On Fast Convergence of Proximal Algorithms for SQRT-Lasso Optimization: Don't Worry About its Nonsmooth Loss Function*

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

Abstract

Many machine learning techniques sacrifice convenient computational structures to gain estimation robustness and modeling flexibility. However, by exploring the modeling structures, we find these "sacrifices" do not always require more computational efforts. To shed light on such a "free-lunch" phenomenon, we study the square-root-Lasso (SQRT-Lasso) type regression problem. Specifically, we show that the nonsmooth loss functions of SQRT-Lasso type regression ease tuning effort and gain adaptivity to inhomogeneous noise, but is not necessarily more challenging than Lasso in computation. We can directly apply proximal algorithms (e.g. proximal gradient descent, proximal Newton, and proximal quasi-Newton algorithms) without worrying about the nonsmoothness of the loss function. Theoretically, we prove that the proximal algorithms enjoy fast local convergence with high probability. Our numerical experiments also show that when further combined with the pathwise optimization scheme, the proximal algorithms significantly outperform other competing algorithms.

1 Introduction

Many statistical machine learning methods can be formulated as optimization problems in the following form

$$\min_{\theta} \mathcal{L}(\theta) + \mathcal{R}(\theta), \tag{1.1}$$

where $\mathcal{L}(\theta)$ is a loss function and $\mathcal{R}(\theta)$ is a regularizer. When the loss function is smooth and has a Lipschitz continuous gradient, (1.1) can be efficiently solved by simple proximal gradient descent and proximal Newton algorithms (also requires a Lipschitz continuous Hessian matrix of $\mathcal{L}(\theta)$).

*Xingguo Li, Jarvis Haupt and Mingyi Hong are affiliated with Department of Electrical and Computer Engineering at University of Minnesota; Raman Arora is affiliated with Department of Computer Science at Johns Hopkins University; Han Liu is affiliated with Department of Electrical Engineering and Computer Science at Northwestern University; Haoming Jiang and Tuo Zhao is affiliated with School of Industrial and Systems Engineering at Georgia Institute of Technology; Tuo Zhao is the corresponding author. Emails: lixx1661@umn.edu, tourzhao@gatech.edu.

Some statistical machine learning methods, however, sacrifice convenient computational structures to gain estimation robustness and modeling flexibility Wang (2013); Belloni et al. (2011); Liu et al. (2015). Taking SVM as an example, the hinge loss function gains estimation robustness, but sacrifices the smoothness (compared with the square hinge loss function). However, by exploring the structure of the problem, we find that these "sacrifices" do not always require more computational efforts.

Advantage of SQRT-Lasso over Lasso. To shed light on such a "free-lunch" phenomenon, we study the high dimensional square-root (SQRT) Lasso regression problem Belloni et al. (2011); Sun and Zhang (2012). Specifically, we consider a sparse linear model in high dimensions,

$$y = X\theta^* + \epsilon,$$

where $X \in \mathbb{R}^{n\times d}$ is the design matrix, $y \in \mathbb{R}^n$ is the response vector, $\epsilon \sim N(0, \sigma^2 I_n)$ is the random noise, and θ^* is the sparse unknown regression coefficient vector. To estimate θ^*, Tibshirani (1996) propose the well-known Lasso estimator by solving

$$\overline{\theta}^{\mathsf{Lasso}} = \operatorname*{argmin}_{\theta} \frac{1}{n}\|y - X\theta\|_2^2 + \lambda_{\mathsf{Lasso}}\|\theta\|_1, \tag{1.2}$$

where λ_{Lasso} is the regularization parameter. Existing literature shows that given

$$\lambda_{\mathsf{Lasso}} \asymp \sigma\sqrt{\frac{\log d}{n}}, \tag{1.3}$$

$\overline{\theta}^{\mathsf{Lasso}}$ is minimax optimal for parameter estimation in high dimensions. Note that the optimal regularization parameter for Lasso in (1.3), however, requires the prior knowledge of the unknown parameter σ. This requires the regularization parameter to be carefully tuned over a wide range of potential values to get a good finite-sample performance.

To overcome this drawback, Belloni et al. (2011) propose the SQRT-Lasso estimator by solving

$$\overline{\theta}^{\mathsf{SQRT}} = \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \frac{1}{\sqrt{n}}\|y - X\theta\|_2 + \lambda_{\mathsf{SQRT}}\|\theta\|_1, \tag{1.4}$$

where λ_{SQRT} is the regularization parameter. They further show that $\overline{\theta}^{\mathsf{SQRT}}$ is also minimax optimal in parameter estimation, but the optimal regularization parameter is

$$\lambda_{\mathsf{SQRT}} \asymp \sqrt{\frac{\log d}{n}}. \tag{1.5}$$

Since (1.5) no longer depends on σ, SQRT-Lasso eases tuning effort.

Extensions of SQRT-Lasso. Besides the tuning advantage, the regularization selection for SQRT-Lasso type methods is also adaptive to inhomogeneous noise. For example, Liu et al. (2015) propose a multivariate SQRT-Lasso for sparse multitask learning. Given a matrix $A \in \mathbb{R}^{d\times d}$, let A_{*k}

denote the k-th column of A, and A_{i*} denote the i-th row of A. Specifically, Liu et al. (2015) consider a multitask regression model

$$Y = X\Theta^* + W,$$

where $X \in \mathbb{R}^{n\times d}$ is the design matrix, $Y \in \mathbb{R}^{n\times m}$ is the response matrix, $W_{*k} \sim N(0, \sigma_k^2 I_n)$ is the random noise, and $\Theta^* \in \mathbb{R}^{d\times m}$ is the unknown row-wise sparse coefficient matrix, i.e., Θ^* has many rows with all zero entries. To estimate Θ^*, Liu et al. (2015) propose a calibrated multivariate regression (CMR) estimator by solving

$$\overline{\theta}^{\mathsf{CMR}} = \underset{\theta\in\mathbb{R}^{d\times m}}{\operatorname{argmin}} \frac{1}{\sqrt{n}} \sum_{k=1}^{m} \|Y_{*k} - X\Theta_{*k}\|_2 + \lambda_{\mathsf{CMR}}\|\Theta\|_{1,2},$$

where $\|\Theta\|_{1,2} = \sum_{j=1}^{d} \|\Theta_{j*}\|_2$. Liu et al. (2015) further shows that the regularization of CMR approach is adaptive to σ_k's for each regression task, i.e., $Y_{*k} = X\Theta^*_{*k} + W_{*k}$, and therefore CMR achieves better performance in parameter estimation and variable selection than its least square loss based counterpart. With a similar motivation, Liu et al. (2017) propose a node-wise SQRT-Lasso approach for sparse precision matrix estimation. Due to space limit, please refer to Liu et al. (2017) for more details.

Existing Algorithms for SQRT-Lasso Optimization. Despite of these good properties, in terms of optimization, (1.4) for SQRT-Lasso is computationally more challenging than (1.2) for Lasso. The ℓ_2 loss in (1.4) is not necessarily differentiable, and does not have a Lipschitz continuous gradient, compared with the least square loss in (1.2). A few algorithms have been proposed for solving (1.4) in existing literature, but none of them are satisfactory when n and d are large. Belloni et al. (2011) reformulate (1.4) as a second order cone program (SOCP) and solve by an interior point method with a computational cost of $\mathcal{O}(nd^{3.5}\log(\epsilon^{-1}))$, where ϵ is a pre-specified optimization accuracy; Li et al. (2015) solve (1.4) by an alternating direction method of multipliers (ADMM) algorithm with a computational cost of $\mathcal{O}(nd^2/\epsilon)$; Sun and Zhang (2012) propose to solve the variational form of (1.4) by an alternating minimization algorithm, and Ndiaye et al. (2016) further develop a coordinate descent subroutine to accelerate its computation. However, no iteration complexity is established in Ndiaye et al. (2016). Our numerical study shows that their algorithm only scales to moderate problems. Moreover, Ndiaye et al. (2016) require a good initial guess for the lower bound of σ. When the initial guess is inaccurate, the empirical convergence can be slow.

Our Motivations. The major drawback of the aforementioned algorithms is that they do not explore the modeling structure of the problem. The ℓ_2 loss function is not differentiable only when the model are overfitted, i.e., the residuals are zero values $y - X\theta = 0$. Such an extreme scenario rarely happens in practice, especially when SQRT-Lasso is equipped with a sufficiently large regularization parameter λ_{SQRT} to yield a sparse solution and prevent overfitting. Thus, we can treat the ℓ_2 loss as an "almost" smooth function. Moreover, our theoretical investigation indicates that the ℓ_2 loss function also enjoys the restricted strong convexity, smoothness, and Hessian smoothness. In other words, the ℓ_2 loss function behaves as a strongly convex and smooth over a sparse domain. An illustration is provided in Figure 1.

Table 1: Comparison with existing algorithms for solving SQRT-Lasso. SOCP: Second-order Cone Programming; TRM: Trust Region Newton; VAM: Variational Alternating Minimization; ADMM: Alternating Direction Method of Multipliers; VCD: Coordinate Descent; Prox-GD: Proximal Gradient Descent; Prox-Newton: Proximal Newton.

	Algorithm	Theoretical Guarantee	Empirical Performance
Belloni et al. (2011)	SOCP + TRM	$\mathcal{O}(nd^{3.5}\log(\epsilon^{-1}))$	Very Slow
Sun and Zhang (2012)	VAM	N.A.	Very Slow
Li et al. (2015)	ADMM	$\mathcal{O}(nd^2/\epsilon)$	Slow
Ndiaye et al. (2016)	VAM + CD	N.A.	Moderate
This paper	Pathwise Prox-GD	$\mathcal{O}(nd\log(\epsilon^{-1}))$	Fast
This paper	Pathwise Prox-Newton + CD	$\mathcal{O}(snd\log\log(\epsilon^{-1}))$	Very Fast

Remark: Ndiaye et al. (2016) requires a good initial guess of σ to achieve moderate performance. Otherwise, its empirical performance is similar to ADMM.

Our Contributions. Given these nice geometric properties of the ℓ_2 loss function, we can directly solve (1.4) by proximal gradient descent (Prox-GD), proximal Newton (Prox-Newton), and proximal Quasi-Newton (Prox-Quasi-Newton) algorithms (Nesterov, 2013; Lee et al., 2014). Existing literature only apply these algorithms to solve optimization problems in statistical machine learning when the loss function is smooth. Our theoretical analysis shows that both algorithms enjoy fast convergence. Specifically, the Prox-GD algorithm achieves a local linear convergence and the Prox-Newton algorithm achieves a local quadratic convergence. The computational performance of these two algorithms can be further boosted in practice, when combined with the pathwise optimization scheme. Specifically, the pathwise optimization scheme solves (1.4) with a decreasing sequence of regularization parameters, $\lambda_0 \geq \ldots \geq \lambda_N$ with $\lambda_N = \lambda_{\mathsf{SQRT}}$. The pathwise optimization scheme helps yield sparse solutions and avoid overfitting throughout all iterations. Therefore, the nonsmooth loss function is differentiable. Besides sparse linear regression, we extend our algorithms and theory to sparse multitask regression and sparse precision matrix estimation. Extensive numerical results show our algorithms uniformly outperform the competing algorithms.

Hardness of Analysis. We highlight that our local analysis with strong convergence guarantees are novel and highly nontrivial for solving the SQRT-Lasso problem using simple and efficient proximal algorithms. First of all, sophisticated analysis is required to demonstrate the restricted strong convexity/smoothness and Hessian smoothness of the ℓ_2 loss function over a neighborhood of the underlying model parameter θ^* in high dimensions. These are key properties for establishing the strong convergence rates of proximal algorithms. Moreover, it is involved to guarantee

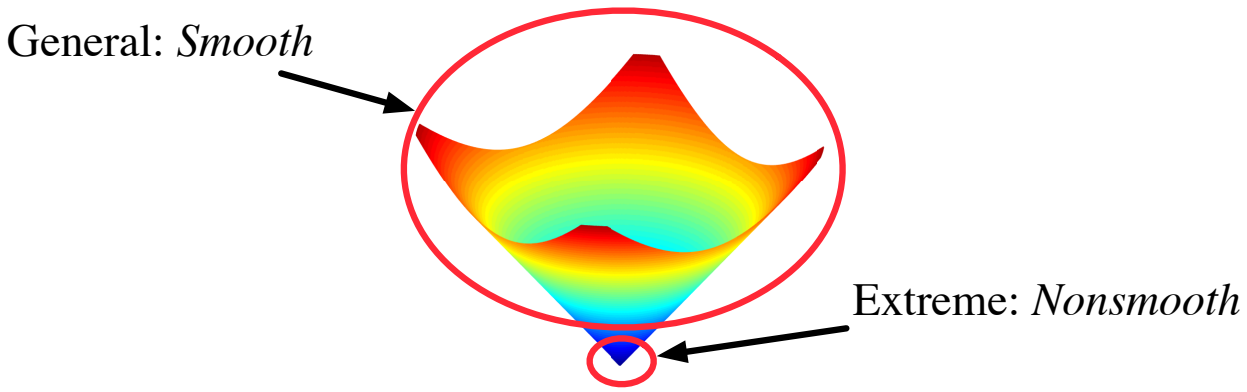

Figure 1: The extreme and general cases of the ℓ_2 loss. The nonsmooth region $\{\theta : y - X\theta = 0\}$ is out of our interest, since it corresponds to those overfitted regression models

that the output solution of the proximal algorithms do not fall in the nonsmooth region of the ℓ_2 loss function. This is important in guaranteeing the favored computational and statistical properties. In addition, it is highly technical to show that the pathwise optimization does enter the strong convergence region at certain stage. We defer all detailed analysis to the appendix.

Notations. Given a vector $v \in \mathbb{R}^d$, we define the subvector of v with the j-th entry removed as $v_{\setminus j} \in \mathbb{R}^{d-1}$. Given an index set $\mathcal{I} \subseteq \{1, ..., d\}$, let $\overline{\mathcal{I}}$ be the complementary set to $\mathcal{I}$ and $v_{\mathcal{I}}$ be a subvector of v by extracting all entries of v with indices in $\mathcal{I}$. Given a matrix $A \in \mathbb{R}^{d\times d}$, we denote A_{*j} (A_{k*}) the j-th column (k-th row), $A_{\setminus i \setminus j}$ as a submatrix of A with the i-th row and the j-th column removed and $A_{\setminus ij}$ ($A_{i\setminus j}$) as the j-th column (i-th row) of A with its i-th entry (j-th entry) removed. Let $\Lambda_{\max}(A)$ and $\Lambda_{\min}(A)$ be the largest and smallest eigenvalues of A respectively. Given an index set $\mathcal{I} \subseteq \{1, ..., d\}$, we use $A_{\mathcal{I}\mathcal{I}}$ to denote a submatrix of A by extracting all entries of A with both row and column indices in $\mathcal{I}$. We denote $A \succ 0$ if A is a positive-definite matrix. Given two real sequences $\{A_n\}, \{a_n\}$, we use conventional notations $A_n = \mathcal{O}(a_n)$ (or $A_n = \Omega(a_n)$) denote the limiting behavior, ignoring constant, $\widetilde{\mathcal{O}}$ to denote limiting behavior further ignoring logarithmic factors, and $\mathcal{O}_P(\cdot)$ to denote the limiting behavior in probability. $A_n \asymp a_n$ if $A_n = \mathcal{O}(a_n)$ and $A_n = \Omega(a_n)$ simultaneously. Given a vector $x \in \mathbb{R}^d$ and a real value $\lambda > 0$, we denote the soft thresholding operator $S_\lambda(x) = [\text{sign}(x_j)\max\{|x_j| - \lambda, 0\}]_{j=1}^d$. We use "w.h.p." to denote "with high probability".

2 Algorithm

We present the Prox-GD and Prox-Newton algorithms. For convenience, we denote

$$\mathcal{F}_\lambda(\theta) = \mathcal{L}(\theta) + \lambda\|\theta\|_1,$$

where $\mathcal{L}(\theta) = \frac{1}{\sqrt{n}}\|y - X\theta\|_2$. Since SQRT-Lasso is equipped with a sufficiently large regularization parameter λ to prevent overfitting, i.e., $y - X\theta \neq 0$, we treat $\mathcal{L}(\theta)$ as a differentiable function in this section. Formal justifications will be provided in the next section.

2.1 Proximal Gradient Desccent Algorithm

Given $\theta^{(t)}$ at t-th iteration, we consider a quadratic approximation of $\mathcal{F}_\lambda(\theta)$ at $\theta = \theta^{(t)}$ as

$$\mathcal{Q}_\lambda(\theta, \theta^{(t)}) = \mathcal{L}(\theta^{(t)}) + \nabla\mathcal{L}(\theta^{(t)})^\top(\theta - \theta^{(t)}) + \frac{L^{(t)}}{2}\|\theta - \theta^{(t)}\|_2^2 + \lambda\|\theta\|_1, \tag{2.1}$$

where $L^{(t)}$ is a step size parameter determined by the backtracking line search. We then take

$$\theta^{(t+1)} = \operatorname*{argmin}_\theta \mathcal{Q}_\lambda(\theta, \theta^{(t)}) = \mathcal{S}_{\frac{\lambda}{L^{(t)}}}\left(\theta^{(t)} - \frac{\nabla\mathcal{L}(\theta^{(t)})}{L^{(t)}}\right).$$

For simplicity, we denote $\theta^{(t+1)} = \mathcal{T}_{L^{(t+1)},\lambda}(\theta^{(t)})$. Given a pre-specified precision ε, we terminate the iterations when the approximate KKT condition holds:

$$\omega_\lambda(\theta^{(t)}) = \min_{g \in \partial\|\theta^{(t)}\|_1} \|\nabla\mathcal{L}(\theta^{(t)}) + \lambda g\|_\infty \le \varepsilon. \tag{2.2}$$

2.2 Proximal Newton Algorithm

Given $\theta^{(t)}$ at t-th iteration, we denote a quadratic term of θ as

$$\|\theta - \theta^{(t)}\|^2_{\nabla^2\mathcal{L}(\theta^{(t)})} = (\theta - \theta^{(t)})^\top\nabla^2\mathcal{L}(\theta^{(t)})(\theta - \theta^{(t)}),$$

and consider a quadratic approximation of $\mathcal{F}_\lambda(\theta)$ at $\theta = \theta^{(t)}$ is

$$\mathcal{Q}_\lambda(\theta, \theta^{(t)}) = \mathcal{L}(\theta^{(t)}) + \nabla\mathcal{L}(\theta^{(t)})^\top(\theta - \theta^{(t)}) + \frac{1}{2}\|\theta - \theta^{(t)}\|^2_{\nabla^2\mathcal{L}(\theta^{(t)})} + \lambda\|\theta\|_1. \tag{2.3}$$

We then take

$$\theta^{(t+0.5)} = \operatorname*{argmin}_\theta \mathcal{Q}_\lambda(\theta, \theta^{(t)}). \tag{2.4}$$

An additional backtracking line search procedure is required to obtain

$$\theta^{(t+1)} = \theta^{(t)} + \eta_t(\theta^{(t+0.5)} - \theta^{(t)}),$$

which guarantees $\mathcal{F}_\lambda(\theta^{(t+1)}) \le \mathcal{F}_\lambda(\theta^{(t)})$. The termination criterion for Prox-Newton is same with (2.2).

Remark 2.1. The ℓ_1 regularized quadratic problem in (2.4) can be solved efficiently by the coordinate descent algorithm combined with the active set strategy. See more details in Zhao et al. (2014). The computational cost is $\widetilde{\mathcal{O}}(snd)$, where $s \ll d$ is the solution sparsity.

Details of Prox-GD and Prox-Newton algorithms are summarized in Algorithms 1 and 2 respectively. To facilitate global fast convergence, we further combine the pathwise optimization Friedman et al. (2007) with the proximal algorithms. See more details in Section 4.

Remark 2.2. We can also apply proximal quasi-Newton method. Accordingly, at each iteration, the Hessian matrix in (2.3) is replaced with an approximation. See Bertsekas (1999) for more details.

Algorithm 1 Prox-GD algorithm for solving the SQRT-Lasso optimization (1.4). We treat $\mathcal{L}(\theta)$ as a differentiable function.

Input: y, X, λ, ε, $L_{\max} > 0$
Initialize: $\theta^{(0)}$, $t \leftarrow 0$, $L^{(0)} \leftarrow L_{\max}$, $\widetilde{L}^{(0)} \leftarrow L^{(0)}$
Repeat: $t \leftarrow t+1$
  Repeat: (Line Search)
    $\theta^{(t)} \leftarrow \mathcal{T}_{\widetilde{L}^{(t)},\lambda}(\theta^{(t-1)})$
    If $\mathcal{F}_\lambda(\theta^{(t)}) < \mathcal{Q}_\lambda(\theta^{(t)}, \theta^{(t-1)})$
    Then $\widetilde{L}^{(t)} \leftarrow \frac{\widetilde{L}^{(t)}}{2}$
  Until: $\mathcal{F}_\lambda(\theta^{(t)}) \geq \mathcal{Q}_\lambda(\theta^{(t)}, \theta^{(t-1)})$
  $L^{(t)} \leftarrow \min\{2\widetilde{L}^{(t)}, L_{\max}\}$, $\widetilde{L}^{(t)} \leftarrow L^{(t)}$
  $\theta^{(t)} \leftarrow \mathcal{T}_{L^{(t)},\lambda}(\theta^{(t-1)})$
Until: $\omega_\lambda(\theta^{(t)}) \leq \varepsilon$
Return: $\widehat{\theta} \leftarrow \theta^{(t)}$

Algorithm 2 Prox-Newton algorithm for solving the SQRT-Lasso optimization (1.4). We treat $\mathcal{L}(\theta)$ as a differentiable function.

Input: y, X, λ, ε
Initialize: $\theta^{(0)}$, $t \leftarrow 0$, $\mu \leftarrow 0.9$, $\alpha \leftarrow \frac{1}{4}$
Repeat: $t \leftarrow t+1$
  $\theta^{(t)} \leftarrow \operatorname{argmin}_\theta \mathcal{Q}_\lambda(\theta, \theta^{(t-1)})$
  $\Delta\theta^{(t)} \leftarrow \theta^{(t)} - \theta^{(t-1)}$
  $\gamma_t \leftarrow \nabla\mathcal{L}\left(\theta^{(t-1)}\right)^\top \Delta\theta^{(t)} + \lambda\left(\|\theta^{(t)}\|_1 - \|\theta^{(t-1)}\|_1\right)$
  $\eta_t \leftarrow 1$, $q \leftarrow 0$
  Repeat: $q \leftarrow q+1$ (Line Search)
    $\eta_t \leftarrow \mu^q$
  Until $\mathcal{F}_\lambda\left(\theta^{(t-1)} + \eta_t\Delta\theta^{(t)}\right) \leq \mathcal{F}_\lambda\left(\theta^{(t-1)}\right) + \alpha\eta_t\gamma_t$
  $\theta^{(t)} \leftarrow \theta^{(t)} + \eta_t\Delta\theta^{(t-1)}$
Until: $\omega_\lambda(\theta^{(t)}) \leq \varepsilon$
Return: $\widehat{\theta} \leftarrow \theta^{(t)}$

3 Theoretical Analysis

We start with defining the locally restricted strong convexity/smoothness and Hessian smoothness.

Definition 3.1. Denote

$$\mathcal{B}_r = \{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_2^2 \leq r\}$$

for some constant $r \in \mathbb{R}^+$. For any $v, w \in \mathcal{B}_r$ satisfying $\|v - w\|_0 \leq s$, $\mathcal{L}$ is *locally restricted strongly convex* (LRSHC), *smooth* (LRSS), and *Hessian smooth* (LRHS) respectively on $\mathcal{B}_r$ at sparsity level s,

if there exist universal constants $\rho_s^-, \rho_s^+, L_s \in (0, \infty)$ such that

$$\begin{aligned}
\text{LRSC:} &\mathcal{L}(v) - \mathcal{L}(w) - \nabla\mathcal{L}(w)^\top(v-w) \geq \frac{\rho_s^-}{2}\|v-w\|_2^2, \\
\text{LRSS:} &\mathcal{L}(v) - \mathcal{L}(w) - \nabla\mathcal{L}(w)^\top(v-w) \leq \frac{\rho_s^+}{2}\|v-w\|_2^2, \\
\text{LRHS:} &u^\top(\nabla^2\mathcal{L}(v) - \nabla^2\mathcal{L}(w))u \leq L_s\|v-w\|_2^2,
\end{aligned} \tag{3.1}$$

for any u satisfying $\|u\|_0 \leq s$ and $\|u\|_2 = 1$. We define the locally restricted condition number at sparsity level s as $\kappa_s = \frac{\rho_s^+}{\rho_s^-}$.

LRSC and LRSS are locally constrained variants of restricted strong convexity and smoothness (Agarwal et al., 2010; Xiao and Zhang, 2013), which are keys to establishing the strong convergence guarantees in high dimensions. The LRHS is parallel to the local Hessian smoothness for analyzing the proximal Newton algorithm in low dimensions (Lee et al., 2014). This is also closely related to the self-concordance (Nemirovski, 2004) in the analysis of Newton method (Boyd and Vandenberghe, 2009). Note that r is associated with the radius of the neighborhood of θ^* excluding the nonsmooth (and overfitted) region of the problem to guarantee strong convergence, which will be quantified below.

