Accelerated Path-following Iterative Shrinkage Thresholding Algorithm with Application to Semiparametric Graph Estimation

Tuo Zhao* Han Liu†

Abstract

We propose an accelerated path-following iterative shrinkage thresholding algorithm (APISTA) for solving high dimensional sparse nonconvex learning problems. The main difference between APISTA and the path-following iterative shrinkage thresholding algorithm (PISTA) is that APISTA exploits an additional coordinate descent subroutine to boost the computational performance. Such a modification, though simple, has profound impact: APISTA not only enjoys the same theoretical guarantee as that of PISTA, i.e., APISTA attains a linear rate of convergence to a unique sparse local optimum with good statistical properties, but also significantly outperforms PISTA in empirical benchmarks. As an application, we apply APISTA to solve a family of nonconvex optimization problems motivated by estimating sparse semiparametric graphical models. APISTA allows us to obtain new statistical recovery results which do not exist in the existing literature. Thorough numerical results are provided to back up our theory.

1 Introduction

High dimensional data challenge both statistics and computation. In the statistics community, researchers have proposed a large family of regularized M-estimators, including Lasso, Group Lasso, Fused Lasso, Graphical Lasso, Sparse Inverse Column Operator, Sparse Multivariate Regression, Sparse Linear Discriminant Analysis (Tibshirani, 1996; Zou and Hastie, 2005; Yuan and Lin, 2005, 2007; Banerjee et al., 2008; Tibshirani et al., 2005; Jacob et al., 2009; Fan et al., 2012; Liu and Luo, 2015; Han et al., 2012; Liu et al., 2015). Theoretical analysis of these methods usually

*Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA; e-mail: tour@cs.jhu.edu.

†Department of Operations Research Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; e-mail: hanliu@princeton.edu. Research supported by NSF IIS1116730, NSF IIS1332109, NSF IIS1408910, NSF IIS1546482-BIGDATA, NSF DMS1454377-CAREER, NIH R01GM083084, NIH R01HG06841, NIH R01MH102339, and FDA HHSF223201000072C.

rely on the sparsity of the parameter space and requires the resulting optimization problems to be strongly convex over a restricted parameter space. More details can be found in Meinshausen and Bühlmann (2006); Zhao and Yu (2006); Zou (2006); Rothman et al. (2008); Zhang and Huang (2008); Van de Geer (2008); Zhang (2009); Meinshausen and Yu (2009); Wainwright (2009); Fan et al. (2009); Zhang (2010a); Ravikumar et al. (2011); Liu et al. (2012a); Negahban et al. (2012); Han et al. (2012); Kim and Kwon (2012); Shen et al. (2012). In the optimization community, researchers have proposed a large variety of computational algorithms including the proximal gradient methods (Nesterov, 1988, 2005; NESTEROV, 2013; Beck and Teboulle, 2009b,a; Zhao and Liu, 2012; Liu et al., 2015) and coordinate descent methods (Fu, 1998; Friedman et al., 2007; Wu and Lange, 2008; Friedman et al., 2008; Meier et al., 2008; Liu et al., 2009; Friedman et al., 2010; Qin et al., 2010; Mazumder et al., 2011; Breheny and Huang, 2011; Shalev-Shwartz and Tewari, 2011; Zhao et al., 2014c).

Recently, Wang et al. (2014) propose the path-following iterative soft shrinkage thresholding algorithm (PISTA), which combines the proximal gradient algorithm with path-following optimization scheme. By exploiting the solution sparsity and restricted strong convexity, they show that PISTA attains a linear rate of convergence to a unique sparse local optimum with good statistical properties for solving a large class of sparse nonconvex learning problems. However, though the PISTA has superior theoretical properties, it is empirical performance is in general not as good as some heuristic competing methods such as the path-following coordinate descent algorithm (PCDA) (Tseng and Yun, 2009b,a; Lu and Xiao, 2013; Friedman et al., 2010; Mazumder et al., 2011; Zhao et al., 2012, 2014a). To address this concern, we propose a new computational algorithm named APISTA (Accelerated Path-following Iterative Shrinkage Thresholding Algorithm). More specifically, we exploit an additional coordinate descent subroutine to assist PISTA to efficiently decrease the objective value in each iteration. This makes APISTA significantly outperform PISTA in practice. Meanwhile, the coordinate descent subroutine preserves the solution sparsity and restricted strong convexity, therefore APISTA enjoys the same theoretical guarantee as those of PISTA, i.e., APISTA attains a linear rate of convergence to a unique sparse local optimum with good statistical properties. As an application, we apply APISTA to a family of nonconvex optimization problems motivated by estimating semiparametric graphical models (Liu et al., 2012b; Zhao and Liu, 2014). PISTA allows us to obtain new sparse recovery results on graph estimation consistency which has not been established before. Thorough numerical results are presented to back up our theory.

NOTATIONS: Let $\boldsymbol{v} = (v_1, \ldots, v_d)^T \in \mathbb{R}^d$, we define $\|\boldsymbol{v}\|_1 = \sum_j |v_j|$, $\|\boldsymbol{v}\|_2^2 = \sum_j v_j^2$, and $\|\boldsymbol{v}\|_\infty = \max_j |v_j|$. We denote the number of nonzero entries in $\boldsymbol{v}$ as $\|\boldsymbol{v}\|_0 = \sum_j \mathbb{1}(v_j \neq 0)$. We define the soft-thresholding operator as $\mathcal{S}_\lambda(\boldsymbol{v}) = [\mathrm{sign}(v_j) \cdot (|v_j| - \lambda)]_{j=1}^d$ for any $\lambda \geq 0$. Given a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, we use $\mathbf{A}_{*j} = (\mathbf{A}_{1j}, ..., \mathbf{A}_{dj})^T$ to denote the j^{th} column of $\mathbf{A}$, and $\mathbf{A}_{k*} = (\mathbf{A}_{k1}, ..., \mathbf{A}_{kd})^T$ to denote the k^{th} row of $\mathbf{A}$. Let $\Lambda_{\max}(\mathbf{A})$ and $\Lambda_{\min}(\mathbf{A})$ denote the largest and smallest eigenvalues of $\mathbf{A}$. Let $\psi_1(\mathbf{A}), \ldots, \psi_d(\mathbf{A})$ be the singular values of $\mathbf{A}$, we define the following matrix norms: $\|\mathbf{A}\|_{\mathrm{F}}^2 = \sum_j \|\mathbf{A}_{*j}\|_2^2$, $\|\mathbf{A}\|_{\max} = \max_j \|\mathbf{A}_{*j}\|_\infty$, $\|\mathbf{A}\|_1 = \max_j \|\mathbf{A}_{*j}\|_1$, $\|\mathbf{A}\|_2 = \max_j \psi_j(\mathbf{A})$, $\|\mathbf{A}\|_\infty = \max_k \|\mathbf{A}_{k*}\|_1$. We denote $\boldsymbol{v}_{\setminus j} = (v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_d)^T$

$\mathbb{R}^{d-1}$ as the subvector of $\boldsymbol{v}$ with the j^{th} entry removed. We denote $\mathbf{A}_{\backslash i \backslash j}$ as the submatrix of $\mathbf{A}$ with the i^{th} row and the j^{th} column removed. We denote $\mathbf{A}_{i \backslash j}$ to be the i^{th} row of $\mathbf{A}$ with its j^{th} entry removed. Let $\mathcal{A} \subseteq \{1, ..., d\}$, we use $\boldsymbol{v}_{\mathcal{A}}$ to denote a subvector of $\boldsymbol{v}$ by extracting all entries of $\boldsymbol{v}$ with indices in $\mathcal{A}$, and $\mathbf{A}_{\mathcal{A}\mathcal{A}}$ to denote a submatrix of $\mathbf{A}$ by extracting all entries of $\mathbf{A}$ with both row and column indices in $\mathcal{A}$.

2 Background and Problem Setup

Let $\boldsymbol{\theta}^* = (\theta_1^*, ..., \theta_d^*)^T$ be a parameter vector to be estimated. We are interested in solving a class of regularized optimization problems in a generic form:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \underbrace{\mathcal{L}(\boldsymbol{\theta}) + \mathcal{R}_\lambda(\boldsymbol{\theta})}_{\mathcal{F}_\lambda(\boldsymbol{\theta})}, \tag{2.1}$$

where $\mathcal{L}(\boldsymbol{\theta})$ is a smooth loss function and $\mathcal{R}_\lambda(\boldsymbol{\theta})$ is a nonsmooth regularization function with a regularization parameter λ.

2.1 Sparsity-inducing Nonconvex Regularization Functions

For high dimensional problems, we exploit sparsity-inducing regularization functions, which are usually continuous and decomposable with respect to each coordinate, i.e., $\mathcal{R}_\lambda(\boldsymbol{\theta}) = \sum_{j=1}^d r_\lambda(\theta_j)$. For example, the widely used ℓ_1 norm regularization decomposes as $\lambda\|\boldsymbol{\theta}\|_1 = \sum_{j=1}^d \lambda|\theta_j|$. One drawback of the ℓ_1 norm is that it incurs large estimation bias when $|\theta_j^*|$ is large. This motivates the usage of nonconvex regularizers. Examples include the SCAD (Fan and Li, 2001) regularization

$$r_\lambda(\theta_j) = \lambda|\theta_j| \cdot \mathbb{1}(|\theta_j| \le \lambda) - \frac{\theta_j^2 - 2\lambda\beta|\theta_j| + \lambda^2}{2(\beta-1)} \cdot \mathbb{1}(\lambda < |\theta_j| \le \lambda\beta) + \frac{(\beta+1)\lambda^2}{2} \cdot \mathbb{1}(|\theta_j| > \lambda\beta) \text{ for } \beta > 2,$$

and MCP (Zhang, 2010a) regularization

$$r_\lambda(\theta_j) = \lambda\left(|\theta_j| - \frac{\theta_j^2}{2\lambda\beta}\right) \cdot \mathbb{1}(|\theta_j| < \lambda\beta) + \frac{\lambda^2\beta}{2} \cdot \mathbb{1}(|\theta_j| \ge \lambda\beta) \text{ for } \beta > 1.$$

Both SCAD and MCP can be written as the sum of an ℓ_1 norm and a concave function $\mathcal{H}_\lambda(\boldsymbol{\theta})$, i.e., $\mathcal{R}_\lambda(\boldsymbol{\theta}) = \lambda\|\boldsymbol{\theta}\|_1 + \mathcal{H}_\lambda(\boldsymbol{\theta})$. It is easy to see that $\mathcal{H}_\lambda(\boldsymbol{\theta}) = \sum_{j=1}^d h_\lambda(\theta_j)$ is also decomposable with respect to each coordinate. More specifically, the SCAD regularization has

$$h_\lambda(\theta_j) = \frac{2\lambda|\theta_j| - \theta_j^2 - \lambda^2}{2(\beta-1)} \cdot \mathbb{1}(\lambda < |\theta_j| \le \lambda\beta) + \frac{(\beta+1)\lambda^2 - 2\lambda|\theta_j|}{2} \cdot \mathbb{1}(|\theta_j| > \lambda\beta),$$
$$h'_\lambda(\theta_j) = \frac{\lambda\,\text{sign}(\theta_j) - \theta_j}{\beta-1} \cdot \mathbb{1}(\lambda < |\theta_j| \le \lambda\beta) - \lambda\,\text{sign}(\theta_j) \cdot \mathbb{1}(|\theta_j| > \lambda\beta),$$

and the MCP regularization has

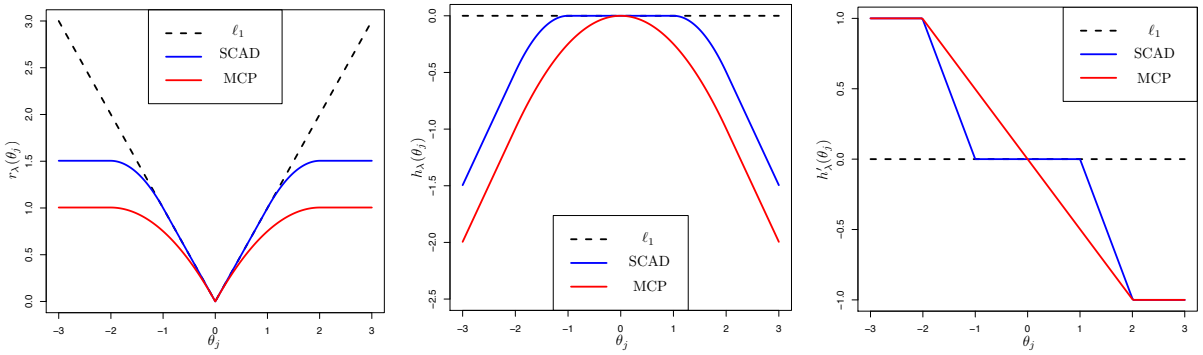
$$h_\lambda(\theta_j) = -\frac{\theta_j^2}{2\beta} \cdot \mathbb{1}(|\theta_j| < \lambda\beta) + \frac{\lambda^2\beta - 2\lambda|\theta_j|}{2} \cdot \mathbb{1}(|\theta_j| \geq \lambda\beta),$$
$$h'_\lambda(\theta_j) = -\frac{\theta_j}{\beta} \cdot \mathbb{1}(|\theta_j| \leq \lambda\beta) - \lambda \operatorname{sign}(\theta_j) \cdot \mathbb{1}(|\theta_j| > \lambda\beta).$$

In general, the concave function $h_\lambda(\cdot)$ is smooth and symmetric about zero with $h_\lambda(0) = 0$ and $h'_\lambda(0) = 0$. Its gradient $h'_\lambda(\cdot)$ is monotone decreasing and Lipschitz continuous, i.e., for any $\theta'_j > \theta_j$, there exists a constant $\alpha \geq 0$ such that

$$-\alpha(\theta_j - \theta'_j) \leq h'_\lambda(\theta_j) - h'_\lambda(\theta'_j) \leq 0. \tag{2.2}$$

Moreover, we require $h'_\lambda(\theta_j) = -\lambda \operatorname{sign}(\theta_j)$ if $|\theta_j| \geq \lambda\beta$, and $h'_\lambda(\theta_j) \in (-\lambda, 0)$ if $|\theta_j| \leq \lambda\beta$.

It is easy to verify that both SCAD and MCP satisfy the above properties. In particular, the SCAD regularization has $\alpha = 1/(\beta - 1)$, and the MCP regularization has $\alpha = 1/\beta$. These nonconvex regularization functions have been shown to achieve better asymptotic behavior than the convex ℓ_1 regularization. More technical details can be found in Fan and Li (2001); Zhang (2010a,b); Zhang and Zhang (2012); Fan et al. (2014); Xue et al. (2012); Wang et al. (2014, 2013); Liu et al. (2014). We present several illustrative examples of the nonconvex regularizers in Figure 2.1.



(a) $r_\lambda(\theta_j)$ (b) $h_\lambda(\theta_j)$ (c) $h'_\lambda(\theta_j)$

Figure 2.1: Two illustrative examples of the nonconvex regularization functions: SCAD and MCP. Here we choose $\lambda = 1$ and $\beta = 2.01$ for both SCAD and MCP.

2.2 Nonconvex Loss Function

A motivating application of the method proposed in this paper is sparse transelliptical graphical model estimation (Liu et al., 2012b). The transelliptical graphical model is a semiparametric graphical modeling tool for exploring the relationships between a large number of variables. We start with a brief review the transelliptical distribution defined below.

Definition 2.1 (Transelliptical Distribution)**.** Let $\{f_j\}_{j=1}^d$ be a set of strictly increasing univariate functions. Given a positive semidefinite matrix $\Sigma^* \in \mathbb{R}^{d\times d}$ with $\text{rank}(\Sigma^*) = r \le d$ and $\Sigma^*_{jj} = 1$ for $j = 1,...,d$, we say that a d-dimensional random vector $\boldsymbol{X} = (X_1,...,X_d)^T$ follows a transelliptical distribution, denoted as $\boldsymbol{X} \sim TE_d(\Sigma^*, \xi, f_1, \ldots, f_d)$, if $\boldsymbol{X}$ has a stochastic representation

$$(f_1(X_1),...,f_d(X_d))^T \overset{d}{=} \xi \mathbf{A}\boldsymbol{U},$$

where $\Sigma^* = \mathbf{A}\mathbf{A}^T$, $\boldsymbol{U} \in \mathbb{S}^{r-1}$ is uniformly distributed on the unit sphere in $\mathbb{R}^r$, and $\xi \ge 0$ is a continuous random variable independent of $\boldsymbol{U}$.

Note that Σ^* in Definition 2.1 is not necessarily the correlation matrix of $\boldsymbol{X}$. To interpret Σ^*, Liu et al. (2012b) provide a latent Gaussian representation for the transelliptical distribution, which implies that the sparsity pattern of $\boldsymbol{\Theta}^* = (\Sigma^*)^{-1}$ encodes the graph structure of some underlying Gaussian distribution. Since Σ^* needs to be invertible, we have $r = d$. To estimate $\boldsymbol{\Theta}^*$, Liu et al. (2012b) suggest to directly plug in the following transformed Kendall's tau estimator into existing gaussian graphical model estimation procedures.

