A General Theory of Pathwise Coordinate Optimization for Nonconvex Sparse Learning*

Tuo Zhao† Han Liu‡ Tong Zhang§

Abstract

The pathwise coordinate optimization is one of the most important computational frameworks for solving high dimensional convex and nonconvex sparse learning problems. It differs from the classical coordinate optimization algorithms in three salient features: warm start initialization, active set updating, and strong rule for coordinate preselection. These three features grant superior empirical performance, but also pose significant challenge to theoretical analysis. To tackle this long lasting problem, we develop a new theory showing that these three features play pivotal roles in guaranteeing the outstanding statistical and computational performance of the pathwise coordinate optimization framework. In particular, we analyze the existing methods for pathwise coordinate optimization and provide new theoretical insights into them. The obtained theory motivates the development of several modifications to improve the pathwise coordinate optimization framework, which guarantees linear convergence to a unique sparse local optimum with optimal statistical properties (e.g. minimax optimality and oracle properties). This is the first result establishing the computational and statistical guarantees of the pathwise coordinate optimization framework in high dimensions. Thorough numerical experiments are provided to back up our theory.

1 Introduction

Modern data acquisition routinely produces massive amount of high dimensional data sets, where the number of variables $d$ greatly exceeds the sample size $n$. Such data include but not limited to chip data from high throughput genomic experiments (Neale et al., 2012), image data from functional Magnetic Resonance Imaging (fMRI, Eloyan et al. (2012)), proteomic data from tandem mass spectrometry analysis (Adams et al., 2007), and climate data from geographically distributed data centers (Hair et al., 2006). To handle the curse of dimensionality, we often assume only a small subset of variables are relevant in modeling (Guyon and Elisseeff, 2003). Such a sparsity assumption motivates various sparse learning approaches. For example, in sparse linear regression, we consider

*The R package picasso implementing the proposed method is available on the Comprehensive R Archive Network http://cran.r-project.org/web/packages/picasso/.
†Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218 USA; e-mail: tour@cs.jhu.edu.
‡Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544 USA; e-mail: hanliu@princeton.edu.
§Department of Statistics, Rutgers University, Piscataway, NJ, 08854 USA; e-mail: tzhang@stat.rutgers.edu.
a linear model \( y = X\theta^* + \epsilon \), where \( y \in \mathbb{R}^n \) is the response vector, \( X \in \mathbb{R}^{n \times d} \) is the design matrix, \( \theta^* = (\theta_1, ..., \theta_d)^\top \in \mathbb{R}^d \) is the unknown sparse regression coefficient vector, and \( \epsilon \sim N(0, \sigma^2 I) \) is the noise vector. Let \( \| \cdot \|_2 \) denote the \( \ell_2 \) norm, and \( R_\lambda(\theta) \) denote a sparsity-inducing regularization function (a.k.a regularizer) with a tuning parameter \( \lambda > 0 \). We estimate \( \theta^* \) by solving the regularized least square problem

\[
\min_{\theta \in \mathbb{R}^d} \mathcal{F}_\lambda(\theta), \quad \text{where} \quad \mathcal{F}_\lambda(\theta) = \frac{1}{2n} \| y - X\theta \|_2^2 + R_\lambda(\theta). \tag{1.1}
\]

Popular choices of \( R_\lambda(\theta) \) are usually coordinate decomposable (i.e., \( R_\lambda(\theta) = \sum_{j=1}^d r_\lambda(\theta_j) \)), including the \( \ell_1 \) regularizer (also known as Lasso, Tibshirani (1996), Chen et al. (1998)), SCAD regularizer (Fan and Li, 2001), and MCP regularizer (Zhang, 2010a).

Lasso is computationally tractable due to its convexity. However, it incurs large estimation bias to parameter estimation, and requires a very restrictive irrepresentable condition to attain variable selection consistency (Zhao and Yu, 2006; Wainwright, 2009; Zou, 2006; Meinshausen and Bühlmann, 2006). To handle this challenge, statisticians have resorted to sparse learning with nonconvex regularization. Fan and Li (2001); Kim et al. (2008); Zhang (2010a); Kim and Kwon (2012) investigate nonconvex regularized least square estimation using SCAD and MCP regularizers. Particularly, they show that there exits a local optimum that attains the oracle properties under much weaker conditions. However, they did not provide specific algorithms that find such a local optimum.

### 1.1 Pathwise Coordinate Optimization

Typical algorithmic solutions to (1.1) developed in the past decade include the homotopy strategy (Efron et al., 2004), proximal gradient (or majorization-minimization) strategy (Nesterov, 2013), and coordinate optimization strategy (Fu, 1998; Friedman et al., 2007; Mazumder et al., 2011; Breheny and Huang, 2011; Tibshirani et al., 2012). Among these methods, the pathwise coordinate optimization framework proposed by Friedman et al. (2007); Mazumder et al. (2011); Tibshirani et al. (2012) gains significant empirical success. Scalable software packages such as GLMNET, SPARSENET, NCVREG, and HUGUE have been developed (Friedman et al., 2010; Mazumder et al., 2011; Breheny and Huang, 2011; Zhao et al., 2012).

When the coordinate optimization strategy was first introduced to solve (1.1), however, its superiority was not widely recognized. The earliest example dates back to Fu (1998), which adopted the classical coordinate minimization algorithm. Similar algorithms were also considered by Shevade and Keerthi (2003), Krishnapuram et al. (2005), Kooij et al. (2007), and Genkin et al. (2007). These algorithms are simple and straightforward: Given a solution \( \theta(t) \) at the \( t \)-th iteration, we randomly select a coordinate \( j \) and take an exact coordinate minimization step

\[
\theta_j^{(t+1)} = \arg\min_{\theta_j} \mathcal{F}_\lambda(\theta_j, \theta_j^{(t)}), \tag{1.2}
\]

where \( \theta_{-j} \) is a subvector of \( \theta \) with the \( j \)-th entry removed. As will be shown in §2, (1.2) admits a closed form solution for the \( \ell_1 \), SCAD, and MCP regularizers. For notational simplicity, we write \( \theta_j^{(t+1)} = T_{\lambda,j}(\theta_j^{(t)}) \). Once \( \theta_j^{(t+1)} \) is obtained, we take \( \theta_{-j}^{(t+1)} = \theta_{-j}^{(t)} \). Such coordinate minimization algorithms, though simple, may not be efficient in theory and practice. Existing computational
theory shows that these algorithms only attain sublinear rates of convergence to local optima in high dimensions (Shalev-Shwartz and Tewari, 2011; Richtárik and Takáč, 2012; Lu and Xiao, 2014). Moreover, no theoretical guarantee has been established on statistical properties of the obtained estimators for nonconvex regularization. Therefore the coordinate optimization was almost neglected until a recent rediscovery by Friedman et al. (2007), Mazumder et al. (2011), and Tibshirani et al. (2012). \(^1\)

As illustrated in Figure 1 and Algorithm 1, Friedman et al. (2007), Mazumder et al. (2011), and Tibshirani et al. (2012) integrate the warm start initialization, active set updating strategy, and strong rule for coordinate preselection into the classical coordinate optimization. The warm start initialization, also referred as the pathwise optimization scheme, optimizes the objective function in a multistage way with a sequence of decreasing regularization parameters corresponding to solutions from sparse to dense. For each stage, the algorithm initializes using the solution from the previous stage; The active set exploits the solution sparsity to speed up computation (Boyd and Vandenberghe, 2009). Within each pathwise optimization stage, the active set updating contains two nested loops: In the outer loop, the algorithm first divides all coordinates into active ones (active set) and inactive ones (inactive set) based on some heuristic coordinate gradient thresholding rule, also known as the strong rule proposed in Tibshirani et al. (2012). Within each outer loop, an inner loop is called to conduct coordinate optimization. In general, the algorithm runs an inner loop on the current active coordinates until convergence, with all inactive coordinates remain zero. The algorithm then exploits some heuristic rule to identify a new active set, which further decreases the objective value and repeats the inner loops. The optimization within each stage terminates when the active set in the outer loop no longer changes. In practice, the warm start initialization, active set updating strategies, and strong rule for coordinate preselection significantly boost the computational performance, making pathwise coordinate optimization one of the most important computational frameworks for solving (1.1).

Despite of the popularity of the pathwise coordinate optimization framework, we are still in lack of adequate theory to justify its superior computational performance due to its complex algorithmic structure. Therefore the warm start initialization, active set updating strategy, and strong rule for coordinate preselection are only considered as engineering heuristics in existing literature (Friedman et al., 2007; Mazumder et al., 2011; Tibshirani et al., 2012). On the other hand, many experimental results have shown that the pathwise coordinate optimization framework is effective at finding local optima with good statistical properties in practice, yet no theoretical guarantee has been established. Thus a gap exists between theory and practice.

### 1.2 Our Contribution

To bridge this gap, we develop new theory and new algorithms to demonstrate the superiority of the pathwise coordinate optimization framework. Specifically, the contribution of this paper can be summarized as follows:

(I) We develop a new theory for analyzing the computational and statistical properties of the

\(^1\)A brief history on applying coordinate optimization to sparse learning problems is presented in Hastie (2009).
Figure 1: The pathwise coordinate optimization framework contains 3 nested loops: (I) Warm start initialization; (II) Active set updating and strong rule for coordinate preselection; (III) Active coordinate minimization. Many empirical results have corroborated its outstanding practical performance.

Algorithm 1: The pathwise coordinate optimization framework contains 3 nested loops: (I) Warm start initialization; (II) Active set updating and strong rule for coordinate preselection; (III) Active coordinate minimization. Many empirical results have corroborated its outstanding practical performance.

| Initialize: $\lambda_0, \hat{\theta}^{(0)}, \eta \in (0, 1)$ |
| For $K \leftarrow 1, 2, ..., N$ |
| $\lambda_K = \eta \lambda_{K-1}, \theta^{[0]} \leftarrow \theta^{(K-1)}, m \leftarrow 0$ (Warm Start Initialization) |
| Initialize an active set $A \subseteq \{1, ..., d\}$ (Strong Rule for Coordinate Preselection) |
| Repeat |
| $\theta^+ \leftarrow \theta^{[m]}$ |
| Repeat |
| $\theta^+_j \leftarrow T_{\lambda_K,j}(\theta^+)$ for all $j \in A$ (Active Coordinate Minimization) |
| Until convergence |
| Update the active set $A \subseteq \{1, ..., d\}$ (Active Set Updating) |
| $\theta^{[m+1]} \leftarrow \theta^+, m \leftarrow m + 1$ |
| Until the active set no longer changes |
| $\hat{\theta}^{(K)} \leftarrow \theta^{[m]}$ |
| Output: $\{\hat{\theta}^{(K)}\}_{K=0}^{N}$ |

pathwise coordinate optimization framework. Particularly, our theory thoroughly analyzes the warm start initialization, active set updating strategy, and strong rule for coordinate preselection. It shows that these strategies play pivotal roles in guaranteeing the solution sparsity and restricted strong convexity, which are the two most important conditions for statistical and computational performance in high dimensions (See more details in §4, Candes and Tao (2005); Zhang and Huang (2008); Bickel et al. (2009); Zhang (2009)). In addition,
our theory provides new insights, and indicates two possible drawbacks of existing pathwise coordinate optimization algorithms: (1) The commonly used heuristic cyclic selection rule for active set updating does not necessarily guarantee the solution sparsity; (2) Using an all zero solution as the initialization may not guarantee the restricted strong convexity, when we extend the algorithms to solving other regularized M-estimation problems with more general loss functions.

(II) To guarantee the solution sparsity, we propose a new algorithm, named PICASSO (PathwIse CalibrAteD Sparse Shooting algOrithm), which improves the pathwise coordinate optimization framework. Particularly, we propose three new active set selection rules, which guarantee the solution sparsity throughout all iterations. The modification, though simple, has a profound impact: Our theory shows that PICASSO attains global linear convergence to a unique sparse local optimum $\theta$ with oracle statistical properties: There exists a universal constant $C$ such that with high probability, we have

$$
\|\hat{\theta} - \theta^*\|_2 \leq C \cdot \sigma \sqrt{\frac{s^*}{n}} \quad \text{and} \quad \text{supp}(\hat{\theta}) = \text{supp} (\theta^*),
$$

where $\text{supp}(\theta) = \{j \mid \theta_j \neq 0\}$ and $s^*$ denotes the number of nonzero entries of $\theta^*$. To the best of our knowledge, this is the first result establishing the computational and statistical properties of the pathwise coordinate optimization framework in high dimensions.

(III) To deal with more general loss functions, we propose a new convex relaxation approach. Particularly, we first approximately solve a convex relaxation of (1.1). We then plug the obtained solution into PICASSO as the initial solution. Theoretically, we show that the proposed convex relaxation approach also guarantees the same computational and statistical performance of PICASSO for solving regularized M-estimation problems with general loss functions.

Several proximal gradient algorithms in exiting literature are closely related to PICASSO. By exploiting similar sparsity structures of the optimization problem, Wang et al. (2014); Zhao and Liu (2014); Loh and Wainwright (2015) show that these proximal gradient algorithms also attain linear or approximately linear rates of convergence to local optima with good statistical properties. We will compare these algorithms with PICASSO in §6 and §7.

The rest of this paper is organized as follows: In §2, we introduce notations and briefly review some regularizers of our interests; In §3, we present the PICASSO algorithm; In §4 we develop a new theory for analyzing the pathwise coordinate optimization framework, and establish the computational and statistical properties of PICASSO for sparse linear regression; In §5, we extend PICASSO to sparse logistic regression and provide theoretical guarantees; In §6, we present thorough numerical experiments to support our theory; In §7, we discuss related work; In §8, we present the technical proof for all theorems.

**Notations:** Given a vector $\mathbf{v} = (v_1, \ldots, v_d)^\top \in \mathbb{R}^d$, we define vector norms: $\|\mathbf{v}\|_1 = \sum_j |v_j|$, $\|\mathbf{v}\|_2^2 = \sum_j v_j^2$, and $\|\mathbf{v}\|_\infty = \max_j |v_j|$. We denote the number of nonzero entries in $\mathbf{v}$ as $\|\mathbf{v}\|_0 = \sum_j \mathbb{1}_{\{v_j \neq 0\}}$. 


We define the matrix norms \( \|A\|_F^2 = \sum_{i=1}^{d} \|A_{i\cdot}\|_2^2 \) and \( \|A\|_2^2 = \sum_{j=1}^{d} \|A_{\cdot j}\|_2^2 \), where \( A_{i\cdot} \) and \( A_{\cdot j} \) are the \( i \)-th row of \( A \) and the \( j \)-th column of \( A \), respectively. We use \( \Lambda_{\text{max}}(A) \) and \( \Lambda_{\text{min}}(A) \) be the largest and smallest eigenvalues of \( A \). Let \( \Phi_1(A) \), \( \Phi_2(A) \) be the singular values of \( A \), we define the matrix norms \( \|A\|_F = \sqrt{\sum_{i=1}^{d} \|A_{i\cdot}\|_2^2} \), \( \|A\|_1 = \sum_{j=1}^{d} \|A_{\cdot j}\|_1 \), \( \|A\|_2 = \max_j \Phi_j(A) \). We denote \( A_{i\cdot} \) as the subvector of \( A \) with the \( i \)-th row removed, and \( A_{\cdot j} \) as the subvector of \( A \) with the \( j \)-th column removed. We denote \( A_{i\cdot j} \) as the submatrix of \( A \) with the \( i \)-th row and the \( j \)-th column removed. Let \( A \subseteq \{1, ..., d\} \) be an index set. We use \( A_{\text{all}} \) to denote a subvector of \( v \) by extracting all entries of \( v \) with indices in \( A \). Given a matrix \( A \in \mathbb{R}^{d \times d} \), we use \( A_{i\cdot j} = (A_{i1}, ..., A_{id})^\top \) to denote the \( j \)-th column of \( A \), and \( A_{k\cdot} = (A_{k1}, ..., A_{kd})^\top \) to denote the \( k \)-th row of \( A \). Let \( \Lambda_{\text{max}}(A) \) and \( \Lambda_{\text{min}}(A) \) be the largest and smallest eigenvalues of \( A \) respectively.

2 Background

We first briefly review some nonconvex regularizers of our interests, and then derive updating formulas for the exact coordinate minimization in (1.2).

2.1 Folded Concave Regularization

For high dimensional problems, we are interested in exploiting sparsity-inducing regularizers to yield sparse estimators \( \hat{\theta} \), where \( \|\hat{\theta}\|_0 \ll n \ll d \). These regularizers are usually coordinate decomposable: \( R_\lambda(\theta) = \sum_{j=1}^{d} r_\lambda(\theta_j) \). For example, the most commonly used \( \ell_1 \) regularization can be written as \( R_\lambda(\theta) = \lambda \|\theta\|_1 = \lambda \sum_{j=1}^{d} |\theta_j| \) with \( r_\lambda(\theta_j) = |\theta_j| \) for \( j = 1, ..., d \).

The \( \ell_1 \) norm is convex and computationally tractable, but often induces large estimation bias (Fan and Li, 2001; Zhang, 2010a). To address this issue, many nonconvex (or folded concave) regularizers have been proposed to obtain nearly unbiased estimators. Two notable examples in existing literature are the SCAD (Smooth Clipped Absolute Deviation, Fan and Li (2001)) regularization with

\[
r_\lambda(\theta_j) = \lambda |\theta_j| \cdot 1(|\theta_j| \leq \lambda) - \frac{\theta_j^2 - 2\lambda \gamma |\theta_j| + \lambda^2}{2(\gamma - 1)} \cdot 1(\lambda < |\theta_j| \leq \lambda \gamma) + \frac{(\gamma + 1)\lambda^2}{2} \cdot 1(|\theta_j| > \lambda \gamma) \text{ for } \gamma > 2, \tag{2.1}
\]

and MCP (Minimax Concavity Penalty, Zhang (2010a)) regularization with

\[
r_\lambda(\theta_j) = \lambda \left( |\theta_j| - \frac{\theta_j^2}{2\lambda \gamma} \right) \cdot 1(|\theta_j| < \lambda \gamma) + \frac{\lambda^2 \gamma}{2} \cdot 1(|\theta_j| \geq \lambda \gamma) \text{ for } \gamma > 1. \tag{2.2}
\]

By examining (2.1) and (2.2), we observe that both regularizers share a generic form

\[
R_\lambda(\theta) = \lambda \|\theta\|_1 + H_\lambda(\theta), \tag{2.3}
\]

where \( H_\lambda(\theta) = \sum_{j=1}^{d} h_\lambda(|\theta_j|) \) is a smooth concave and coordinate decomposable function. Particu-
larly, the SCAD regularization has
\[ h_\lambda(|\theta_j|) = \frac{2\lambda |\theta_j| - |\theta_j|^2 - \lambda^2}{2(\gamma - 1)} \cdot 1(\lambda < |\theta_j| \leq \lambda \gamma) \] (2.4)
and the MCP regularization has
\[ h_\lambda(|\theta_j|) = -\frac{\theta_j^2}{2\gamma} \cdot 1(|\theta_j| < \lambda \gamma) + \frac{\lambda^2 \gamma - 2\lambda |\theta_j|}{2} \cdot 1(|\theta_j| \geq \lambda \gamma). \] (2.5)
We present two examples of the SCAD and MCP regularizers in Figure 2. Existing literature has shown that the SCAD and MCP regularizers effectively reduce the estimation bias, and achieve better performance than that of the \(\ell_1\) regularization in both parameter estimation and support recovery. But their nonconvexity imposes great computational challenge. See more details in Fan and Li (2001); Zhang (2010a,b); Zhang and Zhang (2012); Fan et al. (2014); Xue et al. (2012); Wang et al. (2013, 2014).

2.2 Exact Coordinate Minimization
Recall that each exact coordinate minimization iteration solves the following optimization problem,
\[ \theta_j^{(t+1)} = T_{\lambda,j}(\theta_j^{(t)}) = \arg\min_{\theta_j} F_\lambda(\theta_j, \theta_{\setminus j}^{(t)}) = \arg\min_{\theta_j} \frac{1}{2n} \| y - X_{\setminus j} \theta_{\setminus j}^{(t)} - X_j \theta_j \|_2^2 + r_\lambda(\theta_j). \] (2.6)
For the \(\ell_1\), SCAD, and MCP regularizers, (2.6) admits a closed form solution. More specifically, without loss of generality, we assume that the design matrix \(X\) satisfies the column normalization condition, i.e., \(\|X_j\|_2 = \sqrt{n}\) for all \(j = 1, ..., d\). Let \(\bar{\theta}_j^{(t)} = \frac{1}{n} X_j^\top (y - X_{\setminus j} \theta_{\setminus j}^{(t)})\). Then we can verify:

- For the \(\ell_1\) regularizer, we have \(\theta_j^{(t+1)} = S_\lambda(\bar{\theta}_j^{(t)})\);
For the SCAD regularization, we have
\[ \theta_j^{(t+1)} = \begin{cases} 
\tilde{\theta}_j^{(t)} & \text{if } |\tilde{\theta}_j^{(t)}| \geq \gamma \lambda, \\
S_{\gamma \lambda / (\gamma - 1)}(\tilde{\theta}_j^{(t)}) & \text{if } |\tilde{\theta}_j^{(t)}| \in [2\lambda, \gamma \lambda), \\
S_{\lambda}(\tilde{\theta}_j^{(t)}) / (1 - 1/\gamma) & \text{if } |\tilde{\theta}_j^{(t)}| < 2\lambda; 
\end{cases} \]

For the MCP regularizer, we have
\[ \theta_j^{(t+1)} = \begin{cases} 
\tilde{\theta}_j^{(t)} & \text{if } |\tilde{\theta}_j^{(t)}| \geq \gamma \lambda, \\
S_{\lambda}(\tilde{\theta}_j^{(t)}) & \text{if } |\tilde{\theta}_j^{(t)}| < \gamma \lambda. 
\end{cases} \]

3 Pathwise Calibrated Sparse Shooting Algorithm

We derive the PICASSO algorithm for sparse linear regression. PICASSO is a pathwise coordinate optimization algorithm and contains three nested loops (as illustrated in Algorithm 1). For simplicity, we first introduce its inner loop, then its middle loop, and at last its outer loop.

3.1 Inner Loop: Iterates over Coordinates within an Active Set

We start with the inner loop of PICASSO, which is the active coordinate minimization (ActCooMin) algorithm. Recall that we are interested in the following nonconvex optimization problem
\[ \min_{\theta \in \mathbb{R}^d} F_\lambda(\theta), \quad \text{where } F_\lambda(\theta) = \frac{1}{2n} \|y - X\theta\|_2^2 + R_\lambda(\theta). \tag{3.1} \]

The ActCooMin algorithm solves (3.1) by iteratively conducting exact coordinate minimization, but it is only allowed to iterate over a subset of all coordinates, which is called “the active set”. Accordingly, the complementary set to the active set is called “the inactive set”, because the values of these coordinates do not change throughout all iterations of the inner loop. Since the active set usually contains only a few coordinates, the active set coordinate minimization algorithm is very scalable and efficient.

For notational simplicity, we denote the active and inactive sets by \( A \) and \( \overline{A} \) respectively. Here we select \( A \) and \( \overline{A} \) based on the sparsity pattern of the initial solution of the inner loop \( \theta^{(0)} \),
\[ A = \{ j \mid \theta_j^{(0)} \neq 0 \} \quad \text{and} \quad \overline{A} = \{ j \mid \theta_j^{(0)} = 0 \}. \]

The ActCooMin algorithm then minimizes (3.1) with all coordinates of \( \overline{A} \) staying at zero values,
\[ \min_{\theta \in \mathbb{R}^d} F_\lambda(\theta) \quad \text{subject to } \theta_{\overline{A}} = 0. \tag{3.2} \]

Different from the classical randomized coordinate minimization algorithm (introduced in §1), the ActCooMin algorithm iterates over all active coordinates in a cyclic order at each iteration. Without loss of generality, we assume
\[ |A| = s \quad \text{and} \quad A = \{ j_1, \ldots, j_s \} \subseteq \{ 1, \ldots, d \}, \]
where \( j_1 \leq j_2 \leq \ldots \leq j_s \). Given a solution \( \theta^{(t)} \) at the \( t \)-th iteration, we construct a sequence of auxiliary solutions \( \{ w^{(t+1,k)} \}_{k=0}^s \) to obtain the next solution \( \theta^{(t+1)} \). Particularly, for \( k = 0 \), we take \( w^{(t+1,0)} = \theta^{(t)} \); For \( k = 1, \ldots, s \), we take
\[ w_{jk}^{(t+1,k)} = T_{\lambda, j_k}(w^{(t+1,k-1)}) \quad \text{and} \quad w_{\setminus j_k}^{(t+1,k)} = w_{\setminus j_k}^{(t+1,k-1)}. \]
where $T_{\lambda,j,k}(\cdot)$ is defined in (2.6). Once we obtain $w^{(t+1,s)}$, we set $\theta^{(t+1)} = w^{(t+1,s)}$ for the next iteration. We terminate the ActCooMin algorithm when
\[
\|\theta^{(t+1)} - \theta^{(t)}\|_2 \leq \tau \lambda, \tag{3.3}
\]
where $\tau$ is a small convergence parameter (e.g. $10^{-5}$). We then take the output solution as $\widehat{\theta} = \theta^{(t+1)}$.

The ActCooMin algorithm only converges to a local optimum of (3.2), which is not necessarily a local optimum of (3.1). Thus PICASSO needs to combine this inner loop with some active set updating scheme, which allows the active set to change. This leads to the middle loop of PICASSO.

---

**Algorithm 2:** The active coordinate minimization algorithm (ActCooMin) cyclically iterates over only a small subset of all coordinates. Therefore its computation is scalable and efficient. Without loss of generality, we assume $|A| = s$ and $A = \{j_1, \ldots, j_s\} \subseteq \{1, \ldots, d\}$, where $j_1 \leq j_2 \leq \ldots \leq j_s$.

**Algorithm:**

\[
\widehat{\theta} \leftarrow \text{ActCooMin}(\lambda, \theta^{(0)}, A, \tau)
\]

**Initialize:** $t \leftarrow 0$

**Repeat**

\[
\begin{align*}
& w^{(t+1,0)} \leftarrow \theta^{(t)} \\
& \text{For } k \leftarrow 1, \ldots, s \\
& \quad w^{(t+1,k)} \leftarrow T_{\lambda,j_k}(w^{(t+1,k-1)}) \\
& \quad w^{(t+1,k)} \leftarrow w^{(t+1,k-1)} \\
& \quad \theta^{(t+1)} \leftarrow w^{(t+1,s)} \\
& t \leftarrow t + 1
\end{align*}
\]

**Until** $\|\theta^{(t+1)} - \theta^{(t)}\|_2 \leq \tau \lambda$

**Output:** $\widehat{\theta} \leftarrow \theta^{(t)}$.

---

### 3.2 Middle Loop: Iteratively Updates Active Sets

We then introduce the middle loop of PICASSO, which is the iterative active set updating (IteActUpd) algorithm. The IteActUpd algorithm simultaneously decreases the objective value and iteratively changes the active set to ensure convergence to a local optimum to (3.1). For notational simplicity, we denote the least square loss function and its gradient as
\[
L(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2 \quad \text{and} \quad \nabla L(\theta) = \frac{1}{n}X^\top(X\theta - y).
\]

#### 3.2.1 Active Set Initialization by Strong Rule

We first introduce how PICASSO initializes the active set for each middle loop. Suppose that an initial solution $\theta^{[0]}$ is supplied to the middle loop of PICASSO. Friedman et al. (2007) suggest a “simple rule” to initialize the active set based on the sparsity pattern of $\theta^{[0]}$,
\[
A_0 = \{j \mid \theta^{[0]}_{j} \neq 0\} \quad \text{and} \quad \overline{A}_0 = \{j \mid \theta^{[0]}_{j} = 0\}. \tag{3.4}
\]
But Tibshirani et al. (2012) show that (3.4) is sometimes too conservative, and suggest a more aggressive active set initialization procedure using “strong rule”, which often leads to better computational performance in practice. Specifically, given an active set initialization parameter $\varphi \in (0, 1)$, the strong rule\(^2\) for PICASSO initializes $A_0$ and $A_0^c$ as

$$A_0 = \{ j \mid \theta_j^{[0]} = 0, \ |\nabla_j L(\theta^{[0]})| \geq (1 - \varphi)\lambda \} \cup \{ j \mid \theta_j^{[0]} \neq 0 \},$$  

(3.5)

$$A_0^c = \{ j \mid \theta_j^{[0]} = 0, \ |\nabla_j L(\theta^{[0]})| < (1 - \varphi)\lambda \},$$  

(3.6)

where $\nabla_j L(\theta^{[0]})$ denotes the $j$-th entry of $\nabla L(\theta^{[0]})$. As can be seen from (3.5), the strong rule yields an active set, which is no smaller than the simple rule. Note that we need the initialization parameter $\varphi$ to be a reasonably small value (e.g. 0.1). Otherwise, the strong rule will select too many active coordinates and compromise the restricted strong convexity. We summarize the active set initialization procedure in Algorithm 3.

**Algorithm 3:** The active set of the middle loop can be initialized by either the simple rule (Friedman et al., 2007) or the strong rule (Tibshirani et al., 2012). The strong rule yields a larger active set, and often leads to better computational performance than the simple rule in practice.

**Algorithm:** $(A_0, A_0^c) \leftarrow \text{ActIntProc}(\lambda, \theta^{[0]}, \varphi)$

- **Simple Rule:** $A_0 \leftarrow \{ j \mid \theta_j^{[0]} \neq 0 \}$, $A_0^c = \{ j \mid \theta_j^{[0]} = 0 \}$
- **Strong Rule:** $A_0 \leftarrow \{ j \mid \theta_j^{[0]} = 0, \ |\nabla_j L(\theta^{[0]})| \geq (1 - \varphi)\lambda \} \cup \{ j \mid \theta_j^{[0]} \neq 0 \}$, $A_0^c \leftarrow \{ j \mid \theta_j^{[0]} = 0, \ |\nabla_j L(\theta^{[0]})| < (1 - \varphi)\lambda \}$

**Output:** $(A_0, A_0^c)$

### 3.2.2 Active Set Updating Strategy

We then introduce how PICASSO updates the active set at each iteration of the middle loop. Suppose that at the $m$-th iteration ($m \geq 1$), we are supplied with a solution $\theta^{[m]}$ with a pair of active and inactive sets defined as

$$A_m = \{ j \mid \theta_j^{[m]} \neq 0 \} \quad \text{and} \quad A_m^c = \{ j \mid \theta_j^{[m]} = 0 \}.$$

Each iteration of the IteActUpd algorithm contains two stages. The first stage conducts the active coordinate minimization algorithm over the active set $A_m$ until convergence, and returns a solution $\theta^{[m+0.5]}$. Note that the active coordinate minimization algorithm may yield zero values for some active coordinates. Accordingly, we remove these coordinates from the active set, and obtain a new pair of active and inactive sets as

$$A_{m+0.5} = \{ j \mid \theta_j^{[m+0.5]} \neq 0 \} \quad \text{and} \quad A_{m+0.5}^c = \{ j \mid \theta_j^{[m+0.5]} = 0 \}.$$

The second stage checks which inactive coordinates of $A_{m+0.5}$ should be added into the active set. Existing pathwise coordinate optimization algorithms usually add inactive coordinates into the active set based on a **cyclic selection rule** (Friedman et al., 2007; Mazumder et al., 2011).

