Optimal Computational and Statistical Rates of Convergence for Sparse Nonconvex Learning Problems

Zhaoran Wang* and Han Liu† and Tong Zhang‡

Abstract

We provide theoretical analysis of the statistical and computational properties of penalized M-estimators that can be formulated as the solution to a possibly nonconvex optimization problem. Many important estimators fall in this category, including least squares regression with nonconvex regularization, generalized linear models with nonconvex regularization, and sparse elliptical random design regression. For these problems, it is intractable to calculate the global solution due to the nonconvex formulation. In this paper, we propose an approximate regularization path following method for solving a variety of learning problems with nonconvex objective functions. Under a unified analytic framework, we simultaneously provide explicit statistical and computational rates of convergence of any local solution obtained by the algorithm. Computationally, our algorithm attains a global geometric rate of convergence for calculating the full regularization path, which is optimal among all first-order algorithms. Unlike most existing methods that only attain geometric rates of convergence for one single regularization parameter, our algorithm calculates the full regularization path with the same iteration complexity. In particular, we provide a refined iteration complexity bound to sharply characterize the performance of each stage along the regularization path. Statistically, we provide sharp sample complexity analysis for all the approximate local solutions along the regularization path. In particular, our analysis improves upon existing results by providing a more refined sample complexity bound as well as an exact support recovery result for the final estimator. These results show that the final estimator attains an oracle statistical property due to the usage of nonconvex penalty.

1 Introduction

This paper considers the statistical and computational properties of a family of penalized M-estimators that can be formulated as

$$\widehat{\boldsymbol{\beta}}_\lambda \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d}\Big\{\mathcal{L}(\boldsymbol{\beta}) + \mathcal{P}_\lambda(\boldsymbol{\beta})\Big\}, \tag{1.1}$$

*Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; e-mail: zhaoran@princeton.edu.

†Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; e-mail: hanliu@princeton.edu.

‡Department of Statistics and Biostatistics, Rutgers University, Piscataway, NJ 08854, USA; e-mail: tzhang@stat.rutgers.edu.

where $\mathcal{L}(\boldsymbol{\beta})$ is a loss function while $\mathcal{P}_\lambda(\boldsymbol{\beta})$ is a penalty term with regularization parameter λ. A familiar example is the Lasso estimator (Tibshirani, 1996), in which $\mathcal{L}(\boldsymbol{\beta}) = \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2/(2n)$ and $\mathcal{P}_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$. Here $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \in \mathbb{R}^{n\times d}$ is the design matrix, $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ is the response vector, $\|\cdot\|_2$ is the Euclidean norm, and $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^d |\beta_j|$ is the ℓ_1 norm of $\boldsymbol{\beta}$. In general, we prefer the settings where both the loss function $\mathcal{L}(\boldsymbol{\beta})$ and the penalty term $\mathcal{P}_\lambda(\boldsymbol{\beta})$ in (1.1) are convex, since convexity makes both statistical and computational analysis convenient.

Though significant progress has been made on understanding convex penalized M-estimators (van de Geer, 2000; Bunea et al., 2007; van de Geer, 2008; Rothman et al., 2008; Wainwright, 2009; Bickel et al., 2009; Zhang, 2009; Koltchinskii, 2009b; Raskutti et al., 2011; Negahban et al., 2012), penalized M-estimators with nonconvex loss or penalty functions have recently attracted much interest because of their more attractive statistical properties. Unlike the ℓ_1 penalty, which induces significant estimation bias for parameters with large absolute values (Zhang and Huang, 2008), nonconvex penalties such as the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001) and minimax concave penalty (MCP) (Zhang, 2010a) can eliminate this estimation bias and attain more refined statistical rates of convergence. As another example of penalized M-estimators with nonconvex loss functions, we consider a semiparametric variant of the penalized least squares regression. Recall that a penalized least squares regression estimator can be formulated as

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_\lambda &\in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d} \left\{ \frac{1}{2n}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \mathcal{P}_\lambda(\boldsymbol{\beta}) \right\} \\ &= \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d} \left\{ \frac{1}{2}\left(1, -\boldsymbol{\beta}^T\right) \widehat{\mathbf{S}} \left(1, -\boldsymbol{\beta}^T\right)^T + \mathcal{P}_\lambda(\boldsymbol{\beta}) \right\}, \end{aligned}$$

where $\widehat{\mathbf{S}} = (\mathbf{y}, \mathbf{X})^T(\mathbf{y}, \mathbf{X})/n$ is the sample covariance matrix of a random vector $(Y, \boldsymbol{X}^T)^T \in \mathbb{R}^{d+1}$. When the design matrix $\mathbf{X}$ contains heavy-tail data, we may resort to the elliptical random design regression, which is a semiparametric extension of the Gaussian random design regression. More specifically, we replace the sample covariance matrix $\widehat{\mathbf{S}}$ with a possibly indefinite covariance matrix estimator $\widehat{\mathbf{K}}$ (to be defined in §2.2), which is more robust within the elliptical family. Since $\widehat{\mathbf{K}}$ does not guarantee to be positive semidefinite, the loss function

$$\mathcal{L}(\boldsymbol{\beta}) = (1, -\boldsymbol{\beta}^T)\widehat{\mathbf{K}}(1, -\boldsymbol{\beta}^T)^T$$

could be nonconvex. Another example of nonconvex loss functions is the corrected regression for error-in-variables linear models (Loh and Wainwright, 2012).

Though the global solutions of these nonconvex M-estimators enjoy nice statistical properties, it is in general computationally intractable to obtain the global solutions. Instead, a more realistic approach is to directly leverage standard optimization procedures to obtain a local solution $\widehat{\boldsymbol{\beta}}_\lambda$ that satisfies the first-order Karush-Kuhn-Tucker (KKT) condition

$$\mathbf{0} \in \partial\left\{\mathcal{L}\left(\widehat{\boldsymbol{\beta}}_\lambda\right) + \mathcal{P}_\lambda\left(\widehat{\boldsymbol{\beta}}_\lambda\right)\right\}, \tag{1.2}$$

where $\partial(\cdot)$ denotes the subgradient operator.

In the context of least squares regression with nonconvex penalties, several numerical procedures have been proposed to find the local solutions, including local quadratic approximation (LQA) (Fan

and Li, 2001), minorize-maximize (MM) algorithm (Hunter and Li, 2005), local linear approximation (LLA) (Zou and Li, 2008), and coordinate descent (Breheny and Huang, 2011; Mazumder et al., 2011). The theoretical properties of the local solutions obtained by these numerical procedures are in general unestablished. Only recently Zhang and Zhang (2012) showed that the gradient descent method initialized at a Lasso solution attains a unique local solution that has the same statistical properties as the global solution; Fan et al. (2012) proved that the LLA algorithm initialized with a Lasso solution attains a local solution with oracle statistical properties. Similar conclusion was also obtained by Zhang (2010b, 2012), where the LLA algorithm was referred to as multi-stage convex relaxation. However, each stage of the LLA algorithm requires that we exactly calculate the solution to a Lasso problem, which is not practical in applications. Therefore, the total computational complexity of the LLA algorithm is unclear.

In this paper, we propose an approximate regularization path following method for solving a general family of penalized M-estimators with possibly nonconvex loss or penalty functions. Our algorithm leverages the fast local convergence in the proximity of sparse solutions, which is also observed by Luo and Tseng (1992); Nesterov (2007); Hale et al. (2008); Wright et al. (2009); Agarwal et al. (2012); Xiao and Zhang (2012). More specifically, we consider a decreasing sequence of regularization parameters $\{\lambda_t\}_{t=0}^N$, where λ_0 corresponds to an all-zero solution, and $\lambda_N = \lambda_{\mathrm{tgt}}$ is the target regularization parameter that ensures the obtained estimator to achieve the optimal statistical rate of convergence. For each λ_t, we construct a sequence of local quadratic approximations of the loss function $\mathcal{L}(\boldsymbol{\beta})$, and utilize a variant of Nesterov's proximal-gradient method (Nesterov, 2007), which iterates over the updating step

$$\boldsymbol{\beta}_t^{(k+1)} \leftarrow \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d} \left\{ \mathcal{L}\big(\boldsymbol{\beta}_t^{(k)}\big) + \nabla\mathcal{L}\big(\boldsymbol{\beta}_t^{(k)}\big)^T\big(\boldsymbol{\beta} - \boldsymbol{\beta}_t^{(k)}\big) + \frac{L_t^{(k)}}{2}\big\|\boldsymbol{\beta} - \boldsymbol{\beta}_t^{(k)}\big\|_2^2 + \mathcal{P}_{\lambda_t}(\boldsymbol{\beta}) \right\}, \quad k = 1, 2, \ldots, \tag{1.3}$$

where $\boldsymbol{\beta}_t^{(k)}$ and $L_t^{(k)}$ corresponds to the k-th iteration of the proximal-gradient method for λ_t. Here $L_t^{(k)}$ is chosen by an adaptive line-search method, which will be specified in §3.2. Let $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ be an exact local solution satisfying (1.2) with regularization parameter λ_t. As illustrated in Figure 1, for each λ_t, our algorithm computes an approximation $\widetilde{\boldsymbol{\beta}}_t$ of the exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ up to certain optimization precision. Such an approximate local solution $\widetilde{\boldsymbol{\beta}}_t$ guarantees to be sparse, and therefore falls into the fast convergence region corresponding to λ_{t+1}. In this way, the resulting procedure attains a geometric rate of convergence within each path following stage, and therefore achieves a global geometric rate of convergence for calculating the entire regularization path. Moreover, without relying on the quality of the initial lasso solution as required by Zhang and Zhang (2012) and Fan et al. (2012), we establish the nonasymptotic statistical rates of convergence and oracle properties for all the approximate and exact local solutions along the full regularization path.

The idea of path following has been well-studied for convex sparse recovery problems (Osborne et al., 2000; Efron et al., 2004; Hastie et al., 2005; Park and Hastie, 2007; Zhao and Yu, 2007; Rosset and Zhu, 2007; Hale et al., 2008; Garrigues and Ghaoui, 2008; Wen et al., 2010; Friedman et al., 2010; Xiao and Zhang, 2012; Gärtner et al., 2012; Mairal and Yu, 2012). Among them, Xiao and Zhang (2012) proposed a proximal-gradient homotopy method for the least squares regression with ℓ_1 penalty. Compared to these previous works, we consider a broader family of nonconvex

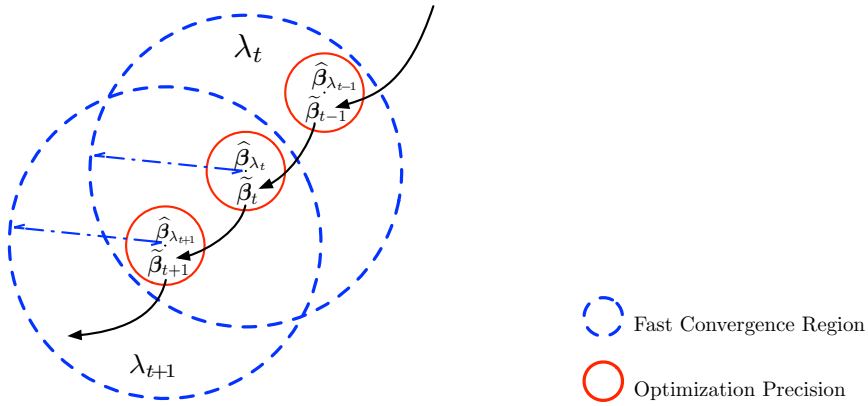


Figure 1: For regularization parameter λ_t, $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ is an exact local solution satisfying (1.2) with regularization parameter λ_t. Within the t-th path following stage, our algorithm achieves an approximate local solution $\widetilde{\boldsymbol{\beta}}_t$, which approximates the exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ up to certain optimization precision. Our approximate path following algorithm ensures that $\widetilde{\boldsymbol{\beta}}_t$ is sparse and therefore falls into the fast convergence region corresponding to regularization parameter λ_{t+1}.

M-estimators, including nonconvex penalty functions such as SCAD and MCP, as well as nonconvex loss functions such as semiparametric elliptical design loss. In particular, we provide sharp computational and statistical analysis for all the approximate and exact local solutions attained by the proposed approximate path following method.