Definition 2.2 (Transformed Kendall's tau Estimator)**.** Let $\boldsymbol{x}_1,...,\boldsymbol{x}_n \in \mathbb{R}^d$ be n independent observations of $\boldsymbol{X} = (X_1,...,X_d)^T$, where $\boldsymbol{x}_i = (x_{i1},...,x_{id})^T$. The transformed Kendall's tau estimator $\widehat{\mathbf{S}} \in \mathbb{R}^{d\times d}$ is defined as $\widehat{\mathbf{S}} = [\widehat{\mathbf{S}}_{kj}] = \left[\sin\left(\frac{\pi}{2}\widehat{\tau}_{kj}\right)\right]$, where $\widehat{\tau}_{kj}$ is the empirical Kendall's tau statistic between X_k and X_j defined as

$$\widehat{\tau}_{kj} = \begin{cases} \dfrac{2}{n(n-1)} \displaystyle\sum_{i<i'} \text{sign}\Big((x_{ij} - x_{i'j})(x_{ik} - x_{i'k})\Big) & \text{if } j \ne k, \\ 1 & \text{otherwise.} \end{cases}$$

We then adopt the sparse column inverse operator to estimate the j^{th} column of $\boldsymbol{\Theta}^*$. In particular, we solve the following regularized quadratic optimization problem (Liu and Luo, 2015),

$$\min_{\boldsymbol{\Theta}_{*j}\in\mathbb{R}^d} \frac{1}{2}\boldsymbol{\Theta}_{*j}^T\widehat{\mathbf{S}}\boldsymbol{\Theta}_{*j} - \mathbf{I}_{*j}^T\boldsymbol{\Theta}_{*j} + \mathcal{R}_\lambda(\boldsymbol{\Theta}_{*j}) \quad \text{for } j = 1,...,d. \tag{2.3}$$

For notational simplicity, we omit the column index j in (2.3), and denote $\boldsymbol{\Theta}_{*j}$ and $\mathbf{I}_{*j}$ by $\boldsymbol{\theta}$ and $\mathbf{e}$ respectively. Throughout the rest of this paper, if not specified, we study the following optimization problem for the transelliptical graph estimation

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{2}\boldsymbol{\theta}^T\widehat{\mathbf{S}}\boldsymbol{\theta} - \mathbf{e}^T\boldsymbol{\theta} + \mathcal{R}_\lambda(\boldsymbol{\theta}). \tag{2.4}$$

The quadratic loss function used in (2.4) is twice differentiable with

$$\nabla\mathcal{L}(\boldsymbol{\theta}) = \widehat{\mathbf{S}}\boldsymbol{\theta} - \mathbf{e}, \quad \nabla^2\mathcal{L}(\boldsymbol{\theta}) = \widehat{\mathbf{S}}.$$

Since the transformed Kendall's tau estimator is rank-based and could be indefinite (Zhao et al., 2014b), the optimization in (2.3) may not be convex even if $\mathcal{R}_\lambda(\boldsymbol{\theta})$ is a convex.

Remark 2.1. It is worth mentioning that the indefiniteness of $\widehat{\mathbf{S}}$ also makes(2.3) unbounded from below, but as will be shown later, our proposed algorithm can still guarantee a unique sparse local solution with optimal statistical properties under suitable solutions.

Remark 2.2. To handle the possible nonconvexity, Liu et al. (2012b) estimate $\mathbf{\Theta}^*_{*j}$ based on a graphical model estimation procedure proposed in Cai et al. (2011) as follows,

$$\min_{\mathbf{\Theta}_{*j}\in\mathbb{R}^d} \|\mathbf{\Theta}_{*j}\|_1 \quad \text{subject to } \|\widehat{\mathbf{S}}\mathbf{\Theta}_{*j} - \mathbf{I}_{*j}\|_\infty \le \lambda \quad \text{for } j = 1,...,d. \tag{2.5}$$

(2.5) is convex regardless the indefiniteness of $\widehat{\mathbf{S}}$. But a major disadvantage of (2.5) is the computation. Existing solvers can only solve (2.5) up to moderate dimensions. We will present more empirical comparison between (2.3) and (2.5) in our numerical experiments.

3 Method

For notational convenience, we rewrite the objective function $\mathcal{F}_\lambda(\boldsymbol{\theta})$ as

$$\mathcal{F}_\lambda(\boldsymbol{\theta}) = \underbrace{\mathcal{L}(\boldsymbol{\theta}) + \mathcal{H}_\lambda(\boldsymbol{\theta})}_{\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\theta})} + \lambda\|\boldsymbol{\theta}\|_1.$$

We call $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\theta})$ the augmented loss function, which is smooth but possibly nonconvex. We first introduce the path-following optimization scheme, which is a multistage optimization framework and also used in PISTA.

3.1 Path-following Optimization Scheme

The path-following optimization scheme solves the regularized optimization problem (2.1) using a decreasing sequence of $N+1$ regularization parameters $\{\lambda_K\}_{K=0}^N$, and yields a sequence of $N+1$ output solutions $\{\widehat{\boldsymbol{\theta}}^{\{K\}}\}_{K=0}^N$ from sparse to dense. We set the initial tuning parameter as $\lambda_0 = \|\nabla\mathcal{L}(\mathbf{0})\|_\infty$. By checking the KKT condition of (2.1) for λ_0, we have

$$\min_{\boldsymbol{\xi}\in\partial\|\mathbf{0}\|_1} \|\nabla\widetilde{\mathcal{L}}_\lambda(\mathbf{0}) + \lambda_0\boldsymbol{\xi}\|_\infty = \min_{\boldsymbol{\xi}\in\partial\|\mathbf{0}\|_1} \|\nabla\mathcal{L}(\mathbf{0}) + \nabla\mathcal{H}_\lambda(\mathbf{0}) + \lambda_0\boldsymbol{\xi}\|_\infty = 0, \tag{3.1}$$

where the second equality comes from $\|\boldsymbol{\xi}\|_\infty \le 1$ and $\nabla\mathcal{H}_\lambda(\mathbf{0}) = (h'_\lambda(0), h'_\lambda(0), ..., h'_\lambda(0))^T = \mathbf{0}$ as introduced in §2.1. Since (3.1) indicates that $\mathbf{0}$ is a local solution to (2.1) for λ_0, we take the leading output solution as $\widehat{\boldsymbol{\theta}}^{\{0\}} = \mathbf{0}$. Let $\eta \in (0,1)$, we set $\lambda_K = \eta\lambda_{K-1}$ for $K = 1,...,N$. We then solve (2.1) for the regularization parameter λ_K with $\widehat{\boldsymbol{\theta}}^{\{K-1\}}$ as the initial solution, which leads to the next output solution $\widehat{\boldsymbol{\theta}}^{\{K\}}$. The path-following optimization scheme is illustrated in Algorithm 1.

Algorithm 1 Path-following optimization. It solves the problem (2.1) using a decreasing sequence of regularization parameters $\{\lambda_K\}_{K=0}^N$. More specifically, $\lambda_0 = \|\mathcal{L}(\mathbf{0})\|_\infty$ yields an all zero output solution $\widehat{\boldsymbol{\theta}}^{\{0\}} = \mathbf{0}$. For $K = 1, ..., N$, we set $\lambda_K = \eta\lambda_{K-1}$, where $\eta \in (0,1)$. We solve (2.1) for λ_K with $\widehat{\boldsymbol{\theta}}^{\{K-1\}}$ as an initial solution. Note that AISTA is the computational algorithm for obtaining $\widehat{\boldsymbol{\theta}}^{K+1}$ using $\widehat{\boldsymbol{\theta}}^K$ as the initial solution. $L_{\min}$ and $\{\widehat{L}^{\{K\}}\}_{K=0}^N$ are corresponding step size parameters. More technical details on AISTA are presented are Algorithm 3.

Algorithm: $\{\widehat{\boldsymbol{\theta}}^{(K)}\}_{K=0}^N \leftarrow APISTA(\{\lambda_K\}_{K=0}^N)$
Parameter: η, $L_{\min}$
Initialize: $\lambda_0 = \|\nabla\mathcal{L}(\mathbf{0})\|_\infty$, $\widehat{\boldsymbol{\theta}}^{\{0\}} \leftarrow \mathbf{0}$, $\widehat{L}^{\{0\}} \leftarrow L_{\min}$
For: $K = 0, ..., N-1$
$\lambda_{K+1} \leftarrow \eta\lambda_K$, $\{\widehat{\boldsymbol{\theta}}^{\{K+1\}}, \widehat{L}^{\{K+1\}}\} \leftarrow AISTA(\lambda_{K+1}, \widehat{\boldsymbol{\theta}}^{\{K\}}, \widehat{L}^{\{K\}})$
End for
Output: $\{\widehat{\boldsymbol{\theta}}^{(K)}\}_{K=0}^N$

3.2 Accelerated Iterative Shrinkage Thresholding Algorithm

We then explain the accelerated iterative shrinkage thresholding (AISTA) subroutine, which solves (2.1) in each stage of the path-following optimization scheme. For notational simplicity, we omit the stage index K, and only consider the iteration index m of AISTA. Suppose that AISTA takes some initial solution $\boldsymbol{\theta}^{[0]}$ and an initial step size parameter $L^{[0]}$, and we want to solve (2.1) with the regularization parameter λ. Then at the m^{th} iteration of AISTA, we already have $L^{[m]}$ and $\boldsymbol{\theta}^{[m]}$. Each iteration of AISTA contains two steps: The first one is the proximal gradient descent iteration, and the second one is the coordinate descent subroutine.

(I) Proximal Gradient Descent Iteration: We consider the following quadratic approximation of $\mathcal{F}_\lambda(\boldsymbol{\theta})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^{[m]}$,

$$\mathcal{Q}_{\lambda, L^{[m+1]}}(\boldsymbol{\theta}; \boldsymbol{\theta}^{[m]}) = \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\theta}^{[m]}) + (\boldsymbol{\theta} - \boldsymbol{\theta}^{[m]})^T \nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\theta}^{[m]}) + \frac{L^{[m+1]}}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^{[m]}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1,$$

where $L^{[m+1]}$ is the step size parameter such that $\mathcal{Q}_{\lambda, L^{[m+1]}}(\boldsymbol{\theta}; \boldsymbol{\theta}^{[m]}) \geq \mathcal{F}_\lambda(\boldsymbol{\theta})$. We then take a proximal gradient descent iteration and obtain $\boldsymbol{\theta}^{[m]}$ by

$$\boldsymbol{\theta}^{[m+0.5]} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\operatorname{argmin}}\ \mathcal{Q}_{\lambda, L^{[m+1]}}(\boldsymbol{\theta}; \boldsymbol{\theta}^{[m]}) = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\operatorname{argmin}} \frac{L^{[m+1]}}{2}\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}^{[m]}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1, \tag{3.2}$$

where $\widetilde{\boldsymbol{\theta}}^{[m]} = \boldsymbol{\theta}^{[m]} - \nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\theta}^{[m]})/L^{[m+1]}$. For notational simplicity, we write

$$\boldsymbol{\theta}^{[m+0.5]} = \mathcal{T}_{\lambda, L^{[m+1]}}(\boldsymbol{\theta}^{[m]}). \tag{3.3}$$

For sparse column inverse operator, we can obtain a closed form solution to (3.2) by soft thresholding

$$\boldsymbol{\theta}^{[m+0.5]} = \mathcal{S}_{\lambda/L^{[m+1]}}(\widetilde{\boldsymbol{\theta}}^{[m+0.5]}).$$

The step size $1/L^{[m+1]}$ can be obtained by the backtracking line search. In particular, we start with a small enough $L^{[0]}$. Then in each iteration of the middle loop, we choose the minimum nonnegative integer z such that $L^{[m+1]} = 2^z L^{[m]}$ satisfies

$$\mathcal{F}_\lambda(\boldsymbol{\theta}^{[m+0.5]}) \leq \mathcal{Q}_{\lambda,L^{[m+1]}}(\boldsymbol{\theta}^{[m+0.5]};\boldsymbol{\theta}^{[m]}) \text{ for } m = 0,1,2,.... \tag{3.4}$$

(II) Coordinate Descent Subroutine: Unlike the proximal gradient algorithm which repeats (3.3) until convergence at each stage of the path-following optimization scheme, AISTA exploits an additional coordinate descent subroutine to further boost the computational performance. More specifically, we define $\mathcal{A}^\perp = \{j \mid \theta_j^{[m+0.5]} = 0\}$ and solve the following optimization problem

$$\min_{\boldsymbol{\theta}} \mathcal{F}_\lambda(\boldsymbol{\theta}) \quad \text{subject to } \boldsymbol{\theta}_{\mathcal{A}^\perp} = \mathbf{0} \tag{3.5}$$

using the cyclic coordinate descent algorithm (CCDA) initiated by $\boldsymbol{\theta}^{[m+0.5]}$. For notational simplicity, we omit the stage index K and iteration index m, and only consider the iteration index t of CCDA. Suppose that the CCDA algorithm takes some initial solution $\boldsymbol{\theta}^{(0)}$ for solving (2.1) with the regularization parameter λ. Without loss of generality, we denote $\mathcal{A} = \{1,...,|\mathcal{A}|\}$. At the t^{th} iteration, we have $\boldsymbol{\theta}^{(t)}$. Then at the $(t+1)^{\text{th}}$ iteration, we conduct the coordinate minimization cyclically over all active coordinates. Let $\boldsymbol{w}^{(t+1,k)}$ be an auxiliary solution of the $(t+1)^{\text{th}}$ iteration with the first $k-1$ coordinates updated. For $k = 1$, we have $\boldsymbol{w}^{(t+1,1)} = \boldsymbol{\theta}^{(t)}$. We then update the k^{th} coordinate to obtain the next auxiliary solution $\boldsymbol{w}^{(t+1,k+1)}$.

More specifically, let $\nabla_k \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\theta})$ be the k^{th} entry of $\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\theta})$. We minimize the objective function with respect to each selected coordinate and keep all other coordinates fixed,

$$w_k^{(t+1,k+1)} = \underset{\theta_k \in \mathbb{R}}{\operatorname{argmin}} \mathcal{L}_\lambda(\theta_k; \boldsymbol{w}_{\setminus k}^{(t+1,k)}) + r_\lambda(\theta_k). \tag{3.6}$$

Once we obtain $w_k^{(t+1,k+1)}$, we can set $\boldsymbol{w}_{\setminus k}^{(t+1,k+1)} = \boldsymbol{w}_{\setminus k}^{(t+1,k)}$ to obtain the next auxiliary solution $\boldsymbol{w}^{(t+1,k+1)}$. For sparse column inverse operator, let $\widetilde{w}_k^{(t+1,k)} = e_k - \widehat{\mathbf{S}}_{\setminus kk}^T \boldsymbol{w}_{\setminus k}^{(t+1,k)}$, we have

$$\begin{aligned} w_k^{(t+1,k+1)} &= \underset{\theta_k \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{2}\widehat{\mathbf{S}}_{kk}\theta_k^2 + e_k - \widehat{\mathbf{S}}_{\setminus kk}^T \boldsymbol{w}_{\setminus k}^{(t+1,k)}\theta_k - e_k\theta_k + r_\lambda(\theta_k) \\ &= \underset{\theta_k \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{2}\left(\theta_k - \widetilde{w}_k^{(t+1,k)}\right)^2 + r_\lambda(\theta_k), \end{aligned} \tag{3.7}$$

where the last equality comes from the fact $\widehat{\mathbf{S}}_{kk} = 1$ for all $k = 1,...,d$. By setting the subgradient of (3.7) equal to zero, we can obtain $w_k^{(t+1,k+1)}$ as follows:

- For the ℓ_1 norm regularization, we have $w_k^{(t+1,k+1)} = \mathcal{S}_\lambda(\widetilde{w}_k^{(t+1,k)})$.
- For the SCAD regularization, we have

$$w_k^{(t+1,k+1)} = \begin{cases} \widetilde{w}_k^{(t+1,k)} & \text{if } |\widetilde{w}_k^{(t+1,k)}| \geq \gamma\lambda, \\ \dfrac{\mathcal{S}_{\gamma\lambda/(\gamma-1)}(\widetilde{w}_k^{(t+1,k)})}{1-1/(\gamma-1)} & \text{if } |\widetilde{w}_k^{(t+1,k)}| \in [2\lambda, \gamma\lambda), \\ \mathcal{S}_\lambda(\widetilde{w}_k^{(t+1,k)}) & \text{if } |\widetilde{w}_k^{(t+1,k)}| < 2\lambda. \end{cases}$$

- For the MCP regularization, we have

$$w_k^{(t+1,k+1)} = \begin{cases} \widetilde{w}_k^{(t+1,k)} & \text{if } |\widetilde{w}_k^{(t+1,k)}| \geq \gamma\lambda, \\ \dfrac{\mathcal{S}_\lambda(\widetilde{w}_k^{(t+1,k)})}{1-1/\gamma} & \text{if } |\widetilde{w}_k^{(t+1,k)}| < \gamma\lambda. \end{cases}$$

When all $|\mathcal{A}|$ coordinate updates in the $(t+1)^{\text{th}}$ iteration of CCDA finish, we set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{w}^{(t+1,|\mathcal{A}|+1)}$. We summarize CCDA in Algorithm 2. Once CCDA terminates, we denote its output solution by $\boldsymbol{\theta}^{[m+1]}$, and start the next iteration of AISTA. We summarize AISTA in Algorithm 3.

Algorithm 2 The cyclic coordinate descent algorithm (CCDA). The cyclic coordinate descent algorithm cyclically iterates over the support of the initial solution. Without loss of generality, we assume $\mathcal{A} = \{1, ..., |\mathcal{A}|\}$.

Algorithm: $\widehat{\boldsymbol{\theta}} \leftarrow CCDA(\lambda, \boldsymbol{\theta}^{(0)})$.
Initialize: $t \leftarrow 0$, $\mathcal{A} = \text{supp}(\boldsymbol{\theta}^{(0)})$
Repeat:
  $\boldsymbol{w}^{(t+1,1)} \leftarrow \boldsymbol{\theta}^{(t)}$
  For $k = 1, ..., |\mathcal{A}|$
    $w_k^{(t+1,k+1)} \leftarrow \text{argmin}_{\theta_k \in \mathbb{R}} \mathcal{L}_\lambda(\theta_k; \boldsymbol{w}_{\setminus k}^{(t+1,k)}) + r_\lambda(\theta_k)$ and $\boldsymbol{w}_{\setminus k}^{(t+1,k+1)} \leftarrow \boldsymbol{w}_{\setminus k}^{(t+1,k)}$
  End for
  $\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{w}^{(t+1,|\mathcal{A}|+1)}$, $t \leftarrow t+1$
Until convergence
$\widehat{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}^{(t)}$

Algorithm 3 The accelerated iterative shrinkage thresholding algorithm (AISTA). Within each iteration, we exploit an additional coordinate descent subroutine to improve the empirical computational performance.