\(^2\)Our proposed strong rule for PICASSO is slightly different from the sequential strong rule proposed in Tibshirani et al. (2012). See more details in Remark 3.2.
Particularly, they conduct exact coordinate minimization over all coordinates of $\mathcal{A}_{m+0.5}$ in a cyclic order. Accordingly, an inactive coordinate is added into the active set if the corresponding exact coordinate minimization yields a nonzero value. Such a cyclic selection rule, however, is heuristic and has no control over the solution sparsity. It may add too many inactive coordinates into the active set, which compromises the restricted strong convexity.

To address this issue, we propose the following three new active set selection rules.

(I) Greedy Selection: We select a coordinate by

$$k_m = \arg\max_{k \in \mathcal{A}_{m+0.5}} |\nabla_k L(\theta^{[m+0.5]})|,$$

where $\nabla_j L(\theta^{[m+0.5]})$ denotes the $j$-th entry of $\nabla L(\theta^{[m+0.5]})$. We terminate the IteActUpd algorithm if

$$|\nabla_{k_m} L(\theta^{[m+0.5]})| \leq (1 + \delta)\lambda,$$

where $\delta$ is a small convergence parameter (e.g. $10^{-5}$). Otherwise, we obtain $\theta^{[m+1]}$ by

$$\theta_{k_m}^{[m+1]} = T_{\lambda,k_m}(\theta^{[m+0.5]}) \quad \text{and} \quad \theta_{\backslash k_m}^{[m+1]} = \theta_{\backslash k_m}^{[m+0.5]},$$

and take the new active and inactive sets as

$$\mathcal{A}_{m+1} = \mathcal{A}_{m+0.5} \cup \{k_m\} \quad \text{and} \quad \mathcal{A}_{m+1} = \mathcal{A}_{m+0.5} \setminus \{k_m\}.$$
(II) Randomized Selection: We randomly select a coordinate $k_m$ from
\[ \mathcal{M}_{m+0.5} = \{ k \mid k \in \mathcal{A}_{m+0.5}, \ |\nabla_k \mathcal{L}(\theta^{[m+0.5]})| \geq (1 + \delta)\lambda \} \] (3.8)
with equal probability, where $\delta$ is defined in (3.7). We terminate the IteActUpd algorithm if $\mathcal{M}_{m+0.5}$ is an empty set, i.e., $\mathcal{M}_{m+0.5} = \emptyset$. Otherwise, we obtain $\theta^{[m+1]}$ by
\[ \theta_k^{[m+1]} = T_{\lambda,k_m}(\theta^{[m+0.5]}) \quad \text{and} \quad \theta_{\setminus k_m}^{[m+1]} = \theta_{\setminus k_m}^{[m+0.5]} , \]
and take the new active and inactive sets as
\[ \mathcal{A}_{m+1} = \mathcal{A}_{m+0.5} \cup \{ k_m \} \quad \text{and} \quad \mathcal{A}_{m+1} = \mathcal{A}_{m+0.5} \setminus \{ k_m \} . \]
We summarize the IteActUpd algorithm using the randomized selection rule in Algorithm 5.

**Remark 3.1.** The randomized selection procedure has an equivalent and efficient implementation as follows: (i) We generate a randomly shuffled order of all inactive coordinates of $\mathcal{A}_{m+0.5}$; (ii) We then cyclically check all inactive coordinates according to the order generated in (i), and select the first inactive coordinate $k$ satisfying $|\nabla_k \mathcal{L}(\theta^{[m+0.5]})| \geq (1 + \delta)\lambda$. If no inactive coordinate satisfies the requirement, i.e.,
\[ \max_{k \in \mathcal{A}_{m+0.5}} |\nabla_k \mathcal{L}(\theta^{[m+0.5]})| \leq (1 + \delta)\lambda , \]
then we have $\mathcal{M}_{m+0.5} = \emptyset$ and terminate the middle loop.

**Algorithm 5:** Similar to the greedy selection rule, the randomized selection rule also moves only one inactive coordinate to the active set in each iteration to encourage the sparsity of the active set.

**Algorithm:** \( \hat{\theta} \leftarrow \text{IteActUpd}(\lambda, \theta^{[0]}, \delta, \tau, \varphi) \)

**Initialize:** \( m \leftarrow 0, (\mathcal{A}_0, \mathcal{A}_0) \leftarrow \text{ActIntProc}(\lambda, \theta^{[0]}, \varphi) \)

**Repeat**

\[ \theta^{[m+0.5]} \leftarrow \text{ActCooMin}(\lambda, \theta^{[m]}, \mathcal{A}_m, \tau) \]
\[ \mathcal{A}_{m+0.5} \leftarrow \{ j \mid \theta_j^{[m+0.5]} \neq 0 \} \]
\[ \overline{\mathcal{A}}_{m+0.5} \leftarrow \{ j \mid \theta_j^{[m+0.5]} = 0 \} \]

Randomly sample a coordinate $k_m$ from
\[ \mathcal{M}_{m+0.5} = \{ k \mid k \in \mathcal{A}_{m+0.5}, \ |\nabla_k \mathcal{L}(\theta^{[m+0.5]})| \geq (1 + \delta)\lambda \} \]
with equal probability
\[ \theta_k^{[m+1]} \leftarrow T_{\lambda,k_m}(\theta^{[m+0.5]}) \]
\[ \theta_{\setminus k_m}^{[m+1]} \leftarrow \theta_{\setminus k_m}^{[m+0.5]} \]
\[ \mathcal{A}_{m+1} \leftarrow \mathcal{A}_{m+0.5} \cup \{ k_m \} \]
\[ \overline{\mathcal{A}}_{m+1} \leftarrow \overline{\mathcal{A}}_{m+0.5} \setminus \{ k_m \} \]
\[ m \leftarrow m + 1 \]

Until \[ \max_{k \in \mathcal{A}_{m+0.5}} |\nabla_k \mathcal{L}(\theta^{[m+0.5]})| \leq (1 + \delta)\lambda \]

**Output:** \( \hat{\theta} \leftarrow \theta^{[m]} \)

(III) Truncated Cyclic Selection: We choose to iterate over all coordinates of $\mathcal{A}_{m+0.5}$ in a cyclic
order. But different from the cyclic selection rule in Friedman et al. (2007); Mazumder et al. (2011), we conduct exact coordinate minimization over an inactive coordinate and add it into the active set only if the corresponding coordinate gradient has a sufficiently large magnitude. Otherwise, we make this coordinate stay in the inactive set.

More specifically, without loss of generality, we assume

\[
|\mathcal{A}_{m+0.5}| = g \quad \text{and} \quad \mathcal{A}_{m+0.5} = \{j_1, \ldots, j_g\} \subseteq \{1, \ldots, d\},
\]

where \(j_1 \leq j_2 \leq \ldots \leq j_g\). We terminate the IteActUpd algorithm if

\[
\max_{k \in \mathcal{A}_{m+0.5}} |\nabla_k \mathcal{L}(\theta)^{m+0.5}| \leq (1 + \delta)\lambda.
\]

Otherwise, we construct a sequence of auxiliary solutions \(\{w^{[m+1,k]}\}_{k=0}^g\) as follows: For \(k = 0\), we set \(w^{[m+1,0]} = \theta^{[m+0.5]}\); For \(k = 1, \ldots, g\), we take \(w^{[m+1,k]} = w^{[m+1,k-1]}\) and

\[
w^{[m+1,k]} = \begin{cases} \mathcal{T}_{\lambda,j_k}(w^{[m+1,k-1]}) & \text{if } |\nabla_{j_k} \mathcal{L}(w^{[m+1,k-1]})| \geq (1 + \delta)\lambda, \\ w^{[m+1,k-1]} & \text{otherwise}, \end{cases}
\]

where \(\delta\) is defined in (3.7). Note that when \(\delta = 0\), the truncated cyclic selection rule is reduced to the cyclic selection rule in Friedman et al. (2007); Mazumder et al. (2011). Once we obtain all auxiliary solutions, we set \(\theta^{[m+1]} = w^{[m+1,g]}\), and take the new active and inactive sets based on the sparsity pattern of \(\theta^{[m+1]}\), i.e.,

\[
\mathcal{A}_{m+1} = \{j \mid \theta_j^{[m+1]} \neq 0\} \quad \text{and} \quad \mathcal{A}_{m+1} = \{j \mid \theta_j^{[m+1]} = 0\}.
\]

We summarize the iterative active set updating algorithm using the truncated cyclic selection rule in Algorithm 6.

The IteActUpd algorithm, though equipped with the three proposed active set selection rules and strong rule for coordinate preselection, can only guarantee the solution sparsity throughout all iterations when specifying an appropriate regularization parameter.\(^3\) Otherwise when an insufficiently large regularization parameter is supplied, the IteActUpd algorithm may still overselect active coordinates. To address this issue, we combine the IteActUpd algorithm with a pathwise regularization scheme, which leads to the outer loop of PICASSO.

### 3.3 Outer Loop: Iterates over Regularization Parameters

The outer loop of PICASSO is the warm start initialization (also known as the pathwise optimization scheme). It solves (1.1) indexed by a geometrically decreasing sequence of regularization parameters \(\{\lambda_K = \lambda_0\eta^K\}_{K=0}^N\) with a common decaying parameter \(\eta \in (0, 1)\), and outputs a sequence of \(N + 1\) solutions \(\{\hat{\theta}^{[K]}\}_{K=0}^N\) (also known as the solution path).

For sparse linear regression, PICASSO chooses the leading regularization parameter \(\lambda_0\) as \(\lambda_0 = \|\nabla \mathcal{L}(\theta)\|_\infty \leq \|rac{1}{n}X^T y\|_\infty\). Recall that \(\mathcal{H}_\lambda(\theta)\) is defined in (2.3). By verifying the KKT condition, we have

\[
\min_{\xi \in \partial \mathcal{H}_\lambda(\theta)} \|\nabla \mathcal{L}(\theta) + \nabla \mathcal{H}_\lambda(\theta) + \lambda_0\xi\|_\infty = \min_{\xi \in \partial \|\theta\|_1} \|\nabla \mathcal{L}(\theta) + \lambda_0\xi\|_\infty = 0,
\]

\(^3\)As will be shown in §4, the choice of the regularization parameter is determined by the initial solution of the middle loop.

13
To encourage the sparsity of the active set, the truncated cyclic selection rule only selects coordinates only when their corresponding coordinate gradients are sufficiently large in magnitude. Without loss of generality, we assume $|A_m| = g$ and $A_m = \{j_1, \ldots, j_g\}$, where $j_1 \leq j_2 \leq \ldots \leq j_g$.

**Algorithm 6:**

```
θ ← IteActUpd(λ, θ[0], δ, τ, ϕ)
```

**Initialize:**

```
m ← 0, (A_0, A_0) ← ActIntProc(λ, θ[0], ϕ)
```

**Repeat**

```
θ[m+0.5] ← ActCooMin(λ, θ[m], A_m, τ)
A_{m+0.5} ← \{j \mid θ_j^{[m+0.5]} \neq 0\}
A_{m+0.5} ← \{j \mid θ_j^{[m+0.5]} = 0\}
w_{[m+1,1]} ← θ_{[m+0.5]}
For k ← 1, ..., g
    If $|∇_{j_k} L(w_{[m+1,k-1]})| \geq (1 + δ)λ$
      $\lambda_{j_k}(w_{[m+1,k-1]} \leftarrow T_{λ,j_k}(w_{[m+1,k-1]})$
    Else
      $w_{[m+1,k]} \leftarrow w_{[m+1,k-1]}$
      $w_{[m+1,k]} \leftarrow w_{[m+1,k-1]}$
      $\theta_{[m+1]} \leftarrow w_{[m+1,g]}$
A_{m+1} ← \{j \mid θ_j^{[m+1]} \neq 0\}
A_{m+1} ← \{j \mid θ_j^{[m+1]} = 0\}
m ← m + 1
```

**Until**

```
max_{k ∈ A_{m+0.5}} |∇_k L(θ^{[m+0.5]})| ≤ (1 + δ)λ
```

**Output:**

```
θ ← θ[m]
```

where the first equality comes the fact $∇H_{λ_0}(0) = 0$ for the MCP and SCAD regularizers (We will elaborate more about this in §4). This indicates that 0 is a local optimum of (1.1). Accordingly, we set $θ^{[0]} = 0$. Then for $K = 1, 2, \ldots, N$, we solve (1.1) for $λ_K$ using $θ^{(K-1)}$ as initialization. We summarize the warm start initialization in Algorithm 7.

The warm start initialization starts with large regularization parameters to suppress the overselection of irrelevant coordinates $\{j \mid θ_j^* = 0\}$ (in conjunction with the IteActUpd algorithm). Therefore the solution sparsity ensures the restricted strong convexity throughout all iterations, making the algorithm behaves as if minimizing a strongly convex function. Though large regularization parameters may also yield zero values for many relevant coordinates $\{j \mid θ_j^* \neq 0\}$ and result in larger estimation errors, this can be compensated by the decreasing regularization sequence. Eventually, PICASSO will gradually recover the relevant coordinates, reduce the estimation error of each output solution, and attain a sparse output solution with good statistical properties.

**Remark 3.2.** Tibshirani et al. (2012) propose a sequential strong rule for coordinate preselection,
Algorithm 7: The warm start initialization solves (1.1) with respect to a decreasing sequence of regularization parameters \( \{\lambda_K\}_{K=0}^N \). The leading regularization parameter \( \lambda_0 \) is chosen as \( \lambda_0 = \|\nabla L(0)\|_\infty \), which yields an all zero output solution \( \hat{\theta}^{(0)} = 0 \). For \( K = 1, \ldots, N \), we solve (1.1) for \( \lambda_K \) using \( \hat{\theta}^{(K-1)} \) as an initial solution. \( \{\tau_K\}_{K=1}^N \) and \( \{\delta_K\}_{K=1}^N \) are two sequence of convergence parameters, where \( \tau_K \) and \( \delta_K \) correspond to the \( K \)-th outer loop iteration with the regularization parameter \( \lambda_K \).

\[
\text{Algorithm: } \{\hat{\theta}^{(K)}\}_{K=0}^N \leftarrow \text{PICASSO}(\{\lambda_K\}_{K=0}^N)
\]

Parameter: \( \eta, \varphi, \{\tau_K\}_{K=1}^N, \{\delta_K\}_{K=1}^N \)

Initialize: \( \lambda_0 \leftarrow \|\nabla L(0)\|_\infty, \hat{\theta}^{(0)} \leftarrow 0 \)

For \( K \leftarrow 1, 2, \ldots, N \)

\[
\lambda_K \leftarrow \eta \lambda_{K-1}
\]

\[\hat{\theta}^{(K)} \leftarrow \text{IteActUpd}(\lambda_K, \hat{\theta}^{(K-1)}, \delta_K, \tau_K, \varphi)\]

Output: \( \{\hat{\theta}^{(K)}\}_{K=0}^N \)

which initializes the active set for the \( K \)-th middle loop as
\[
\mathcal{A}_0 = \{ j \mid \theta_j^{[0]} = 0, |\nabla_j L(\theta^{[0]})| \geq 2\lambda_K - \lambda_{K-1} \} \cup \{ j \mid \theta_j^{[0]} \neq 0 \}, \tag{3.9}
\]
\[
\overline{\mathcal{A}}_0 = \{ j \mid \theta_j^{[0]} = 0, |\nabla_j L(\theta^{[0]})| < 2\lambda_K - \lambda_{K-1} \}. \tag{3.10}
\]

Recall that \( \lambda_K = \eta \lambda_{K-1} \). Then we have
\[
2\lambda_K - \lambda_{K-1} = \left(1 - \frac{1 - \eta}{\eta}\right) \lambda_K.
\]

Thus (3.9) and (3.10) are essentially our proposed strong rule for PICASSO with \( \varphi = \frac{1 - \eta}{\eta} \).

4 Computational and Statistical Theory

We develop a new theory to analyze the pathwise coordinate optimization framework, and establish the computational and statistical properties of PICASSO for sparse linear regression. Recall that in (2.3), we rewrite the nonconvex regularizer as \( \mathcal{R}_\lambda(\theta) = \lambda \|\theta\|_1 + \mathcal{H}_\lambda(\theta) \), where \( \mathcal{H}_\lambda(\theta) = \sum_{j=1}^d h_j(|\theta_j|) \) is a smooth, concave, and coordinate decomposable function. For notational simplicity, we define \( \tilde{L}_\lambda(\theta) = L(\theta) + \mathcal{H}_\lambda(\theta) \). Accordingly we can rewrite \( \mathcal{F}_\lambda(\theta) \) as
\[
\mathcal{F}_\lambda(\theta) = \mathcal{L}(\theta) + \mathcal{R}_\lambda(\theta) = \tilde{L}_\lambda(\theta) + \lambda \|\theta\|_1.
\]

Before we present the main results, we first introduce four assumptions.

4.1 Preliminaries

The first assumption requires \( \lambda_N \) to be sufficiently large.

Assumption 4.1. We require that the geometrically decreasing regularization sequence satisfies
\[
\lambda_N \geq 4\|\nabla L(\theta^*)\|_\infty = \frac{4}{n} \|X^\top \epsilon\|_\infty.
\]

Moreover, we require \( \eta \in [0.96, 1) \).
Assumption 4.1 has been extensively studied in existing literature on high dimensional statistical theories for sparse linear regression (Zhang and Huang, 2008; Bickel et al., 2009; Zhang, 2010a; Negahban et al., 2012). It guarantees that all regularization parameters are sufficiently large for PICASSO to eliminate irrelevant coordinates along the solution path.

**Remark 4.2.** Note that Assumption 4.1 is deterministic for any given \( \lambda_N \). Since \( \|X^T e\|_\infty \) is a random quantity, we need to verify that Assumption 4.1 holds with high probability when applying PICASSO to sparse linear regression.

The second assumption imposes several regularity conditions on the nonconvex regularizer.

**Assumption 4.3.** We require \( h(\cdot) \) and \( h'(\cdot) \) of the nonconvex regularizer to satisfy:

- (R.1) For any \( a > b \geq 0 \), there exists a constant \( \alpha \geq 0 \) such that \( -\alpha(a - b) \leq h_\lambda'(a) - h_\lambda'(b) \leq 0 \).
- (R.2) For some \( \gamma > 0 \) and \( \forall a \geq 0 \), we have \( h_\lambda'(a) \in [-\lambda, 0] \) if \( a \leq \lambda \gamma \), and \( h_\lambda'(a) = -\lambda \) otherwise.
- (R.3) \( h_\lambda(\cdot) \) and \( h_\lambda'(\cdot) \) pass through the origin, i.e., \( h_\lambda(0) = 0 \) and \( h_\lambda'(0) = 0 \).
- (R.4) For \( \forall a \geq 0 \), we have \( |h_\lambda'(a) - h_\lambda'(0)| \leq |\lambda_1 - \lambda_2| \).

For simplicity, we call \( \alpha \) in (R.1) the concavity parameter of \( h_\lambda(\cdot) \), since \( \alpha \) characterizes how concave it is. By examining \( h_\lambda(\cdot) \) defined in (2.4) and (2.5), we can verify that both SCAD and MCP regularizers satisfy Assumption 4.3. Particularly, we have \( \alpha = 1/(\gamma - 1) \) for SCAD and \( \alpha = 1/\gamma \) for MCP.

Before we present the third assumption, we define the largest and smallest \( s \) sparse eigenvalues of the Hessian matrix \( \nabla^2 \mathcal{L}(\theta) = \frac{1}{n} X^T X \) as follows.

**Definition 4.4.** Given an integer \( s \geq 1 \), we define

- Largest \( s \) sparse eigenvalue : \( \rho_+(s) = \sup_{\|v\|_0 \leq s} \frac{\nabla^2 \mathcal{L}(\theta) v}{\|v\|_2^2} \),
- Smallest \( s \) sparse eigenvalue : \( \rho_-(s) = \inf_{\|v\|_0 \leq s} \frac{\nabla^2 \mathcal{L}(\theta) v}{\|v\|_2^2} \).

The next lemma connects the largest and smallest \( s \) sparse eigenvalues to the restricted strong convexity and smoothness.

**Lemma 4.5.** Suppose that there exists an integer \( s \) such that \( 0 \leq \alpha < \rho_-(s) \leq \rho_+(s) < \infty \), where \( \alpha \) is the concavity parameter defined in (R.1) of Assumption 4.3. For any \( \theta, \theta' \in \mathbb{R}^d \) satisfying \( \|\theta - \theta'\|_0 \leq s \), we say that \( \mathcal{L}(\theta) \) is restricted strongly convex and smooth, i.e.,

\[
\frac{\rho_-(s)}{2} \|\theta' - \theta\|_2^2 \leq \mathcal{L}(\theta') - \mathcal{L}(\theta) - (\theta' - \theta)^\top \nabla \mathcal{L}(\theta) \leq \frac{\rho_+(s)}{2} \|\theta' - \theta\|_2^2.
\]

For notational simplicity, we define \( \tilde{\rho}_-(s) = \rho_-(s) - \alpha \). Then we say that \( \tilde{\mathcal{L}}_\lambda(\theta) \) is also restricted strongly convex and smooth, i.e.,

\[
\frac{\tilde{\rho}_-(s)}{2} \|\theta' - \theta\|_2^2 \leq \tilde{\mathcal{L}}_\lambda(\theta') - \tilde{\mathcal{L}}_\lambda(\theta) - (\theta' - \theta)^\top \nabla \tilde{\mathcal{L}}_\lambda(\theta) \leq \frac{\rho_+(s)}{2} \|\theta' - \theta\|_2^2.
\]

Moreover, we have \( \mathcal{F}_\lambda(\theta) \) satisfying the restricted strong convexity, i.e., for any \( \xi \in \partial \|\theta\|_1 \),

\[
\frac{\tilde{\rho}_-(s)}{2} \|\theta' - \theta\|_2^2 \leq \mathcal{F}_\lambda(\theta') - \mathcal{F}_\lambda(\theta) - (\theta' - \theta)^\top (\nabla \tilde{\mathcal{L}}_\lambda(\theta) + \lambda \xi).
\]
The proof of Lemma 4.5 is presented in Appendix A. Lemma 4.5 indicates the importance of the solution sparsity to PICASSO: When $\theta$ is sufficiently sparse, the restricted strong convexity of $\mathcal{L}(\theta)$ can dominate the concavity of $H_\lambda(\theta)$. Thus if an algorithm can guarantee the solution sparsity throughout all iterations, it will behave like minimizing a strongly convex optimization problem. Accordingly, linear convergence can be established. Now we introduce the third assumption on (1.1).

**Assumption 4.6.** Given $\|\theta^*\|_0 \leq s^*$, there exists an integer $\tilde{s}$ such that

$$\tilde{s} \geq (484\kappa^2 + 100\kappa)s^* + \rho_+(s^* + 2\tilde{s} + 1) > +\infty,$$

and $\bar{\rho}_-(s^* + 2\tilde{s} + 1) > 0$,

where $\kappa$ is defined as $\kappa = \frac{\rho_+(s^* + 2\tilde{s} + 1)}{\bar{\rho}_-(1)}$.

Assumption 4.6 guarantees that the optimization problem satisfies the restricted strong convexity as long as the number of active irrelevant coordinates never exceeds $\tilde{s}$ along the solution path. That is closely related to the restricted isometry property (RIP) and restricted eigenvalue (RE) conditions, which have been extensively studied in existing literature. See more details in Candes and Tao (2005); Bickel et al. (2009); Zhang (2010a); Raskutti et al. (2010); Lounici et al. (2011); Negahban et al. (2012).

At last, we introduce the assumption on the computational parameters of PICASSO.

**Assumption 4.7.** Recall that the convergence parameters $\delta_K$’s and $\tau_K$’s are defined in Algorithm 7, and the active set initialization parameter $\varphi$ is defined in (3.5). We require

$$\varphi \leq \frac{1}{8}, \quad \delta_K \leq \frac{1}{8}, \quad \text{and} \quad \tau_K \leq \frac{\delta_K}{\rho_+(s^* + 2\tilde{s})} \sqrt{\frac{\bar{\rho}_-(1)}{\rho_+(1)(s^* + 2\tilde{s})}}$$

for all $K = 1, ..., N$.

Assumption 4.7 guarantees that all middle and inner loops of PICASSO attain adequate precision such that their output solutions satisfy the desired computational and statistical properties.

### 4.2 Active Coordinate Minimization

First, we start with the convergence analysis for the inner loop of PICASSO. The following theorem presents the rate of convergence in terms of the objective value for Algorithm 2. For notational simplicity, we omit the outer loop index $K$, and denote $\lambda_K$ and $\tau_K$ by $\lambda$ and $\tau$.

**Theorem 4.8.** Suppose that Assumptions 4.3 and 4.6 hold. If the initial solution $\theta^{(0)}$ satisfies $\|\theta^{(0)}\|_0 = s \leq s^* + 2\tilde{s}$, then (3.2) is strongly convex with a unique global optimum $\hat{\theta}$. For $t = 1, 2, ...$, we have

$$F_\lambda(\theta^{(t)}) - F_\lambda(\hat{\theta}) \leq \left(\frac{s\rho_+^2(s)}{s\rho_+^2(s) + \bar{\rho}_-(s)\bar{\rho}_-(1)}\right)^t [F_\lambda(\theta^{(0)}) - F_\lambda(\hat{\theta})].$$

Moreover, we need at most

$$\log^{-1}\left(\frac{s\rho_+^2(s)}{s\rho_+^2(s) + \bar{\rho}_-(s)\bar{\rho}_-(1)}\right) \cdot \log\left(\frac{-\bar{\rho}_-(1)\tau^2\lambda^2}{2[F_\lambda(\theta^{(0)}) - F_\lambda(\hat{\theta})]}\right)$$

iterations such that $\|\theta^{(t+1)} - \theta^{(t)}\|_2 \leq \tau\lambda$, where $\tau$ is the convergence parameter defined in (3.3).

The detailed proof of Theorem 4.8 is presented in §8.1. Theorem 4.8 guarantees that given a sparse initial solution, Algorithm 2 essentially minimizes a strongly convex optimization problem,
though (1.1) can be globally nonconvex. Therefore it attains linear convergence to a unique global optimum in terms of the objective value. To the best of our knowledge, Theorem 4.8 is the first result establishing linear rates of convergence for cyclic coordinate optimization algorithms for minimizing nonsmooth strongly convex composite functions.

### 4.3 Active Set Updating

Then, we proceed with the convergence analysis for the middle loop of PICASSO. The following theorem presents the rate of convergence in terms of the objective value. For notational simplicity, we omit the outer loop index \( K \), and denote \( \lambda_K \) and \( \delta_K \) by \( \lambda \) and \( \delta \). Moreover, we define

\[
\Delta_\lambda = \frac{4\lambda^2 s^*}{\bar{\rho}(s^* + \bar{s})}, \quad S = \{ j \mid \theta_j^0 \neq 0 \}, \quad \bar{S} = \{ j \mid \theta_j^0 = 0 \}. \tag{4.1}
\]

**Theorem 4.9.** Suppose that Assumptions 4.1, 4.3, 4.6, and 4.7 hold. For any \( \lambda \geq \lambda_N \), if the initial solution \( \theta^{[0]} \) in Algorithms 4-6 satisfies \( \|\theta^{[0]}\|_0 \leq \bar{s} \) and \( F_\lambda(\theta^{[0]}) \leq F_\lambda(\theta^*) + \Delta_\lambda \), then regarding the active set initialized by the strong rule or simple rule, we have \( |A_0 \cap S| \leq \bar{s} \). Meanwhile, for \( m = 0, 1, 2, \ldots \), we also have \( \|\theta^{[m]}\|_0 \leq 2\bar{s}, \|\theta^{[m+0.5]}\|_0 \leq \bar{s} \), and

**Greedy Selection** :

\[
F_\lambda(\theta^{[m]}) - F_\lambda(\bar{\theta}^\lambda) \leq \left( 1 - \frac{\bar{\rho}(s^* + \bar{s})}{(s^* + \bar{s})\rho_+ (1)} \right)^m [F_\lambda(\theta^{[0]}) - F_\lambda(\bar{\theta}^\lambda)],
\]

**Randomized Selection** :

\[
\mathbb{E} F_\lambda(\theta^{[m]}) - F_\lambda(\bar{\theta}^\lambda) \leq \left( 1 - \frac{\bar{\rho}(s^* + \bar{s})}{(s^* + \bar{s})\rho_+ (1)} \right)^m [F_\lambda(\theta^{[0]}) - F_\lambda(\bar{\theta}^\lambda)],
\]

where \( \bar{\theta}^\lambda \) is a unique sparse local optimum of (1.1), and satisfies the KKT condition

\[
K_\lambda(\bar{\theta}^\lambda) = \min_{\xi \in \partial\|\bar{\theta}^\lambda\|_1} \|\nabla \tilde{\mathcal{L}}_\lambda(\bar{\theta}^\lambda) + \lambda \xi\|_\infty = 0 \quad \text{and} \quad \|\bar{\theta}^\lambda\|_0 \leq \bar{s}. \tag{4.2}
\]

Moreover, when we terminate the IteActUpd algorithm with \( \max_{k \in [m+0.5]} |\nabla_k \mathcal{L}(\theta^{[m+0.5]})| \leq (1 + \delta)\lambda \), where \( \delta \) is the convergence parameter defined in (3.7), the following results hold:

1. The output solution \( \bar{\theta}^\lambda \) satisfies the approximate KKT condition \( K_\lambda(\bar{\theta}^\lambda) \leq \delta \lambda \);
2. For the greedy selection, the number of active set updating iterations is at most

\[
\log^{-1} \left( 1 - \frac{\bar{\rho}(s^* + \bar{s})}{(s^* + \bar{s})\rho_+ (1)} \right) \cdot \log \left( \frac{\delta \lambda}{3\rho_+ (1) [F_\lambda(\theta^{[0]}) - F_\lambda(\theta^*)]} \right);
\]
3. For the randomized selection, given a constant \( \vartheta \in (0, 1) \), the number of active set updating iterations is at most

\[
\log^{-1} \left( 1 - \frac{\bar{\rho}(s^* + \bar{s})}{(s^* + \bar{s})\rho_+ (1)} \right) \cdot \log \left( \frac{\vartheta \delta \lambda}{3\rho_+ (1) [F_\lambda(\theta^{[0]}) - F_\lambda(\theta^*)]} \right)
\]

with probability at least \( 1 - \vartheta \).

\(^{4}\)Since the randomized selection rule randomly selects a coordinate in each active set updating iteration, the objective value of \( F_\lambda(\theta^{[m]}) \) is essentially a random variable. Therefore its rate of convergence is analyzed in terms of the expected objective value.
For the truncated cyclic selection, given $\delta \geq \sqrt{73/(484\kappa + 100)}$, the number of active set updating iterations is at most
\[
\frac{2\rho_+(1)[F_\lambda(\theta_0) - F_\lambda(\bar{\theta}_\lambda)]}{\delta^2\lambda^2}.
\]

The proof of Theorem 4.9 is presented in §8.2. Theorem 4.9 guarantees that when supplied a proper initial solution, the middle loop of PICASSO attains linear convergence to a unique sparse local optimum. Moreover, Theorem 4.9 has three important implications:

(I) The greedy and randomized selection rules are very conservative and only select one coordinate each time. The truncated cyclic selection rule only selects coordinates whose corresponding coordinate gradients are sufficiently large in magnitude. These mechanisms prevent the overselection of irrelevant coordinates and encourage the solution sparsity. In contrast, the cyclic selection used in Friedman et al. (2007); Mazumder et al. (2011) may overselect irrelevant coordinates and compromise the restricted strong convexity. An illustrative example is provided in Figure 3.

(II) Besides decreasing the objective value, the active coordinate minimization algorithm can remove some irrelevant coordinates from the active set. Then in conjunction with our proposed active set selection rules, the solution sparsity is guaranteed throughout all iterations. An illustrative example is provided in Figure 4. To the best of our knowledge, such a “forward-backward” phenomenon has not been discovered and rigorously characterized in existing literature.