The contributions of this paper are two folds: Computationally, we propose an optimization algorithm that ensures a global geometric rate of convergence for nonconvex sparse learning problems, which is the fastest achievable rate among all first-order methods. In detail, recall that N is the total number of path following stages. In the N-th path following stage, let ϵ_{opt} be the desired optimization precision of the approximate local solution $\widetilde{\boldsymbol{\beta}}_N$, we need no more than a logarithmic number of the proximal-gradient update iterations defined in (1.3) to calculate the entire path:

$$\text{Total \# of proximal-gradient iterations} \leq C \log\left(\frac{1}{\epsilon_{\text{opt}}}\right),$$

where $C > 0$ is a constant. Statistically, we prove that along the full regularization path, all the approximate local solutions obtained by our algorithm enjoy desirable statistical rates of convergence for estimating the true parameter vector $\boldsymbol{\beta}^*$. In detail, let s^* be the number of nonzero entries of $\boldsymbol{\beta}^*$, the approximate local solution $\widetilde{\boldsymbol{\beta}}_t$'s satisfy

$$\big\|\widetilde{\boldsymbol{\beta}}_t - \boldsymbol{\beta}^*\big\|_2 \leq C\lambda_t\sqrt{s^*}, \quad \text{for} \;\; t = 1, \ldots, N \tag{1.4}$$

with high probability. In particular, within the N-th path following stage, we have $\lambda_N = \lambda_{\text{tgt}} = C'\sqrt{\log d/n}$. Here C and C' are positive constants that do not dependent on d and n. In certain regimes, the final approximate local solution $\widetilde{\boldsymbol{\beta}}_N$ achieves the optimal statistical rate of convergence. Moreover, we prove that within the t-th path following stage, the iterative solution sequence $\{\boldsymbol{\beta}_t^{(k)}\}_{k=0}^{\infty}$ defined by (1.3) converges towards a unique exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$, which enjoys a more refined oracle statistical property. More specifically, let s_1^* be the number of "large" nonzero

coefficients of $\boldsymbol{\beta}^*$ and $s_2^* = s^* - s_1^*$ be the number of "small" nonzero coefficients (detailed definitions of s_1^* and s_2^* are provided in Theorem 4.7), we have

$$\left\|\widehat{\boldsymbol{\beta}}_{\lambda_t} - \boldsymbol{\beta}^*\right\|_2 \leq C\sqrt{\frac{s_1^*}{n}} + C'\sqrt{s_2^*}\lambda_t, \quad \text{for} \quad t = 1, \ldots, N \tag{1.5}$$

with high probability. In particular, for the final stage we have $\lambda_N = \lambda_{\text{tgt}} = C''\sqrt{\log d/n}$. Here C, C' and C'' are positive constants that don't dependent on d and n. Note that the oracle statistical property in (1.5) is significantly sharper than the rate of convergence in (1.4), e.g., when $s^* = s_1^*$ and $t = N$, the right-hand side of (1.4) is of the order $\sqrt{s^* \log d/n}$, while the right-hand side of (1.5) is of the order $\sqrt{s^*/n}$. Furthermore, we also prove that under suitable conditions, $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ exactly recovers the support of $\boldsymbol{\beta}^*$, i.e.,

$$\operatorname{supp}\left(\widehat{\boldsymbol{\beta}}_{\lambda_t}\right) = \operatorname{supp}(\boldsymbol{\beta}^*).$$

In summary, our joint analysis of the statistical and computational properties provides a theoretical characterization of the entire regularization path.

In an independent work, Loh and Wainwright (2013) considered similar problems and proved that all local solutions of various penalized M-estimators have good statistical properties if the loss and penalty functions satisfy the restricted strong convexity and other regularity conditions. Our results are different from theirs in two aspects: (i) They provided a set of sufficient conditions under which local optima have desired theoretical properties, and verified that the composite gradient descent algorithm satisfies these conditions. However, their conditions can not be applied to analyze our path following method, since we need to simultaneously analyze all the approximate local solutions along the entire regularization path. Our analysis of the full regularization path is a stronger result that requires more sophisticated proof techniques. (ii) Unlike their analysis, which provided a global characterization of local solutions but required additional regularity assumptions, our theoretical analysis of statistical performance is embedded in the analysis of the optimization procedure for the approximate local solutions attained by the procedure. In particular, our statistical results apply to all the approximate local solutions along the full regularization path, which is built upon a more fine-grained analysis of the sparsity pattern of all the intermediate solutions obtained from the proximal-gradient iterations. (iii) Moreover, in the regime where the absolute values of $\boldsymbol{\beta}^*$'s nonzero coefficients are "large", we provide a more refined oracle rate (1.5) of the local solutions along the regularization path, which clearly shows the theoretical benefits of nonconvex penalty functions over ℓ_1 regularization. Our statistical results are sharper than those provided by them, which are the same as using standard ℓ_1 regularization. In addition, we establish the exact support recovery results while they didn't.

The rest of this paper is organized as follows. First we briefly introduce some useful notation. In §2 we introduce sparse learning problems with possibly nonconvex loss and penalty functions. In §3 we introduce our approximate regularization path following method. In §4 we present the main theoretical results concerning the computational efficiency and statistical accuracy of the proposed procedure. In §5 we prove the theoretical results in §4. Numerical results are presented in §6.

Notation: Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T \in \mathbb{R}^d$. For $q \in [1, +\infty)$, we denote the ℓ_q norm of $\boldsymbol{\beta}$ by $\|\boldsymbol{\beta}\|_q = \left(\sum_{j=1}^d |\beta_j|^q\right)^{1/q}$. Specifically, we define $\|\boldsymbol{\beta}\|_\infty = \max_{1\leq j\leq d}\{|\beta_j|\}$ and $\|\boldsymbol{\beta}\|_0 = \operatorname{card}\{\operatorname{supp}(\boldsymbol{\beta})\}$, where $\operatorname{supp}(\boldsymbol{\beta}) = \{j : \beta_j \neq 0\}$ and $\operatorname{card}\{\cdot\}$ is the cardinality of a set. We denote the ℓ_q ball $\{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_q \leq R\}$

by $B_q(R)$. For a set S, we denote its cardinality by $|S|$ and its complement by $\bar{S}$. We define $\boldsymbol{\beta}_S \in \mathbb{R}^d$ and $\boldsymbol{\beta}_{\bar{S}} \in \mathbb{R}^d$ as

$$(\boldsymbol{\beta}_S)_j = \mathbb{I}(j \in S) \cdot \beta_j, \quad (\boldsymbol{\beta}_{\bar{S}})_j = \mathbb{I}(j \notin S) \cdot \beta_j, \quad \text{for } \ j = 1, \ldots, d \ \text{ and } \ S, \bar{S} \subseteq \{1, \ldots, d\},$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. We denote all-zero matrices by $\mathbf{0}$, and the diagonal matrix that has $x_1, \ldots, x_d$ on its diagonal by $\mathrm{diag}\{x_1, \ldots, x_d\}$. Meanwhile, let $\mathbf{M} \in \mathbb{R}^{d \times d}$ be a matrix, we overload $\mathrm{diag}(\mathbf{M})$ to be a diagonal matrix with diagonal entries $[\mathrm{diag}(\mathbf{M})]_{jj} = \mathbf{M}_{jj}$ $(j = 1, \ldots, d)$. For a function $f(\boldsymbol{\beta})$, we denote its gradient by $\nabla f(\boldsymbol{\beta})$ and its subgradient by $\partial f(\boldsymbol{\beta})$. Specifically, the derivative of a differentiable univariate function $f(x)$ is denoted by $f'(x)$. If random vectors $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ have the same distribution, we denote by $\boldsymbol{Z}_1 \overset{d}{=} \boldsymbol{Z}_2$. The d-dimensional ℓ_2 unit sphere is denoted by $\mathbb{S}^{d-1}$. Throughout this paper, we denote $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ to be the exact local solution and approximate local solution respectively. We index $\widehat{\boldsymbol{\beta}}$ with the corresponding regulation parameter λ, e.g., $\widehat{\boldsymbol{\beta}}_\lambda$. In the proposed method, we use subscript t to index the path following stages, e.g, the approximate local solution obtained within the t-th stage is denoted by $\widetilde{\boldsymbol{\beta}}_t$. Within the t-th stage, we index the proximal-gradient iterations with superscript (k), e.g., $\boldsymbol{\beta}_t^{(k)}$. For notational simplicity, we use generic absolute constants $C, C', \ldots$, whose value may change from line to line.

2 Some Nonconvex Sparse Learning Problems

Many theoretical results on penalized M-estimators rely on the condition that the loss and penalty functions are convex, since convexity makes both computational and statistical analysis convenient. However, the statistical performance of the estimator obtained from these convex formulations could be suboptimal in some settings. In the following, we introduce several nonconvex sparse learning problems as motivating examples.

2.1 Nonconvex Penalty

Throughout this paper, we consider decomposable penalty functions

$$\mathcal{P}_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{d} p_\lambda(\beta_j),$$

e.g., the ℓ_1 penalty $\lambda\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^d \lambda|\beta_j|$. When the minimum of $|\beta_j^*| > 0$ is not close to zero, the ℓ_1 penalty introduces large bias in parameter estimation. To remedy this effect, Fan and Li (2001) proposed the SCAD penalty

$$\begin{aligned} p_\lambda(\beta_j) &= \lambda \int_0^{|\beta_j|} \Big\{ \mathbb{I}(z \le \lambda) + \frac{(a\lambda - z)_+}{(a-1)\lambda} \mathbb{I}(z > \lambda) \Big\} \mathrm{d}z, \qquad (2.1) \\ &= \lambda|\beta_j| \cdot \mathbb{I}(|\beta_j| \le \lambda) - \big(\beta_j^2 - 2a\lambda|\beta_j| + \lambda^2\big)/\big(2(a-1)\big) \cdot \mathbb{I}(\lambda < |\beta_j| \le a\lambda) \\ &\quad + \frac{(a+1)\lambda^2}{2} \cdot \mathbb{I}(|\beta_j| > a\lambda), \qquad a > 2, \end{aligned}$$

and Zhang (2010a) proposed the MCP penalty

$$p_\lambda(\beta_j) = \lambda \int_0^{|\beta_j|} \Big(1 - \frac{z}{\lambda b}\Big)_+ \mathrm{d}z, \tag{2.2}$$
$$= \Big(\lambda|\beta_j| - \frac{\beta_j^2}{2b}\Big) \cdot \mathbb{1}(|\beta_j| \le b\lambda) + \frac{b\lambda^2}{2} \cdot \mathbb{1}(|\beta_j| > b\lambda), \qquad b > 0.$$

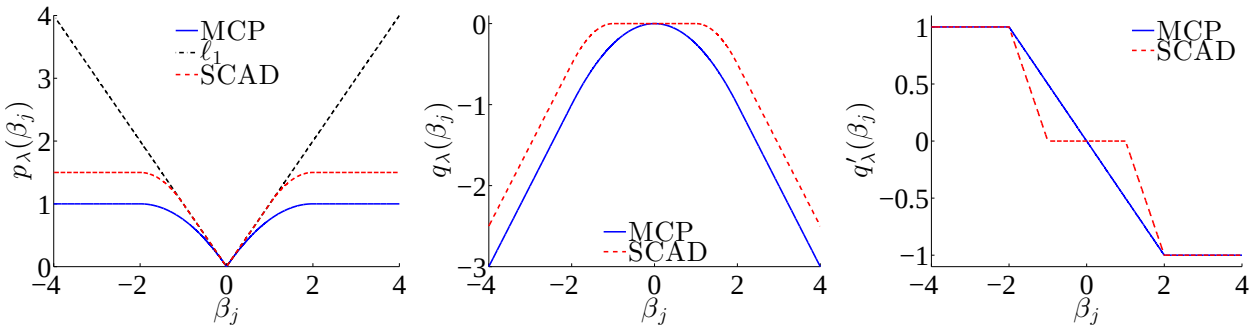
See Zhang and Zhang (2012) for a detailed survey. We illustrate these nonconvex penalty functions in Figure 2(a). These nonconvex penalties can be formulated as the sum of the ℓ_1 penalty and a concave part

$$p_\lambda(\beta_j) = \lambda|\beta_j| + q_\lambda(\beta_j), \tag{2.3}$$

where the specific forms of the concave component $q_\lambda(\beta_j)$ are

$$q_\lambda(\beta_j) = \begin{cases} \dfrac{2\lambda|\beta_j| - \beta_j^2 - \lambda^2}{2(a-1)} \cdot \mathbb{1}(\lambda < |\beta_j| \le a\lambda) + \dfrac{(a+1)\lambda^2 - 2\lambda|\beta_j|}{2} \cdot \mathbb{1}(|\beta_j| > a\lambda), & \text{SCAD}, \\ -\dfrac{\beta_j^2}{2b} \cdot \mathbb{1}(|\beta_j| \le b\lambda) + \Big(\dfrac{b\lambda^2}{2} - \lambda|\beta_j|\Big) \cdot \mathbb{1}(|\beta_j| > b\lambda), & \text{MCP}, \end{cases}$$

which are illustrated in Figure 2(b). The corresponding $q'_\lambda(\beta_j)$'s are also illustrated in Figure 2(c).



(a) (b) (c)

Figure 2: An illustration of nonconvex penalties: (a) Plots of $p_\lambda(\beta_j)$ for MCP, ℓ_1, and SCAD; (b) Plots of $q_\lambda(\beta_j)$ for MCP and SCAD; (c) Plots of $q'_\lambda(\beta_j)$ for MCP and SCAD. Here $p_\lambda(\beta_j)$ is the penalty function evaluated at the j-th dimension of $\boldsymbol{\beta}$, $q_\lambda(\beta_j)$ is the concave component of $p_\lambda(\beta_j)$, and $q'_\lambda(\beta_j)$ is the derivative of $q_\lambda(\beta_j)$. Here we set $a = 2.1$ for SCAD, $b = 2$ for MCP, and $\lambda = 1$.

In fact, our method and theory are not limited to these specific forms of $p_\lambda(\beta_j)$ and $q_\lambda(\beta_j)$. More generally, we only rely on the following regularity conditions on the concave component $q_\lambda(\beta_j)$:

Regularity Conditions on Nonconvex Penalty

(a) $q'_\lambda(\beta_j)$ is monotone and Lipschitz continuous, i.e., for $\beta'_j > \beta_j$, there exist two constants $\zeta_- \ge 0$ and $\zeta_+ \ge 0$ such that

$$-\zeta_- \le \frac{q'_\lambda(\beta'_j) - q'_\lambda(\beta_j)}{\beta'_j - \beta_j} \le -\zeta_+ \le 0;$$

(b) $q_\lambda(\beta_j)$ is symmetric, i.e., $q_\lambda(-\beta_j) = q_\lambda(\beta_j)$ for any β_j;

(c) $q_\lambda(\beta_j)$ and $q'_\lambda(\beta_j)$ pass through the origin, i.e., $q_\lambda(0) = q'_\lambda(0) = 0$;

(d) $q'_\lambda(\beta_j)$ is bounded, i.e., $|q'_\lambda(\beta_j)| \leq \lambda$ for any β_j;

(e) $q'_\lambda(\beta_j)$ has bounded difference with respect to λ: $\left|q'_{\lambda_1}(\beta_j) - q'_{\lambda_2}(\beta_j)\right| \leq |\lambda_1 - \lambda_2|$ for any β_j.