Algorithm: $\{\widehat{\boldsymbol{\theta}}, \widehat{L}\} \leftarrow AISTA(\lambda, \boldsymbol{\theta}^{[0]}, L^{[0]})$
Initialize: $m \leftarrow 0$
Repeat:
  $z \leftarrow 0$
  Repeat:
    $L^{[m+1]} \leftarrow 2^z L^{[m]}$, $\boldsymbol{\theta}^{[m+0.5]} \leftarrow \mathcal{T}_{\lambda,\Omega,L^{[m+1]}}(\boldsymbol{\theta}^{[m]})$, $z \leftarrow z+1$
  Until: $\mathcal{Q}_{\lambda,L^{[m+1]}}(\boldsymbol{\theta}^{[m+0.5]}; \boldsymbol{\theta}^{[m]}) \geq \mathcal{F}_\lambda(\boldsymbol{\theta}^{[m+1]})$
  $\boldsymbol{\theta}^{[m+1]} \leftarrow CCDA(\lambda, \boldsymbol{\theta}^{[m+0.5]})$, $m \leftarrow m+1$
Until convergence
$\widehat{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}^{[m-0.5]}$, $\widehat{L} \leftarrow L^{[m]}$
Output: $\{\widehat{\boldsymbol{\theta}}, \widehat{L}\}$

Remark 3.1. The backtracking line search procedure in PISTA has been extensively studied in existing optimization literature on the adaptive step size selection (Dennis and Schnabel, 1983; Nocedal and Wright, 2006), especially for proximal gradient algorithms (Beck and Teboulle, 2009b,a; NESTEROV, 2013). Many empirical results have corroborated better computational performance than that using a fix step size. But unlike the classical proximal gradient algorithms, APISTA can efficiently reduce the objective value by the coordinate descent subroutine in each iteration. Therefore we can simply choose a constant step size parameter L such that

$$L \geq \sup_{\boldsymbol{\theta} \in \mathbb{R}^d} \Lambda_{\max}(\nabla^2 \mathcal{L}(\boldsymbol{\theta})). \tag{3.8}$$

The step size parameter L in (3.8) guarantees $\mathcal{Q}_{\lambda,L}(\boldsymbol{\theta}; \boldsymbol{\theta}^{[m]}) \geq \mathcal{F}_\lambda(\boldsymbol{\theta})$ in each iteration of AISTA. For sparse column inverse operator, $\nabla^2 \mathcal{L}(\boldsymbol{\theta}) = \widehat{\mathbf{S}}$ does not depend on $\boldsymbol{\theta}$. Therefore we choose $L = \Lambda_{\max}(\widehat{\mathbf{S}})$. Our numerical experiments show that choosing a fixed step not only simplifies the implementation, but also attains better empirical computational performance than the backtracking line search. See more details in §5.

3.3 Stopping Criteria

Since $\boldsymbol{\theta}$ is a local minimum if and only if the KKT condition $\min_{\xi \in \partial \|\boldsymbol{\theta}\|_1} \|\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\theta}) + \lambda \xi\|_\infty = 0$ holds, we terminate AISTA when

$$\omega_\lambda(\boldsymbol{\theta}^{[m+0.5]}) = \min_{\xi \in \partial \|\boldsymbol{\theta}^{[m+0.5]}\|_1} \|\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\theta}^{[m+0.5]}) + \lambda \xi\|_\infty \leq \varepsilon, \tag{3.9}$$

where ε is the target precision and usually proportional to the regularization parameter. More specifically, given the regularization parameter λ_K, we have

$$\varepsilon_K = \delta_K \lambda_K \quad \text{for } K = 1, \ldots, N, \tag{3.10}$$

where $\delta_K \in (0,1)$ is a convergence parameter for the K^{th} stage of the path-following optimization scheme. Moreover, for CCDA, we terminate the iteration when

$$\|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}\|_2^2 \leq \delta_0^2 \lambda^2, \tag{3.11}$$

where $\delta_0 \in (0,1)$ is a convergence parameter. This stopping criterion is natural to the sparse coordinate descent algorithm, since we only need to calculate the value change of each coordinate (not the gradient). We will discuss how to choose δ_K's and δ_0 in §4.1.

4 Theory

Before we present the computational and statistical theories of APISTA, we introduce some additional assumptions. The first one is about the choice of regularization parameters.

Assumption 4.1. Let δ_K's and η saitsify

$$\eta \in [0.9, 1) \quad \text{and} \quad \max_{0 \le K \le N} \delta_K \le \delta_{\max} = 1/4,$$

where η is the rescaling parameter of the path-following optimization scheme, δ_K's are the convergence parameters defined in (3.10), and δ_0 is the convergence parameter defined in (3.11). We have the regularization parameters

$$\lambda_0 > \lambda_1 ... > \lambda_N \ge 8\|\nabla \mathcal{L}(\boldsymbol{\theta}^*)\|_\infty.$$

Assumption 4.1 has been extensively studied in existing literature on high dimensional statistical theory of the regularized M-estimators (Rothman et al., 2008; Zhang and Huang, 2008; Negahban and Wainwright, 2011; Negahban et al., 2012). It requires the regularization parameters to be large enough such that irrelevant variables can be eliminated along the solution path. Though $\|\nabla \mathcal{L}(\boldsymbol{\theta}^*)\|_\infty$ cannot be explicitly calculated ($\boldsymbol{\theta}^*$ is unknown), we can exploit concentration inequalities to show that Assumption 4.1 holds with high probability (Ledoux, 2005). In particular, we will verify Assumption 4.1 for sparse transellpitical graphical model estimation in Lemma 4.8.

Before we proceed with our second assumption, we define the largest and smallest s-sparse eigenvalues of the Hessian matrix of the loss function as follows.

Definition 4.1. Given an integer $s \ge 1$, we define the largest and smallest s-sparse eigenvalues of $\nabla^2 \mathcal{L}(\boldsymbol{\theta})$ as

$$\text{Largest } s\text{-Sparse Eigenvalue} : \quad \rho_+(s) = \sup_{\boldsymbol{v} \in \mathbb{R}^d,\, \|\boldsymbol{v}\|_0 \le s} \frac{\boldsymbol{v}^T \nabla^2 \mathcal{L}(\boldsymbol{\theta}) \boldsymbol{v}}{\|\boldsymbol{v}\|_2^2},$$

$$\text{Smallest } s\text{-Sparse Eigenvalue} : \quad \rho_-(s) = \inf_{\boldsymbol{v} \in \mathbb{R}^d,\, \|\boldsymbol{v}\|_0 \le s} \frac{\boldsymbol{v}^T \nabla^2 \mathcal{L}(\boldsymbol{\theta}) \boldsymbol{v}}{\|\boldsymbol{v}\|_2^2}.$$

Moreover, we define $\widetilde{\rho}_-(s) = \rho_-(s) - \alpha$ and $\rho_+(s) = \rho_+(s)$ for notational simplicity, where α is defined in (2.2).

The next lemma shows the connection between the sparse eigenvalue conditions and restricted strongly convex and smooth conditions.

Lemma 4.1. Given $\rho_-(s) > 0$, for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$ with $|\mathrm{supp}(\boldsymbol{\theta}) \cup \mathrm{supp}(\boldsymbol{\theta}')| \le s$, we have

$$\mathcal{L}(\boldsymbol{\theta}') \le \mathcal{L}(\boldsymbol{\theta}) + (\boldsymbol{\theta}' - \boldsymbol{\theta})^T \nabla \mathcal{L}(\boldsymbol{\theta}) + \frac{\rho_+(s)}{2} \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_2^2,$$

$$\mathcal{L}(\boldsymbol{\theta}') \ge \mathcal{L}(\boldsymbol{\theta}) + (\boldsymbol{\theta}' - \boldsymbol{\theta})^T \nabla \mathcal{L}(\boldsymbol{\theta}) + \frac{\rho_-(s)}{2} \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_2^2.$$

Moreover, if $\rho_-(s) > \alpha$, then we have

$$\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\theta}') \le \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\theta}) + (\boldsymbol{\theta}' - \boldsymbol{\theta})^T \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\theta}) + \frac{\rho_+(s)}{2} \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_2^2,$$

$$\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\theta}') \ge \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\theta}) + (\boldsymbol{\theta}' - \boldsymbol{\theta})^T \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\theta}) + \frac{\widetilde{\rho}_-(s)}{2} \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_2^2,$$

and for any $\xi \in \partial\|\theta\|_1$,

$$\mathcal{F}_\lambda(\theta') \geq \mathcal{F}_\lambda(\theta) + (\nabla\widetilde{\mathcal{L}}_\lambda(\theta) + \lambda\xi)^T(\theta' - \theta) + \frac{\widetilde{\rho}_-(s)}{2}\|\theta' - \theta\|_2^2.$$

The proof of Lemma 4.1 is provided in Wang et al. (2014), therefore omitted. We then introduce the second assumption.

Assumption 4.2. Given $\|\theta^*\|_0 \leq s^*$, there exists an integer $\widetilde{s}$ satisfying

$$\widetilde{s} \geq (144\kappa^2 + 250\kappa)\cdot s^*, \;\; \rho_+(s^* + 2\widetilde{s}) < +\infty, \;\; \text{and} \;\; \widetilde{\rho}_-(s^* + 2\widetilde{s}) > 0,$$

where $\kappa = \rho_+(s + 2\widetilde{s})/\widetilde{\rho}_-(s + 2\widetilde{s})$.

Assumption 4.2 requires that $\widetilde{\mathcal{L}}_\lambda(\theta)$ satisfies the strong convexity and smoothness when θ is sparse. As will be shown later, APISTA can always guarantee the number of irrelevant coordinates with nonzero values not to exceed $\widetilde{s}$. Therefore the restricted strong convexity is preserved along the solution path. We will verify that Assumption 4.2 holds with high probability for the transellpitical graphical model estimation in Lemma 4.9.

Remark 4.2 (Step Size Initialization)**.** We take the initial step size parameter as $L_{\min} \geq \rho_+(1)$. For sparse column inverse operator, we directly choose $L_{\min} = \rho_+(1) = 1$.

4.1 Computational Theory

We develop the computational theory of APISTA. For notational simplicity, we define $\mathcal{S} = \{j \mid \theta_j^* \neq 0\}$ and $\mathcal{S}^\perp = \{j \mid \theta_j^* = 0\}$ for characterizing the the solution sparsity. We first start with the convergence analysis for the cyclic coordinate descent algorithm (CCDA). The next theorem presents its rate of convergence in term of the objective value.

Theorem 4.3 (Geometric Rate of Convergence of CCDA)**.** Suppose that Assumption 4.2 holds. Given a sparse initial solution satisfying $\|\theta^{(0)}_{\mathcal{S}^\perp}\|_0 \leq \widetilde{s}$, (3.5) is a strongly convex optimization problem with a unique global minimizer $\bar{\theta}$. Moreover, for $t = 1, 2...$, we have

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

The proof of Theorems 4.3 is provided in Appendix A. Theorem 4.3 suggests that when the initial solution is sparse, CCDA essentially solves a strongly convex optimization problem with a unique global minimizer. Consequently we can establish the geometric rate of convergence in term of the objective value for CCDA. We then proceed with the convergence analysis of AISTA. The next theorem presents its theoretical rate of convergence in term of the objective value.

Theorem 4.4 (Geometric Rate of Convergence of AISTA)**.** Suppose that Assumptions 4.1 and 4.2 hold. For any $\lambda \geq \lambda_N$, if the initial solution $\theta^{[0]}$ satisfies

$$\|\theta^{[0]}_{\mathcal{S}^\perp}\|_0 \leq \widetilde{s}, \;\; \omega_\lambda(\theta^{[0]}) \leq \lambda/2, \tag{4.1}$$

then we have $\|\boldsymbol{\theta}^{[m]}_{\mathcal{S}^\perp}\|_0 \leq \widetilde{s}$ for $m = 0.5, 1, 1.5, 2, \ldots$. Moreover, for $m = 1, 2, \ldots$, we have

$$\mathcal{F}_\lambda(\boldsymbol{\theta}^{[m]}) - \mathcal{F}_\lambda(\bar{\boldsymbol{\theta}}^\lambda) \leq \left(1 - \frac{1}{8\kappa}\right)^m [\mathcal{F}_\lambda(\boldsymbol{\theta}^{(0)}) - \mathcal{F}_\lambda(\bar{\boldsymbol{\theta}})],$$

where $\bar{\boldsymbol{\theta}}^\lambda$ is a unique sparse local solution to (2.1) satisfying $\omega_\lambda(\bar{\boldsymbol{\theta}}^\lambda) = 0$ and $\|\bar{\boldsymbol{\theta}}^\lambda_{\mathcal{S}^\perp}\|_0 \leq \widetilde{s}$.

The proof of Theorem 4.4 is provided in Appendix B. Theorem 4.4 suggests that all solutions of AISTA are sparse such that the restricted strongly convex and smooth conditions hold for all iterations. Therefore, AISTA attains the geometric rate of convergence in term of the objective value. Theorem 4.4 requires a proper initial solution to satisfy (4.1). This can be verified by the following theorem.

Theorem 4.5 (Path-following Optimization Scheme)**.** Suppose that Assumptions 4.1 and 4.2 hold. Given $\boldsymbol{\theta}$ satisfying

$$\|\boldsymbol{\theta}_{\mathcal{S}^\perp}\|_0 \leq s \quad \text{and} \quad \omega_{\lambda_{K-1}}(\boldsymbol{\theta}) \leq \delta_{K-1}\lambda_{K-1}, \tag{4.2}$$

we have $\omega_{\lambda_K}(\boldsymbol{\theta}) \leq \lambda_K/2$.

The proof of Theorem 4.5 is provided in Wang et al. (2014), therefore omitted. Since $\boldsymbol{\theta}^{\{0\}}$ naturally satisfies (4.2) for λ_1, by Theorem 4.5 and induction, we can show that the path-following optimization scheme always guarantees that the output solution of the $(K-1)^{\text{th}}$ stage is a proper initial solution for the K^{th} stage, where $K = 1, \ldots, N$. Eventually, we combine Theorems 4.3 and 4.4 with Theorem 4.5, and establish the global geometric rate of convergence in term of the objective value for APISTA in the next theorem.

Theorem 4.6 (Global Geometric Rate of Convergence of APISTA)**.** Suppose that Assumptions 4.1 and 4.2 hold. Recall that δ_0 and δ_K's are defined in §3.3, κ and $\widetilde{s}$ are defined in Assumption 4.2, and α is defined in (2.2). We have the following results:

(1) At the K^{th} stage ($K = 1, \ldots, N$), the number of coordinate descent iterations within each CCDA is at most $C_1 \log(C_2/\delta_0)$, where

$$C_1 = 2\log^{-1}\left(\frac{(s^* + \widetilde{s})\rho_+^2(s^* + \widetilde{s})}{(s^* + \widetilde{s})\rho_+^2(s^* + \widetilde{s}) + \widetilde{\rho}_-(1)\widetilde{\rho}_-(s^* + \widetilde{s})}\right) \quad \text{and} \quad C_2 = \sqrt{\frac{21s^*}{\widetilde{\rho}_-(s^* + \widetilde{s})\widetilde{\rho}_-(1)}};$$

(2) At the K^{th} stage ($K = 1, \ldots, N$), the number of the proximal gradient iterations in each AISTA is at most $C_3 \log(C_4/\delta_K)$, where

$$C_3 = 2\log^{-1}\left(1 - \frac{1}{8\kappa}\right) \quad \text{and} \quad C_4 = 10\sqrt{\kappa s^*};$$

(3) To compute all $N + 1$ output solutions, the total number of coordinate descent iterations in APISTA is at most

$$C_1 \log(C_2/\delta_0) \sum_{K=1}^{N} C_3 \log(C_4/\delta_K); \tag{4.3}$$

(4) At the K^{th} stage ($K = 1,...,N$), we have

$$\mathcal{F}_{\lambda_N}(\widehat{\boldsymbol{\theta}}^{[K]}) - \mathcal{F}_{\lambda_N}(\bar{\boldsymbol{\theta}}^{\lambda_N}) \leq [\mathbb{1}(K < N) + \delta_K] \cdot \frac{105\lambda_K^2 s^*}{\widetilde{\rho}_-(s^* + \widetilde{s})};$$

The proof Theorem 4.6 is provided in Appendix C. We then present a more intuitive explanation about Result (3). To secure the generalization performance in practice, we usually tune the regularization parameter over a refined sequence based on cross validation. In particular, we solve (2.1) using partial data with high precision for every regularization parameter. If we set $\delta_K = \delta_{\text{opt}}\lambda_K$ for $K = 1,...N$, where δ_{opt} is a very small value (e.g. 10^{-8}), then we can rewrite (4.3) as

$$NC_1 \log\left(\frac{C_2}{\delta_0}\right) C_3 \log\left(\frac{C_4}{\delta_{\text{opt}}}\right) = \mathcal{O}\left(N \log\left(\frac{1}{\delta_{\text{opt}}}\right)\right), \tag{4.4}$$

where δ_0 is some reasonably large value (e.g. 10^{-2}) defined in §3.3. The iteration complexity in (4.4) depends on N.

Once the regularization parameter is selected, we still need to solve (2.1) using full data with some regularization sequence. But we only need high precision for the selected regularization parameter (e.g., λ_{N}), and for $K = 1,...,N-1$, we only solve (2.1) for λ_K up to an adequate precision, e.g., $\delta_K = \delta_0$ for $K = 1,...,N-1$ and $\delta_N = \delta_{\text{opt}}\lambda_N$. Since $1/\delta_{\text{opt}}$ is much larger than N, we can rewrite (4.3) as

$$C_1 \log\left(\frac{C_2}{\delta_0}\right)\left((N-1)C_3 \log\left(\frac{C_4}{\delta_0}\right) + C_3 \log\left(\frac{C_4}{\delta_{\text{opt}}}\right)\right) = \mathcal{O}\left(\log\left(\frac{1}{\delta_{\text{opt}}}\right)\right). \tag{4.5}$$

Now the iteration complexity in (4.5) does not depend on N.

Remark 4.7. To establish computational theories of APISTA with a fixed step size, we only need to slightly modify the proofs of Theorems 4.4 and 4.6 by replacing $\rho_+(s^* + 2\widetilde{s})$ and $\rho_+(s^* + \widetilde{s})$ by their upper bound L defined in (3.8). Then a global geometric rate of convergence can also be derived, but with a worse constant term.