(III) The strong rule for PICASSO moves inactive coordinates to the active set when their
Figure 4: An illustration of the active set updating algorithm. The active set updating iteration first removes some active coordinates from the active set, then add some inactive coordinates into the active set. Therefore the solution sparsity is guaranteed throughout all iterations. To the best of our knowledge, such a “forward-backward” phenomenon has not been discovered and rigorously characterized in existing literature. The corresponding coordinate gradients are sufficiently large in magnitude. Thus it can also prevent the overselection of irrelevant coordinates and guarantee the restricted strong convexity.

4.4 Warm Start Initialization

Next, we proceed with the convergence analysis for the outer loop of PICASSO. As has been shown in Theorem 4.9, each middle loop of PICASSO requires a proper initialization. Since $\theta^*$ and $\mathcal{S}$ are unknown in practice, it is difficult to manually pick such an initial solution. The next theorem shows that the warm start initialization can guide PICASSO to attain such a proper initialization for every middle loop without any prior knowledge on $\theta^*$ and $\mathcal{S}$.

**Theorem 4.10.** Recall that $\mathcal{K}_{\lambda_K}(\theta)$ is defined in (4.2). Suppose that Assumptions 4.1, 4.3, 4.6, and 4.7 hold. If $\theta$ satisfies $\|\theta_\mathcal{S}\|_0 \leq \bar{s}$ and $\mathcal{K}_{\lambda_{K-1}}(\theta) \leq \delta_{K-1} \lambda_{K-1}$, then we have

$$
\|\hat{\Delta}\|_1 \leq 11\|\hat{\Delta}_\mathcal{S}\|_1 \leq 11\sqrt{\bar{s}}\|\hat{\Delta}\|_2, \quad \mathcal{K}_{\lambda_K}(\theta) \leq \lambda_K/4, \quad \text{and} \quad F_{\lambda_K}(\theta) \leq F_{\lambda_K}(\theta^*) + \Delta_{\lambda_K}.
$$

The proof of Theorem 4.10 is presented in Appendix D.1. The warm start initialization starts with an all zero local optimum and a sufficiently large regularization parameter $\lambda_0$, which naturally satisfy all requirements

$$
\|\theta_\mathcal{S}\|_0 \leq \bar{s} \quad \text{and} \quad \mathcal{K}_{\lambda_0}(\theta) = 0.
$$

Thus $\theta^{[0]} = 0$ is a proper initial solution for $\lambda_1$. Then combining Theorem 4.10 with Theorem 4.9, we can show by induction that the output solution of each middle loop is always a proper initial solution for the next middle loop.

**Remark 4.11.** From a geometric perspective, the warm start initialization yields a sequence
of nested sparse basins of attraction $C_{\lambda_0} \supseteq C_{\lambda_1} \supseteq \ldots \supseteq C_{\lambda_N}$ such that $\hat{\mathbf{\theta}}^{(K-1)} \in C_{\lambda_{K-1}} \cap C_{\lambda_K}$. Therefore PICASSO iterates along these basins, and eventually converges to a “good” local optimum in $C_{\lambda_N}$ (close to $\mathbf{\theta}^*$). An illustration of the warm start initialization is presented in Figure 5.

![Diagram of basin of attraction](image)

Figure 5: An illustration of the warm start initialization (the outer loop). We start with large regularization parameters. This suppresses the overselection of irrelevant coordinates $\{j \mid \theta_j^* = 0\}$ and yields highly sparse solutions. With the decrease of the regularization parameter, the PICASSO algorithm gradually recovers the relevant coordinates, and eventually obtains a sparse output solution with good statistical properties.

Combining Theorems 4.8, 4.9, and 4.10, we establish the global convergence in terms of the objective value for PICASSO in the next theorem.

**Theorem 4.12.** Suppose that Assumptions 4.1, 4.3, 4.6, and 4.7 hold. Recall that $\alpha$ is the concavity parameter defined in (R.1) of Assumption 4.3, $\delta_K$'s and $\tau_K$'s are the convergence parameters defined in Algorithm 7, and $\kappa$ and $\tilde{s}$ are defined in Assumption 4.6. For $K = 1, \ldots, N$, we have:

1. At the $K$-th iteration of the outer loop, the number of exact coordinate minimization iterations within each inner loop is at most
   \[
   T_{\max}(\tau_K) = \log^{-1}\left( \frac{\widetilde{\rho}_-(s^* + 2\tilde{s}) \widetilde{\rho}_-(1) + (s^* + 2\tilde{s}) \rho^2_+(s^* + 2\tilde{s})}{(s^* + 2\tilde{s}) \rho^2_+(s^* + 2\tilde{s})} \right) \cdot \log\left( \frac{\widetilde{\rho}_-(1) \tau^2_K \widetilde{\rho}_-(s^* + 2\tilde{s})}{50s^*} \right);
   \]
2. At the $K$-th iteration of the outer loop, the number of active set updating iterations based on the greedy selection rule is at most
   \[
   M_{\max}^{\text{Grd}}(\delta_K) = \log^{-1}\left( 1 - \frac{\widetilde{\rho}_-(s^* + 2\tilde{s})}{(s^* + 2\tilde{s}) \rho^2_+(1)} \right) \cdot \log\left( \frac{\delta^2_K \widetilde{\rho}_-(s^* + 2\tilde{s})}{75s^* \rho^2_+(1)} \right);
   \]
3. At the $K$-th iteration of the outer loop, given a constant $\vartheta \in (0, 1)$, the number of active set
updating iterations based on the randomized selection rule is at most
\[ M^\text{Rand}_{\max}(\delta_K) = \log^{-1} \left( 1 - \frac{\hat{\rho}_-(s^* + 2\bar{s})}{(s^* + 2\bar{s})\rho_+(1)} \right) \cdot \log \left( \frac{\theta_2 \hat{\rho}_-(s^* + \bar{s})}{75 s^* \rho_+(1)} \right) \]
with probability at least 1 \(- \vartheta;\)

(4) At the \(K\)-th iteration of the outer loop, given \(\delta_K \geq \sqrt{73/(484\kappa + 100)}\), the number of active set updating iterations based on the truncated cyclic selection rule is at most
\[ M^\text{Cyc}_{\max}(\delta_K) = \frac{50 \rho_+(1)s^*}{\hat{\rho}_-(s^* + 2\bar{s})\delta^2_K}; \]

(5) To compute the entire solution path, the total number of exact coordinate minimization iterations in Algorithm 2 is at most \(\sum_{K=1}^{N} T_{\max}(\tau_K) \cdot M^\text{Rand}_{\max}(\delta_K)\);

(6) To compute the entire solution path using the randomized selection rule, the total number of exact coordinate minimization iterations in Algorithm 2 is at most \(\sum_{K=1}^{N} T_{\max}(\tau_K) \cdot M^\text{Rand}_{\max}(\delta_K)\) with probability at least 1 \(- N\vartheta;\)

(7) To compute the entire solution path using the truncated cyclic selection rule, the total number of exact coordinate minimization iterations in Algorithm 2 is at most \(\sum_{K=1}^{N} T_{\max}(\tau_K) \cdot M^\text{Cyc}_{\max}(\delta_K)\);

(8) At the \(K\)-th iteration of the outer loop, we have
\[ F_{\lambda_N}(\hat{\theta}^{(K)}) - F_{\lambda_N}(\hat{\theta}^{\lambda_N}) \leq [\mathbb{1}(K < N) + \mathbb{1}(K = N) \cdot \delta_N] \frac{50\lambda_2^2 s^*}{\hat{\rho}_-(s^* + \bar{s})}. \]

The proof of Theorem 4.12 is presented in §8.3. Theorem 4.12 guarantees that PICASSO attains global linear convergence to a unique sparse local optimum, which is a significant improvement over sublinear convergence of the randomized coordinate minimization algorithms established in existing literature. To the best of our knowledge, this is the first result establishing the convergence properties of the pathwise coordinate optimization framework in high dimensions.

**Remark 4.13.** We provide a simple illustration of Theorem 4.12. When we use PICASSO based on the greedy selection rule to obtain \(\{\hat{\theta}^{(K)}\}_{K=0}^{N}\), the computational complexity of the strong rule and each active set selection step is \(O(d)\), and the computational complexity of calculating each cyclic coordinate minimization iteration is \(O(s^* + 2\bar{s})\). Given \(\bar{s} = O(s^*)\), \((s^*)^2 \log d \ll d\), \(1/\delta_K \leq d\), and \(1/\tau_K \leq d\), the overall computational complexity is \(O(Nd \log d)\).

In contrast, when we use the randomized coordinate minimization algorithm to obtain \(\{\hat{\theta}^{(K)}\}_{K=0}^{N}\), the computational complexity of calculating each randomized coordinate minimization iteration is \(O(1)\), and the iteration complexity is \(O(d/\epsilon)\), where \(\epsilon\) is the gap toward the optimal objective value. Then given \(1/\epsilon \leq d\), the overall computational complexity is \(O(N d^2)\), which is much worse than that of PICASSO.

### 4.5 Statistical Theory

Finally, we analyze the statistical properties of the estimator by PICASSO for sparse linear regression. For compactness, we only consider the greedy selection rule and MCP regularizer, but the extensions
to the SCAD regularization and other selection rules are straightforward. We assume that \( \| \theta^* \|_0 \leq s^* \) and the design matrix \( X \) satisfies

\[
\frac{\| Xv \|_2^2}{n} \geq \psi_{\min}^2 \| v \|_2^2 - \gamma_{\min} \log \frac{d}{n} \| v \|_1^2 \quad \text{and} \quad \frac{\| Xv \|_2^2}{n} \leq \psi_{\max}^2 \| v \|_2^2 + \gamma_{\max} \log \frac{d}{n} \| v \|_1^2, \tag{4.3}
\]

where \( \gamma_{\min}, \gamma_{\max}, \psi_{\min}, \) and \( \psi_{\max} \) are positive constants, and do not scale with \((s^*, n, d)\). Existing literature has shown that (4.3) is satisfied by many common examples of sub-Gaussian random design with high probability (Raskutti et al., 2010; Negahban et al., 2012).

We then verify Assumptions 4.1 and 4.6 by the following lemma.

**Lemma 4.14.** Given \( \lambda_N = 8\sigma \sqrt{\log d/n} \), we have

\[
\mathbb{P} \left( \lambda_N \geq 4\| \nabla \mathcal{L}(\theta^*) \|_{\infty} = 4\| X^T e \|_{\infty} \right) \geq 1 - 2d^{-2}.
\]

Moreover, given \( \alpha = \psi_{\min}/4 \), there exists a universal positive constant \( C_1 \) such that for large enough \( n \), we have

\[
\tilde{s} = C_1 s^* \geq \left[ 484\kappa^2 + 100\kappa \right] \cdot s^* , \quad \tilde{\rho}_-(s^* + 2\tilde{s} + 1) \geq \psi_{\min}/2 , \quad \text{and} \quad \rho_+(s^* + 2\tilde{s} + 1) \leq 5\psi_{\max}/4.
\]

The proof of Lemma 4.14 is presented in Appendix E.1. Lemma 4.14 guarantees that the regularization sequence satisfies Assumption 4.1 with high probability, and Assumption 4.6 holds when the design matrix satisfies (4.3). Thus Theorem 4.12 guarantees that PICASSO attains linear convergence to a unique sparse local optimum in terms of the objective value with high probability for sparse linear regression. That enables us to characterize the statistical rate of convergence for the obtained estimator in the following theorem.

**Theorem 4.15 (Parameter Estimation).** Given \( \alpha = \gamma^{-1} = \psi_{\min}/4 \) and \( \lambda_N = 8\sigma \sqrt{\log d/n} \), for small enough \( \delta_N \) and large enough \( n \) such that \( n \geq C_2 s^* \log d \), where \( C_2 \) is a generic constant, we have

\[
\| \hat{\theta}^{(N)} - \theta^* \|_2 = O_p \left( \sqrt{\frac{s_1^*}{n}} + \sqrt{\frac{s_2^* \log d}{n}} \right),
\]

where \( s_1^* = |\{ j \mid |\theta_j^*| \geq \gamma \lambda_N \}| \) and \( s_2^* = |\{ j \mid 0 < |\theta_j^*| < \gamma \lambda_N \}| \).

The proof of Theorem 4.15 is presented in §8.4. Theorem 4.15 divides the estimation error into two terms: \( V_1 \) for strong signals and \( V_2 \) for weak signals. We then proceed to show that the refined statistical rate of convergence in Theorem 4.15 is minimax optimal. We consider a class of sparse vectors defined as follows:

\[
\Theta_\sigma(s_1^*, s_2^*, d) = \left\{ \theta^* \in \mathbb{R}^d \mid \sum_{j=1}^d \mathbbm{1} \left( \theta_j^* \geq \frac{C \sigma}{\sqrt{s_1^* + s_2^*}} \right) \leq s_1^* , \quad \sum_{j=1}^d \mathbbm{1} \left( 0 < |\theta_j^*| < \frac{C \sigma}{\sqrt{s_1^* + s_2^*}} \right) \leq s_2^* \right\},
\]

where \( C = 32\psi_{\min}^{-1} C_2^{-1/2} \) is a constant. Given \( s^* = s_1^* + s_2^* \) and \( n \geq C_2 s^* \log d \), we have

\[
\frac{C \sigma}{\sqrt{s_1^* + s_2^*}} = \frac{32\sigma}{\psi_{\min} \sqrt{C_2(s_1^* + s_2^*)}} \geq \frac{32\sigma}{\psi_{\min} \sqrt{\log d/n}} = \gamma \lambda_N,
\]

which matches the threshold for dividing signals in Theorem 4.15. The next theorem establishes a minimax lower bound for parameter estimation.
**Theorem 4.16 (Minimax Lower Bound).** Let \( \hat{\theta} \) denote any estimator of \( \theta^* \) based on \( y \sim N(X\theta^*, \sigma^2 I) \), where \( \theta^* \in \Theta_{\sigma}(s^*_1, s^*_2, d) \). Then there exists a universal constant \( C_4 \) such that

\[
\inf_{\hat{\theta}} \sup_{\theta \in \Theta_{\sigma}(s^*_1, s^*_2, d)} \mathbb{E} \| \hat{\theta} - \theta^* \|_2 \geq C_4 \left( \sigma \sqrt{\frac{s^*_1}{n}} + \sigma \sqrt{\frac{s^*_2 \log d}{n}} \right).
\]

The proof of Theorem 4.16 is provided in Appendix 8.5. Theorem 4.16 guarantees that the estimator obtained by PICASSO attains the minimax optimal rates of convergence over \( \Theta_{\sigma}(s^*_1, s^*_2, d) \). The convex \( \ell_1 \) regularizer only attains a suboptimal statistical rate of convergence, \( O_P(\sigma \sqrt{(s^*_1 + s^*_2) \log d/n}) \). See more details in Zhang and Huang (2008); Bickel et al. (2009); Zhang (2009).

We then proceed to analyze the support recovery performance of the obtained estimator. For convenience, we define the oracle least square estimator \( \hat{\theta}_o \) as

\[
\hat{\theta}_o S := \arg\min_{\theta S} \frac{1}{2n} \| y - X \theta \|_2^2 \quad \text{and} \quad \theta \overline{S} = 0,
\]

where \( S \) and \( \overline{S} \) are defined in (4.1). The following theorem shows that \( \hat{\theta}^{\lambda_N} \) is identical to the oracle least square estimator \( \hat{\theta}_o \) with high probability.

**Theorem 4.17 (Support Recovery).** Suppose that there exists a universal constant \( C_5 \) such that

\[
\min_{j \in S} |\theta^*_j| \geq \frac{C_5 \sigma \sqrt{\log d/n}}{\psi_{\min}}.
\]

Given \( \alpha = \psi_{\min}/4 \) and \( \lambda_N = 8\sigma \sqrt{\log d/n} \), for large enough \( n \) such that \( n \geq C_2 s^* \log d \), where \( C_2 \) is a generic constant, we have

\[
P(\hat{\theta}^{\lambda_N} = \hat{\theta}_o) \geq 1 - 4d^{-2}.
\]

Moreover, there exits a universal constant \( C_3 \) such that we have

\[
\| \hat{\theta}^{(N)} - \theta^* \|_2 \leq C_3 \sigma \sqrt{\frac{s^*}{n}} \quad \text{and} \quad \text{supp}(\hat{\theta}^{(N)}) = \text{supp}(\theta^*)
\]

with at least probability \( 1 - 4d^{-2} \).

The proof of Theorem 4.17 is presented in §8.6. Theorem 4.17 implies that when using the MCP regularizer, PICASSO converges to \( \hat{\theta}_o \) with high probability, which is often referred to the oracle property in existing literature (Fan and Li, 2001). Moreover, we guarantee that the obtained estimator \( \hat{\theta}^{(N)} \) also attains the fast statistical rate of convergence in parameter estimation, and correctly identifies the true support with high probability. In contrast, the convex \( \ell_1 \) regularization requires a much stronger irrepresentable condition to correctly identify all relevant coordinates, and the obtained estimator suffers estimation bias. See more details in Zhao and Yu (2006); Zou (2006); Meinshausen and Bühlmann (2006); Wainwright (2009).

## 5 Extension to Generalized Linear Model Estimation

PICASSO can be further extended to other regularized M-estimation problems. Taking sparse logistic regression as an example, we denote the binary response vector by \( y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n \),
and the design matrix by $X \in \mathbb{R}^{n \times d}$. We consider a logistic model $y_i \sim \text{Bernoulli}(\pi_i(\theta^*))$, where $\pi_i(\theta)$ is defined as

$$\pi_i(\theta) = \frac{\exp(X_{i\theta}^T \theta)}{1 + \exp(X_{i\theta}^T \theta)} \text{ for } i = 1, ..., n.$$  

(5.1)

When $\theta^*$ is sparse, we consider the minimization problem

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) + \mathcal{R}_\lambda(\theta), \text{ where } \mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left[ \log \left(1 + \exp(X_{i\theta}^T \theta)\right) - y_i X_{i\theta}^T \theta \right].$$

(5.2)

For notational simplicity, we denote the logistic loss function in (5.2) as $\mathcal{F}(\theta, \theta^*) = \frac{1}{2} \sum_{i=1}^{n} (\pi_i(\theta) - y_i)^2 X_{i\theta}^T X_{i\theta}$, and the design matrix by $X$. We then take $\theta^* = \arg\min_{\theta \in \mathbb{R}^d} \mathcal{F}(\theta, \theta^*) + \lambda \|\theta\|_1$. The logistic loss function is twice differentiable with

$$\nabla \mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} [\pi_i(\theta) - y_i] X_{i\theta} \quad \text{ and } \quad \nabla^2 \mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} [1 - \pi_i(\theta)] \pi_i(\theta) X_{i\theta} X_{i\theta}^T = \frac{1}{n} X^T P X,$$

where $P = \text{diag}([1 - \pi_1(\theta)]\pi_1(\theta), ..., [1 - \pi_n(\theta)]\pi_n(\theta)) \in \mathbb{R}^{n \times n}$. Similar to sparse linear regression, we also assume that the design matrix $X$ satisfies the column normalization condition $\|X_{\cdot j}\|_2 = \sqrt{n}$ for all $j = 1, ..., d$.

### 5.1 Proximal Coordinate Gradient Descent

For sparse logistic regression, directly taking the minimum with respect to a selected coordinate does not admit a closed form solution, and therefore may involve some sophisticated optimization algorithms such as the root-finding method (Boyd and Vandenberghe, 2009).

To address this issue, Shalev-Shwartz and Tewari (2011); Richtárik and Takáč (2012) suggest a more convenient approach, which takes a proximal coordinate gradient descent iteration. Taking the classical coordinate descent algorithm as an example, we randomly select a coordinate $j$ at the $t$-th iteration and consider a quadratic approximation of $\mathcal{F}(\theta_j; \theta^{(t)})$,

$$\mathcal{Q}_{\lambda, j, L}(\theta_j; \theta^{(t)}) = \mathcal{V}_{\lambda, j, L}(\theta_j; \theta^{(t)}) + \lambda \theta_j + \lambda \|\theta_j^{(t)}\|_1,$$

where $L > 0$ is a step size parameter, and $\mathcal{V}_{\lambda, j, L}(\theta_j; \theta^{(t)})$ is defined as

$$\mathcal{V}_{\lambda, j, L}(\theta_j; \theta^{(t)}) = \tilde{\mathcal{L}}(\theta^{(t)}) + (\theta_j - \theta_j^{(t)}) \nabla \tilde{\mathcal{L}}(\theta^{(t)}) + \frac{L}{2} (\theta_j - \theta_j^{(t)})^2.$$

The step size parameter $L$ usually satisfies $L \geq \max_j \nabla^2 \mathcal{L}(\theta) \text{ such that } \mathcal{Q}_{\lambda, j, L}(\theta_j; \theta^{(t)}) \geq \mathcal{F}(\theta_j, \theta^{(t)})$ for all $j = 1, ..., d$. We then take

$$\theta_j^{(t+1)} = \arg\min_{\theta_j} \mathcal{Q}_{\lambda, j, L}(\theta_j; \theta^{(t)}) = \arg\min_{\theta_j} \mathcal{V}_{\lambda, j, L}(\theta_j; \theta^{(t)}) + \lambda |\theta_j|.$$  

(5.3)

Different from the exact coordinate minimization, (5.3) always has a closed form solution obtained by soft thresholding. Particularly, we define $\bar{\theta}_j^{(t)} = \frac{1}{2} (\theta_j - \bar{\theta}_j^{(t)})^2 + \frac{\lambda}{L} |\theta_j| = S_{\lambda/L}(\bar{\theta}_j^{(t)})$ and $\theta_j^{(t+1)} = \theta_j^{(t)}$. For notational convenience, we write $\theta_j^{(t+1)} = T_{\lambda, j, L}(\theta^{(t)})$. When applying PICASSO to solve sparse
logistic regression, we only need to replace $T_{\lambda,j} (\cdot)$ with $T_{\lambda,j,L} (\cdot)$ in Algorithms 2-7.

**Remark 5.1 (Step Size Parameter).** For sparse logistic regression, we have $\nabla^2_{jj} L(\theta) = \frac{1}{n} X^T_{*j} P X_{*j}$.

Since $P$ is a diagonal matrix, and $\pi_i (\theta) \in (0,1)$ for any $\theta \in \mathbb{R}^d$, we have $\|P\|_2 = \max_i P_{ii} \in (0,1/4]$ for all $i = 1, ..., n$. Then we have $X^T_{*j} P X_{*j} \leq \|P\|_2 \|X_{*j}\|_2^2 = n/4$, where the last equality comes from the column normalization condition of $X$. Thus we choose $L = 1/4$.

We then analyze the computational and statistical properties of the estimator obtained by PICASSO for sparse logistic regression. For compactness, we only consider the greedy selection rule and MCP regularizer, but the extensions to the SCAD regularization and other selection rules are straightforward.

### 5.2 Initialization by Convex Relaxation

We assume that $\|\theta^*\|_0 \leq s^*$, and for any $\|\theta - \theta^*\|_2 \leq R$, the design matrix satisfies

$$v^T \nabla^2 L(\theta) v \geq \psi_{\min} \|v\|_2^2 - \gamma_{\min} \sqrt{\frac{\log d}{n}} \|v\|_1 \|v\|_2, \quad (5.4)$$

$$v^T \nabla^2 L(\theta) v \leq \psi_{\max} \|v\|_2^2 + \gamma_{\max} \sqrt{\frac{\log d}{n}} \|v\|_1 \|v\|_2, \quad (5.5)$$

where $\gamma_{\min}, \gamma_{\max}, \psi_{\min}, \psi_{\max}$, and $R$ are positive constants, and do not scale with $(s^*, n, d)$. Existing literature has shown that many common examples of sub-Gaussian random design satisfy (5.4) and (5.5) with high probability (Raskutti et al., 2010; Negahban et al., 2012; Loh and Wainwright, 2015).

Similar to sparse linear regression, we need to verify Assumptions 4.1 and 4.6 for sparse logistic regression by the following lemma.

**Lemma 5.2.** Recall that $\pi_i (\theta)$’s are defined in (5.1). Given $\lambda_N = 16 \sqrt{\log d/n}$, we have

$$\mathbb{P} \left( \lambda_N \geq 4 \|\nabla L(\theta^*)\|_\infty = \frac{4}{n} \|X^T w\|_\infty \right) \geq 1 - d^{-7},$$

where $w = (\pi_1 (\theta^*) - y_1, ..., \pi_n (\theta^*) - y_n)^T$. Moreover, given $\alpha = \psi_{\min}/4$ and $\|\theta - \theta^*\|_2 \leq R$, there exists a universal positive constant $C_1$ such that for large enough $n$, we have

$$\tilde{s} = C_1 s^* \geq [484 \kappa_2^2 + 100 \kappa] \cdot s^*, \quad \bar{\rho}_-(s^* + 2\tilde{s} + 1) \geq \psi_{\min}/2, \quad \text{and} \quad \rho_+(s^* + 2\tilde{s} + 1) \leq 5\psi_{\max}/4.$$

The proof of Lemma 5.2 directly follows Appendix E.1 and Loh and Wainwright (2015), and therefore is omitted. Lemma 5.2 guarantees that the regularization sequence satisfies Assumption 4.1 with high probability, and Assumption 4.6 holds when the design matrix satisfies (5.4) and (5.5).

Different from sparse linear regression, however, the restricted strong convexity and smoothness only hold over an $\ell_2$ ball centered at $\theta^*$ for sparse logistic regression. Thus directly choosing $\tilde{\theta}^{(0)} = 0$ may violate the restricted strong convexity. A simple counter example is $\|\theta^*\|_2 > R$, which results in $\|0 - \theta^*\|_2 > R$. To address this issue, we propose a new convex relaxation approach to obtain an initial solution for $\lambda_0$. Particularly, we solve the following convex relaxation of (1.1):

$$\min_{\theta \in \mathbb{R}^d} \bar{F}_{\lambda_0} (\theta), \quad \text{where} \quad \bar{F}_{\lambda_0} (\theta) = L(\theta) + \lambda_0 \|\theta\|_1 \quad (5.6)$$

26
up to an adequate precision. For example, we choose $\theta^{relax}$ satisfying the approximate KKT condition of (5.6) as follows,

$$
\min_{\xi \in \partial \|\theta^{relax}\|_1} \|\nabla L(\theta^{relax}) + \lambda_0 \xi\|_\infty \leq \delta_0 \lambda_0, \tag{5.7}
$$

where $\delta_0 \in (0, 1)$ is the initial precision parameter for $\lambda_0$. Since $\delta_0$ in (5.7) can be chosen as a sufficiently large value (e.g. $\delta_0 = 1/8$), calculating $\theta^{relax}$ becomes very efficient even for algorithms with only sublinear rates of convergence to global optima, e.g., classical coordinate minimization and proximal gradient algorithms. For notational convenience, we call the above initialization procedure the convex relaxation approach.

**Lemma 5.3.** Suppose that Assumptions 4.3 and 4.6 hold only for $\|\theta - \theta^*\|_2 \leq R$. Given $\rho_{-}(s^* + \tilde{s})R \geq 9\lambda_0 \sqrt{s^*} \geq 18\lambda_N \sqrt{s^*}$ and $\delta_0 = 1/8$, we have

$$
\|\theta_S^{relax}\|_0 \leq \tilde{s}, \quad \|\theta^{relax} - \theta^*\|_2 \leq R, \quad \text{and} \quad F_{\lambda_0}(\theta^{relax}) \leq F_{\lambda_0}(\theta^*) + \Delta_{\lambda_0}.
$$

The proof of Lemma 5.3 is provided in Appendix D.2. Lemma 5.3 guarantees that $\theta^{relax}$ is a proper initial solution for $\lambda_0$. Therefore all convergence results in Theorem 4.12 still hold, and PICASSO attains linear convergence to a unique sparse local optimum in terms of the objective value with high probability. An illustration of the initialization by the convex relaxation approach is provided in Figure 6.

![Figure 6: An illustration of the initialization by the convex relaxation.](image)

**Remark 5.4.** To guarantee the solution always satisfying the restricted strong convexity throughout
all iterations, Wang et al. (2014) solve (1.1) with an additional constraint
\[ \min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) + \mathcal{R}_\lambda(\theta) \quad \text{subject to} \quad \|\theta\|_2 \leq R/2. \] (5.8)

The $\ell_2$ norm constraint in (5.8) introduces an additional tuning parameter. Moreover, their statistical analysis assumes $\|\theta^*\|_2 \leq R/2$. This assumption is very strong, since $R$ is a constant and cannot scale with $(n, s^*, d)$. In contrast, our proposed convex relaxation approach avoids this assumption, and allows $\|\theta^*\|_2$ to be arbitrarily large.

The next theorem presents the statistical rate of convergence for the obtained estimator in parameter estimation.

**Theorem 5.5 (Parameter Estimation).** Given $\alpha = \psi_{\min}/4$ and $\lambda_N = 16\sqrt{\log d/n}$, for large enough $n$ and small enough $\delta_N$ we have

\[ \|\hat{\theta}^{(N)} - \theta^*\|_2 = O_P\left(\sqrt{\frac{s^*_1 \log s^*_1}{n}} + \sqrt{\frac{s^*_2 \log d}{n}}\right), \]

where $s^*_1 = |\{j \mid |\theta^*_j| \geq \gamma \lambda N\}|$ and $s^*_2 = |\{j \mid 0 < |\theta^*_j| < \gamma \lambda N\}|$.

The proof of Theorem 5.5 is presented in Appendix E.6. Theorem 5.5 implies that the MCP regularizer reduces the estimation bias for “strong” signals, since $\log s^*_1 \ll \log d$. Therefore it attains a faster statistical rate of convergence than that of the convex $\ell_1$ regularization, $O_P(\sqrt{s^* \log d/n})$.

See more details in Negahban et al. (2012).

### 6 Numerical Experiment

We evaluate the computational and statistical performance of PICASSO through numerical simulations. We compare the following seven candidates:

1. PICASSO using the greedy selection rule, denoted by “G-PICASSO”.
2. PICASSO using the randomized selection rule, denoted by “R-PICASSO”.
3. PICASSO using the truncated cyclic selection rule, denoted by “TC-PICASSO”.
4. SPARSENET (Mazumder et al., 2011).
5. Path-following Iterative Shrinkage Thresholding Algorithm (PISTA, Wang et al. (2014)).
6. Accelerated PISTA (A-PISTA, Zhao and Liu (2014)).
7. Multistage Convex Relaxation Method (MCVX) (Zhang, 2010b; Zhang et al., 2013).

All experiments are conducted on a PC with Intel Core i5 3.3 GHz and 16GB memory. All programs are coded in double precision C, called from a R wrapper. We optimize the computation since the R package For a fair comparison, we reimplement SPARSENET using C, instead of directly using the fortran-based SPARSENET package.
by exploiting the vector and matrix sparsity, which gains a significant speedup in vector and matrix manipulations (e.g. calculating the gradient and evaluating the objective value). We apply PICASSO to sparse linear regression, and choose MCP as the nonconvex regularizer.

