-regularized sparse covariance selection. Mathematical Programming Computation 2 291–315.

Li, X., Zhao, T., Yuan, X. and Liu, H. (2015). The flare package for high dimensional linear regression and precision matrix estimation in R. The Journal of Machine Learning Research 16 553–557.

Liu, H. and Wang, L. (2012). Tiger: A tuning-insensitive approach for optimally estimating Gaussian graphical models. Tech. rep., Massachusetts Institute of Technology.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. Statistical Science 27 538–557.

Nesterov, Y. (2004). Introductory lectures on convex optimization: A basic course, vol. 87. Springer.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. Mathematical Programming 103 127–152.

Nesterov, Y. (2013). Gradient methods for minimizing composite functions. Mathematical Programming 140 125–161.
Raskutti, G., Wainwright, M. J. and Yu, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *The Journal of Machine Learning Research* **11** 2241–2259.

Raskutti, G., Wainwright, M. J. and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over-balls. *Information Theory, IEEE Transactions on* **57** 6976–6994.

Rudelson, M. and Zhou, S. (2013). Reconstruction from anisotropic random measurements. *Information Theory, IEEE Transactions on* **59** 3434–3447.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58** 267–288.

Wainwright, J. (2015). High-dimensional statistics: A non-asymptotic viewpoint. *preparation. University of California, Berkeley*.

Xiao, L. and Zhang, T. (2013). A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM Journal on Optimization* **23** 1062–1091.

Ye, F. and Zhang, C.-H. (2010). Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *The Journal of Machine Learning Research* **11** 3519–3540.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics* **36** 1567–1594.

Zhang, T. (2009). Some sharp performance bounds for least squares regression with $\ell_1$ regularization. *The Annals of Statistics* **37** 2109–2144.
A Intermediate Results of Theorem 3.4 and Theorem 3.7

We introduce some important implications of the proposed assumptions. Recall that $S^* = \{j : \theta^*_j \neq 0\}$ be the index set of non-zero entries of $\theta^*$ with $s^* = |S^*|$ and $\bar{S}^* = \{j : \theta^*_j = 0\}$ be the complement set. From Lemma 3.6, Assumption 3.3 implies RSC and RSS with parameter $\rho_{s^* + 2\bar{s}}^+$ and $\rho_{s^* + 2\bar{s}}^{+2}$ respectively. By Nesterov (2004), the following conditions are equivalent to RSC and RSS, i.e., for any $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ satisfying $\|\mathbf{v} - \mathbf{w}\|_0 \leq s^* + 2\bar{s}$,

$$\rho_{s^* + 2\bar{s}}^- \|\mathbf{v} - \mathbf{w}\|_2^2 \leq (\mathbf{v} - \mathbf{w})^\top \nabla L_\mu (\mathbf{w}) \leq \rho_{s^* + 2\bar{s}}^+ \|\mathbf{v} - \mathbf{w}\|_2^2, \quad (A.1)$$

$$\frac{1}{\rho_{s^* + 2\bar{s}}^-} \|\nabla L_\mu (\mathbf{v}) - \nabla L_\mu (\mathbf{w})\|_2^2 \leq (\mathbf{v} - \mathbf{w})^\top \nabla L_\mu (\mathbf{w}) \leq \frac{1}{\rho_{s^* + 2\bar{s}}^+} \|\nabla L_\mu (\mathbf{v}) - \nabla L_\mu (\mathbf{w})\|_2^2. \quad (A.2)$$

From the convexity of $\ell_1$ norm, we have

$$\|\mathbf{v}\|_1 - \|\mathbf{w}\|_1 \geq (\mathbf{v} - \mathbf{w})^\top \mathbf{g}, \quad (A.3)$$

where $\mathbf{g} \in \partial \|\mathbf{w}\|_1$. Combining and (A.1) and (A.3), we have for any $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ satisfying $\|\mathbf{v} - \mathbf{w}\|_0 \leq s^* + 2\bar{s}$,

$$\mathcal{F}_{\mu, \lambda} (\mathbf{v}) - \mathcal{F}_{\mu, \lambda} (\mathbf{w}) - (\mathbf{v} - \mathbf{w})^\top \nabla \mathcal{F}_{\mu, \lambda} (\mathbf{w}) \geq \rho_{s^* + 2\bar{s}}^- \|\mathbf{v} - \mathbf{w}\|_2^2. \quad (A.4)$$

**Remark A.1.** For any $t$ and $k$, the line search satisfies

$$\bar{L}_{j[K]}^{(t)} \leq L_{j[K]} \leq \bar{L}_{j[K]}^{(t)} \leq L_{j[K]} \leq 2L_\mu \quad \text{and} \quad \rho_{s^* + 2\bar{s}}^+ \leq \bar{L}_{j[K]}^{(t)} \leq L_{j[K]} \leq 2\rho_{s^* + 2\bar{s}}^+. \quad (A.5)$$

where $L_\mu = \min\{L : \|\nabla L_\mu (\mathbf{v}) - \nabla L_\mu (\mathbf{w})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2, \forall \mathbf{v}, \mathbf{w} \in \mathbb{R}^d\}$.

We first show that when $\mathbf{u}$ is sparse and the approximate KKT condition is satisfied, then both estimation error (in $\ell_2$ norm) and objective error, w.r.t. the true model parameter, are bounded. This characterizes that the initial value $\mathbf{u}^{(0)}_{[K]}$ of $K$-th path following stage has desirable statistical properties if we initialize $\mathbf{u}^{(0)}_{j[K]} = \hat{\theta}_{j[K] - 1}$. This is formalized in Lemma A.2, and its proof is deferred to Appendix N.1.

**Lemma A.2.** Suppose Assumption 3.2 and Assumption 3.3 hold, and $\lambda \geq \lambda_N$. If $\mathbf{u}$ satisfies $\|\mathbf{u}_{\bar{S}^*}\|_0 \leq \bar{s}$ and the approximate KKT condition

$$\min_{\mathbf{g} \in \partial \|\mathbf{u}\|_1} \|\nabla L_\mu (\mathbf{u}) + \lambda \mathbf{g}\|_\infty \leq \lambda / 2, \quad (A.6)$$

then we have

$$\|\mathbf{u} - \mathbf{u}^*\|_1 \leq 5\|\mathbf{u} - \mathbf{u}^*\|_{\bar{S}^*} \|_1, \quad (A.7)$$

$$\|\mathbf{u} - \mathbf{u}^*\|_2 \leq \frac{2\lambda \sqrt{s^*}}{\rho_{s^* + 2\bar{s}}^-}, \quad (A.8)$$

$$\|\mathbf{u} - \mathbf{u}^*\|_1 \leq \frac{12\lambda s^*}{\rho_{s^* + 2\bar{s}}^+}, \quad (A.9)$$

$$\mathcal{F}_{\mu, \lambda} (\mathbf{u}) - \mathcal{F}_{\mu, \lambda} (\mathbf{u}^*) \leq \frac{6\lambda^2 s^*}{\rho_{s^* + 2\bar{s}}^+}. \quad (A.10)$$

15
Next, we show that if $\theta$ is sparse and the objective error is bounded, then the estimation error is also bounded. This characterizes that within the $K$-th path following stage, good statistical performance is preserved after each proximal-gradient update. This is formalized in Lemma A.3, and its proof is deferred to Appendix N.2.

**Lemma A.3.** Suppose Assumption 3.2 and Assumption 3.3 hold, and $\lambda \geq \lambda_N$. If $\theta$ satisfies $\|\theta^*_S\|_0 \leq \tilde{s}$ and the objective satisfies

$$F_{\mu, \lambda}(\theta) - F_{\mu, \lambda}(\theta^*) \leq \frac{6\lambda^2 s^*}{\rho_{s^* + 2\tilde{s}}}$$

then we have

$$\|\theta - \theta^*\|_2 \leq \frac{4\lambda\sqrt{3s^*}}{\rho_{s^* + 2\tilde{s}}}$$  \hspace{1cm} (A.11)

$$\|\theta - \theta^*\|_1 \leq \frac{24\lambda s^*}{\rho_{s^* + 2\tilde{s}}}$$  \hspace{1cm} (A.12)

We then show that if $\theta$ is sparse and the objective error is bounded, then each proximal-gradient update preserves solution to be sparse. This demonstrates that within the $K$-th path following stage, each update of $\theta(t)_{[K]}$ is preserved to be sparse and has good statistical performance. This is formalized in Lemma A.4, and its proof is deferred to Appendix N.3.

**Lemma A.4.** Suppose Assumption 3.2 and Assumption 3.3 hold, and $\lambda \geq \lambda_N$. If $\theta$ satisfies $\|\theta^*_S\|_0 \leq \tilde{s}$, $L$ satisfies $L < 2\rho_{s^* + 2\tilde{s}}$, and the objective satisfies

$$F_{\mu, \lambda}(\theta) - F_{\mu, \lambda}(\theta^*) \leq \frac{6\lambda^2 s^*}{\rho_{s^* + 2\tilde{s}}}$$

then we have

$$\| (T_{L, \lambda}(\theta))_{\tilde{s}}^* \|_0 \leq \tilde{s}. \hspace{1cm} (A.13)$$

Moreover, we show that if $\theta$ satisfies the approximate KKT condition, then the objective has a bounded error w.r.t. the regularization parameter $\lambda$. This characterizes the geometric decrease of the objective error when we choose a geometrically decreasing sequence of regularization parameters. This is formalized in Lemma A.5, and its proof is deferred to Appendix N.4.

**Lemma A.5.** Suppose Assumption 3.2 and Assumption 3.3 hold, and $\lambda \geq \lambda_N$. If $\theta$ satisfies

$$\omega_\lambda(\theta) \leq \lambda/2,$$

For any $\tilde{\lambda} \in [\lambda_N, \lambda]$, let $\overline{\theta} = \arg\min_\theta F_{\mu, \tilde{\lambda}}(\theta)$. Then we have

$$F_{\mu, \lambda}(\theta) - F_{\mu, \tilde{\lambda}}(\overline{\theta}) \leq \frac{12(\lambda + \tilde{\lambda}) (\omega_\lambda(\theta) + \lambda - \tilde{\lambda}) s^*}{\rho_{s^* + 2\tilde{s}}}.$$
Furthermore, we show that each path following stage has a local linear convergence rate if the initial value $\theta^{(0)}$ is sparse and satisfies the approximate KKT condition with adequate precision. Besides, the estimation after each proximal gradient update is also sparse. This is the key result in demonstrating the overall geometric convergence rate of the algorithm. This is formalized in Lemma A.6, and its proof is deferred to Appendix N.5.

Lemma A.6. Suppose Assumption 3.3 holds. If the initialization $\theta^{(0)}$ for every stage with any $\lambda$ in Algorithm 1 satisfies

$$\|\theta^{(0)}\|_0 \leq \tilde{s},$$

Then for any $t = 1, 2, \ldots$, we have $\|\theta^{(t)}\|_0 \leq \tilde{s}$,

$$F_{\mu,\lambda}(\theta^{(t)}) - F_{\mu,\lambda}(\overline{\theta}) \leq \left(1 - \frac{1}{8k_s^* + 2\tilde{s}}\right)^t \left(F_{\mu,\lambda}(\theta^{(0)}) - F_{\mu,\lambda}(\overline{\theta})\right),$$

where $\overline{\theta} = \arg\min_{\theta} F_{\mu,\lambda}(\theta)$.

In addition, we provide the sublinear convergence rate when RSC does not hold based on a refined analysis of the convergence rate for convex objective (not strongly convex) via proximal gradient method with line search Nesterov (2013). Specifically, we provide a sub-linear rate of convergence without the need to classify the distance of the initial objective to the optimal objective. This characterizes the convergence behavior when $\|X(\theta - \theta^*)\|_2$ is large. We formalize this in Lemma A.7, and provide the proof in Appendix N.6.

Lemma A.7 (Refined result of Theorem 4 in Nesterov (2013)). Given the initialization $\theta^{(0)}$, if for any $\theta \in \mathbb{R}^d$ that satisfies $F_{\mu,\lambda}(\theta) \leq F_{\mu,\lambda}(\theta^{(0)})$, denote $R$ as

$$\|\theta - \overline{\theta}\|_2 \leq R.$$

Then for any $t = 1, 2, \ldots$, we have

$$F_{\mu,\lambda}(\theta^{(t)}) - F_{\mu,\lambda}(\overline{\theta}) \leq \frac{4\|X\|_2^2 R^2}{(t + 2)\mu\sqrt{n}}, \quad (A.14)$$

where $\overline{\theta} = \arg\min_{\theta} F_{\mu,\lambda}(\theta)$.

Finally, we introduce two results characterizing the proximal gradient mapping operation, adapted from Nesterov (2013) and Xiao and Zhang (2013) without proof. The first lemma describes sufficient descent of the objective by proximal gradient method.

Lemma A.8 (Adapted from Theorem 2 in Nesterov (2013)). For any $L > 0$,

$$Q_{\mu,\lambda}(T_{L,\lambda}(\theta), \theta) \leq F_{\mu,\lambda}(\theta) - \frac{L}{2}\|T_{L,\lambda}(\theta) - \theta\|_2^2.$$
Besides, if $L_\mu(\theta)$ is convex, we have
\[ Q_{\mu,\lambda}(T_{L,\lambda}(\theta), \theta) \leq \min_x F_{\mu,\lambda}(x) + \frac{L}{2} \|x - \theta\|_2^2. \tag{A.15} \]
Further, we have for any $L \geq L_\mu$,
\[ F_{\mu,\lambda}(T_{L,\lambda}(\theta)) \leq Q_{\mu,\lambda}(T_{L,\lambda}(\theta), \theta) \leq F_{\mu,\lambda}(\theta) - \frac{L}{2} \|T_{L,\lambda}(\theta) - \theta\|_2^2. \tag{A.16} \]

The next lemma provides an upper bound of the optimal residue $\omega(\cdot)$.

**Lemma A.9** (Adapted from Lemma 2 in Xiao and Zhang (2013)). For any $L > 0$, if $L_\mu$ is the Lipschitz constant of $\nabla L_\mu$, then
\[ \omega_\lambda(T_{L,\lambda}(\theta)) \leq (L + S_L(\theta)) \|T_{L,\lambda}(\theta) - \theta\|_2 \leq (L + L_\mu) \|T_{L,\lambda}(\theta) - \theta\|_2, \]
where $S_L(\theta) = \frac{\|\nabla L_\mu(T_{L,\lambda}(\theta)) - \nabla L_\mu(\theta)\|_2}{\|T_{L,\lambda}(\theta) - \theta\|_2}$ is a local Lipschitz constant, which satisfies $S_L(\theta) \leq L_\mu$.

**B Proof of Theorem 3.4**

We first demonstrate the sublinear rate for initial stages when the estimation error $\|\theta - \theta^*\|_2$ is large due large regularization parameter $\lambda_K$. The proof is provided in Appendix F.

**Theorem B.1.** For any $K = 1, \ldots, N$ and $\lambda_K > 0$, let $\overline{\theta}_{[K]} = \arg\min_{\theta} F_{\mu,\lambda_K}(\theta)$ be the optimal solution of $K$-th stage with regularization parameter $\lambda_K$. For any $\theta \in \mathbb{R}^d$ that satisfies $F_{\mu,\lambda_K}(\theta) \leq F_{\mu,\lambda_K}(\theta_{[K]}^{(0)})$, let $\|\theta - \overline{\theta}_{[K]}\|_2 \geq \lambda_K / 2$. If the initial value $\theta_{[K]}^{(0)}$ satisfies $\omega_{\lambda_K}(\theta_{[K]}^{(0)}) \leq \lambda_K / 2$, then within $K$-th stage, for any $t = 1, 2, \ldots$, we have
\[ F_{\mu,\lambda}(\theta_{[K]}^{(t)}) - F_{\mu,\lambda}(\overline{\theta}_{[K]}) \leq \frac{4\|X\|_2^2 R^2}{(t + 2) \mu \sqrt{n}}. \tag{B.1} \]
To achieve the approximate KKT condition $\omega_{\lambda_K}(\theta_{[K]}^{(t)}) \leq \varepsilon_K$, the number of proximal gradient steps is no more than
\[ \frac{18\|X\|_2^2}{\varepsilon_K \mu^{3/2} n^{3/4} \sqrt{L_\mu}} - 3. \tag{B.2} \]

Note that $L_\mu \asymp \|X\|_2^2/(\mu \sqrt{n})$. Then (B.2) can be simplified as $O\left(\frac{\|X\|_2^2 R}{\varepsilon_K \mu \sqrt{n}}\right)$.

Next, we demonstrate the linear rate when the estimator satisfies $\theta \in B_{\rho^*+\bar{\rho}}$. The proof is provided in Appendix G.