4.2 Statistical Theory

We then establish the statistical theory of the SCIO estimator obtained by APISTA under transelliptical models. We use $\boldsymbol{\Theta}^*$ and Σ^* to denote the true latent precision and covariance matrices. We assume that $\boldsymbol{\Theta}^*$ belongs to the following class of sparse, positive definite, and symmetric matrices:

$$\mathcal{U}_{\psi_{\max},\psi_{\min}}(M, s^*) = \Big\{\boldsymbol{\Theta} \in \mathbb{R}^{d\times d} \;\Big|\; \boldsymbol{\Theta} = \boldsymbol{\Theta}^T,\ \max_j \|\boldsymbol{\Theta}_{*j}\|_0 \leq s^*,$$
$$\|\boldsymbol{\Theta}\|_1 \leq M,\ 0 < \psi_{\max}^{-1} \leq \Lambda_{\min}(\boldsymbol{\Theta}) \leq \Lambda_{\max}(\boldsymbol{\Theta}) \leq \psi_{\min}^{-1} < \infty\Big\},$$

where $\psi_{\max}$ and $\psi_{\min}$ are positive constants, and do not scale with (M, s^*, n, d). Since $\Sigma^* = (\boldsymbol{\Theta}^*)^{-1}$, we also have $\psi_{\min} \leq \Lambda_{\min}(\Sigma^*) \leq \Lambda_{\max}(\Sigma^*) \leq \psi_{\max}$.

We first verify Assumptions 4.1 and 4.2 in the next two lemmas for transelliptical models.

Lemma 4.8. Suppose that $\boldsymbol{X} \sim TE_d(\Sigma^*, \xi, \{f_j\}_{j=1}^d)$. Given $\lambda_N = 8\sqrt{2\pi}M\sqrt{\log d/n}$, we have

$$\mathbb{P}(\lambda_N \geq 8\|\nabla \mathcal{L}(\boldsymbol{\theta}^*)\|_\infty) \geq 1 - \frac{1}{d^2}.$$

The proof of Lemma 4.8 is provided in Appendix D. Lemma 4.8 guarantees that the selected regularization parameter λ_N satisfies Assumption 4.1 with high probability.

Lemma 4.9. Suppose that $\boldsymbol{X} \sim TE_d(\Sigma^*, \xi, \{f_j\}_{j=1}^d)$. Given $\alpha = \psi_{\min}/2$, there exist universal positive constants c_1 and c_2 such that for $n \geq 4\psi_{\min}^{-1}c_2(1+2c_1)s^*\log d$, with probability at least $1 - 2/d^2$, we have

$$\widetilde{s} = c_1 s^* \geq (144\kappa^2 + 250\kappa)s^*, \quad \widetilde{\rho}_-(s^* + 2\widetilde{s}) \geq \frac{\psi_{\min}}{4}, \quad \rho_+(s^* + 2\widetilde{s}) \leq \frac{5\psi_{\max}}{4},$$

where κ is defined in Assumption 4.2.

The proof of Lemma 4.9 is provided in Appendix E. Lemma 4.9 guarantees that if the Lipschitz constant of h'_λ defined in (2.2) satisfies $\alpha = \psi_{\min}/2$, then the transformed Kendall's tau estimator $\widehat{\mathbf{S}} = \nabla^2 \mathcal{L}(\boldsymbol{\theta})$ satisfies Assumption 4.2 with high probability.

Remark 4.10. Since Assumptions 4.1 and 4.2 have been verified, by Theorem 4.6, we know that APISTA attains the geometric rate of convergence to a unique sparse local solution to (2.3) in term of the objective value with high probability.

Recall that we use $\boldsymbol{\theta}$ to denote $\boldsymbol{\Theta}_{*j}$ in (2.4), by solving (2.3) with respect to all d columns, we obtain $\widehat{\boldsymbol{\Theta}}^{\{N\}} = \left[\widehat{\boldsymbol{\Theta}}_{*1}^{\{N\}}, ..., \widehat{\boldsymbol{\Theta}}_{*d}^{\{N\}}\right]$ and $\overline{\boldsymbol{\Theta}}^{\lambda_N} = \left[\overline{\boldsymbol{\Theta}}_{*1}^{\lambda_N}, ..., \overline{\boldsymbol{\Theta}}_{*d}^{\lambda_N}\right]$, where $\overline{\boldsymbol{\Theta}}_{*j}^{\lambda_N}$ denotes the output solution of APISTA corresponding to λ_N for the j^{th} column ($j = 1, ..d$), and $\overline{\boldsymbol{\Theta}}_{*j}^{\lambda_N}$ to denote the unique sparse local solution corresponding to λ_N for the j^{th} column ($j = 1, ..d$), which APISTA converges to. We then present concrete rates of convergence of the estimator obtained by APISTA under the matrix ℓ_1 and Frobenius norms in the following theorem.

Theorem 4.11. [Parameter Estimation] Suppose that $\boldsymbol{X} \sim TE_d(\Sigma^*, \xi, \{f_j\}_{j=1}^d)$, and $\alpha = \psi_{\min}/2$. For $n \geq 4\psi_{\min}^{-1}c_2(1+2c_1)s^*\log d$, given $\lambda_N = 8\sqrt{2}\pi M\sqrt{\log d/n}$, we have

$$\|\widehat{\boldsymbol{\Theta}}^{\{N\}} - \boldsymbol{\Theta}^*\|_1 = O_P\left(Ms^*\sqrt{\frac{\log d}{n}}\right), \quad \frac{1}{d}\|\widehat{\boldsymbol{\Theta}}^{\{N\}} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}^2 = O_P\left(\frac{M^2 s^* \log d}{n}\right).$$

The proof of (4.11) is provided in Appendix F. The results in Theorem 4.11 show that the SCIO estimator obtained by APISTA achieves the same rates of convergence as those for subguassian distributions (Liu and Luo, 2015). Moreover, when using the nonconvex regularization such as MCP and SCAD, we can achieve graph estimation consistency under the following assumption.

Assumption 4.3. Suppose that $\boldsymbol{X} \sim TE_d(\Sigma^*, \xi, \{f_j\}_{j=1}^d)$. Define $\mathcal{E}^* = \{(k,j) \mid \boldsymbol{\Theta}_{kj}^* \neq 0\}$ as the support of $\boldsymbol{\Theta}^*$. There exists some universal constant c_3 such that

$$\min_{(k,j)\in\mathcal{E}^*} |\boldsymbol{\Theta}_{kj}^*| \geq c_3 M \cdot \sqrt{\frac{s^*\log d}{n}}.$$

Assumption 4.3 is a sufficient condition for sparse column inverse operator to achieve graph estimation consistency in high dimensions for transelliptical models. The violation of Assumption 4.3 may result in underselection of the nonzero entries in $\boldsymbol{\Theta}^*$.

The next theorem shows that, with high probability, $\overline{\boldsymbol{\Theta}}^{\lambda_N}$ and the oracle solution $\widehat{\boldsymbol{\Theta}}^{\mathrm{o}}$ are identical. More specifically, let $\mathcal{S}_j = \mathrm{supp}(\boldsymbol{\Theta}^*_{*j})$ for $j = 1, ..., d$, $\widehat{\boldsymbol{\Theta}}^{\mathrm{o}} = \big[\widehat{\boldsymbol{\Theta}}^{\mathrm{o}}_{*1}, ..., \widehat{\boldsymbol{\Theta}}^{\mathrm{o}}_{*d}\big]$ defined as follows,

$$\widehat{\boldsymbol{\Theta}}^{\mathrm{o}}_{\mathcal{S}_j j} = \underset{\boldsymbol{\Theta}_{\mathcal{S}_j j} \in \mathbb{R}^{|\mathcal{S}_j|}}{\mathrm{argmin}} \frac{1}{2} \boldsymbol{\Theta}^T_{\mathcal{S}_j j} \widehat{\mathbf{S}}_{\mathcal{S}_j \mathcal{S}_j} \boldsymbol{\Theta}_{\mathcal{S}_j j} - \mathbf{I}^T_{\mathcal{S}_j j} \boldsymbol{\Theta}_{\mathcal{S}_j j} \quad \text{and} \quad \widehat{\boldsymbol{\Theta}}^{\mathrm{o}}_{\mathcal{S}^{\perp}_j j} = \mathbf{0} \text{ for } j = 1, ..., d. \tag{4.6}$$

Theorem 4.12. [Graph Estimation] Suppose that $\boldsymbol{X} \sim TE_d(\Sigma^*, \xi, \{f_j\}_{j=1}^d)$, $\alpha = \psi_{\min}/2$, and Assumption 4.3 holds. There exists a universal constant c_4 such that $n \geq 4\psi^{-1}_{\min} c_2(1 + 2c_1)s^* \log d$, if we choose $\lambda_N = c_4\sqrt{2}\pi M\sqrt{s^* \log d/n}$, then we have

$$\mathbb{P}(\overline{\boldsymbol{\Theta}}^{\lambda_N} = \widehat{\boldsymbol{\Theta}}^{\mathrm{o}}) \geq 1 - \frac{3}{d^2}.$$

The proof of Theorem 4.12 is provided in Appendix G. Since $\widehat{\boldsymbol{\Theta}}^{\mathrm{o}}$ shares the same support with $\boldsymbol{\Theta}^*$, Theorem 4.12 guarantees that the SCIO estimator obtained by APISTA can perfectly recover $\mathcal{E}^*$ with high probability. To the best of our knowledge, Theorem 4.12 is the first graph estimation consistency result for transelliptical models without any post-processing procedure (e.g. thresholding).

Remark 4.13. In Theorem (4.12), we choose $\lambda_N = c_4\sqrt{2}\pi M\sqrt{\log d/n}$, which is different from the selected regularization parameter in Assumption 4.8. But as long as we have $c_4\sqrt{s^*} \geq 8$, which is not an issue under the high dimensional scaling

$$M, s^*, n, d \to \infty \quad \text{and} \quad Ms^* \log d/n \to 0,$$

$\lambda_N \geq 8\|\nabla \mathcal{L}(\boldsymbol{\theta}^*)\|_\infty$ still holds with high probability. Therefore all computational theories in §4.1 hold for $\overline{\boldsymbol{\Theta}}^{\lambda_N}$ in Theorem 4.12.

5 Numerical Experiments

In this section, we study the computational and statistical performance of APISTA method through numerical experiments on sparse transelliptical graphical model estimation. All experiments are conducted on a personal computer with Intel Core i5 3.3 GHz and 16GB memory. All programs are coded in double precision `C`, called from `R`. The computation are optimized by exploiting the sparseness of vector and matrices. Thus we can gain a significant speedup in vector and matrix manipulations (e.g. calculating the gradient and evaluating the objective value). We choose the MCP regularization with varying β's for all simulations.

5.1 Simulated Data

We consider the chain and Erdös-Rényi graph generation schemes with varying $d = 200$, 400, and 800 to obtain the latent precision matrices:

- **Chain**. Each node is assigned a coordinate j for $j = 1, ..., d$. Two nodes are connected by an edge whenever the corresponding points are at distance no more than 1.
- **Erdös-Rényi**. We set an edge between each pair of nodes with probability $1/d$, independently of the other edges.

Two illustrative examples are presented in Figure 5.1. Let $\mathcal{D}$ be the adjacency matrix of the generated graph, and $\mathcal{M}_2$ be the rescaling operator that converts a symmetric positive semidefinite matrix to a correlation matrix. We calculate

$$\Sigma^* = \mathcal{M}_2[(\widetilde{\mathcal{D}} + (1 - \Lambda_{\min}(\mathcal{D}))\mathbf{I})^{-1}].$$

We use Σ^* as the covariance matrix to generate $n = \lceil 60 \log d \rceil$ independent observations from a multivariate t-distribution with mean $\mathbf{0}$ and degrees of freedom 3. We then adopt the power transformation $g(t) = t^5$ to convert to the t-distributed data to the transelliptical data. Note that the corresponding latent precision matrix is $\mathbf{\Omega}^* = (\Sigma^*)^{-1}$. We compare the following five computational methods:

(1) APISTA: The computational algorithm proposed in §3.

(2) F-APISTA: APISTA without the backtracking line search (using a fixed step size instead).

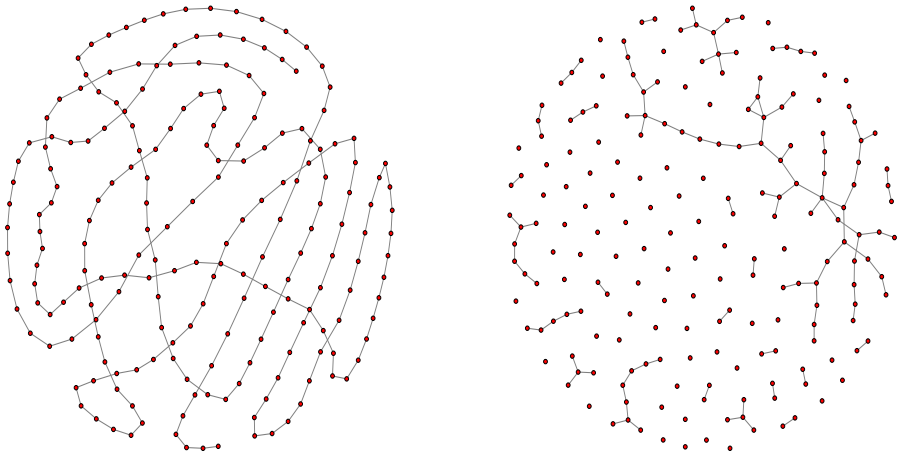
(3) PISTA: The pathwise iterative shrinkage thresholding algoritm proposed in Wang et al. (2014).

(4) CLIME: The sparse latent precision matrix estimation method proposed in Liu et al. (2012b), which solves (2.5) by the ADMM method (Alternating Direction Method of Multipliers, Li et al. (2015); Liu et al. (2014)).

(5) SCIO(P): The SCIO estimator based on the positive semidefinite projection method proposed in Zhao et al. (2014b). More specifically, we first project the possibly indefinite Kendall's tau matrix into the cone of all positive semidefinite matrices. Then we plug the obtained replacement into (2.3), and solve it by the coordinate descent method proposed in Liu and Luo (2015).

Note that (4) and (5) have theoretical guarantees only when the ℓ_1 norm regularization is applied. For (1)-(3), we set $\delta_0 = \delta_K = 10^{-5}$ for $K = 1, ..., N$.

We first compare the statistical performance in parameter estimation and graph estimation of all methods. To meet this end, we generate a validation set of the same size as the training set.



(a) Chain (b) Erdös-Rényi

Figure 5.1: Different graph patterns. To ease the visualization, we only present graphs with $d = 200$.

We use the regularization sequence with $N = 100$ and $\lambda_N = 0.5\sqrt{\log d/n} \approx 0.0645$. The optimal regularization parameter is selected by

$$\widehat{\lambda} = \underset{\lambda \in \{\lambda_1, \ldots, \lambda_N\}}{\operatorname{argmin}} \|\widehat{\mathbf{\Theta}}^{\lambda} \widetilde{\mathbf{S}} - \mathbf{I}\|_{\max},$$

where $\widehat{\mathbf{\Theta}}^{\lambda}$ denotes the estimated latent precision matrix using the training set with the regularization parameter λ, and $\widetilde{\mathbf{S}}$ denotes the estimated latent covariance matrix using the validation set. We repeat the simulation for 100 times, and summarize the averaged results in Tables 5.1 and 5.2. For all settings, we set $\delta_0 = \delta_K = 10^{-5}$. We also vary β of the MCP regularization from 100 to 20/19, thus the corresponding α varies from 0.01 to 0.95. The parameter estimation performance is evaluated by the difference between the obtained estimator and the true latent prediction matrix under the Forbenius and matrix ℓ_1 norms. The graph estimation performance is evaluated by the true positive rate (T. P. R.) and false positive rate (F. P. R.) defined as follows,

$$\text{T. P. R.} = \frac{\sum_{k \neq j} \mathbb{1}(\widehat{\mathbf{\Theta}}^{\widehat{\lambda}}_{kj} \neq 0) \cdot \mathbb{1}(\mathbf{\Theta}^*_{kj} \neq 0)}{\sum_{k \neq j} \mathbb{1}(\mathbf{\Theta}^*_{kj} \neq 0)}, \quad \text{F. P. R.} = \frac{\sum_{k \neq j} \mathbb{1}(\widehat{\mathbf{\Theta}}^{\widehat{\lambda}}_{kj} \neq 0) \cdot \mathbb{1}(\mathbf{\Theta}^*_{kj} = 0)}{\sum_{k \neq j} \mathbb{1}(\mathbf{\Theta}^*_{kj} = 0)}.$$

Since the convergence of PISTA is very slow when α is large, we only present its results for $\alpha = 0.2$. APISTA and F-APISTA can work for larger α's. Therefore they effectively reduces the estimation bias to attain the best statistical performance in both parameter estimation and graph estimation among all estimators. The SCIO(P) and CLIME methods only use ℓ_1 norm without any bias reduction, their performance is worse than the other competitors. Moreover, due to the poor scalability of their solvers, SCIO(P) and CLIME fail to output valid results within 10 hours when $d = 800$.

Table 5.1: Quantitive comparison among different estimators on the chain model. Since APISTA and F-APISTA can output valid results for large α's, their estimator attains better performance than other competitors. The SCIO(P) and CLIME estimators use the ℓ_1 norm regularization with no bias reduction. Thus their performance is worse than the other competitors in both parameter estimation and graph estimation.