In regularity condition (a), ζ_- and ζ_+ are in fact two parameters that control the concavity of $q_\lambda(\beta_j)$. Note that the second order derivative of a function characterizes its convexity/concavity. Taking $\beta'_j \to \beta_j$ in regularity condition (a), we have $q''_\lambda(\beta_j) \in [-\zeta_-, -\zeta_+]$ (here we ignore those β_j's where $q''_\lambda(\beta_j)$ doesn't exist), which suggests larger ζ_- and ζ_+ allow $q_\lambda(\beta_j)$ to be more concave. For SCAD, we take $\zeta_- = 1/(a-1)$ and $\zeta_+ = 0$. For MCP, we take $\zeta_- = 1/b$ and $\zeta_+ = 0$. In Figure 2(b) and Figure 2(c), we can verify that regularity conditions (a)−(d) hold for MCP and SCAD. For MCP, we illustrate regularity condition (e) in Figure 5(a) of Appendix A. For SCAD, we illustrate property (e) in Figure 5(b) (for $\lambda_2 \geq a\lambda_1$) and Figure 5(c) of Appendix A (for $\lambda_2 < a\lambda_1$).

By (2.3) we have $\mathcal{P}_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^d p_\lambda(\beta_j) = \lambda\|\boldsymbol{\beta}\|_1 + \sum_{j=1}^d q_\lambda(\beta_j)$. For notational simplicity, we define

$$\mathcal{Q}_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^d q_\lambda(\beta_j) = \mathcal{P}_\lambda(\boldsymbol{\beta}) - \lambda\|\boldsymbol{\beta}\|_1. \tag{2.4}$$

Hence $\mathcal{Q}_\lambda(\boldsymbol{\beta})$ denotes the decomposable concave component of the nonconvex penalty $\mathcal{P}_\lambda(\boldsymbol{\beta})$.

2.2 Nonconvex Loss Function

In this paper, we focus on an example of nonconvex loss function named semiparametric elliptical design regression. Recall that the elliptical distribution is defined as:

Definition 2.1 (Elliptical distribution)**.** For $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^T \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ with $\mathrm{rank}(\boldsymbol{\Sigma}) = k \leq d$, a random vector $\boldsymbol{W} = (W_1, \ldots, W_d)^T$ follows an elliptical distribution denoted by $\mathrm{EC}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \Xi)$, if and only if

$$\boldsymbol{W} \overset{d}{=} \boldsymbol{\mu} + \Xi\mathbf{A}\boldsymbol{U}.$$

Here $\boldsymbol{U}$ is a random vector uniformly distributed on the unit sphere $\mathbb{S}^{k-1}$; $\Xi \geq 0$ is a scalar random variable independent of $\boldsymbol{U}$; $\mathbf{A} \in \mathbb{R}^{d\times q}$ is a deterministic matrix such that $\mathbf{A}\mathbf{A}^T = \boldsymbol{\Sigma}$. We call $\boldsymbol{\Sigma}$ the scatter matrix. The generalized correlation matrix is defined as $\boldsymbol{\Sigma}^0 = \mathrm{diag}(\boldsymbol{\Sigma})^{-1/2} \cdot \boldsymbol{\Sigma} \cdot \mathrm{diag}(\boldsymbol{\Sigma})^{-1/2}$. When $\mathbb{E}(\Xi^2)$ exists, $\boldsymbol{\Sigma}^0$ is the correlation matrix of $\boldsymbol{W}$.

Remark 2.2. Note that simultaneously scaling Ξ and $\boldsymbol{U}$ (e.g., $\Xi \to \Xi/C$ and $\boldsymbol{U} \to \boldsymbol{U}/C$, where C is a constant) leads to the same elliptical distribution. To make this model identifiable, we assume $\mu_j = \mathbb{E}(W_j)$ and $\boldsymbol{\Sigma}_{jj} = \mathrm{Var}(W_j)$.

Remark 2.3. The elliptical distribution family includes a variety of possibly heavy-tail distributions: multivariate Gaussian, multivariate Cauchy, Student's t, logistic, Kotz, symmetric Pearson type-II and type-VII distributions.

For semiparametric elliptical design regression, we have n pairs of observations $\mathbf{z}_1 = (y_1, \mathbf{x}_1^T)^T$, $\ldots$, $\mathbf{z}_n = (y_n, \mathbf{x}_n^T)^T$ of a random vector $\boldsymbol{Z} = (Y, \boldsymbol{X}^T)^T \in \mathbb{R}^{d+1}$ that follows the $(d+1)$-dimensional elliptical distribution defined in Definition 2.1. We can verify that $(Y|\boldsymbol{X} = \mathbf{x})$ follows a univariate elliptical distribution. We assume $\mathbb{E}(Y|\boldsymbol{X} = \mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}^*$. Then we can define the population version of the semiparametric elliptical design regression estimator as

$$\breve{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2}\mathbb{E}_{\boldsymbol{X},Y}\left(\left(Y - \boldsymbol{X}^T\boldsymbol{\beta}\right)^2 \right) + \mathcal{P}_\lambda(\boldsymbol{\beta}) \right\} = \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2}\left(1, -\boldsymbol{\beta}^T\right) \boldsymbol{\Sigma}_{\boldsymbol{Z}} \left(1, -\boldsymbol{\beta}^T\right)^T + \mathcal{P}_\lambda(\boldsymbol{\beta}) \right\}. \quad (2.5)$$

The above procedure is not practically implementable, since the population covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{Z}}$ is unknown in (2.5). In practice, we need to estimate the population covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{Z}}$. For this, we propose a rank-based covariance matrix estimator $\widehat{\mathbf{K}}_{\boldsymbol{Z}}$, which is obtained in two steps as described below:

Elliptical Covariance Matrix Estimation

S1. In the first step, we define a rank-based estimator $\widehat{\mathbf{R}}_{\boldsymbol{Z}}$ for the generalized correlation matrix $\boldsymbol{\Sigma}^0_{\boldsymbol{Z}}$ using the Kendall's tau statistic. Let $\mathbf{z}_1, \ldots, \mathbf{z}_n \in \mathbb{R}^{d+1}$ with $\mathbf{z}_i = (z_{i1}, \ldots, z_{i(d+1)})^T$ be n independent observations of $\boldsymbol{Z}$. The Kendall's tau correlation coefficient is defined as

$$\widehat{\tau}_{jk}(\mathbf{z}_1, \ldots, \mathbf{z}_n) = \begin{cases} \displaystyle\sum_{1\le i<i'<n} \frac{2\,\mathrm{sign}\left(z_{ij} - z_{i'j}\right)\mathrm{sign}\left(z_{ik} - z_{i'k}\right)}{n(n-1)}, & \text{for } j \neq k, \\ 1, & \text{for } j = k. \end{cases}$$

We define the Kendall's tau correlation matrix estimator as

$$\widehat{\mathbf{R}}_{\boldsymbol{Z}} = \left[\left(\widehat{\mathbf{R}}_{\boldsymbol{Z}}\right)_{jk} \right] = \left[\sin\left(\frac{\pi}{2}\widehat{\tau}_{jk}\left(\mathbf{z}_1, \ldots, \mathbf{z}_n\right) \right) \right]. \quad (2.6)$$

Han and Liu (2012); Liu et al. (2012); Han and Liu (2013) showed that $\widehat{\mathbf{R}}_{\boldsymbol{Z}}$ is a robust estimator of the population generalized correlation matrix $\boldsymbol{\Sigma}^0_{\boldsymbol{Z}}$, and is invariant to different generating variable Ξ within the whole elliptical family.

S2. In the second step, we construct a covariance matrix estimator

$$\widehat{\mathbf{K}}_{\boldsymbol{Z}} = \left[\left(\widehat{\mathbf{K}}_{\boldsymbol{Z}}\right)_{jk} \right] = \left[\left(\widehat{\mathbf{R}}_{\boldsymbol{Z}}\right)_{jk} \cdot \widehat{\sigma}_j\widehat{\sigma}_k \right], \quad (2.7)$$

where $\widehat{\sigma}_1, \ldots, \widehat{\sigma}_{d+1}$ are the estimators of the standard deviations of $Z_1, \ldots, Z_{d+1}$. We calculate $\widehat{\sigma}_1, \ldots, \widehat{\sigma}_{d+1}$ using the Catoni's M-estimator (Catoni, 2012) described in Appendix D. The main advantage of the Cantoni's estimator is that, for a fixed confidence level, it achieves the same deviation behavior as a Gaussian random variable under a weak moment condition.

Note that $\widehat{\mathbf{K}}_{\boldsymbol{Z}}$ is not necessarily positive semidefinite, which implies that the loss function $\mathcal{L}(\boldsymbol{\beta})$ in semiparametric elliptical design regression

$$\mathcal{L}(\boldsymbol{\beta}) = \left(1, -\boldsymbol{\beta}^T\right) \widehat{\mathbf{K}}_{\boldsymbol{Z}} \left(1, -\boldsymbol{\beta}^T\right)^T$$

is possibly nonconvex.

3 Approximate Regularization Path Following Method

Before we get into details, we first present the high level idea of approximate regularization path following. We then introduce the basic building block of our path following method — a proximal-gradient method tailored to nonconvex problems.

3.1 Approximate Regularization Path Following

Fast local geometric convergence in the proximity of sparse solutions has been observed by many authors (Hale et al., 2008; Wright et al., 2009; Wen et al., 2010; Agarwal et al., 2012; Xiao and Zhang, 2012). We exploit such fast local convergence under an approximate path framework to achieve fast global convergence.

Initialization: In (1.1), when the regularization parameter λ is sufficiently large, the solution to sparse learning problems is an all-zero vector. Recall that any exact local solution $\widehat{\boldsymbol{\beta}}_\lambda$ satisfies the first-order optimality condition, $\mathbf{0} \in \partial\{\mathcal{L}(\widehat{\boldsymbol{\beta}}_\lambda) + \mathcal{P}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda)\}$. Since the nonconvex penalty $\mathcal{P}_\lambda(\boldsymbol{\beta})$ can be formulated as $\mathcal{P}_\lambda(\boldsymbol{\beta}) = \mathcal{Q}_\lambda(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1$, where $\mathcal{Q}_\lambda(\boldsymbol{\beta})$ is defined in (2.4), the first-order optimality condition implies there should exist some subgradient $\boldsymbol{\xi} \in \partial\|\widehat{\boldsymbol{\beta}}_\lambda\|_1$ such that

$$\mathbf{0} = \nabla\mathcal{L}(\widehat{\boldsymbol{\beta}}_\lambda) + \mathcal{Q}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda) + \lambda\boldsymbol{\xi}. \tag{3.1}$$

Let λ be chosen such that $\widehat{\boldsymbol{\beta}}_\lambda = \mathbf{0}$. By regularity condition (c), we have $\nabla\mathcal{Q}_\lambda(\mathbf{0}) = \mathbf{0}$. Meanwhile, since $\boldsymbol{\xi} \in \partial\|\mathbf{0}\|_1$, we have $\|\boldsymbol{\xi}\|_\infty \leq 1$, which implies $\|\nabla\mathcal{L}(\mathbf{0})\|_\infty \leq \lambda$ in (3.1). Hence, $\lambda_0 = \|\nabla\mathcal{L}(\mathbf{0})\|_\infty$ is the smallest regularization parameter such that any exact local solution $\widehat{\boldsymbol{\beta}}_\lambda$ to the minimization problem (1.1) is all-zero. We choose this λ_0 to be the initial parameter of our regularization path.

Approximate Path Following: Let $\lambda_{\text{tgt}} \in (0, \lambda_0)$ be the target regularization parameter in (1.1). We consider a decreasing sequence of regularization parameters $\{\lambda_t\}_{t=0}^N$, where

$$\lambda_t = \eta^t\lambda_0 \quad (t = 0, \ldots, N), \qquad \lambda_N = \lambda_{\text{tgt}}, \qquad \text{and} \quad \eta \in [0.9, 1). \tag{3.2}$$

Here η is an absolute constant that doesn't scale with sample size n and dimension d. In §4 and §5 we will show that such a range of η ensures the global geometric rate of convergence. Consequently, since we have $\lambda_{\text{tgt}} = \lambda_0\eta^N$ in (3.2), the number of path following stages is

$$N = \frac{\log(\lambda_0/\lambda_{\text{tgt}})}{\log(\eta^{-1})}. \tag{3.3}$$

Without loss of generality, we assume that η is properly chosen such that N is an integer. We will show in §4 that, λ_{tgt} scales with sample size n and dimension d. Since η is a constant, the number of stages N also scales with n and d. Within the t-th ($t = 1, \ldots, N$) path following stage, we aim to obtain a local solution to the minimization problem $\min\{\mathcal{L}(\boldsymbol{\beta}) + \mathcal{P}_{\lambda_t}(\boldsymbol{\beta})\}$.