**Simulated Data:** We generate each row of the design matrix $\mathbf{X}_i$ independently from a $d$-dimensional Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$, where $\Sigma_{kj} = 0.75$ and $\Sigma_{kk} = 1$ for all $j, k = 1, \ldots, d$ and $k \neq j$. We then normalize each column of the design matrix $\mathbf{X}^*_{ij}$ such that $\|\mathbf{X}^*_{ij}\|_2 = \sqrt{n}$. The response vector is generated from the linear model $y = \mathbf{X} \theta^* + \epsilon$, where $\theta^* \in \mathbb{R}^d$ is the regression coefficient vector, and $\epsilon$ is generated from a $n$-dimensional Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $\sigma^2 \mathbf{I}$. We set $n = 60$, $d = 1000$, $s^* = 3$, and $\sigma^2 = 1$. $\theta^*$ has three nonzero entries, which are $\theta^*_{250} = 3$, $\theta^*_{500} = 2$, and $\theta^*_{750} = 1$.5. We then set $\gamma = 1.05$, $N = 99$, $\lambda_N = 0.25\sigma\sqrt{\log d/n} \approx 0.0657$, $\varphi = 0.05$, $\delta_K = 10^{-3}$, and $\tau_K = 10^{-6}$ for all $1 \leq K \leq N$.

We present the numerical results averaged over multiple simulations. More specifically, we create a validation set using the same design matrix as the training set for regularization parameter selection. We then tune the regularization parameter over the selected regularization sequence. We denote the response vector of the validation set as $\tilde{y} \in \mathbb{R}^n$. Let $\hat{\theta}_\lambda$ denote the obtained estimator using the regularization parameter $\lambda$. We then choose the optimal regularization parameter $\hat{\lambda}$ by

$$
\hat{\lambda} = \arg\min_{\lambda \in \{\lambda_1, \ldots, \lambda_N\}} \|\tilde{y} - \mathbf{X}\hat{\theta}_\lambda\|^2_2. \tag{6.1}
$$

We repeat the simulation for 1000 times and summarize the averaged results in Table 6. We see that PICASSO and SPARSENET attain similar timing results, and both greatly outperform PISTA and MCVX. SPARSENET gets worse statistical performance than the others in both support recovery and parameter estimation. The G-PICASSO slightly outperforms R-PICASSO, TC-PICASSO, A-PISTA, PISTA, and MCVX.

<table>
<thead>
<tr>
<th>Method</th>
<th>$|\hat{\theta} - \theta^*|_2$</th>
<th>$|\hat{\theta}_S|_0$</th>
<th>$|\hat{\theta}_{Sc}|_0$</th>
<th>Correct Selection</th>
<th>Timing</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-PICASSO</td>
<td><strong>0.8003</strong> (0.8908)</td>
<td><strong>2.81</strong> (0.50)</td>
<td><strong>0.84</strong> (2.07)</td>
<td>666/1000</td>
<td>0.0169 (0.0027)</td>
</tr>
<tr>
<td>R-PICASSO</td>
<td>0.8102 (0.9663)</td>
<td>2.79 (0.54)</td>
<td>0.90 (2.35)</td>
<td>653/1000</td>
<td>0.0186 (0.0034)</td>
</tr>
<tr>
<td>TC-PICASSO</td>
<td>0.8057 (0.8374)</td>
<td>2.80 (0.48)</td>
<td>0.89 (2.04)</td>
<td>645/1000</td>
<td>0.0167 (0.0024)</td>
</tr>
<tr>
<td>SPARSENET</td>
<td>1.1260 (1.2708)</td>
<td>2.67 (0.69)</td>
<td>1.68 (3.19)</td>
<td>514/1000</td>
<td>0.0171 (0.0025)</td>
</tr>
<tr>
<td>PISTA</td>
<td>0.8135 (0.8998)</td>
<td>2.80 (0.51)</td>
<td>0.88 (2.11)</td>
<td>664/1000</td>
<td>2.1771 (0.3805)</td>
</tr>
<tr>
<td>A-PISTA</td>
<td>0.8024 (0.9101)</td>
<td>2.79 (0.50)</td>
<td>0.87 (2.07)</td>
<td>663/1000</td>
<td>0.0177 (0.0027)</td>
</tr>
<tr>
<td>MCVX</td>
<td>0.8044 (0.8253)</td>
<td>2.80 (0.51)</td>
<td>0.88 (2.15)</td>
<td>665/1000</td>
<td>8.2474 (0.8550)</td>
</tr>
</tbody>
</table>

Table 1: Quantitative comparison on simulated data. We see that PICASSO and SPARSENET attain similar timing results, and both greatly outperform PISTA and MCVX in computational performance. SPARSENET achieves the worse statistical performance among all candidates in both support recovery and parameter estimation.

**Remark 6.1.** The heuristic cyclic selection rule in SPARSENET always iterates over the first
many irrelevant variables before getting to the relevant variable when identifying a new active set. Since these irrelevant variables are highly correlated with the relevant variables in our experiment, the heuristic cyclic selection may overselect some irrelevant variables and miss relevant variables. In contrast, the other six algorithms have mechanisms to prevent overselecting irrelevant variables when identifying active sets. This eventually makes them outperform SPARSENET.

**Real Data:** We adopt the gene expression data set in Scheetz et al. (2006). The original data set contains 31,042 gene expression values of 120 rats. Our goal is to identify genes with expression values related to that of gene TRIM32, which is known to be associated with human diseases of the retina. Following the same preprocessing procedure as Huang et al. (2008), we remove genes lacking sufficient variation or expression, and then choose 200 genes with the largest sample variances in expression values.

We set $\gamma = 1.05$, $N = 200$, $\lambda_N = 0.01\lambda_0$, $\delta_K = 10^{-3}$, and $\tau_K = 10^{-6}$ for all $1 \leq K \leq N$. We randomly split the 120 rats into a training set of 90 rats for fitting the model, a validation set of 15 rats for tuning parameter selection, and a testing set of 15 rats for evaluating the prediction performance. Similar to simulated data, the optimal tuning parameter is selected based on minimizing the prediction error on the validation set. Table 6 summarizes the numerical results averaged over 100 random splits. We see that PICASSO attains better prediction error and smaller average model sizes than those of the other competing algorithms. Moreover, PICASSO attains much better timing performance than that of PISTA and MCVX.

<table>
<thead>
<tr>
<th>Method</th>
<th>Average model size</th>
<th>Prediction Error</th>
<th>Timing</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-PICASSO</td>
<td>6.24(3.23)</td>
<td>0.1659(0.0671)</td>
<td>0.0481(0.0171)</td>
</tr>
<tr>
<td>R-PICASSO</td>
<td>6.54(3.34)</td>
<td>0.1728(0.0674)</td>
<td>0.0607(0.0267)</td>
</tr>
<tr>
<td>TC-PICASSO</td>
<td>6.61(3.60)</td>
<td>0.1744(0.0672)</td>
<td>0.0544(0.0208)</td>
</tr>
<tr>
<td>SPARSENET</td>
<td>7.12(4.64)</td>
<td>0.1847(0.0753)</td>
<td>0.0586(0.0297)</td>
</tr>
<tr>
<td>PISTA</td>
<td>6.24(3.19)</td>
<td>0.1664(0.0673)</td>
<td>3.7432(0.3515)</td>
</tr>
<tr>
<td>A-PISTA</td>
<td>6.24(3.19)</td>
<td>0.1660(0.0672)</td>
<td>0.0512(0.0272)</td>
</tr>
<tr>
<td>MCVX</td>
<td>6.53(3.44)</td>
<td>0.1711(0.0671)</td>
<td>9.9664(0.7788)</td>
</tr>
</tbody>
</table>

Table 2: Quantitative comparison on real data. We see that PICASSO attains better prediction error and smaller average model sizes than those of other competing algorithms. Moreover, PICASSO attains much better timing performance than that of PISTA and MCVX.

7 Discussions

Here we discuss several existing methods related to PICASSO, including the multistage convex relaxation method, one-step convex relaxation method, path-following iterative shrinkage threshold-
ing algorithm, accelerated path-following iterative shrinkage thresholding algorithm, and proximal gradient algorithm.

The multistage convex relaxation method is proposed in Zhang (2010b); Zhang et al. (2013); Shen et al. (2012). It solves a sequence of convex relaxation problems of (1.1). Zhang (2010b); Zhang et al. (2013) show that the obtained estimator enjoys similar statistical guarantees to those of PICASSO for sparse linear regression. However, there is no computational guarantee on its iteration complexity to a local optimum. Moreover, since each relaxed problem is still nonstrongly convex, the multistage convex relaxation method needs to be combined with some efficient computational algorithms such as PICASSO.

The one-step convex relaxation method is proposed in Zou and Li (2008); Wang et al. (2013); Fan et al. (2014). It is a special case of the multistage convex relaxation with only two iterations. Similar to the multistage convex relaxation method, it also needs an efficient computational algorithm to solve each relaxed problem. Moreover, in order to obtain the variable selection consistency, the one-step convex relaxation method requires a stronger minimum signal strength. Taking sparse linear regression as an example, Wang et al. (2013); Fan et al. (2014) requires a minimum signal strength of order of $\sigma \sqrt{s^* \log d/n}$, while PICASSO only requires a minimum signal strength of order of $\sigma \sqrt{\log d/n}$.

The path-following iterative shrinkage thresholding algorithm (PISTA) is proposed in Wang et al. (2014). It is essentially a proximal gradient algorithm combined with the warm start initialization. Although PISTA and PICASSO enjoy similar theoretical guarantees, PICASSO is computationally more efficient than PISTA in practice, as shown in §6. Besides, the implementation of PISTA requires subtle control over the step size, and may yield slow empirical convergence, as illustrated in §6. An accelerated PISTA algorithm (A-PISTA) is proposed in Zhao and Liu (2014), which uses coordinate minimization algorithms to accelerated PISTA. It shows very competitive computational performance to PICASSO in our numerical simulations, even though its computational rate of convergence is not as good as PISTA or PICASSO.

Alternatively, other researchers focus on solving (1.1) with an additional constraint,

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) + \mathcal{R}_\lambda(\theta) \quad \text{subject to } \|\theta\|_1 \leq M,$$

(7.1)

where $M > 0$ is an extra tuning parameter. Loh and Wainwright (2015) show that the proximal gradient algorithm geometrically converges to a ball centered at the global optimum to (7.1) with a radius approximately equal to the statistical error. However, the analysis of Loh and Wainwright (2015) does not exploit the advantage of nonconvex regularization and only provides a parameter estimation rate of convergence slower than that of PICASSO. In addition, their analysis for generalized linear model estimation requires a restrictive assumption similar to Wang et al. (2014): $\|\theta^*\|_2 \leq R/2$, where $R$ is a constant and does not scale with $(n, s^*, d)$. Nevertheless, PICASSO does not require this assumption.
8 Proof of Main Results

We present the proof sketch of our computational and statistical theories. To unify the convergence analysis of PICASSO using the exact coordinate minimization (2.6) and proximal coordinate gradient descent (5.3), we define two auxiliary parameters $\nu_+(1)$ and $\nu_-(1)$. More specifically, we choose $\nu_+(1) = \nu_-(1) = L$ for the proximal coordinate gradient descent, and $\nu_+(1) = \rho_+(1)$ and $\nu_-(1) = \rho_-(1)$ for the exact coordinate minimization.

8.1 Proof of Theorem 4.8

Proof. Since $\|	heta^{(0)}\|_0 = s \leq s^* + 2\bar{s}$, by Assumption 4.6 and Lemma 4.5, we know that (3.2) is a strongly convex optimization problem. Therefore its minimizer $\hat{\theta}$ is unique. Before we proceed with the proof, we first introduce the following lemmas.

Lemma 8.1. Suppose that Assumptions 4.3 and 4.6 hold, and $\|	heta^{(0)}\|_0 = s \leq s^* + 2\bar{s}$. For $t = 0, 1, 2, \ldots$,

$$F_{\lambda}(\theta^{(t)}) - F_{\lambda}(\theta^{(t+1)}) \geq \frac{\nu_-}{2} \frac{s}{\nu_-(1)} \frac{s}{\nu_-(1)}\|\theta^{(t)} - \theta^{(t+1)}\|_2^2.$$

Lemma 8.2. Suppose that Assumptions 4.3 and 4.6 hold, and $\|	heta^{(0)}\|_0 = s \leq s^* + 2\bar{s}$. For $t = 0, 1, 2, \ldots$,

$$F_{\lambda}(\theta^{(t+1)}) - F_{\lambda}(\bar{\theta}) \leq \frac{s}{\rho_- \rho_-} \|\theta^{(t+1)} - \theta^{(t)}\|_2^2.$$

The proof of Lemmas 8.1 and 8.2 is presented in Appendices B.1 and B.2. Lemmas 8.1 and 8.2 characterize the successive descent in each iteration and the gap towards the optimal objective value after each iteration respectively.

[Linear Convergence]: Combining Lemmas 8.1 and 8.2, we obtain

$$F_{\lambda}(\theta^{(t+1)}) - F_{\lambda}(\bar{\theta}) \leq \frac{s}{\rho_- \rho_-} \|\theta^{(t+1)} - \theta^{(t)}\|_2^2.$$

By simple manipulation, (8.1) implies

$$F_{\lambda}(\theta^{(t+1)}) - F_{\lambda}(\bar{\theta}) \leq \left( \frac{s}{\rho_- \rho_-} \|\theta^{(t+1)} - \theta^{(t)}\|_2^2 \right)^{t+1} F_{\lambda}(\theta^{(0)}) - F_{\lambda}(\bar{\theta}),$$

where (ii) comes from recursively using (i).

[Empirical Iteration Complexity]: Combining (8.2) with Lemma 8.1, we obtain

$$\|\theta^{(t)} - \theta^{(t+1)}\|_2^2 \leq \left( \frac{s}{\rho_- \rho_-} \|\theta^{(t)} - \theta^{(t+1)}\|_2^2 \right)^{t+1} F_{\lambda}(\theta^{(0)}) - F_{\lambda}(\bar{\theta}),$$

where the first inequality comes from $F_{\lambda}(\theta^{(t)}) \geq F_{\lambda}(\bar{\theta})$. Therefore we need at most

$$t = \log \left( \frac{\rho_- \rho_-}{s \rho_- \rho_-} \right) \log \left( \frac{\rho_- \rho_- \rho_- \rho_- \rho_- \rho_-}{s \rho_- \rho_-} \right).$$
iterations such that  \( \|\theta^{(t+1)} - \theta^{(t)}\|_2^2 \leq \left( \frac{s \rho^2_+(s)}{\rho_-(s) \nu_-(1) + s \rho^2_+(s)} \right)^t \frac{2[f_\lambda(\theta^{(0)}) - f_\lambda(\tilde{\theta})]}{\nu_-(1)} \leq \tau^2 \lambda^2. \)

8.2 Proof of Theorem 4.9

Proof. We first analyze the convergence properties of the middle loop using the greedy selection rule. Before we proceed with the proof, we first introduce the following lemmas.

Lemma 8.3. [Existence and Uniqueness of the Sparse Local Optimum] Suppose that Assumptions 4.1, 4.3, 4.6, and 4.7 hold. There exists a local optimum \( \tilde{\theta}^\lambda \) satisfying \( \|\tilde{\theta}^\lambda\|_0 \leq \tilde{s} \) and \( K_\lambda(\tilde{\theta}^\lambda) = 0. \)

Lemma 8.4. Suppose that Assumptions 4.1, 4.3, 4.6, and 4.7 hold. If the initial solution \( \theta^{(0)} \) in Algorithm 2 satisfies \( \|\theta^{(0)}\|_0 \leq 2\tilde{s} \) and \( f_\lambda(\theta^{(0)}) \leq f_\lambda(\theta^*) + \triangle_\lambda \), the output solution \( \hat{\theta} \) satisfies

\[
\min_{\xi_A \in \mathcal{G} \theta_A|_1} \|\nabla_\lambda \tilde{L}_\lambda(\hat{\theta}) + \lambda \xi_A\|_\infty \leq \delta_\lambda \quad \text{and} \quad \|\hat{\theta}\|_0 \leq \tilde{s}. \tag{8.3}
\]

Lemma 8.5 (Sparse Active Set Initialization). Suppose that Assumptions 4.1, 4.3, 4.6, and 4.7 hold. If the initial solution \( \theta^{[0]} \) satisfies \( \|\theta^{[0]}\|_0 \leq \tilde{s} \) and \( f_\lambda(\theta^{[0]}) \leq f_\lambda(\theta^*) + \triangle_\lambda \). Then we have \( |A_0 \cap \tilde{S}| \leq \tilde{s} \).

The proof of Lemmas 8.3, 8.4, and 8.5 is presented in Appendices C.1, C.2, and C.4. Lemma 8.3 verifies the existence of the sparse local optimum. Lemma 8.4 implies that the inner loop of PICASSO removes irrelevant coordinates, and encourages the output solution sparsity. Lemma 8.5 implies that the strong rule initializes a sufficiently sparse active set.

[Solution Sparsity]: Since the objective is monotone decreasing within each middle loop, we have

\[
f_\lambda(\theta^{[m+1]}) \leq f_\lambda(\theta^{[m+0.5]}) \leq f_\lambda(\theta^{[0]}) \leq f_\lambda(\theta^*) + \triangle_\lambda \quad \text{for all} \quad m = 0, 1, 2, \ldots. \tag{8.4}
\]

Since \( \theta^{[0]} \) satisfies \( \|\theta^{[0]}\|_0 \leq \tilde{s} \), if the active set \( A_0 \) is initialized by (3.4), then we have \( |A_0 \cap \tilde{S}| \leq \tilde{s} \). If \( A_0 \) is obtained by the strong rule, by Lemma 8.5, we also have \( |A_0 \cap \tilde{S}| \leq \tilde{s} \). Then by Lemma 8.4, we have \( \|\theta^{[0.5]}\|_0 \leq \tilde{s} \). Moreover, the greedy selection moves only one inactive coordinate to the active set, and therefore guarantees \( \|\theta^{[1]}\|_0 \leq \tilde{s} + 1 \). By induction, we prove \( \|\theta^{[m]}\|_0 \leq \tilde{s} + 1 \) and \( \|\theta^{[m+0.5]}\|_0 \leq \tilde{s} \) for all \( m = 0, 1, 2, \ldots \).

[Linear Convergence]: We first consider the proximal coordinate gradient descent. We need to construct an auxiliary solution \( w^{[m+1]} = \argmin_{w \in \mathbb{R}^d} J_{\lambda,L}(w; \theta^{[m+0.5]}) \), where

\[
J_{\lambda,L}(w; \theta^{[m+1]}) = \tilde{L}_\lambda(\theta^{[m+0.5]}) + (w - \theta^{[m+0.5]})^T \nabla \tilde{L}_\lambda(\theta^{[m+0.5]}) + \frac{L}{2} \|w - \theta^{[m+0.5]}\|_2^2 + \lambda \|w\|_1.
\]

We can verify that for \( j = 1, \ldots, d \), we have \( w_k^{[m+1]} = \arg\min_{\theta_k} Q_{\lambda,k,L}(\theta_k; \theta^{[m+0.5]}) \). For notational simplicity, we define \( w^{[m+1]} = T_{\lambda,L}(\theta^{[m+0.5]}) \). Before we proceed with the proof, we introduce the following lemmas.

Lemma 8.6. Suppose that Assumptions 4.1, 4.3, 4.6, and 4.7 hold. For the proximal coordinate gradient descent and \( m = 0, 1, 2, \ldots \), we have

\[
f_\lambda(\theta^{[m+0.5]}) - f_\lambda(\theta^{[m+1]}) \geq \frac{1}{s^* + 2\tilde{s}} \left[ f_\lambda(\theta^{[m+0.5]}) - J_{\lambda,L}(w^{[m+1]}; \theta^{[m+0.5]}) \right].
\]

"
Lemma 8.7. Suppose that Assumptions 4.1, 4.3, 4.6, and 4.7 hold. For the proximal coordinate gradient descent and \( m = 0, 1, 2, \ldots \), we have

\[
\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\lambda) \leq \frac{L}{\rho_-(s^* + 2\tilde{s})} \left[ \mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_{\lambda,L}(w^{[m+1]}; \theta^{[m+0.5]}) \right].
\]

The proof of Lemmas 8.6 and 8.7 is presented in Appendices C.5 and C.8. Lemmas 8.6 and 8.7 characterize the successive descent in each iteration and the gap towards the optimal objective value after each iteration respectively. Combining Lemmas 8.6 and 8.7, we obtain

\[
\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\hat{\theta}^\lambda) \leq \frac{(s^* + 2\tilde{s})L}{\rho_-(s^* + 2\tilde{s})} \left[ \mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\hat{\theta}^\lambda) \right] - \left( \mathcal{F}_\lambda(\theta^{[m+1]}) - \mathcal{F}_\lambda(\hat{\theta}^\lambda) \right), \tag{8.5}
\]

By simple manipulation, (8.5) implies

\[
\mathcal{F}_\lambda(\theta^{[m+1]}) - \mathcal{F}_\lambda(\hat{\theta}^\lambda) \leq \left( 1 - \frac{\tilde{\rho}_-(s^* + 2\tilde{s})}{(s^* + 2\tilde{s})L} \right) \left[ \mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\hat{\theta}^\lambda) \right] - \left( \mathcal{F}_\lambda(\theta^{[m+1]}) - \mathcal{F}_\lambda(\hat{\theta}^\lambda) \right) \tag{8.6}
\]

where (i) comes from (8.4), and (ii) comes from recursively applying (i).

For the exact coordinate minimization, at the \( m \)-th iteration, we only need to conduct a proximal coordinate gradient descent iteration with \( L = \rho_+(1) \), obtain an auxiliary solution \( \hat{\theta}^{[m+1]} \). Since \( \mathcal{F}_\lambda(\theta^{[m+1]}) \leq \mathcal{F}_\lambda(\hat{\theta}^{[m+1]}) \), by (8.6), we further have

\[
\mathcal{F}_\lambda(\theta^{[m+1]}) - \mathcal{F}_\lambda(\hat{\theta}^\lambda) \leq \left( 1 - \frac{\tilde{\rho}_-(s^* + 2\tilde{s})}{(s^* + 2\tilde{s})L} \right) \left[ \mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\hat{\theta}^\lambda) \right] \tag{8.7}
\]

[Empirical Iteration Complexity]: Before we proceed with the proof, we introduce the following lemma.

Lemma 8.8. Suppose that Assumptions 4.1 and 4.6 hold. For any \( \theta \), we conduct an exact coordinate minimization or proximal coordinate gradient descent iteration over a coordinate \( k \), and obtain \( w \). Then we have \( \mathcal{F}_\lambda(\theta) - \mathcal{F}_\lambda(w) \geq \frac{\nu_-(1)}{2}(w_k - \theta_k)^2 \). Moreover, when \( \theta_k = 0 \) and \( |\nabla_k \tilde{\mathcal{L}}_\lambda(\theta)| \geq (1 + \delta)\lambda \), we have

\[
|w_k| \geq \frac{\delta \lambda}{L} \quad \text{and} \quad \mathcal{F}_\lambda(\theta) - \mathcal{F}_\lambda(w) \geq \frac{\delta^2 \lambda^2}{2\nu_+(1)}.
\]

The proof of Lemma 8.8 is presented in Appendix B.3. Lemma 8.8 characterizes the sufficient descent when adding the selected inactive coordinate into the active set. Assume that the selected coordinate \( k_m \) satisfies \( |\nabla_{k_m} \tilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]})| \geq (1 + \delta)\lambda \). Then by Lemma 8.8, we have

\[
\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\hat{\theta}^\lambda) \geq \mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\theta^{[m+1]}) \geq \frac{\delta^2 \lambda^2}{2\nu_+(1)} \tag{8.8}
\]

Moreover, by (8.6) and (8.7), we need at most

\[
m = \log^{-1} \left( 1 - \frac{\rho_-(s^* + 2\tilde{s})}{(s^* + 2\tilde{s})\nu_+(1)} \right) \log \left( \frac{\delta^2 \lambda^2}{3\nu_+(1) \left[ \mathcal{F}_\lambda(\theta^{[0]}) - \mathcal{F}_\lambda(\hat{\theta}^\lambda) \right]} \right)
\]

iterations such that \( \mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\hat{\theta}^\lambda) \leq \frac{\delta^2 \lambda^2}{3\nu_+(1)} \), which is contradicted by (8.8). Therefore, we must have \( \max_{k \in \Omega_{\lambda_m}} |\nabla_k \tilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]})| \leq (1 + \delta)\lambda \), and the algorithm is terminated.

[Approximate Optimal Output Solution] By Lemma 8.4, we know that when every inner loop terminates, the approximate KKT condition must hold over the active set. Since \( \nabla_{\Omega_{\lambda_m}} H(\theta^{[m+0.5]}) = 0 \), the
stopping criterion \( \max_{k \in \mathcal{A}_m} |\nabla_k \tilde{L}_\lambda(\theta^{[m+0.5]})| \leq (1+\delta)\lambda \) implies that the approximate KKT condition holds over the inactive set,

\[
\min_{\xi \in \partial \lambda(\theta^{[m+0.5]})} \|\nabla_{\mathcal{A}_m} \tilde{L}_\lambda(\theta^{[m+0.5]}) + \lambda \xi \mathcal{A}_m \|_{\infty} \leq \delta \lambda.
\]

Combining the above two approximate KKT conditions implies that \( \theta^{[m+0.5]} \) satisfies the approximate KKT condition \( \mathcal{K}_\lambda(\theta^{[m+0.5]}) \leq \delta \lambda \). Moreover, the randomized selection rule can be analyzed in a similar manner. Due to space limitation, we present its proof in Appendix C.10.

We then analyze the convergence properties of the middle loop using the truncated cyclic selection rule. To guarantee the sparsity of the active set, we need to show that the truncated cyclic selection moves no more than \( \tilde{s} \) inactive coordinates to the active set in each active set updating step. We prove this by contradiction. Before we proceed with the proof, we first introduce the following lemmas.

**Lemma 8.9.** Suppose that Assumptions 4.1, 4.3, 4.6, and 4.7 hold. Given a solution \( \theta^{[m+0.5]} \), we assume that the truncated cyclic selection adds exactly \( \tilde{s} + 1 \) coordinates into the active set. Then we have

\[
\mathcal{F}_\lambda(\theta^{[m]}) \leq \mathcal{F}_\lambda(\theta^{[m+1]}) + \frac{36\lambda^2 s^*}{\bar{\rho}_-(s^* + 2\tilde{s} + 1)}.
\]

**Lemma 8.10.** Suppose that Assumptions 4.1, 4.3, 4.6, and 4.7 hold. Given a solution \( \theta^{[m+0.5]} \), we assume that the truncated cyclic selection adds exactly \( \tilde{s} + 1 \) coordinates into the active set. Then we have

\[
\mathcal{F}_\lambda(\theta^{[m+1]}) \leq \mathcal{F}_\lambda(\theta^{[m+0.5]}) \leq \mathcal{F}_\lambda(\theta^{[m]}) - \frac{(\tilde{s} + 1)\delta^2 \lambda^2}{2\nu_+(1)}.
\]

The proof of Lemmas 8.9 and 8.10 is presented in Appendices C.11 and C.12. Lemma 8.9 characterizes the maximum descent we can gain within each iteration of the middle loop. Lemma 8.10 characterizes the successive descent by adding \( \tilde{s} + 1 \) inactive coordinates into the active set.

**[Solution sparsity]** By simple manipulation of Lemmas 8.9 and 8.10, we have

\[
\tilde{s} \leq \frac{72\nu_+(1)s^*}{\delta^2 \bar{\rho}_-(s^* + 2\tilde{s}) + 1} \leq \frac{72\kappa}{\delta^2 \cdot s^*}. \tag{8.9}
\]

Thus if we have \( \delta \geq \sqrt{73}/(484\kappa + 100) \), then (8.9) implies \( \tilde{s} < (484\kappa^2 + 100\kappa)s^* \), which is contradicted by Assumption 4.6. Thus the truncated cyclic selection cannot add more than \( \tilde{s} \) inactive coordinates into the active set.

**[Empirical Iteration Complexity]** Since the algorithm guarantees the solution sparsity, i.e., \( \|\theta_{\infty}\|_0 \leq \tilde{s} \), by the restricted strong convexity of \( \mathcal{F}_\lambda(\theta) \), we have

\[
\mathcal{F}_\lambda(\theta) - \mathcal{F}_\lambda(\theta^*) \geq (\theta - \theta^*)^\top (\nabla \tilde{L}_\lambda(\theta^*) + \lambda \xi) + \frac{\rho_-(s^* + 2\tilde{s})}{2}\|\theta - \theta^*\|^2 \geq 0,
\]

where \( \xi \in \partial \|\theta^*\|_1 \) satisfies the optimality condition of (1.1), i.e., \( \nabla \tilde{L}_\lambda(\theta^*) + \lambda \xi = 0 \). Thus we know that \( \mathcal{F}_\lambda(\theta) \) is always lower bounded by \( \mathcal{F}_\lambda(\theta^*) \). Moreover, by Lemma 8.8, we have

\[
\mathcal{F}_\lambda(\theta^{[m]}) - \mathcal{F}_\lambda(\theta^{[m+1]}) \geq \frac{\delta^2 \lambda^2}{2\nu_+(1)}.
\]

35
Therefore, the number of the active set updating iterations is at most \( \frac{2\nu_+ (1) [\mathcal{F}_\lambda (\hat{\theta}^{(0)}) - \mathcal{F}_\lambda (\tilde{\theta}^\lambda)]}{\delta^2 \lambda^2} \).

\[ \square \]

### 8.3 Proof of Theorem 4.12

**Proof.** [Result (1)] Before we proceed with the proof, we introduce the following lemma.

**Lemma 8.11.** Suppose that Assumptions 4.1, 4.3, 4.6, and 4.7 hold. For any \( \lambda \geq \lambda_N \), if \( \theta \) satisfies \( \|\theta\|_0 \leq \tilde{s} \) and \( \mathcal{K}_\lambda (\theta) \leq \delta \lambda \), where \( \delta \leq 1/8 \), then for any \( \lambda' \in [\lambda_N, \lambda] \), we have

\[
\mathcal{F}_\lambda (\theta) - \mathcal{F}_\lambda (\tilde{\theta}^{\lambda'}) \leq \frac{40 (\mathcal{K}_\lambda (\theta) + 3 (\lambda - \lambda') (\lambda + \lambda') s^*)}{\bar{\rho}_- (s^* + \tilde{s})}.
\]

The proof of Lemmas 8.11 is presented in Appendix D.3. If we take \( \lambda = \lambda' = \lambda_N \) and \( \theta = \hat{\theta}^{(K-1)} \), then Lemma 8.11 implies

\[
\mathcal{F}_{\lambda_N} (\hat{\theta}^{(K-1)}) - \mathcal{F}_{\lambda_N} (\tilde{\theta}^{\lambda_N}) \leq \frac{25 s^* \lambda_N^2}{\bar{\rho}_- (s^* + \tilde{s})}.
\]  

(8.10)

Since the objective value always decreases in each middle loop, for any inner loop with \( \lambda_N \), we have

\[
\mathcal{F}_{\lambda_K} (\theta^{(0)}) - \mathcal{F}_{\lambda_K} (\tilde{\theta}) \leq \mathcal{F}_{\lambda_K} (\hat{\theta}^{(K-1)}) - \mathcal{F}_{\lambda_K} (\tilde{\theta}^{\lambda_K}).
\]

Therefore by Theorem 4.8 and (8.10), we know that the number of iterations within each inner loop is at most

\[
\log^- \left( \frac{\bar{\rho}_- (s^*) \nu_+ (1) + s^* \bar{\rho}_+^2 (s)}{s^* \bar{\rho}_+^2 (s)} \right) \log \left( \frac{\nu_+ (1) \tau^2 \bar{\rho}_- (s^* + \tilde{s})}{25 s^*} \right),
\]

which is Result (1) in Theorem 4.12.