Method	d	$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_{\mathrm{F}}$	$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_1$	T. P. R.	F. P. R.	α
PISTA	200	4.1112(0.7856)	1.0517(0.1141)	1.0000(0.0000)	0.0048(0.0079)	0.20
	400	6.4507(0.9062)	1.0756(0.0717)	1.0000(0.0000)	0.0007(0.0004)	0.20
	800	8.2640(1.1456)	1.0434(0.0673)	1.0000(0.0000)	0.0003(0.0006)	0.20
APISTA	200	2.5162(0.2677)	0.7665(0.1583)	0.9993(0.0012)	0.0001(0.0001)	0.95
	400	3.3664(0.2735)	0.8298(0.0986)	1.0000(0.0000)	0.0002(0.0000)	0.67
	800	5.0244(0.7984)	0.9312(0.1226)	1.0000(0.0000)	0.0002(0.0004)	0.50
F-APISTA	200	2.5163(0.2670)	0.7658(0.1559)	0.9994(0.0015)	0.0001(0.0002)	0.95
	400	3.3629(0.2702)	0.8253(0.0959)	1.0000(0.0000)	0.0002(0.0000)	0.67
	800	5.0237(0.7963)	0.9373(0.1289)	1.0000(0.0000)	0.0002(0.0005)	0.50
SCIO(P)	200	6.1812(1.2924)	1.2245(0.0777)	1.0000(0.0000)	0.0165(0.0220)	0.00
	400	8.9991(0.9894)	1.2255(0.0785)	1.0000(0.0000)	0.0058(0.0047)	0.00
CLIME	200	6.4771(0.8617)	1.2187(0.0358)	1.0000(0.0000)	0.0126(0.0043)	0.00
	400	9.1221(0.9997)	1.2177(0.0629)	1.0000(0.0000)	0.0043(0.0032)	0.00

We then compare the computational performance of all methods. We use a regularization sequence with $N = 50$, and λ_N is proper selected such that the graphs obtained by all methods have approximately the same number of edges for each regularization parameter. In particular, the obtained graphs corresponding to λ_N have approximately $0.1 \cdot d(d-1)/2$ edges. To make a fair comparison, we choose the ℓ_1 norm regularization for all methods. We repeat the simulation for 100 times, and the timing results are summarized in Tables 5.3 and 5.4. We see that F-APISTA method is up to 10 times faster than PISTA algorithm, and APISTA is up to 5 times after than PISTA. SCIO(P) and CLIME are much slower than the other three competitors.

5.2 Real Data

We present a real data example to demonstrate the usefulness of the transelliptical graph obtained by the sparse column inverse operator (based on the transformed Kendall's tau matrix). We acquire closing prices from all stocks of the S&P 500 for all the days that the market was open between January 1, 2003 and January 1, 2005, which results in 504 samples for the 452 stocks. We transform the dataset by calculating the log-ratio of the price at time $t+1$ to price at time t. The

Table 5.2: Quantitive comparison among different estimators on the Erdös-Rényi model. Since A-PISTA and F-APISTA can output valid results for large α's, their estimators attains better performance than other competitors. The SCIO(P) and CLIME estimators use the ℓ_1 norm regularization with no bias reduction. Thus their performance is worse than the other competitors in both parameter estimation and graph estimation.

Method	d	$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_{\mathrm{F}}$	$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_1$	T. P. R.	F. P. R.	$\widehat{\alpha}$
PISTA	200	3.2647(0.1235)	1.6807(0.2675)	1.0000(0.0000)	0.0587(0.0013)	0.20
	400	4.5609(0.7666)	2.2113(0.3358)	1.0000(0.0000)	0.0295(0.0091)	0.20
	800	5.0751(0.3832)	2.5718(0.2826)	1.0000(0.0000)	0.0099(0.0020)	0.20
APISTA	200	2.2888(0.1141)	1.1644(0.2343)	1.0000(0.0000)	0.0193(0.0005)	0.33
	400	3.2206(0.2733)	1.4974(0.2778)	1.0000(0.0000)	0.0067(0.0100)	0.33
	800	4.0929(0.1862)	1.6347(0.2023)	1.0000(0.0000)	0.0036(0.0008)	0.50
F-APISTA	200	2.2890(0.1161)	1.1647(0.2390)	1.0000(0.0000)	0.0197(0.0007)	0.33
	400	3.2251(0.2702)	1.4928(0.2731)	1.0000(0.0000)	0.0060(0.0102)	0.33
	800	4.0984(0.1891)	1.6397(0.2096)	1.0000(0.0000)	0.0034(0.0009)	0.50
SCIO(P)	200	3.4277(0.5405)	1.5213(0.3223)	1.0000(0.0000)	0.0618(0.0170)	0.00
	400	5.7144(0.8158)	1.9057(0.2933)	0.9994(0.0017)	0.0341(0.0145)	0.00
CLIME	200	3.6297(0.6103)	1.4876(0.2855)	1.0000(0.0000)	0.0581(0.0159)	0.00
	400	5.9206(0.8385)	1.8246(0.2817)	1.0000(0.0000)	0.0320(0.0112)	0.00

Table 5.3: Quantitive comparison of computational performance on the chain model (in seconds). We see that the F-APISTA method attains the best timing performance among all methods. The SCIO(P) and CLIME methods are much slower than the other three methods.

d	PISTA	APISTA	F-APISTA	SCIO(P)	CLIME
200	0.8342(0.0248)	0.2693(0.0031)	0.1013(0.0022)	2.6572(0.1253)	8.5932(0.5396)
400	3.8782(0.0696)	1.2103(0.0368)	0.4559(0.0308)	25.451(2.5752)	48.235(5.3494)
800	30.014(0.3514)	6.5970(0.2338)	2.4283(0.2605)	315.87(34.638)	460.12(45.121)

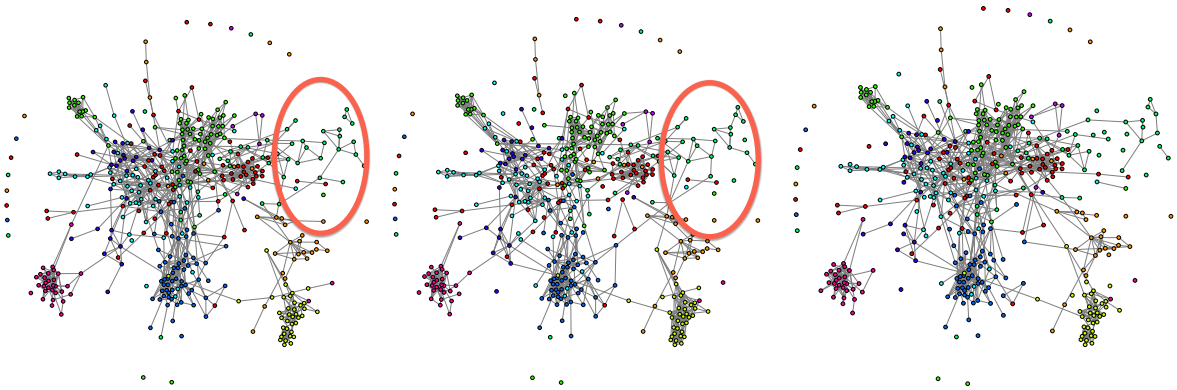
452 stocks are categorized into 10 Global Industry Classification Standard (GICS) sectors.

We adopt the stability graphs obtained by the following procedure (Meinshausen and Bühlmann, 2010; Liu et al., 2010):

(1) Calculate the graph path using all samples, and choose the regularization parameter at the

Table 5.4: Quantitive comparison of computational performance on the Erdös-Rényi model (in seconds). We see that the F-APISTA method attains the best timing performance among all methods. The SCIO(P) and CLIME methods are much slower than the other three methods.

d	PISTA	APISTA	F-APISTA	SCIO(P)	CLIME
200	0.5401(0.0248)	0.2048(0.0056)	0.1063(0.0110)	2.712(0.13558)	7.1325(0.7891)
400	3.0501(0.0829)	0.9982(0.0453)	0.4555(0.0071)	26.140(2.1503)	45.160(4.9026)
800	28.581(0.3517)	6.8417(0.7543)	2.7037(0.2145)	332.90(30.115)	442.57(50.978)



(a) Transelliptical Graphs by SCIO (L) and CLIME (R)

(b) Gaussian Graph by SCIO

Figure 5.2: Stock Graphs. We see that both transelliptical graphs reveal more refined structures than the Gaussian graph.

sparsity level 0.1;

(2) Randomly choose 50% of all samples without replacement using the regularization parameter chosen in (1);

(3) Repeat (2) 100 times and retain the edges that appear with frequencies no less than 95%.

We choose the sparsity level 0.1 in (1) and subsampling ratio 50% in (2) based on two criteria: The resulting graphs need to be sparse to ease visualization, interpretation, and computation; The resulting graphs need to be stable. We then present the obtained stability graphs in Figure 5.2. The nodes are colored according to the GICS sector of the corresponding stock. We highlight a region in the transelliptical graph obtained by the SCIO method and by color coding we see that the nodes in this region belong to the same sector of the market. A similar pattern is also found in the transelliptical graph obtained by the CLIME method. In contrast, this region is shown to be sparse in the Gaussian graph obtained by the SCIO method (based on the Pearson correlation matrix). Therefore we can see that the SCIO method is also capable of generating refined structures as the

CLIME method when estimating the transelliptical graph.

6 Discussions

We compare F-APISTA with a closely related algorithm – the path-following coordinate descent algorithm (PCDA[1]) in timing performance. In particular, we give a failure example of PCDA for solving sparse linear regression. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ denote design matrix and $\boldsymbol{y} \in \mathbb{R}^n$ denote the response vector. We solve the following regularized optimization problem,

$$\min_{\boldsymbol{\theta}} \frac{1}{2n}\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \mathcal{R}_\lambda(\boldsymbol{\theta}).$$

We generate each row of the design matrix $\mathbf{X}_{i*}$ from a d-variate Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, where $\Sigma_{kj} = 0.75$ if $k \neq j$ and $\Sigma_{kk} = 1$ for all $j,k = 1,...,d$. We then normalize each column of the design matrix $\mathbf{X}_{*j}$ such that $\|\mathbf{X}_{*j}\|_2^2 = n$. The response vector is generated from the linear model $\boldsymbol{y} = \mathbf{X}\boldsymbol{\theta}^* + \boldsymbol{\epsilon}$, where $\boldsymbol{\theta}^* \in \mathbb{R}^d$ is the regression coefficient vector, and $\boldsymbol{\epsilon}$ is generated from a n-variate Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}$. We set $n = 60$ and $d = 1000$. We set the coefficient vector as $\theta^*_{250} = 3$, $\theta^*_{500} = 2$, $\theta^*_{750} = 1.5$, and $\theta^*_j = 0$ for all $j \neq 250, 500, 750$. We then set $\alpha = 0.95$, $N = 100$, $\lambda_N = 0.25\sqrt{\log d/n}$, and $\delta_c = \delta_K = 10^{-5}$.

We then generate a validation set using the same design matrix as the training set for the regularization selection. We denote the response vector of the validation set as $\widetilde{\boldsymbol{y}} \in \mathbb{R}^n$. Let $\widehat{\boldsymbol{\theta}}^\lambda$ denote the obtained estimator using the regularization parameter λ. We then choose the optimal regularization parameter $\widehat{\lambda}$ by

$$\widehat{\lambda} = \operatorname*{argmin}_{\lambda\in\{\lambda_1,...,\lambda_N\}} \|\widetilde{\boldsymbol{y}} - \mathbf{X}\widehat{\boldsymbol{\theta}}^\lambda\|_2^2.$$

We repeat 100 simulations, and summarize the average results in Table 6. We see that F-APISTA and PCDA attain similar timing results. But PCDA achieves worse statistical performance than F-APISTA in both support recovery and parameter estimation. This is because PCDA has no control over the solution sparsity. The overselection irrelevant variables compromise the restricted strong convexity, and make PCDA attain some local optima with poor statistical properties.

References

Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research* **9** 485–516.

Beck, A. and Teboulle, M. (2009a). Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *Image Processing, IEEE Transactions on* **18** 2419–2434.

[1] In our numerical experiments, PCDA is implemented by the R package "ncvreg".

Table 6.1: Quantitative comparison between F-APISTA and PCDA. We see that F-APISTA and PCDA attain similar timing results. But PCDA achieves worse statistical performance than F-APISTA in both support recovery and parameter estimation.

Method	$\|\widehat{\theta} - \theta^*\|_2$	$\|\widehat{\theta}_{\mathcal{S}}\|_0$	$\|\widehat{\theta}_{\mathcal{S}^c}\|_0$	Correct Selection	Timing
F-APISTA	0.8001(0.9089)	2.801(0.5123)	0.890(2.112)	667/1000	0.0181(0.0025)
PCDA	1.1275(1.2539)	2.655(0.7051)	1.644(3.016)	517/1000	0.0195(0.0021)

Beck, A. and Teboulle, M. (2009b). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences* **2** 183–202.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics* **5** 232–253.

Cai, T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106** 594—607.

Dennis, J. J. E. and Schnabel, R. B. (1983). *Numerical methods for unconstrained optimization and nonlinear equations*, vol. 16. SIAM.

Fan, J., Feng, Y. and Tong, X. (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B* **74** 745–771.

Fan, J., Feng, Y. and Wu, Y. (2009). Network exploration via the adaptive lasso and scad penalties. *The Annals of Applied Statistics* **3** 521–541.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.

Fan, J., Xue, L. and Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics* **42** 819–849.

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

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* **33** 1–13.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics* **7** 397–416.

Han, F. and Liu, H. (2015). Statistical analysis of latent generalized correlation matrix estimation in transelliptical distribution. *Bernoulli* (Accepted).

Han, F., Zhao, T. and Liu, H. (2012). CODA: High dimensional copula discriminant analysis. *Journal of Machine Learning Research* **14** 629–671.

Jacob, L., Obozinski, G. and Vert, J.-P. (2009). Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*.

Kim, Y. and Kwon, S. (2012). Global optimality of nonconvex penalized estimators. *Biometrika* **99** 315–325.

Ledoux, M. (2005). *The concentration of measure phenomenon*, vol. 89. AMS Bookstore.

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

Liu, H., Han, F., Yuan, M., Lafferty, J. and Wasserman, L. (2012a). High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics* **40** 2293–2326.

Liu, H., Han, F. and Zhang, C.-H. (2012b). Transelliptical graphical models. In *Advances in Neural Information Processing Systems 25*.

Liu, H., Palatucci, M. and Zhang, J. (2009). Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *Proceedings of the 26th Annual International Conference on Machine Learning*.

Liu, H., Roeder, K. and Wasserman, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. In *Advances in Neural Information Processing Systems*.

Liu, H., Wang, L. and Zhao, T. (2014). Sparse covariance matrix estimation with eigenvalue constraints. *Journal of Computational and Graphical Statistics* **23** 439–459.

Liu, H., Wang, L. and Zhao, T. (2015). Calibrated multivariate regression with application to neural semantic basis discovery. *Journal of Machine Learning Research* (Accepted).

Liu, W. and Luo, X. (2015). Fast and adaptive sparse precision matrix estimation in high dimensions. *Journal of Multivariate Analysis* **135** 153–162.

Lu, Z. and Xiao, L. (2013). Randomized block coordinate non-monotone gradient method for a class of nonlinear programming. *arXiv preprint arXiv:1306.5918* .

Mazumder, R., Friedman, J. H. and Hastie, T. (2011). Sparsenet: Coordinate descent with non-convex penalties. *Journal of the American Statistical Association* **106** 1125–1138.

Meier, L., Van De Geer, S. and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B* **70** 53–71.

Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34** 1436–1462.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B* **72** 417–473.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics* **37** 246–270.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics* **39** 1069–1097.

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

Nesterov, Y. (1988). On an approach to the construction of optimal methods of minimization of smooth convex functions. *Ekonomika i Mateaticheskie Metody* **24** 509–517.

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

NESTEROV, Y. (2013). Gradient methods for minimizing composite objective function. *Mathematical Programming Series B* **140** 125–161.

Nocedal, J. and Wright, S. (2006). Numerical optimization, series in operations research and financial engineering. *Springer, New York* .

Qin, Z., Scheinberg, K. and Goldfarb, D. (2010). Efficient block-coordinate descent algorithms for the group lasso. *Mathematical Programming Computation* 1–27.

Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics* **5** 935–980.

Rothman, A. J., Bickel, P. J., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2** 494–515.

Shalev-Shwartz, S. and Tewari, A. (2011). Stochastic methods for ℓ_1-regularized loss minimization. *The Journal of Machine Learning Research* **12** 1865–1892.

Shen, X., Pan, W. and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association* **107** 223–232.

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B* **67** 91–108.

Tseng, P. and Yun, S. (2009a). Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization. *Journal of Optimization Theory and Applications* **140** 513–535.

Tseng, P. and Yun, S. (2009b). A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming* **117** 387–423.

Van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics* **36** 614–645.

Wainwright, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming. *IEEE Transactions on Information Theory* **55** 2183–2201.

Wang, L., Kim, Y. and Li, R. (2013). Calibrating nonconvex penalized regression in ultra-high dimension. *The Annals of Statistics* **41** 2505–2536.

Wang, Z., Liu, H. and Zhang, T. (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *The Annals of Statistics* **42** 2164–2201.

Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics* **2** 224–244.

Xue, L., Zou, H. and Cai, T. (2012). Nonconcave penalized composite conditional likelihood estimation of sparse ising models. *The Annals of Statistics* **40** 1403–1429.

Yuan, M. and Lin, Y. (2005). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B* **68** 49–67.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika* **94** 19–35.

Zhang, C.-H. (2010a). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38** 894–942.

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

Zhang, C.-H. and Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science* **27** 576–593.

Zhang, T. (2009). Some sharp performance bounds for least squares regression with l1 regularization. *The Annals of Statistics* **37** 2109–2144.

Zhang, T. (2010b). Analysis of multi-stage convex relaxation for sparse regularization. *The Journal of Machine Learning Research* **11** 1081–1107.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research* **7** 2541–2563.

Zhao, T. and Liu, H. (2012). Sparse additive machine. In *International Conference on Artificial Intelligence and Statistics*.

Zhao, T. and Liu, H. (2014). Calibrated precision matrix estimation for high-dimensional elliptical distributions. *IEEE transactions on Information Theory* **60** 7874.

Zhao, T., Liu, H., Roeder, K., Lafferty, J. and Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in r. *The Journal of Machine Learning Research* **13** 1059–1062.

Zhao, T., Liu, H. and Zhang, T. (2014a). A general theory of pathwise coordinate optimization. *arXiv preprint arXiv:1412.7477* .