As shown in Lines 5–9 of Algorithm 1, within the t-th ($t = 1, \ldots, N-1$) path following stage, we exploit a variant of proximal-gradient method for nonconvex problems (Algorithm 3) to obtain an approximate solution $\widetilde{\boldsymbol{\beta}}_t$ that corresponds to the regularization parameter $\lambda_t = \eta^t\lambda_0$. To ensure that each path following stage enjoys a fast geometric rate of convergence, we employ an approximation path following strategy. More specifically, we use the approximate local solution $\widetilde{\boldsymbol{\beta}}_{t-1}$ obtained within the $(t-1)$-th path following stage to initialize the t-th stage (Line 8 and Line

12 of Algorithm 1). Recall that we need to adaptively search for the best $L_t^{(k)}$ ($k = 0, 1, \ldots$) in (1.3). To achieve computational efficiency, within the $(t-1)$-th path following stage, we store the chosen $L_{t-1}^{(k)}$ at the last proximal-gradient iteration as L_{t-1}. Within the t-th stage we initialize the search for $L_t^{(0)}$ with L_{t-1} (Line 8 and Line 12 of Algorithm 1), which will be explained in §3.2.

Algorithm 1 The approximate path following method, which solves for a decreasing sequence of regularization parameters $\{\lambda_t\}_{t=0}^N$. Within the t-th path following stage, we employ the proximal-gradient method illustrated in Algorithm 3 to achieve an approximate local solution $\widetilde{\boldsymbol{\beta}}_t$ for λ_t. This approximate local solution is then used to initialize the $(t+1)$-th stage.

1: $\{\widetilde{\boldsymbol{\beta}}_t\}_{t=1}^N \leftarrow$ Approximate-Path-Following$(\lambda_{\text{tgt}}, \epsilon_{\text{opt}})$
2: **input:** $\lambda_{\text{tgt}} > 0, \epsilon_{\text{opt}} > 0$ {Here we set $\epsilon_{\text{opt}} \ll \lambda_{\text{tgt}}/4$.}
3: **parameter:** $\eta \in [0.9, 1), R > 0$, $L_{\min} > 0$, $\lambda_0 = \|\nabla \mathcal{L}(\mathbf{0})\|_\infty$
{For logistic loss, we set $R \in (0, +\infty)$; For other loss functions, we set $R = +\infty$.}
{In practice, we set $L_{\min}$ to be a sufficiently small value, e.g., 10^{-6}.}
4: **initialize:** $\widetilde{\boldsymbol{\beta}}_0 \leftarrow \mathbf{0}, L_0 \leftarrow L_{\min}, N \leftarrow \log(\lambda_0/\lambda_{\text{tgt}})/\log(\eta^{-1})$
5: **for** $t = 1, \ldots, N-1$ **do**
6: $\quad \lambda_t \leftarrow \eta^t \lambda_0$
7: $\quad \epsilon_t \leftarrow \lambda_t/4$
8: $\quad \{\widetilde{\boldsymbol{\beta}}_t, L_t\} \leftarrow$ Proximal-Gradient$\big(\lambda_t, \epsilon_t, \widetilde{\boldsymbol{\beta}}_{t-1}, L_{t-1}, R\big)$ as in Algorithm 3
9: **end for**
10: $\lambda_N \leftarrow \lambda_{\text{tgt}}$
11: $\epsilon_N \leftarrow \epsilon_{\text{opt}}$
12: $\{\widetilde{\boldsymbol{\beta}}_N, L_N\} \leftarrow$ Proximal-Gradient$\big(\lambda_N, \epsilon_N, \widetilde{\boldsymbol{\beta}}_{N-1}, L_{N-1}, R\big)$
13: **return** $\{\widetilde{\boldsymbol{\beta}}_t\}_{t=1}^N$

Configuration of Optimization Precision: We set the optimization precision ϵ_t for the t-th ($t = 1, \ldots, N-1$) stage to be $\lambda_t/4$ (Line 7 of Algorithm 1). Within the N-th path following stage where $\lambda_N = \lambda_{\text{tgt}}$ (Line 10), we solve up to high optimization precision $\epsilon_{\text{opt}} \ll \lambda_{\text{tgt}}/4$ (Line 11). The intuition behind such a configuration of optimization precision is explained as follows:

- For $t = 1, \ldots, N-1$, recall the exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ is an estimator of the true parameter vector $\boldsymbol{\beta}^*$ corresponding to the regularization parameter λ_t. According to high-dimensional statistical theory, the statistical error of $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ should be upper bounded by $C\lambda_t\sqrt{s^*}$ with high probability, where $s^* = \|\boldsymbol{\beta}^*\|_0$. In Lemma 5.1 we will prove that, if the optimization error of the approximate local solution $\widetilde{\boldsymbol{\beta}}_t$ is at most $\lambda_t/4$, then $\widetilde{\boldsymbol{\beta}}_t$ lies within a ball of radius $C'\lambda_t\sqrt{s^*}$ centered at $\boldsymbol{\beta}^*$ with high probability. That is to say, the approximate local solution $\widetilde{\boldsymbol{\beta}}_t$ has the same order of statistical error as the exact solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$, and therefore enjoys certain desired statistical recovery properties. In particular, in Theorem 5.5 we will prove that, $\widetilde{\boldsymbol{\beta}}_t$ is guaranteed to be sparse, and thus falls into the fast convergence region of the next path following stage.

- However, for $t = N$, we need to solve up to high optimization precision $\epsilon_{\text{opt}} \ll \lambda_{\text{tgt}}/4$. This is because, even though $\widetilde{\boldsymbol{\beta}}_t$ and $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ both have statistical error of the order $\lambda_t\sqrt{s^*}$, in certain

regimes the exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ is able to achieve an improved recovery performance due to the usage of nonconvex penalties (as shown in (1.5), which will be proved in Theorem 4.8). Therefore, within the final stage we need to obtain an approximate solution $\widetilde{\boldsymbol{\beta}}_N$ as close to the exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_{\text{tgt}}}$ as possible, so that $\widetilde{\boldsymbol{\beta}}_N$ has a faster statistical rate of convergence.

In Algorithm 1, $R > 0$ (Line 3) is a parameter that decides the radius of the constraint that is used in the proximal-gradient method (Line 8 and Line 12). In detail, for least squares loss and semiparametric elliptical design loss, we do not need any constraint. Therefore, we set $R = +\infty$. However, for logistic loss we need to impose an ℓ_2 constraint of radius $R \in (0, +\infty)$. Here $L_{\min}$ is a parameter used in the proximal-gradient method (Line 3 of Algorithm 3), which is often set to be a sufficiently small value in practice, e.g., $L_{\min} = 10^{-6}$. We will explain with detail in §3.2.

3.2 Proximal-Gradient Method for Nonconvex Problems

Before we introduce our proximal-gradient method that is tailored to nonconvex problems, we first give a brief introduction to Nesterov's proximal-gradient method (Nesterov, 2007), which solves the following convex optimization problem

$$\text{minimize } \phi_\lambda(\boldsymbol{\beta}), \quad \text{where } \phi_\lambda(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \mathcal{P}_\lambda(\boldsymbol{\beta}), \quad \boldsymbol{\beta} \in \Omega. \tag{3.4}$$

Here $\mathcal{L}(\boldsymbol{\beta})$ is convex and differentiable, $\mathcal{P}_\lambda(\boldsymbol{\beta})$ is convex but possibly nonsmooth, and Ω is a closed convex set.

Recall that $\boldsymbol{\beta}_t^{(k)}$ corresponds to the k-th iteration of the proximal-gradient method within the t-th path following stage. Nesterov's proximal-gradient method updates $\boldsymbol{\beta}_t^{(k)}$ to be the minimizer of the following local quadratic approximation of $\phi_{\lambda_t}(\boldsymbol{\beta})$ at $\boldsymbol{\beta}_t^{(k-1)}$

$$\psi_{L_t^{(k)},\lambda_t}(\boldsymbol{\beta}; \boldsymbol{\beta}_t^{(k-1)}) = \mathcal{L}(\boldsymbol{\beta}_t^{(k-1)}) + \nabla\mathcal{L}(\boldsymbol{\beta}_t^{(k-1)})^T(\boldsymbol{\beta} - \boldsymbol{\beta}_t^{(k-1)}) + \frac{L_t^{(k)}}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_t^{(k-1)}\|_2^2 + \mathcal{P}_{\lambda_t}(\boldsymbol{\beta}), \tag{3.5}$$

where $L_t^{(k)} > 0$ is chosen by line search.

However, Nesterov's proximal-gradient method requires both $\mathcal{L}(\boldsymbol{\beta})$ and $\mathcal{P}_\lambda(\boldsymbol{\beta})$ in (3.4) to be convex. However, in the optimization problem (1.1) considered in this paper, $\mathcal{L}(\boldsymbol{\beta})$ and $\mathcal{P}_\lambda(\boldsymbol{\beta})$ may no longer be convex. To extend the proximal-gradient method to nonconvex settings, we adopt an alternative formulation of the objective function.

Recall that the nonconvex penalty can be written as $\mathcal{P}_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1 + \mathcal{Q}_\lambda(\boldsymbol{\beta})$, where $\mathcal{Q}_\lambda(\boldsymbol{\beta})$ is defined in (2.4). For notational simplicity, we denote $\mathcal{L}(\boldsymbol{\beta}) + \mathcal{Q}_\lambda(\boldsymbol{\beta})$ by $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})$. Consequently, the objective function $\phi_\lambda(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \mathcal{P}_\lambda(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \mathcal{Q}_\lambda(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1$ can be reformulated as

$$\phi_\lambda(\boldsymbol{\beta}) = \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1, \tag{3.6}$$

where we can view $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})$ as a surrogate loss function and $\lambda\|\boldsymbol{\beta}\|_1$ as a new penalty function. Such a reformulation ensures the convexity of the new penalty function. Moreover, in Lemma 5.1 we will prove that, the surrogate loss function $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})$ is actually strongly convex under certain conditions, which guarantee to hold along the full regularization path. Correspondingly, we modify Nesterov's proximal-gradient method to minimize the local quadratic approximation defined as

$$\psi_{L_t^{(k)},\lambda_t}(\boldsymbol{\beta}; \boldsymbol{\beta}_t^{(k-1)}) = \widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)}) + \nabla\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)})^T(\boldsymbol{\beta} - \boldsymbol{\beta}_t^{(k-1)}) + \frac{L_t^{(k)}}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_t^{(k-1)}\|_2^2 + \lambda_t\|\boldsymbol{\beta}\|_1. \tag{3.7}$$

Unlike (3.5), we use a quadratical approximation to the surrogate loss function $\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta})$ in (3.7), but instead of the original loss function $\mathcal{L}(\boldsymbol{\beta})$. At the k-th iteration of the proximal-gradient method, we update $\boldsymbol{\beta}_t^{(k)}$ to be the minimizer of the quadratic approximation defined in (3.7), i.e.,

$$\boldsymbol{\beta}_t^{(k)} \leftarrow \underset{\boldsymbol{\beta}\in\Omega}{\operatorname{argmin}} \left\{ \psi_{L_t^{(k)},\lambda_t}\big(\boldsymbol{\beta}; \boldsymbol{\beta}_t^{(k-1)}\big) \right\}. \tag{3.8}$$

Now we specify the constraint set Ω in (3.8). For $\mathcal{L}(\boldsymbol{\beta})$ being least squares or semiparametric elliptical design loss, we set $\Omega = \mathbb{R}^d$. For logistic loss, we set $\Omega = B_2(R)$ with $R \in (0, +\infty)$, where $B_2(R)$ is a centered ℓ_2 ball of radius R. In Lemma 5.1 we will show that, in the setting of logistic loss, the boundedness of $\|\boldsymbol{\beta}_t^{(k)}\|_2$'s is essential for establishing the strong convexity of the surrogate loss function $\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta})$ along the full regularization path. To unify the notations, we consider $\Omega = B_2(R)$ throughout — when the constraint set $\Omega = \mathbb{R}^d$, we set $R = +\infty$. Correspondingly, we denote (3.8) by

$$\boldsymbol{\beta}_t^{(k)} \leftarrow \mathcal{T}_{L_t^{(k)},\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}; R\big). \tag{3.9}$$

In the sequel, we provide the detailed update schemes for the nonconvex problems discussed in §2:

Update Schemes of Proximal-Gradient Method for Nonconvex Problems

- When $\Omega = \mathbb{R}^d$, $\mathcal{T}_{L_t^{(k)},\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}; +\infty\big)$ is a soft-thresholding operator taking the form

$$\Big(\mathcal{T}_{L_t^{(k)},\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}; +\infty\big)\Big)_j = \begin{cases} 0 & \text{if } \big|\bar{\beta}_j\big| \le \lambda_t / L_t^{(k)}, \\ \operatorname{sign}\big(\bar{\beta}_j\big)\big(\big|\bar{\beta}_j\big| - \lambda_t / L_t^{(k)}\big) & \text{if } \big|\bar{\beta}_j\big| > \lambda_t / L_t^{(k)}, \end{cases} \tag{3.10}$$

  for $j = 1, \ldots, d$, where

$$\bar{\boldsymbol{\beta}} = \boldsymbol{\beta}_t^{(k-1)} - \frac{1}{L_t^{(k)}} \nabla \widetilde{\mathcal{L}}_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}\big) = \boldsymbol{\beta}_t^{(k-1)} - \frac{1}{L_t^{(k)}} \Big(\nabla \mathcal{L}\big(\boldsymbol{\beta}_t^{(k-1)}\big) + \nabla \mathcal{Q}_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}\big) \Big), \tag{3.11}$$

  and $\bar{\beta}_j$ is the j-th dimension of $\bar{\boldsymbol{\beta}}$.