[Results (2)–(4)] Combining Result (2) in Theorem 4.9 with (8.10), we know that the number of active set updating iterations within each middle loop is at most

\[
\log^- \left( 1 - \frac{\bar{\rho}_- (s^* + 2 \tilde{s})}{(s^* + 2 \tilde{s}) \nu_+ (1)} \right) \log \left( \frac{\delta K \bar{\rho}_- (s^* + \tilde{s})}{75 \nu_+ (1) s^*} \right),
\]

which is Result (2) in Theorem 4.12. Similarly, combining (8.10) with Results (3) and (4) in Theorem 4.9, we obtain Results (3) and (4) in Theorem 4.12 respectively.

[Result (5)–(7)] They are straightforward combinations of Results (1)–(4).

[Result (8)] For \( K < N \), we take \( \lambda' = \lambda_N \), \( \lambda = \lambda_K \), and \( \theta = \hat{\theta}^{(K)} \). Then by Lemma 8.11, we have

\[
\mathcal{F}_{\lambda_N} (\hat{\theta}^{(K)}) - \mathcal{F}_{\lambda_N} (\tilde{\theta}^{\lambda_N}) \leq \frac{25 (\lambda_K + \lambda_N)(\mathcal{K}_{\lambda_K} (\hat{\theta}^{(K)}) + 3 (\lambda_K - \lambda_N) s^*)}{\bar{\rho}_- (s^* + \tilde{s})}.
\]

(8.11)

Since we have \( \lambda_K > \lambda_N \) for \( K = 0, ..., N - 1 \), we directly obtain Result (8) using (8.11).

\[ \square \]

### 8.4 Proof of Theorem 4.15

Before we proceed with the main proof, we first introduce the following lemmas.
Lemma 8.12. Suppose that Assumptions 4.1, 4.3, 4.6, and 4.7 hold. Then we have
\[
\| \hat{\theta}^{(N)} - \theta^* \|_2 = O \left( \frac{\| \nabla S_1 \mathcal{L}(\theta^*) \|_2}{\rho_- (s^* + 2\delta)} + \frac{\lambda \sqrt{|S_2|}}{\rho_- (s^* + 2\delta)} + \frac{\delta_N \lambda \sqrt{s^*}}{\rho_- (s^* + 2\delta)} \right),
\]
where \( S_1 = \{ j \mid | \theta^*_j | \geq \gamma \lambda_N \} \) and \( S_2^* = \{ j \mid 0 < | \theta^*_j | < \gamma \lambda_N \} \).

The proof of Lemma 8.12 is presented in Appendix E.2. Lemma 8.12 divides the estimation error of \( \hat{\theta}^{(N)} \) into three parts: \( V_1 \) is the error for strong signals; \( V_2 \) is the error for weak signals; \( V_3 \) is the optimization error.

Lemma 8.13. Suppose that Assumptions 4.3 and 4.6 hold, \( X \) satisfies the column normalization condition, and the observation noise \( \epsilon \sim N(0, \sigma^2 I) \) is Gaussian. We then have
\[
P \left( \frac{1}{n} \| X_{S_1}^T \epsilon \|_2 \geq 3 \sigma \sqrt{\frac{\rho_+ (|S_1|) \cdot |S_1|}{n}} \right) \leq 2 \exp (-2 \|S_1\|).
\]
The proof of Lemma 8.13 is provided in Appendix E.3. Lemma 8.13 characterizes the large deviation properties of \( \| \nabla S_1 \mathcal{L}(\theta^*) \|_2 \) in Lemma 8.12 for sparse linear regression.

We proceed with the main proof. For notational simplicity, we omit the index \( N \) and denote \( \hat{\theta}^{(N)} \), \( \lambda_N \), and \( \delta_N \) by \( \hat{\theta} \), \( \lambda \), and \( \delta \) respectively. If we choose a sufficiently small \( \delta \leq \frac{1}{40 \sqrt{s^*}} \), then we apply Lemmas 8.12 and 8.13, and obtain
\[
\| \hat{\Delta} \|_2 \leq \frac{3 \sqrt{|S_1|} \sigma}{\rho_- (s^* + 2\delta)} \sqrt{\frac{\rho_+ (|S_1|) \cdot |S_1|}{n}} + \frac{3 \lambda \sqrt{|S_2|}}{\rho_- (s^* + 2\delta)} + \frac{0.3 \lambda}{\rho_- (s^* + 2\delta)}.
\]
Since all above results rely on Assumptions 4.1 and 4.6, by Lemma 8.14, we have
\[
\| \hat{\Delta} \|_2 \leq \frac{15 \sqrt{|S_1|} \sigma}{\psi_{\min}} \sqrt{\frac{\rho_+ (|S_1|) \cdot |S_1|}{n}} + \frac{(96 \sqrt{|S_2|} + 10) \sigma}{\psi_{\min}} \frac{\log d}{n}
\]
with probability at least \( 1 - 2 \exp (-2 \log d) - 2 \exp (-2 \|S_1\|) \), which completes the proof.

8.5 Proof of Theorem 4.16

Proof. For any \( \theta^* \), we consider a partition of \( \mathbb{R}^d \) as
\[
S_1 = \{ j \mid \theta^*_j \geq \frac{C \sigma}{s_1^* + s_2^*} \}, \quad \text{and} \quad S_3 = \{ j \mid \theta^*_j < \frac{C \sigma}{s_1^* + s_2^*} \}.
\]
We consider the first scenario, where \( \theta^*_S \) is known. Then we establish the minimax lower bound for estimating \( \theta^*_S \). Particularly, let \( \hat{\theta}_{S_1} \) denote any estimator of \( \theta^*_S \) based on \( y - X_S \theta^*_S \sim N(X_S \theta^*, \sigma^2 I) \). This is essentially a low dimension linear regression problem since \( s_1^* \ll n \). By the minimax lower bound for standard linear regression model established in Chapter 2 of Duchi (2015), we have
\[
\inf_{\hat{\theta}_{S_1}} \sup_{\theta \in \Theta_{(s_1^*, s_2^*), 1}} \mathbb{E} \| \hat{\theta}_{S_1} - \theta^*_S \|_2 \geq C_6 \sigma \sqrt{\frac{s_1^*}{n}}
\]
for a universal constant \( C_6 \). We then consider a second scenario, where \( \theta^*_{S_1} \) is known. Then we establish the minimax lower bound for estimating \( \theta^*_{S_2} \). Particularly, let \( \hat{\theta}_{S_2} \) denote any estimator of \( \theta^*_{S_2} \) based on \( y - X_{S_1} \theta^*_{S_1} \sim N(X_{S_2} \theta^*_{S_2}, \sigma^2 I) \). This is essentially a high dimensional sparse linear
regression problem. By the minimax lower bound for sparse linear regression model established in Raskutti et al. (2011), we have
\[
\inf_{\tilde{\theta}_{S_2}} \sup_{\theta \in \Theta_x(s_1^*, s_2^*, d)} \mathbb{E}\|\tilde{\theta}_{S_2} - \theta^*\|_2 \geq 2C_7\sqrt{\frac{s_2^2 \log(d - s_2^*)}{n}} \geq C_7\sqrt{\frac{s_2^2 \log d}{n}},
\]
where \(C_7\) is a universal constant, and the last inequality comes from the fact \(s_2^* \ll d\). Combining two scenarios, we can have
\[
\inf_{\tilde{\theta}} \sup_{\theta \in \Theta_x(s_1^*, s_2^*, d)} \mathbb{E}\|\tilde{\theta} - \theta^*\|_2 \geq \max \left\{ \inf_{\tilde{\theta}_{S_1}} \sup_{\theta \in \Theta_x(s_1^*, s_2^*, d)} \mathbb{E}\|\tilde{\theta}_{S_1} - \theta^*\|_2 + \inf_{\tilde{\theta}_{S_2}} \sup_{\theta \in \Theta_x(s_1^*, s_2^*, d)} \mathbb{E}\|\tilde{\theta}_{S_2} - \theta^*\|_2 \right\}
\[
\geq \frac{1}{2} \inf_{\tilde{\theta}_{S_1}} \sup_{\theta \in \Theta_x(s_1^*, s_2^*, d)} \mathbb{E}\|\tilde{\theta}_{S_1} - \theta^*\|_2 + \frac{1}{2} \inf_{\tilde{\theta}_{S_2}} \sup_{\theta \in \Theta_x(s_1^*, s_2^*, d)} \mathbb{E}\|\tilde{\theta}_{S_2} - \theta^*\|_2
\[
\geq C_6 \frac{1}{2} \sigma \sqrt{\frac{s_1^*}{n}} + \frac{C_7}{2} \sigma \sqrt{\frac{s_2^2 \log d}{n}} \geq C_4 \left( \sigma \sqrt{\frac{s_1^*}{n}} + \sigma \sqrt{\frac{s_2^2 \log d}{n}} \right),
\]
where \(C_4 = \min\{C_6/2, C_7/2\}\).}

### 8.6 Proof of Theorem 4.17

**Proof.** For notational simplicity, we denote \(\lambda_N\) by \(\lambda\), and \(\bar{\theta}^{\lambda_N}\) by \(\bar{\theta}\). Before we proceed with the proof, we first introduce the following lemmas.

**Lemma 8.14.** Suppose that \(X\) satisfies the column normalization condition, and the observation noise \(\epsilon \sim N(0, \sigma^2 I)\) is Gaussian. Then we have
\[
P \left( \frac{1}{n} \|X^T \epsilon\|_\infty \geq 2\sigma \sqrt{\frac{\log d}{n}} \right) \leq 2d^{-2}.
\]

**Lemma 8.15.** Suppose that Assumptions 4.1, 4.3, and 4.6, and the following event
\[
\mathcal{E}_1 = \left\{ \frac{1}{n} \|X^T \epsilon\|_\infty \geq 2\sigma \sqrt{\frac{\log d}{n}} \right\}
\]
hold. We have
\[
\frac{1}{n} X_{\ast S}(Y - X\hat{\theta}^o) + \nabla_S \mathcal{H}_\lambda(\hat{\theta}^o) + \lambda \nabla \|\hat{\theta}^o_S\|_1 = 0.
\]

**Lemma 8.16.** Suppose that Assumptions 4.1, 4.3, and 4.6, and the following event
\[
\mathcal{E}_2 = \left\{ \frac{1}{n} \|U^T \epsilon\|_\infty \geq 2\sigma \sqrt{\frac{\log d}{n}} \right\}
\]
hold, where \(U = X^\top (I - X_{\ast S}(X_{\ast S}^\top X_{\ast S})^{-1} X_{\ast S}^\top)\). There exists some \(\hat{\xi}_S^o \in \partial \|\hat{\theta}^o_S\|_1\) such that
\[
\frac{1}{n} X_{\ast S}^\top (Y - X\hat{\theta}^o) + \nabla_S \mathcal{H}_\lambda(\hat{\theta}^o) + \lambda \hat{\xi}_S^o = 0.
\]

The proof of Lemma 8.14 is provided in Negahban et al. (2012), therefore is omitted. The proof of Lemmas 8.15 and 8.16 is presented in Appendices E.4 and E.5. Lemmas 8.15 and 8.16 imply that \(\hat{\theta}^o\) satisfies the KKT condition of (1.1) over \(S\) and \(\overline{S}\) respectively. Note that the above results only
depend on Conditions $\mathcal{E}_1$ and $\mathcal{E}_2$. Meanwhile, we also have
\[
\|U_{ij}\|_2 = \|X_{ij}^T(I - X_{S}^T(X_{S}^TX_{S})^{-1}X_{S}^T)\|_2 \\
\leq \|I - X_{S}^T(X_{S}^TX_{S})^{-1}X_{S}^T\|_2\|X_{ij}\|_2 \leq \|X_{ij}\|_2 = \sqrt{n},
\]
(8.12)
where the last inequality comes from $\|I - X_{S}^T(X_{S}^TX_{S})^{-1}X_{S}^T\|_2 \leq 1$. Therefore (8.12) implies that Lemma 8.14 can be also applied to verify $\mathcal{E}_2$. Moreover, since both $\hat{\theta}^{(N)}$ and $\hat{\theta}^0$ are sparse local optima, by Lemma C.1, we further have $P(\hat{\theta}^0 = \theta) \geq 1 - 4d^{-2}$.

Moreover, since $\hat{\theta}$ converges to $\theta$, given a sufficiently small $\delta_N$, we have
\[
\|\nabla \tilde{L}_\lambda(\theta) - \nabla \tilde{L}_\lambda(\hat{\theta})\|_\infty \leq \|\tilde{L}_\lambda(\theta) - \tilde{L}_\lambda(\hat{\theta})\|_2 \leq \rho_+(s^*)\|\theta - \hat{\theta}\|_2 \leq \omega \ll \frac{\lambda}{4}.
\]
Since we have proved $\|\nabla \tilde{L}_\lambda(\theta)\|_\infty \leq \lambda/4$ in Lemma 8.16, we have
\[
\tilde{L}_\lambda(\hat{\theta}) \leq \|\nabla \tilde{L}_\lambda(\theta)\|_\infty + \|\nabla \tilde{L}_\lambda(\theta) - \nabla \tilde{L}_\lambda(\hat{\theta})\|_\infty \leq \frac{\lambda}{4} + \omega.
\]
Since $\hat{\theta}$ also satisfies the approximate KKT condition and $\delta \leq 1/8$, then we must have $\hat{\theta}^T_s = 0$.
Moreover, since we have also proved that there exists some constant $C_8$ such that $\min_{j \in S} |\bar{\theta}_j| \geq C_8 \sigma \sqrt{\log d/n}$ in Lemma 8.15, then for $\omega/\rho_-(s^*) \ll C_8 \sigma \sqrt{\log d/n}$, we have
\[
\min_{j \in S} |\hat{\theta}_j| = \min_{j \in S} |\bar{\theta}_j| - \omega \geq C_8 \sigma \sqrt{\frac{\log d}{n}} > 0.
\]
Combining with the fact $\hat{\theta}^T_s = 0$, we have $\text{supp}(\hat{\theta}) = \text{supp}(\bar{\theta}) = \text{supp}(\theta^*)$. Meanwhile, since all signals are strong enough, then by Theorem 4.15, we also have
\[
\|\hat{\theta} - \theta^*\|_2 \leq C_3 \sigma \sqrt{\frac{s^*}{n}},
\]
which completes the proof. \qed

References


Neale, B. M., Kou, Y., Liu, L., Ma’ayan, A., Samocha, K. E., Sabo, A., Lin, C.-F., Stevens, C.,


Supplementary Materials to “A General Theory of Pathwise Coordinate Optimization”

Tuo Zhao, Han Liu, and Tong Zhang

Abstract

The Supplementary Materials contain the supplementary proof of the theoretical results in the paper “A General Theory of Pathwise Coordinate Optimization” authored by Tuo Zhao, Han Liu, and Tong Zhang. Particularly, Appendix A provides the proof of Lemma 4.5; Appendix B provides the proof of all lemmas related to Theorem 4.8; Appendix C provides the proof of all lemmas related to Theorem 4.9; Appendix D provides the proof of all lemmas related to Theorem 4.12; Appendix E provides the proof of all lemmas related to statistical theory.

A Proof of Lemma 4.5

Proof. Since $\mathcal{L}(\theta)$ is twice differentiable and $\|\theta - \theta\|_0 \leq s$, by the mean value theorem, we have

$$\mathcal{L}(\theta') - \mathcal{L}(\theta) - (\theta' - \theta)^\top \nabla \mathcal{L}(\theta) = \frac{1}{2} (\theta' - \theta)^\top \nabla^2 \mathcal{L}(\tilde{\theta})(\theta' - \theta), \quad (A.1)$$

where $\tilde{\theta} = (1 - \beta)\theta' + \beta \theta$ for some $\beta \in (0, 1)$. By Definition 4.4, we have

$$\frac{\rho_-(s)}{2} \|\theta' - \theta\|_2^2 \leq \frac{1}{2} (\theta' - \theta)^\top \nabla^2 \mathcal{L}(\tilde{\theta})(\theta' - \theta) \leq \frac{\rho_+(s)}{2} \|\theta' - \theta\|_2^2. \quad (A.2)$$

Combining (A.1) with (A.2), we have

$$\frac{\rho_-(s)}{2} \|\theta' - \theta\|_2^2 \leq \mathcal{L}(\theta') - \mathcal{L}(\theta) - (\theta' - \theta)^\top \nabla \mathcal{L}(\theta) \leq \frac{\rho_+(s)}{2} \|\theta' - \theta\|_2^2. \quad (A.3)$$

By (R.1) in Assumption 4.3, we have

$$-\frac{\alpha}{2} \|\theta' - \theta\|_2^2 \leq \mathcal{H}_\lambda(\theta') - \mathcal{H}_\lambda(\theta) - (\theta' - \theta)^\top \nabla \mathcal{H}_\lambda(\theta) \leq 0. \quad (A.4)$$

Combining (A.3) with (A.4), we have

$$\frac{\rho_-(s) - \alpha}{2} \|\theta' - \theta\|_2^2 \leq \mathcal{L}_\lambda(\theta') - \mathcal{L}_\lambda(\theta) - (\theta' - \theta)^\top \nabla \mathcal{L}_\lambda(\theta) \leq \frac{\rho_+(s)}{2} \|\theta' - \theta\|_2^2. \quad (A.5)$$

By the convexity of $\|\theta\|_1$, we have

$$\|\theta'\|_1 \geq \|\theta\|_1 + (\theta' - \theta)^\top \xi \quad \text{for any } \xi \in \partial \|\theta\|_1.$$  

Combining (A.6) with (A.5), we obtain

$$\mathcal{F}_\lambda(\theta') \geq \mathcal{F}_\lambda(\theta) + (\theta' - \theta)^\top (\nabla \mathcal{L}_\lambda(\theta) + \lambda \xi) + \frac{\rho_-(s)}{2} \|\theta' - \theta\|_2^2.$$  

\[\square\]
B Lemmas for Proving Theorem 4.8

B.1 Proof of Lemma 8.1

Proof. By Lemma 8.8, we have
\[ \mathcal{F}_\lambda(w^{(t+1,k-1)}) - \mathcal{F}_\lambda(w^{(t+1,k)}) \geq \frac{\nu_\lambda}{2} (w_k^{(t+1,k-1)} - w_k^{(t+1,k)})^2 \]
which further implies
\[ \mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\theta^{(t+1)}) = \sum_{k=1}^{s} |\mathcal{F}_\lambda(w^{(t+1,k-1)}) - \mathcal{F}_\lambda(w^{(t+1,k)})| \geq \frac{\nu_\lambda}{2} \|\theta^{(t)} - \theta^{(t+1)}\|^2. \]

\[ \square \]

B.2 Proof of Lemma 8.2

Proof. We first analyze the gap for the proximal coordinate gradient descent. Let \( \theta \in \mathbb{R}^d \) be a vector satisfying \( \theta_\mathcal{A} = 0 \). By the restricted strong convexity of \( \mathcal{F}_\lambda(\theta) \), we have
\[ \mathcal{F}_\lambda(\theta) \geq \mathcal{F}_\lambda(\theta^{(t+1)}) + (\nabla_\lambda \tilde{L}_\lambda(\theta^{(t+1)}) + \lambda \xi_A^{(t+1)})^\top (\theta - \theta^{(t+1)}) + \frac{\rho_\lambda(s)}{2} \|\theta - \theta^{(t+1)}\|^2, \] (B.1)
where \( \xi_A^{(t+1)} \) satisfies the optimality condition of the proximal coordinate gradient descent,
\[ \nabla V_{\lambda,k,L}(\theta_k^{(t+1)}; w^{(t+1,k-1)}; \lambda_k^{(t+1)}) = 0 \text{ for any } k \in \mathcal{A}. \] (B.2)
We minimize both sides of (B.1) with respect to \( \theta_\mathcal{A} \) and set \( \theta_\overline{\mathcal{A}} = 0 \). We then obtain
\[ \mathcal{F}_\lambda(\theta^{(t+1)}) - \mathcal{F}_\lambda(\theta) \leq \frac{1}{2 \rho_\lambda(s)} \|\nabla_\lambda \mathcal{L}_\lambda(\theta^{(t+1)}) + \lambda \xi_A^{(t+1)}\|^2 \]
\[ \leq \frac{1}{2 \rho_\lambda(s)} \sum_{k=1}^{s} \|\nabla_k \tilde{L}_\lambda(\theta^{(t+1)}) - \nabla V_{\lambda,k,L}(\theta_k^{(t+1)}; w^{(t+1,k-1)})\|^2 \]
\[ \leq \frac{s \rho_\lambda^2(s)}{2 \rho_\lambda(s)} \|\theta^{(t+1)} - \theta^{(t)}\|^2, \] (B.3)
where (i) comes from (B.2), and (ii) comes from \( \nabla V_{\lambda,k,L}(\theta_k^{(t+1)}; w^{(t+1,k-1)}) = \nabla \tilde{L}_\lambda(w^{(t+1,k-1)}) \) and the restricted strong smoothness of \( \tilde{L}_\lambda(\theta) \).

For the exact coordinate minimization, we have \( \nabla V_{\lambda,k,L}(\theta_k^{(t+1)}; w^{(t+1,k-1)}) = \nabla \mathcal{Y}_{\lambda,k}(\theta_k^{(t+1)}; w^{(t+1,k-1)}) \).
Therefore (B.3) also holds for the exact coordinate minimization.

\[ \square \]

B.3 Proof of Lemma 8.8

Proof. For the proximal coordinate gradient descent, we have
\[ \mathcal{F}_\lambda(\theta) = V_{\lambda,k,L}(\theta_k; \theta) + \lambda |\theta_k| + \lambda \|\theta_\mathcal{A}\|_1, \] (B.4)
\[ \mathcal{F}_\lambda(w) \leq V_{\lambda,k,L}(w_k; \theta) + \lambda |\theta_k| + \lambda \|\theta_\mathcal{A}\|_1. \] (B.5)
Since \( V_{\lambda,k,L}(\theta_k; \theta) \) is strongly convex in \( \theta_k \), we have
\[ V_{\lambda,k,L}(\theta_k; \theta) - V_{\lambda,k,L}(w_k; \theta) \geq (\theta_k - w_k) \nabla V_{\lambda,k,L}(w_k; \theta) + \frac{L}{2} (w_k - \theta_k)^2. \] (B.6)
By the convexity of the absolute value function, we have
\[ |\theta_k| - |w_k| \geq (\theta_k - w_k)\xi_k, \tag{B.7} \]
where \( \xi_k \in \partial |w_k| \) satisfies the optimality condition of the proximal coordinate gradient descent,
\[ \nabla V_{\lambda,k,L}(w_k; \theta) + \lambda \xi_k = 0. \tag{B.8} \]
Subtracting (B.4) by (B.5), we have
\[ F_{\lambda}(\theta) - F_{\lambda}(w) \geq V_{\lambda,k,L}(\theta_k; \theta) - V_{\lambda,k,L}(w_k; \theta) + \lambda |\theta_k| - \lambda |w_k| \]
\[ \geq (\theta_k - w_k)(\nabla V_{\lambda,k,L}(w_k; \theta) + \lambda \xi_k) + \frac{L}{2}(w_k - \theta_k)^2 \geq (w_k - \theta_k)^2. \]
where (i) comes from (B.6) and (B.7), and (ii) comes from (B.8).

For the exact coordinate minimization, we only need to slightly trim the above analysis. More specifically, we replace \( V_{\lambda,k,L}(w_k; \theta) \) with
\[ \mathcal{Y}_{\lambda,k}(w_k; \theta) = \tilde{L}_{\lambda}(w_k, \theta_k). \]
Since \( \tilde{L}_{\lambda}(\theta) \) is restricted strongly convex, we have
\[ \mathcal{Y}_{\lambda,k}(\theta_k; \theta) - \mathcal{Y}_{\lambda,k}(w_k; \theta) \geq (\theta_k - w_k)\nabla \mathcal{Y}_{\lambda,k}(\theta_k; \theta) + \frac{\tilde{\rho}_-(1)}{2}(w_k - \theta_k)^2. \]
Eventually, we can obtain
\[ F_{\lambda}(w) - F_{\lambda}(\theta) \geq \frac{\tilde{\rho}_-(1)}{2}(w_k - \theta_k)^2. \]

We then proceed to analyze the descent for the proximal coordinate gradient descent when \( \theta_k = 0 \) and \( |\nabla_k \tilde{L}_{\lambda}(\theta)| \geq (1 + \delta)\lambda \). Then we have
\[ |w_k| = |S_{\lambda/L}(-\nabla_k \tilde{L}_{\lambda}(\theta)/L)| \geq \frac{\delta \lambda}{L}, \]
where the last inequality comes from the definition of the soft thresholding function. Thus we obtain
\[ F_{\lambda}(\theta) - F_{\lambda}(w) \geq \frac{L}{2} w_k^2 \geq \frac{\delta^2 \lambda^2}{2L}. \]

For the exact coordinate minimization, we construct an auxiliary solution \( w' \) by a proximal coordinate gradient descent iteration using \( L = \rho_+(1) \). Since \( w \) is obtained by the exact minimization, we have
\[ F_{\lambda}(\theta) - F_{\lambda}(w) \geq F_{\lambda}(\theta) - F_{\lambda}(w') \geq \frac{\delta^2 \lambda^2}{2\rho_+(1)}. \]

\[ \Box \]

C Lemmas for Proving Theorem 4.9

C.1 Proof of Lemma 8.3

Proof. Before we proceed with the proof, we first introduce the following lemma.

Lemma C.1. Suppose that Assumptions 4.3 and (4.6) hold. If \( \hat{\theta}^\lambda \) satisfies
\[ \|\hat{\theta}^\lambda\|_0 \leq \tilde{s} \quad \text{and} \quad K_{\lambda}(\hat{\theta}^\lambda) = 0, \]
then \( \hat{\theta}^\lambda \) is a unique sparse local optimum to (1.1).
The proof of Lemma is presented in Appendix C.9. We then proceed with the proof. We consider a sequence of auxiliary solutions obtained by the proximal gradient algorithm. The details for generating such a sequence can be found in Wang et al. (2014). By Theorem 5.1 in Wang et al. (2014), we know that such a sequence of solutions converges to a sparse local optimum \( \tilde{\theta}^\lambda \). By Lemma C.1, we know that the sparse local optimum is unique.

\[ \tilde{\theta}^\lambda \]

\section{Proof of Lemma 8.4}

Proof. Before we proceed with the proof, we first introduce the following lemma.

**Lemma C.2.** Suppose that Assumptions 4.1, 4.3, 4.6, and 4.7 hold. For any \( \lambda \geq \lambda_N \), if \( \theta \) satisfies
\[
\|\theta_S\|_0 \leq s \quad \text{and} \quad F_\lambda(\theta) \leq F_\lambda(\theta^*) + \frac{4\lambda^2s^*}{\bar{\rho}_-(s^* + s)},
\]
where \( s \leq 2\bar{s} + 1 \), then we have
\[
\|\theta - \theta^*\|_2 \leq \frac{9\lambda\sqrt{s^*}}{\bar{\rho}_-(s^* + s)} \quad \text{and} \quad \|\theta - \theta^*\|_1 \leq \frac{25\lambda s^*}{\bar{\rho}_-(s^* + s)}.
\]

The proof of Lemma C.2 is presented in Appendix C.3. Lemma C.2 characterizes the estimation errors of any sufficiently sparse solution with a sufficiently small objective value.