Zhao, T., Roeder, K. and Liu, H. (2014b). Positive semidefinite rank-based correlation matrix estimation with application to semiparametric graph estimation. *Journal of Computational and Graphical Statistics* **23** 895–922.

Zhao, T., Yu, M., Wang, Y., Arora, R. and Liu, H. (2014c). Accelerated mini-batch randomized block coordinate descent method. In *Advances in neural information processing systems*.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B* **67** 301–320.

A Proof of Theorem 4.3

Proof. Since $\|\boldsymbol{\theta}^{(0)}\|_0 \le s^* + \widetilde{s}$ implies that $|\mathcal{A}| \le s^* + \widetilde{s}$, by Assumption 4.2 and Lemma 4.1, we know that (3.5) is strongly convex over $\boldsymbol{\theta}_{\mathcal{A}}$. Thus it has a unique global minimizer. We then analyze the amount of successive decrease. By the restricted strong convexity of $\mathcal{F}_\lambda(\boldsymbol{\theta})$, we have

$$\begin{aligned}
&\mathcal{F}_\lambda(\boldsymbol{w}^{(t+1,k)}) - \mathcal{F}_\lambda(\boldsymbol{w}^{(t+1,k+1)}) \\
&\quad \ge (\nabla_k \mathcal{L}_\lambda(\theta_k^{(t+1)}, \boldsymbol{w}_{\setminus k}^{(t+1,k)}) + \lambda \xi_k^{(t+1)})(\theta_k^{(t)} - \theta_k^{(t+1)}) + \frac{\widetilde{\rho}_-(1)}{2}(\theta_k^{(t)} - \theta_k^{(t+1)})^2,
\end{aligned} \tag{A.1}$$

where $\xi_k^{(t+1)} \in \partial|\theta_k^{(t+1)}|$ satisfies the optimality condition of (3.6),

$$\nabla_k \widetilde{\mathcal{L}}_\lambda(\theta_k^{(t+1)}, \boldsymbol{w}_{\setminus k}^{(t+1,k)}) + \lambda \xi_k^{(t+1)} = 0. \tag{A.2}$$

By combining (A.1) with (A.2), we have

$$\mathcal{F}_\lambda(\boldsymbol{w}^{(t+1,k)}) - \mathcal{F}_\lambda(\boldsymbol{w}^{(t+1,k+1)}) \ge \frac{\widetilde{\rho}_-(1)}{2}(\theta_k^{(t+1)} - \theta_k^{(t)})^2,$$

which further implies

$$\mathcal{F}_\lambda(\boldsymbol{\theta}^{(t)}) - \mathcal{F}_\lambda(\boldsymbol{\theta}^{(t+1)}) \ge \frac{\widetilde{\rho}_-(1)}{2}\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t+1)}\|_2^2. \tag{A.3}$$

We then analyze the gap in the objective value yet to be minimized after each iteration. For any $\boldsymbol{\theta}', \boldsymbol{\theta} \in \mathbb{R}^d$ with $\boldsymbol{\theta}'_{\mathcal{A}^\perp} = \boldsymbol{\theta}_{\mathcal{A}^\perp} = \mathbf{0}$, by the restricted strong convexity of $\mathcal{F}_\lambda(\boldsymbol{\theta})$, we have

$$\mathcal{F}_\lambda(\boldsymbol{\theta}') \ge \mathcal{F}_\lambda(\boldsymbol{\theta}) + (\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\theta}) + \lambda \boldsymbol{\xi})^T(\boldsymbol{\theta}' - \boldsymbol{\theta}) + \frac{\widetilde{\rho}_-(s^* + \widetilde{s})}{2}\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_2^2, \tag{A.4}$$

where $\boldsymbol{\xi} \in \mathbb{R}^d$ with $\boldsymbol{\xi}_{\mathcal{A}} \in \partial\|\boldsymbol{\theta}_{\mathcal{A}}\|_1$ and $\boldsymbol{\xi}_{\mathcal{A}^\perp} = \mathbf{0}$. We then minimize both sides of (A.4) with respect to $\boldsymbol{\theta}'_{\mathcal{A}}$ and obtain

$$\begin{aligned}
\mathcal{F}_\lambda(\boldsymbol{\theta}^{(t+1)}) - \mathcal{F}_\lambda(\bar{\boldsymbol{\theta}}) &\le \frac{1}{2\widetilde{\rho}_-(s^* + \widetilde{s})}\|\nabla_{\mathcal{A}} \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\theta}^{(t+1)}) + \lambda \boldsymbol{\xi}_{\mathcal{A}}^{(t+1)}\|_2^2 \\
&\overset{\text{(i)}}{=} \frac{1}{2\widetilde{\rho}_-(s^* + \widetilde{s})} \sum_{k=1}^{|\mathcal{A}|} \Big[\nabla_k \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\theta}^{(t+1)}) - \nabla_k \widetilde{\mathcal{L}}_\lambda(\theta_k^{(t+1)}, \boldsymbol{w}_{\setminus k}^{(t+1,k)})\Big]^2 \\
&\overset{\text{(ii)}}{\le} \frac{\rho_+^2(s^* + \widetilde{s})}{2\widetilde{\rho}_-(s^* + \widetilde{s})} \sum_{k=1}^{|\mathcal{A}|} \|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{w}^{(t+1,k)}\|^2 \\
&\le \frac{(s^* + \widetilde{s})\rho_+^2(s^* + \widetilde{s})}{2\widetilde{\rho}_-(s^* + \widetilde{s})}\|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}\|^2,
\end{aligned} \tag{A.5}$$

where (i) comes from (A.2) and (ii) comes from the restricted strong smoothness of $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\theta})$.

Eventually, by combing (A.5) with (A.3), we obtain

$$\begin{aligned}\mathcal{F}_\lambda(\boldsymbol{\theta}^{(t+1)}) - \mathcal{F}_\lambda(\bar{\boldsymbol{\theta}}) &\leq \frac{(s^*+\widetilde{s})\rho_+^2(s^*+\widetilde{s})}{\widetilde{\rho}_-(1)\widetilde{\rho}_-(s^*+\widetilde{s})}[\mathcal{F}_\lambda(\boldsymbol{\theta}^{(t)}) - \mathcal{F}_\lambda(\boldsymbol{\theta}^{(t+1)})] \\ &\leq \frac{(s^*+\widetilde{s})\rho_+^2(s^*+\widetilde{s})}{\widetilde{\rho}_-(1)\widetilde{\rho}_-(s^*+\widetilde{s})}([\mathcal{F}_\lambda(\boldsymbol{\theta}^{(t)}) - \mathcal{F}_\lambda(\bar{\boldsymbol{\theta}})] - [\mathcal{F}_\lambda(\boldsymbol{\theta}^{(t+1)}) - \mathcal{F}_\lambda(\bar{\boldsymbol{\theta}})]),\end{aligned}$$

which further implies

$$\mathcal{F}_\lambda(\boldsymbol{\theta}^{(t+1)}) - \mathcal{F}_\lambda(\bar{\boldsymbol{\theta}}) \leq \left(\frac{s^*+\widetilde{s})\rho_+^2(s^*+\widetilde{s})}{(s^*+\widetilde{s})\rho_+^2(s^*+\widetilde{s}) + \widetilde{\rho}_-(1)\widetilde{\rho}_-(s^*+\widetilde{s})}\right)[\mathcal{F}_\lambda(\boldsymbol{\theta}^{(t)}) - \mathcal{F}_\lambda(\bar{\boldsymbol{\theta}})]. \tag{A.6}$$

By recursively applying (A.6), we complete the proof. □

B Proof of Theorem 4.4

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

Lemma B.1. Suppose that Assumptions 4.1 and 4.2 hold. For any $\lambda \geq \lambda_N$, if $\boldsymbol{\theta}$ satisfies,

$$\|\boldsymbol{\theta}_{\mathcal{S}^\perp}\|_0 \leq \widetilde{s} \quad \text{and} \quad \omega_\lambda(\boldsymbol{\theta}) \leq \lambda/2, \tag{B.1}$$

then we have

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq \frac{21\lambda\sqrt{s^*}}{8\widetilde{\rho}_-(s^*+\widetilde{s})}, \ \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_1 \leq \frac{21\lambda s^*}{\widetilde{\rho}_-(s^*+\widetilde{s})}, \ \text{and} \ \mathcal{F}_\lambda(\boldsymbol{\theta}) \leq \mathcal{F}_\lambda(\boldsymbol{\theta}^*) + \frac{21\lambda^2 s^*}{2\widetilde{\rho}_-(s^*+\widetilde{s})}.$$

Lemma B.2. Suppose that Assumptions 4.1 and 4.2 hold. For any $\lambda \geq \lambda_N$, if $\boldsymbol{\theta}$ satisfies,

$$\|\boldsymbol{\theta}_{\mathcal{S}^\perp}\|_0 \leq \widetilde{s} \quad \text{and} \quad \mathcal{F}_\lambda(\boldsymbol{\theta}) \leq \mathcal{F}_\lambda(\boldsymbol{\theta}^*) + \frac{21\lambda^2 s^*}{2\widetilde{\rho}_-(s^*+\widetilde{s})},$$

then we have $\|[\mathcal{T}_{L,\lambda}(\boldsymbol{\theta})]_{\mathcal{S}^\perp}\|_0 \leq \widetilde{s}$ for any $L \leq 2\rho_+(s^*+2\widetilde{s})$.

The proofs of Lemmas B.1 and B.2 are provided in Wang et al. (2014), therefore omitted. Since the initial solution $\boldsymbol{\theta}^{[0]}$ satisfies the approximate KKT condition. By Lemma B.1, we know that $\boldsymbol{\theta}^{[0]}$ satisfies

$$\mathcal{F}_\lambda(\boldsymbol{\theta}^{[0]}) \leq \mathcal{F}_\lambda(\boldsymbol{\theta}^*) + \frac{21\lambda^2 s^*}{2\widetilde{\rho}_-(s^*+\widetilde{s})}. \tag{B.2}$$

We assume $L^{[m]} \leq 2\rho_+(s^*+2\widetilde{s})$. Since $\|\boldsymbol{\theta}^{[0]}_{\mathcal{S}^\perp}\|_0 \leq \widetilde{s}$, by (B.2) and Lemma B.2, we have $\boldsymbol{\theta}^{[0.5]} = \mathcal{T}_{L,\lambda}(\boldsymbol{\theta}^{[0]})$ and $\|\boldsymbol{\theta}^{[0.5]}_{\mathcal{S}^\perp}\|_0 \leq \widetilde{s}$. Since the coordinate descent subroutine iterates over $\mathcal{A} = \text{supp}(\boldsymbol{\theta}^{[0.5]})$, its output solution $\boldsymbol{\theta}^{[1]}$ also satisfies $\|\boldsymbol{\theta}^{[1]}_{\mathcal{S}^\perp}\|_0 \leq \widetilde{s}$. Since the proximal gradient descent iteration and coordinate descent subroutine decrease the objective value, by (B.2), we also have

$$\mathcal{F}_\lambda(\boldsymbol{\theta}^{[1]}) \leq \mathcal{F}_\lambda(\boldsymbol{\theta}^{[0.5]}) \leq \mathcal{F}_\lambda(\boldsymbol{\theta}^{[0]}) \leq \mathcal{F}_\lambda(\boldsymbol{\theta}^*) + \frac{21\lambda^2 s^*}{2\widetilde{\rho}_-(s^*+\widetilde{s})}.$$

Then by induction, we know that all successive $\boldsymbol{\theta}^{[m]}$'s satisfy $\|\boldsymbol{\theta}^{[m]}_{\mathcal{S}^\perp}\|_0 \le \widetilde{s}$ for $m = 1.5, 2, 2.5, \ldots$.

Now we verify $L^{[m]} \le 2\rho_+(s^* + 2\widetilde{s})$. Since we start with a small enough $L = \rho_+(1) \le 2\rho_+(s^* + 2\widetilde{s})$. If L does not satisfy the stopping criterion for the backtracking line search in (3.4), then we multiply L by 2. Once L attains the interval $\in [\rho_+(s^* + 2\widetilde{s}), 2\rho_+(s^* + 2\widetilde{s})]$, it stops increasing. Because by the restricted strong smoothness of $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\theta})$, such a step size parameter always guarantees that the algorithm iterates from a sparse $\boldsymbol{\theta}^{[m]}$ to a sparse $\boldsymbol{\theta}^{[m+0.5]}$, and meanwhile satisfies the stopping criterion of the backtracking line search. Thus $L^{[m]} \le 2\rho_+(s^* + 2\widetilde{s})$ is verified.

The existence and uniqueness of $\bar{\boldsymbol{\theta}}^\lambda$ has been verified in Wang et al. (2014). Therefore the proof is omitted. We then proceed to derive the geometric rate of convergence to $\bar{\boldsymbol{\theta}}^\lambda$ by the next lemma.

Lemma B.3. Suppose that Assumptions 4.1 and 4.2 hold. For any $\lambda \ge \lambda_N$, if $\boldsymbol{\theta}$ satisfies

$$\|\boldsymbol{\theta}_{\mathcal{S}^\perp}\|_0 \le \widetilde{s} \quad \text{and} \quad \mathcal{F}_\lambda(\boldsymbol{\theta}) \le \mathcal{F}_\lambda(\boldsymbol{\theta}^*) + \frac{21\lambda^2 s^*}{2\widetilde{\rho}_-(s^* + \widetilde{s})}, \tag{B.3}$$

given $L \le 2\rho_+(s^* + 2\widetilde{s})$, then we have

$$\mathcal{F}_\lambda(\mathcal{T}_{\lambda,L}(\boldsymbol{\theta})) - \mathcal{F}_\lambda(\bar{\boldsymbol{\theta}}^\lambda) \le \left(1 - \frac{1}{8\kappa}\right)[\mathcal{F}_\lambda(\boldsymbol{\theta}) - \mathcal{F}_\lambda(\bar{\boldsymbol{\theta}}^\lambda)].$$

The proof of Lemma B.3 is provided in Wang et al. (2014), therefore omitted. Since we have verified that all $\boldsymbol{\theta}^{[m]}$'s satisfy (B.3) and all $L^{[m]}$'s satisfy $L^{[m]} \le 2\rho_+(s^* + 2\widetilde{s})$ for $m = 0, 1, 2, \ldots$, Lemma B.3 implies

$$\mathcal{F}_\lambda(\boldsymbol{\theta}^{[m+1]}) - \mathcal{F}_\lambda(\bar{\boldsymbol{\theta}}^\lambda) \le \mathcal{F}_\lambda(\boldsymbol{\theta}^{[m+0.5]}) - \mathcal{F}_\lambda(\bar{\boldsymbol{\theta}}^\lambda) \le \left(1 - \frac{1}{8\kappa}\right)[\mathcal{F}_\lambda(\boldsymbol{\theta}^{[m]}) - \mathcal{F}_\lambda(\bar{\boldsymbol{\theta}}^\lambda)], \tag{B.4}$$

where the first inequality holds because the coordinate descent subroutine decreases the objective value. Then by recursively applying (B.4), we compete the proof. □

C Proof of Theorem 4.6

Proof. Before we proceed with the proof of Result (1), we first introduce the following lemma.

Lemma C.1. Suppose that Assumptions 4.1 and 4.2 hold. For any $\lambda \ge \lambda_N$, if $\boldsymbol{\theta}$ satisfies

$$\|\boldsymbol{\theta}_{\mathcal{S}^\perp}\|_0 \le \widetilde{s} \quad \text{and} \quad \omega_\lambda(\boldsymbol{\theta}) \le \delta_{\max}\lambda,$$

then for any $\lambda' \in [\lambda_N, \lambda]$, we have

$$\mathcal{F}_{\lambda'}(\boldsymbol{\theta}) - \mathcal{F}_{\lambda'}(\bar{\boldsymbol{\theta}}^{\lambda'}) \le \frac{21[\delta_{\max}\lambda + 2(\lambda - \lambda')](\lambda + \lambda')s^*}{\widetilde{\rho}_-(s^* + \widetilde{s})}.$$

The proof of Lemmas C.1 is provided in Wang et al. (2014), therefore omitted. If we take $\lambda = \lambda' = \lambda_K$ and $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{\{K-1\}}$, then Lemma C.1 implies

$$\mathcal{F}_{\lambda_K}(\widehat{\boldsymbol{\theta}}^{\{K-1\}}) - \mathcal{F}_{\lambda_K}(\bar{\boldsymbol{\theta}}^{\lambda_K}) \leq \frac{21 s^* \lambda_K^2}{2\widetilde{\rho}_-(s^* + \widetilde{s})}. \tag{C.1}$$

Recall (A.3) in Appendix A. Within each coordinate descent subroutine for λ_K, we have

$$\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t+1)}\|_2^2 \leq \frac{2\left[\mathcal{F}_{\lambda_K}(\boldsymbol{\theta}^{(t)}) - \mathcal{F}_{\lambda_K}(\boldsymbol{\theta}^{(t+1)})\right]}{\widetilde{\rho}_-(1)} \leq \frac{2\left[\mathcal{F}_{\lambda_K}(\boldsymbol{\theta}^{(t)}) - \mathcal{F}_{\lambda_K}(\bar{\boldsymbol{\theta}})\right]}{\widetilde{\rho}_-(1)}. \tag{C.2}$$

By combining Theorem 4.3 with (C.2), we have

$$\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t+1)}\|_2^2 \leq 2\left(\frac{(s^* + \widetilde{s})\rho_+^2(s^* + \widetilde{s})}{(s^* + \widetilde{s})\rho_+^2(s^* + \widetilde{s}) + \widetilde{\rho}_-(1)\widetilde{\rho}_-(s^* + \widetilde{s})}\right)^t \frac{\left[\mathcal{F}_{\lambda_K}(\boldsymbol{\theta}^{(0)}) - \mathcal{F}_{\lambda_K}(\bar{\boldsymbol{\theta}})\right]}{\widetilde{\rho}_-(1)}.$$

Therefore given

$$t \geq \log\left(\frac{2\left[\mathcal{F}_{\lambda_K}(\boldsymbol{\theta}^{(0)}) - \mathcal{F}_{\lambda_K}(\bar{\boldsymbol{\theta}})\right]}{\widetilde{\rho}_-(1)\delta_0^2\lambda_K^2}\right) \Big/ \log^{-1}\left(\frac{(s^* + \widetilde{s})\rho_+^2(s^* + \widetilde{s})}{(s^* + \widetilde{s})\rho_+^2(s^* + \widetilde{s}) + \widetilde{\rho}_-(1)\widetilde{\rho}_-(s^* + \widetilde{s})}\right), \tag{C.3}$$

we have

$$\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t+1)}\|_2^2 \leq 2\left(\frac{s^* + \widetilde{s})\rho_+^2(s^* + \widetilde{s})}{(s^* + \widetilde{s})\rho_+^2(s^* + \widetilde{s}) + \widetilde{\rho}_-(1)\widetilde{\rho}_-(s^* + \widetilde{s})}\right)^t \frac{\left[\mathcal{F}_{\lambda_K}(\boldsymbol{\theta}^{(0)}) - \mathcal{F}_{\lambda_K}(\bar{\boldsymbol{\theta}})\right]}{\widetilde{\rho}_-(1)} \leq \delta_0^2\lambda_K^2,$$

which satisfies the stopping criterion of CCDA for λ_K. Since both the proximal gradient descent iteration and coordinate descent subroutine decrease the objective value, we have

$$\mathcal{F}_{\lambda_K}(\widehat{\boldsymbol{\theta}}^{\{K-1\}}) \geq \mathcal{F}_{\lambda_K}(\boldsymbol{\theta}^{(0)}) \geq \mathcal{F}_{\lambda_K}(\bar{\boldsymbol{\theta}}) \geq \mathcal{F}_{\lambda_K}(\bar{\boldsymbol{\theta}}^{\lambda_K}) \tag{C.4}$$

within each coordinate descent subroutine for the K^{th} stage. By combining (C.1) and (C.3) with (C.4), we have

$$t \geq \log\left(\frac{21 s^*}{\widetilde{\rho}_-(s^* + \widetilde{s})\widetilde{\rho}_-(1)\delta_0^2}\right) \Big/ \log^{-1}\left(\frac{(s^* + \widetilde{s})\rho_+^2(s^* + \widetilde{s})}{(s^* + \widetilde{s})\rho_+^2(s^* + \widetilde{s}) + \widetilde{\rho}_-(1)\widetilde{\rho}_-(s^* + \widetilde{s})}\right).$$

Before we proceed with the proof of Result (2), we first introduce the following lemma.