- When $\Omega = B_2(R)$, $\mathcal{T}_{L_t^{(k)},\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}; R\big)$ is obtained by projecting $\mathcal{T}_{L_t^{(k)},\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}; +\infty\big)$ defined in (3.10) onto $B_2(R)$, i.e.,

$$\mathcal{T}_{L_t^{(k)},\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}; R\big) = \begin{cases} \mathcal{T}_{L_t^{(k)},\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}; +\infty\big) & \text{if } \big\|\mathcal{T}_{L_t^{(k)},\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}; +\infty\big)\big\|_2 < R, \\ \dfrac{R \cdot \mathcal{T}_{L_t^{(k)},\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}; +\infty\big)}{\big\|\mathcal{T}_{L_t^{(k)},\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}; +\infty\big)\big\|_2} & \text{if } \big\|\mathcal{T}_{L_t^{(k)},\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}; +\infty\big)\big\|_2 \ge R. \end{cases} \tag{3.12}$$

See Appendix B for a detailed derivation. In the following, we specify $\nabla \mathcal{L}(\boldsymbol{\beta})$ and $\nabla \mathcal{Q}_{\lambda_t}(\boldsymbol{\beta})$ in (3.11) for the nonconvex problems discussed in §2:

- For the (nonconvex) loss functions discussed in §2, $\nabla\mathcal{L}(\boldsymbol{\beta})$ takes the forms of

$$\nabla\mathcal{L}(\boldsymbol{\beta}) = \begin{cases} \frac{1}{n}\mathbf{X}^T\big(\mathbf{X}\boldsymbol{\beta}-\mathbf{y}\big), & \text{least squares loss,} \\ \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\Big(\frac{\exp\big(\mathbf{x}_i^T\boldsymbol{\beta}\big)}{1+\exp\big(\mathbf{x}_i^T\boldsymbol{\beta}\big)}-y_i\Big), & \text{logistic loss,} \\ \widehat{\mathbf{K}}_{\boldsymbol{X}}\boldsymbol{\beta}-\widehat{\mathbf{K}}_{\boldsymbol{X},Y}, & \text{semiparametric elliptical design loss,} \end{cases}$$

  where $\widehat{\mathbf{K}}_{\boldsymbol{X}} \in \mathbb{R}^{d\times d}$ and $\widehat{\mathbf{K}}_{\boldsymbol{X},Y} \in \mathbb{R}^{d\times 1}$ are defined as the submatrices of $\widehat{\mathbf{K}}_{\boldsymbol{Z}}$, i.e.,

$$\widehat{\mathbf{K}}_{\boldsymbol{Z}} = \begin{pmatrix} \widehat{\mathbf{K}}_Y & \widehat{\mathbf{K}}_{\boldsymbol{X},Y}^T \\ \widehat{\mathbf{K}}_{\boldsymbol{X},Y} & \widehat{\mathbf{K}}_{\boldsymbol{X}} \end{pmatrix}. \tag{3.13}$$

- For the nonconvex penalty functions discussed in §2, $\nabla\mathcal{Q}_{\lambda_t}(\boldsymbol{\beta})$ takes the forms of

$$\big(\nabla\mathcal{Q}_{\lambda_t}(\boldsymbol{\beta})\big)_j = \begin{cases} \frac{\lambda_t\,\mathrm{sign}(\beta_j)-\beta_j}{a-1}\cdot\mathbb{I}\big(\lambda_t<|\beta_j|\le a\lambda_t\big) - \lambda_t\,\mathrm{sign}(\beta_j)\cdot\mathbb{I}\big(|\beta_j|>a\lambda_t\big), & \text{SCAD,} \\ -\frac{\beta_j}{b}\lambda_t\,\mathrm{sign}(\beta_j)\cdot\mathbb{I}\big(|\beta_j|\le b\lambda_t\big) - \lambda_t\,\mathrm{sign}(\beta_j)\cdot\mathbb{I}\big(|\beta_j|>b\lambda_t\big), & \text{MCP,} \end{cases}$$

  where $a>2$, $b>0$.

Line-Search Method: Before we present the proposed proximal-gradient method in detail, we briefly introduce a line-search algorithm, which adaptively searches for the best quadratic coefficient $L_t^{(k)}$ of the local quadratic approximation (3.7). As shown in Lines 4−7 of Algorithm 2, the main idea of line-search is to iteratively increase $L_t^{(k)}$ by a factor of two and compute the corresponding $\boldsymbol{\beta}_t^{(k)}$, until the local approximation $\psi_{L_t^{(k)},\lambda_t}\big(\boldsymbol{\beta}_t^{(k)};\boldsymbol{\beta}_t^{(k-1)}\big)$ becomes a tight upper bound of the objective function $\phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k)}\big)$. We will theoretically characterize the computational complexity of this line-search method in Remark 4.6 and specify the range of $L_t^{(k)}$ in Theorem 5.5.

Algorithm 2 The line-search method used to search for the best $L_t^{(k)}$ and compute the corresponding $\boldsymbol{\beta}_t^{(k)}$. Here $\phi_{\lambda_t}(\boldsymbol{\beta})$ is the objective function defined in (3.4), and $\psi_{L_t^{(k)},\lambda_t}\big(\boldsymbol{\beta};\boldsymbol{\beta}_t^{(k-1)}\big)$ is the local quadratic approximation of $\phi_{\lambda_t}(\boldsymbol{\beta})$ defined in (3.7).

1: $\big\{\boldsymbol{\beta}_t^{(k)}, L_t^{(k)}\big\} \leftarrow$ Line-Search$\big(\lambda_t, \boldsymbol{\beta}_t^{(k-1)}, L_{\text{init}}, R\big)$
2: **input:** $\lambda_t>0, \boldsymbol{\beta}_t^{(k-1)}\in\mathbb{R}^d, L_{\text{init}}>0, R>0$
3: **initialize:** $L_t^{(k)} \leftarrow L_{\text{init}}$
4: **repeat**
5: $\quad\boldsymbol{\beta}_t^{(k)} \leftarrow \mathcal{T}_{L_t^{(k)},\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}; R\big)$ as defined in (3.9)
6: $\quad$**if** $\phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k)}\big) > \psi_{L_t^{(k)},\lambda_t}\big(\boldsymbol{\beta}_t^{(k)};\boldsymbol{\beta}_t^{(k-1)}\big)$ **then** $L_t^{(k)} \leftarrow 2L_t^{(k)}$
7: **until** $\phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k)}\big) \le \psi_{L_t^{(k)},\lambda_t}\big(\boldsymbol{\beta}_t^{(k)};\boldsymbol{\beta}_t^{(k-1)}\big)$
8: **return** $\big\{\boldsymbol{\beta}_t^{(k)}, L_t^{(k)}\big\}$

Stopping Criterion: Now we introduce the stopping criterion of our proximal-gradient method. In other words, we specify the optimality conditions that should be satisfied by the approximate solution $\widetilde{\boldsymbol{\beta}}_t$ attained by our proximal-gradient method.

It is known that any exact local solution $\widehat{\boldsymbol{\beta}}_\lambda$ to the optimization problem

$$\text{minimize } \phi_\lambda(\boldsymbol{\beta}), \quad \text{where } \phi_\lambda(\boldsymbol{\beta}) = \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1, \quad \boldsymbol{\beta} \in \Omega$$

satisfies the optimality condition, i.e, there exists some $\boldsymbol{\xi} \in \partial\big\|\widehat{\boldsymbol{\beta}}_\lambda\big\|_1$ such that

$$\big(\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\big)^T\Big(\nabla\widetilde{\mathcal{L}}_\lambda\big(\widehat{\boldsymbol{\beta}}_\lambda\big) + \lambda\boldsymbol{\xi}\Big) \le 0, \quad \text{for any } \boldsymbol{\beta} \in \Omega. \tag{3.14}$$

We can understand this optimality condition as follows: Locally at $\widehat{\boldsymbol{\beta}}_\lambda$, any feasible direction pointed at $\widehat{\boldsymbol{\beta}}_\lambda$, i.e., $\big(\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\big)$ where $\boldsymbol{\beta} \in \Omega$, leads to a decrease in the objective function value $\phi_\lambda(\boldsymbol{\beta})$, because as shown in (3.14), such a direction forms an obtuse angle with the (sub)gradient vector of $\phi_\lambda(\boldsymbol{\beta})$ evaluated at $\widehat{\boldsymbol{\beta}}_\lambda$. If $\widehat{\boldsymbol{\beta}}_\lambda$ lies in the interior of Ω, e.g., $\Omega = \mathbb{R}^d$, then (3.14) reduces to the known first-order KKT condition,

$$\nabla\widetilde{\mathcal{L}}_\lambda\big(\widehat{\boldsymbol{\beta}}_\lambda\big) + \lambda\boldsymbol{\xi} = \mathbf{0}, \quad \text{where } \boldsymbol{\xi} \in \partial\big\|\widehat{\boldsymbol{\beta}}_\lambda\big\|_1. \tag{3.15}$$

To see this, given $\widehat{\boldsymbol{\beta}}_\lambda$ lies in the interior of Ω, we have $\big(\widehat{\boldsymbol{\beta}}_\lambda + C\boldsymbol{v}\big) \in \Omega$ and $\big(\widehat{\boldsymbol{\beta}}_\lambda - C\boldsymbol{v}\big) \in \Omega$ for any fixed $\boldsymbol{v} \in \mathbb{R}^d$ and $C > 0$ sufficiently small. Setting $\boldsymbol{\beta}$ in (3.14) to be these two values, we obtain $\boldsymbol{v}^T\big(\nabla\widetilde{\mathcal{L}}_\lambda\big(\widehat{\boldsymbol{\beta}}_\lambda\big) + \boldsymbol{\xi}\big) = 0$, which further implies (3.15) since $\boldsymbol{v}$ is arbitrarily chosen.

Based on the optimality condition in (3.17), we measure the suboptimality of a $\boldsymbol{\beta} \in \Omega$ with

$$\omega_\lambda(\boldsymbol{\beta}) = \min_{\boldsymbol{\xi}' \in \partial\|\boldsymbol{\beta}\|_1} \max_{\boldsymbol{\beta}' \in \Omega} \left\{ \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}')^T}{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_1}\big(\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda\boldsymbol{\xi}'\big) \right\}. \tag{3.16}$$

To understand this measure of suboptimality, first note that, if $\boldsymbol{\beta}$ is an exact local solution, then we have $\omega_\lambda(\boldsymbol{\beta}) \le 0$ by (3.14). Otherwise, if $\boldsymbol{\beta}$ is close to some exact local solution, then $\omega_\lambda(\boldsymbol{\beta})$ is some small positive value. When $\boldsymbol{\beta}$ lies in the interior of Ω, then (3.16) reduces to a more straightforward

$$\omega_\lambda(\boldsymbol{\beta}) = \min_{\boldsymbol{\xi}' \in \partial\|\boldsymbol{\beta}\|_1} \Big\{ \big\|\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda\boldsymbol{\xi}'\big\|_\infty \Big\}. \tag{3.17}$$

This is because for any fixed $\boldsymbol{v} \in \mathbb{R}^d$, we have $(\boldsymbol{\beta} + C\boldsymbol{v}) \in \Omega$ for $C > 0$ sufficiently small. Setting $\boldsymbol{\beta}$ to be this value in (3.16), we have

$$\omega_\lambda(\boldsymbol{\beta}) = \min_{\boldsymbol{\xi}' \in \partial\|\boldsymbol{\beta}\|_1} \max_{\boldsymbol{v} \in \mathbb{R}^d} \left\{ \frac{\boldsymbol{v}^T}{\|\boldsymbol{v}\|_1}\big(\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda\boldsymbol{\xi}'\big) \right\} = \min_{\boldsymbol{\xi}' \in \partial\|\boldsymbol{\beta}\|_1} \Big\{ \big\|\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda\boldsymbol{\xi}'\big\|_\infty \Big\},$$

where the second equality follows from the duality between ℓ_1 and ℓ_∞ norm.

Equipped with the suboptimality measure $\omega_\lambda(\boldsymbol{\beta})$ defined in (3.16), we can define the stopping criterion of our proximal-gradient method within the t-th path following stage to be $\omega_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k)}\big) \le \epsilon_t$, where $\epsilon_t > 0$ is the desired optimization precision (Line 9 of Algorithm 3). Therefore, the proximal-gradient method achieves an approximate local solution $\widetilde{\boldsymbol{\beta}}_t$ with suboptimality ϵ_t. Recall that within the t-th path following stage ($t = 1, \ldots, N-1$), we set ϵ_t to be $\lambda_t/4$ (Line 7 of Algorithm 1), while within the N-th path following stage, we set $\epsilon_t = \epsilon_{\text{opt}} \ll \lambda_{\text{tgt}}/4$ (Line 11 of Algorithm 1).

Algorithm 3 The proximal-gradient method for nonconvex problems, which iteratively leverages the line-search method illustrated in Algorithm 2 at each iteration.