When the inner loop terminates, we have the output solution as \( \hat{\theta} = \theta^{(t+1)} \). Since both the exact and proximal coordinate gradient descent iterations always decrease the objective value, we have
\[
F_\lambda(\theta^{(t+1)}) \leq F_\lambda(\theta^*) + \frac{4\lambda^2s^*}{\bar{\rho}_-(s^* + 2\bar{s})}.
\]

By (B.3) in Appendix B.2, we have shown
\[
\|\nabla_A \tilde{L}_\lambda(\theta^{(t+1)}) + \mathbf{\xi}_A^{(t+1)}\|_2^2 \leq (s^* + 2\bar{s})\rho_+^2(s^* + \bar{s})\|\theta^{(t+1)} - \theta^{(t)}\|_2^2.
\]

Since Assumption 4.7 holds and \( \bar{\rho}_-(1) \leq \nu_+(1) \), we have
\[
\|\theta^{(t+1)} - \theta^{(t)}\|_2 \leq \tau^2\lambda^2 \leq \frac{\delta^2\lambda^2}{(s^* + 2\bar{s})\rho_+(s^* + 2\bar{s})}.
\]

Combining (C.3) with (C.4), we have \( \theta^{(t+1)} \) satisfying the approximate KKT condition over the active set,
\[
\min_{\mathbf{\xi}_A \in \partial F_\lambda(\theta^{(t+1)})} \|\nabla_A \tilde{L}_\lambda(\theta^{(t+1)}) + \lambda\mathbf{\xi}_A\|_\infty \leq \|\nabla_A \tilde{L}_\lambda(\theta^{(t+1)}) + \lambda\mathbf{\xi}_A^{(t+1)}\|_2 \leq \delta\lambda.
\]

We now proceed to characterize the sparsity of \( \hat{\theta} = \theta^{(t+1)} \) by exploiting the above approximate KKT condition. By Assumption 4.1, we have \( \lambda \geq 4\|\nabla \tilde{L}_\lambda(\theta^*)\|_\infty \), which implies
\[
\left\{ j \mid |\nabla_j \tilde{L}_\lambda(\theta^*)| \geq \frac{\lambda}{4}, \ j \in \mathcal{S} \cap \mathcal{A} \right\} = 0.
\]

We then consider an arbitrary set \( S' \) such that
\[
S' = \left\{ j \mid |\nabla_j \tilde{L}_\lambda(\hat{\theta}) - \nabla_j \tilde{L}_\lambda(\theta^*)| \geq \frac{\lambda}{2}, \ j \in \mathcal{S} \cap \mathcal{A} \right\}.
\]

Let \( s' = |S'| \). Then there exists a \( \mathbf{v} \in \mathbb{R}^d \) such that
\[
\|\mathbf{v}\|_\infty = 1, \ \|\mathbf{v}\|_0 \leq s', \ \text{and} \ s'\lambda/2 \geq \mathbf{v}^\top (\nabla \tilde{L}_\lambda(\hat{\theta}) - \nabla \tilde{L}_\lambda(\theta^*)).
\]

4
By Cauchy-Schwarz inequality, (C.6) implies
\[
\frac{s^/\lambda}{2} \leq \|v\|_2 \|\nabla \tilde{L}_\lambda(\hat{\theta}) - \nabla \tilde{L}_\lambda(\hat{\theta}^*)\|_2 \leq \sqrt{s'} \|\nabla \tilde{L}_\lambda(\hat{\theta}) - \nabla \tilde{L}_\lambda(\hat{\theta}^*)\|_2
\]
\[
\leq \rho_+(s^* + 2\bar{s}) \sqrt{s'} \|\hat{\theta} - \theta^*\|_2 \leq \rho_+(s^* + 2\bar{s}) \sqrt{s'} \frac{9\lambda \sqrt{s'}}{\rho_-(s^* + 2\bar{s})}, \tag{C.7}
\]
where (i) comes from the restricted strong smoothness of \(\tilde{L}_\lambda(\theta)\), and (ii) comes from (C.2) and Lemma C.2. By simple manipulation, (C.7) can be rewritten as
\[
\sqrt{s'} \leq \frac{18\rho_+(s^* + 2\bar{s}) \sqrt{s'}}{\rho_-(s^* + 2\bar{s})}. \tag{C.8}
\]
Since \(S'\) is arbitrary defined, by simple manipulation, (C.8) implies
\[
\left\{ j \mid \|\nabla_j \tilde{L}_\lambda(\hat{\theta}) - \nabla_j \tilde{L}_\lambda(\hat{\theta}^*)\| \geq \frac{\lambda}{2}, j \in \mathcal{S} \cap \mathcal{A} \right\} \leq 364\kappa^2 s^*. \tag{C.9}
\]
Combining (C.5) with (C.9), we have
\[
\left\{ j \mid \|\nabla_j \tilde{L}_\lambda(\hat{\theta})\| \geq \frac{3\lambda}{4}, j \in \mathcal{S} \cap \mathcal{A} \right\} \leq \left\{ j \mid \|\nabla_j \tilde{L}_\lambda(\theta^*)\| \geq \frac{\lambda}{4}, j \in \mathcal{S} \cap \mathcal{A} \right\}
\]
\[
+ \left\{ j \mid \|\nabla_j \tilde{L}_\lambda(\hat{\theta}) - \nabla_j \tilde{L}_\lambda(\theta^*)\| \geq \frac{\lambda}{2}, j \in \mathcal{S} \cap \mathcal{A} \right\} \leq 364\kappa^2 s^* < \bar{s}, \tag{C.10}
\]
where the last inequality comes from Assumption 4.6. Since we require \(\delta \leq 1/8\) in Assumption 4.7, (C.10) implies that for any \(u \in \mathbb{R}^d\) satisfying \(\|u\|_\infty \leq 1\), we have
\[
\left\{ j \mid \|\nabla_j \tilde{L}_\lambda(\hat{\theta}) + \delta \lambda u_j \| \geq \frac{7\lambda}{8}, j \in \mathcal{S} \cap \mathcal{A} \right\} \leq \bar{s}.
\]
Then for any \(j \in \mathcal{S} \cap \mathcal{A}\) satisfying \(\|\nabla_j \tilde{L}_\lambda(\hat{\theta}) + \delta \lambda u_j \| \leq 7\lambda/8\), there exists a \(\xi_j\) such that
\[
|\xi_j| \leq 1 \quad \text{and} \quad \nabla_j \tilde{L}_\lambda(\hat{\theta}) + \delta \lambda u_j + \lambda \xi_j = 0,
\]
which further implies \(\hat{\theta}_j = 0\). Therefore we must have \(\|\hat{\theta}_\mathcal{S}\|_0 \leq \bar{s}\). \qed

C.3 Proof of Lemma C.2

\textbf{Proof.} For notational simplicity, we define \(\Delta = \theta - \theta^*\). We first rewrite (C.1) as
\[
\lambda\|\theta^*\|_1 - \lambda\|\theta\|_1 + \frac{4\lambda^2 s^*}{\rho_-(s^* + s)} \geq \bar{L}_\lambda(\theta) - \bar{L}_\lambda(\theta^*). \tag{C.11}
\]
By the restricted strong convexity of \(\tilde{L}_\lambda(\theta)\), we have
\[
\tilde{L}_\lambda(\theta) - \tilde{L}_\lambda(\theta^*) - \frac{\rho_-(s^* + s)}{2} \|\Delta\|_2^2 \geq \delta \frac{1}{\delta} \|\nabla \tilde{L}(\theta^*)\|_2 \geq -\|\Delta\|_2 \|\nabla \tilde{L}(\theta^*)\|_2 - \|\Delta\|_1 \|\nabla \tilde{L}(\theta^*)\|_\infty - \|\Delta\|_1 \|\nabla \mathcal{H}_\lambda(\theta^*)\|_\infty, \tag{C.12}
\]
where (i) comes from \(\nabla \mathcal{H}_\lambda(\theta^*) = 0\) by (R.3) of Assumption 4.3, and (ii) comes from Hölder’s inequality. By Assumption 4.1 and (R.2) of Assumption 4.3, we have
\[
\|\nabla \tilde{L}(\theta^*)\|_\infty \leq \frac{\lambda}{4} \quad \text{and} \quad \|\nabla \mathcal{H}_\lambda(\theta^*)\|_\infty \leq \lambda. \tag{C.13}
\]
Combining (C.12) with (C.13), we obtain
\[
\widetilde{\mathcal{L}}_\lambda(\theta) - \widetilde{\mathcal{L}}_\lambda(\theta^*) \geq -\frac{5\lambda}{4} \|\Delta_s\|_1 - \frac{\lambda}{4} \|\Delta_{\overline{s}}\|_1 + \frac{\tilde{\rho}_-(s^* + s)}{2} \|\Delta\|_2^2. \tag{C.14}
\]
Plugging (C.14) and
\[
\|\theta^*\|_1 - \|\theta\|_1 = \|\theta^*_S\|_1 - (\|\theta_S\|_1 + \|\Delta_{\overline{s}}\|_1) \leq \|\Delta_s\|_1 - \|\Delta_{\overline{s}}\|_1
\]
into (C.11), we obtain
\[
\frac{9\lambda}{4} \|\Delta_s\|_1 + \frac{4\lambda^2 s^*}{\tilde{\rho}_-(s^* + s)} \geq \frac{3\lambda}{4} \|\Delta_S\|_1 + \frac{\tilde{\rho}_-(s^* + s)}{2} \|\Delta\|_2^2. \tag{C.15}
\]
We consider the first case: \(\tilde{\rho}_-(s^* + s)\|\Delta\|_1 > 16\lambda s^*.\) Then we have
\[
\frac{5\lambda}{2} \|\Delta_s\|_1 \geq \frac{\lambda}{2} \|\Delta_S\|_1 + \frac{\tilde{\rho}_-(s^* + s)}{2} \|\Delta\|_2^2. \tag{C.16}
\]
By simple manipulation, (C.16) implies
\[
\frac{\tilde{\rho}_-(s^* + s)}{2} \|\Delta\|_2^2 \leq \frac{5\lambda}{2} \|\Delta_s\|_1 \leq \frac{5\lambda}{2} \sqrt{s^*} \|\Delta_s\|_2 \leq \frac{5\lambda}{2} \sqrt{s^*} \|\Delta\|_2, \tag{C.17}
\]
where the second inequality comes from the fact that \(\Delta_S\) only contains \(s^*\) entries. By simple manipulation, (C.17) further implies
\[
\|\Delta\|_2 \leq \frac{5\lambda \sqrt{s^*}}{\tilde{\rho}_-(s^* + s)}. \tag{C.18}
\]
Meanwhile, (C.16) also implies
\[
\|\Delta_S\|_1 \leq 5 \|\Delta_s\|_1. \tag{C.19}
\]
Combining (C.18) with (C.19), we obtain
\[
\|\Delta\|_1 \leq 5 \|\Delta_s\|_1 \leq 5 \sqrt{s^*} \|\Delta_s\|_2 \leq 5 \sqrt{s^*} \|\Delta\|_2 \leq \frac{25\lambda s^*}{\tilde{\rho}_-(s^* + s)}. \tag{C.20}
\]
We then consider the second case: \(\tilde{\rho}_-(s^* + s)\|\Delta\|_1 \leq 16\lambda s^*.\) Then (C.15) implies
\[
\|\Delta\|_2 \leq \frac{9\lambda \sqrt{s^*}}{\tilde{\rho}_-(s^* + s)}. \tag{C.21}
\]
Combining two cases, we obtain
\[
\|\Delta\|_2 \leq \frac{9\lambda \sqrt{s^*}}{\tilde{\rho}_-(s^* + s)} \quad \text{and} \quad \|\Delta\|_1 \leq \frac{25\lambda s^*}{\tilde{\rho}_-(s^* + s)}. \tag{C.22}
\]

C.4 Proof of Lemma 8.5

Proof. By Assumption 4.1, we have \(\lambda \geq 4\|\nabla \tilde{\mathcal{L}}_\lambda(\theta^*)\|_\infty,\) which implies
\[
\left| \left\{ j \mid |\nabla_j \tilde{\mathcal{L}}_\lambda(\theta^*)| \geq \frac{\lambda}{4}, \ j \in \overline{S} \cap A \right\} \right| = 0. \tag{C.21}
\]
We then consider an arbitrary set \(S'\) such that
\[
S' = \left\{ j \mid |\nabla_j \tilde{\mathcal{L}}_\lambda(\theta^0) - \nabla_j \tilde{\mathcal{L}}_\lambda(\theta^*)| \geq \frac{\lambda}{2}, \ j \in \overline{S} \right\}.
\]
Let \( s' = |S'| \). Then there exists a \( v \in \mathbb{R}^d \) such that
\[
\|v\|_\infty = 1, \quad \|v\|_0 \leq s', \quad \text{and} \quad s'\lambda/2 \geq v^\top (\nabla \tilde{L}_\lambda(\theta^{[0]}) - \nabla \tilde{L}_\lambda(\theta^*)).
\] (C.22)

By Cauchy-Schwarz inequality, (C.22) implies
\[
\frac{s'\lambda}{2} \leq \|v\|_2\|\nabla \tilde{L}_\lambda(\theta^{[0]}) - \nabla \tilde{L}_\lambda(\theta^*)\|_2 \leq \sqrt{s'}\|\nabla \tilde{L}_\lambda(\theta^{[0]}) - \nabla \tilde{L}_\lambda(\theta^*)\|_2
\]
\[
\leq \rho_+(s^* + 2\bar{s})\sqrt{s'}\|\theta^{[0]} - \theta^*\|_2 \leq \rho_+(s^* + 2\bar{s})\sqrt{s'}\frac{9\lambda\sqrt{s'}}{\rho_-(s^* + 2\bar{s})},
\] (C.23)

where (i) comes from the restricted strong smoothness of \( \tilde{L}_\lambda(\theta) \), and (ii) comes from Lemma C.2. By simple manipulation, (C.23) can be rewritten as
\[
\sqrt{s'} \leq \frac{18\rho_+(s^* + 2\bar{s})\sqrt{s'}}{\rho_-(s^* + 2\bar{s})}.
\] (C.24)

Since \( S' \) is arbitrary defined, by simple manipulation, (C.8) implies
\[
\left\{ j \mid |\nabla_j \tilde{L}_\lambda(\theta^{[0]}) - \nabla_j \tilde{L}_\lambda(\theta^*)| \geq \frac{\lambda}{2}, \ j \in S \cap A \right\} \leq 364\kappa^2 s^*.
\] (C.25)

Combining (C.21) with (C.25), we have
\[
\left\{ j \mid |\nabla_j \tilde{L}_\lambda(\theta^{[0]})| \geq \frac{3\lambda}{4}, \ j \in S \cap A \right\} \leq \left\{ j \mid |\nabla_j \tilde{L}_\lambda(\theta^*)| \geq \frac{\lambda}{4}, \ j \in S \cap A \right\} + \left\{ j \mid |\nabla_j \tilde{L}_\lambda(\theta^{[0]}) - \nabla_j \tilde{L}_\lambda(\theta^*)| \geq \frac{\lambda}{2}, \ j \in S \cap A \right\} \leq 364\kappa^2 s^* < \bar{s},
\] (C.26)

where the last inequality comes from Assumption 4.6. Since Assumption 4.7 requires \( \varphi \leq 1/8 \), we have \( (1 - \varphi)\lambda > \frac{3}{4}\lambda \). Thus (C.26) implies that the strong rule selects at most \( \bar{s} \) irrelevant coordinates.

\[ \square \]

C.5 Proof of Lemma 8.6

Proof. Before we proceed with the proof, we first introduce the following lemmas.

Lemma C.3. Suppose that Assumptions 4.1, 4.3, 4.6, and 4.7 hold. For any \( \lambda \geq \lambda_N \), if \( \theta \) satisfies
\[
\|\theta_{S'}\|_0 \leq \bar{s} \quad \text{and} \quad F_\lambda(\theta) \leq F_\lambda(\theta^*) + \frac{4\lambda^2 s^*}{\rho_-(s^* + \bar{s})},
\] (C.27)

then we have \( |||\bar{T}_{\lambda,L}(\theta)|||_0 \leq \bar{s} \).

The proof of Lemma C.3 is presented in Appendix C.6. Since \( \theta^{[m+0.5]} \) satisfies (C.27) for all \( m = 0, 1, 2, \ldots \), by Lemma C.3, we have \( \|w_{S}^{[m+0.5]}\|_0 \leq \bar{s} \) for all \( m = 0, 1, 2, \ldots \).

Lemma C.4. Suppose that Assumptions 4.1, 4.3, 4.6, and 4.7 hold. For every active set updating iteration, if we select a coordinate as
\[
k_m = \arg\max_{k \in \mathcal{A}_m} |\nabla_k \tilde{L}_\lambda(\theta^{[m+0.5]})|,
\]
then we have
\[
k_m = \arg\min_k Q_{\lambda,k,L}(\bar{T}_{\lambda,k,L}(\theta^{[m+0.5]}); \theta^{[m+0.5]}).
\]
The proof of Lemma C.4 is presented in Appendix C.7. Lemma C.4 guarantees that our selected coordinate $k_m$ leads to a sufficient descent in the objective value. Therefore we have

$$F_\lambda(\theta^{m+0.5}) - F_\lambda(\theta^{m+1}) \geq F_\lambda(\theta^{m+0.5}) - Q_{\lambda,k_m,L}(\theta_k^{m+0.5}; \theta^{m+0.5}) \geq F_\lambda(\theta^{m+0.5}) - \frac{1}{|B_m|} \sum_{k \in B_m} Q_{\lambda,k,L}(w_k^{m+0.5}; \theta^{m+0.5})$$

(C.28)

where $B_m = \{k \mid w_k^{m+1} \neq 0 \text{ or } \theta_k^{m+0.5} \neq 0\}$ and $|B_m| \leq s^* + 2\bar{s}$. By rearranging (C.28), we obtain

$$F_\lambda(\theta^{m+0.5}) - F_\lambda(\theta^{m+1}) \geq \frac{1}{s^* + 2\bar{s}} \left[ F_\lambda(\theta^{m+0.5}) - J_{\lambda,L}(w^{m+1}; \theta^{m+0.5}) \right].$$

□

C.6 Proof of Lemma C.3

Proof. We define an auxiliary solution

$$\tilde{\theta} = \theta - \frac{1}{L} \nabla \tilde{\ell}_\lambda(\theta) = \theta - \frac{1}{L} \nabla \tilde{\ell}_\lambda(\theta^*) + \frac{1}{L} (\nabla \tilde{\ell}_\lambda(\theta) - \nabla \tilde{\ell}_\lambda(\theta^*)).$$

For notational simplicity, we denote $\Delta = \theta - \theta^*$. We first consider

$$\left\{ j \in \mathcal{S} : |\theta_j| \geq \frac{\lambda}{4} \right\} \leq \left\{ j \in \mathcal{S} : |\Delta_j| \geq \frac{\lambda}{4L} \right\} \leq \frac{4L}{\lambda} \|\Delta\|_1 \leq \frac{4L}{\lambda} \|\Delta\|_1 \leq \frac{100Ls^*}{\rho_- (s^* + \bar{s})},$$

(C.29)

where the last inequality comes from Lemma C.2. By Assumption 4.1, we have $\|\nabla \tilde{\ell}_\lambda(\theta^*)\|_{\infty,2} \leq \lambda/4$, which implies

$$\left\{ j \in \mathcal{S} : |\nabla_j \tilde{\ell}_\lambda(\theta^*)| \geq \frac{\lambda}{4} \right\} = 0.$$

(C.30)

Recall that in Appendix C.2, we have shown that

$$\left\{ j \mid |\nabla_j \tilde{\ell}_\lambda(\theta)| \geq \frac{\lambda}{2}, j \in \mathcal{S} \cap \mathcal{A} \right\} \leq 364\kappa^2 s^*.$$

(C.31)

Combining (C.29) and (C.30) with (C.31), we have

$$\left\{ j \in \mathcal{S} : |\bar{\theta}_j| \geq \frac{\lambda}{L} \right\} \leq \left\{ j \in \mathcal{S} : |\theta_j| \geq \frac{\lambda}{4L} \right\} + \left\{ j \in \mathcal{S} : |\nabla_j \tilde{\ell}_\lambda(\theta^*)| \geq \frac{\lambda}{4} \right\}$$

$$+ \left\{ j \mid |\nabla_j \tilde{\ell}_\lambda(\theta)| \geq \frac{\lambda}{2}, j \in \mathcal{S} \cap \mathcal{A} \right\} \leq \left( 364\kappa^2 + \frac{100Ls^*}{\rho_- (s^* + \bar{s})} \right) s^* \leq \bar{s},$$

(C.32)

where the last inequality comes from $L \leq \rho_+(s^* + 2\bar{s})$ and Assumption 4.6. By definition of the soft thresholding operator, we have $[T_{\lambda,L}(\theta)]_j = \mathcal{S}_{\lambda/L}(\bar{\theta}_j)$. Therefore (C.32) further implies $\|T_{\lambda,L}(\theta)\|_0 \leq \bar{s}$.

□

C.7 Proof of Lemma C.4

Proof. Suppose that there exists a coordinate $k$ such that

$$\theta_k^{m+0.5} = 0 \text{ and } |\nabla_k \tilde{\ell}_\lambda(\theta^{m+0.5})| \geq (1 + \delta)\lambda.$$

(C.33)
We conduct a proximal coordinate gradient descent iteration over the coordinate \( k \), and obtain an auxiliary solution \( w_k^{[m+1]} \). Since \( w_k^{[m+1]} \) is obtained by the proximal coordinate gradient descent over the coordinate \( k \), we have

\[ w_k^{[m+1]} = \arg\min_{w_k} Q_{\lambda,k,L}(w_k; \theta^{[m+0.5]}) \tag{C.34} \]

We then proceed to derive an upper bound for \( Q_{\lambda,k,L}(w_k^{[m+1]}; \theta^{[m+0.5]}) \). We consider

\[
Q_{\lambda,k,L}(w_k^{[m+1]}; \theta^{[m+0.5]}) = \overline{L}_k(\theta^{[m+0.5]}) + (w_k^{[m+1]} - \theta_k^{[m+0.5]}) \nabla_k \overline{L}_k(\theta^{[m+0.5]}) + \frac{L}{2} (w_k^{[m+1]} - \theta_k^{[m+0.5]})^2 + \lambda |w_k^{[m+1]}| + \lambda \|\theta_k^{[m+0.5]}\|_1. \tag{C.35}
\]

By the convexity of the absolute value function, we have

\[
|\theta_k^{[m+0.5]}| \geq |w_k^{[m+1]}| + (\theta_k^{[m+0.5]} - w_k^{[m+1]}) \xi_k, \tag{C.36}
\]

where \( \xi_k \in \partial|w_k^{[m+1]}| \) satisfies the optimality condition of (C.34), i.e.,

\[
w_k^{[m+1]} - \theta_k^{[m+0.5]} + \frac{1}{L} \nabla_k \overline{L}_k(\theta^{[m+0.5]}) + \frac{\lambda}{L} \xi_k = 0 \quad \text{for some} \quad \xi_k \in \partial|w_k^{[m+1]}|. \tag{C.37}
\]

Combining (C.36) with (C.35), we have

\[
Q_{\lambda,k,L}(w_k^{[m+1]}; \theta^{[m+0.5]}) - \mathcal{F}_\lambda(\theta^{[m+0.5]}) \leq (w_k^{[m+1]} - \theta_k^{[m+0.5]}) (\nabla_k \overline{L}_k(\theta^{[m+0.5]}) + \lambda \xi_k) + \frac{L}{2} (w_k^{[m+1]} - \theta_k^{[m+0.5]})^2 \tag{C.38}
\]

where (i) comes from (C.37) and (ii) comes from Lemma 8.8 and (C.33).

We assume that there exists another coordinate \( j \) with \( \theta_j^{[m+0.5]} = 0 \) such that

\[
|\nabla_k \overline{L}_k(\theta^{[m+0.5]})| > |\nabla_j \overline{L}_k(\theta^{[m+0.5]})|. \tag{C.39}
\]

Similarly, we conduct a proximal coordinate gradient descent iteration over the coordinate \( j \), and obtain an auxiliary solution \( w_j^{[m+1]} \). By definition of the soft thresholding function, we can rewrite \( w_k^{[m+1]} \) and \( w_j^{[m+1]} \) as

\[
w_k^{[m+1]} = -\frac{z_k}{L} \nabla_k \overline{L}_k(\theta^{[m+0.5]}) \quad \text{and} \quad w_j^{[m+1]} = -\frac{z_j}{L} \nabla_j \overline{L}_k(\theta^{[m+0.5]}),
\]

where \( z_k \) and \( z_j \) are defined as

\[
z_k = 1 - \frac{\lambda}{|\nabla_k \overline{L}_k(\theta^{[m+0.5]})|} \quad \text{and} \quad z_j = 1 - \frac{\lambda}{|\nabla_j \overline{L}_k(\theta^{[m+0.5]})|}.
\]

By (C.39), we know \( z_k \geq z_j \). Moreover, we define

\[
z = \frac{|\nabla_j \overline{L}_k(\theta^{[m+0.5]})|}{|\nabla_k \overline{L}_k(\theta^{[m+0.5]})|}, \quad z_j \quad \text{and} \quad w_k^{[m+1]} = -\frac{z}{L} \nabla_k \overline{L}_k(\theta^{[m+0.5]}). \tag{C.40}
\]
Note that we have \( |\tilde{w}_k^{[m+1]}| = |w_j^{[m+1]}| \). We then consider
\[
\mathcal{Q}_{\lambda,k,L}(w_k^{[m+1]}; \theta^{[m+0.5]}) - \tilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]})
\]
\[
= -\frac{z}{L} |\nabla_k \tilde{\mathcal{L}}_\lambda(\theta^{m+0.5})|^2 + \frac{L}{2} |\tilde{w}_k^{[m+1]}|^2 + \lambda |\tilde{w}_k^{[m+1]}| + \lambda \|\theta_k^{m+0.5}\|_1
\]
\[
\leq -\frac{z}{L} |\nabla_k \tilde{\mathcal{L}}_\lambda(\theta^{m+0.5})| \cdot |\nabla_k \tilde{\mathcal{L}}_\lambda(\theta^{m+0.5})| + \frac{L}{2} |\tilde{w}_k^{[m+1]}|^2 + \lambda |\tilde{w}_k^{[m+1]}| + \lambda \|\theta_k^{m+0.5}\|_1
\]
\[
= \mathcal{Q}_{\lambda,k,L}(w_j^{[m+1]}; \theta^{m+0.5}) - \tilde{\mathcal{L}}_\lambda(\theta^{m+0.5}),
\]
where (i) comes from (C.40) and (ii) comes from (C.33). We then have
\[
\mathcal{Q}_{\lambda,k,L}(w_k^{[m+1]}; \theta^{m+0.5}) \leq \mathcal{Q}_{\lambda,k,L}(\tilde{w}_k^{[m+1]}; \theta^{[m+0.5]}) \leq \mathcal{Q}_{\lambda,j,L}(w_j^{[m+1]}; \theta^{[m+0.5]}),
\]
where the least inequality comes from (C.34). Thus (C.41) guarantees
\[
\mathcal{Q}_{\lambda,k,m}(w_k^{[m+0.5]}; \theta^{[m+0.5]}) = \min_{j \in \mathcal{A}_m} \mathcal{Q}_{\lambda,j,L}(w_j^{[m+1]}; \theta^{[m+0.5]}),
\]
where \( k_m = \arg\max_{k \in \mathcal{A}_m} |\nabla \tilde{\mathcal{L}}_k(\theta^{[m+0.5]})| \).

For any \( j \in \mathcal{A}_m \), we construct two auxiliary solutions \( w_j^{[m+1]} \) and \( v_j^{[m+1]} \) by
\[
 w_j^{[m+1]} = \arg\min_{v_j} \mathcal{Q}_{\lambda,j,L}(v_j; \theta^{[m+0.5]}) \quad \text{and} \quad v_j^{[m+1]} = \arg\min_{v_j} \mathcal{F}_\lambda(v_j, \theta^{[m+0.5]}).
\]
Recall that \( \theta^{[m+0.5]} \) is the output solution of the previous inner loop, i.e., \( \theta^{[m+0.5]} = \theta^{(t+1)} \). By the restricted strong convexity of \( \mathcal{F}_\lambda(\theta) \), we have
\[
\mathcal{F}_\lambda(\theta^{(t+1)}) - \mathcal{F}_\lambda(v_j^{[m+1]}, \theta^{(t+1)}) \leq \left( \frac{\nabla_j \tilde{\mathcal{L}}_\lambda(\theta^{(t+1)}) + \lambda \xi_j}{2\rho_-} \right)^2 \leq \frac{\| \nabla_j \mathcal{L}_\lambda(\theta^{(t+1)}) + \lambda \xi_j \|_2^2}{2\rho_-(1)}
\]
for some \( \xi \in \partial\|\theta^{(t+1)}\|_1 \). Since the inner loop terminates when \( \|\theta^{(t+1)} - \theta^{(t)}\|_2^2 \leq \tau^2 \lambda^2 \), we have
\[
\mathcal{F}_\lambda(\theta^{(t+1)}) - \mathcal{F}_\lambda(v_j^{[m+1]}, \theta^{(t+1)}) \leq \frac{(s^* + 2\bar{s})\rho_+^2(s^* + 2\bar{s})\|\theta^{(t+1)} - \theta^{(t)}\|_2^2}{2\rho_-} \leq \frac{\delta^2 \lambda^2}{2L},
\]
where the equality comes from Assumption 4.7. Therefore (C.44) further implies
\[
\mathcal{Q}_{\lambda,j,L}(w_j^{[m+1]}; \theta^{[m+0.5]}) - \mathcal{F}_\lambda(\theta^{[m+0.5]}) \geq \mathcal{F}_\lambda(\theta^{(t+1)}) - \mathcal{F}_\lambda(v_j^{[m+1]}, \theta^{(t+1)}) \geq -\frac{\delta^2 \lambda^2}{2L}.
\]
Since \( j \) is arbitrarily selected from \( \mathcal{A}_m \), combining (C.38) with (C.45), we have
\[
\mathcal{Q}_{\lambda,k,m}(w_k^{[m+0.5]}; \theta^{[m+0.5]}) \leq \min_{j \in \mathcal{A}_m} \mathcal{Q}_{\lambda,j,L}(w_j^{[m+1]}; \theta^{[m+0.5]}).
\]
Combining (C.42) with (C.46), we have
\[
\mathcal{Q}_{\lambda,k,m,L}(w_k^{[m+0.5]}; \theta^{[m+0.5]}) = \min_j \mathcal{Q}_{\lambda,j,L}(w_j^{[m+1]}; \theta^{[m+0.5]}),
\]
which completes the proof.
C.8 Proof of Lemma 8.7

Proof. Define \( \mathcal{D}_m = \{ w \mid w \in \mathbb{R}^d, w_{\mathbb{B}_m} = 0 \} \), we have

\[
J_{\lambda,L}(w^{[m+1]}; \theta^{[m+0.5]}) = \min_{w \in \mathcal{D}_m} J_{\lambda,L}(w; \theta^{[m+0.5]})
\]

\[
= \min_{w \in \mathcal{D}_m} \tilde{L}_\lambda(w^{[m+0.5]}) + (w - \theta^{[m+0.5]})^T \nabla \tilde{L}_\lambda(\theta^{[m+0.5]}) + \lambda \| w \|_1 + \frac{L}{2} \| w - \theta^{[m+0.5]} \|_2^2
\]

\[
\leq \min_{w \in \mathcal{D}_m} \mathcal{F}_\lambda(w) + \frac{(L - \rho_-(s^* + 2\bar{s}))}{2} \| w - \theta^{[m+0.5]} \|_2^2,
\]

where the last inequality comes from the restricted strong convexity of \( \tilde{L}_\lambda(\theta) \), i.e.,

\[
\tilde{L}_\lambda(w) \leq \tilde{L}_\lambda(\theta^{[m+0.5]}) + (w - \theta^{[m+0.5]})^T \nabla \tilde{L}_\lambda(\theta^{[m+0.5]}) + \frac{\rho_-(s^* + 2\bar{s})}{2} \| w - \theta^{[m+0.5]} \|_2^2.
\]

Let \( w = z\bar{\theta} + (1 - z)\theta^{[m+0.5]} \) for \( z \in [0, 1] \). Then we have

\[
J_{\lambda,L}(w^{[m+1]}; \theta^{[m+0.5]}) \leq \min_{z \in [0, 1]} \mathcal{F}_\lambda(z\bar{\theta} + (1 - z)\theta^{[m+0.5]}) + \frac{z^2(L - \rho_-(s^* + 2\bar{s}))}{2} \| \bar{\theta}^\lambda - \theta^{[m+0.5]} \|_2^2
\]

\[
\leq \mathcal{F}_\lambda(\theta^{[m+0.5]}) + \min_{z \in [0, 1]} z[\mathcal{F}_\lambda(\bar{\theta}^\lambda) - \mathcal{F}_\lambda(\theta^{[m+0.5]})]
\]

\[
+ \frac{(z^2L - z\rho_-(s^* + 2\bar{s}))}{2} \| \bar{\theta}^\lambda - \theta^{[m+0.5]} \|_2^2,
\]

(C.47)

where the last inequality comes from the restricted strong convexity of \( \mathcal{F}_\lambda(\theta) \), i.e.,

\[
\mathcal{F}_\lambda(z\bar{\theta}^\lambda + (1 - z)\theta^{[m+0.5]}) \leq \frac{z(1 - z)\rho_-(s^* + 2\bar{s})}{2} \| \bar{\theta}^\lambda - \theta^{[m+0.5]} \|_2^2 \leq z\mathcal{F}_\lambda(\bar{\theta}^\lambda) + (1 - z)\mathcal{F}_\lambda(\theta^{[m+0.5]})
\]

By the restricted strong convexity of \( \mathcal{F}_\lambda(\theta) \), we have

\[
\| \bar{\theta}^\lambda - \theta^{[m+0.5]} \|_2^2 \leq \frac{2[\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\bar{\theta}^\lambda)]}{\rho_-(s^* + 2\bar{s})}.
\]

(C.48)

Combining (C.48) with (C.47), we obtain

\[
J_{\lambda,L}(w^{[m+1]}; \theta^{(t)}) - \mathcal{F}_\lambda(\theta^{[m+0.5]}) \leq \min_{z \in [0, 1]} \left( \frac{z^2L}{\rho_-(s^* + 2\bar{s})} - 2z \right) \left[ \mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\bar{\theta}^\lambda) \right].
\]

(C.49)

By setting \( z = \frac{\rho_-(s^* + 2\bar{s})}{L} \), we minimize the R.H.S of (C.49) and obtain

\[
\mathcal{F}_\lambda(\theta^{[m+0.5]}) - J_{\lambda,L}(w^{[m+1]}; \theta^{(t)}) \geq \frac{\rho_-(s^* + 2\bar{s})}{L} \left[ \mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\bar{\theta}^\lambda) \right].
\]

C.9 Proof of Lemma C.1

Proof. We prove the uniqueness of \( \bar{\theta}^\lambda \) by contradiction. Assume that there exist two different local optima \( \bar{\theta}^\lambda \) and \( \bar{\theta}^\lambda \). Let \( \bar{\xi} \in \partial \| \bar{\theta}^\lambda \|_1 \) and \( \bar{\xi} \in \partial \| \bar{\theta}^\lambda \|_1 \) be two subgradient vectors satisfying

\[
\nabla \tilde{L}_\lambda(\bar{\theta}^\lambda) + \lambda \bar{\xi} = 0 \quad \text{and} \quad \nabla \tilde{L}_\lambda(\bar{\theta}^\lambda) + \lambda \bar{\xi} = 0.
\]

(C.50)

Since \( \| \bar{\theta}_0^\lambda \|_0 \leq \bar{s} \) and \( \| \bar{\theta}_0^\lambda \|_0 \leq \bar{s} \), by the restricted strong convexity of \( \mathcal{F}_\lambda(\theta) \), we have

\[
\mathcal{F}_\lambda(\bar{\theta}^\lambda) \geq \mathcal{F}_\lambda(\bar{\theta}^\lambda) + (\bar{\theta}^\lambda - \bar{\theta}^\lambda)^T (\nabla \tilde{L}_\lambda(\bar{\theta}^\lambda) + \lambda \bar{\xi}) + \frac{\rho_-(s^* + 2\bar{s})}{2} \| \bar{\theta}^\lambda - \bar{\theta}^\lambda \|_2^2,
\]

(C.51)

\[
\mathcal{F}_\lambda(\bar{\theta}^\lambda) \geq \mathcal{F}_\lambda(\bar{\theta}^\lambda) + (\bar{\theta}^\lambda - \bar{\theta}^\lambda)^T (\nabla \tilde{L}_\lambda(\bar{\theta}^\lambda) + \lambda \bar{\xi}) + \frac{\rho_-(s^* + 2\bar{s})}{2} \| \bar{\theta}^\lambda - \bar{\theta}^\lambda \|_2^2.
\]

(C.52)

11
Combining (C.50) and (C.51) with (C.52), we have $\|\tilde{\theta}^\lambda - \bar{\theta}^\lambda\|^2 = 0$ implying $\tilde{\theta}^\lambda = \bar{\theta}^\lambda$. That is contradicted by the assumption. Thus the local optimum $\bar{\theta}^\lambda$ is unique.

\[ \]

C.10 Proof of Theorem 4.9 (cont’d)

Proof. We can prove similar results for the randomized selection rule by slightly trimming the proof of the greedy selection rule. We first follow similar lines to show $\|\theta[m]_S\|_0 \leq \tilde{s}$ and $\|\theta[55]_S\|_0 \leq \tilde{s} + 1$ for all $m = 0, 1, 2, \ldots$, since the randomized selection also moves only one inactive coordinate into the active set in each iteration.