Lemma C.2. Suppose that Assumptions 4.1 and 4.2 hold. For any $\lambda \geq \lambda_N$, if $\boldsymbol{\theta}$ satisfies,

$$\|\boldsymbol{\theta}_{\mathcal{S}^\perp}\|_0 \leq \widetilde{s} \quad \text{and} \quad \mathcal{F}_\lambda(\boldsymbol{\theta}) \leq \mathcal{F}_\lambda(\boldsymbol{\theta}^*) + \frac{21\lambda^2 s^*}{2\widetilde{\rho}_-(s^* + \widetilde{s})}, \tag{C.5}$$

given $L \leq 2\rho_+(s^* + 2\widetilde{s})$, we have

$$\omega_\lambda(\mathcal{T}_{\lambda,L}(\boldsymbol{\theta})) \leq 3\sqrt{\rho_+(s^* + 2\widetilde{s})\left[\mathcal{F}_\lambda(\mathcal{T}_{\lambda,L}(\boldsymbol{\theta})) - \mathcal{F}_\lambda(\boldsymbol{\theta})\right]}.$$

The proof of Lemma C.2 is provided in Wang et al. (2014), therefore omitted. Recall that in Appendix B, we have shown that at the K^{th} stage, $\boldsymbol{\theta}^{[m]}$ satisfies (C.5). The backtracking line search guarantees $L^{[m+1]} \leq 2\rho_+(s^* + 2\widetilde{s})$. Thus by Lemma C.2, we have

$$\begin{aligned}\omega_{\lambda_K}(\boldsymbol{\theta}^{[m+0.5]}) &\leq 3\sqrt{\rho_+(s^* + 2\widetilde{s})\left[\mathcal{F}_{\lambda_K}(\boldsymbol{\theta}^{[m+0.5]}) - \mathcal{F}_{\lambda_K}(\boldsymbol{\theta}^{[m]})\right]} \\ &\leq 3\sqrt{\rho_+(s^* + 2\widetilde{s})\left[\mathcal{F}_{\lambda_K}(\boldsymbol{\theta}^{[m+1]}) - \mathcal{F}_{\lambda_K}(\bar{\boldsymbol{\theta}}^{\lambda_K})\right]}, \end{aligned} \tag{C.6}$$

where the last inequality holds since the coordinate descent subroutine decreases the objective value. By combining (C.6) with Theorem 4.4, we obtain

$$\omega_{\lambda_K}(\boldsymbol{\theta}^{[m+0.5]}) \leq 3\sqrt{\rho_+(s^* + 2\widetilde{s})\left(1 - \frac{1}{8\kappa}\right)^{m+1}\left[\mathcal{F}_{\lambda_K}(\boldsymbol{\theta}^{[0]}) - \mathcal{F}_{\lambda_K}(\bar{\boldsymbol{\theta}}^{\lambda_K})\right]}.$$

Thus as long as

$$m \geq \log\left(\frac{9\rho_+(s^* + 2\widetilde{s})\left[\mathcal{F}_{\lambda_K}(\boldsymbol{\theta}^{[0]}) - \mathcal{F}_{\lambda_K}(\bar{\boldsymbol{\theta}}^{\lambda_K})\right]}{\delta_K^2\lambda_K^2}\right)\Big/\log^{-1}\left(1 - \frac{1}{8\kappa}\right), \tag{C.7}$$

we have

$$\omega_{\lambda_K}(\boldsymbol{\theta}^{[m+0.5]}) \leq 3\sqrt{\rho_+(s^* + 2\widetilde{s})\left(1 - \frac{1}{8\kappa}\right)^{m}\left[\mathcal{F}_{\lambda_K}(\boldsymbol{\theta}^{[0]}) - \mathcal{F}_{\lambda_K}(\bar{\boldsymbol{\theta}}^{\lambda_K})\right]} \leq \delta_K\lambda_K,$$

which satisfies the stopping criterion of AISTA at the K^{th} stage. By combining (C.1) with (C.7), we have

$$m \geq \log\left(\frac{189\kappa\lambda_K^2 s^*}{2\delta_K^2\lambda_K^2}\right)\Big/\log^{-1}\left(1 - \frac{1}{8\kappa}\right).$$

Result (3) is just a straightforward combination of Results (1) and (2).

To prove Result (4), we need to use Lemma C.1 again. In particular, for $K < N$, we take $\lambda' = \lambda_N$, $\lambda = \lambda_K$ and $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{\{K\}}$. We then have

$$\mathcal{F}_{\lambda_N}(\widehat{\boldsymbol{\theta}}^{\{K\}}) - \mathcal{F}_{\lambda_N}(\bar{\boldsymbol{\theta}}^{\lambda_N}) \leq \frac{21(\lambda_K + \lambda_N)(\omega_{\lambda_K}(\widehat{\boldsymbol{\theta}}^{\{K\}}) + 2(\lambda_K - \lambda_N))s^*)}{\widetilde{\rho}_-(s^* + \widetilde{s})}. \tag{C.8}$$

Since we have $\lambda_K > \lambda_N$ for $K = 1, ..., N-1$, (C.8) implies

$$\mathcal{F}_{\lambda_N}(\widehat{\boldsymbol{\theta}}^{\{K\}}) - \mathcal{F}_{\lambda_N}(\bar{\boldsymbol{\theta}}^{\lambda_N}) \leq \frac{105\lambda_K^2 s^*}{\widetilde{\rho}_-(s^* + \widetilde{s})}. \tag{C.9}$$

For $K = N$, (C.8) implies

$$\mathcal{F}_{\lambda_N}(\widehat{\boldsymbol{\theta}}^{\{N\}}) - \mathcal{F}_{\lambda_N}(\bar{\boldsymbol{\theta}}^{\lambda_N}) \leq \frac{105\delta_N\lambda_N^2 s^*}{\widetilde{\rho}_-(s^* + \widetilde{s})}. \tag{C.10}$$

By combining (C.9) with (C.10), we prove Result (4). □

D Proof of Lemma 4.8

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

Lemma D.1. Suppose that $\boldsymbol{X} \sim TE_d(\Sigma^*, \xi, \{f_j\}_{j=1}^d)$. We have

$$\mathbb{P}\left(\|\widehat{\mathbf{S}} - \Sigma^*\|_{\max} \leq \sqrt{2}\pi\sqrt{\frac{\log d}{n}}\right) \geq 1 - \frac{1}{d^2}. \tag{D.1}$$

The proof of Lemma D.1 is provided in Liu et al. (2012a), therefore omitted. We consider the following decomposition,

$$\|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\|_\infty = \|\widehat{\mathbf{S}}\boldsymbol{\theta}^* - \boldsymbol{e}\|_\infty = \|(\widehat{\mathbf{S}} - \Sigma^*)\boldsymbol{\theta}^*\|_\infty \leq \|\boldsymbol{\theta}^*\|_1 \|\widehat{\mathbf{S}} - \Sigma^*\|_{\max}. \tag{D.2}$$

Then by combining (D.1) and (D.2) with the fact $\|\boldsymbol{\theta}^*\|_1 \leq \|\boldsymbol{\Theta}^*\|_1 \leq M$, we have

$$\mathbb{P}\left(\|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\|_\infty \leq \sqrt{2}\pi M\sqrt{\frac{\log d}{n}}\right) \geq 1 - \frac{1}{d^2},$$

which completes the proof. □

E Proof of Lemma 4.9

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

Lemma E.1. Suppose that $\boldsymbol{X} \sim TE_d(\Sigma^*, \xi, \{f_j\}_{j=1}^d)$. There exists a universal constant c_2 such that

$$\mathbb{P}\left(\sup_{\|\boldsymbol{\theta}\|_0 \leq s} |\boldsymbol{\theta}^T(\widehat{\mathbf{S}} - \Sigma^*)\boldsymbol{\theta}| \leq \frac{c_2 s\log d}{n}\|\boldsymbol{\theta}\|_2^2\right) \geq 1 - \frac{2}{d^2}. \tag{E.1}$$

The proof of Lemma E.1 is provided in Han and Liu (2015), therefore omitted. We consider the decomposition

$$\boldsymbol{\theta}^T\widehat{\mathbf{S}}\boldsymbol{\theta} = \boldsymbol{\theta}^T\Sigma^*\boldsymbol{\theta} + \boldsymbol{\theta}^T(\widehat{\mathbf{S}} - \Sigma^*)\boldsymbol{\theta}. \tag{E.2}$$

By assuming $\|\boldsymbol{\theta}\|_0 \leq s^* + 2\widetilde{s}$ and

$$|\boldsymbol{\theta}^T(\widehat{\mathbf{S}} - \overline{\Sigma})\boldsymbol{\theta}| \leq c_2\frac{(s^* + 2\widetilde{s})\log d}{n}\|\boldsymbol{\theta}\|_2^2/n,$$

we further have

$$\boldsymbol{\theta}^T\widehat{\mathbf{S}}\boldsymbol{\theta} \leq \Lambda_{\max}(\Sigma^*)\cdot\|\boldsymbol{\theta}\|_2^2 + |\boldsymbol{\theta}^T(\widehat{\mathbf{S}} - \Sigma^*)\boldsymbol{\theta}| \leq \psi_{\max}\|\boldsymbol{\theta}\|_2^2 + c_2\frac{(s^* + 2\widetilde{s})\log d}{n}\|\boldsymbol{\theta}\|_2^2, \tag{E.3}$$

$$\boldsymbol{\theta}^T\widehat{\mathbf{S}}\boldsymbol{\theta} \geq \Lambda_{\min}(\Sigma^*)\cdot\|\boldsymbol{\theta}\|_2^2 - |\boldsymbol{\theta}^T(\widehat{\mathbf{S}} - \Sigma^*)\boldsymbol{\theta}| \leq \psi_{\min}\|\boldsymbol{\theta}\|_2^2 - c_2\frac{(s^* + 2\widetilde{s})\log d}{n}\|\boldsymbol{\theta}\|_2^2. \tag{E.4}$$

Thus for $n \geq 4\psi_{\min}^{-1} c_2(s^* + 2\widetilde{s})\log d$, we have

$$3\psi_{\min}\|\boldsymbol{\theta}\|_2^2/4 \leq \boldsymbol{\theta}^T \widehat{\mathbf{S}}\boldsymbol{\theta} \leq 5\psi_{\max}\|\boldsymbol{\theta}\|_2^2/4.$$

Given $\alpha = \psi_{\min}/2$, we have

$$\rho_+(s^* + 2\widetilde{s}) \leq 5\psi_{\max}/4, \quad \widetilde{\rho}_-(s^* + 2\widetilde{s}) \geq \psi_{\min}/4, \quad \kappa \leq 5\psi_{\max}/\psi_{\min}. \tag{E.5}$$

Since we need to secure $\widetilde{s} = c_1 s^* \geq (144\kappa^2 + 250\kappa)s^*$, we take

$$c_1 = 3600\psi_{\max}^2/\psi_{\min}^2 + 1250\psi_{\max}/\psi_{\min} \geq 72(1+\gamma)\kappa^2 + 250\kappa. \tag{E.6}$$

In another word, we need

$$n \geq 4\psi_{\min}^{-1} c_2(1 + 2c_1)s^* \log d \geq 4\psi_{\min}^{-1} c_2(s^* + 2\widetilde{s})\log d.$$

Eventually by combining (E.1) and (E.5) with (E.6), we complete the proof. □

F Proof of Theorem 4.11

Proof. Recall that the output solution $\widehat{\boldsymbol{\theta}}^{\{N\}}$ satisfies $\|\widehat{\boldsymbol{\theta}}^{\{N\}}_{\mathcal{S}^\perp}\|_0 \leq \widetilde{s}$ and $\omega_{\lambda_N} \leq \delta_N \lambda_N$. By Lemma B.1, we have

$$\|\widehat{\boldsymbol{\theta}}^{\{N\}} - \boldsymbol{\theta}^*\|_1 \leq \frac{21\lambda_N s^*}{\widetilde{\rho}_-(s^* + \widetilde{s})} \quad \text{and} \quad \|\widehat{\boldsymbol{\theta}}^{\{N\}} - \boldsymbol{\theta}^*\|_2^2 \leq \frac{7\lambda_N^2 s^*}{\widetilde{\rho}_-(s^* + \widetilde{s})}. \tag{F.1}$$

By the definition of the matrix ℓ_1 and Frobenius norms, we have

$$\|\widehat{\boldsymbol{\Theta}}^{\{N\}} - \boldsymbol{\Theta}^*\|_1 = \max_{1\leq j\leq d} \|\boldsymbol{\Theta}_{*j}^{\{N\}} - \boldsymbol{\Theta}_{*j}^*\|_1 \quad \text{and} \quad \|\widehat{\boldsymbol{\Theta}}^{\{N\}} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}^2 = \sum_{j=1}^{d} \|\boldsymbol{\Theta}_{*j}^{\{N\}} - \boldsymbol{\Theta}_{*j}^*\|_2^2. \tag{F.2}$$

Recall that we use $\widehat{\boldsymbol{\theta}}^{\{N\}}$ to denote arbitrary column of $\widehat{\boldsymbol{\Theta}}^{\{N\}}$. By combining (F.2) with (F.1), we have

$$\|\widehat{\boldsymbol{\Theta}}^{\{N\}} - \boldsymbol{\Theta}^*\|_1 \leq \frac{21\lambda_N s^*}{\widetilde{\rho}_-(s^* + \widetilde{s})} \quad \text{and} \quad \frac{1}{d}\|\widehat{\boldsymbol{\Theta}}^{\{N\}} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}^2 \leq \frac{7\lambda_N^2 s^*}{\widetilde{\rho}_-(s^* + \widetilde{s})}.$$

Since all above results rely on Assumptions 4.1 and 4.2, by Lemma 4.8 and 4.9, we have

$$\|\widehat{\boldsymbol{\Theta}}^{\{N\}} - \boldsymbol{\Theta}^*\|_1 \leq \frac{168\sqrt{2}\pi s^* M}{\widetilde{\rho}_-(s^* + \widetilde{s})}\sqrt{\frac{\log d}{n}} \quad \text{and} \quad \frac{1}{d}\|\widehat{\boldsymbol{\Theta}}^{\{N\}} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}^2 \leq \frac{896\pi^2 s^* M^2 \log d}{\widetilde{\rho}_-(s^* + \widetilde{s})n}$$

with probability $1 - 3d^{-2}$, which completes the proof. □

G Proof of Theorem 4.12

Proof. For notational simplicity, we omit the column index j, and use $\mathcal{S}$ and $\widehat{\boldsymbol{\theta}}^{\mathrm{o}} \in \mathbb{R}^d$ to denote the true support $\mathcal{S}_j$ and corresponding oracle estimator $\widehat{\boldsymbol{\Theta}}^{\mathrm{o}}$ respectively for the j^{th} column. In particular, we can rewrite (4.6) as follows,

$$\widehat{\boldsymbol{\theta}}^{\mathrm{o}}_{\mathcal{S}} = \underset{\boldsymbol{\theta}_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}}{\operatorname{argmin}} \frac{1}{2} \boldsymbol{\theta}_{\mathcal{S}}^T \widehat{\mathbf{S}}_{\mathcal{S}\mathcal{S}} \boldsymbol{\theta}_{\mathcal{S}} - \boldsymbol{e}_{\mathcal{S}}^T \boldsymbol{\theta}_{\mathcal{S}} \quad \text{and} \quad \widehat{\boldsymbol{\theta}}^{\mathrm{o}}_{\mathcal{S}^\perp} = \mathbf{0}. \tag{G.1}$$