1: $\{\widetilde{\boldsymbol{\beta}}_t, L_t\} \leftarrow \text{Proximal-Gradient}\big(\lambda_t, \epsilon_t, \boldsymbol{\beta}_t^{(0)}, L_t^{(0)}, R\big)$
2: **input:** $\lambda_t > 0, \epsilon_t > 0, \boldsymbol{\beta}_t^{(0)} \in \mathbb{R}^d, L_t^{(0)} > 0, R > 0$
3: **parameter:** $L_{\min} > 0$
4: **initialize:** $k \leftarrow 0$
5: **repeat**
6: $\quad k \leftarrow k + 1$
7: $\quad L_{\text{init}} \leftarrow \max\big\{L_{\min}, L_t^{(k-1)}/2\big\}$
8: $\quad \boldsymbol{\beta}_t^{(k)}, L_t^{(k)} \leftarrow \text{Line-Search}\big(\lambda_t, \boldsymbol{\beta}_t^{(k-1)}, L_{\text{init}}, R\big)$ as in Algorithm 2
9: **until** $\omega_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k)}\big) \le \epsilon_t$ as defined in (3.16)
10: $\widetilde{\boldsymbol{\beta}}_t \leftarrow \boldsymbol{\beta}_t^{(k)}$
11: $L_t \leftarrow L_t^{(k)}$
12: **return** $\{\widetilde{\boldsymbol{\beta}}_t, L_t\}$

Proposed Proximal-Gradient Method: We are now ready to present the proposed proximal-gradient method in detail. Recall that, within the t-th stage of our path following algorithm, we employ the proximal-gradient method to obtain a desired approximate local solution $\widetilde{\boldsymbol{\beta}}_t$ (Line 8 and Line 12 of Algorithm 1). As shown in Line 8 of Algorithm 3, at the k-th iteration of our proximal-gradient method, we employ the line-search method (Algorithm 2) to search for the best $L_t^{(k)}$ and calculate the corresponding $\boldsymbol{\beta}_t^{(k)}$.

At the k-th iteration of the proximal-gradient method, we set the initial value L_{init} of the line-search procedure to be $\max\big\{L_{\min}, L_t^{(k-1)}/2\big\}$ (Line 7 of Algorithm 3). Here $L_{\min} > 0$ is a parameter used to prevent L_{init} from being too small. In practice, $L_{\min}$ is often set to be a sufficiently small value, e.g., $L_{\min} = 10^{-6}$. The intuition behind such initialization can be understood as follows: As shown in (3.7), $L_t^{(k-1)}$ and $L_t^{(k)}$ are the quadratic coefficients of the local quadratic approximations of the objective function at $\boldsymbol{\beta}_t^{(k-2)}$ and $\boldsymbol{\beta}_t^{(k-1)}$ respectively. Intuitively speaking, $\boldsymbol{\beta}_t^{(k-2)}$ and $\boldsymbol{\beta}_t^{(k-1)}$ are close to each other, which implies that $L_t^{(k-1)}$ is a good guess for $L_t^{(k)}$. Hence we can initialize the line-search method for $L_t^{(k)}$ with a value slightly smaller than $L_t^{(k-1)}$, e.g., $L_t^{(k-1)}/2$.

When the stopping criterion $\omega_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k)}\big) \le \epsilon_t$ is satisfied, the proximal-gradient method stops and outputs the approximate local solution $\widetilde{\boldsymbol{\beta}}_t = \boldsymbol{\beta}_t^{(k)}$ (Line 10 of Algorithm 3). We also keep track of $L_t = L_t^{(k)}$ to accelerate the line-search procedure within the next path following stage.

The reason we employ the line-search method instead of using a fixed $L_t^{(k)}$ is that, the adaptive line-search algorithm enables us to automatically exploit the strong convexity of $\phi_{\lambda_t}(\boldsymbol{\beta})$. In other words, in §4 we will show that, as long as $\phi_{\lambda_t}(\boldsymbol{\beta})$ is strongly convex, the proximal-gradient method within the t-th path following stage adapts to attain a fast geometric rate of convergence without manually choosing a fixed $L_t^{(k)}$. Here geometric convergence means that we need at most $C\log(1/\epsilon_t)$ proximal-gradient steps to obtain an ϵ_t-suboptimal approximate local solution.

4 Theoretical Results

We establish theoretical results on the iteration complexity and statistical performance of our approximate regularization path following method for nonconvex learning problems.

4.1 Assumptions

We first list the required assumptions. The first assumption is about the relationship between λ_{tgt} and $\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty$.

Assumption 4.1. For least squares loss and logistic loss, we set $\lambda_{\text{tgt}} = C\sqrt{\log d/n}$. For semiparametric elliptical design loss, we set $\lambda_{\text{tgt}} = C'\|\boldsymbol{\beta}^*\|_1\sqrt{\log d/n}$. We assume

$$\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty \le \lambda_{\text{tgt}}/8. \tag{4.1}$$

Assumption 4.1 is a common condition that λ_{tgt} should be large enough to dominate the noise. For instance, for least squares loss we have

$$\nabla\mathcal{L}(\boldsymbol{\beta}^*) = \frac{1}{n}\mathbf{X}^T\big(\mathbf{X}\boldsymbol{\beta}^* - \mathbf{y}\big),$$

where $\mathbf{X}\boldsymbol{\beta}^* - \mathbf{y}$ is in fact the noise vector. In Lemma C.3 we will show that, for least squares loss and logistic loss, we have that $\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty \le C\sqrt{\log d/n}$ holds with high probability under certain conditions. Similarly, in Lemma C.4 we will prove that, for semiparametric elliptical design loss, $\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty \le C'\|\boldsymbol{\beta}^*\|_1\sqrt{\log d/n}$ holds with high probability under certain conditions. Therefore, our assumption about λ_{tgt} and $\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty$ holds with high probability.

In the sequel, we lay out another assumption on the sparse eigenvalues of $\nabla^2\mathcal{L}(\boldsymbol{\beta})$, which are defined as follows.

Definition 4.2 (Sparse Eigenvalues)**.** Let s be a positive integer. We define the largest and smallest s-sparse eigenvalues of the Hessian matrix $\nabla^2\mathcal{L}(\boldsymbol{\beta})$ to be

$$\begin{aligned}
\rho_+\big(\nabla^2\mathcal{L}, s\big) &= \sup\Big\{\boldsymbol{v}^T\nabla^2\mathcal{L}(\boldsymbol{\beta})\boldsymbol{v} : \|\boldsymbol{v}\|_0 \le s,\ \|\boldsymbol{v}\|_2 = 1,\ \boldsymbol{\beta}\in\mathbb{R}^d\Big\},\\
\rho_-\big(\nabla^2\mathcal{L}, s\big) &= \inf\Big\{\boldsymbol{v}^T\nabla^2\mathcal{L}(\boldsymbol{\beta})\boldsymbol{v} : \|\boldsymbol{v}\|_0 \le s,\ \|\boldsymbol{v}\|_2 = 1,\ \boldsymbol{\beta}\in\mathbb{R}^d\Big\}.
\end{aligned}$$

For least squares loss and semiparametric elliptical design loss, $\nabla^2\mathcal{L}(\boldsymbol{\beta})$ does not depend on $\boldsymbol{\beta}$. However, for logistic loss we have

$$\nabla^2\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T\cdot\frac{1}{1+\exp(-\mathbf{x}_i^T\boldsymbol{\beta})}\cdot\frac{1}{1+\exp(\mathbf{x}_i^T\boldsymbol{\beta})}, \tag{4.2}$$

which depends on $\boldsymbol{\beta}$. In Definition 4.2, the smallest s-sparse eigenvalue $\rho_-\big(\nabla^2\mathcal{L}, s\big)$ is obtained by taking infimum over all $\boldsymbol{\beta}\in\mathbb{R}^d$. Consequently, for logistic loss, $\rho_-\big(\nabla^2\mathcal{L}, s\big)$ is always zero, because in (4.2) we can take $\boldsymbol{\beta}$ such that $|\mathbf{x}_i^T\boldsymbol{\beta}| \to +\infty$ for all nonzero $\mathbf{x}_i$'s, which implies that $\nabla^2\mathcal{L}(\boldsymbol{\beta})$ goes to an all-zero matrix. To avoid this degenerate case, for logistic loss we define the sparse eigenvalues by taking infimum/supremum over all $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\|_2$ bounded instead of over all $\boldsymbol{\beta}\in\mathbb{R}^d$. To unify the later analysis for different loss functions, we overload the definition of sparse eigenvalues for logistic loss as follows.

Definition 4.3 (Sparse Eigenvalues for Logistic Loss)**.** Let s be a positive integer. For logistic loss, we define the largest and smallest s-sparse eigenvalues of $\nabla^2\mathcal{L}(\boldsymbol{\beta})$ to be

$$\begin{aligned}\rho_+\big(\nabla^2\mathcal{L}, s\big) &= \sup\left\{\boldsymbol{v}^T\nabla^2\mathcal{L}(\boldsymbol{\beta})\boldsymbol{v} : \|\boldsymbol{v}\|_0 \le s,\ \|\boldsymbol{v}\|_2 = 1,\ \|\boldsymbol{\beta}\|_2 \le R\right\},\\ \rho_-\big(\nabla^2\mathcal{L}, s\big) &= \inf\left\{\boldsymbol{v}^T\nabla^2\mathcal{L}(\boldsymbol{\beta})\boldsymbol{v} : \|\boldsymbol{v}\|_0 \le s,\ \|\boldsymbol{v}\|_2 = 1,\ \|\boldsymbol{\beta}\|_2 \le R\right\},\end{aligned}$$

where $R \in (0, +\infty)$ is an absolute constant such that $\|\boldsymbol{\beta}^*\|_2 \le R$.

Note that in Definition 4.3, we implicitly assume that $\|\boldsymbol{\beta}^*\|_2$ is upper bounded by some known absolute constant. Although it seems rather restrictive, this assumption is essential for logistic loss. Otherwise, $\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)$ might go to an all-zero matrix when $\|\boldsymbol{\beta}^*\|_2 \to +\infty$. When the curvature of the objective function at $\boldsymbol{\beta}^*$ is zero, a consistent estimation of $\boldsymbol{\beta}^*$ is impossible. Although this assumption is necessary for theoretical purposes, we require no prior knowledge about the exact value of $\|\boldsymbol{\beta}^*\|_2$ in practice, since we can always set R to be a sufficiently large constant in our algorithm (Line 3 of Algorithm 1).

Recall that, as shown in Line 8 and Line 12 of Algorithm 1, we impose an ℓ_2 constraint of radius R for all the proximal-gradient iterations at each path following stage. Therefore we have $\big\|\boldsymbol{\beta}_t^{(k)}\big\|_2 \le R$ during the whole iterative procedure of our approximate path following method. Now we are ready to present the sparse eigenvalue assumption on the Hessian matrix.

Assumption 4.4. Let $s^* = \|\boldsymbol{\beta}^*\|_0$, where $\boldsymbol{\beta}^*$ is the true parameter vector. We assume there exists an integer $\widetilde{s} > Cs^*$ such that

$$\rho_+\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big) < +\infty, \quad \rho_-\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big) > 0$$

are two absolute constants. The constant $C > 0$ is constant specified as follows.

In Assumption 4.4, the constant

$$C = 144\kappa^2 + 250\kappa, \tag{4.3}$$

where κ is a condition number defined as

$$\kappa = \frac{\rho_+\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big) - \zeta_+}{\rho_-\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big) - \zeta_-} \in [1, +\infty). \tag{4.4}$$

Here recall that $\zeta_+ \ge 0$ and $\zeta_+ \ge 0$ are the two concavity parameters of the nonconvex penalty as defined in regularity condition (a). To ensure that $\kappa \in [1, +\infty)$, it is necessary to choose

$$\zeta_- \le C'\rho_-\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big) \tag{4.5}$$

with constant $C' < 1$, which automatically implies

$$\zeta_+ \le C'\rho_+\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big), \tag{4.6}$$

because regularity condition (a) implies $\zeta_+ \le \zeta_-$, and we have $\rho_-\big(\nabla^2\mathcal{L}, s^*+2\widetilde{s}\big) \le \rho_+\big(\nabla^2\mathcal{L}, s^*+2\widetilde{s}\big)$ by definition. Such a restriction on the concavity parameters suggests that the concave component $\mathcal{Q}_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^d q_\lambda(\beta_j)$ of the nonconvex penalty is not allowed to be arbitrarily concave.

Assumption 4.4 is a standard condition in high-dimensional statistical theory, which is closely related to the restricted isometry property (RIP) proposed by Candés and Tao (2005). Similar conditions have been studied by Bickel et al. (2009); Raskutti et al. (2010); Negahban et al. (2012); Zhang (2012); Xiao and Zhang (2012). More specifically, for least squares loss, the RIP condition assumes that there exists an integer s and some constant $\delta \in (0, 1)$ such that

$$1 - \delta \leq \rho_-\big(\nabla^2\mathcal{L}, s\big) \leq \rho_+\big(\nabla^2\mathcal{L}, s\big) \leq 1 + \delta. \tag{4.7}$$

In the following, we justify Assumption 4.4 for least squares loss with an example.

To illustrate that Assumption 4.4 is well defined, we assume that the RIP condition in (4.7) holds with $s = 877s^*$ and $\delta = 0.01$. We set the concavity parameters of the nonconvex penalty in (a) to be $\zeta_+ = 0$ and $\zeta_- = \rho_-\big(\nabla^2\mathcal{L}, s\big)/20$, e.g., for MCP defined in (2.2), we set $b = 1/\zeta_- = 20/\rho_-\big(\nabla^2\mathcal{L}, s\big)$. In the following, we verify that there exists an integer $\widetilde{s} = 438s^*$ that satisfies Assumption 4.4.