For the proximal coordinate gradient descent, we construct the same auxiliary solution $w^{[m+1]} = (w_1^{[m+0.5]}, \ldots, w_d^{[m+0.5]})^\top$, where

\[
w_k^{[m+1]} = \arg\min_{\theta_k} q_{\lambda,k,L}(\theta_k; \theta^{[m+0.5]}).
\]

We define $B = \{j \mid w_k^{[m+1]} \neq \theta_k^{[m+0.5]})$. Similarly, Lemma 8.4 guarantees $|B| \leq s^* + 2\tilde{s}$. We then divide all coordinates into three subsets

\[
M_1 = B \cap \{ j \mid j \in A_m, |\nabla_j L(\theta^{[m+0.5]})| \geq (1 + \delta)\lambda \},
\]

\[
M_2 = B \cap \{ j \mid j \in A_m, |\nabla_j L(\theta^{[m+0.5]})| \leq (1 + \delta)\lambda \},
\]

\[
M_3 = B \cap \{ j \mid j \in A_m \}.
\]

Following similar lines to the proof of Lemma C.4, we have

\[
\max_{k \in M_1} q_{\lambda,k,L}(w_k^{[m+1]}; \theta^{[m+0.5]}) \leq \min_{j \in M_2 \cup M_3} q_{\lambda,j,L}(w_j^{[m+1]}; \theta^{[m+0.5]}),
\]

which implies that

\[
\frac{1}{|B|} \sum_{j \in |B|} q_{\lambda,j,L}(w_j^{[m+1]}; \theta^{[m+0.5]}) \geq \frac{1}{|M_1|} \sum_{k \in |M_1|} q_{\lambda,k,L}(w_k^{[m+1]}; \theta^{[m+0.5]}), \tag{C.53}
\]

Then conditioning on $\theta^{[m+0.5]}$, we have

\[
E F_\lambda(\theta^{[m+0.5]})|\theta^{[m+0.5]} = \frac{1}{|M_1|} \sum_{k \in |M_1|} F_\lambda(w_k^{[m+1]}, \theta_k^{[m+0.5]})
\]

\[
\leq \frac{1}{|M_1|} \sum_{k \in |M_1|} q_{\lambda,k,L}(w_k^{[m+1]}; \theta^{[m+0.5]}) \leq \frac{1}{|B|} \sum_{j \in |B|} q_{\lambda,j,L}(w_j^{[m+1]}; \theta^{[m+0.5]}), \tag{C.54}
\]

where the last inequality comes from (C.53). By rearranging (C.54), we have

\[
F_\lambda(\theta^{[m+0.5]}) - E[F_\lambda(\theta^{[m+1]})|\theta^{[m+0.5]}] \geq \frac{1}{s^* + 2\tilde{s}}[F_\lambda(w^{[m+1]}; \theta^{[m+0.5]}) - F_\lambda(\theta^{[m+0.5]})].
\]

We then follow similar lines to show

\[
E F_\lambda(\theta^{[m+1]}) - F_\lambda(\tilde{\theta}^\lambda) \leq \left(1 - \tilde{\rho}_-(s^* + 2\tilde{s}) \right) \left[F_\lambda(\theta^{[m]}) - F_\lambda(\tilde{\theta}^\lambda) \right]. \tag{C.55}
\]

For the exact coordinate minimization, we construct a similar auxiliary solution $\theta^{[m+0.75]}$ obtain by
the proximal coordinate gradient descent using \( L = \rho_+(1) \), then follow similar lines to show
\[
\mathbb{E}_{\lambda} \mathcal{F}_{\lambda}(\theta^{[m+1]}) - \mathcal{F}_{\lambda}(\tilde{\theta}^{\lambda}) \leq \left[ \mathbb{E}_{\lambda} \mathcal{F}_{\lambda}(\theta^{[m+0.75]]) - \mathcal{F}_{\lambda}(\hat{\theta}^{\lambda}) \right] \\
\leq \left( 1 - \frac{\bar{\mu} - (s^* + 2\bar{s})}{(s^* + 2\bar{s})\rho_+(1)} \right) \left[ \mathcal{F}_{\lambda}(\theta^{[m]}) - \mathcal{F}_{\lambda}(\hat{\theta}^{\lambda}) \right]. 
\]  
(C.56)

We then combine (C.55) with (C.56), and take the expectation over \( m = 0, 1, 2, \ldots \), and then obtain
\[
\mathbb{E}_{\lambda} \mathcal{F}_{\lambda}(\theta^{[m+1]}) - \mathcal{F}_{\lambda}(\hat{\theta}^{\lambda}) \leq \left( 1 - \frac{\bar{\mu} - (s^* + 2\bar{s})}{(s^* + 2\bar{s})\rho_+(1)} \right)^m \left[ \mathcal{F}_{\lambda}(\theta^{[0]}) - \mathcal{F}_{\lambda}(\hat{\theta}^{\lambda}) \right] 
\]  
(C.57)

The iteration complexity can also be derived in a similar fashion. When we have
\[
\mathcal{F}_{\lambda}(\theta^{[m]}) - \mathcal{F}_{\lambda}(\hat{\theta}^{\lambda}) \leq \frac{\delta^2\lambda^2}{3\nu_+(1)}, 
\]  
(C.58)

we must have \( |\nabla_{k_n} \tilde{\lambda}(\theta^{[m+0.5]})| \leq (1 + \delta)\lambda \). That implies that the algorithm must terminate when (C.58) holds. Applying the Markov inequality to (C.57), we have
\[
\mathbb{P} \left( \mathcal{F}_{\lambda}(\theta^{[m]}) - \mathcal{F}_{\lambda}(\hat{\theta}^{\lambda}) \geq \frac{\delta^2\lambda^2}{3\nu_+(1)} \right) \leq \frac{3\nu_+(1)}{\delta^2\lambda^2} \left[ \mathbb{E}_{\lambda} \mathcal{F}_{\lambda}(\theta^{[m]}) - \mathcal{F}_{\lambda}(\hat{\theta}^{\lambda}) \right] \\
\leq \frac{3\nu_+(1)}{\delta^2\lambda^2} \left( 1 - \frac{\bar{\mu} - (s^* + 2\bar{s})}{(s^* + 2\bar{s})\nu_+(1)} \right)^m \left[ \mathcal{F}_{\lambda}(\theta^{[0]}) - \mathcal{F}_{\lambda}(\hat{\theta}^{\lambda}) \right] \leq \vartheta. 
\]

By simple manipulation, we need
\[
m \geq \log^{-1} \left( 1 - \frac{\bar{\mu} - (s^* + 2\bar{s})}{(s^* + 2\bar{s})\nu_+(1)} \right) \log \left( \frac{\vartheta \delta^2\lambda^2}{3\nu_+(1) \left[ \mathcal{F}_{\lambda}(\theta^{[0]}) - \mathcal{F}_{\lambda}(\hat{\theta}^{\lambda}) \right]} \right) 
\]
iterations such that
\[
\mathbb{P} \left( \mathcal{F}_{\lambda}(\theta^{[m]}) - \mathcal{F}_{\lambda}(\hat{\theta}^{\lambda}) \leq \frac{\delta^2\lambda^2}{3\nu_+(1)} \right) \geq 1 - \vartheta. 
\]

\( \Box \)

C.11 Proof of Lemma 8.9

**Proof.** We assume that the truncated cyclic selection adds exactly \( \bar{s} + 1 \) inactive coordinates into the active set, and obtain a new active set \( A^{[m+1]} \). We define an auxiliary set \( B = A^{[m+1]} \cup S \). Since the objective value always decreases in each middle loop, we have
\[
\mathcal{F}_{\lambda}(\theta^{[m+1]}) \leq \mathcal{F}_{\lambda}(\theta^*) + \frac{4\lambda^2 s^*}{\bar{\mu} - (s^* + \bar{s})}. 
\]  
(C.59)

We then define \( w^{[m+1]} \) as a local optimum to the following optimization problem,
\[
\min_{\theta \in \mathbb{R}^d} \mathcal{F}_{\lambda}(\theta) \quad \text{subject to} \quad \theta_B = 0. 
\]  
(C.60)

By Assumption 4.6, we know that (C.60) is a strongly convex optimization problem. Therefore \( w^{[m+1]} \) is a unique global optimum. Moreover, since \( |B \cap \mathcal{B}| \leq 2\bar{s} + 1 \), by Lemma C.2, we have
\[
\|w^{[m+1]} - \theta^*\|_1 \leq \frac{25\lambda s^*}{\bar{\mu} - (s^* + 2\bar{s} + 1)}. 
\]
By the restricted strong convexity of $\mathcal{F}_\lambda(\theta)$, for any $\xi \in \partial \|\theta^*\|_1$ such that
\[
\mathcal{F}_\lambda(\theta^*) - \mathcal{F}_\lambda(w^{[m+1]}) \leq -(w^{[m+1]} - \theta^*)^\top (\nabla \tilde{L}_\lambda(\theta^*) + \lambda \xi)
\leq \|w^{[m+1]} - \theta^*\|_1 \cdot \|\nabla \tilde{L}_\lambda(\theta^*) + \lambda \xi\|_\infty
\]
\[
\leq \frac{25\lambda s^*}{\rho_-(s^* + 2\bar{s} + 1)} \left(\frac{\lambda}{4} + \lambda\right) \leq \frac{125\lambda^2 s^*}{4\rho_-(s^* + 2\bar{s} + 1)},
\]
where (i) comes from Hölder’s inequality, and (ii) comes from Assumption 4.1 and the fact $\|\xi\|_\infty \leq 1$. Since $w^{[m+1]}$ is the global optimum to (C.60), (C.61) further implies
\[
\mathcal{F}_\lambda(\theta^*) \leq \mathcal{F}_\lambda(\theta^{[m+1]}) + \frac{125\lambda^2 s^*}{4\rho_-(s^* + 2\bar{s} + 1)} \leq \mathcal{F}_\lambda(\theta^{[m+1]}) + \frac{125\lambda^2 s^*}{4\rho_-(s^* + 2\bar{s} + 1)}.
\]
Combining (C.62) with (C.59), we have
\[
\mathcal{F}_\lambda(\theta^{[m]}) \leq \mathcal{F}_\lambda(\theta^{[m+1]}) + \frac{36\lambda^2 s^*}{\rho_-(s^* + 2\bar{s} + 1)}.
\]

C.12 Proof of Lemma 8.10

Proof. Since the truncated cyclic selection only selects an inactive coordinate when its corresponding coordinate gradient is sufficiently large in magnitude, by Lemma 8.8, we know that adding $\bar{s} + 1$ inactive coordinates leads to
\[
\mathcal{F}_\lambda(\theta^{[m+1]}) \leq \mathcal{F}_\lambda(\theta^{[m+0.5]}) \leq \mathcal{F}_\lambda(\theta^{[m]}) - \frac{(\bar{s} + 1)\delta^2\lambda^2}{2\nu_+(1)}.
\]

D Lemmas for Proving Theorem 4.12

D.1 Proof of Theorem 4.10

Proof. For notational simplicity, we define $\Delta = \theta - \theta^*$. Let $\tilde{\xi} \in \partial \|\theta\|_1$ be a subgradient vector satisfying
\[
K_{\lambda_{K-1}}(\theta) = \|\nabla \tilde{L}_{\lambda_{K-1}}(\theta) + \lambda_{K-1} \tilde{\xi}\|_\infty.
\]
We then consider the following decomposition
\[
K_{\lambda_{K}}(\theta) \leq \|\nabla \tilde{L}_{\lambda_{K}}(\theta) + \lambda_{K} \tilde{\xi}\|_\infty \leq \|\nabla \tilde{L}_{\lambda_{K-1}}(\theta) + \lambda_{K-1} \tilde{\xi}\|_\infty + \|\lambda_{K} \tilde{\xi} - \lambda_{K-1} \tilde{\xi}\|_\infty
\]
\[
+ \|\nabla H_{\lambda_{K}}(\theta) - \nabla H_{\lambda_{K-1}}(\theta)\|_\infty \leq \frac{\delta_{K-1} \lambda_{K-1} + 3(1 - \eta)\lambda_{K-1} \leq \frac{\lambda_{K}}{4},
\]
where (i) comes from (R.4) in Assumption 4.3, and (ii) comes from $\delta_{K-1} \leq 1/8$ and $1 - \eta \leq 1/24$ in Assumption 4.1.

We then proceed to characterize the statistical error of $\theta$ in terms of $\lambda_{K}$. For notational simplicity, we omit the index $K$ and denote $\lambda_{K}$ by $\lambda$. Since (D.1) implies that $\theta$ satisfies the approximate
KKT condition for $\lambda$, then by the restricted strong convexity of $\tilde{L}_\lambda(\theta)$, we have
\[
F_\lambda(\theta^*) - \frac{\bar{\rho}-(s^*+\bar{s})}{2} \|\Delta\|^2 \geq F_\lambda(\theta) - \Delta^\top (\nabla \tilde{L}_\lambda(\theta) + \lambda \tilde{\xi})
\]
\[
\geq F_\lambda(\theta) - \|\nabla \tilde{L}_\lambda(\theta) + \lambda \tilde{\xi}\|_\infty \cdot \|\Delta\|_1 \geq F_\lambda(\theta) - \frac{\lambda}{4} \|\Delta\|_1,
\]
where (i) comes from Hölder’s inequality and (ii) comes from (D.1). We then rewrite (D.2) as
\[
\lambda \|\theta^*\|_1 - \lambda \|\theta\|_1 + \frac{\lambda}{4} \|\Delta\|_1 \geq \tilde{L}_\lambda(\theta) - \tilde{L}_\lambda(\theta^*) + \frac{\bar{\rho}-(s^*+\bar{s})}{2} \|\Delta\|^2.
\]
By the restricted strong convexity of $\tilde{L}_\lambda(\theta)$, we have
\[
\tilde{L}_\lambda(\theta) - \tilde{L}_\lambda(\theta^*) \geq \frac{3}{2} \lambda \|\Delta_s\|_1 - \frac{\lambda}{2} \|\Delta_{\overline{S}}\|_1 + \bar{\rho}-(s^*+\bar{s}) \|\Delta\|^2.
\]
Plugging (D.6) and
\[
\|\theta^*\|_1 - \|\theta\|_1 = \|\theta_{\overline{S}}\|_1 - (\|\theta_S\|_1 + \|\Delta_{\overline{S}}\|_1) \leq \|\Delta_S\|_1 - \|\Delta_{\overline{S}}\|_1
\]
into (D.3), we obtain
\[
\frac{11\lambda}{4} \|\Delta_S\|_1 \geq \frac{\lambda}{4} \|\Delta_{\overline{S}}\|_1 + \bar{\rho}-(s^*+\bar{s}) \|\Delta\|^2.
\]
By simple manipulation, (D.7) implies
\[
\bar{\rho}-(s^*+\bar{s}) \|\Delta\|^2 \leq \frac{11\lambda}{4} \|\Delta_S\|_1 \leq \frac{11\lambda}{4} \sqrt{s^*} \|\Delta_S\|_2 \leq \frac{11\lambda}{4} \sqrt{s^*} \|\Delta\|_2,
\]
where the second inequality comes from the fact that $\Delta_S$ only contains $s^*$ rows. By simple manipulation again, (D.8) implies
\[
\|\Delta\|_2 \leq \frac{11\lambda \sqrt{s^*}}{4\bar{\rho}-(s^*+\bar{s})}.
\]
Meanwhile, (D.7) also implies
\[
\|\Delta_{\overline{S}}\|_1 \leq 11 \|\Delta_S\|_1.
\]
Combining (D.9) with (D.10), we obtain
\[
\|\Delta\|_1 \leq 11 \|\Delta_S\|_1 \leq 11 \sqrt{s^*} \|\Delta_S\|_2 \leq 11 \sqrt{s^*} \|\Delta\|_2 \leq \frac{31\lambda s^*}{\bar{\rho}-(s^*+\bar{s})}.
\]
Plugging (D.11) and (D.9) into (D.2), we have
\[
F_\lambda(\theta) - F_\lambda(\theta^*) \leq \delta \lambda \|\Delta\|_1 \leq \frac{4\lambda^2 s^*}{\bar{\rho}-(s^*+\bar{s})}.
\]
D.2 Proof of Lemma 5.3

Proof. For notational simplicity, we denote $\theta^{\text{relax}}$ by $\theta$ and write $\tilde{F}_\lambda(\theta) = \mathcal{L}(\theta) + \lambda \|\theta\|_1$. Let $\tilde{\xi} \in \partial \|\theta\|_1$ be a subgradient vector satisfying

$$\|\nabla \mathcal{L}(\theta) + \lambda \tilde{\xi}\|_\infty = \min_{\xi \in \partial \|\theta\|_1} \|\nabla \mathcal{L}(\theta) + \lambda \xi\|_\infty.$$  

For notational simplicity, we define $\Delta = \theta^* - \theta$. Since $\tilde{F}_\lambda(\theta)$ is a convex function, we have

$$\tilde{F}_\lambda(\theta^*) \geq \tilde{F}_\lambda(\theta) - \Delta^\top (\nabla \mathcal{L}(\theta) + \lambda \tilde{\xi})$$

$$\geq \tilde{F}_\lambda(\theta) - \|\Delta\|_1\|\nabla \mathcal{L}(\theta) + \lambda \tilde{\xi}\|_\infty \geq \tilde{F}_\lambda(\theta) - \frac{\lambda}{8} \|\Delta\|_1,$$  \hspace{1cm} (D.12)

where the second inequality comes from Hölder’s inequality, and the last inequality comes from (5.7).

To establish the statistical properties of $\theta$, we need to verify that $\theta$ satisfies $\|\theta - \theta^*\|_2 \leq R$ such that the restricted strong convexity holds for $\theta$. We prove it by contradiction. We first assume $\|\theta - \theta^*\|_2 \geq R$. Then there exists some $z \in (0, 1)$ such that

$$\tilde{\theta} = (1 - z) \theta + z \theta^* \quad \text{and} \quad \|\tilde{\theta} - \theta^*\|_2 = R.$$  \hspace{1cm} (D.13)

Then by the convexity of $\tilde{F}_\lambda(\theta)$ again, (D.12) and (D.13) imply

$$\tilde{F}_\lambda(\tilde{\theta}) \leq (1 - z) \tilde{F}_\lambda(\theta) + z \tilde{F}_\lambda(\theta^*)$$

$$\leq (1 - z) \tilde{F}_\lambda(\theta^*) + \frac{(1 - z)\lambda}{8} \|\Delta\|_1 + z \tilde{F}_\lambda(\theta^*) \leq \tilde{F}_\lambda(\theta^*) + \frac{\lambda}{8} \|\Delta\|_1,$$  \hspace{1cm} (D.14)

where the last inequality comes from the fact

$$\|\Delta\|_1 = \|\tilde{\theta} - \theta^*\|_1 = \|(1 - z) \theta + z \theta^* - \theta^*\|_1 = (1 - z) \|\Delta\|_1.$$  

By simple manipulation, we can rewrite (D.14) as

$$\mathcal{L}(\tilde{\theta}) - \mathcal{L}(\theta^*) \leq \lambda \|\theta^*\|_1 - \lambda \|\tilde{\theta}\|_1 + \frac{\lambda}{8} \|\Delta\|_1.$$  \hspace{1cm} (D.15)

By the convexity of $\mathcal{L}(\theta)$, we have

$$\mathcal{L}(\tilde{\theta}) - \mathcal{L}(\theta^*) \geq \Delta^\top \nabla \mathcal{L}(\theta^*) \geq -\|\Delta\|_1 \|\nabla \mathcal{L}(\theta^*)\|_\infty \geq -\frac{\lambda}{8} \|\Delta\|_1 - \frac{\lambda}{8} \|\Delta_S\|_1,$$  \hspace{1cm} (D.16)

where the last inequality comes from our assumption $\lambda \geq 8 \|\nabla \mathcal{L}(\theta^*)\|_\infty$. By the decomposability of the $\ell_1$ norm, we have

$$\|\theta^*\|_1 - \|\theta\|_1 + \frac{1}{8} \|\Delta\|_1 = \|\theta_S^\perp\|_1 - (\|\theta_S\|_1 + \|\Delta_S\|_1) + \frac{1}{8} \|\Delta_S\|_1 + \frac{1}{8} \|\Delta_S^\perp\|_1$$

$$\leq \frac{9}{8} \|\Delta_S\|_1 - (1 - \delta) \|\Delta_S^\perp\|_1 \leq \frac{9}{8} \|\Delta_S\|_1 - \frac{7}{8} \|\Delta_S^\perp\|_1.$$  \hspace{1cm} (D.17)

Combining (D.15) with (D.16) and (D.17), we obtain

$$\|\Delta_S^\perp\|_1 \leq \frac{5}{3} \|\Delta_S\|_1.$$  \hspace{1cm} (D.18)
To establish the statistical properties of $\tilde{\theta}$, we define the following sets:

$$S_0 = \left\{ j \mid j \in \mathcal{S}, \sum_{k \in \mathcal{S}} \mathbb{1}(|\tilde{\theta}_k| \geq |\tilde{\theta}_j|) \leq \tilde{s} \right\},$$

$$S_1 = \left\{ j \mid j \in \mathcal{S} \setminus S_0, \sum_{k \in \mathcal{S} \setminus S_0} \mathbb{1}(|\tilde{\theta}_k| \geq |\tilde{\theta}_j|) \leq \tilde{s} \right\},$$

$$S_2 = \left\{ j \mid j \in \mathcal{S} \setminus (S_0 \cup S_1), \sum_{k \in \mathcal{S} \setminus (S_0 \cup S_1)} \mathbb{1}(|\tilde{\theta}_k| \geq |\tilde{\theta}_j|) \leq \tilde{s} \right\},$$

$$S_3 = \left\{ j \mid j \in \mathcal{S} \setminus (S_0 \cup S_1 \cup S_2), \sum_{k \in \mathcal{S} \setminus (S_0 \cup S_1 \cup S_2)} \mathbb{1}(|\tilde{\theta}_k| \geq |\tilde{\theta}_j|) \leq \tilde{s} \right\}, \ldots$$

Before we proceed with the proof, we introduce the following lemma.

**Lemma D.1** (Lemma 6.9 in Bühlmann and van de Geer (2011)). Let $b_1 \geq b_2 \geq \ldots \geq 0$. For $s \in \{1, 2, \ldots\}$, we have

$$\sqrt{\sum_{j \geq 1} b_j^2} \leq \sum_{k=1}^{\infty} \left(\sum_{j=k+1}^{\infty} b_j^2\right)^{1/2} \leq \sqrt{s} \sum_{k=1}^{\infty} b_j.$$

The proof of Lemma D.1 is provided in Bühlmann and van de Geer (2011), and therefore is omitted. By Lemma D.1 and (D.18), we have

$$\sum_{j \geq 1} \|\Delta S_j\|_1 \leq \frac{1}{\sqrt{s}} \|\Delta \mathcal{S}\|_1 \leq \frac{5}{3} \sqrt{\frac{s^*}{s}} \|\Delta \mathcal{S}\|_2 \leq \frac{5}{3} \sqrt{\frac{s^*}{s}} \|\Delta A\|_2,$$

where $A = \mathcal{S} \cup S_0$. By definition of the largest sparse eigenvalue and Assumption 4.6, given $\tilde{\theta} = z\tilde{\theta} + (1 - z)\theta^*$ for any $z \in [0, 1]$ and $j \geq 1$, we have

$$\left| \Delta_{S_j}^\top \nabla_{A}^2 \mathcal{L}(\tilde{\theta}) \Delta_A \right| \leq \rho_+(s^* + \tilde{s}) \|\Delta S_j\|_2 \|\Delta A\|_2,$$

which further implies

$$\left| \Delta_{A}^\top \nabla_{A}^2 \mathcal{L}(\tilde{\theta}) \Delta_A \right| \leq \sum_{j \geq 1} \left| \Delta_{S_j}^\top \nabla_{A}^2 \mathcal{L}(\tilde{\theta}) \Delta_A \right| = \frac{5\rho_+(s^* + 2\tilde{s})}{3} \|\Delta A\|_2^2 \left(\sqrt{\frac{s^*}{s}}\right). \quad (D.19)$$

By definition of the smallest sparse eigenvalue and Assumption 4.6 again, we have

$$\frac{\Delta_{A}^\top \nabla_{A}^2 \mathcal{L}(\tilde{\theta}) \Delta_A}{\|\Delta A\|_2^2} \geq \rho_-(s^* + \tilde{s}). \quad (D.20)$$

Combining (D.19) with (D.20), we have

$$\left| \Delta_{A}^\top \nabla_{A}^2 \mathcal{L}(\tilde{\theta}) \Delta_A \right| \leq \frac{5\rho_+(s^* + 2\tilde{s})}{3\rho_-(s^* + \tilde{s})} \sqrt{\frac{s^*}{s}} \Delta_{A}^\top \nabla_{A}^2 \mathcal{L}(\tilde{\theta}) \Delta_A,$$

which further implies

$$\left| \Delta_{A}^\top \nabla_{A}^2 \mathcal{L}(\tilde{\theta}) \Delta_A \right| \leq \frac{5\rho_+(s^* + 2\tilde{s})}{3\rho_-(s^* + \tilde{s})} \sqrt{\frac{s^*}{s}}. \Delta_{A}^\top \nabla_{A}^2 \mathcal{L}(\tilde{\theta}) \Delta_A.$$
Eventually, we have
\[
\frac{\Delta^\top \nabla^2 \mathcal{L}(\bar{\theta}) \Delta}{\|\Delta \|_2^2} \geq \left( 1 - \frac{9 \rho_+(s^* + \bar{s})}{7 \rho_-(s^* + \bar{s})} \frac{\| \bar{s} \|}{\| \bar{s} \|} \right) \rho_-(s^* + \bar{s}) \geq \frac{7 \rho_-(s^* + \bar{s})}{8},
\]
where the last inequality comes from Assumption 4.6. Then by the mean value theorem, we choose some \( z \) such that
\[
\mathcal{L}(\bar{\theta}) - \mathcal{L}(\theta^*) - \bar{\Delta}^\top \nabla \mathcal{L}(\theta^*) = \frac{1}{2} \bar{\Delta}^\top \nabla^2 \mathcal{L}(\bar{\theta}) \bar{\Delta} \geq \frac{7 \rho_-(s^* + \bar{s})}{16} \| \bar{\Delta} \|_2^2,
\]
which implies
\[
\mathcal{L}(\bar{\theta}) - \mathcal{L}(\theta^*) \geq \bar{\Delta}^\top \nabla \mathcal{L}(\theta^*) + \frac{7 \rho_-(s^* + \bar{s})}{16} \| \bar{\Delta} \|_2^2 \geq \frac{7 \rho_-(s^* + \bar{s})}{16} \bar{\Delta} \|_2^2 - \frac{\lambda}{8} \| \bar{\Delta} \|_1 - \frac{\lambda}{8} \| \bar{\Delta} \|_1.
\]
Then by (D.15) and (D.17), we have
\[
\rho_-(s^* + \bar{s}) \| \bar{\Delta} \|_2^2 \leq \rho_-(s^* + \bar{s}) \| \bar{\Delta} \|_2^2 \leq \frac{20}{7} \lambda \| \Delta \|_1
\]
\[
\leq \frac{20}{7} \sqrt{s^* \lambda} \| \Delta \|_2 \leq \frac{20}{7} \sqrt{s^* \lambda} \| \Delta \|_2,
\]
which implies
\[
\| \Delta \|_2 \leq \frac{20 \sqrt{s^* \lambda}}{7 \rho_-(s^* + \bar{s})} \quad \text{and} \quad \| \Delta \|_2 \leq \frac{20 \sqrt{s^* \lambda}}{7 \rho_-(s^* + \bar{s})}.
\]
By Lemma D.1, (22) implies
\[
\| \bar{\Delta} \|_2 \leq \frac{\| \bar{\Delta} \|_1}{\sqrt{s^*}} \leq \frac{5 \| \bar{\Delta} \|_1}{3 \sqrt{s^*}} = \frac{24 \sqrt{s^* \lambda}}{5 \rho_-(s^* + \bar{s})}.
\]
Combining the above results, we have
\[
\| \bar{\Delta} \|_2 = \sqrt{\| \bar{\Delta} \|_2^2 + \| \bar{\Delta} \|_2^2} \leq \frac{17 \sqrt{s^* \lambda}}{3 \rho_-(s^* + \bar{s})} < R.
\]
where the last inequality comes from the initial condition of \( \theta \). This conflicts with our assumption \( \| \bar{\Delta} \|_2 = R \). Therefore we must have \( \| \theta - \theta^* \|_2 \leq R \). Consequently, we repeat the above proof for \( \theta \), and obtain
\[
\| \Delta \|_2 \leq \frac{17 \sqrt{s^* \lambda}}{3 \rho_-(s^* + \bar{s})} \quad \text{and} \quad \| \Delta \|_2 \leq \frac{23 \sqrt{s^* \lambda}}{3 \rho_-(s^* + \bar{s})}.
\]
We now proceed to characterize the sparsity of \( \theta \). By Assumption 4.1 and the initial condition of \( \theta \), we have \( \lambda = 2 \lambda_N \geq 8 \| \Delta \|_2 \), which further implies
\[
\left\{ j \mid |\nabla_j \mathcal{L}(\theta^*)| \geq \frac{\lambda}{8}, \ j \in \mathcal{S} \right\} = 0.
\]
We then consider an arbitrary set \( S' \) such that
\[
S' = \left\{ j \mid |\nabla_j \mathcal{L}(\theta) - \nabla_j \mathcal{L}(\theta^*)| \geq \frac{5 \lambda}{8}, \ j \in \mathcal{S} \right\}.
\]
Let $s' = |S'|$. Then there exists $v$ such that
\[
\|v\|_{\infty} = 1, \quad \|v\|_0 \leq s', \quad \text{and} \quad 5s'\lambda/8 \leq v^T(\nabla L(\theta) - \nabla L(\theta^*)).
\]
Since $L(\theta)$ is twice differentiable, then by the mean value theorem, there exists a convex combination of $\theta$ and $\theta^*$ such that for some $z_1 \in [0, 1]$, we have
\[
\tilde{\theta} = z_1 \theta + (1 - z_1)\theta^* \quad \text{and} \quad \nabla L(\theta) - \nabla L(\theta^*) = \nabla^2 L(\tilde{\theta}) \Delta.
\]
Then we have
\[
\frac{5s'\lambda}{8} \leq v^T \nabla^2 L(\tilde{\theta}) \Delta \leq \sqrt{v^T \nabla^2 L(\tilde{\theta})} v \sqrt{\Delta^T \nabla^2 L(\tilde{\theta}) \Delta}.
\]
Since we have $\|v\|_0 \leq s'$, then we obtain
\[
\frac{3s'\lambda}{4} \leq \sqrt{\rho_+(s')} \sqrt{s} \sqrt{\Delta^T (\nabla L(\theta) - \nabla L(\theta^*))}
\leq \sqrt{\rho_+(s')} \sqrt{s} \sqrt{\|\Delta\|_1 \cdot \|\nabla L(\theta) - \nabla L(\theta^*)\|_{\infty}}
\leq \sqrt{\rho_+(s')} \sqrt{s} \sqrt{\|\Delta\|_1 (\|\nabla L(\theta)\|_{\infty} + \|\nabla L(\theta^*)\|_{\infty})}
\leq \sqrt{\rho_+(s')} \sqrt{s} \sqrt{115s^2 \lambda^2 / 12\rho_-(s^2 + \tilde{s})}.
\]
By simple manipulation, we have $\frac{5\sqrt{s^2}}{8} \leq \sqrt{\rho_+(s')} \sqrt{\frac{115s^2}{12\rho_-(s^2 + \tilde{s})}}$, which implies $s' \leq \frac{184\rho_+(s')}{15\rho_-(s^2 + \tilde{s})} \cdot s^*.$

Since $S'$ is arbitrary defined, by simple manipulation, we have
\[
\left\{ j \mid |\nabla_j L(\theta) - \nabla_j L(\theta^*)| \geq \frac{5\lambda}{8}, \quad j \in S' \right\} \leq 13\kappa s^* < \tilde{s}.
\]
Therefore (D.23) and (D.24) imply that for any $u \in \mathbb{R}^d$ satisfying $\|u\|_{\infty} \leq 1$, we have
\[
\left\{ j \mid |\nabla_j L(\hat{\theta}) + \frac{\lambda}{8} u_j| \geq \frac{7\lambda}{8}, \quad j \in S' \cap A \right\} \leq \tilde{s}.
\]
Then for any $j \in S' \cap A$ satisfying $|\nabla_j L(\hat{\theta}) + \frac{\lambda}{8} u_j| \leq 7\lambda/8$, there exists a $\xi_j \in \mathbb{R}^{q_j}$ satisfying
\[
|\xi_j| \leq 1 \quad \text{and} \quad \nabla_j L(\hat{\theta}) + \frac{\lambda}{8} u_j + \lambda \xi_j = 0,
\]
which implies $\theta_j = 0$. Therefore we have $\|\theta_{-S'}\|_0 \leq \tilde{s}$.