Next, we prove that the ℓ_2 loss of SQRT-Lasso enjoys the good geometric properties defined in Definition 3.1 under mild modeling assumptions.

Lemma 3.2. Suppose ϵ has i.i.d. sub-Gaussian entries with $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i^2] = \sigma^2$, $\|\theta^*\|_0 = s^*$. Then for any $\lambda \geq C_1\sqrt{\frac{\log d}{n}}$, w.h.p. we have

$$\lambda \geq \frac{C_1}{4}\|\nabla\mathcal{L}(\theta^*)\|_\infty.$$

Moreover, given each row of the design matrix X independently sampled from a sub-Gaussian distribution with the positive definite covariance matrix $\Sigma_X \in \mathbb{R}^{d\times d}$ with bounded eigenvalues. Then for

$$n \geq C_2 s^* \log d,$$

$\mathcal{L}(\theta)$ satisfies LRSC, LRSS, and LRHS properties on $\mathcal{B}_r$ at sparse level $s^* + 2\widetilde{s}$ with high probability. Specifically, (3.1) holds with

$$\rho^+_{s^*+2\widetilde{s}} \leq \frac{C_3}{\sigma},\ \rho^-_{s^*+2\widetilde{s}} \geq \frac{C_4}{\sigma} \text{ and } L_{s^*+2\widetilde{s}} \leq \frac{C_5}{\sigma},$$

where $C_1, \ldots, C_5 \in \mathbb{R}^+$ are generic constants, and r and $\widetilde{s}$ are sufficiently large constants, i.e., $\widetilde{s} > (196\kappa^2_{s^*+2\widetilde{s}} + 144\kappa_{s^*+2\widetilde{s}})s^*$.

The proof is provided in Appendix A. Lemma 3.2 guarantees that with high probability:
(i) λ is sufficiently large to eliminate the irrelevant variables and yields sufficiently sparse solutions (Bickel et al., 2009; Negahban et al., 2012);

(ii) LRSC, LRSS, and LRHS hold for the ℓ_2 loss of SQRT-Lasso such that fast convergence of the proximal algorithms can be established in a sufficiently large neighborhood of θ^* associated with r;
(iii) (3.1) holds in $\mathcal{B}_r$ at sparsity level $s^*+2\widetilde{s}$. Such a property is another key to the fast convergence of the proximal algorithms, because the algorithms can not ensure that the nonzero entries exactly falling in the true support set of θ^*.

3.1 Local Linear Convergence of Prox-GD

For notational simplicity, we denote

$$\mathcal{S}^* = \{j \mid \theta_j^* \neq 0\}, \quad \overline{\mathcal{S}}^* = \{j \mid \theta_j^* = 0\}, \text{ and} \mathcal{B}_r^{s^*+\widetilde{s}} = \mathcal{B}_r \cap \{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_0 \leq s^* + \widetilde{s}\}.$$

To ease the analysis, we provide a local convergence analysis when $\theta \in \mathcal{B}_r^{s^*+\widetilde{s}}$ is sufficiently close to θ^*. The convergence of Prox-GD is presented as follows.

Theorem 3.3. Suppose X and n satisfy conditions in Lemma 3.2. Given λ and $\theta^{(0)}$ such that $\lambda \geq \frac{C_1}{4}\|\nabla\mathcal{L}(\theta^*)\|_\infty$, $\|\theta^{(0)} - \theta^*\|_2^2 \leq s^*\left(8\lambda/\rho^-_{s^*+\widetilde{s}}\right)^2$ and $\theta^{(0)} \in \mathcal{B}_r^{s^*+\widetilde{s}}$, we have sufficiently sparse solutions throughout all iterations, i.e.,

$$\|[\theta^{(t)}]_{\overline{\mathcal{S}}^*}\|_0 \leq \widetilde{s}.$$

Moreover, given $\varepsilon > 0$, we need at most

$$T = \mathcal{O}\left(\kappa_{s^*+2\widetilde{s}} \log\left(\frac{\kappa^3_{s^*+2\widetilde{s}} s^* \lambda^2}{\varepsilon^2}\right)\right)$$

iterations to guarantee that the output solution $\widehat{\theta}$ satisfies

$$\|\widehat{\theta} - \overline{\theta}\|_2^2 = \mathcal{O}\left(\left(1 - \tfrac{1}{8\kappa_{s^*+2\widetilde{s}}}\right)^T \varepsilon\lambda s^*\right) \quad \text{and} \quad \mathcal{F}_\lambda(\widehat{\theta}) - \mathcal{F}_\lambda(\overline{\theta}) = \mathcal{O}\left(\left(1 - \tfrac{1}{8\kappa_{s^*+2\widetilde{s}}}\right)^T \varepsilon\lambda s^*\right),$$

where $\overline{\theta}$ is the unique sparse global optimum to (1.4) with $\|[\overline{\theta}]_{\overline{\mathcal{S}}^*}\|_0 \leq \widetilde{s}$.

The proof is provided in Appendix C. Theorem 3.3 guarantees that when properly initialized, the Prox-GD algorithm iterates within the smooth region, maintains the solution sparsity, and achieves a local linear convergence to the unique sparse global optimum to (1.4).

3.2 Local Quadratic Convergence of Prox-Newton

We then present the convergence analysis of the Prox-Newton algorithm as follows.

Theorem 3.4. Suppose X and n satisfy conditions in Lemma 3.2. Given λ and $\theta^{(0)}$ such that $\lambda \geq \frac{C_1}{4}\|\nabla\mathcal{L}(\theta^*)\|_\infty$, $\|\theta^{(0)} - \theta^*\|_2^2 \leq s^*\left(8\lambda/\rho^-_{s^*+\widetilde{s}}\right)^2$ and $\theta^{(0)} \in \mathcal{B}_r^{s^*+\widetilde{s}}$, we have sufficiently sparse solutions throughout all iterations, i.e.,

$$\|[\theta^{(t)}]_{\overline{\mathcal{S}}^*}\|_0 \leq \widetilde{s}.$$

Moreover, given $\varepsilon > 0$, we need at most

$$T = \mathcal{O}\left(\log\log\left(\frac{3\rho^{+}_{s^*+2\bar{s}}}{\varepsilon}\right)\right)$$

iterations to guarantee that the output solution $\widehat{\theta}$ satisfies

$$\|\widehat{\theta} - \overline{\theta}\|_2^2 = \mathcal{O}\left(\left(\frac{L_{s^*+2\bar{s}}}{2\rho^{-}_{s^*+2\bar{s}}}\right)^{2^T} \varepsilon\lambda s^*\right) \quad \text{and} \quad \mathcal{F}_\lambda(\widehat{\theta}) - \mathcal{F}_\lambda(\overline{\theta}) = \mathcal{O}\left(\left(\frac{L_{s^*+2\bar{s}}}{2\rho^{-}_{s^*+2\bar{s}}}\right)^{2^T} \varepsilon\lambda s^*\right),$$

where $\overline{\theta}$ is the unique sparse global optimum to (1.4).

The proof is provided in Appendix D. Theorem 3.4 guarantees that when properly initialized, the Prox-Newton algorithm also iterates within the smooth region, maintains the solution sparsity, and achieves a local quadratic convergence to the unique sparse global optimum to (1.4).

Remark 3.5. Our analysis can be further extended to the proximal quasi-Newton algorithm. The only technical difference is controlling the error of the Hessian approximation under restricted spectral norm.

3.3 Statistical Properties

Next, we characterize the statistical properties for the output solutions of the proximal algorithms.

Theorem 3.6. Suppose X, and n satisfy conditions in Lemma 3.2. Given $\lambda = C_1\sqrt{\log d/n}$, if the output solution $\widehat{\theta}$ obtained from Algorithm 1 and 2 satisfies the approximate KKT condition,

$$\omega_\lambda(\widehat{\theta}) \leq \varepsilon = \mathcal{O}\left(\frac{\sigma s^* \log d}{n}\right),$$

then we have:

$$\|\widehat{\theta} - \theta^*\|_2 = \mathcal{O}_P\left(\sigma\sqrt{\frac{s^*\log d}{n}}\right) \quad \text{and} \quad \|\widehat{\theta} - \theta^*\|_1 = \mathcal{O}_P\left(\sigma s^*\sqrt{\frac{\log d}{n}}\right).$$

Moreover, we have

$$|\widehat{\sigma} - \sigma| = \mathcal{O}_P\left(\frac{\sigma s^* \log d}{n}\right), \text{ where } \widehat{\sigma} = \frac{\|y - X\widehat{\theta}\|_2}{\sqrt{n}}.$$

The proof is provided in Appendix E. Recall that we use $\mathcal{O}_P(\cdot)$ to denote the limiting behavior in probability. Theorem 3.6 guarantees that the output solution $\widehat{\theta}$ obtained from Algorithm 1 and 2 achieves the minimax optimal rate of convergence in parameter estimation (Raskutti et al., 2011; Ye and Zhang, 2010). Note that in the stopping criteria $\omega_\lambda(\widehat{\theta}) \leq \varepsilon$, ε is not a tuning parameter, where $\mathcal{O}\big(\frac{\sigma s^* \log d}{n}\big)$ only serves as an upper bound and we can choose a small ε as desired. This is fundamentally different with the optimal λ_{Lasso} that tightly depends on σ.

4 Boosting Performance via Pathwise Optimization Scheme

We then apply the pathwise optimization scheme to the proximal algorithms, which extends the local fast convergence established in Section 3 to the global setting[1]. The pathwise optimization is essentially a multistage optimization scheme for boosting the computational performance Friedman et al. (2007); Xiao and Zhang (2013); Zhao et al. (2014).

Specifically, we solve (1.4) using a geometrically decreasing sequence of regularization parameters

$$\lambda_{[0]} > \lambda_{[1]} > \ldots > \lambda_{[N]},$$

where $\lambda_{[N]}$ is the target regularization parameter of SQRT-Lasso. This yields a sequence of output solutions

$$\widehat{\theta}_{[0]}, \widehat{\theta}_{[1]}, \ldots, \widehat{\theta}_{[N]},$$

also known as the solution path. At the K-th optimization stage, we choose $\widehat{\theta}_{[K-1]}$ (the output solution of the $(K-1)$-th stage) as the initial solution, and solve (1.4) with $\lambda = \lambda_{[K]}$ using the proximal algorithms. This is also referred as the warm start initialization in existing literature (Friedman et al., 2007). Details of the pathwise optimization is summarized in Algorithm 3. In terms of $\epsilon_{[K]}$, because we only need high precision for the final stage, we set $\epsilon_{[K]} = \lambda_{[K]}/4 \gg \epsilon_{[N]}$ for $K < N$.

Algorithm 3 The pathwise optimization scheme for the proximal algorithms. We solve the optimization problem using a geometrically decreasing sequence of regularization parameters.

Input: y, X, N, $\lambda_{[N]}$, $\varepsilon_{[N]}$

Initialize: $\widehat{\theta}_{[0]} \leftarrow 0$, $\lambda_{[0]} \leftarrow \|\nabla \mathcal{L}(0)\|_\infty$, $\eta_\lambda \leftarrow \left(\frac{\lambda_{[N]}}{\lambda_{[0]}}\right)^{\frac{1}{N}}$

For: $K = 1, \ldots, N$

$\lambda_{[K]} \leftarrow \eta_\lambda \lambda_{[K-1]}$, $\theta^{(0)}_{[K]} \leftarrow \widehat{\theta}_{[K-1]}$, $\varepsilon_{[K]} \leftarrow \varepsilon_{[N]}$

$\widehat{\theta}_{[K]} \leftarrow \text{Prox-Alg}\left(y, X, \lambda_{[K]}, \theta^{(0)}_{[K]}, \varepsilon_{[K]}\right)$

End For

Return: $\widehat{\theta}_{[N]}$

As can be seen in Algorithm 3, the pathwise optimization scheme starts with

$$\lambda_{[0]} = \|\nabla \mathcal{L}(0)\|_\infty = \left\| \frac{X^\top y}{\sqrt{n}\|y\|_2} \right\|_\infty,$$

which yields an all zero solution $\widehat{\theta}_{[0]} = 0$ (null fit). We then gradually decrease the regularization parameter, and accordingly, the number of nonzero coordinates gradually increases.

The next theorem proves that there exists an $N_1 < N$ such that the fast convergence of the proximal algorithms holds for all $\lambda_{[K]}$'s, where $K \in [N_1 + 1, .., N]$.

[1] We only provide partial theoretical guarantees.

Theorem 4.1. Suppose the design matrix X is sub-Gaussian, and $\lambda_{[N]} = C_1\sqrt{\log d/n}$. For $n \geq C_2 s^* \log d$ and $\eta_\lambda \in (\frac{5}{6}, 1)$, the following results hold:
(I) There exists an $N_1 < N$ such that

$$r > s^*\left(8\lambda_{N_1}/\rho^-_{s^*+\widetilde{s}}\right)^2;$$

(II) For any $K \in [N_1+1, .., N]$, we have $\|\theta^{(0)}_{[K]} - \theta^*\|_2^2 \leq s^*\left(8\lambda_{[K]}/\rho^-_{s^*+\widetilde{s}}\right)^2$, $\theta^{(0)}_{[K]} \in \mathcal{B}_r^{s^*+\widetilde{s}}$ w.h.p.;
(III) Theorems 3.3 and 3.4 hold for all λ_K's, where $K \in [N_1+1, .., N]$ w.h.p..

The proof is provided in Appendix G. Theorem 4.1 implies that for all $\lambda_{[K]}$'s, where $K \in [N_1, N_1+1, .., N]$, the regularization parameter is large enough for ensuring the solution sparsity and preventing overfitting. Therefore, the fast convergence of proximal algorithms can be guaranteed. For $\lambda_{[0]}$ to $\lambda_{[N_1]}$, we do not have theoretical justification for the fast convergence due to the limit of our proof technique. However, as $\lambda_{[0]}$,..., $\lambda_{[N_1]}$ are all larger than $\lambda_{[N_1+1]}$, we can expect that the obtained model is very unlikely to be overfitted. Accordingly, we can also expect that all intermediate solutions $\widehat{\theta}_{[K]}$'s stay out of the nonsmooth region, and LRSC, LRSS, and LRHS properties should also hold along the solution path. Therefore, the proximal algorithms achieve fast convergence in practice. Note that when the design X is normalized, we have $\lambda_{[0]} = \mathcal{O}(d)$, which implies that the total number N of regularization parameter satisfies

$$N = \mathcal{O}(\log d).$$

A geometric illustration of the pathwise optimization is provided in Figure 2. The supporting numerical experiments are provided in Section 6.

5 Extension to CMR and SPME

We extend our algorithm and theory to calibrated multivariate regression (CMR, Liu et al. (2015)) and sparse precision matrix estimation (SPME, Liu et al. (2017)). Due to space limit, we only provide a brief discussion and omit the detailed theoretical deviation.
Extension to CMR. Recall that CMR solves

$$\overline{\Theta}^{\mathsf{CMR}} = \underset{\theta \in \mathbb{R}^{d\times m}}{\operatorname{argmin}} \frac{1}{\sqrt{n}} \sum_{k=1}^{m} \|Y_{*k} - X\Theta_{*k}\|_2 + \lambda_{\mathsf{CMR}}\|\Theta\|_{1,2}.$$

Similar to SQRT-Lasso, we choose a sufficiently large λ_{CMR} to prevent overfitting. Thus, we can expect

$$\|Y_{*k} - X\Theta_{*k}\|_2 \neq 0 \text{ for all } k = 1, ..., m,$$

and treat the nonsmooth loss of CMR as a differentiable function. Accordingly, we can trim our algorithms and theory for the nonsmooth loss of CMR, and establish fast convergence guarantees, as we discussed in §4.

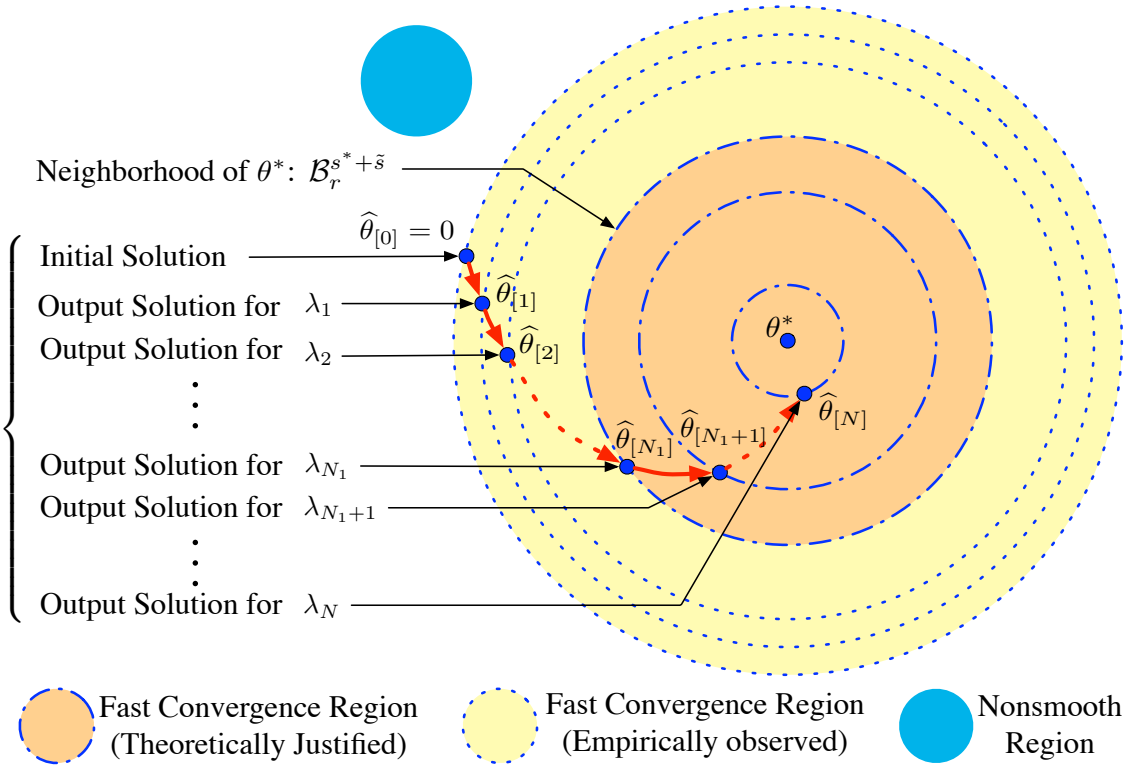

Figure 2: A geometric illustration for the fast convergence of the proximal algorithms. The proximal algorithms combined with the pathwise optimization scheme suppress the overfitting and yield sparse solutions along the solution path. Therefore, the nonsmooth region of the ℓ_2 loss, i.e., the set $\{\theta : y - X\theta = 0\}$, is avoided, and LRSC, LRSS, and LRHS enable the proximal algorithms to achieve fast convergence.

Extension to SPME. Liu et al. (2017) show that a $d \times d$ sparse precision matrix estimation problem is equivalent to a collection of d sparse linear model estimation problems. For each linear model, we apply SQRT-Lasso to estimate the regression coefficient vector and the standard deviation of the random noise. Since SQRT-Lasso is adaptive to imhomogenous noise, we can use one singular regularization parameter to prevent overfitting for all SQRT-Lasso problems. Accordingly, we treat the nonsmooth loss function in every SQRT-Lasso problem as a differentiable function, and further establish fast convergence guarantees for the proximal algorithms combined with the pathwise optimization scheme.

6 Numerical Experiments

We compare the computational performance of the proximal algorithms with other competing algorithms using both synthetic and real data. All algorithms are implemented in `C++` with double precision using a PC with an Intel 2.4GHz Core i5 CPU and 8GB memory. All algorithms are

combined with the pathwise optimization scheme to boost the computational performance. Due to space limit, we omit some less important details.

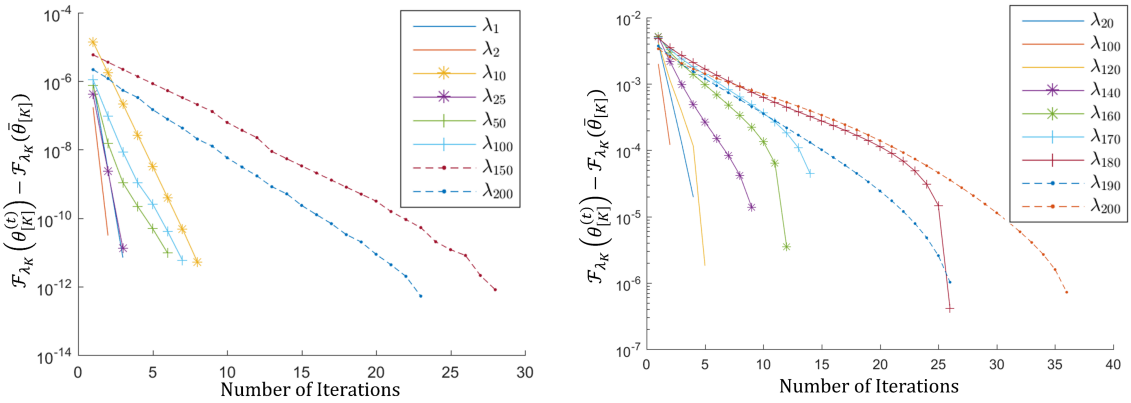

Figure 3: The objective gap v.s. the number of iterations. We can see that the Prox-GD (Left) and Prox-Newton (Right) algorithms achieve linear and quadratic convergence at every stage respectively.

Synthetic Data: For synthetic data, we generate a training dataset of 200 samples, where each row of the design matrix $X_{i*}$is 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 $s^* = 3$ with $\theta^*_1 = 3$, $\theta^*_2 = -2$, and $\theta^*_4 = 1.5$, and $\theta^*_j = 0$ for all $j \neq 1, 2, 4$. The response vector is generated by $y = X\theta^* + \epsilon$, where ϵ is sampled from $N(0, \sigma^2 I)$.

We first show the **fast convergence** of the proximal algorithms at **every stage** of the pathwise optimization scheme. Here we set $\sigma = 0.5$, $N = 200$, $\lambda_N = \sqrt{\log d/n}$, $\varepsilon_K = 10^{-6}$ for all $K = 1, \ldots, N$. Figure 3 presents the objective gap versus the number of iterations. We can see that the proximal algorithms achieves linear (prox-GD) and quadratic (prox-Newton) convergence at every stage. Since the solution sparsity levels are different at each stage, the slopes of these curves are also different.

Next, we show that the computational performance of the pathwise optimization scheme under different settings. Table 2 presents the timing performance of Prox-GD combined with the pathwise optimization scheme. We can see that $N = 10$ actually leads to better timing performance than $N = 1$. That is because when $N = 1$, the solution path does not fall into the local fast convergence region as illustrated in Figure 2. We can also see that the timing performance of Prox-GD is not sensitive to σ. Moreover, we see that the minimal residual sum of squares along the solution path is much larger than 0, thus the overfitting is prevented and the Prox-GD algorithm enjoys the smoothness of the ℓ_2 loss.

Real Data: We adopt two data sets. The first one is the Greenhouse Gas Observing Network Data Set Lucas et al. (2015), which contains 2921 samples and 5232 variables. The second one is the DrivFace data set, which contains 606 samples and 6400 variables. We compare our proximal algorithms with ADMM in Li et al. (2015), Coordinate Descent (CD) in Ndiaye et al. (2016), Prox-

Table 2: Computational performance of Prox-GD on synthetic data under different choices of variance σ, the number of stages N, and the stopping criterion ε_N. The training time is presented, where each entry is the mean execution time in seconds over 100 random trials. The minimal mean square error (MSE) is $\frac{1}{n}\|y - X\widehat{\theta}_{[K]}\|_2^2$, where $\widehat{\theta}_{[K]}$ is the optimal solution that attains $\min \mathcal{F}_{\lambda_K}(\theta)$ for all stages $K = 1, \ldots, N$.

σ	N	ε_N			Minimal MSE	σ	ε_N			Minimal MSE
		10^{-4}	10^{-5}	10^{-6}			10^{-4}	10^{-5}	10^{-6}	
0.1	1	0.3718	0.3721	0.3647	0.0132	0.5	0.2850	0.2951	0.2886	0.3054
	10	**0.2749**	**0.2764**	**0.2804**			**0.1646**	**0.1698**	**0.1753**	
	30	0.3364	0.3452	0.3506			0.2207	0.2247	0.2285	
1	1	0.2347	0.2478	0.2618	1.1833	2	0.4317	0.4697	0.4791	4.2197
	10	**0.1042**	**0.1031**	**0.1091**			**0.1661**	**0.1909**	**0.2110**	
	30	0.2172	0.2221	0.2199			0.2701	0.2955	0.3134	

Table 3: Timing comparison between multiple algorithms on real data. Each entry is the execution time in seconds. All experiments are conducted to achieve similar suboptimality.

Data Set	SQRT-Lasso						Lasso
	Prox-GD	Newton	ADMM	ScalReg	CD	Alt.Min	PISTA
Greenhouse	5.812	**1.708**	1027.590	3180.747	14.311	99.814	5.113
DrivFace	**0.421**	0.426	18.879	124.032	3.138	17.691	0.414

GD (solving Lasso) in Xiao and Zhang (2013) and Alternating Minimization (Alt.Min.) Sun and Zhang (2012) and ScalReg (a simple variant of Alt. Min) in Sun and Sun (2013). Table 3 presents the timing performance of the different algorithms. We can see that Prox-GD for solving SQRT-Lasso significantly outperforms the competitors, and is almost as efficient as Prox-GD for solving Lasso. Prox-Newton is even more efficient than Prox-GD.