First, according to the RIP condition, we have

$$\rho_+\big(\nabla^2\mathcal{L}, s^*+2\widetilde{s}\big) = \rho_+\big(\nabla^2\mathcal{L}, 877s^*\big) = \rho_+\big(\nabla^2\mathcal{L}, s\big) \leq (1 + \delta) = 1.01 < +\infty, \tag{4.8}$$

$$\rho_-\big(\nabla^2\mathcal{L}, s^*+2\widetilde{s}\big) = \rho_-\big(\nabla^2\mathcal{L}, 877s^*\big) = \rho_-\big(\nabla^2\mathcal{L}, s\big) \geq (1 - \delta) = 0.99 > 0. \tag{4.9}$$

Second, we calculate the value of $\widetilde{s}$ in detail. Since the condition number κ defined in (4.4) satisfies

$$1 \leq \kappa = \frac{\rho_+\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big) - \zeta_+}{\rho_-\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big) - \zeta_-} = \frac{\rho_+\big(\nabla^2\mathcal{L}, s\big) - \zeta_+}{\rho_-\big(\nabla^2\mathcal{L}, s\big) - \zeta_-} = \frac{20}{19} \cdot \frac{\rho_+\big(\nabla^2\mathcal{L}, s\big)}{\rho_-\big(\nabla^2\mathcal{L}, s\big)} \leq \frac{20}{19} \cdot \frac{1+\delta}{1-\delta} < 1.08.$$

We now verify that $\widetilde{s}$ satisfies $\widetilde{s} > Cs^*$ in Assumption 4.4, where C is defined in (4.3). Plugging the range $1 \leq \kappa < 1.08$ into the definition of C, we obtain $C = 144\kappa^2 + 250\kappa < 438$. Therefore, as long as the RIP condition holds with $s = 877s^*$ and $\delta = 0.01$, we can find an integer $\widetilde{s} = 438s^*$ that satisfies Assumption 4.4. For least squares loss, the RIP condition is known to hold for a variety of design matrices with high probability, which implies that Assumption 4.4 also holds with high probability for these designs.

It is worth noting that the constants in this example are rather large for practical purposes. We could expect that these constants would be much smaller if we manage to get a small constant C in (4.3). However, we mainly focus on providing novel theoretical insights in this paper, without paying too much effort on optimizing constants.

Furthermore, we will justify Assumption 4.4 for $\mathcal{L}(\boldsymbol{\beta})$ being semiparametric elliptical design loss and logistic loss in Appendix C.3. In Lemma 5.1 we will show that Assumption 4.4 actually implies the strong convexity and smoothness of $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \mathcal{Q}_\lambda(\boldsymbol{\beta})$ for $\boldsymbol{\beta}$ on a sparse set, which are essential for establishing the fast geometric rate of convergence of the proposed optimization algorithm and achieving the desired statistical properties of the local solutions. Hereafter, we use the shorthands

$$\rho_+ = \rho_+\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big), \quad \rho_- = \rho_-\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big) \tag{4.10}$$

for notational simplicity.

4.2 Main Theorems

We first provide the main results about the computational rate of convergence. We then establish the statistical properties of the local solutions obtained by our approximate path following method.

4.2.1 Computational Theory

The next theorem shows that the proposed regularization path following method attains a global geometric rate of convergence for calculating the full regularization path. Such a rate of convergence is the fastest achievable rate among all first-order optimization methods.

Theorem 4.5 (Geometric Rate of Convergence)**.** Under Assumption 4.1 and Assumption 4.4, we have the following results:

1. **Geometric Rate of Convergence within the t-th Stage:** Within the t-th ($t = 1, \ldots, N$) path following stage (Line 8 and Line 12 of Algorithm 1), the iterative sequence $\{\boldsymbol{\beta}_t^{(k)}\}_{k=0}^{\infty}$ generated by the proximal-gradient method (Algorithm 3) converges to a unique local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$.

   - Within the t-th stage ($t = 1, \ldots, N-1$), the total number of proximal-gradient iterations (Lines 5–9 of Algorithm 3) is no more than $C' \log\left(4C\sqrt{s^*}\right)$.
   - Within the N-th stage, the total number of proximal-gradient iterations (Lines 5–9 of Algorithm 3) is no more than $\max\left\{0, C' \log\left(C\lambda_{\text{tgt}}\sqrt{s^*}/\epsilon_{\text{opt}}\right)\right\}$.

   Here s^* is the number of nonzero entries of the true parameter vector $\boldsymbol{\beta}^*$,

   $$C = 2\sqrt{21} \cdot \sqrt{\kappa}(1+\kappa), \quad C' = 2 \Big/ \log\left(\frac{1}{1 - 1/(8\kappa)}\right), \tag{4.11}$$

   where $\kappa \in [1, +\infty)$ is the condition number defined in (4.4).

2. **Geometric Rate of Convergence over the Full Path:** To compute the entire path, we need no more than

   $$\underbrace{(N-1)C' \log\left(4C\sqrt{s^*}\right)}_{1, \ldots, (N-1)-\text{th Stages}} + \underbrace{C' \log\left(\frac{C\lambda_{\text{tgt}}\sqrt{s^*}}{\epsilon_{\text{opt}}}\right)}_{N-\text{th Stage}} \tag{4.12}$$

   proximal-gradient update iterations (Lines 5–9 of Algorithm 3), where C, C' are specified in (4.11). Here $\epsilon_{\text{opt}} \ll \lambda_{\text{tgt}}/4$ is the optimization precision of the final path following stage (Line 12 of Algorithm 1), and $N = \log(\lambda_0/\lambda_{\text{tgt}})/\log(\eta^{-1})$ is the total number of approximate path following stages, where $\eta \in [0.9, 1)$ is an absolute constant.

3. **Geometric Rate of Convergence of the Objective Function Values:** Let $\widetilde{\boldsymbol{\beta}}_t$ be the approximate local solution obtained within the t-th stage (Line 8 and Line 12 of Algorithm 1).

   - For $t = 0, \ldots, N-1$, the value of the objective function decays exponentially towards the value at the final exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_{\text{tgt}}}$, i.e.,

   $$\phi_{\lambda_{\text{tgt}}}\left(\widetilde{\boldsymbol{\beta}}_t\right) - \phi_{\lambda_{\text{tgt}}}\left(\widehat{\boldsymbol{\beta}}_{\lambda_{\text{tgt}}}\right) \leq Cs^* \cdot \eta^{2(t+1)}, \quad \text{where } C = \frac{105\lambda_0^2}{\rho_- - \zeta_-}. \tag{4.13}$$

- For $t = N$, we have

$$\phi_{\lambda_{\text{tgt}}}\big(\widetilde{\boldsymbol{\beta}}_N\big) - \phi_{\lambda_{\text{tgt}}}\big(\widehat{\boldsymbol{\beta}}_{\lambda_{\text{tgt}}}\big) \le \big(C' \lambda_{\text{tgt}} s^*\big) \cdot \epsilon_{\text{opt}}, \quad \text{where } C' = \frac{21}{\rho_- - \zeta_-}. \tag{4.14}$$

  Here $\rho_- = \rho_-\big(\nabla^2 \mathcal{L}, s^* + 2\widetilde{s}\big) > 0$ is the smallest sparse eigenvalue specified in Assumption 4.4; As defined in regularity condition (a), $\zeta_- > 0$ is the concavity parameter of the nonconvex penalty, which satisfies (4.5).

Proof. See the next section for a detailed proof. □

Result 1 suggests that within each path following stage the proximal-gradient algorithm attains a geometric rate of convergence. More specifically, within the t-th ($t = 1, \ldots, N$) stage (Line 8 and Line 12 of Algorithm 1), we only need a logarithmic number of proximal-gradient update iterations (Lines 5–9 of Algorithm 3) to compute an approximate local solution $\widetilde{\boldsymbol{\beta}}_t$. Furthermore, within the t-th path following stage, the iterative sequence $\big\{\boldsymbol{\beta}_t^{(k)}\big\}_{k=0}^{\infty}$ produced by Algorithm 3 converges towards a unique local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$. In Theorem 4.8, we will show that $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ enjoys a more refined statistical rate of convergence due to the usage of nonconvex penalty.

Result 2 suggests that our approximate path following method attains a global geometric rate of convergence. From the perspective of high-dimensional statistics, the total number of stages N scales with dimension d and sample size n, because $N = \log(\lambda_0/\lambda_{\text{tgt}})/\log(\eta^{-1})$, where η is an absolute constant. From the perspective of optimization, given dimension d and sample size n, when the optimization precision ϵ_{opt} is sufficiently small such that in (4.12) the second term dominates the first term, then the total iteration complexity is $C \log(1/\epsilon_{\text{opt}})$. In other words, we only need to conduct a logarithmic number of proximal-gradient iterations (Lines 5–9 of Algorithm 3) to compute the full regularization path.

Recall that we measure the suboptimality of an approximate solution with $\omega_\lambda(\boldsymbol{\beta})$ defined in (3.16), which does not directly reflect the optimality of the objective function value. Hence we provide result 3 to characterize the decay of the objective gap $\phi_{\lambda_{\text{tgt}}}\big(\widetilde{\boldsymbol{\beta}}_t\big) - \phi_{\lambda_{\text{tgt}}}\big(\widehat{\boldsymbol{\beta}}_{\lambda_{\text{tgt}}}\big)$. In detail, (4.13) illustrates the exponential decay of the objective gap along the regularization path, i.e., $t = 1, \ldots, N-1$, while (4.14) suggests that, the final objective function value evaluated at $\widetilde{\boldsymbol{\beta}}_N$ is close to the value at the exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_{\text{tgt}}}$, as long as the optimization precision ϵ_{opt} is sufficiently small.

Remark 4.6. Nesterov (2007) showed that the total number of line-search steps (Lines 4–7 of Algorithm 2) within the k-th proximal-gradient iteration (Line 8 of Algorithm 3) is no more than

$$2(k+1) + \max\left\{0, \frac{\log(\rho_+ - \zeta_+) - \log L_{\min}}{\log 2}\right\},$$

where the sparse eigenvalue $\rho_+ = \rho_+\big(\nabla^2 \mathcal{L}, s^* + 2\widetilde{s}\big) > 0$ is specified in Assumption 4.4; As defined in regularity condition (a), $\zeta_+ > 0$ is the concavity parameter of the nonconvex penalty that satisfies (4.6); $L_{\min}$ is a parameter of Algorithm 3 (Line 3). Piecing the above results together, we conclude that the total number of line-search iterations (Lines 4–7 of Algorithm 2) required to compute the full regularization path is of the same order as (4.12).

4.2.2 Statistical Theory

We present two types of statistical results. Recall that $\widetilde{\boldsymbol{\beta}}_t$ is the approximate local solution obtained within the t-th path following stage, while $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ is the corresponding exact local solution that satisfies the exact optimality condition in (3.14). In Theorem 4.7, we will provide a statistical characterization of all the approximate local solutions $\{\widetilde{\boldsymbol{\beta}}_t\}_{t=1}^N$ attained along the full regularization path. Remind in Theorem 4.5 we prove that within the t-th stage, the iterative sequence $\{\boldsymbol{\beta}_t^{(k)}\}_{k=0}^\infty$ produced by the proximal-gradient method (Algorithm 3) converges towards a unique exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$. In Theorem 4.8, we will provide more refined statistical properties of these exact local solutions $\{\widehat{\boldsymbol{\beta}}_{\lambda_t}\}_{t=1}^N$ along the full regularization path. Since $\widehat{\boldsymbol{\beta}}_{\lambda_N} = \widehat{\boldsymbol{\beta}}_{\lambda_{\text{tgt}}}$, this result justifies the statistical property of the final estimator.

Theorem 4.7 (Statistical Rates of Convergence of Approximate Local Solutions)**.** Let d be the dimension of $\boldsymbol{\beta}$ and n be the sample size. Recall that $\widetilde{\boldsymbol{\beta}}_t$ is the approximate local solution obtained within the t-th path following stage (Line 8 and Line 12 of Algorithm 1). Under Assumption 4.1 and Assumption 4.4, we have

$$\big\|\widetilde{\boldsymbol{\beta}}_t - \boldsymbol{\beta}^*\big\|_2 \le C\lambda_t\sqrt{s^*}, \quad \text{for} \;\; t = 1, \ldots, N, \tag{4.15}$$

where $s^* = \|\boldsymbol{\beta}^*\|_0$. Here $N = \log(\lambda_0/\lambda_{\text{tgt}})/\log(\eta^{-1})$ is the total number of path following stages, where $\eta \in [0.9, 1)$ is a constant and $\lambda_t = \eta^t\lambda_0$. In (4.15), the constant $C = (21/8)/(\rho_- - \zeta_-)$, where $\rho_- = \rho_-\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big) > 0$ is the smallest sparse eigenvalue specified in Assumption 4.4. As defined in regularity condition (a), $\zeta_- > 0$ is the concavity parameter of the nonconvex penalty, which satisfies (4.5).