Since $\theta$ is sufficiently sparse, we know that the restricted strong convexity holds for $\theta$ and $\theta^*$. Then we can refine our analysis for $\theta$. By the restricted strong convexity of $\tilde{F}_\lambda(\theta)$, we have
\[
\tilde{F}_\lambda(\theta^*) - \frac{\rho_-(s^2 + \tilde{s})}{2} \|\Delta\|_2^2 \geq \tilde{F}_\lambda(\theta) - \Delta^T (\nabla L(\theta) + \lambda \xi) \geq \tilde{F}_\lambda(\theta) - \frac{\lambda}{8} \|\Delta\|_1.
\]
By simple manipulation, we can rewrite (D.25) as
\[
L(\theta) - L(\theta^*) \leq \lambda \|\theta^*\|_1 - \lambda \|\theta\|_1 + \frac{\lambda}{8} \|\Delta\|_1.
\]
By the restricted strong convexity of $L(\theta)$, we have
\[
L(\theta) - L(\theta^*) - \rho_-(s^2 + \tilde{s}) \|\Delta\|_2^2 \geq - \frac{\lambda}{8} \|\Delta_S\|_1 - \frac{\lambda}{8} \|\Delta_{\bar{S}}\|_1,
\]
where the last inequality comes from our assumption $\lambda \geq 8 \|\nabla L(\theta^*)\|_{\infty}$. By the decomposability of
the $\ell_1$ norm, we have
\[
\|\theta^*\|_1 - \|\theta\|_1 + \frac{1}{8}\|\Delta\|_1 = \|\theta^*_s\|_1 - (\|\theta_s\|_1 + \|\Delta_s\|_1) + \frac{1}{8}\|\Delta_s\|_1 + \frac{1}{8}\|\Delta_{\overline{s}}\|_1
\leq \frac{9}{8}\|\Delta_s\|_1 - (1 - \delta)\|\Delta_{\overline{s}}\|_1 \leq \frac{9}{8}\|\Delta_s\|_1 - \frac{7}{8}\|\Delta_{\overline{s}}\|_1,
\]
where the last inequality comes from $\delta < 1/8$ in Assumption 4.1. Combining (D.18) and (D.15) with (D.26) and (D.27), we obtain
\[
\rho_-(s^* + \overline{s})\|\Delta\|^2 \leq \frac{5\lambda}{4}\|\Delta_s\|_1 \leq \frac{5\lambda\sqrt{s}}{4}\|\Delta_s\|_2 \leq \frac{5\lambda\sqrt{s}}{4}\|\Delta_s\|_2,
\]
which implies that
\[
\|\Delta\|_2 \leq \frac{5\lambda\sqrt{s}}{4\rho_-(s^* + \overline{s})} \quad \text{and} \quad \|\Delta_s\|_1 \leq \sqrt{s}\|\Delta_s\|_2 \leq \frac{5\lambda s^*}{4\rho_-(s^* + \overline{s})}.
\]
By (D.18), we further have
\[
\|\Delta\|_1 \leq \frac{8}{3}\|\Delta_s\|_1 \leq \frac{10\lambda s^*}{3\rho_-(s^* + \overline{s})}.
\]
Plugging (D.28) into (D.25), we have
\[
\bar{F}_\lambda(\theta^*) \geq \bar{F}_\lambda(\theta) + \frac{8s^*}{7\rho_-(s^* + \overline{s})}.
\]
By the concavity of $H_\lambda(\theta)$ and Hölder’s inequality, we have
\[
H_\lambda(\theta^{\text{relax}}) \leq H_\lambda(\theta^*) + (\theta^{\text{relax}} - \theta^*)^\top \nabla H_\lambda(\theta^*) \leq H_\lambda(\theta^*) + \|\theta^{\text{relax}} - \theta^*\|_1 \|\nabla H_\lambda(\theta^*)\|_\infty.
\]
Since we have $\|H_\lambda(\theta)\|_\infty \leq \lambda$, by Lemma 5.3, we have
\[
H_\lambda(\theta^{\text{relax}}) \leq H_\lambda(\theta^*) + \lambda\|\theta^{\text{relax}} - \theta^*\|_1 \leq H_\lambda(\theta^*) + \Delta_{\lambda_0}.
\]
Since $F_{\lambda_0}(\theta) = \bar{F}_{\lambda_0}(\theta) + H_{\lambda_0}(\theta)$, by Lemma 5.3 again, we have $F_{\lambda_0}(\theta^{\text{relax}}) \leq F_{\lambda_0}(\theta^*) + \Delta_{\lambda_0}$, Therefore $\theta^{\text{relax}}$ is a proper initial solution for solving (1.1) with $\lambda_0$ by PICASSO. 

D.3 Proof of Theorem 8.11

Proof. Let $\overline{\xi} \in \partial\|\theta\|_1$ be a subgradient vector satisfying $K_\lambda(\theta) = \|\nabla \overline{L}_\lambda(\theta) + \lambda\overline{\xi}\|_\infty$. By the restricted strong convexity of $\overline{L}_\lambda(\theta)$, we have
\[
F_{\lambda'}(\theta) - F_{\lambda'}(\theta') \leq (\theta - \theta')^\top (\nabla L_{\lambda'}(\theta) + \nabla H_{\lambda'}(\theta) + \lambda'\overline{\xi})
\leq \|\theta - \theta'\|_1 (\|\nabla L_{\lambda'}(\theta) + \nabla H_{\lambda'}(\theta) + \lambda\overline{\xi} + \lambda'\overline{\xi} - \nabla H_{\lambda'}(\theta) + \nabla H_{\lambda'}(\theta))
\leq \|\theta - \theta'\|_1 \|\nabla L_{\lambda'}(\theta) + \nabla H_{\lambda'}(\theta) + \lambda\overline{\xi}\|_\infty + (\lambda - \lambda') + \|\nabla H_{\lambda'}(\theta) - \nabla H_{\lambda'}(\theta)\|_\infty + (\lambda - \lambda') + \|\nabla H_{\lambda'}(\theta) - \nabla H_{\lambda'}(\theta)\|_\infty
\leq (K_\lambda(\theta) + 3(\lambda - \lambda'))\|\theta - \theta'\|_1,
\]
where (i) comes from Hölder’s inequality and $\|\overline{\xi}\|_\infty \leq 1$, and (ii) comes from (R.3) of Assumption 4.3. Meanwhile, since we have
\[
\|\nabla L_{\lambda'}(\theta)\|_\infty \leq \overline{s}, \quad K_{\lambda'}(\theta') = 0 \leq \lambda' / 4, \quad K_{\lambda'}(\theta) \leq \lambda / 4,
\]

20
following similar lines to the proof of Theorem 4.10, we have
\[ \|\theta^\gamma - \theta^*\|_1 \leq \frac{25 \lambda s^*}{\tilde{\rho}_-(s^* + \tilde{s})} \quad \text{and} \quad \|\theta - \theta^*\|_1 \leq \frac{25 \lambda s^*}{\tilde{\rho}_-(s^* + \tilde{s})}, \]
which further implies
\[ \|\theta - \theta^\gamma\|_1 \leq \|\theta^* - \theta\|_1 + \|\theta^* - \theta^\gamma\|_1 \leq \frac{50(\lambda + \lambda')s^*}{\tilde{\rho}_-(s^* + \tilde{s})}. \quad \text{(D.30)} \]
Plugging (D.30) into (D.29), we obtain
\[ F_{\lambda'}(\theta) - F_{\lambda'}(\bar{\theta}_{\lambda'}) \leq \frac{50(\kappa_{\lambda}(\theta) + 3(\lambda - \lambda'))(\lambda + \lambda')s^*}{\tilde{\rho}_-(s^* + \tilde{s})}. \]

E Lemmas for Statistical Theory

E.1 Proof of Lemma 4.14

Proof. By Lemma 8.14, we have
\[ \|\nabla L(\theta^*)\|_\infty = \left\| \frac{1}{n} X^\top (y - X\theta^*) \right\|_\infty = \left\| \frac{1}{n} X^\top \epsilon \right\|_\infty. \quad \text{(E.1)} \]
Since we take \( \lambda = 8\sigma \sqrt{\log d/n} \), combining (E.1) with Lemma 8.14, we obtain
\[ \mathbb{P}(\lambda \geq 4\|\nabla L(\theta^*)\|_\infty) \leq 1 - \frac{2}{d^2}. \]
Moreover, for any \( v \in \mathbb{R}^d \) such that \( \|v\|_0 \leq s \), we have \( \|v\|_1 \leq \sqrt{s}\|v\|_2 \). Then (4.3) implies
\[ \frac{\|Xv\|_2^2}{n} \geq \psi_{\min}\|v\|_2^2 - \gamma_{\min}s\log d/n \|v\|_2^2. \quad \text{(E.2)} \]
By simple manipulation, (E.2) implies
\[ \frac{\|Xv\|_2^2}{n} \geq \frac{3\psi_{\min}^2}{4} \|v\|_2^2 \quad \text{(E.3)} \]
for \( n \) large enough such that
\[ \gamma_{\min}s\log d/n \leq \psi_{\min}/4. \quad \text{(E.4)} \]
Similarly, we can show that (4.3) implies
\[ \frac{\|Xv\|_2^2}{n} \leq \frac{5\psi_{\max}^2}{4} \|v\|_2^2 \quad \text{(E.5)} \]
for \( n \) large enough such that
\[ \gamma_{\max}s\log d/n \leq \psi_{\max}/4. \quad \text{(E.6)} \]
Since \( v \) is an arbitrary sparse vector, for \( \alpha = \psi_{\min}/4 \), (E.3) and (E.5) guarantee
\[ \tilde{\rho}_-(s) = \rho_-(s) - \alpha \geq \psi_{\min}/2 \quad \text{and} \quad \rho_+(s) = \rho_-(s) \leq 5\psi_{\max}/4. \quad \text{(E.7)} \]
Let \( s = s^* + 2\tilde{s} + 1 \). (E.7) implies
\[ 484\kappa^2 + 100\kappa \leq 484 \cdot \frac{25\psi_{\max}^2}{4\psi_{\min}^2} + 100 \cdot \frac{5\psi_{\max}}{2\psi_{\min}}. \]
Then we can choose $C_1$ as

$$C_1 = 3025 \cdot \frac{\psi_{\max}}{\psi_{\min}}^2 + 250 \cdot \frac{\psi_{\max}}{\psi_{\min}}$$

such that $\bar{s} = C_1 s^* \geq (484k^2 + 100k) s^*$. Meanwhile, we also need a large enough $n$ satisfying

$$\frac{\log d}{n} \leq \frac{\psi_{\min}}{4\gamma_{\min}(s^* + 2C_1 s^*)}$$

and

$$\frac{\log d}{n} \leq \frac{\psi_{\max}}{4\gamma_{\max}s^* + 2C_1 s^* + 1}$$

such that (E.6) and (E.4) hold.

**E.2 Proof of Lemma 8.12**

*Proof.* For notational simplicity, we omit the index $N$ and denote $\tilde{\theta}^{(N)}$, $\lambda_N$, and $\delta_N$ by $\tilde{\theta}$, $\lambda$, and $\delta$ respectively. We define $\hat{\Delta} = \tilde{\theta} - \theta^*$. Let $\hat{\xi} \in \partial \|\theta\|_1$ be a subgradient vector satisfying $K_\lambda(\tilde{\theta}) = \|\nabla \hat{L}_\lambda(\theta) + \lambda \hat{\xi}\|_\infty \leq \delta \lambda$. Then by the restricted strong convexity of $F_\lambda(\theta)$, we have

$$F_\lambda(\tilde{\theta}) \geq F_\lambda(\theta^*) + \hat{\Delta}^\top (\nabla \hat{L}_\lambda(\theta^*) + \lambda \hat{\xi}) + \frac{\bar{\rho}_*(s^* + \bar{s})}{2} \|\hat{\Delta}\|_2^2,$$  \hspace{1cm} (E.8)

$$F_\lambda(\theta^*) \geq F_\lambda(\hat{\theta}) - \hat{\Delta}^\top (\nabla \hat{L}_\lambda(\theta) + \lambda \hat{\xi}) + \frac{\bar{\rho}_*(s^* + \bar{s})}{2} \|\hat{\Delta}\|_2^2,$$  \hspace{1cm} (E.9)

where $\hat{\xi} \in \partial \|\theta^*\|_1$. Combining (E.8) with (E.9), we have

$$\bar{\rho}_*(s^* + \bar{s}) \|\hat{\Delta}\|_2^2 \leq \|\hat{\Delta}\|_1 \|\nabla \hat{L}_\lambda(\hat{\theta}) + \lambda \hat{\xi}\|_{\infty} - \hat{\Delta}^\top (\nabla \hat{L}(\theta^*) + \nabla H_\lambda(\theta^*) + \lambda \hat{\xi}) \leq \|\hat{\Delta}^\top (\nabla \hat{L}(\theta^*) + \nabla H_\lambda(\theta^*) + \lambda \hat{\xi})\|_1 + \delta \lambda \|\hat{\Delta}\|_1.$$  \hspace{1cm} (E.10)

[Bounding $V_0$] We consider the following decomposition

$$|\hat{\Delta}^\top (\nabla \hat{L}(\theta^*) + \nabla H_\lambda(\theta^*) + \lambda \hat{\xi})| \leq \sum_{A \in \{S_1, S_2, S\}} |\hat{\Delta}_A^\top (\nabla_A \hat{L}(\theta^*) + \nabla_A H_\lambda(\theta^*) + \lambda \hat{\xi}_A)|,$$

where $S_1 = \{j \mid |\theta_j^*| \geq \gamma \lambda\}$ and $S_2 = \{j \mid 0 < |\theta_j^*| < \gamma \lambda\}$. For $\bar{S}$, we have $\|\nabla \bar{S} \hat{L}(\theta^*)\|_{\infty} \leq \lambda / 4$ and $\nabla \bar{S} H_\lambda(\theta^*) = 0$. Thus there exists some $\tilde{\xi}_{\bar{S}} \in \partial \|\theta_{\bar{S}}\|_1$ such that $\nabla \bar{S} \hat{L}(\theta^*) + \nabla \bar{S} H_\lambda(\theta^*) + \lambda \tilde{\xi}_{\bar{S}} = 0$, which implies

$$|\hat{\Delta}^\top (\nabla \bar{S} \hat{L}(\theta^*) + \nabla \bar{S} H_\lambda(\theta^*) + \lambda \tilde{\xi}_{\bar{S}})| = 0.$$  \hspace{1cm} (E.11)

For all $j \in S_1$, we have $|\theta_j^*| \geq \gamma \lambda$ and $|\theta_j|$ is smooth at $\theta_j = \theta_j^*$. Thus by (R.2) of Assumption 4.3, we have $\nabla S_1 H_\lambda(\theta^*) + \lambda \tilde{\xi}_{S_1} = 0$, which implies

$$|\hat{\Delta}_{S_1}^\top (\nabla S_1 \hat{L}(\theta^*) + \nabla S_1 H_\lambda(\theta^*) + \lambda \tilde{\xi}_{S_1})| = |\hat{\Delta}_{S_1}^\top \nabla S_1 \hat{L}(\theta^*)| \leq \|\hat{\Delta}_{S_1}\|_2 \|\nabla S_1 \hat{L}(\theta^*)\|_2.$$  \hspace{1cm} (E.12)

We then consider $S_2$. Then we have

$$|\hat{\Delta}_{S_2}^\top (\nabla S_2 \hat{L}(\theta^*) + \nabla S_2 H_\lambda(\theta^*) + \lambda \tilde{\xi}_{S_2})| \leq \|\hat{\Delta}_{S_2}\|_1 (\|\nabla S_2 \hat{L}(\theta^*)\|_{\infty} + \|\nabla S_2 H_\lambda(\theta^*)\|_{\infty} + \|\lambda \tilde{\xi}_{S_2}\|_{\infty}) \leq 3 \lambda \sqrt{|S_2|} \|\hat{\Delta}\|_2.$$  \hspace{1cm} (E.13)

Combining (E.11) and (E.12) with (E.13), we have

$$V_0 \leq \|\nabla \hat{S}_1 \hat{L}(\theta^*)\|_2 \|\hat{\Delta}\|_2 + 3 \lambda \sqrt{|S_2|} \|\hat{\Delta}\|_2.$$  \hspace{1cm} (E.14)

[Bounding $V_4$] We then proceed to bound $V_4$. Since $\theta$ satisfies the approximate KKT condition, by
Theorem 4.10, we have \( \|\hat{\Delta}\|_1 \leq 11 \sqrt{s^2 \|\hat{\Delta}\|_2} \). Therefore by (E.14) into (E.10), we have
\[
\rho_-(s^* + \bar{s}) \|\hat{\Delta}\|_2^2 \leq \|\nabla S_1 \mathcal{L}(\theta^*)\|_2 \|\hat{\Delta}\|_2 + 11 \lambda \sqrt{|S_2|} \|\hat{\Delta}\|_2 + 11 \delta \lambda \sqrt{s^2 \|\hat{\Delta}\|_2}.
\]
Solving the above inequality, we complete the proof. \( \square \)

### E.3 Proof of Lemma 8.13

**Proof.** For any \( \epsilon, \bar{\epsilon} \in \mathbb{R}^n \), we have
\[
\left\| \frac{1}{n} X_{sS_1}^\top (\epsilon - \bar{\epsilon}) \right\|_2 = \frac{1}{n} \sqrt{(\epsilon - \bar{\epsilon})^\top X_{sS_1} X_{sS_1}^\top (\epsilon - \bar{\epsilon})} \leq \sqrt{\frac{1}{n} \Lambda_{\max} \left( \frac{1}{n} X_{sS_1} X_{sS_1}^\top \right) \|\epsilon - \bar{\epsilon}\|_2}. \tag{E.15}
\]

Since we have
\[
\Lambda_{\max} \left( \frac{1}{n} X_{sS_1} X_{sS_1}^\top \right) = \Lambda_{\max} \left( \frac{1}{n} X_{sS_1} X_{sS_1}^\top \right) = \rho_+ (|S_1|),
\]
then (E.15) implies
\[
\left\| \frac{1}{n} X_{sS_1}^\top (\epsilon - \bar{\epsilon}) \right\|_2 \leq \sqrt{\frac{\rho_+ (|S_1|)}{n} \|\epsilon - \bar{\epsilon}\|_2}.
\]
Therefore \( \|\frac{1}{n} X_{sS_1}^\top \epsilon\|_2 \) is a Lipschitz continuous function with a Lipschitz constant \( \sqrt{\frac{\rho_+ (|S_1|)}{n}} \). By Talagrand’s inequality (Ledoux and Talagrand, 2011), we have
\[
\mathbb{P} \left( \left\| \frac{1}{n} X_{sS_1}^\top \epsilon \right\|_2 \geq \mathbb{E}\left[ \left\| \frac{1}{n} X_{sS_1}^\top \epsilon \right\|_2 \right] + t \sqrt{\frac{\rho_+ (|S_1|)}{n}} \right) \leq 2 \exp \left( -\frac{-nt^2}{2} \right).
\]
For any vector \( \mathbf{v} \in \mathbb{R}^{|S_1|} \), we define a zero mean Gaussian random variable \( Z(\mathbf{v}) = \frac{1}{n} \mathbf{v}^\top X_{sS_1}^\top \epsilon \). Then we have
\[
\left\| \frac{1}{n} X_{sS_1}^\top \epsilon \right\|_2 = \max_{\|\mathbf{v}\|_2 = 1} Z(\mathbf{v}),
\]
which is the supremum of a Gaussian process, and can be upper bounded by the Gaussian comparison principle. For any \( \mathbf{v}, \bar{\mathbf{v}} \in \mathbb{R}^{|S_1|} \) satisfying \( \|\mathbf{v}\|_2 \leq 1 \) and \( \|\bar{\mathbf{v}}\|_2 \leq 1 \), we have
\[
\mathbb{E}(Z(\mathbf{v}) - Z(\bar{\mathbf{v}}))^2 = \frac{1}{n^2} \|X_{sS_1}(\mathbf{v} - \bar{\mathbf{v}})\|_2^2 \leq \frac{1}{n} \Lambda_{\min} \left( \frac{1}{n} X_{sS_1} X_{sS_1}^\top \right) \|\mathbf{v} - \bar{\mathbf{v}}\|_2^2 \leq \frac{\rho_+ (|S_1|)}{n} \|\mathbf{v} - \bar{\mathbf{v}}\|_2^2.
\]

We then define a second Gaussian process
\[
Y(\mathbf{v}) = \sqrt{\frac{\rho_+ (|S_1|)}{n}} \mathbf{v}^\top \mathbf{w},
\]
where \( \mathbf{w} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \) is standard Gaussian. By construction, for any pair of \( \mathbf{v} \) and \( \bar{\mathbf{v}} \), we have
\[
\mathbb{E}(Y(\mathbf{v}) - Y(\bar{\mathbf{v}}))^2 = \frac{\rho_+ (|S_1|)}{n} \|\mathbf{v} - \bar{\mathbf{v}}\|_2^2 \geq \mathbb{E}(Z(\mathbf{v}) - Z(\bar{\mathbf{v}}))^2,
\]
then by the Sudakov-Fernique comparison principle, we have
\[
\mathbb{E} \left[ \left\| \frac{1}{n} X_{sS_1}^\top \epsilon \right\|_2 \right] = \mathbb{E} \max_{\|\mathbf{v}\|_2 = 1} Z(\mathbf{v}) \leq \mathbb{E} \max_{\|\mathbf{v}\|_2 = 1} Y(\mathbf{v}).
\]
By definition of \( Y(v) \), we have
\[
\mathbb{E} \max_{\|v\|_2=1} Y(v) = \sqrt{\frac{\rho_+ (|S_1|)}{n}} \mathbb{E} \|w\|_2 = \sqrt{\frac{\rho_+ (|S_1|)}{n}} \mathbb{E} \sqrt{\|w\|_2^2} \\
\leq \sqrt{\frac{\rho_+ (|S_1|)}{n}} \sqrt{\mathbb{E} \|w\|_2^2} = \sqrt{\frac{\rho_+ (|S_1|)}{n}} |S_1|.
\]
Taking \( t = 2\sigma \sqrt{|S_1|}/n \), we have
\[
P \left( \frac{1}{n} X_{iS_1}^T \epsilon \|_2 \geq 3\sigma \sqrt{\frac{\rho_+ (|S_1|)|S_1|}{n}} \right) \leq 2 \exp \left( -\frac{2}{\rho_-(s^*)} \sqrt{\frac{\log d}{n}} \right),
\]
which completes the proof.

### E.4 Proof of Lemma 8.15

**Proof.** We then proceed to establish the error bound of the oracle estimator under the \( \ell_\infty \) norm. Since Lemma 4.14 guarantees that \( \rho_-(s) > 0 \), (4.4) is a strongly convex problem over \( \theta_S \) with a unique optimum
\[
\hat{\theta}^o_S = (X_{iS}^T X_{iS})^{-1} X_{iS}^T y. \tag{E.16}
\]
Then conditioning on the event \( \mathcal{E}_1 = \left\{ \frac{1}{n} \|X_{iS} \epsilon\|_\infty \leq 2\sigma \sqrt{\frac{\log d}{n}} \right\} \), we can rewrite (E.16) as
\[
\|\hat{\theta}^o_S - \theta^*_S\|_\infty = \|(X_{iS}^T X_{iS})^{-1} X_{iS}^T (y - \mathbb{E} y)\|_\infty = \|(X_{iS}^T X_{iS})^{-1} X_{iS}^T \epsilon\|_\infty \\
\leq \max_{j \in S} \frac{1}{\rho_-(s^*)} \frac{1}{n} \|X_{iS} \epsilon\|_\infty \leq \frac{1}{\rho_-(s^*)} \frac{2\sigma}{\rho_- (s^*)} \sqrt{\frac{\log d}{n}}. \tag{E.17}
\]
Since \( \theta^* \) satisfies (4.5), (E.17) implies
\[
\min_{j \in \hat{S}} |\hat{\theta}^o_j| = \min_{j \in \hat{S}} |\hat{\theta}^o_j - \theta^*_j + \theta^*_j| \geq \min_{j \in \hat{S}} |\theta^*_j| - \|\hat{\theta}^o_S - \theta^*_S\|_\infty \\
\geq \left( \frac{C_5 \sigma}{\psi_{\min}} - \frac{2\sigma}{\rho_-(s^*)} \right) \sqrt{\frac{\log d}{n}} \geq \left( \frac{C_5 \sigma}{\psi_{\min}} - \frac{4\sigma}{\psi_{\min}} \right) \sqrt{\frac{\log d}{n}}, \tag{E.18}
\]
where the last inequality comes from Lemma 4.14. We then take \( C_5 = 260 \), and (E.18) implies
\[
\min_{j \in \hat{S}} |\hat{\theta}^o_j| \geq \sigma \left( \frac{C_5}{8\psi_{\min}} - \frac{1}{2\psi_{\min}} \right) \sqrt{\frac{\log d}{n}} \geq \frac{32\sigma}{\psi_{\min}} \sqrt{\frac{\log d}{n}} = \gamma \lambda,
\]
where the last equality comes from \( \gamma = 1/\alpha = 4/\psi_{\min} \). Then by (R.2) of Assumption 4.3, we have
\[
\nabla_S \mathcal{H}_\lambda (\hat{\theta}^o) + \lambda \nabla \|\hat{\theta}^o_S\|_1 = 0. \tag{E.19}
\]
Combining (E.19) with the optimality condition of (4.4), we have
\[
\frac{1}{n} X_{iS} (Y - \hat{X}\hat{\theta}^o) + \nabla_S \mathcal{H}_\lambda (\hat{\theta}^o) + \lambda \nabla \|\hat{\theta}^o_S\|_1 = 0. \tag{E.20}
\]
\[\square\]
E.5 Proof of Lemma 8.16

Proof. We then consider the decomposition

\[
\|X^\top_s(Y - X\hat{\theta})\|_\infty = \|X^\top_s(Y - X_s\hat{\theta}^\top)\|_\infty \\
= \|X^\top_s(X_s\theta + \epsilon + X^\top_sX_s)^{-1}X^\top_s(X_s\theta + \epsilon)\|_\infty \\
= \|X^\top_s(I - X^\top_sX_s)^{-1}X^\top_s\epsilon\|_\infty \leq \|U^\top\epsilon\|_\infty, \tag{E.21}
\]

where \(U = X^\top(I - X^\top_sX_s)^{-1}X^\top_s\). Conditioning on the event \(E_2 = \left\{ \frac{1}{n}\|U^\top\epsilon\|_\infty \leq 2\sigma\sqrt{\log d/n} \right\}\), (E.21) implies

\[
\|\frac{1}{n}X^\top_s(Y - X\hat{\theta})\|_\infty \leq \frac{\lambda}{4}. \tag{E.22}
\]

By (R.3) of Assumption 4.3, we have \(\nabla H_\lambda(\hat{\theta}_S^\top) = 0\). Since \(|\theta_j|\) is non-differentiable when \(|\theta_j| = 0\), then (E.22) implies that there exists some \(\hat{\xi}_S \in \partial \|\hat{\theta}_S^\top\|_1\) such that

\[
\frac{1}{n}X^\top_s(Y - X\hat{\theta}) + \nabla_\lambda S H_\lambda(\hat{\theta}) + \lambda \hat{\xi}_S = 0. \tag{E.23}
\]

\(\square\)

E.6 Proof of Theorem 5.5

Proof. For notational simplicity, we omit the index \(N\) and denote \(\hat{\theta}^{(N)}\), \(\lambda_N\), and \(\delta_N\) by \(\Delta\), \(\lambda\), and \(\delta\) respectively. According to Lemma 8.12, we only need to bound \(\|\nabla S_1 \ell(\theta^*)\|_2\). We have

\[
\|\nabla S_1 \ell(\theta^*)\|_2 = \|\frac{1}{n}\sum_{i=1}^{n}(y_i - \pi_i(\theta^*))X_{iS\ell}\|_2 \leq \sqrt{|S|} \cdot \|\frac{1}{n}\sum_{i=1}^{n}(y_i - \pi_i(\theta^*))X_{iS}\|_\infty.
\]

By Lemma 5.2, we obtain

\[
P\left(\|\nabla S_1 \ell(\theta^*)\|_2 \leq 4\sqrt{\frac{|S| \log |S|}{n}}\right) \geq 1 - \frac{3}{|S|^3}.
\]

We then follow similar lines in Appendix 8.4 and derive the rate of convergence under the \(\ell_2\) norm. \(\square\)