Sparse Precision Matrix Estimation. We compare the proximal algorithms with ADMM and CD over real data sets for precision matrix estimation. Particularly, we use four real world biology data sets preprocessed by Li and Toh (2010): Arabidopsis ($d = 834$), Lymph ($d = 587$), Estrogen ($d = 692$), Leukemia ($d = 1,225$). We set three different values for λ_N such that the obtained estimators achieve different levels of sparse recovery. We set $N = 10$, and $\varepsilon_K = 10^{-4}$ for all K's. The timing performance is summarized in Table 4. Prox-GD for solving SQRT-Lasso significantly outperforms the competitors, and is almost as efficient as Prox-GD for solving Lasso. Prox-Newton is even more efficient than Prox-GD.

Calibrated Multivariate Regression. We compare the proximal algorithms with ADMM and CD for CMR on both synthetic data and DrivFace data. For synthetic data, the data generating scheme

Table 4: Timing comparison between multiple algorithms for sparse precision matrix estimation on biology data under different levels of sparsity recovery. Each entry is the execution time in seconds. All experiments are conducted to achieve similar suboptimality. Here CD failed to converge and the program aborted before reaching the desired suboptimality. Scalreg failed to terminate in 1 hour for Estrogen.

Sparsity	Arabidopsis					
	Prox-GD	Newton	ADMM	ScalReg	CD	Alt.Min
1%	5.099	**1.264**	292.05	411.74	12.02	183.63
3%	6.201	**2.088**	339.22	426.08	18.18	217.72
5%	7.122	**2.258**	366.67	435.50	28.60	256.97
Sparsity	Estrogen					
1%	108.24	**3.099**	1597.41	>3600	136.181	634.128
3%	130.93	**7.101**	1845.60	>3600	332.028	662.232
5%	143.54	**10.120**	2029.61	>3600	588.407	739.464
Sparsity	Lymph					
1%	3.709	**0.625**	256.43	354.93	7.208	120.25
3%	4.819	**0.905**	289.08	355.30	10.51	130.61
5%	4.891	**1.123**	310.16	358.70	14.95	148.92
Sparsity	Leukemia					
1%	8.542	**2.715**	331.28	610.147	173.319	239.247
3%	10.562	**3.935**	384.74	766.072	174.295	285.127
5%	10.768	**4.712**	442.54	1274.38	288.884	333.611

Table 5: Timing comparison between multiple algorithms for calibrated multivariate regression on synthetic and real data with different values of λ_N. Each entry is the execution time in seconds. All experiments are conducted to achieve similar suboptimality. Here CD failed to converge and the program aborted before reaching the desired suboptimality.

λ_N	Synthetic ($\sigma = 1$)				DrivFace			
	Prox-GD	Newton	ADMM	CD	Prox-GD	Newton	ADMM	CD
$\sqrt{\log d/n}$	0.2964	**0.0320**	14.8307	2.4098	9.5621	**0.2186**	158.8559	12.7693
$2\sqrt{\log d/n}$	0.1725	**0.0213**	2.2307	2.2272	8.6883	**0.1603**	129.3729	20.4183
$4\sqrt{\log d/n}$	0.0478	**0.0112**	1.8683	1.3656	1.8236	**0.0924**	94.3733	19.1710

is the same as Liu et al. (2015). Table 5 presents the timing performance. Prox-GD for solving SQRT-Lasso significantly outperforms the competitors, and is almost as efficient as Prox-GD for solving Lasso. Prox-Newton is even more efficient than Prox-GD. CD failed to converge and the program aborted before reaching the desired suboptimality.

7 Discussion and Conclusions

This paper shows that although the loss function in the SQRT-Lasso optimization problem is nonsmooth, we can directly apply the proximal gradient and Newton algorithms with fast convergence. First, the fast convergence rate can be established locally in a neighborhood of θ^*. Note that, due to the limited analytical tools, we are not able to directly extend the analysis to establish a global fast convergence rate. Instead, we resort to the pathwise optimization scheme, which helps establishing empirical global fast convergence for the proximal algorithms as illustrated in Figure 2. Specifically, in the early stage of pathwise scheme, with a large regularization parameter λ, the solution quickly falls into the neighborhood of θ^*, where the problem enjoys good properties. After that, the algorithm can quickly converges to θ^* thanks to the fast local convergence property. Our results corroborate that exploiting modeling structures of machine learning problems is of great importance from both computational and statistical perspectives.

Moreover, we remark that to establish the local fast convergence rate, we prove the restricted strong convexity, smoothness, and Hessian smoothness hold over a neighborhood of θ^*. Rigorously establishing the global fast convergence, however, requires these conditions to hold along the solution path. We conjecture that these conditions do hold because our empirical results show the proximal algorithms indeed achieve fast convergence along the entire solution path of the pathwise optimization. We will look for more powerful analytic tools and defer a sharper characterization to the future effort.

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*.

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

Bertsekas, D. P. (1999). *Nonlinear programming*. Athena scientific Belmont.

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

Boyd, S. and Vandenberghe, L. (2009). *Convex Optimization*. Cambridge University Press.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

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

Fan, J., Liu, H., Sun, Q. and Zhang, T. (2015). Tac for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *arXiv preprint arXiv:1507.01037* .

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics* **1** 302–332.

Lee, J. D., Sun, Y. and Saunders, M. A. (2014). Proximal newton-type methods for minimizing composite functions. *SIAM Journal on Optimization* **24** 1420–1443.

Li, L. and Toh, K.-C. (2010). An inexact interior point method for ℓ_1-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., Wang, L. and Zhao, T. (2015). Calibrated multivariate regression with application to neural semantic basis discovery. *Journal of Machine Learning Research* **16** 1579–1606.

Liu, H., Wang, L. et al. (2017). Tiger: A tuning-insensitive approach for optimally estimating gaussian graphical models. *Electronic Journal of Statistics* **11** 241–294.

Lucas, D. D., Yver Kwok, C., Cameron-Smith, P., Graven, H., Bergmann, D., Guilderson, T. P., Weiss, R. and Keeling, R. (2015). Designing optimal greenhouse gas observing networks that

consider performance and cost. *Geoscientific Instrumentation, Methods and Data Systems* **4** 121–137.
URL https://www.geosci-instrum-method-data-syst.net/4/121/2015/

Ndiaye, E., Fercoq, O., Gramfort, A., Leclère, V. and Salmon, J. (2016). Efficient smoothed concomitant lasso estimation for high dimensional regression. *arXiv preprint arXiv:1606.02702* .

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.

Nemirovski, A. (2004). Interior point polynomial time methods in convex programming. *Lecture Notes* .

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

Nesterov, Y. (2013). Gradient methods for minimizing composite functions. *Mathematical Programming* **140** 125–161.

Ning, Y., Zhao, T. and Liu, H. (2014). A likelihood ratio framework for high dimensional semiparametric regression. *arXiv preprint arXiv:1412.2295* .

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.

Sun, T. and Sun, M. T. (2013). Package 'scalreg' .

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898.

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

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

Wang, L. (2013). The l1 penalized lad estimator for high dimensional linear regression. *Journal of Multivariate Analysis* **120** 135–151.

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 ℓ_q loss in ℓ_r balls. *The Journal of Machine Learning Research* **11** 3519–3540.

Zhao, T., Liu, H. and Zhang, T. (2014). Pathwise coordinate optimization for sparse learning: Algorithm and theory. *arXiv preprint arXiv:1412.7477* .

A Proof of Lemma 3.2

Part 1. We first show the claim on λ. By $y = X\theta^* + \epsilon$ and (A.5), we have

$$\nabla \mathcal{L}(\theta^*) = \frac{X^\top (X\theta^* - y)}{\sqrt{n}\|y - X\theta^*\|_2} = -\frac{X^\top \epsilon}{\sqrt{n}\|\epsilon\|_2}. \tag{A.1}$$

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

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

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

Lemma A.1. Assume X satisfies $\|\mathbf{x}_j\|_2 \leq \sqrt{n}$ for all $j \in \{1, \ldots, d\}$ and ϵ 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 $\mathbb{P}\left[\frac{1}{n}\|X^\top \epsilon\|_\infty \geq 2\sigma\sqrt{\frac{\log d}{n}}\right] \leq 2d^{-1}$.

Combining (A.1), (A.2) and Lemma A.1, we have $\|\nabla \mathcal{L}(\theta^*)\|_\infty \leq 4\sqrt{\log d/n}$ with probability at least $1 - 2d^{-1} - \exp\left(-\frac{n}{32}\right)$.

Part 2. Next, we show that LRSC, LRSS, and LRHS holds. First, for correlated sub-Gaussian random design with the covariance satisfying the bounded eigenvalues, we have from (Rudelson and Zhou, 2013) that the design matrix X satisfies the RE condition with high probability given $n \geq cs^* \log d$, i.e., for any $v \in \mathcal{B}_r^{s^*+\widetilde{s}} = \mathcal{B}_r \cap \{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_0 \leq s^* + \widetilde{s}\}$,

$$\psi_{\min}\|v\|_2^2 - \varphi_{\min}\frac{\log d}{n}\|v\|_1^2 \leq \frac{\|Xv\|_2^2}{n} \leq \psi_{\max}\|v\|_2^2 + \varphi_{\max}\frac{\log d}{n}\|v\|_1^2, \tag{A.3}$$

where $\psi_{\min}, \psi_{\max}, \varphi_{\min}, \varphi_{\max} \in (0, \infty)$ are generic constants. The RE condition has been extensively studied for sparse recovery (Candès and Tao, 2005; Bickel et al., 2009; Raskutti et al., 2010).

We divide the proof into three steps.

Step 1. When X satisfies the RE condition, i.e.

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

Denote $s = s^* + 2\widetilde{s}$. Since $\|v\|_0 \leq s$, which implies $\|v\|_1^2 \leq s\|v\|_2^2$, then we have

$$\left(\psi_{\min} - \varphi_{\min}\frac{s\log d}{n}\right)\|v\|_2^2 \leq \frac{\|Xv\|_2^2}{n}$$
$$\left(\psi_{\max} + \varphi_{\max}\frac{s\log d}{n}\right)\|v\|_2^2 \geq \frac{\|Xv\|_2^2}{n},$$

Then there exists a universal constant c_1 such that if $n \geq c_1 s^* \log d$, we have

$$\frac{1}{2}\psi_{\min}\|v\|_2^2 \leq \frac{\|Xv\|_2^2}{n} \leq 2\psi_{\max}\|v\|_2^2. \tag{A.4}$$

Step 2. Conditioning on (A.4), we show that $\mathcal{L}$ satisfies LRSC and LRSS with high probability. The gradient of $\mathcal{L}(\theta)$ is

$$\nabla\mathcal{L}(\theta) = \frac{1}{\sqrt{n}}\left(\left(\frac{\partial\|y - X\theta\|_2}{\partial(y - X\theta)}\right)^\top \left(\frac{\partial(y - X\theta)}{\partial\theta}\right)^\top\right)^\top = \frac{X^\top(X\theta - y)}{\sqrt{n}\|y - X\theta\|_2}. \tag{A.5}$$

The Hessian of $\mathcal{L}(\theta)$ is

$$\nabla^2\mathcal{L}(\theta) = \frac{1}{n}\frac{\partial(-X^\top\widetilde{z})}{\partial\theta} = \frac{1}{\sqrt{n}\|y - X\theta\|_2}X^\top\left(I - \frac{(y - X\theta)(y - X\theta)^\top}{\|y - X\theta\|_2^2}\right)X. \tag{A.6}$$

For notational convenience, we define $\Delta = v - w$ for any $v, w \in \mathcal{B}_r^{s^*+\widetilde{s}}$. Also denote the residual of the first order Taylor expansion as $\delta\mathcal{L}(w + \Delta, w) = \mathcal{L}(w + \Delta) - \mathcal{L}(w) - \nabla\mathcal{L}(w)^\top\Delta$. Using the first order Taylor expansion of $\mathcal{L}(\theta)$ at w and the Hessian of $\mathcal{L}(\theta)$ in (A.6), we have from mean value theorem that there exists some $\alpha \in [0, 1]$ such that $\delta\mathcal{L}(w + \Delta, w) = \frac{1}{\sqrt{n}\|\xi\|_2}\Delta^\top X^\top\left(I - \frac{\xi\xi^\top}{\|\xi\|_2^2}\right)X\Delta$, where $\xi = y - X(w + \alpha\Delta)$. For notational simplicity, let's denote $\dot{z} = X(v - \theta^*)$ and $\ddot{z} = X(w - \theta^*)$, which can be considered as two fixed vectors in $\mathbb{R}^n$. Without loss of generality, assume $\|\dot{z}\|_2 \leq \|\ddot{z}\|_2$. Then we have

$$\|\dot{z}\|_2^2 \leq \|\ddot{z}\|_2^2 \leq 2\psi_{\max}n\|w - \theta^*\|_2^2 \leq \frac{n\sigma^2}{4}.$$

Further, we have

$$\xi = y - X(w + \alpha\Delta) = \epsilon - X(w + \alpha\Delta - \theta^*) = \epsilon - \alpha\dot{z} - (1 - \alpha)\ddot{z}, \text{ and } X\Delta = \dot{z} - \ddot{z}.$$

We have from Wainwright (2015) that

$$\mathbb{P}\left[\|\epsilon\|_2^2 \leq n\sigma^2(1 - \delta)\right] \leq \exp\left(-\frac{n\delta^2}{16}\right), \tag{A.7}$$

Then by taking $\delta = 1/3$ in (A.7), we have with probability $1 - \exp\left(-\frac{n}{144}\right)$,

$$\|\xi\|_2 \geq \|\epsilon\|_2 - \alpha\|\dot{z}\|_2 - (1 - \alpha)\|\ddot{z}\|_2 \geq \|\epsilon\|_2 - \|\ddot{z}\|_2 \geq \frac{4}{5}\sqrt{n}\sigma - \frac{1}{2}\sqrt{n}\sigma \geq \frac{1}{4}\sqrt{n}\sigma. \tag{A.8}$$

We first discuss the RSS property. From (A.8), we have

$$\delta\mathcal{L}(w + \Delta, w) = \frac{\Delta^\top X^\top\left(I - \frac{\xi\xi^\top}{\|\xi\|_2^2}\right)X\Delta}{\sqrt{n}\|\xi\|_2} = \frac{\left(\|X\Delta\|_2^2 - \frac{(\xi^\top X\Delta)^2}{\|\xi\|_2^2}\right)}{\sqrt{n}\|\xi\|_2} \leq \frac{\|X\Delta\|_2^2}{\sqrt{n}\|\xi\|_2} \leq \frac{8\psi_{\max}}{\sigma}\|\Delta\|_2^2$$

Next, we verify the RSC property. We want to show that with high probability, for any constant $a \in (0, 3/5]$

$$\left| \frac{\xi^\top}{\|\xi\|_2} X\Delta \right| \leq \sqrt{1-a} \|X\Delta\|_2. \tag{A.9}$$

Consequently, we have

$$\Delta^\top X^\top \left(I - \frac{\xi\xi^\top}{\|\xi\|_2^2} \right) X\Delta = \|X\Delta\|_2^2 - \left(\frac{\xi^\top}{\|\xi\|_2} X\Delta \right)^2 \geq a\|X\Delta\|_2^2.$$

This further implies

$$\delta\mathcal{L}(w+\Delta, w) = \frac{1}{\sqrt{n}\|\xi\|_2} \Delta^\top X^\top \left(I - \frac{\xi\xi^\top}{\|\xi\|_2^2} \right) X\Delta \geq \frac{a\psi_{\min}}{2\|\xi\|_2/\sqrt{n}} \|\Delta\|_2^2. \tag{A.10}$$

Since $\|\dot{z}\|_2 \leq \|\ddot{z}\|_2$, then for any real constant $a \in (0,1)$,

$$\begin{aligned}
&\mathbb{P}\left[\left| \frac{\xi^\top}{\|\xi\|_2} X\Delta \right| \leq \sqrt{1-a} \|X\Delta\|_2 \right] = \mathbb{P}\left[\left| \frac{(\epsilon - \alpha\dot{z} - (1-\alpha)\ddot{z})^\top}{\|\epsilon - \alpha\dot{z} - (1-\alpha)\ddot{z}\|_2} (\dot{z} - \ddot{z}) \right| \leq \sqrt{1-a} \|\dot{z} - \ddot{z}\|_2 \right] \\
&\overset{(i)}{\geq} \mathbb{P}\left[\left| \frac{(\epsilon - \dot{z})^\top (\dot{z} - \ddot{z})}{\|\epsilon - \dot{z}\|_2} \right| \leq \sqrt{1-a} \|\dot{z} - \ddot{z}\|_2 \right] = \mathbb{P}\left[\left(\epsilon^\top(\dot{z} - \ddot{z}) - \dot{z}^\top(\dot{z} - \ddot{z}) \right)^2 \leq (1-a)\|\epsilon - \dot{z}\|_2^2 \|\dot{z} - \ddot{z}\|_2^2 \right] \\
&\overset{(ii)}{\geq} \mathbb{P}\left[\left| \left(\frac{\epsilon^\top(\dot{z} - \ddot{z})}{\|\dot{z} - \ddot{z}\|_2} \right)^2 + \|\dot{z}\|_2^2 - \frac{2\epsilon^\top(\dot{z} - \ddot{z}) z^\top (\dot{z} - \ddot{z})}{\|\dot{z} - \ddot{z}\|_2^2} \right| \leq (1-a)(\|\epsilon\|_2^2 + \|\dot{z}\|_2^2 - 2\epsilon^\top \dot{z}) \right],
\end{aligned} \tag{A.11}$$

where (i) is from a geometric inspection and the randomness of ϵ, i.e., for any $\alpha \in [0,1]$ and $\|\dot{z}\|_2 \leq \|\ddot{z}\|_2$, we have $\left| \frac{-\dot{z}^\top}{\|-\dot{z}\|_2} (\dot{z} - \ddot{z}) \right| \leq \left| \frac{(-\alpha\dot{z} - (1-\alpha)\ddot{z})^\top}{\|-\alpha\dot{z} - (1-\alpha)\ddot{z}\|_2} (\dot{z} - \ddot{z}) \right|$, and (ii) is from dividing both sides by $\|\dot{z} - \ddot{z}\|_2^2$. The random vector ϵ with i.i.d. entries does not affect the inequality above. Let's first discuss one side of the probability in (A.11),

$$\begin{aligned}
&\mathbb{P}\left[\left(\frac{\epsilon^\top(\dot{z} - \ddot{z})}{\|\dot{z} - \ddot{z}\|_2} \right)^2 + \|\dot{z}\|_2^2 - \frac{2\epsilon^\top(\dot{z} - \ddot{z}) z^\top (\dot{z} - \ddot{z})}{\|\dot{z} - \ddot{z}\|_2^2} \leq (1-a)(\|\epsilon\|_2^2 + \|\dot{z}\|_2^2 - 2\epsilon^\top \dot{z}) \right] \\
&= \mathbb{P}\left[(1-a)\|\epsilon\|_2^2 \geq \left(\frac{\epsilon^\top(\dot{z} - \ddot{z})}{\|\dot{z} - \ddot{z}\|_2} \right)^2 - 2(1-a)\epsilon^\top \dot{z} + a\|\dot{z}\|_2^2 + \frac{2\epsilon^\top(\dot{z} - \ddot{z}) z^\top (\dot{z} - \ddot{z})}{\|\dot{z} - \ddot{z}\|_2^2} \right].
\end{aligned} \tag{A.12}$$

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

$$\begin{aligned}
&\mathbb{P}\left[\|\epsilon\|_2^2 \leq n\sigma^2(1-\delta) \right] \leq \exp\left(-\frac{n\delta^2}{16} \right), \mathbb{P}\left[\left(\frac{\epsilon^\top(\dot{z} - \ddot{z})}{\|\dot{z} - \ddot{z}\|_2} \right)^2 \geq n\sigma^2\delta \right] \leq \exp\left(-\frac{n\delta^2}{2} \right), \\
&\mathbb{P}\left[\epsilon^\top \dot{z} \leq -n\sigma^2\delta \right] \leq \exp\left(-\frac{n^2\sigma^2\delta^2}{2\|\dot{z}\|_2^2} \right), \mathbb{P}\left[\frac{\epsilon^\top(\dot{z} - \ddot{z}) z^\top (\dot{z} - \ddot{z})}{\|\dot{z} - \ddot{z}\|_2^2} \geq n\sigma^2\delta \right] \leq \exp\left(-\frac{n^2\sigma^2\delta^2}{2\|\dot{z}\|_2^2} \right).
\end{aligned} \tag{A.13}$$

Combining (A.13) with $\|\dot{z}\|_2^2 \le n\sigma^2/4$, we have from union bound that with probability at least $1-\exp\left(-\frac{n}{144}\right)-\exp\left(-\frac{n}{128}\right)-\exp\left(-\frac{n}{128}\right) \ge 1-3\exp\left(-\frac{n}{144}\right)$,

$$\|\epsilon\|_2^2 \ge \frac{2}{3}n\sigma^2,\ \left(\frac{\epsilon^\top(\dot{z}-\ddot{z})}{\|\dot{z}-\ddot{z}\|_2}\right)^2 \le \frac{1}{64}n\sigma^2,\ -\epsilon^\top\dot{z} \le \frac{1}{16}n\sigma^2, \frac{\epsilon^\top(\dot{z}-\ddot{z})z^\top(\dot{z}-\ddot{z})}{\|\dot{z}-\ddot{z}\|_2^2} \le \frac{1}{16}n\sigma^2.$$

This implies for $a \le 3/5$, we have $\frac{\xi^\top}{\|\xi\|_2}X\Delta \le \sqrt{1-a}\|X\Delta\|_2$. For the other side of (A.11), we have

$$\begin{aligned}
&\mathbb{P}\begin{bmatrix}\left(\frac{\epsilon^\top(\dot{z}-\ddot{z})}{\|\dot{z}-\ddot{z}\|_2}\right)^2 + \|\dot{z}\|_2^2 - 2\epsilon^\top\dot{z} \ge \\ -(1-a)(\|\epsilon\|_2^2 + \|\dot{z}\|_2^2 - 2\epsilon^\top\dot{z})\end{bmatrix} \\
&\overset{(i)}{\ge} \mathbb{P}\begin{bmatrix}-\left(\frac{\epsilon^\top(\dot{z}-\ddot{z})}{\|\dot{z}-\ddot{z}\|_2}\right)^2 - (\|\dot{z}\|_2^2 - 2\epsilon^\top\dot{z}) \ge \\ -(1-a)(\|\epsilon\|_2^2 + \|\dot{z}\|_2^2 - 2\epsilon^\top\dot{z})\end{bmatrix} = \mathbb{P}\left[(1-a)\|\epsilon\|_2^2 \ge \left(\frac{\epsilon^\top(\dot{z}-\ddot{z})}{\|\dot{z}-\ddot{z}\|_2}\right)^2 + a(\|\dot{z}\|_2^2 - 2\epsilon^\top\dot{z})\right]. \qquad \text{(A.14)}
\end{aligned}$$

where (i) is from the fact that $\mathbb{P}[c_1 \ge -c_2] \ge \mathbb{P}[-c_1 \ge -c_2]$ for $c_1, c_2 \ge 0$.

Combining (A.11), (A.12) and (A.14), we have (A.9) holds with high probability, i.e., for any $r > 0$,

$$\mathbb{P}\left[\left|\frac{\xi^\top}{\|\xi\|_2}X\Delta\right| \le \sqrt{1-a}\|X\Delta\|_2\right] \ge 1-6\exp\left(-\frac{n}{144}\right).$$

Now we bound $\|\xi\|_2$ to obtain the desired result. From Wainwright (2015), we have

$$\mathbb{P}\left[\|\epsilon\|_2^2 \ge n\sigma^2(1+\delta)\right] \le \exp\left(-\frac{n\delta^2}{18}\right) = \exp\left(-\frac{n}{72}\right), \qquad \text{(A.15)}$$

where we take $\delta = 1/2$. From $\xi = \epsilon - \alpha\dot{z} - (1-\alpha)\ddot{z}$, we have

$$\|\xi\|_2 \le \|\epsilon\|_2 + \alpha\|\dot{z}\|_2 + (1-\alpha)\|\ddot{z}\|_2 \overset{(i)}{\le} \|\epsilon\|_2 + \|\ddot{z}\|_2 \overset{(ii)}{\le} \sqrt{\frac{3n}{2}}\sigma + \frac{1}{2}\sqrt{n}\sigma. \qquad \text{(A.16)}$$

where (i) is from $\|\dot{z}\|_2 \le \|\ddot{z}\|_2$ and (ii) is from (A.15) and $\|\dot{z}\|_2^2 \le n\sigma^2/4$. Then by the union bound setting $a = 1/2$, with probability at least $1-7\exp\left(-\frac{n}{144}\right)$, we have $\delta\mathcal{L}(w+\Delta, w) \ge \frac{\psi_{\min}}{8\sigma}\|\Delta\|_2^2$. Moreover, we also have $r = \frac{\sigma^2}{8\psi_{\max}}$ for large enough $n \ge c_1 s^* \log d$.