**Theorem B.2.** Suppose Assumption 3.2 and Assumption 3.3 hold, and $\lambda_K > 0$ for any $K = N_1, \ldots, N$. Let $\overline{\theta}_{[K]} = \arg\min_{\theta} F_{\mu,\lambda_K}(\theta)$ be the optimal solution of $K$-th stage with regularization
parameter $\lambda_K$. If the initial value $\theta_{[K]}^{(0)}$ satisfies $\omega_{\lambda_K}(\theta_{[K]}^{(0)}) \leq \lambda_K/2$ with $\|\theta_{[K]}^{(0)}\|_0 \leq \tilde{s}$, then within $K$-th stage, for any $t = 1, 2, \ldots$, we have

$$
\|\theta_{[K]}^{(t)}\|_{\mathcal{S}^*} \leq \tilde{s}, \quad \|\theta_{[K]}^{(t)} - \overline{\theta}_{[K]}\|_2^2 \leq \left(1 - \frac{1}{8\kappa_{s^*+2\tilde{s}}}\right)^t \frac{24\lambda_K s^* \omega_{\lambda_K}(\theta_{[K]}^{(t)})}{(\rho_{s^*+2\tilde{s}}^*)^2} \
\mathcal{F}_{\mu, \lambda_K}(\theta_{[K]}^{(t)}) - \mathcal{F}_{\mu, \lambda_K}(\overline{\theta}_{[K]}) \leq \left(1 - \frac{1}{8\kappa_{s^*+2\tilde{s}}}\right)^t \frac{24\lambda_K s^* \omega_{\lambda_K}(\theta_{[K]}^{(t)})}{\rho_{s^*+2\tilde{s}}^*},
$$

(B.3)

(1) For $K = N_1, \ldots, N - 1$, to achieve the approximate KKT condition $\omega_{\lambda_K}(\theta_{[K]}^{(t)}) \leq \lambda_K/4$, the number of proximal gradient steps is no more than

$$
\frac{\log \left(1536 \left(1 + \kappa_{s^*+2\tilde{s}}\right)^2 s^* \kappa_{s^*+2\tilde{s}} \right)}{\log \left(8\kappa_{s^*+2\tilde{s}}/(8\kappa_{s^*+2\tilde{s}} - 1)\right)}.
$$

(B.4)

(2) For $K = N$, to achieve the approximate KKT condition $\omega_{\lambda_N}(\theta_{[N]}^{(t)}) \leq \varepsilon_N$, the number of proximal gradient steps is no more than

$$
\frac{\log \left(96 \left(1 + \kappa_{s^*+2\tilde{s}}\right)^2 \lambda_N^2 s^* \kappa_{s^*+2\tilde{s}} \varepsilon_N^2 \right)}{\log \left(8\kappa_{s^*+2\tilde{s}}/(8\kappa_{s^*+2\tilde{s}} - 1)\right)}.
$$

(B.5)

From basic inequalities, since $\kappa_{s^*+2\tilde{s}} \geq 1$, we have

$$
\log \left(\frac{8\kappa_{s^*+2\tilde{s}}}{8\kappa_{s^*+2\tilde{s}} - 1}\right) \geq \log \left(1 + \frac{1}{8\kappa_{s^*+2\tilde{s}} - 1}\right) \geq \frac{1}{8\kappa_{s^*+2\tilde{s}}}.
$$

Then (B.4) and (B.5) can be simplified as $O \left(\kappa_{s^*+2\tilde{s}} \left(\log \left(\kappa_{s^*+2\tilde{s}} s^*\right)\right)\right)$ and $O \left(\kappa_{s^*+2\tilde{s}} \left(\log \left(\kappa_{s^*+2\tilde{s}}^3 \lambda_N^2 s^*/\varepsilon_N^2\right)\right)\right)$ respectively.

As can be seen from Theorem B.2, when the initial value $\theta_{[K]}^{(0)}$ satisfies the approximate KKT condition $\omega_{\lambda_K}(\theta_{[K]}^{(0)}) \leq \lambda_K/2$ with $\theta_{[K]}^{(0)} \in \mathcal{B}_{r^*+\tilde{s}}$, then we can guarantee the geometric convergence rate of the estimated objective value towards the minimal objective. Next, we show that if the optimal solution $\hat{\theta}_{[K-1]}$ from $K - 1$-th path following stage satisfies the approximate KKT condition and the regularization parameter $\lambda_K$ in the $K$-th path following stage is chosen properly, then $\hat{\theta}_{[K-1]}$ satisfies the approximate KKT condition for $\lambda_K$ with a slightly larger bound. This characterizes that good computational properties are preserved by using the warm start $\theta_{[K]}^{(0)} = \hat{\theta}_{[K-1]}$ and geometric sequence of regularization parameters $\lambda_K$. We formalize this notion in Lemma B.3, and its proof is deferred to Appendix H.

**Lemma B.3.** Let $\hat{\theta}_{[K-1]}$ be the approximate solution of $K - 1$-th path following state, which satisfies the approximate KKT condition $\omega_{\lambda_{K-1}}(\hat{\theta}_{[K-1]}) \leq \lambda_{K-1}/4$. Then we have

$$
\omega_{\lambda_K}(\hat{\theta}_{[K-1]}) \leq \lambda_K/2,
$$

where $\lambda_K = \eta \lambda_{K-1}$ with $\eta \in (5/6, 1)$. 

19
Combining Theorem (B.2) and Lemma (B.3), we can achieve the global convergence in terms of the objective value using the path following proximal gradient method. We have the bounds of iterations $T_K$ in Phase 2 directly from (B.4) and (B.5) of Theorem B.2.

Finally, we obtain the objective gap of $N$-th stage via analogous argument. Specifically, in the $N$-th (final) path following stage, when the number of iterations for proximal method is large enough such that $\omega_{\lambda N}(\theta_{[N]}^{(t)}) \leq \epsilon_N$ holds, then we obtain the result from Lemma A.5 with $\lambda = \tilde{\lambda} = \lambda_N$.

Finally, we need to show that there exists some $N_1 \in \{1, \ldots, N\}$ such that $\hat{\theta}_{[N_1]} \in B^*_r + \tilde{s}$. We demonstrate this result in Lemma B.4 and provide its proof in Appendix I.

**Lemma B.4.** Suppose Assumption 3.2 and Assumption 3.3 holds, and the approximate KKT satisfies $\omega_{\lambda}(\theta) \leq \lambda/4$. If $\mu \leq \sqrt{n}\sigma/4$ and $\frac{\rho_8 + \tilde{s}}{\sigma} > \lambda > \lambda_N$, then we have

$$\|\theta - \theta^*\|_2 \leq r \quad \text{and} \quad \|\theta_{\tilde{s}}\|_0 \leq \tilde{s}.$$ 

Lemma B.4 guarantees that there exits some $N_1 < N$ such that for some $\lambda_{N_1} > \lambda_N$, the approximate solution $\hat{\theta}_{[N_1]}$ satisfies the approximate KKT condition and $\hat{\theta}_{[N_1]} \in B^*_r + \tilde{s}$, then we enters the Phase 2 of strong linear convergence by Theorem B.2. Thus we finish the proof.

We further provide a bound of the total number of proximal gradient steps for each $\lambda_K$ in the following lemma for interested readers. The proof is provided in Appendix J.

**Lemma B.5.** For each $\lambda_K$, $K = 1, \ldots, N$, if we restart the line search with a large enough $L_{\text{max}}$, then the total number of proximal gradient steps is no more than $2(T_K + 1) + \max\{\log_2 L_{\text{max}} - \log_2 \rho_N^k + 2s\}, 0}.

### C Proof of Lemma 3.6

**Part 1.** We first show that Assumption 3.2 holds. By $y = X\theta^* + \epsilon$ and (C.4), we have

$$\nabla L_{\mu}(\theta^*) = \frac{X^T(X\theta^* - y)}{\max\{\sqrt{n}\mu, \sqrt{n}\|y - X\theta^*\|_2\}} = -\frac{X^T\epsilon}{\max\{\sqrt{n}\mu, \sqrt{n}\|\epsilon\|_2\}}. \quad (C.1)$$

Since $\epsilon$ has i.i.d. sub-Gaussian entries and $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i^2] = \sigma^2$ for all $i = 1, \ldots, n$, then we have from Wainwright (2015) that

$$\mathbb{P}\left[\|\epsilon\|_2^2 \leq \frac{1}{4} n\sigma^2\right] \leq \exp\left(-\frac{n}{32}\right), \quad (C.2)$$

By Negahban et al. (2012), we have the following result.

**Lemma C.1.** Assume $X$ satisfies $\|x_j\|_2 \leq \sqrt{n}$ for all $j = 1, \ldots, d$ and $\epsilon$ has i.i.d. zero-mean sub-Gaussian entries with $\mathbb{E}[w_i^2] = \sigma^2$ for all $i = 1, \ldots, n$, then we have

$$\mathbb{P}\left[\frac{1}{n}\|X^T\epsilon\|_\infty \geq 2\sigma\sqrt{\frac{\log d}{n}}\right] \leq 2d^{-1}.$$
Combining (C.1), (C.2) and Lemma C.1, we have with probability at least $1 - 2d^{-1} - \exp\left(-\frac{n}{32}\right)$, 
\[
\|\nabla L_\mu(\theta^*)\|_\infty \leq \frac{4 \sqrt{\log d/n}}{\max\{\sqrt{2} \mu/\sqrt{n} \sigma, 1\}}.
\]

**Part 2.** Next, we show that Assumption 3.3 holds. We divide the proof into two steps.

**Step 1.** When $X$ satisfies the RE condition, i.e.,
\[
\psi_{\min} \|v\|_2^2 - \varphi_{\min} \frac{\log d}{n}\|v\|_1^2 \leq \frac{\|Xv\|_2^2}{n} \leq \psi_{\max} \|v\|_1^2 + \varphi_{\max} \frac{\log d}{n}\|v\|_1^2,
\]

Denote $s = s^* + 2\tilde{s}$. Since $\|v\|_0 \leq s$, which implies $\|v\|_1^2 \leq s\|v\|_2^2$, then we have
\[
\left(\psi_{\min} - \varphi_{\min} \frac{s \log d}{n}\right) \|v\|_2^2 \leq \frac{\|Xv\|_2^2}{n} \leq \left(\psi_{\max} + \varphi_{\max} \frac{s \log d}{n}\right) \|v\|_2^2.
\]

Then there exists a universal constant $c_1$ such that if $n \geq c_1 s^* \log d$, we have
\[
\frac{1}{2} \psi_{\min} \|v\|_2^2 \leq \frac{\|Xv\|_2^2}{n} \leq 2 \psi_{\max} \|v\|_2^2.
\]  
(C.3)

**Step 2.** Conditioning on (C.3), we show that $L_\mu$ satisfies LRSC and LRSS with high probability. The gradient of $L_\mu(\theta)$ is

\[
\nabla L_\mu(\theta) = \frac{1}{\sqrt{n}} \begin{pmatrix}
\partial \|y - X\theta\|_\mu \\
\partial (y - X\theta)
\end{pmatrix}^\top \begin{pmatrix}
\partial (y - X\theta) \\
\partial \theta
\end{pmatrix}^\top = \frac{X^\top (X\theta - y)}{\max\{\sqrt{n} \mu, \sqrt{n} \|y - X\theta\|_2\}}.
\]  
(C.4)

The Hessian of $L_\mu(\theta)$ is

\[
\nabla^2 L_\mu(\theta) = \frac{1}{n} \frac{\partial (X^\top z)}{\partial \theta} = \begin{cases}
\frac{X^\top X}{\sqrt{n} \mu}, & \text{if } \|y - X\theta\|_2 < \mu \\
\frac{1}{\sqrt{n} \|y - X\theta\|_2} X^\top \left(I - \frac{(y - X\theta)(y - X\theta)^\top}{\|y - X\theta\|_2^2}\right) X, & \text{o.w.}
\end{cases}
\]  
(C.5)

For notational convenience, we define $\Delta = v - w$ for any $v, w \in \mathcal{B}_s^\ast$. Also denote the residual of the first order Taylor expansion as

\[
\delta L_\mu(w + \Delta, w) = L_\mu(w + \Delta) - L_\mu(w) - \nabla L_\mu(w)^\top \Delta.
\]

Using the first order Taylor expansion of $L_\mu(\theta)$ at $w$ and the Hessian of $L_\mu(\theta)$ in (C.5), we have from mean value theorem that there exists some $\alpha \in [0, 1]$ such that

\[
\delta L_\mu(w + \Delta, w) = \begin{cases}
\frac{\Delta^\top X^\top X \Delta}{\sqrt{n} \mu}, & \text{if } \|\xi\|_2 < \mu \\
\frac{1}{\sqrt{n} \|\xi\|_2} \Delta^\top X^\top \left(I - \frac{\xi \xi^\top}{\|\xi\|_2^2}\right) X \Delta, & \text{o.w.}
\end{cases}
\]

where $\xi = y - X(w + \alpha \Delta)$. For notational simplicity, let’s denote $\hat{z} = X(v - \theta^*)$ and $\tilde{z} = X(w - \theta^*)$, which can be considered as two fixed vectors in $\mathbb{R}^n$. Without loss of generality, assume $\|\hat{z}\|_2 \leq \|\tilde{z}\|_2$. Then we have

\[
\|\hat{z}\|_2^2 \leq \|\tilde{z}\|_2^2 \leq 2 \psi_{\max} n \|w - \theta^*\|_2^2 \leq \frac{n \sigma^2}{4}.
\]
Further, we have
\[\xi = y - X(w + \alpha \Delta) = \epsilon - X(w + \alpha \Delta - \theta^*) = \epsilon - \alpha \dot{z} - (1 - \alpha) \ddot{z},\] and \(X\Delta = \dot{z} - \ddot{z}\).

We have from Wainwright (2015) that
\[
P \left[ \|\epsilon\|_2^2 \leq n\sigma^2 (1 - \delta) \right] \leq \exp \left( -\frac{n\delta^2}{16} \right), \tag{C.6}
\]
Then by taking \(\delta = 1/3\) in (C.6), we have with probability \(1 - \exp \left( -\frac{n}{144} \right)\),
\[
\|\xi\|_2 \geq \|\epsilon\|_2 - \alpha \|\dot{z}\|_2 - (1 - \alpha) \|\ddot{z}\|_2 \geq \frac{4}{5} \sqrt{n\sigma} - \frac{1}{2} \sqrt{n\sigma} \geq \frac{1}{4} \sqrt{n\sigma}. \tag{C.7}
\]
We first discuss the RSS property. From (C.7), we have
\[
\delta L_\mu(w + \Delta, w) = \frac{1}{\sqrt{n}\|\xi\|_2} \Delta^T X^T \left( I - \frac{\xi \xi^T}{\|\xi\|_2^2} \right) X \Delta = \frac{1}{\sqrt{n}\|\xi\|_2} \left( \|X\Delta\|^2 - \frac{(\xi^T X \Delta)^2}{\|\xi\|_2^2} \right)
\leq \frac{\|X\Delta\|^2}{\sqrt{n}\|\xi\|_2^2} \leq \frac{8\psi_{max}}{\sigma} \|\Delta\|_2^2.
\]
Next, we verify the RSC property. From (C.7), we have \(\|\xi\|_2 \geq \mu\). We want to show that with high probability, for any constant \(a \in (0, 1)\)
\[
\left| \frac{\xi^T X \Delta}{\|\xi\|_2} \right| \leq \sqrt{1 - a} \|X\Delta\|_2. \tag{C.8}
\]
Consequently, we have
\[
\Delta^T X^T \left( I - \frac{\xi \xi^T}{\|\xi\|_2^2} \right) X \Delta = \|X\Delta\|^2 - \left( \frac{\xi^T X \Delta}{\|\xi\|_2} \right)^2 \geq a \|X\Delta\|^2_2.
\]
This further implies
\[
\delta L_\mu(w + \Delta, w) = \frac{1}{\sqrt{n}\|\xi\|_2} \Delta^T X^T \left( I - \frac{\xi \xi^T}{\|\xi\|_2^2} \right) X \Delta \geq \frac{a\psi_{min}}{2\|\xi\|_2/\sqrt{n}} \|\Delta\|_2^2. \tag{C.9}
\]
Since \(\|\dot{z}\|_2 \leq \|\ddot{z}\|_2\), then for any real constant \(a \in (0, 1)\),
\[
P \left[ \left| \frac{\xi^T X \Delta}{\|\xi\|_2} \right| \leq \sqrt{1 - a} \|X\Delta\|_2 \right]
= P \left[ \left| \frac{(\epsilon - \alpha \dot{z} - (1 - \alpha) \ddot{z})^T}{\|\epsilon - \alpha \dot{z} - (1 - \alpha) \ddot{z}\|_2} (\dot{z} - \ddot{z}) \right| \leq \sqrt{1 - a} \|\dot{z} - \ddot{z}\|_2 \right] \geq (i) P \left[ \left| \frac{(\epsilon - \ddot{z})^T (\ddot{z} - \ddot{z})}{\|\epsilon - \ddot{z}\|_2} \right| \leq \sqrt{1 - a} \|\dot{z} - \ddot{z}\|_2 \right]
= P \left[ \left| \frac{(\epsilon^T (\ddot{z} - \ddot{z}) - (\dot{z} - \ddot{z})}{\|\ddot{z} - \ddot{z}\|_2} \right|^2 \leq (1 - a) \|\epsilon - \ddot{z}\|_2^2 \|\ddot{z} - \ddot{z}\|_2^2 \right]
\leq (ii) P \left[ \left| \frac{(\epsilon^T (\ddot{z} - \ddot{z})}{\|\ddot{z} - \ddot{z}\|_2} \right|^2 + \|\ddot{z}\|^2_2 - 2\epsilon^T \ddot{z} \right| \leq (1 - a)(\|\epsilon\|_2^2 + \|\ddot{z}\|^2_2 - 2\epsilon^T \ddot{z}) \right], \tag{C.10}
\]
22
where (ii) is from dividing both sides by $\|v\|_2^2$, and (i) is from a geometric inspection and the randomness of $e$, i.e., for any $\alpha \in [0, 1]$ and $\|\dot{z}\|_2 \leq \|\ddot{z}\|_2$,

$$\left| \frac{-\ddot{z}^T}{\|\ddot{z}\|_2} (\dot{z} - \ddot{z}) \right| \leq \left| \frac{-(\alpha \ddot{z} - (1 - \alpha)\dot{z})^T}{\|\ddot{z}^T - (1 - \alpha)\dot{z}\|_2^2} (\dot{z} - \ddot{z}) \right|.$$  