Suppose that Assumption 4.2 holds. We have

$$\Lambda_{\min}(\mathbf{S}_{\mathcal{S}\mathcal{S}}) \geq \rho_-(s^*) \geq \rho_-(s^* + 2\widetilde{s}) = \widetilde{\rho}_-(s^* + 2\widetilde{s}) + \alpha > \alpha,$$

which implies that $\mathbf{S}_{\mathcal{S}\mathcal{S}}$ is positive definite. Thus (G.1) is strongly convex and $\widehat{\boldsymbol{\theta}}^{\mathrm{o}}$ is a unique minimizer. In our following analysis, we also assume

$$\|\widehat{\mathbf{S}} - \Sigma^*\|_{\max} \leq \sqrt{2}\pi \sqrt{\frac{\log d}{n}}. \tag{G.2}$$

By the strong convexity of (G.1), we have

$$\begin{aligned} 0 &\overset{\text{(i)}}{\geq} \frac{1}{2} (\widehat{\boldsymbol{\theta}}^{\mathrm{o}}_{\mathcal{S}})^T \widehat{\mathbf{S}}_{\mathcal{S}\mathcal{S}} \widehat{\boldsymbol{\theta}}^{\mathrm{o}}_{\mathcal{S}} - \boldsymbol{e}_{\mathcal{S}}^T \widehat{\boldsymbol{\theta}}^{\mathrm{o}}_{\mathcal{S}} - \frac{1}{2} (\boldsymbol{\theta}^*_{\mathcal{S}})^T \widehat{\mathbf{S}}_{\mathcal{S}\mathcal{S}} \boldsymbol{\theta}^*_{\mathcal{S}} + \boldsymbol{e}_{\mathcal{S}}^T \boldsymbol{\theta}^*_{\mathcal{S}} \\ &\geq (\widehat{\mathbf{S}}_{\mathcal{S}\mathcal{S}} \boldsymbol{\theta}^*_{\mathcal{S}} - \boldsymbol{e})^T (\widehat{\boldsymbol{\theta}}^{\mathrm{o}}_{\mathcal{S}} - \boldsymbol{\theta}^*_{\mathcal{S}}) + \frac{\rho_-(s^*)}{2} \|\boldsymbol{\theta}^*_{\mathcal{S}} - \widehat{\boldsymbol{\theta}}^{\mathrm{o}}_{\mathcal{S}}\|_2^2, \end{aligned} \tag{G.3}$$

where (i) comes from the fact that $\widehat{\boldsymbol{\theta}}^{\mathrm{o}}$ is the minimizer to (G.1). For notational simplicity, we denote $\widehat{\boldsymbol{\Delta}}^{\mathrm{o}}_{\mathcal{S}} = \widehat{\boldsymbol{\theta}}^{\mathrm{o}}_{\mathcal{S}} - \boldsymbol{\theta}^*_{\mathcal{S}}$. By the Cauchy-Schwarz inequality, (G.3) can be rewritten as

$$\frac{\rho_-(s^*)}{2} \|\widehat{\boldsymbol{\Delta}}^{\mathrm{o}}_{\mathcal{S}}\|_2^2 \leq -(\widehat{\mathbf{S}}_{\mathcal{S}\mathcal{S}} \boldsymbol{\theta}^*_{\mathcal{S}} - \boldsymbol{e}_{\mathcal{S}})^T \widehat{\boldsymbol{\Delta}}^{\mathrm{o}}_{\mathcal{S}} \leq \|\widehat{\mathbf{S}}_{\mathcal{S}\mathcal{S}} \boldsymbol{\theta}^*_{\mathcal{S}} - \boldsymbol{e}_{\mathcal{S}}\|_{\max} \|\widehat{\boldsymbol{\Delta}}^{\mathrm{o}}_{\mathcal{S}}\|_1 \leq \|\widehat{\mathbf{S}}\boldsymbol{\theta}^* - \boldsymbol{e}\|_{\max} \sqrt{s^*} \|\widehat{\boldsymbol{\Delta}}^{\mathrm{o}}_{\mathcal{S}}\|_2,$$

where the last inequality comes from (G.2) and the fact that $\widehat{\boldsymbol{\Delta}}^{\mathrm{o}}$ contains at most s^* entries. By simple manipulations, we obtain

$$\|\widehat{\boldsymbol{\Delta}}^{\mathrm{o}}_{\mathcal{S}}\|_2 \leq \frac{2\sqrt{s^*} \|\widehat{\mathbf{S}}\boldsymbol{\theta}^* - \boldsymbol{e}\|_{\max}}{\rho_-(s^*)} \leq \frac{2\sqrt{s^*} \|\boldsymbol{\theta}^*_{\mathcal{S}}\|_1 \|\widehat{\mathbf{S}}_{\mathcal{S}\mathcal{S}} - \Sigma^*_{\mathcal{S}\mathcal{S}}\|_{\max}}{\rho_-(s^*)} \leq \frac{2\sqrt{2}\pi M}{\rho_-(s^*)} \sqrt{\frac{s^* \log d}{n}}, \tag{G.4}$$

where the last inequality comes from the fact $\|\boldsymbol{\theta}^*\|_1 \leq \|\boldsymbol{\Theta}^*\|_1 \leq M$. By combining (G.4) with Assumption 4.3, we obtain

$$\min_{j \in \mathcal{S}} |\widehat{\theta}^{\mathrm{o}}_j| \geq \min_{j \in \mathcal{S}} |\theta^*_j| - \|\widehat{\boldsymbol{\Delta}}^{\mathrm{o}}_{\mathcal{S}}\|_\infty \overset{\text{(i)}}{\geq} \min_{j \in \mathcal{S}} |\theta^*_j| - \|\widehat{\boldsymbol{\Delta}}^{\mathrm{o}}_{\mathcal{S}}\|_2 = \left(c_3 - \frac{2\sqrt{2}\pi}{\rho_-(s^*)}\right) M \sqrt{\frac{s^* \log d}{n}},$$

where (i) comes from the fact $\|\widehat{\boldsymbol{\Delta}}^{\mathrm{o}}_{\mathcal{S}}\|_\infty \leq \|\widehat{\boldsymbol{\Delta}}^{\mathrm{o}}_{\mathcal{S}}\|_2$. Now we assume $c_3 \geq 2\sqrt{2}\pi\rho_-^{-1}(s^*) + c_4\sqrt{2}\pi\beta$ for some constant c_4 (will be discussed later). We then have

$$\min_{j \in \mathcal{S}} |\widehat{\theta}^{\mathrm{o}}_j| \geq c_4 \sqrt{2}\pi M \sqrt{\frac{s^* \log d}{n}} \geq \lambda_N \beta.$$

Now we show that $\widehat{\theta}^{\mathrm{o}}$ is a sparse local solution to (2.4). In particular, we have the following decomposition,

$$\mathcal{L}\widehat{\theta}^{\mathrm{o}} = \widehat{\mathbf{S}}\widehat{\theta}^{\mathrm{o}} - \boldsymbol{e} = \begin{bmatrix} \widehat{\mathbf{S}}_{\mathcal{S}\mathcal{S}} & \widehat{\mathbf{S}}_{\mathcal{S}\mathcal{S}^{\perp}} \\ \widehat{\mathbf{S}}_{\mathcal{S}^{\perp}\mathcal{S}} & \widehat{\mathbf{S}}_{\mathcal{S}^{\perp}\mathcal{S}^{\perp}} \end{bmatrix} \begin{bmatrix} \widehat{\theta}^{\mathrm{o}}_{\mathcal{S}} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{e}_{\mathcal{S}} \\ \mathbf{0} \end{bmatrix}.$$

Since $\widehat{\theta}^{\mathrm{o}}_{\mathcal{S}}$ is the minimizer to (G.1), by the KKT condition of (G.1), we have

$$\widehat{\mathbf{S}}_{\mathcal{S}\mathcal{S}}\widehat{\theta}^{\mathrm{o}}_{\mathcal{S}} - \boldsymbol{e}_{\mathcal{S}} = \mathbf{0}. \tag{G.5}$$

Moreover, since $\min_{j\in\mathcal{S}}|\widehat{\theta}^{\mathrm{o}}_j| \geq \lambda_N\beta$, we have

$$\partial\mathcal{R}_{\lambda_N}(\widehat{\theta}^{\mathrm{o}}_{\mathcal{S}}) = -\nabla\mathcal{H}_{\lambda_N}(\widehat{\theta}^{\mathrm{o}}_{\mathcal{S}}) + \lambda_N\partial\|\widehat{\theta}^{\mathrm{o}}_{\mathcal{S}}\|_1 = \mathbf{0}. \tag{G.6}$$

By combining (G.5) with (G.6), we have

$$\widehat{\mathbf{S}}_{\mathcal{S}\mathcal{S}}\widehat{\theta}^{\mathrm{o}}_{\mathcal{S}} - \boldsymbol{e}_{\mathcal{S}} - \nabla\mathcal{H}_{\lambda_N}(\widehat{\theta}^{\mathrm{o}}_{\mathcal{S}}) + \lambda_N\partial\|\widehat{\theta}^{\mathrm{o}}_{\mathcal{S}}\|_1 = \mathbf{0}. \tag{G.7}$$

Now we consider

$$\begin{aligned}
\|\widehat{\mathbf{S}}_{\mathcal{S}^{\perp}\mathcal{S}}\widehat{\theta}^{\mathrm{o}}_{\mathcal{S}}\|_{\infty} &= (\widehat{\mathbf{S}}_{\mathcal{S}^{\perp}\mathcal{S}} - \overline{\Sigma}_{\mathcal{S}^{\perp}\mathcal{S}} + \overline{\Sigma}_{\mathcal{S}^{\perp}\mathcal{S}})(\widehat{\theta}^{\mathrm{o}}_{\mathcal{S}} - \theta^*_{\mathcal{S}} + \theta^*_{\mathcal{S}}) \\
&= \|(\widehat{\mathbf{S}}_{\mathcal{S}^{\perp}\mathcal{S}} - \overline{\Sigma}_{\mathcal{S}^{\perp}\mathcal{S}})\widehat{\Delta}^{\mathrm{o}}_{\mathcal{S}}\|_{\infty} + \|\overline{\Sigma}_{\mathcal{S}^{\perp}\mathcal{S}}\widehat{\Delta}^{\mathrm{o}}_{\mathcal{S}}\|_{\infty} + \|\widehat{\mathbf{S}}_{\mathcal{S}^{\perp}\mathcal{S}} - \overline{\Sigma}_{\mathcal{S}^{\perp}\mathcal{S}}\theta^*\|_{\infty} \\
&= \sqrt{s^*}\|\widehat{\mathbf{S}}_{\mathcal{S}^{\perp}\mathcal{S}} - \overline{\Sigma}_{\mathcal{S}^{\perp}\mathcal{S}}\|_{\max}\|\widehat{\Delta}^{\mathrm{o}}_{\mathcal{S}}\|_2 + \|\widehat{\Delta}^{\mathrm{o}}_{\mathcal{S}}\|_{\infty} + \|\widehat{\mathbf{S}}_{\mathcal{S}^{\perp}\mathcal{S}} - \overline{\Sigma}_{\mathcal{S}^{\perp}\mathcal{S}}\|_{\infty}\|\theta^*\|_1 \\
&= \frac{4\pi^2 M s^*\log d}{\rho_-(s^*)n} + \frac{2\sqrt{2}\pi M}{\rho_-(s^*)}\sqrt{\frac{s^*\log d}{n}} + \sqrt{2}\pi M\sqrt{\frac{\log d}{n}} \\
&= \left(\frac{\sqrt{2}\pi\psi_{\min}c_2}{(1+2c_3)\rho_-(s^*)} + \frac{2}{\rho_-(s^*)} + 1\right)\sqrt{2}\pi M\sqrt{\frac{s^*\log d}{n}}.
\end{aligned}$$

Therefore as long as

$$c_4 \geq \frac{\sqrt{2}\pi\psi_{\min}c_2}{(1+2c_3)\rho_-(s^*)} + \frac{2}{\rho_-(s^*)} + 1,$$

we have $\|\widehat{\mathbf{S}}_{\mathcal{S}^{\perp}\mathcal{S}}\widehat{\theta}^{\mathrm{o}}_{\mathcal{S}}\|_{\infty} \leq \lambda_N$, which implies that there exists $\boldsymbol{\xi} \in \partial\|\mathbf{0}\|_1$ such that

$$\widehat{\mathbf{S}}_{\mathcal{S}^{\perp}\mathcal{S}}\widehat{\theta}^{\mathrm{o}}_{\mathcal{S}} - \nabla\mathcal{H}_{\lambda_N}(\mathbf{0}) + \lambda_N\boldsymbol{\xi} = \mathbf{0}. \tag{G.8}$$

By combining (G.7) with (G.8), we know that $\widehat{\theta}^{\mathrm{o}}$ satisfies the KKT condition and is a local solution to (2.4).

Now we will show that $\widehat{\theta}^{\mathrm{o}}$ and $\bar{\theta}^{\lambda_N}$ are identical. Since $\|\bar{\theta}^{\lambda_N}_{\mathcal{S}^{\perp}}\| \leq \widetilde{s}$ and $\|\widehat{\theta}^{\mathrm{o}}_{\mathcal{S}^{\perp}}\| = 0$, we have

$$|\mathrm{supp}(\widehat{\theta}^{\mathrm{o}}) \cup \mathrm{supp}(\bar{\theta}^{\lambda_N})| \leq s^* + \widetilde{s}.$$

By the restricted strong convexity of $\mathcal{F}_{\lambda_N}$, we have

$$\begin{aligned}\mathcal{F}_{\lambda_N}(\bar{\boldsymbol{\theta}}^{\lambda_N}) &\geq \mathcal{F}_{\lambda_N}(\widehat{\boldsymbol{\theta}}^{\mathrm{o}}) + (\nabla\widetilde{\mathcal{L}}_{\lambda_N}(\widehat{\boldsymbol{\theta}}^{\mathrm{o}}) + \lambda_N\widetilde{\boldsymbol{\xi}}^{\mathrm{o}})^T(\bar{\boldsymbol{\theta}}^{\lambda_N} - \widehat{\boldsymbol{\theta}}^{\mathrm{o}}) + \frac{\widetilde{\rho}_-(s^*+\widetilde{s})}{2}\|\bar{\boldsymbol{\theta}}^{\lambda_N} - \widehat{\boldsymbol{\theta}}^{\mathrm{o}}\|_2^2, \\ &= \mathcal{F}_{\lambda_N}(\bar{\boldsymbol{\theta}}^{\lambda_N}) + \frac{\widetilde{\rho}_-(s^*+\widetilde{s})}{2}\|\bar{\boldsymbol{\theta}}^{\lambda_N} - \widehat{\boldsymbol{\theta}}^{\mathrm{o}}\|_2^2, \qquad \text{(G.9)} \\ \mathcal{F}_{\lambda_N}(\widehat{\boldsymbol{\theta}}^{\mathrm{o}}) &\geq \mathcal{F}_{\lambda_N}(\bar{\boldsymbol{\theta}}^{\lambda_N}) + (\nabla\widetilde{\mathcal{L}}_{\lambda_N}(\bar{\boldsymbol{\theta}}^{\lambda_N}) + \lambda_N\widetilde{\boldsymbol{\xi}})^T(\widehat{\boldsymbol{\theta}}^{\mathrm{o}} - \bar{\boldsymbol{\theta}}^{\lambda_N}) + \frac{\widetilde{\rho}_-(s^*+\widetilde{s})}{2}\|\widehat{\boldsymbol{\theta}}^{\mathrm{o}} - \bar{\boldsymbol{\theta}}^{\lambda_N}\|_2^2, \\ &= \mathcal{F}_{\lambda_N}(\bar{\boldsymbol{\theta}}^{\lambda_N}) + \frac{\widetilde{\rho}_-(s^*+\widetilde{s})}{2}\|\bar{\boldsymbol{\theta}}^{\lambda_N} - \widehat{\boldsymbol{\theta}}^{\mathrm{o}}\|_2^2, \qquad \text{(G.10)}\end{aligned}$$

where $\widetilde{\boldsymbol{\xi}}$ and $\widetilde{\boldsymbol{\xi}}^{\mathrm{o}}$ are defined as

$$\widetilde{\boldsymbol{\xi}} = \underset{\xi\in\partial\|\bar{\boldsymbol{\theta}}^{\lambda_N}\|_1}{\operatorname{argmin}} \|\nabla\widetilde{\mathcal{L}}_{\lambda_N}(\bar{\boldsymbol{\theta}}^{\lambda_N}) + \lambda_N\xi\|_\infty \quad \text{and} \quad \widetilde{\boldsymbol{\xi}}^{\mathrm{o}} = \underset{\xi\in\partial\|\widehat{\boldsymbol{\theta}}^{\mathrm{o}}\|_1}{\operatorname{argmin}} \|\nabla\widetilde{\mathcal{L}}_{\lambda_N}(\widehat{\boldsymbol{\theta}}^{\mathrm{o}}) + \lambda_N\xi\|_\infty.$$

By combining (G.9) with (G.10), we have $\|\widehat{\boldsymbol{\theta}}^{\mathrm{o}} - \bar{\boldsymbol{\theta}}^{\lambda_N}\|_2^2 = 0$, i.e., $\widehat{\boldsymbol{\theta}}^{\mathrm{o}} = \bar{\boldsymbol{\theta}}^{\lambda_N}$. Note that we choose $\lambda_N = c_4\sqrt{2}\pi M\sqrt{\log d/n}$, which is different from the selected regularization parameter in Assumption 4.8. But as long as we have $c_4\sqrt{s^*} \geq 8$, which is not an issue under the high dimensional scaling

$$M, s^*, n, d \to \infty \ \text{ and } \ Ms^*\log d/n \to 0,$$

$\lambda_N \geq 8\|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\|_\infty$ still holds with high probability. Since the above results universally hold over all columns of $\overline{\boldsymbol{\Theta}}^{\lambda_N}$ and $\boldsymbol{\Theta}^*$ under Assumptions 4.1 and (4.2), by Lemmas 4.8 and 4.9, we obtain $\widehat{\boldsymbol{\Theta}}^{\mathrm{o}} = \overline{\boldsymbol{\Theta}}^{\lambda_N}$, which completes the proof. □