Step 3. Given the proposed conditions, we have that $\mathcal{L}$ satisfies the LRHS property by combining the analysis in Ning et al. (2014).

B Intermediate Results of Theorem 3.3

We introduce some important implications of the proposed assumptions. Recall that $\mathcal{S}^* = \{j : \theta_j^* \ne 0\}$ be the index set of non-zero entries of θ^* with $s^* = |\mathcal{S}^*|$ and $\overline{\mathcal{S}^*} = \{j : \theta_j^* = 0\}$ be the complement set. Lemma 3.2 implies RSC and RSS hold with parameter $\rho^-_{s^*+2\widetilde{s}}$ and $\rho^+_{s^*+2\widetilde{s}}$ respectively. By

Nesterov (2004), the following conditions are equivalent to RSC and RSS, i.e., for any $v, w \in \mathbb{R}^d$ satisfying $\|v - w\|_0 \leq s^* + 2\widetilde{s}$,

$$\rho^-_{s^*+2\widetilde{s}}\|v-w\|_2^2 \leq (v-w)^\top \nabla\mathcal{L}(w) \text{ and } \rho^+_{s^*+2\widetilde{s}}\|v-w\|_2^2 \geq (v-w)^\top \nabla\mathcal{L}(w), \tag{B.1}$$

$$\frac{1}{\rho^+_{s^*+2\widetilde{s}}}\|\nabla\mathcal{L}(v)-\nabla\mathcal{L}(w)\|_2^2 \leq (v-w)^\top \nabla\mathcal{L}(w) \text{ and } \frac{1}{\rho^-_{s^*+2\widetilde{s}}}\|\nabla\mathcal{L}(v)-\nabla\mathcal{L}(w)\|_2^2 \geq (v-w)^\top \nabla\mathcal{L}(w). \tag{B.2}$$

From the convexity of ℓ_1 norm, we have

$$\|v\|_1 - \|w\|_1 \geq (v-w)^\top g, \tag{B.3}$$

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

$$\mathcal{F}_\lambda(v) - \mathcal{F}_\lambda(w) - (v-w)^\top \nabla\mathcal{F}_\lambda(w) \geq \rho^-_{s^*+2\widetilde{s}}\|v-w\|_2^2, \tag{B.4}$$

Remark B.1. For any t and k, the line search satisfies

$$\widetilde{L}^{(t)} \leq L^{(t)} \leq L_{\max},\ L \leq \widetilde{L}^{(t)} \leq L^{(t)} \leq 2L \text{ and } \rho^+_{s^*+2\widetilde{s}} \leq \widetilde{L}^{(t)} \leq L^{(t)} \leq 2\rho^+_{s^*+2\widetilde{s}}, \tag{B.5}$$

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

We first show that when θ is sparse and the approximate KKT condition is satisfied, then both estimation error and objective error, w.r.t. the true model parameter, are bounded. This is formalized in Lemma B.2, and its proof is deferred to Appendix J.1.

Lemma B.2. Suppose conditions in Lemma 3.2 hold and $s^*\left(8\lambda/\rho^-_{s^*+\widetilde{s}}\right)^2 < r$. If θ satisfies $\|\theta_{\overline{\mathcal{S}}^*}\|_0 \leq \widetilde{s}$ and the approximate KKT condition $\min_{g\in\partial\|\theta\|_1}\|\nabla\mathcal{L}(\theta) + \lambda g\|_\infty \leq \lambda/2$, then we have

$$\|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1 \leq 5\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1, \tag{B.6}$$

$$\|\theta-\theta^*\|_2 \leq \frac{2\lambda\sqrt{s^*}}{\rho^-_{s^*+\widetilde{s}}} \leq \frac{2\lambda\sqrt{s^*}}{\rho^-_{s^*+2\widetilde{s}}}, \tag{B.7}$$

$$\|\theta-\theta^*\|_1 \leq \frac{12\lambda s^*}{\rho^-_{s^*+\widetilde{s}}} \leq \frac{12\lambda s^*}{\rho^-_{s^*+2\widetilde{s}}}, \tag{B.8}$$

$$\mathcal{F}_\lambda(\theta) - \mathcal{F}_\lambda(\theta^*) \leq \frac{6\lambda^2 s^*}{\rho^-_{s^*+\widetilde{s}}} \leq \frac{6\lambda^2 s^*}{\rho^-_{s^*+2\widetilde{s}}}. \tag{B.9}$$

Next, we show that if θ is sparse and the objective error is bounded, then the estimation error is also bounded. This is formalized in Lemma B.3, and its proof is deferred to Appendix J.2.

Lemma B.3. Suppose conditions in Lemma 3.2 hold and $s^*\left(8\lambda/\rho^-_{s^*+\widetilde{s}}\right)^2 < r$. If θ satisfies $\|\theta_{\overline{\mathcal{S}}^*}\|_0 \leq \widetilde{s}$ and the objective satisfies $\mathcal{F}_\lambda(\theta) - \mathcal{F}_\lambda(\theta^*) \leq \frac{6\lambda^2 s^*}{\rho^-}$, where ρ^- can be either $\rho^-_{s^*+\widetilde{s}}$ or $\rho^-_{s^*+2\widetilde{s}}$, then we have

$$\|\theta-\theta^*\|_2 \leq \frac{4\lambda\sqrt{3s^*}}{\rho^-}, \tag{B.10}$$

$$\|\theta-\theta^*\|_1 \leq \frac{24\lambda s^*}{\rho^-}. \tag{B.11}$$

We then show that if θ is sparse and the objective error is bounded, then each proximal-gradient update preserves solution to be sparse. This is formalized in Lemma B.4, and its proof is deferred to Appendix J.3.

Lemma B.4. Suppose conditions in Lemma 3.2 hold and $s^*\left(8\lambda/\rho^-_{s^*+\widetilde{s}}\right)^2 < r$. If θ satisfies $\|\theta_{\overline{\mathcal{S}}^*}\|_0 \le \widetilde{s}$, L satisfies $L < 2\rho^+_{s^*+2\widetilde{s}}$, and the objective satisfies $\mathcal{F}_\lambda(\theta) - \mathcal{F}_\lambda(\theta^*) \le \frac{6\lambda^2 s^*}{\rho^-_{s^*+2\widetilde{s}}}$, then we have $\|(\mathcal{T}_{L,\lambda}(\theta))_{\overline{\mathcal{S}}^*}\|_0 \le \widetilde{s}$.

Moreover, we show that if θ satisfies the approximate KKT condition, then the objective has a bounded error w.r.t. the regularizaSuppose conditions in Lemma 3.2 with parameter λ. This characterizes the geometric decrease of the objective error when we choose a geometrically decreasing sequence of regularization parameters. This is formalized in Lemma B.5, and its proof is deferred to Appendix J.4.

Lemma B.5. . If θ satisfies $\omega_\lambda(\theta) \le \lambda/2$, then for $\overline{\theta} = \operatorname{argmin}_\theta \mathcal{F}_\lambda(\theta)$, we have $\mathcal{F}_\lambda(\theta) - \mathcal{F}_\lambda(\overline{\theta}) \le \frac{24\lambda\omega_\lambda(\theta)s^*}{\rho^-_{s^*+2\widetilde{s}}}$.

Furthermore, we show 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 B.6, and its proof is deferred to Appendix J.5.

Lemma B.6. Suppose conditions in Lemma 3.2 hold and $s^*\left(8\lambda/\rho^-_{s^*+\widetilde{s}}\right)^2 < r$s. If the initialization $\theta^{(0)}$ satisfies $\|\theta^{(0)}\|_0 \le \widetilde{s}$. Then with $\overline{\theta} = \operatorname{argmin}_\theta \mathcal{F}_\lambda(\theta)$, for any $t = 1, 2, \ldots$, we have $\|\theta^{(t)}\|_0 \le \widetilde{s}$ and $\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta}) \le \left(1 - \frac{1}{8\kappa_{s^*+2\widetilde{s}}}\right)^t \left(\mathcal{F}_\lambda(\theta^{(0)}) - \mathcal{F}_\lambda(\overline{\theta})\right)$.

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 B.7 (Adapted from Theorem 2 in Nesterov (2013))**.** For any $L > 0$,

$$\mathcal{Q}_\lambda\left(\mathcal{T}_{L,\lambda}(\theta), \theta\right) \le \mathcal{F}_\lambda\left(\theta\right) - \frac{L}{2}\|\mathcal{T}_{L,\lambda}(\theta) - \theta\|_2^2.$$

Besides, if $\mathcal{L}(\theta)$ is convex, we have

$$\mathcal{Q}_\lambda\left(\mathcal{T}_{L,\lambda}(\theta), \theta\right) \le \min_{\mathbf{x}} \mathcal{F}_\lambda\left(\mathbf{x}\right) + \frac{L}{2}\|\mathbf{x} - \theta\|_2^2. \tag{B.12}$$

Further, we have for any $L \ge L$,

$$\mathcal{F}_\lambda\left(\mathcal{T}_{L,\lambda}(\theta)\right) \le \mathcal{Q}_\lambda\left(\mathcal{T}_{L,\lambda}(\theta), \theta\right) \le \mathcal{F}_\lambda\left(\theta\right) - \frac{L}{2}\|\mathcal{T}_{L,\lambda}(\theta) - \theta\|_2^2. \tag{B.13}$$

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

Lemma B.8 (Adapted from Lemma 2 in Xiao and Zhang (2013))**.** For any $L > 0$, if L is the Lipschitz constant of $\nabla\mathcal{L}$, then

$$\omega_\lambda\left(\mathcal{T}_{L,\lambda}(\theta)\right) \le (L + S_L(\theta))\|\mathcal{T}_{L,\lambda}(\theta) - \theta\|_2 \le 2L\|\mathcal{T}_{L,\lambda}(\theta) - \theta\|_2,$$

where $S_L(\theta) = \frac{\|\nabla\mathcal{L}(\mathcal{T}_{L,\lambda}(\theta)) - \nabla\mathcal{L}(\theta)\|_2}{\|\mathcal{T}_{L,\lambda}(\theta) - \theta\|_2}$ is a local Lipschitz constant, which satisfies $S_L(\theta) \le L$.

C Proof of Theorem 3.3

We demonstrate the linear rate when the initial value $\theta^{(0)}$ satisfies $\omega_\lambda(\theta^{(0)}) \le \frac{\lambda}{2}$ with $\|(\theta^{(0)})_{\overline{\mathcal{S}}^*}\|_0 \le \widetilde{s}$. The proof is provided in Appendix H.

Theorem C.1. Suppose conditions in Lemma 3.2 hold and $s^*\left(8\lambda/\rho^-_{s^*+\widetilde{s}}\right)^2 < r$. Let $\overline{\theta} = \operatorname{argmin}_\theta \mathcal{F}_\lambda(\theta)$ be the optimal solution with regularization parameter λ. If the initial value $\theta^{(0)}$ satisfies $\omega_\lambda(\theta^{(0)}) \le \frac{\lambda}{2}$ with $\|(\theta^{(0)})_{\overline{\mathcal{S}}^*}\|_0 \le \widetilde{s}$, then for any $t = 1, 2, \ldots$, we have $\|(\theta^{(t)})_{\overline{\mathcal{S}}^*}\|_0 \le \widetilde{s}$,

$$\|\theta^{(t)} - \overline{\theta}\|_2^2 \le \left(1 - \frac{1}{8\kappa_{s^*+2\widetilde{s}}}\right)^t \frac{24\lambda s^*\omega_\lambda(\theta^{(t)})}{(\rho^-_{s^*+2\widetilde{s}})^2} \quad \text{and} \tag{C.1}$$

$$\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta}) \le \left(1 - \frac{1}{8\kappa_{s^*+2\widetilde{s}}}\right)^t \frac{24\lambda s^*\omega_\lambda(\theta^{(t)})}{\rho^-_{s^*+2\widetilde{s}}}, \tag{C.2}$$

In addition, to achieve the approximate KKT condition $\omega_\lambda(\theta^{(t)}) \le \varepsilon$, the number of proximal gradient steps is no more than

$$\frac{\log\left(96\left(1 + \kappa_{s^*+2\widetilde{s}}\right)^2\lambda^2 s^*\kappa_{s^*+2\widetilde{s}}/\varepsilon^2\right)}{\log\left(8\kappa_{s^*+2\widetilde{s}}/(8\kappa_{s^*+2\widetilde{s}} - 1)\right)}. \tag{C.3}$$

From basic inequalities, since $\kappa_{s^*+2\widetilde{s}} \ge 1$, we have $\log\left(\frac{8\kappa_{s^*+2\widetilde{s}}}{8\kappa_{s^*+2\widetilde{s}}-1}\right) \ge \log\left(1 + \frac{1}{8\kappa_{s^*+2\widetilde{s}}-1}\right) \ge \frac{1}{8\kappa_{s^*+2\widetilde{s}}}$. Then (C.3) can be simplified as $\mathcal{O}\left(\kappa_{s^*+2\widetilde{s}}\left(\log\left(\kappa^3_{s^*+2\widetilde{s}}\lambda^2 s^*/\varepsilon^2\right)\right)\right)$.

As can be seen from Theorem C.1, when the initial value $\theta^{(0)}$ satisfies $\omega_\lambda(\theta^{(0)}) \le \frac{\lambda}{2}$ with $\|(\theta^{(0)})_{\overline{\mathcal{S}}^*}\|_0 \le \widetilde{s}$, then we can guarantee the geometric convergence rate of the estimated objective value towards the minimal objective.

Next, we need to show that when $\theta_{(0)} \in \mathcal{B}_r$, the approximate KKT holds for $\theta_{(1)}$, which is also sparse. We demonstrate this result in Lemma C.2 and provide its proof in Appendix I.

Lemma C.2. Suppose conditions in Lemma 3.2 hold and $s^*\left(8\lambda/\rho^-_{s^*+\widetilde{s}}\right)^2 < r$.s. If $\frac{\rho^-_{s^*+\widetilde{s}}}{8}\sqrt{\frac{r}{s^*}} > \lambda$ and $\|\theta - \theta^*\|_2^2 \le r$ holds, then we have $\omega_\lambda(\theta) \le 4\sqrt{r}$ and $\|\theta_{\overline{\mathcal{S}}^*}\|_0 \le \widetilde{s}$.

Combining the results above, we finish the proof.

D Proof of Theorem 3.4

We present a few important intermediate results that are key components of our main proof. The first result shows that in a neighborhood of the true model parameter θ^*, the sparsity of the solution is preserved when we use a sparse initialization. The proof is provided in Appendix K.1.

Lemma D.1 (Sparsity Preserving Lemma)**.** Suppose conditions in Lemma 3.2 hold and $s^*\left(8\lambda/\rho^-_{s^*+\widetilde{s}}\right)^2 < r$ with $\varepsilon \le \frac{\lambda}{8}$. Given $\theta^{(t)} \in \mathcal{B}(\theta^*, R)$ and $\|\theta^{(t)}_{\overline{\mathcal{S}}}\|_0 \le \widetilde{s}$, there exists a generic constant C_1 such that

$$\|\theta^{(t+1)}_{\overline{\mathcal{S}}}\|_0 \le \widetilde{s}, \ \ \|\theta^{(t+1)} - \theta^*\|_2 \le \frac{C_1\lambda\sqrt{s^*}}{\rho^-_{s^*+2\widetilde{s}}} \quad \text{and} \quad \mathcal{F}_\lambda(\theta^{(t)}) \le \mathcal{F}_\lambda(\theta^*) + \frac{15\lambda^2 s^*}{4\rho^-_{s^*+2\widetilde{s}}}..$$

Denote $\mathcal{B}(\theta, r) = \left\{\phi \in \mathbb{R}^d \mid \|\phi - \theta\|_2 \le r\right\}$. We then show that every step of proximal Newton updates within each stage has a quadratic convergence rate to a local minimizer, if we start with a sparse solution in the refined region. The proof is provided in Appendix K.2.

Lemma D.2. Suppose conditions in Lemma 3.2 hold and $s^*\left(8\lambda/\rho^-_{s^*+\widetilde{s}}\right)^2 < r$. If $\theta^{(t)} \in \mathcal{B}(\theta^*, r)$ and $\left\|\theta^{(t)}_{\overline{\mathcal{S}}}\right\|_0 \le \widetilde{s}$, then for each stage $K \ge 2$, we have

$$\|\theta^{(t+1)} - \overline{\theta}\|_2 \le \frac{L_{s^*+2\widetilde{s}}}{2\rho^-_{s^*+2\widetilde{s}}}\|\theta^{(t)} - \overline{\theta}\|_2^2.$$

In the following, we need to use the property that the iterates $\theta^{(t)} \in \mathcal{B}(\overline{\theta}, 2r)$ instead of $\theta^{(t)} \in \mathcal{B}(\theta^*, r)$ for convergence analysis of the proximal Newton method. This property holds since we have $\theta^{(t)} \in \mathcal{B}(\theta^*, r)$ and $\overline{\theta} \in \mathcal{B}(\theta^*, r)$ simultaneously. Thus $\theta^{(t)} \in \mathcal{B}\left(\overline{\theta}, 2r\right)$, where $2r = \frac{\rho^-_{s^*+2\widetilde{s}}}{L_{s^*+2\widetilde{s}}}$ is the radius for quadratic convergence region of the proximal Newton algorithm.

The following lemma demonstrates that the step size parameter is simply 1 if the the sparse solution is in the refined region. The proof is provided in Appendix K.3.

Lemma D.3. Suppose conditions in Lemma 3.2 hold and $s^*\left(8\lambda/\rho^-_{s^*+\widetilde{s}}\right)^2 < r$. If $\theta^{(t)} \in \mathcal{B}(\overline{\theta}, 2r)$ and $\|\theta^{(t)}_{\overline{\mathcal{S}}}\|_0 \le \widetilde{s}$ at each stage $K \ge 2$ with $\frac{1}{4} \le \alpha < \frac{1}{2}$, then $\eta_t = 1$. Further, we have

$$\mathcal{F}_\lambda(\theta^{(t+1)}) \le \mathcal{F}_\lambda(\theta^{(t)}) + \frac{1}{4}\gamma_t.$$

Moreover, we present a critical property of γ_t. The proof is provided in Appendix K.4.

Lemma D.4. Denote $\Delta\theta^{(t)} = \theta^{(t)} - \theta^{(t+1)}$ and

$$\gamma_t = \nabla\mathcal{L}(\theta^{(t)})^\top\Delta\theta^{(t)} + \left\|\lambda(\theta^{(t)} + \Delta\theta^{(t)})\right\|_1 - \left\|\lambda(\theta^{(t)})\right\|_1.$$

Then we have $\gamma_t \le -\|\Delta\theta^{(t)}\|^2_{\nabla^2\mathcal{L}(\theta^{(t)})}$.

In addition, we present the sufficient number of iterations for each convex relaxation stage to achieve the approximate KKT condition. The proof is provided in Appendix K.5.

Lemma D.5. Suppose conditions in Lemma 3.2 hold and $s^*\left(8\lambda/\rho^-_{s^*+\widetilde{s}}\right)^2 < r$. To achieve the approximate KKT condition $\omega_\lambda\left(\theta^{(t)}\right) \le \varepsilon$ for any $\varepsilon > 0$ at each stage $K \ge 2$, the number of iteration for proximal Newton updates is at most

$$\log\log\left(\frac{3\rho^+_{s^*+2\widetilde{s}}}{\varepsilon}\right).$$

Combining the results above, we have desired results in Theorem 3.4.

E Proof of Theorem 3.6

Part 1. We first show that estimation errors are as claimed. We have that $\omega_\lambda(\widehat{\theta}^{(0)}) \le \lambda/2$. By Theorem C.1, we have for any $t = 1, 2, \ldots$, $\|(\theta^{(t)}_{[K+1]})_{\overline{\mathcal{S}}^*}\|_0 \le \widetilde{s}$. Applying Lemma B.2 recursively, we have

$$\|\widehat{\theta} - \theta^*\|_2 \le \frac{2\lambda\sqrt{s^*}}{\rho^-_{s^*+2\widetilde{s}}} \quad \text{and} \quad \|\widehat{\theta} - \theta^*\|_1 \le \frac{12\lambda s^*}{\rho^-_{s^*+2\widetilde{s}}}.$$

Applying Lemma 3.2 with $\lambda \le 24\sqrt{\log d/n}$ and $\rho^-_{s^*+2\widetilde{s}} = \frac{\psi_{\min}}{8\sigma}$, then by union bound, with probability at least $1 - 8\exp\left(-\frac{n}{144}\right) - 2d^{-1}$, we have

$$\|\widehat{\theta} - \theta^*\|_2 \le \frac{384\sigma\sqrt{s^*\log d/n}}{\psi_{\min}} \text{ and } \|\widehat{\theta} - \theta^*\|_1 \le \frac{2304\sigma s^*\sqrt{\log d/n}}{\psi_{\min}}.$$

Part 2. Next, we demonstrate the result of the estimation of variance. Let $\overline{\theta} = \operatorname{argmin}_\theta \mathcal{F}_\lambda(\theta)$ be the optimal solution. Apply the argument in Part 1 recursively, we have

$$\|\overline{\theta} - \theta^*\|_1 \le \frac{2304\sigma s^*\sqrt{\log d/n}}{\psi_{\min}}. \tag{E.1}$$

Denote $c_1, c_2, \ldots$ as positive universal constants. Then we have

$$\begin{aligned}\mathcal{L}(\overline{\theta}) - \mathcal{L}(\theta^*) &\le \lambda(\|\theta^*\|_1 - \|\overline{\theta}\|_1) \le \lambda(\|\theta^*_{\mathcal{S}^*}\|_1 - \|(\overline{\theta})_{\mathcal{S}^*}\|_1 - \|(\overline{\theta})_{\overline{\mathcal{S}}^*}\|_1) \\ &\le \lambda\|(\overline{\theta} - \theta^*)_{\mathcal{S}^*}\|_1 \le \lambda\|\overline{\theta} - \theta^*\|_1 \overset{(i)}{\le} c_1\frac{\sigma s^*\log d}{n},\end{aligned} \tag{E.2}$$

where (i) is from the value of λ and ℓ_1 error bound in (E.1).

On the other hand, from the convexity of $\mathcal{L}(\theta)$, we have

$$\begin{aligned}\mathcal{L}(\overline{\theta}) - \mathcal{L}(\theta^*) &\ge (\overline{\theta} - \theta^*)^\top\nabla\mathcal{L}(\theta^*) \ge -\|\nabla\mathcal{L}(\theta^*)\|_\infty\|\widehat{\theta} - \theta\|_1 \\ &\overset{(i)}{\ge} -c_2\lambda\|\overline{\theta} - \theta\|_1 \overset{(ii)}{\ge} -c_3\frac{\sigma s^*\log d}{n},\end{aligned} \tag{E.3}$$

where (i) is from Lemma 3.2 and (ii) value of λ and ℓ_1 error bound in (E.1). By definition, we have

$$\mathcal{L}(\overline{\theta}) - \mathcal{L}(\theta^*) = \frac{\|y - X\overline{\theta}\|_2}{\sqrt{n}} - \frac{\|\epsilon\|_2}{\sqrt{n}}. \tag{E.4}$$

From Wainwright (2015), we have for any $\delta > 0$,

$$\mathbb{P}\left[\left|\frac{\|\epsilon\|_2^2}{n} - \sigma^2\right| \geq \sigma^2\delta\right] \leq 2\exp\left(-\frac{n\delta^2}{18}\right). \tag{E.5}$$

Combining (E.2), (E.3), (E.4) and (E.5) with $\delta^2 = \frac{c_3 s^* \log d}{n}$, we have with high probability,

$$\left|\frac{\|y - X\overline{\theta}\|_2}{\sqrt{n}} - \sigma\right| = \mathcal{O}\left(\frac{\sigma s^* \log d}{n}\right). \tag{E.6}$$

From Part 1, for $n \geq c_4 s^* \log d$, with high probability, we have $\|\overline{\theta} - \theta^*\|_2 \leq \frac{384\sigma\sqrt{s^* \log d/n}}{\psi_{\min}} \leq \frac{\sigma}{2\sqrt{2\psi_{\max}}}$, then $\overline{\theta} \in \mathcal{B}_r^{s^*+\widetilde{s}}$ and $\|\widehat{\theta} - \overline{\theta}\|_0 \leq s^* + 2\widetilde{s}$. Then from the analysis of Theorem C.1, we have

$$\omega_\lambda(\theta^{(t+1)}) \leq (1 + \kappa_{s^*+2\widetilde{s}})\sqrt{4\rho^+_{s^*+2\widetilde{s}}\left(\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta})\right)} \leq \varepsilon.$$

This implies

$$\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta}) \leq \frac{\epsilon^2}{4\rho^+_{s^*+2\widetilde{s}}(1 + \kappa_{s^*+2\widetilde{s}})^2}. \tag{E.7}$$

On the other hand, from the LRSC property of $\mathcal{L}$, convexity of ℓ_1 norm and optimality of $\overline{\theta}$, we have

$$\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta}) \geq \rho^-_{s^*+2\widetilde{s}}\|\widehat{\theta} - \overline{\theta}\|_2^2. \tag{E.8}$$

Combining (E.7), (E.8) and Lemma 3.2, we have

$$\frac{\|X(\widehat{\theta} - \overline{\theta})\|_2}{\sqrt{n}} \leq \sqrt{\frac{8\rho^+_{s^*+2\widetilde{s}}}{\sigma}}\|\widehat{\theta} - \theta^*\|_2 \leq \sqrt{\frac{2}{\sigma\rho^-_{s^*+2\widetilde{s}}}}\frac{\epsilon}{(1 + \kappa_{s^*+2\widetilde{s}})} \leq \frac{4\epsilon}{(1 + \kappa_{s^*+2\widetilde{s}})\sqrt{\psi_{\min}}}. \tag{E.9}$$

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

$$\left|\frac{\|y - X\widehat{\theta}\|_2}{\sqrt{n}} - \frac{\|y - X\overline{\theta}\|_2}{\sqrt{n}}\right| \leq \frac{\|X(\widehat{\theta} - \overline{\theta})\|_2}{\sqrt{n}} \leq \frac{4\epsilon}{(1 + \kappa_{s^*+2\widetilde{s}})\sqrt{\psi_{\min}}}.$$

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

F Intermediate Results of Theoremm 4.1

We first characterize the sparsity of $\widehat{\theta}$ and its distance to θ^* when approximate KKT condition holds in Lemma F.1 and provide its proof in Appendix L.1.