The random vector $e$ with i.i.d. entries does not affect the inequality above. Let's first discuss one side of the probability in (C.10), i.e.,

$$\mathbb{P} \left[ \left( \frac{e^T (\dot{z} - \ddot{z})}{\|\dot{z} - \ddot{z}\|_2} \right)^2 + \|\ddot{z}\|_2^2 - 2e^T \dot{z} \leq (1 - a)(\|e\|_2^2 + \|\dot{z}\|_2^2 - 2e^T \dot{z}) \right]$$  

$$= \mathbb{P} \left[ (1 - a)\|e\|_2^2 \geq \left( \frac{e^T (\dot{z} - \ddot{z})}{\|\dot{z} - \ddot{z}\|_2} \right)^2 + a(\|\dot{z}\|_2^2 - 2e^T \dot{z}) \right]. \quad \text{(C.11)}$$

Since $e$ has i.i.d. sub-Gaussian entries with $\mathbb{E}[e_i] = 0$ and $\mathbb{E}[e_i^2] = \sigma^2$ for all $i = 1, \ldots, n$, then $\frac{e^T (\dot{z} - \ddot{z})}{\|\dot{z} - \ddot{z}\|_2}$ and $e^T \dot{z}$ are also zero-mean sub-Gaussians with variances $\sigma^2$ and $\sigma^2 \|\dot{z}\|_2^2$ respectively. We have from Wainwright (2015) that

$$\mathbb{P} \left[ \|e\|_2^2 \leq n\sigma^2(1 - \delta) \right] \leq \exp \left( -\frac{n\delta^2}{16} \right), \quad \text{(C.12)}$$

$$\mathbb{P} \left[ \left( \frac{e^T (\dot{z} - \ddot{z})}{\|\dot{z} - \ddot{z}\|_2} \right)^2 \geq n\sigma^2 \delta^2 \right] \leq \exp \left( -\frac{n\delta^2}{2} \right), \quad \text{(C.13)}$$

$$\mathbb{P} \left[ e^T \dot{z} \leq -n\sigma^2 \delta \right] \leq \exp \left( -\frac{n^2 \sigma^2 \delta^2}{2\|\dot{z}\|_2^2} \right). \quad \text{(C.14)}$$

Combining (C.12) – (C.14) with $\|\ddot{z}\|_2^2 \leq n\sigma^2/4$, we have from union bound that with probability at least $1 - \exp \left( -\frac{n}{144} \right) - \exp \left( -\frac{n}{128} \right) - \exp \left( -\frac{n}{128} \right) \geq 1 - 3\exp \left( -\frac{n}{144} \right)$,

$$\|e\|_2^2 \geq \frac{2}{3} n\sigma^2, \quad \left( \frac{e^T (\dot{z} - \ddot{z})}{\|\dot{z} - \ddot{z}\|_2} \right)^2 \leq \frac{1}{64} n\sigma^2, \quad -e^T \dot{z} \leq \frac{1}{16} n\sigma^2.$$ 

This implies for $a \leq 3/5$, we have

$$\xi^T \frac{X\Delta}{\|\xi\|_2} \leq \sqrt{1 - a}\|X\Delta\|_2.$$ 

For the other side of the probability in (C.10), we have

$$\mathbb{P} \left[ \left( \frac{e^T (\dot{z} - \ddot{z})}{\|\dot{z} - \ddot{z}\|_2} \right)^2 + \|\ddot{z}\|_2^2 - 2e^T \dot{z} \geq -(1 - a)(\|e\|_2^2 + \|\dot{z}\|_2^2 - 2e^T \dot{z}) \right]$$

$$= \mathbb{P} \left[ (1 - a)\|e\|_2^2 \geq -\left( \frac{e^T (\dot{z} - \ddot{z})}{\|\dot{z} - \ddot{z}\|_2} \right)^2 - (2 - a)(\|\dot{z}\|_2^2 - 2e^T \dot{z}) \right]$$

$$\geq \mathbb{P} \left[ (1 - a)\|e\|_2^2 \geq \left( \frac{e^T (\dot{z} - \ddot{z})}{\|\dot{z} - \ddot{z}\|_2} \right)^2 + a(\|\dot{z}\|_2^2 - 2e^T \dot{z}) \right]. \quad \text{(C.15)}$$
Combining (C.10), (C.11) and (C.15), we have (C.8) holds with high probability, i.e., for any \( r > 0 \),
\[
\mathbb{P} \left[ \frac{\xi^T}{\|\xi\|_2} X \Delta \leq \sqrt{1-a\|X \Delta\|_2} \right] \geq 1 - 6 \exp\left(-\frac{n}{144}\right).
\]

Now we bound \( \|\xi\|_2 \) to obtain the desired result. From Wainwright (2015), we have
\[
\mathbb{P} \left[ \|e\|_2^2 \geq n\sigma^2(1+\delta) \right] \leq \exp\left(-\frac{n\delta^2}{18}\right) = \exp\left(-\frac{n}{72}\right),
\]
where we take \( \delta = 1/2 \). From \( \xi = \epsilon - \alpha \tilde{z} - (1-\alpha)\hat{z} \), we have
\[
\|\xi\|_2 \leq \|e\|_2 + \alpha \|\tilde{z}\|_2 + (1-\alpha)\|\hat{z}\|_2 \leq \|e\|_2 + \|\hat{z}\|_2 \leq \sqrt{\frac{3n}{2}\sigma + \frac{1}{2}\sqrt{n}\sigma}.
\]
where (i) is from \( \|\hat{z}\|_2 \leq \|\tilde{z}\|_2 \) and (ii) is from (C.16) and \( \|\tilde{z}\|^2 \leq n\sigma^2/4 \). Then by the union bound setting \( a = 1/2 \), with probability at least \( 1 - 7 \exp\left(-\frac{n}{144}\right) \), we have
\[
\delta \mathcal{L}_\mu(w + \Delta, w) \geq \frac{\psi_{\min}}{8\sigma}\|\Delta\|^2_2.
\]
Moreover, we also have \( r = \frac{\sigma^2}{8\psi_{\max}} > s^* (64\sigma \lambda_{N1}/\psi_{\min})^2 \geq s^* \left(8\lambda_{N1}/\rho_{s^*+\bar{s}}\right)^2 \) for large enough \( n \geq c_1 s^* \log d \), where \( \lambda_{N1} \geq 2\lambda_N = 48\sqrt{\log d/n} \). The choice of the constant “2” in \( \lambda_{N1} \geq 2\lambda_N \) is somewhat arbitrary, which can be any fixed constant larger than \( 1/\eta \) such that the existence of \( \lambda_{N1} \) is guaranteed.

\section{Proof of Theorem 3.7}

**Part 1.** We first show that estimation errors are as claimed. Since \( \hat{\theta}_{[K]} \) is the approximate solution of \( K \)-th path following stage, it satisfies \( \omega_{\lambda_K}(\hat{\theta}_{[K]}) \leq \lambda_K/4 \leq \lambda_{K+1}/2 \) for \( t \in [N_1 + 1, T - 1] \), then we have from Lemma B.3 that
\[
\omega_{\lambda_{K+1}}(\hat{\theta}_{[K+1]}) \leq \lambda_{K+1}/2.
\]
By Theorem B.2, we have for any \( t = 1, 2, \ldots \),
\[
\|\theta_{[t]} - \theta^*\|_0 \leq \bar{s}.
\]
Applying Lemma A.2 recursively, we have
\[
\|\hat{\theta}_{[N]} - \theta^*\|_2 \leq \frac{2\lambda_{N} \sqrt{s^*}}{P_{s^*+2\bar{s}}} \quad \text{and} \quad \|\hat{\theta}_{[N]} - \theta^*\|_1 \leq \frac{12\lambda_{N}s^*}{P_{s^*+2\bar{s}}}.
\]
Applying Lemma 3.6 with \( \lambda_N = 24\sqrt{\log d/n} / n \) and \( \rho_{s^*+2\bar{s}} = \frac{\psi_{\min}}{8\sigma} \), then by union bound, with probability at least \( 1 - 8 \exp\left(-\frac{n}{144}\right) - 2d^{-1} \), we have
\[
\|\hat{\theta}_{[N]} - \theta^*\|_2 \leq \frac{384\sigma \sqrt{s^* \log d/n}}{\psi_{\min}}, \quad \|\hat{\theta}_{[N]} - \theta^*\|_1 \leq \frac{2304\sigma s^* \sqrt{\log d/n}}{\psi_{\min}}.
\]
Part 2. Next, we demonstrate the result of the estimation of variance. Let \( \hat{\theta}_N = \text{argmin}_\theta F_{\mu, \lambda N}(\theta) \) be the optimal solution of \( K \)-th stage. Apply the argument in Part recursively, we have
\[
\|\hat{\theta}_N - \theta^*\|_1 \leq \frac{2304\sigma s^* \sqrt{\log d/n}}{\psi_{\min}}. \tag{D.1}
\]

Denote \( c_1, c_2, \ldots \) as positive universal constants. Then we have
\[
L_{\mu}(\hat{\theta}_N) - L_{\mu}(\theta^*) \leq \lambda_N \|\theta^*\|_1 - \|\hat{\theta}_N\|_1 \leq \lambda_N \left( \|\theta^*_S\|_1 - \|\hat{\theta}_N S^*\|_1 - \|\hat{\theta}_N - \theta^*_S\|_1 \right)
\leq \lambda_N \|\hat{\theta}_N - \theta^*\|_{S^*} \leq \lambda_N \|\hat{\theta}_N - \theta^*\|_1 \leq \frac{c_1 \sigma s^* \log d}{n}, \tag{D.2}
\]
where (i) is from the value of \( \lambda_N \) and \( \ell_1 \) error bound in (D.1).

On the other hand, from the convexity of \( L_{\mu}(\theta) \), we have
\[
L_{\mu}(\hat{\theta}_N) - L_{\mu}(\theta^*) \geq \left( \hat{\theta}_N - \theta^* \right)^\top \nabla L_{\mu}(\theta^*) \geq -\|\nabla L_{\mu}(\theta^*)\|_\infty \|\hat{\theta}_N - \theta\|_1
\geq -c_2 \lambda_N \|\hat{\theta}_N - \theta\|_1 \geq -c_3 \frac{s^* \log d}{n}, \tag{D.3}
\]
where (i) is from Assumption 3.2 and (ii) value of \( \lambda_N \) and \( \ell_1 \) error bound in (D.1).

For our choice of \( \mu \) and \( n \), we have \( L_{\mu}(\theta) = \frac{1}{\sqrt{n}} \| y - X\theta \|_2 - \frac{1}{2} \) by Proposition 3.8, then
\[
L_{\mu}(\hat{\theta}_N) - L_{\mu}(\theta^*) = \frac{\| y - X\hat{\theta}_N \|_2}{\sqrt{n}} - \frac{\| \epsilon \|_2}{\sqrt{n}}. \tag{D.4}
\]

From Wainwright (2015), we have for any \( \delta > 0 \),
\[
\mathbb{P} \left[ \left| \frac{\| \epsilon \|_2^2}{n} - \sigma^2 \right| \leq \sigma^2 \delta \right] \leq 2 \exp \left( -\frac{n\delta^2}{18} \right). \tag{D.5}
\]

Combining (D.2), (D.3), (D.4) and (D.5) with \( \delta^2 = \frac{c_3 s^* \log d}{n} \), we have with high probability,
\[
\left| \frac{\| y - X\hat{\theta}_N \|_2}{\sqrt{n}} - \sigma \right| = O \left( \frac{\sigma s^* \log d}{n} \right). \tag{D.6}
\]

From Part 1, for \( n \geq c_4 s^* \log d \), we have with high probability,
\[
\|\hat{\theta}_N - \theta^*\|_2 \leq \frac{384 \sigma \sqrt{s^* \log d/n}}{\psi_{\min}} \leq \frac{\sigma}{2\sqrt{2\psi_{\max}}},
\]
then \( \hat{\theta}_N \in B_{r^* + \tilde{s}} \) and \( \|\hat{\theta}_N - \bar{\theta}_N\|_0 \leq s^* + 2\tilde{s} \). Then from the analysis of Theorem B.2, we have
\[
\omega_{\lambda_K}(\theta_{[K]}^{(t+1)}) \leq (1 + \kappa_{s^* + 2\tilde{s}}) \sqrt{4 \rho_{s^* + 2\tilde{s}}^+ \left( F_{\mu, \lambda_K}(\theta_{[K]}^{(t)}) - F_{\mu, \lambda_K}(\hat{\theta}_{[K]}) \right)} \leq \varepsilon_N.
\]

This implies
\[
F_{\mu, \lambda_K}(\theta_{[K]}^{(t)}) - F_{\mu, \lambda_K}(\hat{\theta}_{[K]}) \leq \frac{\varepsilon_N^2}{4 \rho_{s^* + 2\tilde{s}}^+ (1 + \kappa_{s^* + 2\tilde{s}})^2}. \tag{D.7}
\]
On the other hand, from the LRSC property of \( L_\mu \), convexity of \( \ell_1 \) norm and optimality of \( \bar{\theta} \), we have

\[
\mathcal{F}_{\mu, \lambda K}(\theta^{(t)}_{[K]}) - \mathcal{F}_{\mu, \lambda K}(\bar{\theta}_{[K]}) \geq \rho_{s^*+2\bar{s}}\| \hat{\theta}_{[N]} - \bar{\theta}_{[N]} \|^2.
\]  

(D.8)

Combining (D.7), (D.8) and Assumption 3.5, we have

\[
\frac{\|X(\hat{\theta}_{[N]} - \bar{\theta}_{[N]})\|_2}{\sqrt{n}} \leq \sqrt{\frac{8\rho_{s^*+2\bar{s}}^2}{\sigma}} \| \hat{\theta}_{[N]} - \theta^* \|_2 \leq \frac{2}{\sigma \rho_{s^*+2\bar{s}}(1 + \kappa_{s^*+2\bar{s}})} \epsilon_N
\]

\[
\leq \frac{4\epsilon_N}{(1 + \kappa_{s^*+2\bar{s}}) \sqrt{\psi_{\min}}}.
\]  

(D.9)

Combining (D.6) and (D.9), we have

\[
\left| \left| \frac{y - X\hat{\theta}_{[N]}}{\sqrt{n}} \right| \right|_2 \leq \left| \left| \frac{y - X\bar{\theta}_{[N]}}{\sqrt{n}} \right| \right|_2 + \left| \left| \frac{X(\hat{\theta}_{[N]} - \bar{\theta}_{[N]})}{\sqrt{n}} \right| \right|_2 \leq \left| \left| \frac{y - X\bar{\theta}_{[N]}}{\sqrt{n}} \right| \right|_2 + \frac{4\epsilon_N}{(1 + \kappa_{s^*+2\bar{s}}) \sqrt{\psi_{\min}}}.
\]

If \( \epsilon_N \leq c_5 \frac{s^* \log d}{n} \) for some constant \( c_5 \), then we have the desired result.

**E  Proof of Proposition 3.8**

Let \( \tilde{\theta}_{[K]} \) and \( \bar{\theta}_{[K]} \) be the unique global optima of (1.3) and (1.4) respectively for all \( K = N_1 + 1, \ldots, N \). Then, we show \( \tilde{\theta}_{[K]} = \bar{\theta}_{[K]} \) under the proposed conditions. Apply the argument of the proof of Theorem 3.7 recursively, we have with probability at least \( 1 - 8 \exp \left( - \frac{n}{144} \right) - 2d^{-1} \),

\[
\| \bar{\theta}_{[K]} - \theta^* \|^2 \leq \frac{384\sigma \sqrt{s^* \log d / n}}{\psi_{\min}}.
\]

By SE condition of \( X \) in Assumption 3.3, this implies

\[
\| X(\bar{\theta}_{[K]} - \theta^*) \|^2 \leq \frac{384\sigma \sqrt{2\psi_{\max} s^* \log d}}{\psi_{\min}}.
\]  

(E.1)

On the other hand, we have

\[
\| y - X\bar{\theta}_{[K]} \|^2 = \| X(\bar{\theta}_{[K]} - \theta^*) \| + \| \epsilon \|^2 \geq \| \epsilon \|^2 - \| X(\bar{\theta}_{[K]} - \theta^*) \|^2.
\]  

(E.2)

Since \( \epsilon \) has i.i.d. sub-Gaussian entries with \( \mathbb{E}[\epsilon_i] = 0 \) and \( \mathbb{E}[\epsilon_i^2] = \sigma^2 \) for all \( i = 1, \ldots, n \), we have from Wainwright (2015) that

\[
\mathbb{P} \left[ \| \epsilon \|^2 \leq \frac{2}{3} n \sigma^2 \right] \leq \exp \left( - \frac{n}{144} \right),
\]

(E.3)
Combining (E.1) and (E.2), (E.3) and the condition on \( n \geq c_4 s^* \log d \) for some constant \( c_4 \), we have with probability at least \( 1 - 9\exp\left(-\frac{n}{144}\right) - 2d^{-1} \),

\[
\|y - X\hat{\theta}_{[K]}\|_2 \geq \sqrt{n}\sigma \left( \sqrt{\frac{2}{3}} - \frac{384\sigma \sqrt{2\psi_{\max} s^* \log d/n}}{\psi_{\min}} \right) > \frac{\sqrt{n}\sigma}{4} \geq \mu.
\]

This implies \( F_{\mu,\lambda}(\theta) = F_{\mu}(\theta) + \frac{\sigma^2}{s^*} \), thus \( \arg\min_{\theta} F_{\mu,\lambda}(\theta) = \arg\min_{\theta} F_{\mu}(\theta) \), i.e., \( \hat{\theta}_T = \tilde{\theta}_{[K]} \). Besides this also implies \( \|y - X\theta^*\|_2 \geq \frac{\sqrt{n}\sigma}{4} \geq \mu \), i.e., \( \theta^* \) is not in the smoothed region.