Lemma F.1. Suppose conditions in Lemma 3.2 hold and $s^*\left(8\lambda/\rho^-_{s^*+\widetilde{s}}\right)^2 < r$, and the approximate KKT satisfies $\omega_\lambda(\theta) \leq \lambda/4$. If $\frac{\rho^-_{s^*+\widetilde{s}}}{8}\sqrt{\frac{r}{s^*}} > \lambda > \lambda_{[N]}$, then we have

$$\|\theta - \theta^*\|_2^2 \leq r \quad \text{and} \quad \|\theta_{\overline{\mathcal{S}}^*}\|_0 \leq \widetilde{s}.$$

Next, we show that if the optimal solution $\widehat{\theta}_{[K-1]}$ 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 $\widehat{\theta}_{[K-1]}$ 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 $\theta^{(0)}_{[K]} = \widehat{\theta}_{[K-1]}$ and geometric sequence of regularization parameters $\lambda_{[K]}$. We formalize this notion in Lemma F.2, and its proof is provided in Appendix L.2.

Lemma F.2. Let $\widehat{\theta}_{[K-1]}$ be the approximate solution of $K-1$-th path following state, which satisfies the approximate KKT condition $\omega_{\lambda_{[K-1]}}(\widehat{\theta}_{[K-1]}) \le \lambda_{[K-1]}/4$. Then we have

$$\omega_{\lambda_{[K]}}(\widehat{\theta}_{[K-1]}) \le \lambda_{[K]}/2,$$

where $\lambda_{[K]} = \eta_\lambda \lambda_{[K-1]}$ with $\eta_\lambda \in (5/6, 1)$.

G Proof of Theorem 4.1

Part 1. We first show the existence of N_1. Following the notation in Appendix A, $r = \frac{\sigma^2}{8\psi_{\max}}$ is a constant independent of n. As a result for a large enough $n > C_2 s^* \log d$, we have

$$r = \frac{\sigma^2}{8\psi_{\max}} \overset{(i)}{>} s^* \left(64\sigma\lambda_{[N_1]}/\psi_{\min}\right)^2 \overset{(ii)}{\ge} s^* \left(8\lambda_{[N_1]}/\rho^-_{s^*+\widetilde{s}}\right)^2,$$

where (i) is from $n > C_2 s^* \log d$ with a sufficiently large constant C_2 and $\lambda_{[N_1]} = \frac{1}{\eta_\lambda}^{N-N_1} \lambda_{[N]} = \frac{1}{\eta_\lambda}^{N-N_1} C_1 \sqrt{\frac{\log d}{n}}$, and (ii) is from $\rho^-_{s^*+2\widetilde{s}} \ge \frac{\psi_{\min}}{8\sigma}$.

Part 2. We next show that for $K \in [N_1, ..., N-1]$, $\lambda_{[K]}$, $\widehat{\theta}_{[K]}$ is a good initial for $\theta^{(0)}_{[K+1]}$, i.e., satisfies

$$\|\widehat{\theta}_{[K]} - \theta^*\|_2^2 \le s^* \left(8\lambda_{[K+1]}/\rho^-_{s^*+\widetilde{s}}\right)^2 \text{ and } \|[\widehat{\theta}_{[K]}]_{\overline{\mathcal{S}}^*}\|_0 \le \widetilde{s}.$$

Lemma F.1 directly implies $\|[\widehat{\theta}_{[K]}]_{\overline{\mathcal{S}}^*}\|_0 \le \widetilde{s}$. Applying Lemma F.2, we have $\omega_{\lambda_{[K+1]}}(\widehat{\theta}_{[K]}) \le \lambda_{[K+1]}/2$. Then we can apply Lemma B.2, we have $\|\widehat{\theta}_{[K]} - \theta^*\|_2^2 \le (2\lambda_{[K+1]}\sqrt{s^*}/\rho^-_{s^*+\widetilde{s}})^2 \le s^* \left(8\lambda_{[K+1]}/\rho^-_{s^*+\widetilde{s}}\right)^2$

Part 3. So far we prove that for $K \in [N_1 + 1, ..., N-1]$,

$$\|\theta^{(0)}_{[K]} - \theta^*\|_2^2 \le s^* \left(8\lambda_{[K]}/\rho^-_{s^*+\widetilde{s}}\right)^2 \le s^* \left(8\lambda_{[N_1]}/\rho^-_{s^*+\widetilde{s}}\right)^2 < r \text{ and } \|[\theta^{(0)}_{[K]}]_{\overline{\mathcal{S}}^*}\|_0 \le \widetilde{s}.$$

So the fast convergence rate in Theorems 3.3 and 3.4 hold for λ_K.

H Proof of Theorem C.1

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

$$\rho^+_{s^*+2\widetilde{s}} \le \widetilde{L}^{(t)} \le 2\rho^+_{s^*+2\widetilde{s}}. \tag{H.1}$$

Since the initialization $\theta^{(0)}$ satisfies $\omega_\lambda(\theta^{(0)}) \le \frac{\lambda}{2}$ with $\|(\theta^{(0)})_{\overline{\mathcal{S}}^*}\|_0 \le \widetilde{s}$, then by Lemma B.2, we have $\mathcal{F}_\lambda(\theta^{(0)}) - \mathcal{F}_\lambda(\theta^*) \le \frac{6\lambda^2 s^*}{\rho^-_{s^*+2\widetilde{s}}}$. Then by Lemma B.4, we have $\|(\theta^{(1)})_{\overline{\mathcal{S}}^*}\|_0 \le \widetilde{s}$.

By monotone decrease of $\mathcal{F}_\lambda(\theta^{(t)})$ from (B.13) in Lemma B.7 and recursively applying Lemma B.4, $\|(\theta^{(t)})_{\overline{\mathcal{S}}^*}\|_0 \le \widetilde{s}$ holds in (C.2) for all $t = 1, 2, \ldots$.

For the objective error, we have

$$\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta}) \overset{(i)}{\le} \left(1 - \frac{1}{8\kappa_{s^*+2\widetilde{s}}}\right)^t \left(\mathcal{F}_\lambda(\theta^{(0)}) - \mathcal{F}_\lambda(\overline{\theta})\right) \overset{(ii)}{\le} \left(1 - \frac{1}{8\kappa_{s^*+2\widetilde{s}}}\right)^t \frac{24\lambda s^* \omega_\lambda(\theta^{(t)})}{\rho^-_{s^*+2\widetilde{s}}}, \tag{H.2}$$

where (i) is from Lemma B.6, and (ii) is from Lemma B.5 and $\omega_\lambda(\theta^{(t+1)}) \le \lambda/2 \le \lambda$, which results in (C.2).

Combining (H.2), (B.4) with $\nabla\mathcal{F}_\lambda(\overline{\theta}) = 0$, we have

$$\|\theta^{(t)} - \overline{\theta}\|_2^2 \le \frac{1}{\rho^-_{s^*+2\widetilde{s}}}\left(\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta}) - \nabla\mathcal{F}_\lambda(\overline{\theta})\right) \le \left(1 - \frac{1}{8\kappa_{s^*+2\widetilde{s}}}\right)^t \frac{24\lambda s^* \omega_\lambda(\theta^{(t)})}{(\rho^-_{s^*+2\widetilde{s}})^2}$$

For $\omega_\lambda(\theta^{(t+1)})$ of $(t+1)$-th iteration, we have

$$\begin{aligned}
&\omega_\lambda(\theta^{(t+1)}) \\
&\overset{(i)}{\le} \left(\widetilde{L}^{(t)} + S_{\widetilde{L}^{(t)}}(\theta^{(t)})\right)\|\theta^{(t+1)} - \theta^{(t)}\|_2 \overset{(ii)}{\le} \left(\widetilde{L}^{(t)} + \rho^+_{s^*+2\widetilde{s}}\right)\|\theta^{(t+1)} - \theta^{(t)}\|_2 \\
&\overset{(iii)}{\le} \widetilde{L}^{(t)}\left(1 + \frac{\rho^+_{s^*+2\widetilde{s}}}{\rho^-_{s^*+2\widetilde{s}}}\right)\|\theta^{(t+1)} - \theta^{(t)}\|_2 \overset{(iv)}{\le} \widetilde{L}^{(t)}\left(1 + \frac{\rho^+_{s^*+2\widetilde{s}}}{\rho^-_{s^*+2\widetilde{s}}}\right)\sqrt{\frac{2\left(\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\theta^{(t+1)})\right)}{\widetilde{L}^{(t)}}} \\
&\overset{(v)}{\le} (1 + \kappa_{s^*+2\widetilde{s}})\sqrt{4\rho^+_{s^*+2\widetilde{s}}\left(\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta})\right)} \overset{(vi)}{\le} (1 + \kappa_{s^*+2\widetilde{s}})\sqrt{96\lambda^2 s^* \kappa_{s^*+2\widetilde{s}}\left(1 - \frac{1}{8\kappa_{s^*+2\widetilde{s}}}\right)^t},
\end{aligned} \tag{H.3}$$

where (i) is from Lemma B.8, (ii) is from $S_{\widetilde{L}^{(t)}}(\theta^{(t)}) \le \rho^+_{s^*+2\widetilde{s}}$, (iii) is from $\rho^-_{s^*+2\widetilde{s}} \le \widetilde{L}^{(t)}$ in (H.1), (iv) is from (B.13) in Lemma B.7, (v) is from $\widetilde{L}^{(t)} \le 2\rho^+_{s^*+2\widetilde{s}}$ in (H.1) and monotone decrease of $\mathcal{F}_\lambda(\theta^{(t)})$ from (B.13) in Lemma B.7, and (vi) is from (H.2) and $\kappa_{s^*+2\widetilde{s}} = \frac{\rho^+_{s^*+2\widetilde{s}}}{\rho^-_{s^*+2\widetilde{s}}}$.

Then we need $\omega_\lambda(\widehat{\theta}) \le \varepsilon \le \lambda/4$. Set the R.H.S. of (H.3) to be no greater than ε, which is equivalent to require the number of iterations k to be an upper bound of (C.3).

I Proof of Lemma C.2

Part 1. We first show that given $\|\theta^{(0)} - \theta^*\|_2^2 \le r$, $\omega_\lambda(\theta^{(1)}) \le 4\sqrt{r}$ holds. From Lemma B.8, we have

$$\omega_\lambda(\theta^{(1)}) \le 2L\|\theta^{(1)} - \theta^{(0)}\|_2 \le 4\|\theta^{(1)} - \theta^*\|_2 \le 4\sqrt{r}.$$

Part 2. We next demonstrate the sparsity of θ. From $\lambda \ge 6\|\nabla\mathcal{L}(\theta^*)\|_\infty$, then we have

$$\left|\left\{i \in \overline{\mathcal{S}}^* : |\nabla_i\mathcal{L}(\theta^*)| \ge \frac{\lambda}{6}\right\}\right| = 0. \tag{I.1}$$

Denote $\check{\mathcal{S}}_1 = \left\{i \in \overline{\mathcal{S}}^* : |\nabla_i \mathcal{L}(\theta) - \nabla_i \mathcal{L}(\theta^*)| \geq \frac{2\lambda}{3}\right\}$ and $\check{s}_1 = |\check{\mathcal{S}}_1|$. Then there exists some $\boldsymbol{b} \in \mathbb{R}^d$ such that $\|\boldsymbol{b}\|_\infty = 1$, $\|\boldsymbol{b}\|_0 \leq \check{s}_1$ and $\boldsymbol{b}^\top(\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta^*)) \geq \frac{2\lambda\check{s}_1}{3}$. Then by the mean value theorem, we have for some $\check{\theta} = (1-\alpha)\theta + \alpha\theta^*$ with $\alpha \in [0,1]$, $\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta^*) = \nabla^2\mathcal{L}(\check{\theta})\Delta$, where $\Delta = \theta - \theta^*$. Then we have

$$\frac{2\lambda\check{s}_1}{3} \leq \boldsymbol{b}^\top\nabla^2\mathcal{L}(\check{\theta})\Delta \overset{(i)}{\leq} \sqrt{\boldsymbol{b}^\top\nabla^2\mathcal{L}(\check{\theta})\boldsymbol{b}}\sqrt{\Delta^\top\nabla^2\mathcal{L}(\check{\theta})\Delta} \overset{(ii)}{\leq} \sqrt{\check{s}_1\rho^+_{\check{s}_1}}\sqrt{\Delta^\top(\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta^*))}, \tag{I.2}$$

where (i) is from the generalized Cauchy-Schwarz inequality, (ii) is from the definition of RSS and the fact that $\|\boldsymbol{b}\|_2 \leq \sqrt{\check{s}_1}\|\boldsymbol{b}\|_\infty = \sqrt{\check{s}_1}$. Let g achieve $\min_{g\in\partial\|\theta\|_1} \mathcal{F}_\lambda(\theta)$. Further, we have

$$\begin{aligned}
&\Delta^\top(\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta^*)) \leq \|\Delta\|_1\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta^*)\|_\infty \leq \|\Delta\|_1(\|\nabla\mathcal{L}(\theta^*)\|_\infty + \|\nabla\mathcal{L}(\theta)\|_\infty) \\
&\leq \|\Delta\|_1(\|\nabla\mathcal{L}(\theta^*)\|_\infty + \|\nabla\mathcal{L}(\theta) + \lambda g\|_\infty + \lambda\|g\|_\infty) \overset{(i)}{\leq} \frac{28\lambda s^*}{3\rho^-_{s^*+\widetilde{s}}}(\frac{\lambda}{6} + \frac{\lambda}{4} + \lambda) \leq \frac{14\lambda^2 s^*}{\rho^-_{s^*+\widetilde{s}}},
\end{aligned} \tag{I.3}$$

where (i) is from $\|\widetilde{\Delta}_{\overline{\mathcal{S}}^*}\|_1 \leq \frac{5}{2}\|\widetilde{\Delta}_{\mathcal{S}^*}\|_1$ and $\|\widetilde{\Delta}_{\mathcal{S}^*}\|_1 \leq \frac{8\lambda s^*}{3\rho^-_{s^*+\widetilde{s}}}$, condition on λ, approximate KKT condition and $\|g\|_\infty \leq 1$. Combining (I.2) and (I.3), we have $\frac{2\sqrt{\check{s}_1}}{3} \leq \sqrt{\frac{14\rho^+_{\check{s}_1}s^*}{\rho^-_{s^*+\widetilde{s}}}}$, which further implies

$$\check{s}_1 \leq \frac{32\rho^+_{\check{s}_1}s^*}{\rho^-_{s^*+\widetilde{s}}} \leq 32\kappa_{s^*+2\widetilde{s}}s^* \leq \widetilde{s}. \tag{I.4}$$

For any $v \in \mathbb{R}^d$ that satisfies $\|v\|_0 \leq 1$, we have

$$\check{\mathcal{S}}_2 = \left\{i \in \overline{\mathcal{S}}^* : \left|\nabla_i\mathcal{L}(\theta) + \frac{\lambda}{4}v_i\right| \geq \frac{5\lambda}{6}\right\} \subseteq \left\{i \in \overline{\mathcal{S}}^* : |\nabla_i\mathcal{L}(\theta^*)| \geq \frac{\lambda}{6}\right\} \bigcup \check{\mathcal{S}}_1.$$

Then we have $|\check{\mathcal{S}}_2| \leq |\check{\mathcal{S}}_1| \leq \widetilde{s}$. Since for any $i \in \overline{\mathcal{S}}^*$ and $\left|\nabla_i\mathcal{L}(\theta) + \frac{\lambda}{4}v_i\right| < \frac{5\lambda}{6}$, we can find g_i that satisfies $|g_i| \leq 1$ such that $\nabla_i\mathcal{L}(\theta) + \frac{\lambda}{4}v_i + \lambda g_i = 0$ which implies $\theta_i = 0$, then we have

$$\left|\left\{i \in \overline{\mathcal{S}}^* : \left|\nabla_i\mathcal{L}(\theta) + \frac{\lambda}{4}v_i\right| < \frac{5\lambda}{6}\right\}\right| = 0.$$

This implies $\|\theta_{\overline{\mathcal{S}}^*}\|_0 \leq |\check{\mathcal{S}}_2| \leq \widetilde{s}$.

J Proofs of Intermediate Lemmas in Appendix B

J.1 Proof of Lemma B.2

We first bound the estimation error. From Lemma 3.2, we have the RSC property, which indicates

$$\mathcal{L}(\theta) \geq \mathcal{L}(\theta^*) + (\theta - \theta^*)^\top\nabla\mathcal{L}(\theta^*) + \frac{\rho^-_{s^*+\widetilde{s}}}{2}\|\theta - \theta^*\|_2^2, \tag{J.1}$$

$$\mathcal{L}(\theta^*) \geq \mathcal{L}(\theta) + (\theta^* - \theta)^\top\nabla\mathcal{L}(\theta) + \frac{\rho^-_{s^*+\widetilde{s}}}{2}\|\theta - \theta^*\|_2^2, \tag{J.2}$$

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

$$(\theta - \theta^*)^\top\nabla\mathcal{L}(\theta) \geq (\theta - \theta^*)^\top\nabla\mathcal{L}(\theta^*) + \rho^-_{s^*+\widetilde{s}}\|\theta - \theta^*\|_2^2. \tag{J.3}$$

Let $g \in \partial\|\theta\|_1$ be the subgradient that achieves the approximate KKT condition, then

$$(\theta-\theta^*)^\top(\nabla\mathcal{L}(\theta)+\lambda g) \le \|\theta-\theta^*\|_1 \|\nabla\mathcal{L}(\theta)+\lambda g\|_\infty \le \frac{1}{2}\lambda\|\theta-\theta^*\|_1. \tag{J.4}$$

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

$$(\theta-\theta^*)^\top(\nabla\mathcal{L}(\theta)+\lambda g) \ge (\theta-\theta^*)^\top\nabla\mathcal{L}(\theta^*)+\rho^-_{s^*+\widetilde{s}}\|\theta-\theta^*\|_2^2+\lambda g^\top(\theta-\theta^*), \tag{J.5}$$

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

$$(\theta-\theta^*)^\top\nabla\mathcal{L}(\theta^*) \ge -\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1\|\mathcal{L}(\theta^*)\|_\infty - \|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1\|\mathcal{L}(\theta^*)\|_\infty. \tag{J.6}$$

Besides, we have

$$\begin{aligned}(\theta-\theta^*)^\top g &= g_{\mathcal{S}^*}^\top(\theta-\theta^*)_{\mathcal{S}^*} + g_{\overline{\mathcal{S}}^*}^\top(\theta-\theta^*)_{\overline{\mathcal{S}}^*} \overset{(i)}{\ge} -\|g_{\mathcal{S}^*}\|_\infty\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1 + g_{\overline{\mathcal{S}}^*}^\top\theta_{\overline{\mathcal{S}}^*} \\ &\overset{(ii)}{\ge} -\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1 + \|g_{\overline{\mathcal{S}}^*}\|_1 \overset{(iii)}{=} -\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1 + \|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1,\end{aligned} \tag{J.7}$$

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

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

$$\frac{1}{2}\lambda\|\theta-\theta^*\|_1 \ge \rho^-_{s^*+\widetilde{s}}\|\theta-\theta^*\|_2^2 - (\lambda+\|\mathcal{L}(\theta^*)\|_\infty)\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1 + (\lambda-\|\mathcal{L}(\theta^*)\|_\infty)\|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1.$$

This implies

$$\rho^-_{s^*+\widetilde{s}}\|\theta-\theta^*\|_2^2 + (\frac{1}{2}\lambda-\|\mathcal{L}(\theta^*)\|_\infty)\|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1 \le (\frac{3}{2}\lambda+\|\mathcal{L}(\theta^*)\|_\infty)\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1, \tag{J.8}$$

which results in (B.6) from $\rho^-_{s^*+2\widetilde{s}}>0$ and Lemma 3.2 as

$$\|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1 \le \frac{\frac{3}{2}\lambda+\|\mathcal{L}(\theta^*)\|_\infty}{\frac{1}{2}\lambda-\|\mathcal{L}(\theta^*)\|_\infty}\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1.$$

Combining $\frac{1}{2}\lambda-\|\mathcal{L}(\theta^*)\|_\infty \ge 0$, $\frac{3}{2}\lambda+\|\mathcal{L}(\theta^*)\|_\infty \le 2\lambda$ and (J.8), we have estimation errors in (B.7) and (B.8) as

$$\rho^-_{s^*+2\widetilde{s}}\|\theta-\theta^*\|_2^2 \le 2\lambda\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1 \le 2\lambda\sqrt{s^*}\|\theta-\theta^*\|_2 \ \text{ and } \ \|\theta-\theta^*\|_1 \le 6\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1 \le 6\sqrt{s^*}\|\theta-\theta^*\|_2.$$

Next, we bound the objective error in (B.9). We have

$$\begin{aligned}&\mathcal{F}_\lambda(\theta)-\mathcal{F}_\lambda(\theta^*) \overset{(i)}{\le} -(\nabla\mathcal{L}(\theta)+\lambda g)^\top(\theta^*-\theta) \le \|\nabla\mathcal{L}(\theta)+\lambda g\|_\infty\|\theta^*-\theta\|_1 \le \frac{1}{2}\lambda\|\theta^*-\theta\|_1 \\ &= \frac{1}{2}\lambda(\|(\theta^*-\theta)_{\mathcal{S}^*}\|_1 + \|(\theta^*-\theta)_{\overline{\mathcal{S}}^*}\|_1) \overset{(ii)}{\le} 3\lambda\|(\theta^*-\theta)_{\mathcal{S}^*}\|_1 \le 3\lambda\sqrt{s^*}\|(\theta^*-\theta)_{\mathcal{S}^*}\|_2 \overset{(iii)}{\le} \frac{6\lambda^2 s^*}{\rho^-_{s^*+2\widetilde{s}}},\end{aligned}$$

where (i) is from the convexity of $\mathcal{F}_\lambda(\theta)$ with subgradient $\nabla\mathcal{L}(\theta)+\lambda g$, (ii) is from (B.6), and (iii) is from (B.7).

J.2 Proof of Lemma B.3

Recall that ρ^- can be either $\rho^-_{s^*+\widetilde{s}}$ or $\rho^-_{s^*+2\widetilde{s}}$. Assumption $\mathcal{F}_\lambda(\theta)-\mathcal{F}_\lambda(\theta^*)\le 6\lambda^2 s^*/\rho^-$ implies

$$\mathcal{L}(\theta)-\mathcal{L}(\theta^*)+\lambda(\|\theta\|_1-\|\theta^*\|_1)\le\frac{6\lambda^2 s^*}{\rho^-}. \tag{J.9}$$

We have from the RSC property that

$$\mathcal{L}(\theta)\ge\mathcal{L}(\theta^*)+(\theta-\theta^*)^\top\nabla\mathcal{L}(\theta^*)+\frac{\rho^-}{2}\|\theta-\theta^*\|_2^2, \tag{J.10}$$

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

$$\frac{\rho^-}{2}\|\theta-\theta^*\|_2^2\le\frac{6\lambda^2 s^*}{\rho^-}-(\theta-\theta^*)^\top\nabla\mathcal{L}(\theta^*)+\lambda(\|\theta^*\|_1-\|\theta\|_1). \tag{J.11}$$

Besides, we have

$$(\theta-\theta^*)^\top\nabla\mathcal{L}(\theta^*)\ge-\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1\|\mathcal{L}(\theta^*)\|_\infty-\|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1\|\mathcal{L}(\theta^*)\|_\infty,\ \text{and} \tag{J.12}$$

$$\|\theta^*\|_1-\|\theta\|_1=\|\theta^*_{\mathcal{S}^*}\|_1-\|\theta_{\mathcal{S}^*}\|_1-\|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1\le\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1-\|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1. \tag{J.13}$$

Combining (J.11), (J.12) and (J.13), we have

$$\frac{\rho^-}{2}\|\theta-\theta^*\|_2^2\le\frac{6\lambda^2 s^*}{\rho^-}+(\|\nabla\mathcal{L}(\theta^*)\|_\infty+\lambda)\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1+(\|\nabla\mathcal{L}(\theta^*)\|_\infty-\lambda)\|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1. \tag{J.14}$$

We discuss two cases as following:

Case 1. We first assume $\|\theta-\theta^*\|_1\le\frac{12\lambda s^*}{\rho^-}$. Then (J.14) implies

$$\frac{\rho^-}{2}\|\theta-\theta^*\|_2^2\overset{(i)}{\le}\frac{6\lambda^2 s^*}{\rho^-}+(\|\nabla\mathcal{L}(\theta^*)\|_\infty+\lambda)\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1\overset{(ii)}{\le}\frac{6\lambda^2 s^*}{\rho^-}+\frac{3}{2}\lambda\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1\le\frac{24\lambda^2 s^*}{\rho^-}.$$

where (i) is from $\|\nabla\mathcal{L}(\theta^*)\|_\infty-\lambda\le 0$ and (ii) is from $\|\nabla\mathcal{L}(\theta^*)\|_\infty+\lambda\le\frac{3}{2}\lambda$. This indicates

$$\|\theta-\theta^*\|_2\le\frac{4\sqrt{3s^*}\lambda}{\rho^-}. \tag{J.15}$$

Case 2. Next, we assume $\|\theta-\theta^*\|_1>\frac{12\lambda s^*}{\rho^-}$. Then (J.14) implies

$$\begin{aligned}
&\frac{\rho^-}{2}\|\theta-\theta^*\|_2^2\\
&\le(\|\nabla\mathcal{L}(\theta^*)\|_\infty+\lambda)\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1+(\|\nabla\mathcal{L}(\theta^*)\|_\infty-\lambda)\|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1+\frac{1}{2}\lambda\|\theta-\theta^*\|_1\\
&=(\|\nabla\mathcal{L}(\theta^*)\|_\infty+\frac{3}{2}\lambda)\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1+(\|\nabla\mathcal{L}(\theta^*)\|_\infty-\frac{1}{2}\lambda)\|(\theta-\theta^*)_{\overline{\mathcal{S}}^*}\|_1\\
&\overset{(i)}{\le}2\lambda\|(\theta-\theta^*)_{\mathcal{S}^*}\|_1\le 2\sqrt{s^*}\lambda\|(\theta-\theta^*)_{\mathcal{S}^*}\|_2,
\end{aligned} \tag{J.16}$$

where (i) is from $\|\nabla\mathcal{L}(\theta^*)\|_\infty + \frac{3}{2}\lambda \le 2\lambda$ and $\|\nabla\mathcal{L}(\theta^*)\|_\infty - \frac{1}{2}\lambda \le 0$. This indicates

$$\|\theta - \theta^*\|_2 \le \frac{4\sqrt{s^*}\lambda}{\rho^-}. \tag{J.17}$$

Besides, we have

$$\|\theta - \theta^*\|_1 \overset{(i)}{\le} 6\|(\theta - \theta^*)_{\mathcal{S}^*}\|_1 \le 6\sqrt{s^*}\|(\theta - \theta^*)_{\mathcal{S}^*}\|_2 \le \frac{24\lambda s^*}{\rho^-}, \tag{J.18}$$

where (i) is from $\|\nabla\mathcal{L}(\theta^*)\|_\infty + \frac{3}{2}\lambda \le 2\lambda$ and (J.16).

Combining (J.15) and (J.17), we have desired result (B.10). Combining the assumption in Case 1 and (J.18), we have desired result (B.11).

J.3 Proof of Lemma B.4

Recall that the proximal-gradient update can be computed by the soft-thresholding operation,

$$(\mathcal{T}_{L,\lambda}(\theta))_i = \operatorname{sign}(\check{\theta}_i)\max\left\{|\check{\theta}_i| - \lambda/L, 0\right\} \ \forall i = 1,\dots,d,$$

where $\check{\theta} = \theta - \nabla\mathcal{L}(\theta)/L$. To bound $\|(\mathcal{T}_{L,\lambda}(\theta))_{\overline{\mathcal{S}}^*}\|_0$, we consider

$$\check{\theta} = \theta - \frac{1}{L}\nabla\mathcal{L}(\theta) = \theta - \frac{1}{L}\nabla\mathcal{L}(\theta^*) + \frac{1}{L}(\nabla\mathcal{L}(\theta^*) - \nabla\mathcal{L}(\theta)).$$

We then consider the following three events:

$$A_1 = \left\{i \in \overline{\mathcal{S}}^* : |\theta_i| \ge \lambda/(3L)\right\}, \tag{J.19}$$

$$A_2 = \left\{i \in \overline{\mathcal{S}}^* : |(\nabla\mathcal{L}(\theta^*)/L)_i| > \lambda/(6L)\right\}, \tag{J.20}$$

$$A_3 = \left\{i \in \overline{\mathcal{S}}^* : \left|(\nabla\mathcal{L}(\theta^*)/L - \nabla\mathcal{L}(\theta)/L)_i\right| \ge \lambda/(2L)\right\}, \tag{J.21}$$

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

$$|A_1| \le \sum_{i\in\overline{\mathcal{S}}^*} \frac{3L}{\lambda}|\theta_i - \theta_i^*| \cdot \mathbb{1}(|\theta_i - \theta_i^*| \ge \lambda/(3L)) \le \frac{3L}{\lambda}\sum_{i\in\overline{\mathcal{S}}^*}|\theta_i - \theta_i^*| \le \frac{3L}{\lambda}\|\theta - \theta^*\|_1 \overset{(i)}{\le} \frac{72Ls^*}{\rho^-_{s^*+2\widetilde{s}}}, \tag{J.22}$$

where (i) is from (B.11) in Lemma B.3.

Event A_2. By Lemma 3.2, we have

$$0 \le |A_2| \le \sum_{i\in\overline{\mathcal{S}}^*} \frac{6L}{\lambda}|(\nabla\mathcal{L}(\theta^*)/L)_i| \cdot \mathbb{1}(|(\nabla\mathcal{L}(\theta^*)/L)_i| > \lambda/(6L)) = \sum_{i\in\overline{\mathcal{S}}^*} \frac{6L}{\lambda}|(\nabla\mathcal{L}(\theta^*)/L)_i| \cdot 0 = 0, \tag{J.23}$$

which indicates that $|A_2| = 0$.

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

$$v^\top(\nabla\mathcal{L}(\theta^*) - \nabla\mathcal{L}(\theta)) = \sum_{i\in\widetilde{A}} v_i(\nabla\mathcal{L}(\theta^*)/L - \nabla\mathcal{L}(\theta)/L)_i = \sum_{i\in\widetilde{A}}\left|(\nabla\mathcal{L}(\theta^*) - \nabla\mathcal{L}(\theta))_i\right| \geq \lambda|\widetilde{A}|/2. \tag{J.24}$$

On the other hand, we have

$$\begin{aligned} &v^\top(\nabla\mathcal{L}(\theta^*) - \nabla\mathcal{L}(\theta)) \leq \|v\|_2\|\nabla\mathcal{L}(\theta^*) - \nabla\mathcal{L}(\theta)\|_2 \\ &\overset{(i)}{\leq} \sqrt{|\widetilde{A}|}\cdot\|\nabla\mathcal{L}(\theta^*) - \nabla\mathcal{L}(\theta)\|_2 \overset{(ii)}{\leq} \rho^+_{s^*+2\widetilde{s}}\sqrt{|\widetilde{A}|}\cdot\|\theta - \theta^*\|_2, \end{aligned} \tag{J.25}$$

where (i) is from $\|v\|_2 \leq \sqrt{|\widetilde{A}|}\max\{i : |A_i|\} \leq \sqrt{|\widetilde{A}|}$, and (ii) is from (B.1) and (B.2).

Combining (J.24) and (J.25), we have

$$\lambda|\widetilde{A}| \leq 2\rho^+_{s^*+2\widetilde{s}}\sqrt{|\widetilde{A}|}\cdot\|\theta - \theta^*\|_2 \overset{(i)}{\leq} 8\lambda\kappa_{s^*+2\widetilde{s}}\sqrt{3s^*|\widetilde{A}|}$$

where (i) is from (B.10) in Lemma B.3 and definition of $\kappa_{s^*+2\widetilde{s}} = \frac{\rho^+_{s^*+2\widetilde{s}}}{\rho^-_{s^*+2\widetilde{s}}}$. Considering $A_3 \subseteq \widetilde{A}$, this implies

$$|A_3| \leq |\widetilde{A}| \leq 196\kappa^2_{s^*+2\widetilde{s}}s^*. \tag{J.26}$$

Now combining Even A_1, A_2, A_3 and $L \leq 2\rho^+_{s^*+2\widetilde{s}}$ in assumption, we close the proof as

$$\|(\mathcal{T}_{L,\lambda}(\theta))_{\overline{\mathcal{S}}^*}\|_0 \leq |A_1| + |A_2| + |A_3| \leq \frac{72Ls^*}{\rho^-_{s^*+2\widetilde{s}}} + 196\kappa^2_{s^*+2\widetilde{s}}s^* \leq (144\kappa_{s^*+2\widetilde{s}} + 196\kappa^2_{s^*+2\widetilde{s}})s^* \leq \widetilde{s}.$$

J.4 Proof of Lemma B.5

Let $g = \operatorname{argmin}_{g\in\partial\|\theta\|_1}\mathcal{L} + \lambda\|\theta\|_1$, then $\omega_\lambda = \|\nabla\mathcal{L} + \lambda g\|_\infty$. By the optimality of $\overline{\theta}$ and convexity of $\mathcal{F}_\lambda$, we have

$$\mathcal{F}_\lambda(\theta) - \mathcal{F}_\lambda(\overline{\theta}) \leq (\nabla\mathcal{L} + \lambda g)^\top(\theta - \overline{\theta}) \leq \|\nabla\mathcal{L} + \lambda g\|_\infty\|\theta - \overline{\theta}\|_1 \leq (\omega_\lambda(\theta))\|\theta - \overline{\theta}\|_1. \tag{J.27}$$

Besides, we have

$$\begin{aligned} \|\theta - \overline{\theta}\|_1 &\leq \|\theta - \theta^*\|_1 + \|\overline{\theta} - \theta^*\|_1 \overset{(i)}{\leq} 6\left(\|(\theta - \theta^*)_{\mathcal{S}^*}\|_1 + \|(\overline{\theta} - \theta^*)_{\mathcal{S}^*}\|_1\right) \\ &\leq 6\sqrt{s^*}\left(\|(\theta - \theta^*)_{\mathcal{S}^*}\|_2 + \|(\overline{\theta} - \theta^*)_{\mathcal{S}^*}\|_2\right) \overset{(ii)}{\leq} \frac{24\lambda s^*}{\rho^-_{s^*+2\widetilde{s}}}. \end{aligned} \tag{J.28}$$

where (i) and (ii) are from (B.6) and (B.7) in Lemma B.2 respectively. Combining (J.27) and (J.28), we have desired result.

J.5 Proof of Lemma B.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 $\theta^{(t+1)} = \mathcal{T}_{L,\lambda}(\theta^{(t)})$. Since $\mathcal{F}_\lambda(\theta)$ is convex in θ (but not strongly convex), the sub-level set $\{\theta : \mathcal{F}_\lambda(\theta) \le \mathcal{F}_\lambda(\theta^{(0)})\}$ is bounded. By the monotone decrease of $\mathcal{F}_\lambda(\theta^{(t)})$ from (B.13) in Lemma B.7, $\{\theta^{(t)}\}_{t=0}^{\infty}$ is also bounded. By BolzanoWeierstrass theorem, it has a convergent subsequence and we will show that $\overline{\theta}$ is the unique accumulation point.

Since $\mathcal{F}_{\lambda(\theta)}$ is bounded below,

$$\lim_{k\to\infty} \|\theta^{(t+1)} - \theta^{(t)}\|_2 \le \frac{2}{L^{(t)}} \cdot \lim_{k\to\infty} \left[\mathcal{F}_\lambda\left(\theta^{(t+1)}\right) - \mathcal{F}_\lambda\left(\theta^{(t)}\right)\right] = 0.$$

By Lemma B.8, 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 $\overline{\theta} = \operatorname{argmin}_\theta \mathcal{F}_\lambda(\theta)$, then there exists some $g \in \partial\|\overline{\theta}\|_1$ such that

$$\nabla\mathcal{F}_\lambda(\overline{\theta}) = \mathcal{L}_\lambda(\overline{\theta}) + \lambda g = 0. \tag{J.29}$$

By Lemma B.4, every proximal update is sparse, hence $\|\overline{\theta}_{\overline{\mathcal{S}}^*}\|_0 \le \widetilde{s}$. By RSC property in (3.1), if $\|\theta_{\overline{\mathcal{S}}^*}\|_0 \le \widetilde{s}$, i.e.,$\|(\theta - \overline{\theta})_{\overline{\mathcal{S}}^*}\|_0 \le \widetilde{s}$, then we have

$$\mathcal{L}(\theta) - \mathcal{L}(\overline{\theta}) \ge (\theta - \overline{\theta})^\top \nabla\mathcal{L}(\overline{\theta}) + \frac{\bar{\rho}^-_{s^*+2\widetilde{s}}}{2}\|\theta - \overline{\theta}\|_2^2, \tag{J.30}$$

From the convexity of $\|\theta\|_1$ and $g \in \partial\|\overline{\theta}\|_1$, we have

$$\|\theta\|_1 - \|\overline{\theta}\|_1 \ge (\theta - \overline{\theta})^\top g. \tag{J.31}$$

Combining (J.30) and (J.31), we have for any $\|\theta_{\overline{\mathcal{S}}^*}\|_0 \le \widetilde{s}$,

$$\begin{aligned}\mathcal{F}_\lambda(\theta) - \mathcal{F}_\lambda(\overline{\theta}) &= \mathcal{L}(\theta) + \lambda\|\theta\|_1 - \left(\mathcal{L}(\overline{\theta}) - \lambda\|\overline{\theta}\|_1\right) \ge (\theta - \overline{\theta})^\top\left(\mathcal{L}(\overline{\theta}) + \lambda g\right) + \frac{\bar{\rho}^-_{s^*+2\widetilde{s}}}{2}\|\theta - \overline{\theta}\|_2^2 \\ &\overset{(i)}{=} \frac{\bar{\rho}^-_{s^*+2\widetilde{s}}}{2}\|\theta - \overline{\theta}\|_2^2 \ge 0,\end{aligned} \tag{J.32}$$

where (i) is from (J.29). Therefore, $\overline{\theta}$ is the unique accumulation point, i.e. $\lim_{k\to\infty} \theta^{(t)} = \overline{\theta}$.

Step 2. The objective $\mathcal{F}_\lambda(\theta^{(t+1)})$ satisfies

$$\mathcal{F}_\lambda(\theta^{(t+1)}) \overset{(i)}{\le} \mathcal{Q}_\lambda\left(\theta^{(t+1)}, \theta^{(t)}\right) \overset{(ii)}{=} \min_\theta \mathcal{L}(\theta^{(t)}) + \nabla\mathcal{L}(\theta^{(t)})^\top(\theta - \theta^{(t)}) + \frac{\widetilde{L}_\lambda^{(t)}}{2}\|\theta - \theta^{(t)}\|_2^2 + \lambda\|\theta\|_1. \tag{J.33}$$

where (i) is from (B.13) in Lemma B.7, (ii) is from the definition of $\mathcal{O}_\lambda$ in (2.1). To further bound R.H.S. of (J.33), 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 $S(\overline{\theta}, \theta^{(t)})$,

$$\mathcal{F}_\lambda(\theta^{(t+1)}) - \mathcal{L}(\theta^{(t)}) \le \min_{\theta \in S(\overline{\theta}, \theta^{(t)})} \nabla \mathcal{L}(\theta^{(t)})^\top (\theta - \theta^{(t)}) + \frac{\widetilde{L}_\lambda^{(t)}}{2} \|\theta - \theta^{(t)}\|_2^2 + \lambda \|\theta\|_1. \tag{J.34}$$

Since $\|\overline{\theta}_{\overline{\mathcal{S}}^*}\|_0 \le \widetilde{s}$ and $\|\theta^{(t)}_{\overline{\mathcal{S}}^*}\|_0 \le \widetilde{s}$, then for any $\theta \in S(\overline{\theta}, \theta^{(t)})$, we have $\|\theta_{\overline{\mathcal{S}}^*}\|_0 \le \widetilde{s}$ and $\|(\theta - \theta^{(t)})_{\overline{\mathcal{S}}^*}\|_0 \le 2\widetilde{s}$. By RSC property, we have

$$\mathcal{L}(\theta) \ge \mathcal{L}(\theta^{(t)}) + \nabla \mathcal{L}(\theta^{(t)})^\top (\theta - \theta^{(t)}) + \frac{\rho^-_{s^*+2\widetilde{s}}}{2} \|\theta - \theta^{(t)}\|_2^2 \ge \mathcal{L}(\theta^{(t)}) + \nabla \mathcal{L}(\theta^{(t)})^\top (\theta - \theta^{(t)}). \tag{J.35}$$

Combining (J.34) and (J.35), we have

$$\begin{aligned}
&\mathcal{F}_\lambda(\theta^{(t+1)}) \\
&\le \min_{\theta \in S(\overline{\theta}, \theta^{(t)})} \mathcal{L}(\theta) + \frac{\widetilde{L}_\lambda^{(t)}}{2} \|\theta - \theta^{(t)}\|_2^2 + \lambda \|\theta\|_1 \\
&= \min_{\alpha \in [0,1]} \mathcal{F}_\lambda(\alpha \overline{\theta} + (1-\alpha)\theta^{(t)}) + \frac{\alpha^2 \widetilde{L}_\lambda^{(t)}}{2} \|\overline{\theta} - \theta^{(t)}\|_2^2 \\
&\overset{(i)}{\le} \min_{\alpha \in [0,1]} \alpha \mathcal{F}_\lambda(\overline{\theta}) + (1-\alpha)\mathcal{F}_\lambda(\theta^{(t)}) + \frac{\alpha^2 \widetilde{L}_\lambda^{(t)}}{2} \|\overline{\theta} - \theta^{(t)}\|_2^2 \\
&\overset{(ii)}{\le} \min_{\alpha \in [0,1]} \mathcal{F}_\lambda(\theta^{(t)}) - \alpha \Big(\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta})\Big) + \frac{\alpha^2 \widetilde{L}_\lambda^{(t)}}{\rho^-_{s^*+2\widetilde{s}}} \Big(\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta})\Big) \\
&= \min_{\alpha \in [0,1]} \mathcal{F}_\lambda(\theta^{(t)}) - \alpha \left(1 - \frac{\alpha \widetilde{L}_\lambda^{(t)}}{\rho^-_{s^*+2\widetilde{s}}}\right) \Big(\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta})\Big),
\end{aligned} \tag{J.36}$$

where (i) is from the convexity of $\mathcal{F}_\lambda$ and (ii) is from (J.32).

Minimize the R.H.S. of (J.36) w.r.t. α, the optimal value $\alpha = \frac{\rho^-_{s^*+2\widetilde{s}}}{2\widetilde{L}_\lambda^{(t)}}$ results in

$$\mathcal{F}_\lambda(\theta^{(t+1)}) \le \mathcal{F}_\lambda(\theta^{(t)}) - \frac{\rho^-_{s^*+2\widetilde{s}}}{4\widetilde{L}_\lambda^{(t)}} \Big(\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta})\Big). \tag{J.37}$$

Subtracting both sides of (J.37) by $\mathcal{F}_\lambda(\overline{\theta})$, we have

$$\mathcal{F}_\lambda(\theta^{(t+1)}) - \mathcal{F}_\lambda(\overline{\theta}) \le \left(1 - \frac{\rho^-_{s^*+2\widetilde{s}}}{4\widetilde{L}_\lambda^{(t)}}\right) \Big(\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta})\Big) \overset{(i)}{\le} \left(1 - \frac{\rho^-_{s^*+2\widetilde{s}}}{8\rho^+_{s^*+2\widetilde{s}}}\right) \Big(\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\overline{\theta})\Big), \tag{J.38}$$

where (i) is from Remark B.1. Apply (J.38) recursively, we have the desired result.

K Proof of Intermediate Results for Theorem 3.4

We also introduce an important notion as follows, which is closely related with the SE properties.

Definition K.1. We denote the local ℓ_1 cone as

$$\mathcal{C}(s, vartheta, r) = \Big\{ v, \theta : \mathcal{S} \subseteq \mathcal{M}, |\mathcal{M}| \le s, \|v_{\mathcal{M}_\perp}\|_1 \le vartheta \|v\mathcal{M}\|_1, \\ \|\theta - \theta^*\|_2 \le r \Big\}.$$

Then we define the largest and smallest **localized restricted eigenvalues** (LRE) as

$$\psi^+_{s,vartheta,r} = \sup_{u,\theta} \left\{ \frac{v^\top \nabla^2 \mathcal{L}(\theta) v}{v^\top v} : (v, \theta) \in \mathcal{C}(s, vartheta, r) \right\},$$
$$\psi^-_{s,vartheta,r} = \inf_{u,\theta} \left\{ \frac{v^\top \nabla^2 \mathcal{L}(\theta) v}{v^\top v} : (v, \theta) \in \mathcal{C}(s, vartheta, r) \right\}.$$

The following proposition demonstrates the relationships between SE and LRE. The proof can be found in Bühlmann and Van De Geer (2011), thus is omitted here.

Proposition K.2. Given any $\theta, \theta' \in \mathcal{C}(s, vartheta, r) \cap \mathcal{B}(\theta^*, r)$, we have

$$c_1 \psi^-_{s,vartheta,r} \le \rho^-_s \le c_2 \psi^-_{s,vartheta,r}, \text{ and } c_3 \psi^+_{s,vartheta,r} \le \rho^+_s \le c_4 \psi^+_{s,vartheta,r}.$$

where c_1, c_2, c_3, and c_4 are constants.

K.1 Proof of Lemma D.1

We first demonstrate the sparsity of the update. Since $\theta^{(t+1)}$ is the minimizer to the proximal Newton problem, we have

$$\nabla^2 \mathcal{L}(\theta^{(t)})(\theta^{(t+1)} - \theta^{(t)}) + \nabla \mathcal{L}(\theta^{(t)}) + \lambda \xi^{(t+1)} = 0,$$

where $\xi^{(t+1)} \in \partial \|\theta^{(t+1)}\|_1$.

It follows from Fan et al. (2015) that if conditions in Lemma 3.2 holds, then we have $\min_{j \in \overline{\mathcal{S}}'} \{\lambda_j\} \ge \lambda/2$ for some set $\mathcal{S}' \supset \mathcal{S}$ with $|\mathcal{S}'| \le 2s^*$. Then the analysis of sparsity of can be performed through λ directly.

We then consider the following decomposition

$$\begin{aligned} &\nabla^2 \mathcal{L}(\theta^{(t)})(\theta^{(t+1)} - \theta^{(t)}) + \nabla \mathcal{L}(\theta^{(t)}) \\ &= \underbrace{\nabla^2 \mathcal{L}(\theta^{(t)})(\theta^{(t+1)} - \theta^*)}_{V_1} + \underbrace{\nabla^2 \mathcal{L}(\theta^{(t)})(\theta^* - \theta^{(t)})}_{V_2} + \underbrace{\nabla \mathcal{L}(\theta^{(t)}) - \nabla \mathcal{L}(\theta^*)}_{V_3} + \underbrace{\nabla \mathcal{L}(\theta^*)}_{V_4}. \end{aligned}$$

Consider the following sets: $\mathcal{A}_i = \Big\{ j \in \overline{\mathcal{S}}' \ : \ |(V_i)_j| \ge \lambda/4 \Big\}$, for all $i \in \{1, 2, 3, 4\}$.

Set $\mathcal{A}_2$. Suppose we choose a vector $v \in \mathbb{R}^d$ such that $v_j = \operatorname{sign}\Big\{ (\nabla^2 \mathcal{L}(\theta^{(t)})(\theta^* - \theta^{(t)}))_j \Big\}$ for all $j \in \mathcal{A}_2$ and $v_j = 0$ for $j \notin \mathcal{A}_2$. Then we have

$$v^\top \nabla^2 \mathcal{L}(\theta^{(t)})(\theta^* - \theta^{(t)}) = \sum_{j \in \mathcal{A}_2} v_j (\nabla^2 \mathcal{L}(\theta^{(t)})(\theta^* - \theta^{(t)}))_j = \sum_{j \in \mathcal{A}_2} |(\nabla^2 \mathcal{L}(\theta^{(t)})(\theta^* - \theta^{(t)}))_j| \ge \lambda |\mathcal{A}_2|/4. \quad \text{(K.1)}$$

On the other hand, we have

$$v^\top \nabla^2 \mathcal{L}(\theta^{(t)})(\theta^* - \theta^{(t)}) \le \|v(\nabla^2 \mathcal{L}(\theta^{(t)}))^{1/2}\|_2 \|(\nabla^2 \mathcal{L}(\theta^{(t)}))^{1/2}(\theta^* - \theta^{(t)})\|_2$$
$$\overset{(i)}{\le} \rho^+_{s^*+2\widetilde{s}} \|v\|_2 \|\theta^* - \theta^{(t)}\|_2 \overset{(ii)}{\le} \sqrt{|\mathcal{A}_2|}\rho^+_{s^*+2\widetilde{s}} \|\theta^* - \theta^{(t)}\|_2 \overset{(iii)}{\le} C'\sqrt{|\mathcal{A}_2|}\kappa_{s^*+2\widetilde{s}}\lambda\sqrt{s^*}, \tag{K.2}$$

where (i) is from the SE properties, (ii) is from the definition of v, and (iii) is from $\|\theta^{(t)} - \theta^*\|_2 \le C'\lambda\sqrt{s^*}/\rho^-_{s^*+2\widetilde{s}}$. Combining (K.1) and (K.2), we have $|\mathcal{A}_2| \le C_2\kappa^2_{s^*+2\widetilde{s}}s^*$.