Applying the same argument again to \( \hat{\theta}_{[K]} \), we have that for large enough \( n \), with high probability,

\[
\|y - X\hat{\theta}_{[K]}\|_2 > \frac{\sqrt{n}\sigma}{4} \geq \mu,
\]

and \( r = \frac{\sigma^2}{s^* \psi_{\max}} > s^*(64\sigma_1 \lambda \sqrt{n} / \psi_{\min})^2 \geq s^* \left( 8\lambda \sqrt{n} / \rho \right)^2 \) is guaranteed, where \( \lambda \geq 2\lambda_N = 48\sqrt{\log d/n} \). By Lemma B.4, this implies the existence of the linear convergence region, which does not fall into the smoothed region. Besides, \( \hat{\theta}_{[K]} \) is not in the smoothed region.

F Proof of Theorem B.1

The sub-linear rate of convergence (B.1) follows directly from Lemma A.7. In terms of the optimal residual, we have

\[
\omega_{\delta_k}^2(\theta^{(t+1)}_{[K]}) \leq \left( L_{[K]}^{(t)} + S_{L_{[K]}^{(t)}}(\theta^{(t)}_{[K]}) \right)^2 \|\theta^{(t+1)}_{[K]} - \hat{\theta}^{(t)}_{[K]}\|_2^2
\]

\[
\leq \left( \frac{\|X\|_2^2}{\sqrt{n}\mu} \right)^2 \|\theta^{(t+1)}_{[K]} - \hat{\theta}^{(t)}_{[K]}\|_2^2
\]

\[
\leq \frac{2 \left( \frac{\|X\|_2^2}{\sqrt{n}\mu} \right)^2}{(k - m + 1)} \cdot \left( \sum_{i=m}^{k} F_{\mu,\lambda}(\theta^{(m)}_{[K]}) - F_{\mu,\lambda}(\theta^{(m+1)}_{[K]}) \right)
\]

\[
\leq \frac{18 \|X\|_2^4}{(k - m + 1)n\mu^2} \cdot \frac{F_{\mu,\lambda}(\theta^{(m)}_{[K]}) - F_{\mu,\lambda}(\theta^{(m+1)}_{[K]})}{L_{\mu}}
\]

\[
\leq \frac{18 \|X\|_2^4}{(k - m + 1)n\mu^2} \cdot \frac{F_{\mu,\lambda}(\theta^{(m)}_{[K]}) - F_{\mu,\lambda}(\theta^{(m+1)}_{[K]})}{L_{\mu}}
\]

\[
\leq \frac{288R^2\|X\|_2^6}{L_{\mu}(k - m + 1)(m + 2)n^{3/2}\mu^3} \leq 288R^2\|X\|_2^6.
\]

where (i) is from Lemma A.9, (ii) is from \( S_{L_{[K]}^{(t)}}(\theta^{(t)}_{[K]}) \leq L_{\mu} \leq \|X\|_2^2/\sqrt{n}\mu \) in Lemma A.9 and Lemma A.7, (iii) is from (A.16) in Lemma A.8, (iv) is from \( L_{\mu} \leq L_{[K]}^{(t)} \leq 2L_{\mu} \leq 2\|X\|_2^2/\sqrt{n}\mu \) in Remark A.1 and Lemma 3.6, (v) is from Lemma A.7 and (vi) is obtained by choosing \( m = \lfloor k/2 \rfloor \). To achieve the approximate KKT condition \( \omega_{\mu,\lambda}(\theta^{(t)}_{[K]}) \leq \varepsilon_K \), we require the R.H.S. of (F.1) to be no greater than \( \varepsilon^2 K \), then we have the desired result (B.2).
G Proof of Theorem B.2

Note that the RSS property implies that line search terminate when \( \tilde{L}(t) \) satisfies

\[
\rho^+_{s^*+2\bar{s}} \leq \tilde{L}(t) \leq 2\rho^+_{s^*+2\bar{s}}. \tag{G.1}
\]

Since the initialization \( \theta^{(0)}_{[K]} \) satisfies \( \omega_{\lambda_K}(\theta^{(0)}_{[K]}) \leq \frac{\lambda_K}{2} \) with \( \|\theta^{(0)}_{[K]} - \theta^*\|_0 \leq \bar{s} \), then by Lemma A.2, the objective satisfies

\[
F_{\mu,\lambda_K}(\theta^{(0)}_{[K]}) - F_{\mu,\lambda}(\theta^*) \leq \frac{6\lambda^2_K s^*}{\rho_{s^*+2\bar{s}}}.
\]

Then by Lemma A.4, we have

\[
\|\theta^{(1)}_{[K]} - \theta^*\|_0 \leq \bar{s}.
\]

By monotone decrease of \( F_{\mu,\lambda_K}(\theta^{(t)}_{[K]}) \) from (A.16) in Lemma A.8 and recursively applying Lemma A.4, \( \|\theta^{(t)}_{[K]} - \theta^*\|_0 \leq \bar{s} \) holds in (B.3) for any \( t = 1, 2, \ldots \).

For the objective error, we have

\[
F_{\mu,\lambda_K}(\theta^{(t)}_{[K]}) - F_{\mu,\lambda}(\theta^{(t)}_{[K]}) \leq \left(1 - \frac{1}{8\kappa_{s^*+2\bar{s}}}\right)^t \left( F_{\mu,\lambda_K}(\theta^{(0)}_{[K]}) - F_{\mu,\lambda}(\theta^{(0)}_{[K]}) \right)
\]

\[
\leq \left(1 - \frac{1}{8\kappa_{s^*+2\bar{s}}}\right)^t \frac{24\lambda_K s^* \omega_{\lambda_K}(\theta^{(t)}_{[K]})}{\rho_{s^*+2\bar{s}}}, \tag{G.2}
\]

where (i) is from Lemma A.6, and (ii) is from Lemma A.5 with \( \bar{\lambda} = \lambda = \lambda_K \) and \( \omega_{\lambda_K}(\theta^{(t+1)}_{[K]}) \leq \lambda_K/2 \leq \lambda_K \), which results in (B.3).

Combining (G.2) and (A.4), we have

\[
\|\theta^{(t)}_{[K]} - \bar{\theta}_{[K]}\|^2 \leq \frac{1}{\rho_{s^*+2\bar{s}}} \left( F_{\mu,\lambda_K}(\theta^{(t)}_{[K]}) - F_{\mu,\lambda_K}(\bar{\theta}_{[K]}) - \nabla F_{\mu,\lambda_K}(\bar{\theta}_{[K]}) \right)
\]

\[
= \frac{1}{\rho_{s^*+2\bar{s}}} \left( F_{\mu,\lambda_K}(\theta^{(t)}_{[K]}) - F_{\mu,\lambda_K}(\bar{\theta}_{[K]}) \right) \leq \left(1 - \frac{1}{8\kappa_{s^*+2\bar{s}}}\right)^t \frac{24\lambda_K s^* \omega_{\lambda_K}(\theta^{(t)}_{[K]})}{(\rho_{s^*+2\bar{s})^2}}
\]

28
For the optimal residue $\omega_{\lambda K}(\theta_{[K]}^{(t+1)})$ of $(t+1)$-th iteration of $K$-th path following stage, we have

$$\omega_{\lambda K}(\theta_{[K]}^{(t+1)}) \overset{(i)}{\leq} \left( \tilde{L}_{[K]}^{(t)} + S_{L_{[K]}^{(t)}}(\theta_{[K]}^{(t)}) \right) \|\theta_{[K]}^{(t+1)} - \theta_{[K]}^{(t)}\|_2 \overset{(ii)}{\leq} \left( \tilde{L}_{[K]}^{(t)} + \rho_{s^*+2\tilde{s}}^+ \right) \|\theta_{[K]}^{(t+1)} - \theta_{[K]}^{(t)}\|_2$$

$$\overset{(iii)}{\leq} \left( \tilde{L}_{[K]}^{(t)} \left( 1 + \frac{\rho_{s^*+2\tilde{s}}^+}{\rho_{s^*+2\tilde{s}}} \right) \right) \|\theta_{[K]}^{(t+1)} - \theta_{[K]}^{(t)}\|_2$$

$$\overset{(iv)}{\leq} \left( 1 + \kappa_{s^*+2\tilde{s}} \right) \sqrt{\frac{2 \left( \mathcal{F}_{\mu,\lambda K}(\theta_{[K]}^{(t)}) - \mathcal{F}_{\mu,\lambda K}(\tilde{\theta}_{[K]}^{(t+1)}) \right)}{\tilde{L}_{[K]}^{(t)}}} \overset{(v)}{\leq} \left( 1 + \kappa_{s^*+2\tilde{s}} \right) \sqrt{\frac{4 \rho_{s^*+2\tilde{s}}^+ \left( \mathcal{F}_{\mu,\lambda K}(\theta_{[K]}^{(t)}) - \mathcal{F}_{\mu,\lambda K}(\tilde{\theta}_{[K]}^{(t+1)}) \right)}{\tilde{L}_{[K]}^{(t)}}} \overset{(vi)}{\leq} \left( 1 + \kappa_{s^*+2\tilde{s}} \right) \sqrt{\frac{96 \lambda_{K}^2 s^* \kappa_{s^*+2\tilde{s}} \left( 1 - \frac{1}{8\kappa_{s^*+2\tilde{s}}} \right)}{\tilde{L}_{[K]}^{(t)}}}, \quad (G.3)$$

where (i) is from Lemma A.9, (ii) is from $S_{L_{[K]}^{(t)}}(\theta_{[K]}^{(t)}) \leq \rho_{s^*+2\tilde{s}}^+$, (iii) is from $\rho_{s^*+2\tilde{s}}^+ \leq \tilde{L}_{[K]}^{(t)}$ in (G.1), (iv) is from (A.16) in Lemma A.8, (v) is from $\tilde{L}_{[K]}^{(t)} \leq 2\rho_{s^*+2\tilde{s}}^+$ in (G.1) and monotone decrease of $\mathcal{F}_{\mu,\lambda K}(\theta_{[K]}^{(t)})$ from (A.16) in Lemma A.8, and (vi) is from (G.2) and $\kappa_{s^*+2\tilde{s}} = \frac{\rho_{s^*+2\tilde{s}}^+}{\rho_{s^*+2\tilde{s}}}$.

For $K$-th path following stage, $K = 1, \ldots, N - 1$, to have $\omega_{\lambda K}(\theta_{[K]}^{(t+1)}) \leq \lambda_K/4$, we set the R.H.S. of (G.3) to be no greater than $\lambda_K/4$, which is equivalent to requiring the number of iterations $k$ to be an upper bound of (B.4). For the last $N$-th path following stage, we need $\omega_{\lambda_N}(\tilde{\theta}_{[N]}) \leq \varepsilon_N \leq \lambda_N/4$. Set the R.H.S. of (G.3) to be no greater than $\varepsilon_N$, which is equivalent to requiring the number of iterations $k$ to be an upper bound of (B.5).

**H Proof of Lemma B.3**

Since $\omega_{\lambda_{K-1}}(\tilde{\theta}_{[K-1]}) \leq \lambda_{K-1}/4$, there exists some subgradient $g \in \partial\|\tilde{\theta}_{[K-1]}\|_1$ such that

$$\|\nabla L_{\mu}(\tilde{\theta}_{[K-1]}) + \lambda_{K-1}g\|_\infty \leq \lambda_{K-1}/4. \quad (H.1)$$

By the definition of $\omega_{\lambda K}(\cdot)$, we have

$$\omega_{\lambda K}(\tilde{\theta}_{[K-1]}) \leq \|\nabla L_{\mu}(\tilde{\theta}_{[K-1]}) + \lambda_K g\|_\infty = \|\nabla L_{\mu}(\tilde{\theta}_{[K-1]}) + \lambda_{K-1}g + (\lambda_K - \lambda_{K-1})g\|_\infty$$

$$\overset{(i)}{\leq} \|\nabla L_{\mu}(\tilde{\theta}_{[K-1]}) + \lambda_{K-1}g\|_\infty + |\lambda_K - \lambda_{K-1}| \cdot \|g\|_\infty \overset{(ii)}{\leq} \lambda_{K-1}/4 + (1 - \eta)\lambda_{K-1} \leq \lambda_K/2,$$

where (i) is from (H.1) and choice of $\lambda_K$, (ii) is from the condition on $\eta$. 

29
I Proof of Lemma B.4

Part 1. We first show $\|\theta - \theta^*\|_2^2 \leq r$ by contradiction. Suppose $\|\theta - \theta^*\|_2 > \sqrt{r}$. Let $\alpha \in [0, 1]$ such that $\tilde{\theta} = (1 - \alpha) \theta + \alpha \theta^*$ and

$$\|\tilde{\theta} - \theta^*\|_2 = \sqrt{r}. \quad (I.1)$$

Let $\tilde{g} = \arg\min_{g \in \theta} \|\nabla L_\mu(\theta) + \lambda g\|_\infty$ and $\Delta = \theta - \theta^*$, then we have

$$F_{\mu,\lambda}(\theta^*) \geq F_{\mu,\lambda}(\theta) - \langle \nabla L_\mu(\theta) + \lambda \tilde{g} \rangle^\top \Delta \geq F_{\mu,\lambda}(\theta) - \|\nabla L_\mu(\theta) + \lambda \tilde{g}\|_\infty \|\Delta\|_1 \ \ (I.2)$$

where (i) is from the convexity of $F_{\mu,\lambda}(\theta)$ and (ii) is from the approximate KKT condition.

Denote $\tilde{\Delta} = \tilde{\theta} - \theta^*$. Combining (I.2) and (I.1), we have

$$F_{\mu,\lambda}(\tilde{\theta}) \leq (1 - \alpha) F_{\mu,\lambda}(\theta) + \alpha F_{\mu,\lambda}(\theta^*) \leq (1 - \alpha) F_{\mu,\lambda}(\theta^*) + (1 - \alpha) \lambda \frac{\|\Delta\|_1}{4} + \alpha F_{\mu,\lambda}(\theta^*)$$

$$\leq F_{\mu,\lambda}(\theta^*) + \frac{\lambda}{4} \|\tilde{\theta} - \theta^*\|_1 = F_{\mu,\lambda}(\theta^*) + \frac{\lambda}{4} \|\tilde{\Delta}\|_1. \quad (I.3)$$

where (i) is from the convexity of $F_{\mu,\lambda}(\theta)$. This indicates

$$L_\mu(\tilde{\theta}) - L_\mu(\theta^*) \leq \lambda (\|\theta^*\|_1 - \|\tilde{\theta}\|_1 + \frac{1}{4} \|\tilde{\Delta}\|_1)$$

$$= \lambda (\|\theta^*_S\|_1 - \|\tilde{\theta}^*_S\|_1 + \frac{1}{4} \|\tilde{\Delta}^*_S\|_1 + \frac{1}{4} \|\tilde{\Delta}^*_S\|_1)$$

$$\leq \lambda (\|\theta^*_S - \tilde{\theta}^*_S\|_1 + \frac{1}{4} \|\tilde{\Delta}^*_S\|_1 + \frac{1}{4} \|\tilde{\Delta}^*_S\|_1)$$

$$= \frac{5\lambda}{4} \|\tilde{\Delta}^*_S\|_1 - \frac{3\lambda}{4} \|\tilde{\Delta}^*_S\|_1. \quad (I.3)$$

On the other hand, we have

$$L_\mu(\tilde{\theta}) - L_\mu(\theta^*) \geq \nabla L_\mu(\theta^*) \tilde{\Delta} \geq -\|\nabla L_\mu(\theta^*)\|_\infty \|\tilde{\Delta}\|_1 \ \ (I.4)$$

where (i) is from the convexity of $L_\mu(\theta)$, (ii) is from Assumption 3.2. Combining (I.3) and (I.4), we have

$$\|\tilde{\Delta}^*_S\|_1 \leq \frac{5}{2} \|\tilde{\Delta}^*_S\|_1. \quad (I.5)$$
Next, we consider the following sequence of sets:

\[ S_0 = \left\{ j \in \mathcal{S}^* : \sum_{m \in \mathcal{S}} \mathbf{1}(\tilde{\theta}_m \geq \tilde{\theta}_j) \leq \tilde{s} \right\} \]

and

\[ S_i = \left\{ j \in \mathcal{S}^* \setminus \bigcup_{k<i} \mathcal{S}_k : \sum_{m \in \mathcal{S}^* \setminus \bigcup_{k<i} \mathcal{S}_k} \mathbf{1}(\tilde{\theta}_m \geq \tilde{\theta}_j) \leq \tilde{s} \right\} \text{ for all } i = 1, 2, \ldots. \]

We introduce a result from Bühlmann and van de Geer (2011) with its proof provided therein.