Set $\mathcal{A}_3$. Consider the event $\widetilde{A} = \left\{i : \left|\left(\nabla\mathcal{L}(\theta^{(t)}) - \nabla\mathcal{L}(\theta^*)\right)_i\right| \ge \lambda/4\right\}$, which satisfies $\mathcal{A}_3 \subseteq \widetilde{A}$. We will provide an upper bound of $|\widetilde{A}|$, which is also an upper bound of $|\mathcal{A}_3|$. Let $v \in \mathbb{R}^d$ be chosen such that $v_i = \operatorname{sign}\left\{\left(\nabla\mathcal{L}(\theta^{(t)}) - \nabla\mathcal{L}(\theta^*)\right)_i\right\}$ for any $i \in \widetilde{A}$, and $v_i = 0$ for any $i \notin \widetilde{A}$. Then we have

$$v^\top\left(\nabla\mathcal{L}(\theta^{(t)}) - \nabla\mathcal{L}(\theta^*)\right) = \sum_{i\in\widetilde{A}} v_i\left(\nabla\mathcal{L}(\theta^{(t)}) - \nabla\mathcal{L}(\theta^*)\right)_i = \sum_{i\in\widetilde{A}}\left|\left(\nabla\mathcal{L}(\theta^{(t)}) - \nabla\mathcal{L}(\theta^*)\right)_i\right| \ge \lambda|\widetilde{A}|/4. \tag{K.3}$$

On the other hand, we have

$$v^\top\left(\nabla\mathcal{L}(\theta^{(t)}) - \nabla\mathcal{L}(\theta^*)\right) \le \|v\|_2\|\nabla\mathcal{L}(\theta^{(t)}) - \nabla\mathcal{L}(\theta^*)\|_2$$
$$\overset{(i)}{\le} \sqrt{|\widetilde{A}|}\cdot\|\nabla\mathcal{L}(\theta^{(t)}) - \nabla\mathcal{L}(\theta^*)\|_2 \overset{(ii)}{\le} \rho^+_{s^*+2\widetilde{s}}\sqrt{|\widetilde{A}|}\cdot\|\theta^{(t)} - \theta^*\|_2, \tag{K.4}$$

where (i) is from $\|v\|_2 \le \sqrt{|\widetilde{A}|}\max\{i : |\mathcal{A}_i|\} \le \sqrt{|\widetilde{A}|}$, and (ii) is from the mean value theorem and the SE properties.

Combining (K.3) and (K.4), we have

$$\lambda|\widetilde{A}| \le 4\rho^+_{s^*+2\widetilde{s}}\sqrt{|\widetilde{A}|}\cdot\|\theta - \theta^*\|_2 \overset{(i)}{\le} 8\lambda\kappa_{s^*+2\widetilde{s}}\sqrt{3s^*|\widetilde{A}|}$$

where (i) is from $\|\theta^{(t)} - \theta^*\|_2 \le C'\lambda\sqrt{s^*}/\rho^-_{s^*+2\widetilde{s}}$ and definition of $\kappa_{s^*+2\widetilde{s}} = \rho^+_{s^*+2\widetilde{s}}/\rho^-_{s^*+2\widetilde{s}}$. Considering $\mathcal{A}_3 \subseteq \widetilde{A}$, this implies $|\mathcal{A}_3| \le |\widetilde{A}| \le C_3\kappa^2_{s^*+2\widetilde{s}}s^*$.

Set $\mathcal{A}_4$. By conditions in Lemma 3.2 and $\lambda \ge 4\|\nabla\mathcal{L}(\theta^*)\|_\infty$, we have

$$0 \le |V_4| \le \sum_{i\in\overline{\mathcal{S}}^*}\frac{4}{\lambda}|(\nabla\mathcal{L}(\theta^*))_i|\cdot\mathbb{1}(|(\nabla\mathcal{L}(\theta^*))_i| > \lambda/(4)) = \sum_{i\in\overline{\mathcal{S}}^*}\frac{4}{\lambda}|(\nabla\mathcal{L}(\theta^*))_i|\cdot 0 = 0,$$

Set $\mathcal{A}_1$. From Lemma K.3, we have $\mathcal{F}_\lambda(\theta^{(t+1)}) \le \mathcal{F}_\lambda(\theta^*) + \frac{\lambda}{4}\|\theta^{(t+1)} - \theta^*\|_1$. This implies

$$\mathcal{L}(\theta^{(t+1)}) - \mathcal{L}(\theta^*) \le \lambda(\|\theta^*\|_1 - \|\theta^{(t+1)}\|_1) + \frac{\lambda}{4}\|\theta^{(t+1)} - \theta^*\|_1$$
$$= \lambda(\|\theta^*_{\mathcal{S}'}\|_1 - \|\theta^{(t+1)}_{\mathcal{S}'}\|_1 - \|\theta^{(t+1)}_{\mathcal{S}'_\perp}\|_1) + \frac{\lambda}{4}\|\theta^{(t+1)} - \theta^*\|_1$$
$$\le \frac{5\lambda}{4}\|\theta^{(t+1)}_{\mathcal{S}'} - \theta^*_{\mathcal{S}'}\|_1 - \frac{3\lambda}{4}\|\theta^{(t+1)}_{\mathcal{S}'_\perp} - \theta^*_{\mathcal{S}'_\perp}\|_1. \tag{K.5}$$

where the equality holds since $\theta^*_{\mathcal{S}'_\perp} = 0$. On the other hand, we have

$$\mathcal{L}(\theta^{(t+1)}) - \mathcal{L}(\theta^*) \overset{(i)}{\geq} \nabla\mathcal{L}(\theta^*)(\theta^{(t+1)} - \theta^*) \geq -\|cL(\theta^*)\|_\infty \|\theta^{(t+1)} - \theta^*\|_1 \overset{(ii)}{\geq} -\frac{\lambda}{4}\|\theta^{(t+1)} - \theta^*\|_1$$
$$= -\frac{\lambda}{4}\|\theta^{(t+1)}_{\mathcal{S}'} - \theta^*_{\mathcal{S}'}\|_1 - \frac{\lambda}{4}\|\theta^{(t+1)}_{\mathcal{S}'_\perp} - \theta^*_{\mathcal{S}'_\perp}\|_1, \tag{K.6}$$

where (i) is from the convexity of $\mathcal{L}$ and (ii) is from conditions of Lemma 3.2. Combining (K.5) and (K.6), we have

$$\|\theta^{(t+1)}_{\mathcal{S}'_\perp} - \theta^*_{\mathcal{S}'_\perp}\|_1 \leq 3\|\theta^{(t+1)}_{\mathcal{S}'} - \theta^*_{\mathcal{S}'}\|_1,$$

which implies that $(\theta^{(t+1)} - \theta^*, \theta^{(t+1)}) \in \mathcal{C}(s^*, 3, r)$ with respect to the set $\mathcal{S}'$. Then we choose a vector $v \in \mathbb{R}^d$ such that $v_j = \operatorname{sign}\left\{(\nabla^2\mathcal{L}(\theta^{(t)})(\theta^{(t+1)} - \theta^*))_j\right\}$ for all $j \in \mathcal{A}_1$ and $v_j = 0$ for $j \notin \mathcal{A}_1$. Then we have

$$v^\top\nabla^2\mathcal{L}(\theta^{(t)})(\theta^{(t+1)} - \theta^*) = \sum_{j\in\mathcal{A}_2} v_j(\nabla^2\mathcal{L}(\theta^{(t)})(\theta^{(t+1)} - \theta^*))_j$$
$$= \sum_{j\in\mathcal{A}_2} |(\nabla^2\mathcal{L}(\theta^{(t)})(\theta^{(t+1)} - \theta^*))_j| \geq \lambda|\mathcal{A}_1|/4. \tag{K.7}$$

On the other hand, we have

$$v^\top\nabla^2\mathcal{L}(\theta^{(t)})(\theta^{(t+1)} - \theta^*) \leq \|v(\nabla^2\mathcal{L}(\theta^{(t)}))^{1/2}\|_2 \|(\nabla^2\mathcal{L}(\theta^{(t)}))^{1/2}(\theta^{(t+1)} - \theta^*)\|_2$$
$$\overset{(i)}{\leq} c_1\rho^+_{s^*+2\widetilde{s}}\|v\|_2\|\theta^{(t+1)} - \theta^*\|_2 \overset{(ii)}{\leq} c_1\sqrt{|\mathcal{A}_2|}\rho^+_{s^*+2\widetilde{s}}\|\theta^{(t+1)} - \theta^*\|_2$$
$$\overset{(iii)}{\leq} c_2\sqrt{|\mathcal{A}_1|}\kappa_{s^*+2\widetilde{s}}\lambda\sqrt{s^*}, \tag{K.8}$$

where (i) is from the SE properties and Proposition K.2, (ii) is from the definition of v, and (iii) is from $\|\theta^{(t+1)} - \theta^*\|_2 \leq C'\lambda\sqrt{s^*}/\rho^-_{s^*+2\widetilde{s}}$. Combining (K.7) and (K.8), we have $|\mathcal{A}_1| \leq C_1\kappa^2_{s^*+2\widetilde{s}}s^*$.

Combining the results for Set $\mathcal{A}_1 \sim \mathcal{A}_4$, we have that there exists some constant C_0 such that

$$\|\theta^{(t+1)}_{\widetilde{\mathcal{S}}}\|_0 \leq C_0\kappa^2_{s^*+2\widetilde{s}}s^* \leq \widetilde{s}.$$

This finishes the first part. The estimation error follows directly from Lemma K.4.

K.2 Proof of Lemma D.2

For notational simplicity, we introduce the following proximal operator,

$$\operatorname{prox}^{H,g}_r(\theta) = \operatorname{argmin}_{\theta'} r(\theta') + g^\top(\theta' - \theta) + \frac{1}{2}\|\theta' - \theta\|^2_H.$$

Then we have

$$\theta^{(t+1)} = \operatorname{prox}^{\nabla^2\mathcal{L}(\theta^{(t)}),\nabla\mathcal{L}(\theta^{(t)})}_{\mathcal{R}^{\ell_1}_\lambda(\theta^{(t)})}\left(\theta^{(t)}\right).$$

By Lemma D.1, we have

$$\|\theta^{(t+1)}_{\overline{\mathcal{S}}}\|_0 \le \widetilde{s}.$$

By the KKT condition of function $\min \mathcal{F}_\lambda$, i.e., $-\nabla\mathcal{L}(\overline{\theta}) \in \partial\mathcal{R}^{\ell_1}_\lambda(\overline{\theta})$, we also have

$$\overline{\theta} = \mathrm{prox}^{\nabla^2\mathcal{L}(\theta^{(t)}),\nabla\mathcal{L}(\overline{\theta})}_{\mathcal{R}^{\ell_1}_\lambda(\overline{\theta})}\left(\overline{\theta}\right).$$

By monotonicity of sub-gradient of a convex function, we have the *strictly non-expansive* property: for any $\theta, \theta' \in \mathbb{R}$, let $u = \mathrm{prox}^{H,g}_r(\theta)$ and $v = \mathrm{prox}^{H,g'}_r(\theta')$, then

$$(u-v)^\top H(\theta-\theta') - (u-v)^\top(g-g') \ge \|u-v\|^2_H.$$

Thus by the strictly non-expansive property of the proximal operator, we obtain

$$\begin{aligned}
\|\theta^{(t+1)} - \overline{\theta}\|^2_{\nabla^2\mathcal{L}(\overline{\theta})} &\le \left(\theta^{(t+1)} - \overline{\theta}\right)^\top \left[\nabla^2\mathcal{L}(\theta^{(t)})\left(\theta^{(t)} - \overline{\theta}\right) + \left(\nabla\mathcal{L}(\overline{\theta}) - \nabla\mathcal{L}(\theta^{(t)})\right)\right] \\
&\le \|\theta^{(t+1)} - \overline{\theta}\|_2 \left\|\nabla^2\mathcal{L}(\theta^{(t)})\left(\theta^{(t)} - \overline{\theta}\right) + \left(\nabla\mathcal{L}(\overline{\theta}) - \nabla\mathcal{L}(\theta^{(t)})\right)\right\|_2. \qquad \text{(K.9)}
\end{aligned}$$

Note that both $\|\theta^{(t+1)}\|_0 \le \widetilde{s}$ and $\|\overline{\theta}\|_0 \le \widetilde{s}$. On the other hand, from the SE properties, we have

$$\|\theta^{(t+1)} - \overline{\theta}\|^2_{\nabla^2\mathcal{L}(\overline{\theta})} = (\theta^{(t+1)} - \overline{\theta})^\top \nabla^2\mathcal{L}(\overline{\theta})(\theta^{(t+1)} - \overline{\theta}) \ge \rho^-_{s^*+2\widetilde{s}}\|\theta^{(t+1)} - \overline{\theta}\|_2^2. \qquad \text{(K.10)}$$

Combining (K.9) and (K.10), we have

$$\begin{aligned}
\rho^-_{s^*+2\widetilde{s}}\left\|\theta^{(t+1)} - \overline{\theta}\right\|_2 &\le \left\|\nabla^2\mathcal{L}(\theta^{(t)})\left(\theta^{(t)} - \overline{\theta}\right) + \left(\nabla\mathcal{L}(\overline{\theta}) - \nabla\mathcal{L}(\theta^{(t)})\right)\right\|_2 \\
&= \left\|\int_0^1 \left[\nabla^2\mathcal{L}\left(\theta^{(t)} + \tau\left(\overline{\theta} - \theta^{(t)}\right)\right) - \nabla^2\mathcal{L}\left(\theta^{(t)}\right)\right]\cdot\left(\overline{\theta} - \theta^{(t)}\right)d\tau\right\|_2 \\
&\le \int_0^1 \left\|\left[\nabla^2\mathcal{L}\left(\theta^{(t)} + \tau\left(\overline{\theta} - \theta^{(t)}\right)\right) - \nabla^2\mathcal{L}\left(\theta^{(t)}\right)\right]\cdot\left(\overline{\theta} - \theta^{(t)}\right)\right\|_2 d\tau \\
&\le L_{s^*+2\widetilde{s}}\left\|\theta^{(t)} - \overline{\theta}\right\|_2^2,
\end{aligned}$$

where the last inequality is from the local restricted Hessian smoothness of $\mathcal{L}$. Then we finish the proof by the definition of r.

K.3 Proof of Lemma D.3

Suppose the step size $\eta_t < 1$. Note that we do not need the step size to be $\eta_t = 1$ in Lemma D.1 and Lemma D.2. We denote $\Delta\theta^{(t)} = \theta^{(t+1/2)} - \theta^{(t)}$. Then we have

$$\left\|\Delta\theta^{(t)}\right\|_2 \overset{(i)}{\le} \left\|\theta^{(t)} - \overline{\theta}\right\|_2 + \left\|\theta^{(t+1/2)} - \overline{\theta}\right\|_2 \overset{(ii)}{\le} \left\|\theta^{(t)} - \overline{\theta}\right\|_2 + \frac{L_{s^*+2\widetilde{s}}}{2\rho^-_{s^*+2\widetilde{s}}}\left\|\theta^{(t)} - \overline{\theta}\right\|_2^2 \overset{(iii)}{\le} \frac{3}{2}\left\|\theta^{(t)} - \overline{\theta}\right\|_2, \quad \text{(K.11)}$$

where (i) is from triangle inequality, (ii) is from Lemma D.2, and (iii) is from $\left\|\theta^{(t)} - \overline{\theta}\right\|_2 \le r \le \frac{\rho^-_{s^*+2\widetilde{s}}}{L_{s^*+2\widetilde{s}}}$.

By Lemma D.1, we have $\left\|\Delta\theta^{(t)}_{\overline{\mathcal{S}}}\right\|_0 \le 2\widetilde{s}$. To show $\eta_t = 1$, it is now suffice to demonstrate that

$$\mathcal{F}_\lambda(\theta^{(t+1/2)}) - \mathcal{F}_\lambda(\theta^{(t)}) \le \frac{1}{4}\gamma_t.$$

By expanding $\mathcal{F}_\lambda$, we have

$$\begin{aligned}
&\mathcal{F}_\lambda(\theta^{(t)} + \Delta\theta^{(t)}) - \mathcal{F}_\lambda(\theta^{(t)}) = \mathcal{L}(\theta^{(t)} + \Delta\theta^{(t)}) - \mathcal{L}(\theta^{(t)}) + \mathcal{R}^{\ell_1}_\lambda(\theta^{(t)} + \Delta\theta^{(t)}) - \mathcal{R}^{\ell_1}_\lambda(\theta^{(t)}) \\
&\overset{(i)}{\le} \nabla\mathcal{L}(\theta^{(t)})^\top\Delta\theta^{(t)} + \frac{1}{2}\Delta(\theta^{(t)})^\top\nabla^2\mathcal{L}(\theta)\Delta\theta^{(t)} + \frac{L_{s^*+2\widetilde{s}}}{6}\left\|\Delta\theta^{(t)}\right\|_2^3 + \mathcal{R}^{\ell_1}_\lambda(\theta^{(t)} + \Delta\theta^{(t)}) - \mathcal{R}^{\ell_1}_\lambda(\theta^{(t)}) \\
&\overset{(ii)}{\le} \gamma_t - \frac{1}{2}\gamma_t + \frac{L_{s^*+2\widetilde{s}}}{6}\left\|\Delta\theta^{(t)}\right\|_2^3 \overset{(iii)}{\le} \frac{1}{2}\gamma_t + \frac{L_{s^*+2\widetilde{s}}}{6\rho^-_{s^*+2\widetilde{s}}}\left\|\Delta\theta^{(t)}\right\|_{\nabla^2\mathcal{L}(\theta)}\left\|\Delta\theta^{(t)}\right\|_2 \overset{(iv)}{\le} \left(\frac{1}{2} - \frac{L_{s^*+2\widetilde{s}}}{6\rho^-_{s^*+2\widetilde{s}}}\left\|\Delta\theta^{(t)}\right\|_2\right)\gamma_t \\
&\overset{(v)}{\le} \frac{1}{4}\gamma_t,
\end{aligned}$$

where (i) is from the restricted Hessian smooth condition, (ii) and (iv) are from Lemma D.4, (iii) is from the same argument of (K.10), and (v) is from (K.11), $\gamma_t < 0$, and $\left\|\theta^{(t)} - \overline{\theta}\right\|_2 \le r \le \frac{\rho^-_{s^*+2\widetilde{s}}}{L_{s^*+2\widetilde{s}}}$. This implies $\theta^{(t+1)} = \theta^{(t+1/2)}$.

K.4 Proof of Lemma D.4

We denote $H = \nabla^2\mathcal{L}(\theta^{(t)})$. Since $\Delta\theta^{(t)}$ is the solution for

$$\min_{\Delta\theta^{(t)}} \nabla\mathcal{L}\left(\theta^{(t)}\right)^\top \cdot \Delta\theta^{(t)} + \frac{1}{2}\left\|\Delta\theta^{(t)}\right\|_H^2 + \mathcal{R}^{\ell_1}_\lambda\left(\theta^{(t)} + \Delta\theta^{(t)}\right)$$

then for any $\eta_t \in (0, 1]$, we have

$$\begin{aligned}
&\eta_t\nabla\mathcal{L}\left(\theta^{(t)}\right)^\top \cdot \Delta\theta^{(t)} + \frac{\eta_t^2}{2}\left\|\Delta\theta^{(t)}\right\|_H^2 + \mathcal{R}^{\ell_1}_\lambda\left(\theta^{(t)} + \eta_t\Delta\theta^{(t)}\right) \\
&\ge \nabla\mathcal{L}\left(\theta^{(t)}\right)^\top \cdot \Delta\theta^{(t)} + \frac{1}{2}\left\|\Delta\theta^{(t)}\right\|_H^2 + \mathcal{R}^{\ell_1}_\lambda\left(\theta^{(t)} + \Delta\theta^{(t)}\right)
\end{aligned}$$

By the convexity of $\mathcal{R}^{\ell_1}_\lambda$, we have

$$\begin{aligned}
&\eta_t\nabla\mathcal{L}\left(\theta^{(t)}\right)^\top \cdot \Delta\theta^{(t)} + \frac{\eta_t^2}{2}\left\|\Delta\theta^{(t)}\right\|_H^2 + \eta_t\mathcal{R}^{\ell_1}_\lambda\left(\theta^{(t)} + \Delta\theta^{(t)}\right) + (1-\eta_t)\mathcal{R}^{\ell_1}_\lambda(\theta^{(t)}) \\
&\ge \nabla\mathcal{L}\left(\theta^{(t)}\right)^\top \cdot \Delta\theta^{(t)} + \frac{1}{2}\left\|\Delta\theta^{(t)}\right\|_H^2 + \mathcal{R}^{\ell_1}_\lambda\left(\theta^{(t)} + \Delta\theta^{(t)}\right).
\end{aligned}$$

Rearranging the terms, we obtain

$$(1-\eta_t)\left(\nabla\mathcal{L}\left(\theta^{(t)}\right)^\top \cdot \Delta\theta^{(t)} + \mathcal{R}^{\ell_1}_\lambda\left(\theta^{(t)} - \Delta\theta^{(t)}\right) - \mathcal{R}^{\ell_1}_\lambda(\theta^{(t)})\right) \le -\frac{1-\eta_t^2}{2}\left\|\Delta\theta^{(t)}\right\|_H^2$$

Canceling the $(1-\eta_t)$ factor from both sides and let $\eta_t \to 1$, we obtain the desired inequality,

$$\gamma_t \le -\left\|\Delta\theta^{(t)}\right\|_H^2.$$

K.5 Proof of Lemma D.5

We first demonstrate an upper bound of the approximate KKT parameter ω_λ. Given the solution $\theta^{(t-1)}$ from the $(t-1)$-th iteration, the optimal solution at t-th iteration satisfies the KKT condition:

$$\nabla^2\mathcal{L}(\theta^{(t-1)})(\theta^{(t)}-\theta^{(t-1)})+\nabla\mathcal{L}(\theta^{(t-1)})+\lambda\xi^{(t)}=0,$$

where $\xi^{(t)}\in\partial\|\theta^{(t)}\|_1$. Then for any vector v with $\|v\|_2\le\|v\|_1=1$ and $\|v\|_0\le s^*+2\widetilde{s}$, by taking $\Delta\theta^{(t-1)}=\theta^{(t)}-\theta^{(t-1)}$, we have

$$\begin{aligned}
&(\nabla\mathcal{L}(\theta^{(t)})+\lambda\xi^{(t)})^\top v\\
&=(\nabla\mathcal{L}(\theta^{(t)})-\nabla^2\mathcal{L}(\theta^{(t-1)})\Delta\theta^{(t-1)}-\nabla\mathcal{L}(\theta^{(t-1)}))^\top v\\
&=(\nabla\mathcal{L}(\theta^{(t)})-\nabla\mathcal{L}(\theta^{(t-1)}))^\top v-(\nabla^2\mathcal{L}(\theta^{(t-1)})\Delta\theta^{(t-1)})^\top v\\
&\overset{(i)}{\le}\left\|(\nabla^2\mathcal{L}(\widetilde{\theta}))^{1/2}\Delta\theta^{(t-1)}\right\|_2\left\|v^\top(\nabla^2\mathcal{L}(\widetilde{\theta}))^{1/2}\right\|_2+\left\|(\nabla^2\mathcal{L}(\theta^{(t-1)}))^{1/2}\Delta\theta^{(t-1)}\right\|_2\left\|v^\top(\nabla^2\mathcal{L}(\theta^{(t-1)}))^{1/2}\right\|_2\\
&\overset{(ii)}{\le}2\rho^+_{s^*+2\widetilde{s}}\left\|\Delta\theta^{(t-1)}\right\|_2,
\end{aligned}\tag{K.12}$$

where (i) is from mean value theorem with some $\widetilde{\theta}=(1-a)\theta^{(t-1)}+a\theta^{(t)}$ for some $a\in[0,1]$ and Cauchy-Schwarz inequality, and (ii) is from the SE properties. Take the supremum of the L.H.S. of (K.12) with respect to v, we have

$$\left\|\nabla\mathcal{L}(\theta^{(t)})+\lambda\xi^{(t)}\right\|_\infty\le2\rho^+_{s^*+2\widetilde{s}}\left\|\Delta\theta^{(t-1)}\right\|_2.\tag{K.13}$$

Then from Lemma D.2, we have

$$\left\|\theta^{(t+1)}-\overline{\theta}\right\|_2\le\left(\frac{L_{s^*+2\widetilde{s}}}{2\rho^-_{s^*+2\widetilde{s}}}\right)^{1+2+4+\ldots+2^{t-1}}\left\|\theta^{(0)}-\overline{\theta}\right\|_2^{2^\top}\le\left(\frac{L_{s^*+2\widetilde{s}}}{2\rho^-_{s^*+2\widetilde{s}}}\left\|\theta^{(0)}-\overline{\theta}\right\|_2\right)^{2^t}.$$

By (K.13) and (K.11), we obtain

$$\omega_\lambda\left(\theta^{(t)}\right)\le2\rho^+_{s^*+2\widetilde{s}}\left\|\Delta\theta^{(t-1)}\right\|_2\le3\rho^+_{s^*+2\widetilde{s}}\left\|\theta^{(t-1)}-\overline{\theta}\right\|_2\le3\rho^+_{s^*+2\widetilde{s}}\left(\frac{L_{s^*+2\widetilde{s}}}{2\rho^-_{s^*+2\widetilde{s}}}\left\|\theta^{(0)}-\overline{\theta}\right\|_2\right)^{2^t}.$$

By requiring the R.H.S. equal to ε we obtain

$$\begin{aligned}
t&=\log\frac{\log\left(\frac{3\rho^+_{s^*+2\widetilde{s}}}{\varepsilon}\right)}{\log\left(\frac{2\rho^-_{s^*+2\widetilde{s}}}{L_{s^*+2\widetilde{s}}\left\|\theta^{(0)}-\overline{\theta}\right\|_2}\right)}=\log\log\left(\frac{3\rho^+_{s^*+2\widetilde{s}}}{\varepsilon}\right)-\log\log\left(\frac{2\rho^-_{s^*+2\widetilde{s}}}{L_{s^*+2\widetilde{s}}\left\|\theta^{(0)}-\overline{\theta}\right\|_2}\right)\\
&\overset{(i)}{\le}\log\log\left(\frac{3\rho^+_{s^*+2\widetilde{s}}}{\varepsilon}\right)-\log\log4\le\log\log\left(\frac{3\rho^+_{s^*+2\widetilde{s}}}{\varepsilon}\right),
\end{aligned}$$

where (i) is from the fact that $\left\|\theta^{(0)}-\overline{\theta}\right\|_2\le r=\frac{\rho^-_{s^*+2\widetilde{s}}}{2L_{s^*+2\widetilde{s}}}$.

Lemma K.3. Given $\omega_\lambda(\theta^{(t)})\le\frac{\lambda}{4}$, we have

$$\mathcal{F}_\lambda(\theta^{(t)})\le\mathcal{F}_\lambda(\theta^*)+\frac{\lambda}{4}\|\theta^{(t)}-\theta^*\|_1.$$

Proof. For some $\xi^{(t)} = \operatorname{argmin}_{\xi \in \partial \|\theta^{(t)}\|_1} \|\nabla \mathcal{L}(\theta^{(t)}) + \lambda \xi\|_\infty$, we have

$$\begin{aligned}\mathcal{F}_\lambda(\theta^*) &\overset{(i)}{\geq} \mathcal{F}_\lambda(\theta^{(t)}) - (\nabla \mathcal{L}(\theta^{(t)}) + \lambda \xi^{(t)})^\top (\theta^{(t)} - \theta^*) \geq \mathcal{F}_\lambda(\theta^{(t)}) - \|\nabla \mathcal{L}(\theta^{(t)}) + \lambda \xi^{(t)}\|_\infty \|\theta^{(t)} - \theta^*\|_1 \\ &\overset{(ii)}{\geq} \mathcal{F}_\lambda(\theta^{(t)}) - \frac{\lambda}{4}\|\theta^{(t)} - \theta^*\|_1\end{aligned}$$

where (i) is from the convexity of $\mathcal{F}_\lambda$ and (ii) is from the fact that for all $t \geq 0$, $\mathcal{F}_\lambda(\theta^{(t)}) \leq \mathcal{F}_\lambda(\theta^{(t-1)})$ and $\omega_\lambda(\theta^{(t)}) \leq \frac{\lambda}{4}$. This finishes the proof. □

Lemma K.4 (Adapted from Fan et al. (2015))**.** Suppose $\|\theta^{(t)}_{\widetilde{\mathcal{S}}}\|_0 \leq \widetilde{s}$ and $\omega_\lambda(\theta^{(t)}) \leq \frac{\lambda}{4}$. Then there exists a generic constant c_1 such that $\|\theta^{(t)} - \theta^*\|_2 \leq \frac{c_1 \lambda \sqrt{s^*}}{\rho_{s^*+2\widetilde{s}}}$.

L Proofs of Intermediate Lemmas in Appendix F

L.1 Proof of Lemma F.1

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 $\widetilde{\theta} = (1-\alpha)\theta + \alpha\theta^*$ and

$$\|\widetilde{\theta} - \theta^*\|_2 = \sqrt{r}. \tag{L.1}$$

Let $\widetilde{g} = \operatorname{argmin}_{g \in \partial \|\theta\|_1} \|\nabla \mathcal{L}(\theta) + \lambda g\|_\infty$ and $\Delta = \theta - \theta^*$, then we have

$$\begin{aligned}\mathcal{F}_\lambda(\theta^*) &\overset{(i)}{\geq} \mathcal{F}_\lambda(\theta) - (\nabla \mathcal{L}(\theta) + \lambda \widetilde{g})^\top \Delta \geq \mathcal{F}_\lambda(\theta) - \|\nabla \mathcal{L}(\theta) + \lambda \widetilde{g}\|_\infty \|\Delta\|_1 \\ &\overset{(ii)}{\geq} \mathcal{F}_\lambda(\theta) - \frac{\lambda}{4}\|\Delta\|_1,\end{aligned} \tag{L.2}$$

where (i) is from the convexity of $\mathcal{F}_\lambda(\theta)$ and (ii) is from the approximate KKT condition.

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

$$\begin{aligned}\mathcal{F}_\lambda(\widetilde{\theta}) &\overset{(i)}{\leq} (1-\alpha)\mathcal{F}_\lambda(\theta) + \alpha \mathcal{F}_\lambda(\theta^*) \leq (1-\alpha)\mathcal{F}_\lambda(\theta^*) + \frac{(1-\alpha)\lambda}{4}\|\Delta\|_1 + \alpha \mathcal{F}_\lambda(\theta^*) \\ &\leq \mathcal{F}_\lambda(\theta^*) + \frac{\lambda}{4}\|(1-\alpha)(\theta - \theta^*)\|_1 = \mathcal{F}_\lambda(\theta^*) + \frac{\lambda}{4}\|(1-\alpha)\theta + \alpha\theta^* - \theta^*)\|_1 \\ &= \mathcal{F}_\lambda(\theta^*) + \frac{\lambda}{4}\|\widetilde{\theta} - \theta^*\|_1 = \mathcal{F}_\lambda(\theta^*) + \frac{\lambda}{4}\|\widetilde{\Delta}\|_1.\end{aligned}$$

where (i) is from the convexity of $\mathcal{F}_\lambda(\theta)$. This indicates

$$\begin{aligned}\mathcal{L}(\widetilde{\theta}) - \mathcal{L}(\theta^*) &\leq \lambda(\|\theta^*\|_1 - \|\widetilde{\theta}\|_1 + \frac{1}{4}\|\widetilde{\Delta}\|_1) \\ &= \lambda(\|\theta^*_{\mathcal{S}^*}\|_1 - \|\widetilde{\theta}_{\mathcal{S}^*}\|_1 - \|\widetilde{\theta}_{\overline{\mathcal{S}}^*}\|_1 + \frac{1}{4}\|\widetilde{\Delta}_{\mathcal{S}^*}\|_1 + \frac{1}{4}\|\widetilde{\Delta}_{\overline{\mathcal{S}}^*}\|_1) \\ &\leq \lambda(\|\theta^*_{\mathcal{S}^*} - \widetilde{\theta}_{\mathcal{S}^*}\|_1 - \|\widetilde{\theta}_{\overline{\mathcal{S}}^*} - \theta^*_{\overline{\mathcal{S}}^*}\|_1 + \frac{1}{4}\|\widetilde{\Delta}_{\mathcal{S}^*}\|_1 + \frac{1}{4}\|\widetilde{\Delta}_{\overline{\mathcal{S}}^*}\|_1) \\ &= \frac{5\lambda}{4}\|\widetilde{\Delta}_{\mathcal{S}^*}\|_1 - \frac{3\lambda}{4}\|\widetilde{\Delta}_{\overline{\mathcal{S}}^*}\|_1.\end{aligned} \tag{L.3}$$

On the other hand, we have

$$\mathcal{L}(\widetilde{\theta}) - \mathcal{L}(\theta^*) \overset{(i)}{\geq} \nabla\mathcal{L}(\theta^*)^\top \widetilde{\Delta} \geq -\|\nabla\mathcal{L}(\theta^*)\|_\infty \|\widetilde{\Delta}\|_1 \overset{(ii)}{\geq} -\frac{\lambda}{6}\|\widetilde{\Delta}\|_1$$
$$= -\frac{\lambda}{6}\|\widetilde{\Delta}_{\mathcal{S}^*}\|_1 - \frac{\lambda}{6}\|\widetilde{\Delta}_{\overline{\mathcal{S}}^*}\|_1, \tag{L.4}$$

where (i) is from the convexity of $\mathcal{L}(\theta)$, (ii) is from Lemma 3.2. Combining (L.3) and (L.4), we have

$$\|\widetilde{\Delta}_{\overline{\mathcal{S}}^*}\|_1 \leq \frac{5}{2}\|\widetilde{\Delta}_{\mathcal{S}^*}\|_1. \tag{L.5}$$

Next, we consider the following sequence of sets:

$$\mathcal{S}_0 = \left\{ j \in \overline{\mathcal{S}}^* : \sum_{m \in \overline{\mathcal{S}}^*} \mathbb{1}(|\widetilde{\theta}_m| \geq |\widetilde{\theta}_j|) \leq \widetilde{s} \right\} \text{ and}$$
$$\mathcal{S}_i = \left\{ j \in \overline{\mathcal{S}}^* \setminus \bigcup_{k<i} \mathcal{S}_k : \sum_{m \in \overline{\mathcal{S}}^* \setminus \bigcup_{k<i} \mathcal{S}_k} \mathbb{1}(|\widetilde{\theta}_m| \geq |\widetilde{\theta}_j|) \leq \widetilde{s} \right\} \text{ for all } i = 1, 2, \ldots.$$

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

Lemma L.1 (Adapted from Lemma 6.9 in Bühlmann and Van De Geer (2011) by setting $q = 2$)**.** Let $v = [v_1, v_2, \ldots]^\top$ with $v_1 \geq v_2 \geq \ldots \geq 0$. For any $s \in \{1, 2, \ldots\}$, we have

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

Denote $\mathcal{A} = \mathcal{S}^* \cup \mathcal{S}_0$. Then we have

$$\sum_{i \geq 1} \|\widetilde{\Delta}_{\mathcal{S}_i}\|_2 \overset{(i)}{\leq} \frac{1}{\sqrt{\widetilde{s}}}\|\widetilde{\Delta}_{\overline{\mathcal{S}}^*}\|_1 \overset{(ii)}{\leq} \frac{5}{2}\sqrt{\frac{s^*}{\widetilde{s}}}\|\widetilde{\Delta}_{\mathcal{S}^*}\|_2 \leq \frac{5}{2}\sqrt{\frac{s^*}{\widetilde{s}}}\|\widetilde{\Delta}_{\mathcal{A}}\|_2, \tag{L.6}$$

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

$$\|\check{\theta} - \theta^*\|_2 = (1-\beta)\|\widetilde{\theta} - \theta^*\|_2 \leq \sqrt{r},$$

which implies $\mathcal{L}(\check{\theta})$ satisfies RSC/RSS for $\check{\theta}$ restricted on a sparse set by Lemma 3.2. Then we have

$$|\widetilde{\Delta}_{\mathcal{A}}^\top \nabla_{\mathcal{A},\mathcal{A}}\mathcal{L}(\check{\theta})\widetilde{\Delta}_{\mathcal{A}}| \leq \sum_{i \geq 1} |\widetilde{\Delta}_{\mathcal{S}_i}^\top \nabla_{\mathcal{S}_i,\mathcal{A}}\mathcal{L}(\check{\theta})\widetilde{\Delta}_{\mathcal{A}}| \leq \rho^+_{s^*+\widetilde{s}}\|\widetilde{\Delta}_{\mathcal{A}}\|_2 \sum_{i \geq 1}\|\widetilde{\Delta}_{\mathcal{S}_i}\|_2$$
$$\overset{(i)}{\leq} \frac{5}{2}\sqrt{\frac{s^*}{\widetilde{s}}}\rho^+_{s^*+\widetilde{s}}\|\widetilde{\Delta}_{\mathcal{A}}\|_2^2, \tag{L.7}$$

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

$$\widetilde{\Delta}_{\mathcal{A}}^{\top}\nabla_{\mathcal{A},\mathcal{A}}\mathcal{L}(\check{\theta})\widetilde{\Delta}_{\mathcal{A}} \geq \rho^{-}_{s^*+\widetilde{s}}\|\widetilde{\Delta}_{\mathcal{A}}\|_2^2. \tag{L.8}$$

Then we have w.h.p.

$$\begin{aligned}
\widetilde{\Delta}\nabla\mathcal{L}(\check{\theta})\widetilde{\Delta} &= \widetilde{\Delta}_{\mathcal{A}}^{\top}\nabla_{\mathcal{A},\mathcal{A}}\mathcal{L}(\check{\theta})\widetilde{\Delta}_{\mathcal{A}} + 2\widetilde{\Delta}_{\overline{\mathcal{A}}}^{\top}\nabla_{\overline{\mathcal{A}},\mathcal{A}}\mathcal{L}(\check{\theta})\widetilde{\Delta}_{\mathcal{A}} + \widetilde{\Delta}_{\overline{\mathcal{A}}}^{\top}\nabla_{\overline{\mathcal{A}},\overline{\mathcal{A}}}\mathcal{L}(\check{\theta})\widetilde{\Delta}_{\overline{\mathcal{A}}} \\
&\geq \widetilde{\Delta}_{\mathcal{A}}^{\top}\nabla_{\mathcal{A},\mathcal{A}}\mathcal{L}(\check{\theta})\widetilde{\Delta}_{\mathcal{A}} - 2|\widetilde{\Delta}_{\overline{\mathcal{A}}}^{\top}\nabla_{\overline{\mathcal{A}},\mathcal{A}}\mathcal{L}(\check{\theta})\widetilde{\Delta}_{\mathcal{A}}| \\
&\overset{(i)}{\geq} \left(\rho^{-}_{s^*+\widetilde{s}} - 5\sqrt{\frac{s^*}{\widetilde{s}}}\rho^{+}_{s^*+\widetilde{s}}\right)\|\widetilde{\Delta}_{\mathcal{A}}\|_2^2 \overset{(ii)}{\geq} \frac{9}{14}\rho^{-}_{s^*+\widetilde{s}}\|\widetilde{\Delta}_{\mathcal{A}}\|_2^2,
\end{aligned}$$

where (i) is from (L.7) and (L.8), (ii) is from Lemma 3.2. This implies

$$\begin{aligned}
\mathcal{L}(\widetilde{\theta}) - \mathcal{L}(\theta^*) &= \nabla\mathcal{L}(\theta^*)^{\top}\widetilde{\Delta} + \frac{1}{2}\widetilde{\Delta}\nabla\mathcal{L}(\check{\theta})\widetilde{\Delta} \geq \nabla\mathcal{L}(\theta^*)^{\top}\widetilde{\Delta} + \frac{9}{28}\rho^{-}_{s^*+\widetilde{s}}\|\widetilde{\Delta}_{\mathcal{A}}\|_2^2 \\
&\overset{(i)}{\geq} \frac{9}{28}\rho^{-}_{s^*+\widetilde{s}}\|\widetilde{\Delta}_{\mathcal{A}}\|_2^2 - \frac{\lambda}{6}\|\widetilde{\Delta}_{\mathcal{S}^*}\|_1 - \frac{\lambda}{6}\|\widetilde{\Delta}_{\overline{\mathcal{S}}^*}\|_1,
\end{aligned} \tag{L.9}$$

where (i) is from $\lambda \geq \lambda_{[N]} \geq 6\|\nabla\mathcal{L}(\theta^*)\|_\infty$. Combining (L.3) and (L.9), we have

$$\rho^{-}_{s^*+\widetilde{s}}\|\widetilde{\Delta}_{\mathcal{S}^*}\|_2^2 \leq \rho^{-}_{s^*+\widetilde{s}}\|\widetilde{\Delta}_{\mathcal{A}}\|_2^2 \leq \frac{8}{3}\lambda\|\widetilde{\Delta}_{\mathcal{S}^*}\|_1 \leq \frac{8}{3}\lambda\sqrt{s^*}\|\widetilde{\Delta}_{\mathcal{S}^*}\|_2 \leq \frac{8}{3}\lambda\sqrt{s^*}\|\widetilde{\Delta}_{\mathcal{A}}\|_2.$$

This implies

$$\|\widetilde{\Delta}_{\mathcal{S}^*}\|_2 \leq \|\widetilde{\Delta}_{\mathcal{A}}\|_2 \leq \frac{8\lambda\sqrt{s^*}}{3\rho^{-}_{s^*+\widetilde{s}}} \text{ and } \|\widetilde{\Delta}_{\mathcal{S}^*}\|_1 \leq \frac{8\lambda s^*}{3\rho^{-}_{s^*+\widetilde{s}}}. \tag{L.10}$$

Then we have

$$\|\widetilde{\Delta}_{\overline{\mathcal{A}}}\|_2 \leq \sum_{i\geq 1}\|\widetilde{\Delta}_{\mathcal{S}_i}\|_2 \overset{(i)}{\leq} \frac{1}{\sqrt{\widetilde{s}}}\|\widetilde{\Delta}_{\overline{\mathcal{S}}^*}\|_1 \overset{(ii)}{\leq} \frac{5}{2}\sqrt{\frac{1}{s^*}}\|\widetilde{\Delta}_{\mathcal{S}^*}\|_1 \overset{(iii)}{\leq} \frac{20\lambda\sqrt{s^*}}{3\rho^{-}_{s^*+\widetilde{s}}}, \tag{L.11}$$

where (i) is rom Lemma L.1 with $s = \widetilde{s}$, (ii) is from (L.5) and $\widetilde{s} \geq s^*$ and (iii) is from (L.10). Combining (L.10) and (L.11), we have

$$\|\widetilde{\Delta}\|_2 = \sqrt{\|\widetilde{\Delta}_{\mathcal{A}}\|_2^2 + \|\widetilde{\Delta}_{\overline{\mathcal{A}}}\|_2^2} \leq \frac{8\lambda\sqrt{s^*}}{\rho^{-}_{s^*+\widetilde{s}}} < \sqrt{r},$$

where the last inequality is from the condition $\frac{\rho^{-}_{s^*+\widetilde{s}}}{8}\sqrt{\frac{r}{s^*}} > \lambda$. This conflicts with (L.1), which indicates that $\|\theta - \theta^*\|_2 \leq \sqrt{r}$.

Part 2. We next demonstrate the sparsity of θ. From $\lambda > \lambda_{[N]} \geq 6\|\nabla\mathcal{L}(\theta^*)\|_\infty$, we have

$$\left|\left\{i \in \overline{\mathcal{S}}^* : |\nabla_i\mathcal{L}(\theta^*)| \geq \frac{\lambda}{6}\right\}\right| = 0. \tag{L.12}$$

Denote $\check{\mathcal{S}}_1 = \left\{i \in \overline{\mathcal{S}}^* : |\nabla_i \mathcal{L}(\theta) - \nabla_i \mathcal{L}(\theta^*)| \geq \frac{\lambda}{2}\right\}$ and $\check{s}_1 = |\check{\mathcal{S}}_1|$. Then there exists some $b \in \mathbb{R}^d$ such that $\|b\|_\infty = 1$, $\|b\|_0 \leq \check{s}_1$ and $b^\top(\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta^*)) \geq \frac{\lambda\check{s}_1}{2}$. Then by the mean value theorem, we have for some $\check{\theta} = (1-\alpha)\theta + \alpha\theta^*$ with $\alpha \in [0,1]$, $\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta^*) = \nabla^2\mathcal{L}(\check{\theta})\Delta$, where $\Delta = \theta - \theta^*$. Then we have

$$\begin{aligned}\frac{\lambda\check{s}_1}{2} \leq b^\top\nabla^2\mathcal{L}(\check{\theta})\Delta &\overset{(i)}{\leq} \sqrt{b^\top\nabla^2\mathcal{L}(\check{\theta})b}\sqrt{\Delta^\top\nabla^2\mathcal{L}(\check{\theta})\Delta} \\ &\overset{(ii)}{\leq} \sqrt{\check{s}_1\rho^+_{\check{s}_1}}\sqrt{\Delta^\top(\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta^*))},\end{aligned} \tag{L.13}$$

where (i) is from the generalized Cauchy-Schwarz inequality, (ii) is from the definition of RSS and the fact that $\|b\|_2 \leq \sqrt{\check{s}_1}\|b\|_\infty = \sqrt{\check{s}_1}$. Let g achieve $\min_{g\in\partial\|\theta\|_1}\mathcal{F}_\lambda(\theta)$. Further, we have

$$\begin{aligned}\Delta^\top(\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta^*)) &\leq \|\Delta\|_1\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta^*)\|_\infty \\ &\leq \|\Delta\|_1(\|\nabla\mathcal{L}(\theta^*)\|_\infty + \|\nabla\mathcal{L}(\theta)\|_\infty) \\ &\leq \|\Delta\|_1(\|\nabla\mathcal{L}(\theta^*)\|_\infty + \|\nabla\mathcal{L}(\theta) + \lambda g\|_\infty + \lambda\|g\|_\infty) \\ &\overset{(i)}{\leq} \frac{28\lambda s^*}{3\rho^-_{s^*+\widetilde{s}}}(\frac{\lambda}{6} + \frac{\lambda}{4} + \lambda) \leq \frac{14\lambda^2 s^*}{\rho^-_{s^*+\widetilde{s}}},\end{aligned} \tag{L.14}$$

where (i) is from combining (L.5) and (L.10), condition on λ, approximate KKT condition and $\|g\|_\infty \leq 1$. Combining (L.13) and (L.14), we have $\frac{\sqrt{\check{s}_1}}{2} \leq \sqrt{\frac{14\rho^+_{\check{s}_1}s^*}{\rho^-_{s^*+\widetilde{s}}}}$, which further implies

$$\check{s}_1 \leq \frac{56\rho^+_{\check{s}_1}s^*}{\rho^-_{s^*+\widetilde{s}}} \leq 56\kappa_{s^*+2\widetilde{s}}s^* \leq \widetilde{s}. \tag{L.15}$$

For any $v \in \mathbb{R}^d$ that satisfies $\|v\|_\infty \leq 1$, we have

$$\check{\mathcal{S}}_2 = \left\{i \in \overline{\mathcal{S}}^* : \left|\nabla_i\mathcal{L}(\theta) + \frac{\lambda}{6}v_i\right| \geq \frac{2\lambda}{3}\right\} \subseteq \left\{i \in \overline{\mathcal{S}}^* : |\nabla_i\mathcal{L}(\theta^*)| \geq \frac{\lambda}{6}\right\}\bigcup\check{\mathcal{S}}_1 = \check{\mathcal{S}}_1.$$

Then we have $|\check{\mathcal{S}}_2| \leq |\check{\mathcal{S}}_1| \leq \widetilde{s}$. Since for any $i \in \overline{\mathcal{S}}^*$ and $\left|\nabla_i\mathcal{L}(\theta) + \frac{\lambda}{6}v_i\right| < \frac{2\lambda}{3}$, we can find g_i that satisfies $|g_i| \leq 1$ such that $\nabla_i\mathcal{L}(\theta) + \frac{\lambda}{6}v_i + \lambda g_i = 0$ which implies $\theta_i = 0$, then we have

$$\left|\left\{i \in \overline{\mathcal{S}}^* : \left|\nabla_i\mathcal{L}(\theta) + \frac{\lambda}{6}v_i\right| \geq \frac{2\lambda}{3}\right\}\right| \leq \check{s}_1.$$

Therefore, we have $\|\theta_{\overline{\mathcal{S}}^*}\|_0 \leq |\check{\mathcal{S}}_2| \leq \widetilde{s}$.

L.2 Proof of Lemma F.2

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

$$\|\nabla\mathcal{L}(\widehat{\theta}_{[K-1]}) + \lambda_{[K-1]}g\|_\infty \leq \lambda_{[K-1]}/4. \tag{L.16}$$

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

$$\begin{aligned}\omega_{\lambda_{[K]}}(\widehat{\theta}_{[K-1]}) &\le \|\nabla\mathcal{L}(\widehat{\theta}_{[K-1]}) + \lambda_{[K]}g\|_\infty = \|\nabla\mathcal{L}(\widehat{\theta}_{[K-1]}) + \lambda_{[K-1]}g + (\lambda_{[K]} - \lambda_{[K-1]})g\|_\infty \\ &\le \|\nabla\mathcal{L}(\widehat{\theta}_{[K-1]}) + \lambda_{[K-1]}g\|_\infty + |\lambda_{[K]} - \lambda_{[K-1]}| \cdot \|g\|_\infty \overset{(i)}{\le} \lambda_{[K-1]}/4 + (1-\eta_\lambda)\lambda_{[K-1]} \\ &\overset{(ii)}{\le} \lambda_{[K]}/2,\end{aligned}$$

where (i) is from (L.16) and choice of $\lambda_{[K]}$, (ii) is from the condition on η_λ.