**Lemma I.1** (Adapted from Lemma 6.9 in Bühlmann and van de Geer (2011) by setting \( q = 2 \)). Let \( \mathbf{v} = [v_1, v_2, \ldots]^T \) with \( v_1 \geq v_2 \geq \ldots \geq 0 \). For any \( s \in \{1, 2, \ldots\} \), we have

\[
\left( \sum_{j \geq s+1} v_j^2 \right)^{1/2} \leq \sum_{k=1}^{\infty} \left( \sum_{j=ks+1}^{(k+1)s} v_j^2 \right)^{1/2} \leq \frac{\|\mathbf{v}\|_1}{\sqrt{s}}.
\]

Denote \( \mathcal{A} = \mathcal{S}^* \cup S_0 \). Then we have

\[
\sum_{i \geq 1} \|\tilde{\Delta}_{i} S_i\|_2 \overset{(i)}{\leq} \frac{1}{\sqrt{s}} \|\tilde{\Delta}_{\mathcal{S}^*}\|_1 \overset{(ii)}{\leq} \frac{5}{2} \sqrt{\frac{s+1}{s}} \|\tilde{\Delta}_{\mathcal{S}^*}\|_2 \leq \frac{5}{2} \sqrt{\frac{s+1}{s}} \|\tilde{\Delta}_{\mathcal{A}}\|_2,
\]

where (i) is from Lemma I.1 with \( s = \tilde{s} \) and (ii) is from (I.5). Let \( \tilde{\theta} = (1 - \beta)\tilde{\theta} + \beta\theta^* \) for any \( \beta \in [0, 1] \). Then we have

\[
\|\tilde{\theta} - \theta^*\|_2 = (1 - \beta)\|\tilde{\theta} - \theta^*\|_2 \leq \sqrt{r},
\]

which implies \( \mathcal{L}_\mu(\tilde{\theta}) \) satisfies RSC/RSS for \( \tilde{\theta} \) restricted on a sparse set by Assumption 3.3. Then we have

\[
|\tilde{\Delta}_{\mathcal{A}}^\top \nabla_{\mathcal{A}} \mathcal{L}_\mu(\tilde{\theta}) \tilde{\Delta}_{\mathcal{A}}| \leq \sum_{i \geq 1} |\tilde{\Delta}_{\mathcal{S}_i}^\top \nabla_{\mathcal{S}_i} \mathcal{L}_\mu(\tilde{\theta}) \tilde{\Delta}_{\mathcal{A}}| \leq \rho_{s^*+\tilde{s}}^+ \|\tilde{\Delta}_{\mathcal{A}}\|_2 \sum_{i \geq 1} \|\tilde{\Delta}_{S_i}\|_2 \overset{(i)}{\leq} \frac{5}{2} \sqrt{\frac{s^*+1}{s}} \|\rho_{s^*+\tilde{s}}^+ \tilde{\Delta}_{\mathcal{A}}\|_2,
\]

where (i) is from (I.6). On the other hand, we have from RSC

\[
\tilde{\Delta}_{\mathcal{A}}^\top \nabla_{\mathcal{A}} \mathcal{L}_\mu(\theta) \tilde{\Delta}_{\mathcal{A}} \geq \rho_{s^*+\tilde{s}}^- \|\tilde{\Delta}_{\mathcal{A}}\|_2^2.
\]

Then we have w.h.p.

\[
\tilde{\Delta}^\top \mathcal{L}_\mu(\tilde{\theta}) \tilde{\Delta} = \tilde{\Delta}_{\mathcal{A}}^\top \nabla_{\mathcal{A}} \mathcal{L}_\mu(\tilde{\theta}) \tilde{\Delta}_{\mathcal{A}} + 2 \tilde{\Delta}_{\mathcal{A}}^\top \nabla_{\mathcal{A}} \mathcal{L}_\mu(\tilde{\theta}) \tilde{\Delta}_{\mathcal{A}} + \tilde{\Delta}_{\mathcal{A}}^\top \nabla_{\mathcal{A}} \mathcal{L}_\mu(\tilde{\theta}) \tilde{\Delta}_{\mathcal{A}}
\geq \tilde{\Delta}_{\mathcal{A}}^\top \nabla_{\mathcal{A}} \mathcal{L}_\mu(\tilde{\theta}) \tilde{\Delta}_{\mathcal{A}} + 2 |\tilde{\Delta}_{\mathcal{A}}^\top \nabla_{\mathcal{A}} \mathcal{L}_\mu(\tilde{\theta}) \tilde{\Delta}_{\mathcal{A}}|
\overset{(i)}{\geq} \left( \rho_{s^*+\tilde{s}}^- - 5 \sqrt{\frac{s^*}{s}} \rho_{s^*+\tilde{s}}^+ \right) \|\tilde{\Delta}_{\mathcal{A}}\|_2^2 \overset{(ii)}{\geq} \frac{9}{14} \rho_{s^*+\tilde{s}}^- \|\tilde{\Delta}_{\mathcal{A}}\|_2^2,
\]

31
where (i) is from (I.7) and (I.8), (ii) is from Assumption 3.3. This implies
\[
L_\mu(\tilde{\theta}) - L_\mu(\theta^*) = \nabla L_\mu(\theta^*)^\top \Delta + \frac{1}{2} \Delta \nabla L_\mu(\theta) \Delta \geq \nabla L_\mu(\theta^*)^\top \Delta + \frac{9}{28} \rho_{s^* + \tilde{s}} \|\Delta_A\|_2^2
\]
\[
\geq \frac{9}{28} \rho_{s^* + \tilde{s}} \|\Delta_A\|_2^2 - \frac{\lambda}{6} \|\Delta S^*\|_1 - \frac{\lambda}{6} \|\Delta_S^*\|_1,
\]
where (i) is from Assumption 3.2. Combining (I.3) and (I.9), we have
\[
\rho_{s^* + \tilde{s}} \|\Delta S^*\|_2^2 \leq \rho_{s^* + \tilde{s}} \|\Delta_A\|_2^2 \leq \frac{8}{3} \lambda \|\Delta S^*\|_1 \leq \frac{8}{3} \lambda \sqrt{s^*} \|\Delta_S^*\|_2 \leq \frac{8}{3} \lambda \sqrt{s^*} \|\Delta_A\|_2.
\]
This implies
\[
\|\Delta S^*\|_2 \leq \|\Delta A\|_2 \leq \frac{8\lambda \sqrt{s^*}}{3\rho_{s^* + \tilde{s}}} \quad \text{and} \quad \|\Delta S^*\|_1 \leq \frac{8\lambda s^*}{3\rho_{s^* + \tilde{s}}}. \tag{I.10}
\]
Then we have
\[
\|\Delta\|_2 \leq \sqrt{\|\Delta A\|_2^2 + \|\Delta_A\|_2^2} \leq \frac{8\lambda \sqrt{s^*}}{\rho_{s^* + \tilde{s}}} < \sqrt{7}.
\]
This conflicts with (I.1), which indicates that \(\|\theta - \theta^*\|_2 \leq \sqrt{7}\).

**Part 2.** We next demonstrate the sparsity of \(\theta\). From \(\lambda > \lambda_N \geq 6\|\nabla L_\mu(\theta^*)\|_\infty\), then we have
\[
\left\{ i \in \mathcal{S}^* : |\nabla_i L_\mu(\theta^*)| \geq \frac{\lambda}{6} \right\} = 0. \tag{I.12}
\]
Denote \(\mathcal{S}_1 = \left\{ i \in \mathcal{S}^* : |\nabla_i L_\mu(\theta) - \nabla_i L_\mu(\theta^*)| \geq \frac{2\lambda}{3} \right\}\) and \(\hat{s}_1 = |\mathcal{S}_1|\). Then there exists some \(b \in \mathbb{R}^d\) such that \(\|b\|_\infty = 1\), \(\|b\|_0 \leq \hat{s}_1\) and \(b^\top (\nabla L_\mu(\theta) - \nabla L_\mu(\theta^*)) \geq \frac{2\lambda}{3}\hat{s}_1\). Then by the mean value theorem, we have for some \(\tilde{\theta} = (1 - \alpha) \theta + \alpha \theta^*\) with \(\alpha \in [0, 1]\), \(\nabla L_\mu(\theta) - \nabla L_\mu(\theta^*) = \nabla^2 L_\mu(\tilde{\theta}) \Delta\), where \(\Delta = \theta - \theta^*\). Then we have
\[
\frac{2\lambda \hat{s}_1}{3} \leq b^\top \nabla^2 L_\mu(\tilde{\theta}) \Delta \leq \sqrt{b^\top \nabla^2 L_\mu(\tilde{\theta}) b \sqrt{\Delta^\top \nabla^2 L_\mu(\tilde{\theta}) \Delta}} \tag{I.13}
\]
where (i) is from the generalized Cauchy-Schwarz inequality, (ii) is from the definition of RSS and the fact that \(\|b\|_2 \leq \sqrt{\hat{s}_1}\|b\|_\infty \leq \sqrt{\hat{s}_1}\). Let \(g\) achieve \(\min_{g \in \partial \|\theta\|_1} \mathcal{R}_{\mu, \lambda}(\theta)\). Further, we have
\[
\Delta^\top (\nabla L_\mu(\theta) - \nabla L_\mu(\theta^*)) \leq \|\Delta\|_1 \|\nabla L_\mu(\theta) - \nabla L_\mu(\theta^*)\|_\infty
\]
\[
\leq \|\Delta\|_1 (\|\nabla L_\mu(\theta^*)\|_\infty + \|\nabla L_\mu(\theta)\|_\infty)
\]
\[
\leq \|\Delta\|_1 (\|\nabla L_\mu(\theta^*)\|_\infty + \|\nabla L_\mu(\theta) + \lambda g\|_\infty + \lambda \|g\|_\infty)
\]
\[
\leq \frac{28\lambda s^*}{3\rho_{s^* + \tilde{s}}} \left[ \frac{\lambda}{6} + \frac{\lambda}{4} + \lambda \right] \leq \frac{14\lambda^2 s^*}{\rho_{s^* + \tilde{s}}}. \tag{I.14}
\]
where (i) is from combining (I.5) and (I.10), condition on \( \lambda \), approximate KKT condition and \( \|g\|_\infty \leq 1 \). Combining (I.13) and (I.14), we have
\[
\frac{32 \rho^{+}_{s^{*}} s^{*}}{\rho_{s^{*} + \tilde{s}}} \leq 32 \kappa_{s^{*} + 2s^{*}} \leq \tilde{s}.
\]

For any \( v \in \mathbb{R}^d \) that satisfies \( \|v\|_0 \leq 1 \), we have
\[
\tilde{S}_2 = \left\{ i \in \mathcal{S}^* : \left| \nabla_i \mathcal{L}_\mu(\theta) + \frac{\lambda}{4} v_i \right| \geq \frac{5 \lambda}{6} \right\} \subseteq \left\{ i \in \mathcal{S}^* : \left| \nabla_i \mathcal{L}_\mu(\theta^*) \right| \geq \frac{\lambda}{6} \right\} \bigcup \tilde{S}_1.
\]
Then we have \( |\tilde{S}_2| \leq |\tilde{S}_1| \leq \tilde{s} \). Since for any \( i \in \mathcal{S}^* \) and \( \left| \nabla_i \mathcal{L}_\mu(\theta) + \frac{\lambda}{4} v_i \right| < \frac{5 \lambda}{6} \), we can find \( g_i \) that satisfies \( |g_i| \leq 1 \) such that \( \nabla_i \mathcal{L}_\mu(\theta) + \frac{\lambda}{4} v_i + \lambda g_i = 0 \) which implies \( \theta_i = 0 \), then we have
\[
\left| \left\{ i \in \mathcal{S}^* : \left| \nabla_i \mathcal{L}_\mu(\theta) + \frac{\lambda}{4} v_i \right| < \frac{5 \lambda}{6} \right\} \right| = 0.
\]
Therefore, we have \( \|\theta_{\mathcal{S}^*}\|_0 \leq |\tilde{S}_2| \leq \tilde{s} \).

### J Proof of Lemma B.5

Let \( n^{(t)} \) be the number of proximal gradient steps in \( t \)-th iteration of PIS\(^2\)TA. Then we have
\[
L^{(t+1)} \leq 2L^{(t)} \left( \frac{1}{2} \right)^{n^{(t)}-1}.
\]
This indicates
\[
n^{(t)} \leq 2 + \log_2 \frac{L^{(t)}}{L^{(t+1)}}.
\]
Then we have
\[
\frac{t = 0}{T_K} n^{(t)} \leq 2(T_K + 1) + \log_2 \frac{L^{(0)}}{L^{(T_K+1)}}
\]

We obtain the desired result by \( L^{(0)} = L_{\max} \) and \( L^{(T_K+1)} \geq \rho_{s^{*} + 2\tilde{s}}^{+} \).

### K Intermediate Results of Theorem 4.3

We start with some preliminaries. For any \( S \subset \{1, \ldots, d\} \) with \( |S| \leq s^{*} \), we denote the set of cone
\[
\mathcal{C}_S^\nu = \left\{ x \in \mathbb{R}^d : \|x_S\|_1 \leq \nu \|x_S\|_1 \right\} \quad \text{and} \quad \mathcal{C}_s^\nu = \bigcup_{S \subset \{1, \ldots, d\} : |S| \leq s^{*}} \mathcal{C}_S^\nu.
\]
Besides, since $\Theta^* = \Sigma^{*-1} \in \mathcal{M}(\kappa, s^*)$, we have
\[
\frac{\lambda_{\text{max}}(\Theta^*)}{\lambda_{\text{min}}(\Theta^*)} = \frac{\lambda_{\text{max}}(\Sigma^{*-1})}{\lambda_{\text{min}}(\Sigma^{*-1})} = \frac{\lambda_{\text{max}}(\Sigma^*)}{\lambda_{\text{min}}(\Sigma^*)} \leq \kappa.
\]

We first introduce some important results on characterizing the data matrix $X$. These are adapted from intermediate lemmas in Liu and Wang (2012), which we refer to interested readers for detailed proofs.

The first lemma provides the bounds of entry-wise difference between sample and population correlation matrices.

**Lemma K.1.** Let $\hat{R}$ and $R$ be the sample and population correlation matrices. Then for event
\[
\mathcal{E}_1 = \left\{ \|\hat{R} - R\|_\infty \leq 18 \sqrt{\frac{\log d}{n}} \right\},
\]
we have $\mathbb{P}[\mathcal{E}_1] \geq 1 - d^{-1}$.

The second lemma provides the bounds of normalized model noise $\epsilon$.

**Lemma K.2.** Let $\epsilon_i \in \mathbb{R}^d$ follows $\epsilon_i \sim \mathcal{N}_n(0, \sigma_i^2 I_n)$. Then for event
\[
\mathcal{E}_2 = \left\{ \max_{i \in \{1, \ldots, d\}} \frac{\|\epsilon_i\|_2^2}{n\sigma_i^2} \leq 1.4 \quad \text{and} \quad \max_{i \in \{1, \ldots, d\}} \left\| \frac{\|\epsilon_i\|_2^2}{n\sigma_i^2} - 1 \right\| \leq 3.5 \sqrt{\frac{\log d}{n}} \right\},
\]
we have $\mathbb{P}[\mathcal{E}_1] \geq 1 - d^{-1} - d \exp(-100/n)$.

The third lemma provides the bounds of sample standard deviation of the marginal univariate Gaussian random variables.

**Lemma K.3.** Let $\hat{\Sigma}$ be the sample covariance matrix. Suppose Assumption (A3) holds, then for event
\[
\mathcal{E}_3 = \left\{ \frac{1}{2} \lambda_{\text{min}}(\Sigma) \leq \min_{i \in \{1, \ldots, d\}} \hat{\Sigma}_{ii} \leq \max_{i \in \{1, \ldots, d\}} \hat{\Sigma}_{ii} \leq \frac{3}{2} \lambda_{\text{max}}(\Sigma) \right\},
\]
we have $\mathbb{P}[\mathcal{E}_1] \geq 1 - d^{-1} - d \exp(-100/n)$.

Next, we verify Assumption 3.2 in the following lemma, and provide its proof in Appendix N.7.

**Lemma K.4.** Denote $\mathcal{L}_{\mu,i}(\theta^*_i) = \|z_i - Z_{*,\lambda_i}^*\|_\mu / \sqrt{n}$. Let $\lambda_N = \frac{6 \sqrt{5 \log d / n}}{\max\{\sqrt{5/3} \mu \min_{i} \{\Gamma^{-1/2}_{i} / (\sqrt{\pi} \sigma_i)\}, 1\}}$, then for event
\[
\mathcal{E}_4 = \left\{ \lambda_N \geq 6 \max_{i \in \{1, \ldots, d\}} \|\nabla \mathcal{L}_{\mu,i}(\theta^*_i)\|_\infty \right\},
\]
we have $\mathbb{P}[\mathcal{E}_1] \geq 1 - d \exp\left(-\frac{n}{25}\right) - d^{0.6} \frac{0.6 \pi a \log d}{n}$.

Then we provide the bound of the restricted eigenvalue of the sample correlation matrix.
Lemma K.5. Suppose $E_3$ and Assumption 4.1 (A2) hold, then for event 
\[ E_5 = \left\{ \inf_{\theta \in C_{\nu}} \frac{\sqrt{s^* \theta^\top R \theta}}{\|R\|_1} \geq \frac{1}{5(1 + \nu) \sqrt{\kappa}} \right\}, \]
there exists constants $c_1$ and $c_2$ such that $\mathbb{P}[E_5|E_3] \geq 1 - c_2 \exp(-c_2n)$.

Further, we provide the prediction error bound for the approximate solution. It follows directly from Theorem 3.7.

Lemma K.6. Suppose $E_5$ holds, then for event 
\[ E_6 = \left\{ \max_{i \in \{1, \ldots, d\}} \left\| Z_{\cdot i} (\hat{\theta}_i - \theta_i^\ast) \right\| \leq c_3 \sqrt{s^* \log d} \right\}, \]
there exists constants $c_1$, $c_2$ and $c_3$ such that $\mathbb{P}[E_6|E_5] \geq 1 - c_1 \exp(-c_2n) - c_3d^{-1}$.

L Proof of Lemma 4.2

Using the result in Liu and Wang (2012) (Lemma 12) and Agarwal et al. (2010) (Proposition 1), we have with probability at least $1 - c_1 \exp(-c_2n)$ for some universal constants $c_1$ and $c_2$, for any $v \in \mathbb{R}^d$,
\[ \frac{1}{3 \Lambda_{\min}(\Sigma)} \|v\|_2^2 - \frac{9 \Lambda_{\max}(\Sigma)}{\Lambda_{\min}(\Sigma)} \cdot \log d \cdot \frac{n}{\|v\|_2^2} \leq \frac{\|Zv\|_2^2}{n} \leq \frac{2 \Lambda_{\max}(\Sigma)}{\Lambda_{\min}(\Sigma)} \cdot \frac{\|v\|_2^2}{n} + \frac{9 \Lambda_{\max}(\Sigma)}{\Lambda_{\min}(\Sigma)} \cdot \frac{\log d}{n} \cdot \|v\|_2^2. \]

Applying the same analysis for Theorem 3.4, we have for some constant $c$, $\|v\|_2^2 \leq c \|v_{S^\ast}\|_2^2 \leq cs^* \|v_{S^\ast}\|_2^2$ and we have
\[ \left( \frac{1}{3 \Lambda_{\min}(\Sigma)} - \frac{9 \Lambda_{\max}(\Sigma)}{\Lambda_{\min}(\Sigma)} \cdot s^* \log d \cdot \frac{n}{\|v\|_2^2} \right) \|v\|_2^2 \leq \frac{\|Zv\|_2^2}{n} \leq \frac{2 \Lambda_{\max}(\Sigma)}{\Lambda_{\min}(\Sigma)} \cdot \frac{\|v\|_2^2}{n} + \frac{9 \Lambda_{\max}(\Sigma)}{\Lambda_{\min}(\Sigma)} \cdot \frac{\log d}{n} \cdot \|v\|_2^2. \]

Since $n \geq \frac{54 \Lambda_{\max}(\Sigma)s^* \log d}{\Lambda_{\min}(\Sigma)}$, we have
\[ \frac{1}{3 \Lambda_{\max}(\Sigma)} \|v\|_2^2 \leq \frac{\|Zv\|_2^2}{n} \leq \frac{3 \Lambda_{\max}(\Sigma)}{\Lambda_{\min}(\Sigma)} \|v\|_2^2. \]

From $\Lambda_{\max}(\Sigma) = 1/\Lambda_{\min}(\Theta)$ and $\Lambda_{\min}(\Sigma) = 1/\Lambda_{\max}(\Theta)$, we have
\[ \frac{1}{3 \kappa \Theta} \|v\|_2^2 \leq \frac{\|Zv\|_2^2}{n} \leq 3 \kappa \Theta \|v\|_2^2. \]
To satisfy the condition for the computational theory, we require \( \mu \leq \frac{n \Var(\hat{\Gamma}_{ii}^{-1/2} \epsilon_i)}{4} \) for any \( i \in \{1, \ldots, d\} \). From \( \sigma_i = \Theta_i^{-1/2} \) and \( \min_{i \in \{1, \ldots, d\}} \hat{\Sigma}_{ii} \leq \frac{3}{2} \Lambda_{\max}(\Sigma) \) with high probability in Lemma K.3, we have

\[
\sqrt{\Var(\hat{\Gamma}_{ii}^{-1/2} \epsilon_i)} = \hat{\Gamma}_{ii}^{-1/2} \sigma_i = \frac{1}{\sqrt{\Theta_{ii} \hat{\Sigma}_{ii}}} \geq \frac{1}{\sqrt{\frac{3}{2} \Theta_{ii} \Lambda_{\max}(\Sigma)}}
\]

Since \( \Sigma = \Theta^{-1} \), for any \( i \in \{1, \ldots, d\} \), \( \mu \) need to satisfy

\[
\mu \leq \frac{\sqrt{n}}{4\sqrt{\frac{3}{2} \max_i \Theta_{ii} \Lambda_{\max}(\Sigma)}} \leq \frac{1}{5} \sqrt{\frac{n}{\Lambda_{\max}(\Theta) \Lambda_{\max}(\Sigma)}} = \frac{1}{5} \sqrt{\frac{n \Lambda_{\min}(\Theta)}{\Lambda_{\max}(\Theta)}} = \frac{1}{5} \sqrt{\frac{n}{\kappa_{\Theta}}}.
\]

By Lemma K.4, the condition on \( \lambda_N \) guarantees that Assumption 3.2 holds for each \( i = 1, \ldots, d \). Besides, when \( n \) is large enough, it can be guaranteed that there exists \( N_1 < N \), \( N_1 \in \mathbb{Z}^+ \), such that \( \frac{\sqrt{\sigma_i^2 + 2}}{8} > \lambda_{N_1} \geq 2 \lambda_N \geq 12 \|\nabla L_{\mu,i}(\theta_i)\|_\infty \).

## M Proof of Theorem 4.3

The analysis here follows directly from our analysis in the linear model and the analysis in Liu and Wang (2012). Let \( \mathcal{E} = \cap_{i=1}^6 \mathcal{E}_i \). Combining Lemma K.4 and our choice of \( \mu \), we have \( \lambda_N = 6\sqrt{\frac{5 \log d}{n}} \).

We first show that the estimation error of diagonal elements are bounded. suppose Assumption 4.1 and the event \( \mathcal{E} \) hold, the we have

\[
\max_{i \in \{1, \ldots, d\}} |\hat{\Theta}_{ii} - \Theta^*_{ii}| \leq c_4 \|\Theta^*\|_2 \frac{\log d}{n}.
\]

Besides, we have the \( \ell_1 \) norm error bounded for the estimation of off-diagonal elements each column.

**Lemma M.1** (Adapted from Lemma 14 in Liu and Wang (2012)). Suppose Assumption 4.1 and the event \( \mathcal{E} \) hold, the we have

\[
\max_{i \in \{1, \ldots, d\}} \|\hat{\Theta}_{\setminus i,i} - \Theta^*_{\setminus i,i}\|_1 \leq c_5 (s^* \|\Theta^*\|_2 + \|\Theta^*\|_1) \frac{\log d}{n}.
\]

Combining Lemma M.1 and Lemma M.2, we have

\[
\|\hat{\Theta} - \Theta^*\|_1 = \max_{i \in \{1, \ldots, d\}} \|\hat{\Theta}_{\setminus i,i} - \Theta^*_{\setminus i,i}\|_1 \leq \max_{i \in \{1, \ldots, d\}} \|\hat{\Theta}_{ii} - \Theta^*_{ii}\| + \|\hat{\Theta}_{\setminus i,i} - \Theta^*_{\setminus i,i}\|_1 \leq c_6 (s^* \|\Theta^*\|_2 + \|\Theta^*\|_1) \frac{\log d}{n} \leq c_7 (s^* \|\Theta^*\|_2) \frac{\log d}{n},
\]

where \( (i) \) is from \( \|\Theta^*\|_1 \leq s^* \|\Theta^*\|_2 \). Then we finish the proof from

\[
\|\hat{\Theta} - \Theta^*\|_2 \leq \|\hat{\Theta} - \Theta^*\|_1.
\]

36
Proofs of Intermediate Lemmas in Appendix A and Appendix K

N.1 Proof of Lemma A.2

We first bound the estimation error. From Assumption 3.3, we have the RSC property, which indicates

\[ \mathcal{L}_\mu(\theta) \geq \mathcal{L}_\mu(\theta^*) + (\theta - \theta^*)^\top \nabla \mathcal{L}_\mu(\theta^*) + (\rho_{s^* + 2\delta}/2)\|\theta - \theta^*\|_2^2, \quad (N.1) \]

\[ \mathcal{L}_\mu(\theta^*) \geq \mathcal{L}_\mu(\theta) + (\theta^* - \theta)^\top \nabla \mathcal{L}_\mu(\theta) + (\rho_{s^* + 2\delta}/2)\|\theta - \theta^*\|_2^2, \quad (N.2) \]

Adding (N.2) and (N.1), we have

\[ (\theta - \theta^*)^\top \nabla \mathcal{L}_\mu(\theta) \geq (\theta - \theta^*)^\top \nabla \mathcal{L}_\mu(\theta^*) + \rho_{s^* + 2\delta}\|\theta - \theta^*\|_2^2. \quad (N.3) \]

Let \( g \in \partial\|\theta\|_1 \) be the subgradient that achieves the approximate KKT condition of the L.H.S of (A.6), then we have

\[ (\theta - \theta^*)^\top (\nabla \mathcal{L}_\mu(\theta) + \lambda g) \leq \|\theta - \theta^*\|_1 \|\nabla \mathcal{L}_\mu(\theta) + \lambda g\|_\infty \leq \frac{1}{2}\lambda\|\theta - \theta^*\|_1. \quad (N.4) \]

On the other hand, we have from (N.3)

\[ (\theta - \theta^*)^\top (\nabla \mathcal{L}_\mu(\theta) + \lambda g) \geq (\theta - \theta^*)^\top \nabla \mathcal{L}_\mu(\theta^*) + \rho_{s^* + 2\delta}\|\theta - \theta^*\|_2^2 + \lambda g^\top (\theta - \theta^*), \quad (N.5) \]

Since \( \|\theta - \theta^*\|_1 = \|(\theta - \theta^*)_{S^*}\|_1 + \|\theta - \theta^*\|_{\overline{S^*}} \), then

\[ (\theta - \theta^*)^\top \nabla \mathcal{L}_\mu(\theta) \geq -\|(\theta - \theta^*)_{S^*}\|_1 \mathcal{L}_\mu(\theta^*)\|_\infty - \|(\theta - \theta^*)_{\overline{S^*}}\|_1 \mathcal{L}_\mu(\theta^*)\|_\infty. \quad (N.6) \]

Besides, we have

\[ (\theta - \theta^*)^\top g = g_{S^*}^\top (\theta - \theta^*)_{S^*} + g_{\overline{S^*}}^\top (\theta - \theta^*)_{\overline{S^*}} \overset{(i)}{=} -\|g_{S^*}\|_\infty \|(\theta - \theta^*)_{S^*}\|_1 + g_{S^*}^\top \theta_{S^*}, \]

\[ \overset{(ii)}{\geq} -\|(\theta - \theta^*)_{S^*}\|_1 + \|g_{S^*}\|_1 \overset{(iii)}{=} -\|(\theta - \theta^*)_{S^*}\|_1 + \|(\theta - \theta^*)_{\overline{S^*}}\|_1, \quad (N.7) \]

where (i) and (iii) is from \( \theta_{\overline{S^*}} = 0 \), (ii) is from \( \|g_{S^*}\|_\infty \leq 1 \) and \( g \in \partial\|\theta\|_1 \).

Combining (N.4), (N.5), (N.6) and (N.7), we have

\[ \frac{1}{2}\lambda\|\theta - \theta^*\|_1 = \frac{1}{2}\lambda \|(\theta - \theta^*)_{S^*}\|_1 + \lambda \|(\theta - \theta^*)_{\overline{S^*}}\|_1 \]

\[ \geq \rho_{s^* + 2\delta}\|\theta - \theta^*\|_2^2 - (\lambda + \|\mathcal{L}_\mu(\theta^*)\|_\infty)\|(\theta - \theta^*)_{S^*}\|_1 \]

\[ + (\lambda - \|\mathcal{L}_\mu(\theta^*)\|_\infty)\|(\theta - \theta^*)_{\overline{S^*}}\|_1. \]

This implies

\[ \rho_{s^* + 2\delta}\|\theta - \theta^*\|_2^2 + (\frac{1}{2}\lambda - \|\mathcal{L}_\mu(\theta^*)\|_\infty)\|(\theta - \theta^*)_{S^*}\|_1 \]

\[ \leq (\frac{3}{2}\lambda + \|\mathcal{L}_\mu(\theta^*)\|_\infty)\|(\theta - \theta^*)_{S^*}\|_1, \quad (N.8) \]

37
which results in (A.7) from $\rho_{s^*+2\tilde{s}} > 0$ and Assumption 3.2 as

$$\|L_{s^*}^\perp\|_1 \leq \frac{3}{2} \lambda + \|L_{\mu}(\theta^*)\|_\infty \|((\theta - \theta^*)_{s^*})\|_1.$$ 

Combining $\frac{1}{2} \lambda - \|L_{\mu}(\theta^*)\|_\infty \geq 0$, $\frac{3}{2} \lambda + \|L_{\mu}(\theta^*)\|_\infty \leq 2\lambda$ and (N.8), we have estimation error bound in (A.8) and (A.9) as

$$\rho_{s^*+2\tilde{s}} \|\theta - \theta^*\|^2_2 \leq 2\lambda \|((\theta - \theta^*)_{s^*})\|_1 \leq 2\lambda \sqrt{s^*} \|\theta - \theta^*\|_2.$$ 

Next, we bound the objective error in (A.10). We have

\[
\begin{align*}
F_{\mu,\lambda}(\theta) - F_{\mu,\lambda}(\theta^*) &\leq (\nabla L_{\mu}(\theta) + \lambda g)^\top (\theta^* - \theta) \leq \|\nabla L_{\mu}(\theta) + \lambda g\|_\infty \|\theta^* - \theta\|_1 \\
&\leq \frac{1}{2} \lambda \|\theta^*-\theta\|_1 = \frac{1}{2} \lambda \|((\theta^* - \theta)_{s^*})\|_1 + \|((\theta^* - \theta)_{s^*})\|_1 \\
&\leq 3\lambda \|((\theta^* - \theta)_{s^*})\|_1 \leq 3\lambda \sqrt{s^*} \|((\theta^* - \theta)_{s^*})\|_2 \leq \frac{6\lambda^2 s^*}{\rho_{s^*+2\tilde{s}}},
\end{align*}
\]

where (i) is from the convexity of $F_{\mu,\lambda}(\theta)$ with $\nabla L_{\mu}(\theta) + \lambda g$ as its subgradient, (ii) is from (A.7), and (iii) is from (A.8).

### N.2 Proof of Lemma A.3

Assumption $F_{\mu,\lambda}(\theta) - F_{\mu,\lambda}(\theta^*) \leq 6\lambda^2 s^*/\rho_{s^*+2\tilde{s}}$ implies

$$L_{\mu}(\theta) - L_{\mu}(\theta^*) + \lambda\|\theta\|_1 - \|\theta^*\|_1 \leq \frac{6\lambda^2 s^*}{\rho_{s^*+2\tilde{s}}}. \tag{N.9}$$

We have from the RSC property that

$$L_{\mu}(\theta) \geq L_{\mu}(\theta^*) + (\theta - \theta^*)^\top \nabla L_{\mu}(\theta^*) + \frac{\rho_{s^*+2\tilde{s}}}{2} \|\theta - \theta^*\|^2_2, \tag{N.10}$$

Then we have (N.9) and (N.10),

$$\frac{\rho_{s^*+2\tilde{s}}}{2} \|\theta - \theta^*\|^2_2 \leq \frac{6\lambda^2 s^*}{\rho_{s^*+2\tilde{s}}} - (\theta - \theta^*)^\top \nabla L_{\mu}(\theta^*) + \lambda\|\theta^*\|_1 - \|\theta\|_1. \tag{N.11}$$

Besides, we have

$$\|\theta^*\|_1 - \|\theta\|_1 = \|\theta_{s^*}\|_1 - \|\theta_{s^*}\|_1 - \|((\theta - \theta^*)_{s^*})\|_1 \leq \|((\theta - \theta^*)_{s^*})\|_1 - \|((\theta - \theta^*)_{s^*})\|_1. \tag{N.12}$$

and

$$\|\theta^*\|_1 - \|\theta\|_1 \leq \|\theta_{s^*}\|_1 - \|\theta_{s^*}\|_1 - \|((\theta - \theta^*)_{s^*})\|_1 \leq \|((\theta - \theta^*)_{s^*})\|_1 - \|((\theta - \theta^*)_{s^*})\|_1. \tag{N.13}$$

38
Combining (N.11), (N.12) and (N.13), we have
\[
\frac{\rho_{s^*+2\bar{s}}}{2} \| \theta - \theta^* \|_2^2 \leq \frac{6\lambda^2 s^*}{\rho_{s^*+2\bar{s}}} + (\| \nabla L_\mu(\theta^*) \|_\infty + \lambda)(\theta - \theta^*)_{S^*} \|_1 \\
+ (\| \nabla L_\mu(\theta^*) \|_\infty - \lambda)(\theta - \theta^*)_{\bar{S}^*} \|_1.
\]
(N.14)

We discuss two cases as following:

Case 1. We first assume \( \| \theta - \theta^* \|_1 \leq \frac{12\lambda s^*}{\rho_{s^*+2\bar{s}}} \). Then (N.14) implies
\[
\frac{\rho_{s^*+2\bar{s}}}{2} \| \theta - \theta^* \|_2^2 \leq (i) \frac{6\lambda^2 s^*}{\rho_{s^*+2\bar{s}}} + (\| \nabla L_\mu(\theta^*) \|_\infty + \lambda)(\theta - \theta^*)_{S^*} \|_1 \\
+ (ii) \frac{6\lambda^2 s^*}{\rho_{s^*+2\bar{s}}} + \frac{3}{2} \lambda(\theta - \theta^*)_{S^*} \|_1 \\
\leq \frac{6\lambda^2 s^*}{\rho_{s^*+2\bar{s}}} + \frac{18\lambda^2 s^*}{\rho_{s^*+2\bar{s}}} = \frac{24\lambda^2 s^*}{\rho_{s^*+2\bar{s}}}.
\]
where (i) is from \( \| \nabla L_\mu(\theta^*) \|_\infty - \lambda \leq 0 \) and (ii) is from \( \| \nabla L_\mu(\theta^*) \|_\infty + \lambda \leq \frac{3}{2} \lambda \). This indicates
\[
\| \theta - \theta^* \|_2 \leq \frac{4\sqrt{3s^*} \lambda}{\rho_{s^*+2\bar{s}}}.
\]
(N.15)

Case 2. Next, we assume \( \| \theta - \theta^* \|_1 > \frac{12\lambda s^*}{\rho_{s^*+2\bar{s}}} \). Then (N.14) implies
\[
\frac{\rho_{s^*+2\bar{s}}}{2} \| \theta - \theta^* \|_2^2 \\
\leq (\| \nabla L_\mu(\theta^*) \|_\infty + \lambda)(\theta - \theta^*)_{S^*} \|_1 + (\| \nabla L_\mu(\theta^*) \|_\infty - \lambda)(\theta - \theta^*)_{\bar{S}^*} \|_1 + \frac{1}{2} \lambda(\theta - \theta^*) \|_1 \\
= (\| \nabla L_\mu(\theta^*) \|_\infty + \frac{3}{2} \lambda)(\theta - \theta^*)_{S^*} \|_1 + (\| \nabla L_\mu(\theta^*) \|_\infty - \frac{1}{2} \lambda)(\theta - \theta^*)_{\bar{S}^*} \|_1 \\
(i) \leq 2\lambda(\theta - \theta^*)_{S^*} \|_1 \leq 2\sqrt{s^*} \lambda(\theta - \theta^*)_{S^*} \|_2.
\]
(N.16)
where (i) is from \( \| \nabla L_\mu(\theta^*) \|_\infty + \frac{3}{2} \lambda \leq 2 \lambda \) and \( \| \nabla L_\mu(\theta^*) \|_\infty - \frac{1}{2} \lambda \leq 0 \). This indicates
\[
\| \theta - \theta^* \|_2 \leq \frac{4\sqrt{s^*} \lambda}{\rho_{s^*+2\bar{s}}}.
\]
(N.17)

Besides, we have
\[
\| \theta - \theta^* \|_1 \leq 6(\theta - \theta^*)_{S^*} \|_1 \leq 6\sqrt{s^*}(\theta - \theta^*)_{S^*} \|_2 \leq \frac{24\lambda s^*}{\rho_{s^*+2\bar{s}}}.
\]
(N.18)

where (i) is from \( \| \nabla L_\mu(\theta^*) \|_\infty + \frac{3}{2} \lambda \leq 2 \lambda \) and (N.16).

Combining (N.15) and (N.17), we have desired result (A.11). Combining the assumption in Case 1 and (N.18), we have desired result (A.12).
N.3 Proof of Lemma A.4

Recall that the proximal-gradient update can be computed by the soft-thresholding operation, i.e., for all $i = 1, \ldots, d$,

$$ (T_{L,\lambda}(\theta))_i = \text{sign}(\tilde{\theta}_i) \max \left\{ |\tilde{\theta}_i| - \lambda/L, 0 \right\} $$ \hspace{1cm} (N.19)

where $\theta = \theta - \nabla L_\mu(\theta)/L$. To bound $\| (T_{L,\lambda}(\theta))_{S'} \|_0$, we consider

$$ \theta = \theta - \frac{1}{L} \nabla L_\mu(\theta) = \theta - \frac{1}{L} \nabla L_\mu(\theta^*) + \frac{1}{L} (\nabla L_\mu(\theta^*) - \nabla L_\mu(\theta)) . $$ \hspace{1cm} (N.20)

We then consider the following three events:

$$ A_1 = \left\{ i \in S^* : |\theta_i| \geq \lambda/(3L) \right\} , $$ \hspace{1cm} (N.21)
$$ A_2 = \left\{ i \in S^* : |(\nabla L_\mu(\theta^*)/L)_i| > \lambda/(6L) \right\} , $$ \hspace{1cm} (N.22)
$$ A_3 = \left\{ i \in S^* : |(\nabla L_\mu(\theta^*)/L - \nabla L_\mu(\theta)/L)_i| \geq \lambda/(2L) \right\} , $$ \hspace{1cm} (N.23)

**Event** $A_1$. Note that for any $i \in S^*$, $|\theta_i| = |\theta_i - \theta_i^*|$, then we have

$$ |A_1| \leq \sum_{i \in S^*} \frac{3L}{\lambda} |\theta_i - \theta_i^*| \cdot 1(|\theta_i - \theta_i^*| \geq \lambda/(3L)) \leq \frac{3L}{\lambda} \sum_{i \in S^*} |\theta_i - \theta_i^*| $$

$$ \leq \frac{3L}{\lambda} \| \theta - \theta^* \|_1 \leq \frac{72Ls^*}{\rho_{s^*+2\bar{s}}}, $$ \hspace{1cm} (N.24)

where (i) is from (A.12) in Lemma A.3.

**Event** $A_2$. By Assumption 3.2 and $\lambda \geq \lambda_N$, we have

$$ 0 \leq |A_2| \leq \sum_{i \in S^*} \frac{6L}{\lambda} |(\nabla L_\mu(\theta^*)/L)_i| \cdot 1(|(\nabla L_\mu(\theta^*)/L)_i| > \lambda/(6L)) $$

$$ = \sum_{i \in S^*} \frac{6L}{\lambda} |(\nabla L_\mu(\theta^*)/L)_i| \cdot 0 = 0, $$ \hspace{1cm} (N.25)

which indicates that $|A_2| = 0$.

**Event** $A_3$. Consider the event $\tilde{A} = \left\{ i : |(\nabla L_\mu(\theta^*) - \nabla L_\mu(\theta))/L_i| \geq \lambda/2 \right\}$, which satisfies $A_3 \subseteq \tilde{A}$. We will provide an upper bound of $|\tilde{A}|$, which is also an upper bound of $|A_3|$. Let $v \in \mathbb{R}^d$ be chosen such that, $v_i = \text{sign} \left\{ (\nabla L_\mu(\theta^*)/L - \nabla L_\mu(\theta)/L)_i \right\}$ for any $i \in \tilde{A}$, and $v_i = 0$ for any $i \notin \tilde{A}$. Then we have

$$ v^T (\nabla L_\mu(\theta^*) - \nabla L_\mu(\theta)) = \sum_{i \in \tilde{A}} v_i (\nabla L_\mu(\theta^*)/L - \nabla L_\mu(\theta)/L)_i $$

$$ = \sum_{i \in \tilde{A}} |(\nabla L_\mu(\theta^*) - \nabla L_\mu(\theta))/L_i| \geq \lambda |\tilde{A}|/2. $$ \hspace{1cm} (N.26)
On the other hand, we have
\[
\mathbf{v}^\top (\nabla L_\mu(\theta^*) - \nabla L_\mu(\theta)) \leq \|\mathbf{v}\|_2 \|\nabla L_\mu(\theta^*) - \nabla L_\mu(\theta)\|_2 \leq \sqrt{|\tilde{A}|} \cdot \|\nabla L_\mu(\theta^*) - \nabla L_\mu(\theta)\|_2 \leq \rho_{s^*+2\bar{s}}^+ \sqrt{|\tilde{A}|} \cdot \|\theta - \theta^*\|_2,
\]
where (i) is from (A.11) in Lemma A.3 and definition of \(\kappa_{s^*+2\bar{s}} = \rho_{s^*+2\bar{s}}^+ / \rho_{s^*+2\bar{s}}^-\). Considering \(A_3 \subseteq \tilde{A}\), this implies
\[
|A_3| \leq |\tilde{A}| \leq 196\kappa_{s^*+2\bar{s}}^2 s^*.
\]

Now combining Even A1, A2, A3 and \(L \leq 2\rho_{s^*+2\bar{s}}^+\) in assumption, we close the proof as
\[
\| (T_{L,\lambda}(\theta))_{S^c} \|_0 \leq |A_1| + |A_2| + |A_3| \leq \frac{72 L s^*}{\rho_{s^*+2\bar{s}}^+} + 196\kappa_{s^*+2\bar{s}}^2 s^* \leq (144\kappa_{s^*+2\bar{s}}^2 + 196\kappa_{s^*+2\bar{s}}^2) s^* \\
\leq \bar{s}.
\]

N.4 Proof of Lemma A.5

Let \(g = \arg\min_{g \in \partial \lambda \|\theta\|_1} L_\mu + \lambda \|\theta\|_1\), then \(\omega_\lambda = \|\nabla L_\mu + \lambda g\|_\text{\infty}\). By the optimality of \(\overline{\theta}\) and convexity of \(\mathcal{F}_{L,\lambda}\), we have
\[
\mathcal{F}_{L,\lambda}(\theta) - \mathcal{F}_{L,\lambda}(\overline{\theta}) \leq (\nabla L_\mu + \lambda g)^\top (\theta - \overline{\theta}) \leq \|\nabla L_\mu + \lambda g\|_\text{\infty} \|\theta - \overline{\theta}\|_1 \\
\leq (\omega_\lambda(\theta) + \lambda - \overline{\lambda}) \|\theta - \overline{\theta}\|_1.
\]

Besides, we have
\[
\|\theta - \overline{\theta}\|_1 \leq \|\theta - \theta^*\|_1 + \|\overline{\theta} - \theta^*\|_1 \leq 6 \left(\| (\theta - \theta^*)_{S^c}\|_1 + \| (\overline{\theta} - \theta^*)_{S^c}\|_1\right) \leq 6\sqrt{s^*} \| (\theta - \theta^*)_{S^c}\|_2 + \| (\overline{\theta} - \theta^*)_{S^c}\|_2 \leq \frac{12(\lambda + \overline{\lambda})s^*}{\rho_{s^*+2\bar{s}}^+}.
\]

where (i) and (ii) are from (A.7) and (A.8) in Lemma A.2 respectively. Combining (N.29) and (N.30), we have desired result.
N.5 Proof of Lemma A.6

Our analysis has two steps. In the first step, we show that \( \{\theta(t)\}_{t=0}^{\infty} \) converges to the unique limit point \( \overline{\theta} \). In the second step, we show that the proximal gradient method has linear convergence rate.

**Step 1.** Note that \( \theta^{(t+1)} = T_{\mu,\lambda}(\theta^{(t)}) \). Since \( F_{\mu,\lambda}(\theta) \) is convex in \( \theta \) (but not strongly convex), the sub-level set \( \{\theta : F_{\mu,\lambda}(\theta) \leq F_{\mu,\lambda}(\theta^{(0)})\} \) is bounded. By the monotone decrease of \( F_{\mu,\lambda}(\theta^{(t)}) \) from (A.16) in Lemma A.8, \( \{\theta^{(t)}\}_{t=0}^{\infty} \) is also bounded. By Bolzano-Weierstrass theorem, it has a convergent subsequence and we will show that \( \overline{\theta} \) is the unique accumulation point.

Since \( F_{\mu,\lambda}(\theta) \) is bounded below,
\[
\lim_{k \to \infty} \|\theta^{(t+1)} - \theta^{(t)}\|_2 \leq \frac{2}{L^{(t)}_\mu} \cdot \lim_{k \to \infty} \left[ F_{\mu,\lambda}(\theta^{(t+1)}) - F_{\mu,\lambda}(\theta^{(t)}) \right] = 0.
\]

By Lemma A.9, we have
\[
\lim_{k \to \infty} \omega_{\lambda}(\theta^{(t)}) = 0,
\]
This implies \( \lim_{k \to \infty} \theta^{(t)} \) satisfies the KKT condition, hence is an optimal solution.

Let \( \overline{\theta} \) be an accumulation point. Since \( \overline{\theta} = \arg\min_{\theta} F_{\mu,\lambda}(\theta) \), then there exists some \( g \in \partial \|\overline{\theta}\|_1 \) such that
\[
\nabla F_{\mu,\lambda}(\overline{\theta}) = L_{\mu,\lambda}(\overline{\theta}) + \lambda g = 0. \tag{N.31}
\]

By Lemma A.4, every proximal update is sparse, hence \( \|\overline{\theta}_{S^c}\|_0 \leq \overline{s} \). By RSC property in (3.1), if \( \|\theta_{S^c}\|_0 \leq \overline{s} \), i.e., \( \|\theta - \overline{\theta}_{S^c}\|_0 \leq \overline{s} \), then we have
\[
L_{\mu}(\theta) - L_{\mu}(\overline{\theta}) \geq (\theta - \overline{\theta})^\top L_{\mu}(\overline{\theta}) + \frac{\rho_{s^*} + 2\overline{s}}{2} \|\theta - \overline{\theta}\|_2^2, \tag{N.32}
\]
From the convexity of \( \|\theta\|_1 \) and \( g \in \partial \|\overline{\theta}\|_1 \), we have
\[
\|\theta\|_1 - \|\overline{\theta}\|_1 \geq (\theta - \overline{\theta})^\top g. \tag{N.33}
\]
Combining (N.32) and (N.33), we have for any \( \|\theta_{S^c}\|_0 \leq \overline{s} \),
\[
F_{\mu,\lambda}(\theta) - F_{\mu,\lambda}(\overline{\theta}) = L_{\mu}(\theta) + \lambda \|\theta\|_1 - (L_{\mu}(\overline{\theta}) - \lambda \|\overline{\theta}\|_1)
\geq (\theta - \overline{\theta})^\top (L_{\mu}(\overline{\theta}) + \lambda g) + \frac{\rho_{s^*} + 2\overline{s}}{2} \|\theta - \overline{\theta}\|_2^2
\leq (\theta - \overline{\theta})^\top (L_{\mu}(\overline{\theta}) + \lambda g) + \frac{\rho_{s^*} + 2\overline{s}}{2} \|\theta - \overline{\theta}\|_2^2 \geq 0, \tag{N.34}
\]
where (i) is from (N.31). Therefore, \( \overline{\theta} \) is the unique accumulation point, i.e. \( \lim_{k \to \infty} \theta^{(t)} = \overline{\theta} \).

**Step 2.** The objective \( F_{\mu,\lambda}(\theta^{(t+1)}) \) satisfies
\[
F_{\mu,\lambda}(\theta^{(t+1)}) \overset{(i)}{\leq} Q_{\mu,\lambda}(\theta^{(t+1)}, \theta^{(t)})
\overset{(ii)}{=} \min_{\theta} L_{\mu}(\theta) + \nabla L_{\mu}(\theta^t)^\top (\theta - \theta^{(t)}) + \frac{L^{(t)}_\lambda}{2} \|\theta - \theta^{(t)}\|_2^2 + \lambda \|\theta\|_1. \tag{N.35}
\]
where (i) is from (A.16) in Lemma A.8, (ii) is from the definition of $O_{\mu,\lambda}$ in (2.3). To further bound R.H.S. of (N.35), we consider the line segment

$$S(\overline{\theta}, \theta^{(t)}) = \{ \theta : \theta = \alpha \overline{\theta} + (1 - \alpha)\theta^{(t)}, \alpha \in [0, 1] \}.$$  

Then we restrict the minimization over the line segment $S(\overline{\theta}, \theta^{(t)})$,

$$\mathcal{F}_{\mu,\lambda}(\theta^{(t+1)}) \leq \min_{\theta \in S(\overline{\theta}, \theta^{(t)})} \mathcal{L}_{\mu}(\theta^{(t)}) + \nabla \mathcal{L}_{\mu}(\theta^{(t)})^\top (\theta - \theta^{(t)}) + \frac{\overline{L}_\lambda}{2} \|\theta - \theta^{(t)}\|^2 + \lambda \|\theta\|_1. \quad (N.36)$$

Since $\|\overline{\theta}\|_1 \leq \tilde{s}$ and $\|\theta^{(t)}\|_1 \leq \tilde{s}$, then for any $\theta \in S(\overline{\theta}, \theta^{(t)})$, we have $\|\theta\|_1 \leq \tilde{s}$ and $\|\theta - \theta^{(t)}\|_1 \leq 2\tilde{s}$. By RSC property, we have

$$\mathcal{L}_{\mu}(\theta) \geq \mathcal{L}_{\mu}(\theta^{(t)}) + \nabla \mathcal{L}_{\mu}(\theta^{(t)})^\top (\theta - \theta^{(t)}) + \frac{\overline{L}_\lambda}{2} \|\theta - \theta^{(t)}\|^2$$

$$\geq \mathcal{L}_{\mu}(\theta^{(t)}) + \nabla \mathcal{L}_{\mu}(\theta^{(t)})^\top (\theta - \theta^{(t)}). \quad (N.37)$$

Combining (N.36) and (N.37), we have

$$\mathcal{F}_{\mu,\lambda}(\theta^{(t+1)}) \leq \min_{\theta \in S(\overline{\theta}, \theta^{(t)})} \mathcal{L}_{\mu}(\theta) + \frac{\overline{L}_\lambda}{2} \|\theta - \theta^{(t)}\|^2 + \lambda \|\theta\|_1$$

$$= \min_{\theta \in S(\overline{\theta}, \theta^{(t)})} \mathcal{F}_{\mu,\lambda}(\theta) + \frac{\overline{L}_\lambda}{2} \|\theta - \theta^{(t)}\|^2$$

$$= \min_{\alpha \in [0, 1]} \mathcal{F}_{\mu,\lambda}(\alpha \overline{\theta} + (1 - \alpha)\theta^{(t)}) + \frac{\alpha^2 \overline{L}_\lambda}{2} \|\theta - \theta^{(t)}\|^2$$

$$\leq \min_{\alpha \in [0, 1]} \alpha \mathcal{F}_{\mu,\lambda}(\overline{\theta}) + (1 - \alpha) \mathcal{F}_{\mu,\lambda}(\theta^{(t)}) + \frac{\alpha^2 \overline{L}_\lambda}{2} \|\theta - \theta^{(t)}\|^2$$

$$\leq \min_{\alpha \in [0, 1]} \mathcal{F}_{\mu,\lambda}(\theta^{(t)}) - \alpha \left( \mathcal{F}_{\mu,\lambda}(\theta^{(t)}) - \mathcal{F}_{\mu,\lambda}(\overline{\theta}) \right) + \frac{\alpha^2 \overline{L}_\lambda}{\rho \tilde{s}^* + 2\tilde{s}} \left( \mathcal{F}_{\mu,\lambda}(\theta^{(t)}) - \mathcal{F}_{\mu,\lambda}(\overline{\theta}) \right)$$

$$= \min_{\alpha \in [0, 1]} \mathcal{F}_{\mu,\lambda}(\theta^{(t)}) - \alpha \left( 1 - \frac{\alpha \overline{L}_\lambda}{\rho \tilde{s}^* + 2\tilde{s}} \right) \left( \mathcal{F}_{\mu,\lambda}(\theta^{(t)}) - \mathcal{F}_{\mu,\lambda}(\overline{\theta}) \right), \quad (N.38)$$

where (i) is from the convexity of $\mathcal{F}_{\mu,\lambda}$ and (ii) is from (N.34).

Minimize the R.H.S. of (N.38) w.r.t. $\alpha$, the optimal value $\alpha = \frac{\rho \tilde{s}^* + 2\tilde{s}}{2\overline{L}_\lambda}$ results in

$$\mathcal{F}_{\mu,\lambda}(\theta^{(t+1)}) \leq \mathcal{F}_{\mu,\lambda}(\theta^{(t)}) - \frac{\rho \tilde{s}^* + 2\tilde{s}}{4\overline{L}_\lambda} \left( \mathcal{F}_{\mu,\lambda}(\theta^{(t)}) - \mathcal{F}_{\mu,\lambda}(\overline{\theta}) \right). \quad (N.39)$$

Subtracting both sides of (N.39) by $\mathcal{F}_{\mu,\lambda}(\overline{\theta})$, we have

$$\mathcal{F}_{\mu,\lambda}(\theta^{(t+1)}) - \mathcal{F}_{\mu,\lambda}(\overline{\theta}) \leq \left( 1 - \frac{\rho \tilde{s}^* + 2\tilde{s}}{4\overline{L}_\lambda} \right) \left( \mathcal{F}_{\mu,\lambda}(\theta^{(t)}) - \mathcal{F}_{\mu,\lambda}(\overline{\theta}) \right)$$

$$\leq \left( 1 - \frac{\rho \tilde{s}^* + 2\tilde{s}}{8\rho \tilde{s}^* + 2\tilde{s}} \right) \left( \mathcal{F}_{\mu,\lambda}(\theta^{(t)}) - \mathcal{F}_{\mu,\lambda}(\overline{\theta}) \right), \quad (N.40)$$

43
where (i) is from Remark A.1. Apply (N.40) recursively, we have the desired result.

N.6 Proof of Lemma A.7

We first show an upper bound of $L_\mu$. Recall from the analysis in Appendix C of Lemma 3.6, there exists some $\alpha \in [0, 1]$ such that

$$
\nabla^2 L_\mu = \begin{cases} 
\frac{X^T X}{\sqrt{\mu}} , \\
\frac{1}{\sqrt{\mu} \|\xi\|_2} X^T \left( I - \frac{\xi \xi^T}{\|\xi\|_2^2} \right) X ,
\end{cases}
$$

if $\|\xi\|_2 < \mu$

$$
\text{or w.}
$$

where $\xi = y - X(w + \alpha \Delta)$. We discuss two cases depending on $\|\xi\|_2 < \mu$ and $\|\xi\|_2 \geq \mu$.

Case 1. For $\|\xi\|_2 < \sqrt{\mu}$, we have from the definition of $L_\mu$ in Remark A.1 that

$$
L_\mu \leq \|\nabla^2 L_\mu\|_2 = \frac{|X^T|}{\sqrt{\mu}} \|X\|_2 = \frac{|X^T|}{\sqrt{\mu}}.
$$

Case 2. For $\|\xi\|_2 \geq \mu$, we have

$$
L_\mu \leq \|\nabla^2 L_\mu\|_2 = \frac{1}{\sqrt{\mu} \|\xi\|_2} \left| X^T \left( I - \frac{\xi \xi^T}{\|\xi\|_2^2} \right) X \right|_2 \leq \frac{|X^T|}{\|\xi\|_2} = \frac{|X^T|}{\sqrt{\mu} \|\xi\|_2} \leq \frac{|X^T|}{\sqrt{\mu}}.
$$

Combining the two cases, we have

$$
L_\mu \leq \frac{|X^T|}{\sqrt{\mu}}. \quad (N.41)
$$

Applying the analogous argument in Step 1 of the proof of Lemma A.6, we have that $\{\theta(t)\}_{t=0}^{\infty}$ converges to the unique limit point $\tilde{\theta}$. By the monotonicity of $F_{\mu, \lambda}(\theta(t))$ from (A.16) in Lemma A.8 and convexity of $F_{\mu, \lambda}(\theta)$, we have $\|\theta(t) - \tilde{\theta}\|_2 \leq R$ for all $t = 1, 2, \ldots$. Then we have

$$
F_{\mu, \lambda}(\theta(1 + 1)) \leq Q_{\mu, \lambda}(\theta(1), \theta(t)) \leq \min_{\theta} F_{\mu, \lambda}(\theta) + \frac{L(t)}{2} \|\theta - \theta(t)\|_2^2
$$

$$
\leq \min_{\theta = a\tilde{\theta} + (1-a)\theta(t), a \in [0,1]} F_{\mu, \lambda}(\theta) + \frac{L(t)}{2} \|\theta - \theta(t)\|_2^2
$$

$$
= \min_{a \in [0,1]} F_{\mu, \lambda}(a\tilde{\theta} + (1-a)\theta(t)) + \frac{L(t)}{2} \|\theta(t) - \tilde{\theta}\|_2^2
$$

$$
\leq \min_{a \in [0,1]} F_{\mu, \lambda}(\theta(t)) - \alpha \left( F_{\mu, \lambda}(\theta(t)) - F_{\mu, \lambda}(\tilde{\theta}) \right) + \frac{2\|X\|_2^2 R^2 \alpha^2}{2\sqrt{\mu}}. \quad (N.42)
$$

where (i) and (ii) are from (A.16) and (A.15) in Lemma A.8 respectively, (iii) is from the convexity of $F_{\mu, \lambda}(\theta)$, $\|\theta(t) - \tilde{\theta}\|_2 \leq R$ for all $t = 1, 2, \ldots$ and $L(t) \leq 2\mu \leq 2\|X\|_2^2/(\sqrt{\mu})$ in Remark A.1 and Lemma 3.6. We discuss in two cases to provide an upper bound of R.H.S. (N.42).

Case 1: Suppose $F_{\mu, \lambda}(\theta(0)) - F_{\mu, \lambda}(\tilde{\theta}) \leq 2\|X\|_2^2 R^2/(\sqrt{\mu})$. Minimizing the R.H.S. of (N.42) w.r.t. $\alpha$, then the optimal value is

$$
\alpha = \frac{F_{\mu, \lambda}(\theta(t)) - F_{\mu, \lambda}(\tilde{\theta})}{2\|X\|_2^2 R^2/(\sqrt{\mu})} \leq \frac{F_{\mu, \lambda}(\theta(0)) - F_{\mu, \lambda}(\tilde{\theta})}{2\|X\|_2^2 R^2/(\sqrt{\mu})} \leq 1.
$$
Then we have
\[
F_{\mu,\lambda}(\theta^{(t+1)}) \leq F_{\mu,\lambda}(\theta^{(t)}) - \frac{(F_{\mu,\lambda}(\theta^{(t)}) - F_{\mu,\lambda}(\bar{\theta}))^2}{4\|X\|^2 R^2/(\sqrt{n}\mu)}
\]
Equivalently, we have
\[
F_{\mu,\lambda}(\theta^{(t+1)}) - F_{\mu,\lambda}(\bar{\theta}) \leq F_{\mu,\lambda}(\theta^{(t)}) - F_{\mu,\lambda}(\bar{\theta}) - \frac{(F_{\mu,\lambda}(\theta^{(t)}) - F_{\mu,\lambda}(\bar{\theta}))^2}{4\|X\|^2 R^2/(\sqrt{n}\mu)}
\]
Denote \(f_k = F_{\mu,\lambda}(\theta^{(t)}) - F_{\mu,\lambda}(\bar{\theta})\). Then we have
\[
\frac{1}{f_{k+1}} \leq \frac{1}{f_k} - \frac{1}{4f_k^2\|X\|^2 R^2/(\sqrt{n}\mu)},
\]
which results in
\[
f_{k+1} \geq f_k + \frac{f_{k+1}}{4f_k\|X\|^2 R^2/(\sqrt{n}\mu)} \geq f_k + \frac{1}{4\|X\|^2 R^2/(\sqrt{n}\mu)},
\]
where (i) is from the monotonicity of \(F_{\mu,\lambda}(\theta^{(t)})\) (A.16) in Lemma A.8. Applying (N.43) recursively, we have
\[
f_k \geq f_0 + \frac{k}{4\|X\|^2 R^2/(\sqrt{n}\mu)} \geq \frac{t + 2}{4\|X\|^2 R^2/(\sqrt{n}\mu)},
\]
where (i) is from \(F_{\mu,\lambda}(\theta^{(0)}) - F_{\mu,\lambda}(\bar{\theta}) < 2\|X\|^2 R^2/(\sqrt{n}\mu)\). Then we have the desired result (A.14).

Case 2: Suppose \(F_{\mu,\lambda}(\theta^{(0)}) - F_{\mu,\lambda}(\bar{\theta}) > 2\|X\|^2 R^2/(\sqrt{n}\mu)\). Minimize the R.H.S. of (N.42) w.r.t. \(\alpha\), then the optimal value is \(\alpha = 1\) and
\[
F_{\mu,\lambda}(\theta^{(1)}) - F_{\mu,\lambda}(\bar{\theta}) \leq \frac{2\|X\|^2 R^2}{2\sqrt{n}\mu}.
\]
We claim that for all \(t = 1, 2, \ldots\),
\[
F_{\mu,\lambda}(\theta^{(t)}) - F_{\mu,\lambda}(\bar{\theta}) = c_t 2\|X\|^2 R^2/(\sqrt{n}\mu),
\]
where \(\frac{1}{t+2} \leq c_t \leq \frac{2}{t+2}\). We prove the claim by induction.

This obviously holds when \(t = 1\). Assume \(\frac{1}{t+2} \leq c_t \leq \frac{2}{t+2}\) holds when \(t = T\). For \(t = T + 1\), minimize the R.H.S. of (N.42) w.r.t. \(\alpha\), then the optimal value is \(\alpha = c_T \leq 1\) and convergence rate for \(T + 1\)-th iteration is \(c_{T+1} = c_T - c_T^2/2\). Since \(c_{T+1}\) is a increasing function of \(c_T\) in \(c_T \in [0, 1/2]\) and \(c_t\) is monotone decreasing, i.e., \(c_t \leq c_1\) for all \(k > 1\), then we verifies the claim since
\[
c_{T+1} = \frac{2}{T+2} - \frac{1}{2} \left( \frac{2}{T+2} \right)^2 \leq \frac{2}{T+3}, \quad \text{and}
\]
\[
c_{T+1} \geq \frac{1}{T+2} - \frac{1}{2} \left( \frac{1}{T+2} \right)^2 \leq \frac{1}{T+3}.
\]
Combining the two cases, we have the desired result (A.14).
N.7 Proof of Lemma K.4

Recall that the model is
\[ z_i = Z_{*,\setminus i}^\star \theta_i^* + \hat{\Gamma}_{ii}^{-1/2} \epsilon_i, \]
where \( z_i = \hat{\Gamma}_{ii}^{-1/2} x_i \), \( Z_{*,\setminus i}^\star = X_{*,\setminus i} \hat{\Gamma}_{\setminus i,\setminus i}^{-1/2} \) and \( \epsilon_i \sim \mathcal{N}_n(0, \sigma_i^2 I_n) \). Then we have
\[
\nabla L_{\mu,i}(\theta_i^*) = Z_{*,\setminus i}^\top (Z_{*,\setminus i}^\star \theta_i^* - z_i) = -Z_{*,\setminus i}^\top \epsilon_i \max\{\sqrt{n\mu}, \sqrt{n\|Z_{*,\setminus i}^\star \theta_i^* - z_i\|_2}\}.
\]

Since \( \frac{\|\epsilon_i\|_2}{n\sigma_i} \sim \chi_n^2 \), we have from Johnstone (2001) that for any \( \delta \in [0, 1/2) \),
\[
\mathbb{P}\left[ \max_{i \in \{1, \ldots, d\}} \frac{\|\epsilon_i\|_2^2}{n\sigma_i^2} \leq 1 - \delta \right] \leq d \exp\left( -\frac{n\delta^2}{4} \right). \tag{N.44}
\]

Besides, \( Z_{*,\setminus i}^\top \epsilon_i \sim \mathcal{N}(0, n\sigma_i^2) \). Then we have from Liu and Wang (2012) that for any \( \delta \in [0, 1/2) \) and \( c > 2 \),
\[
\mathbb{P}\left[ \max_{i \in \{1, \ldots, d\}} \|Z_{*,\setminus i}^\top \epsilon_i\|_\infty > \sigma_i \sqrt{2cn \log d(1 - \delta)} \right] \leq \frac{d^{2-c(1-\delta)}}{\sqrt{\pi a \log d(1 - \delta)}}. \tag{N.45}
\]

Combining (N.44) and (N.45), we have with probability at least \( 1 - d \exp\left( -\frac{n\delta^2}{4} \right) - \frac{d^{2-c(1-\delta)}}{\sqrt{\pi a \log d(1 - \delta)}} \),
\[
\max_{i \in \{1, \ldots, d\}} \frac{\|Z_{*,\setminus i}^\top \epsilon_i\|_\infty}{\max\{\min_{i \in \{1, \ldots, d\}} \mu \hat{\Gamma}_{ii}^{1/2}, \sqrt{n}\|\epsilon_i\|_2\}} \leq \frac{\sqrt{2c \log d(1 - \delta)/n}}{\max\{\min_{i \in \{1, \ldots, d\}} \mu \hat{\Gamma}_{ii}^{1/2}/(\sqrt{n}\sigma_i), \sqrt{1 - \delta}\}}.
\]

Take \( \delta = 2/5 \) and \( c = 7/3 \), then we have the desired result.