Proof. See the next section for a detailed proof. □

Theorem 4.7 provides statistical rates of convergence of all the approximate local solutions attained by our algorithm along the regularization path. Recall that in Assumption 4.1, we set $\lambda_{\text{tgt}} = C\sqrt{\log d/n}$ for least squares and logistic loss, and $\lambda_{\text{tgt}} = C'\|\boldsymbol{\beta}^*\|_1\sqrt{\log d/n}$ for semiparametric elliptical design loss. For least squares and logistic loss, taking $t = N$ in Theorem 4.7, we have

$$\big\|\widetilde{\boldsymbol{\beta}}_N - \boldsymbol{\beta}^*\big\|_2 \le \frac{21/8}{\rho_- - \zeta_-}\lambda_{\text{tgt}}\sqrt{s^*} = \frac{21/8 \cdot C}{\rho_- - \zeta_-}\sqrt{\frac{s^*\log d}{n}}.$$

Hence, the final approximate local solution $\widetilde{\boldsymbol{\beta}}_N$ attains the minimax rate of convergence for parameter estimation. Similarly, for semiparametric elliptical design loss, we have

$$\big\|\widetilde{\boldsymbol{\beta}}_N - \boldsymbol{\beta}^*\big\|_2 \le \frac{21/8 \cdot C'}{\rho_- - \zeta_-}\|\boldsymbol{\beta}^*\|_1\sqrt{\frac{s^*\log d}{n}},$$

which suggests that the rate of convergence of the final approximate solution is also optimal in the regime where $\|\boldsymbol{\beta}^*\|_1$ is upper bounded by a constant. Moreover, since η is an absolute constant, for $\widetilde{\boldsymbol{\beta}}_{N-K}$ with K being a positive integer constant, Theorem 4.7 gives

$$\big\|\widetilde{\boldsymbol{\beta}}_{N-K} - \boldsymbol{\beta}^*\big\|_2 \le \frac{21/8}{\rho_- - \zeta_-}\lambda_{N-K}\sqrt{s^*} \le \frac{21/8 \cdot \eta^{-K}}{\rho_- - \zeta_-}\lambda_{\text{tgt}}\sqrt{s^*},$$

which suggests that, the approximate local solution $\widetilde{\boldsymbol{\beta}}_{N-K}$ enjoys the same rate of convergence as the final approximate local solution $\widetilde{\boldsymbol{\beta}}_N$, but with a looser constant $C = (21/8) \cdot \eta^{-K}/(\rho_- - \zeta_-) > (21/8)/(\rho_- - \zeta_-)$.

In the next theorem, we provide a refined statistical rate of convergence. Remind that, within the t-th path following stage, the iterative sequence $\{\boldsymbol{\beta}_t^{(k)}\}_{k=0}^{\infty}$ produced by the proximal-gradient method (Algorithm 3) converges to a unique exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$. The next theorem states that $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ benefits from nonconvex penalty functions and possesses an improved statistical rate of convergence.

Theorem 4.8 (Refined Statistical Rates of Convergence of Exact Local Solutions)**.** For the regularization parameter λ_t, we assume that the nonconvex penalty $\mathcal{P}_{\lambda_t}(\boldsymbol{\beta}) = \sum_{j=1}^{d} p_{\lambda_t}(\beta_j)$ satisfies

$$p'_{\lambda_t}(\beta_j) = 0, \quad \text{for} \quad |\beta_j| \geq \nu_t, \tag{4.16}$$

for some $\nu_t > 0$. Let $S_1^* \cup S_2^* = S^* = \mathrm{supp}(\boldsymbol{\beta}^*)$ with $|S_1^*| = s_1^*$, $|S_2^*| = s_2^*$ and $|S^*| = s^* = s_1^* + s_2^*$. For $j \in S_1^* \subseteq S^*$, we assume $|\beta_j^*| \geq \nu_t$, while for $j \in S_2^* \subseteq S^*$, we assume $|\beta_j^*| < \nu_t$. Within the t-th path following stage, let $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ be the unique local solution that $\{\boldsymbol{\beta}_t^{(k)}\}_{k=0}^{\infty}$ converges towards (as has been shown in Theorem 4.5). Under Assumption 4.1 and Assumption 4.4, we have

$$\big\|\widehat{\boldsymbol{\beta}}_{\lambda_t} - \boldsymbol{\beta}^*\big\|_2 \leq \underbrace{C\big\|\big(\nabla\mathcal{L}(\boldsymbol{\beta}^*)\big)_{S_1^*}\big\|_2}_{S_1^*:\ \text{Large } |\beta_j|'\text{s}} + \underbrace{C'\lambda_t\sqrt{s_2^*}}_{S_2^*:\ \text{Small } |\beta_j|'\text{s}}, \quad \text{for} \quad t = 1, \ldots, N, \tag{4.17}$$

where $C = 1/(\rho_- - \zeta_-)$ and $C' = 3/(\rho_- - \zeta_-)$.

Proof. See the next section for a detailed proof. □

In Theorem 4.8, the assumption in (4.16) applies to a variety of nonconvex penalty functions. For SCAD in (2.1), we have $\nu_t = a\lambda_t$; While for MCP in (2.2), we have $\nu_t = b\lambda_t$. Theorem 4.8 suggests that, for "small" coefficients such that $|\beta_j| < \nu_t$, the second part on the right-hand side of (4.17) has the same recovery performance as in Theorem 4.7, while for "large" coefficients such that $|\beta_j| \geq \nu_t$, the first part in (4.17) possesses a more refined rate of convergence. To understand this, we consider an example with $\mathcal{L}(\boldsymbol{\beta})$ being least squares loss. We assume that $(Y|\boldsymbol{X} = \mathbf{x}_i)$ follows a sub-Gaussian distribution with mean $\mathbf{x}_i^T\boldsymbol{\beta}^*$ and variance proxy σ^2. Moreover, we assume that the columns of $\mathbf{X}$ are normalized in such a way that $\max_{j\in\{1,\ldots,d\}}\{\|\mathbf{X}_j\|_2\} \leq \sqrt{n}$. Then we have

$$\big\|\big(\nabla\mathcal{L}(\boldsymbol{\beta}^*)\big)_{S_1^*}\big\|_2 \leq C\sigma\sqrt{\frac{s_1^*}{n}} \tag{4.18}$$

with high probability. Clearly, such a $\sqrt{s_1^*/n}$ rate of convergence on the right-hand side of (4.18) is significantly faster than the usual $\sqrt{s^* \log d/n}$ rate, since it gets rid of the $\log d$ term, and $s_1^* \leq s^*$. In fact, ν_t is the minimum signal strength above which we are able to obtain such a refined rate of convergence. In the examples of SCAD and MCP, we have $\nu_t = C\lambda_t$. Recall that $\{\lambda_t\}_{t=0}^N$ is a decreasing sequence. Hence, we are able to achieve this more refined rate of convergence for smaller and smaller signal strength along the full regularization path. Moreover, for $t = N$, the minimum signal strength $\nu_N = \lambda_N = \lambda_{\text{tgt}} = C\sqrt{\log d/n}$. Hence, the required minimum signal strength goes

to zero as the sample size increases. Following a similar proof of Lemma C.3 and Lemma C.4 in Appendix C, we can obtain similar results for logistic loss and semiparametric elliptical design loss. This refined rate of convergence is sharper than the results in Loh and Wainwright (2013), in which they didn't sharply characterize the different conditions of S_1^* and S_2^*. Thus their obtained rate is suboptimal compared to ours in the regime where all the nonzero coefficients of $\boldsymbol{\beta}^*$ are relatively large (i.e., the signal strength is strong).

Besides the refined rate of convergence for parameter estimation in Theorem 4.8, in the next theorem we prove that, the exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ also recovers the support of $\boldsymbol{\beta}^*$ under suitable conditions. Before we present the next theorem, we introduce the definition of an oracle estimator, denoted by $\widehat{\boldsymbol{\beta}}_{\mathrm{O}}$. Recall that $S^* = \mathrm{supp}(\boldsymbol{\beta}^*)$. The oracle estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{O}}$ is defined as

$$\widehat{\boldsymbol{\beta}}_{\mathrm{O}} = \underset{\substack{\mathrm{supp}\,(\boldsymbol{\beta})\subseteq S^* \\ \boldsymbol{\beta}\in\Omega}}{\mathrm{argmin}} \mathcal{L}(\boldsymbol{\beta}), \tag{4.19}$$

where $\Omega = \mathbb{R}^d$ for least squares loss and semiparametric elliptical design loss, while $\Omega = B_2(R)$ for logistic loss with R specified in Definition 4.3. In the next Lemma, we show that $\widehat{\boldsymbol{\beta}}_{\mathrm{O}}$ is the unique global solution to the minimization problem in (4.19) even for nonconvex loss functions, and has nice statistical recovery properties.

Lemma 4.9. Under Assumption 4.4, the oracle estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{O}}$ is the unique global minimizer of (4.19). If $\mathcal{L}(\boldsymbol{\beta})$ is least squares loss, we assume that $(Y|\boldsymbol{X} = \mathbf{x}_i)$ follows a sub-Gaussian distribution with mean $\mathbf{x}_i^T\boldsymbol{\beta}^*$ and variance proxy σ^2, then the oracle estimator satisfies

$$\big\|\widehat{\boldsymbol{\beta}}_{\mathrm{O}} - \boldsymbol{\beta}^*\big\|_\infty \leq C\sigma\sqrt{2/\rho_-}\cdot\sqrt{\frac{\log s^*}{n}} \tag{4.20}$$

with high probability for some constant C, where $\rho_- = \rho_-\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big) > 0$ is the smallest sparse eigenvalue specified in Assumption 4.4.

Proof. See Appendix C.12 for a detailed proof. □

Statistical recovery results similar to (4.20) also hold for logistic loss and semiparametric elliptical design loss under different conditions. These results are omitted here for simplicity. Lemma 4.9 suggests that, for a sufficiently large n and suitable minimum signal strength, the oracle estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{O}}$ exactly recovers the support of $\boldsymbol{\beta}^*$. More specifically, if the minimum signal strength satisfies $\min_{j\in S^*}|\beta_j^*| \geq 2\nu$ for some $\nu > 0$, then we have

$$\min_{j\in S^*}\big|\big(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}\big)_j\big| \geq \min_{j\in S^*}|\beta_j^*| - \big\|\widehat{\boldsymbol{\beta}}_{\mathrm{O}} - \boldsymbol{\beta}^*\big\|_\infty \geq 2\nu - \sigma\sqrt{2/\rho_-}\cdot\sqrt{\frac{\log s^*}{n}},$$

which implies that $\min_{j\in S^*}\big|\big(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}\big)_j\big| \geq \nu > 0$ for a sufficiently large n. Meanwhile, recall $\mathrm{supp}\big(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}\big) \subseteq S^*$. Therefore we have $\mathrm{supp}\big(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}\big) = S^*$.

Remind that, within the t-th approximate path following stage, the sequence $\big\{\boldsymbol{\beta}_t^{(k)}\big\}_{k=0}^{\infty}$ produced by the proximal-gradient method (Algorithm 3) converges to a unique exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$. In the next theorem, we prove that under Assumption 4.1 and Assumption 4.4, $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ is the oracle estimator, and exactly recovers the support of $\boldsymbol{\beta}^*$ under suitable conditions.

Theorem 4.10 (Support Recovery)**.** For the regularization parameter λ_t, suppose that the nonconvex penalty $\mathcal{P}_{\lambda_t}(\boldsymbol{\beta}) = \sum_{j=1}^{d} p_{\lambda_t}(\beta_j)$ satisfies (4.16) for some $\nu_t > 0$. We assume the oracle estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{O}}$ defined in (4.19) satisfies $\min_{j \in S^*} \big|\big(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}\big)_j\big| \geq \nu_t$. Under Assumption 4.1 and Assumption 4.4, we have $\widehat{\boldsymbol{\beta}}_{\lambda_t} = \widehat{\boldsymbol{\beta}}_{\mathrm{O}}$, which implies $\mathrm{supp}\big(\widehat{\boldsymbol{\beta}}_{\lambda_t}\big) = \mathrm{supp}\big(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}\big) = \mathrm{supp}(\boldsymbol{\beta}^*)$.

Proof. See the next section for a detailed proof. □

Recall that the assumption in (4.16) applies to a variety of nonconvex penalties including SCAD and MCP, for which we have $\nu_t = C\lambda_t$ with $C > 0$. According to our discussion for Lemma 4.9, if the minimum signal strength satisfies $\min_{j \in S^*} |\beta_j^*| \geq 2\nu_t$, then for a sufficiently large sample size n, the oracle estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{O}}$ satisfies $\min_{j \in S^*} \big|\big(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}\big)_j\big| \geq \nu_t$. In this situation, Theorem (4.10) holds, i.e., the exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ exactly recovers the support of $\boldsymbol{\beta}^*$. Since $\nu_t = C\lambda_t$, the minimum signal strength required for exact support recovery also shrinks with the decreasing sequence $\{\lambda_t\}_{t=0}^{N}$ along the regularization path. In the examples of least squares and logistic loss, for $t = N$ we have $\nu_N = C\lambda_{\mathrm{tgt}} = C'\sqrt{\log d/n}$. Therefore, for $t = N$ the required minimum signal strength goes to zero as sample size n goes to infinity.

5 Proof of Main Results

In this section we present the proof sketch of the main results. The desired computational and statistical results rely on the strong convexity of the surrogate loss function $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})$, e.g., we need $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})$ to be strongly convex to establish the geometric rate of convergence of the proximal-gradient method within each path following stage. However, $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})$ is nonconvex in general, since $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \mathcal{Q}_\lambda(\boldsymbol{\beta})$, where $\mathcal{L}(\boldsymbol{\beta})$ is possibly nonconvex and $\mathcal{Q}_\lambda(\boldsymbol{\beta})$ is concave. In the following lemma, we prove that $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \mathcal{Q}_\lambda(\boldsymbol{\beta})$ is strongly convex for $\boldsymbol{\beta}$ on a sparse set. This property is also referred to as restricted strongly convexity in the literature (Negahban et al., 2012; Xiao and Zhang, 2012; Zhang and Zhang, 2012). In a similar way, we establish the restricted strong smoothness of $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})$.

Lemma 5.1. Let $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^d$ be two sparse vectors that satisfy $\|(\boldsymbol{\beta} - \boldsymbol{\beta}')_{\overline{S^*}}\|_0 \leq 2\widetilde{s}$, where $\widetilde{s}$ is specified in Assumption 4.4 and $S^* = \mathrm{supp}(\boldsymbol{\beta}^*)$. For $\mathcal{L}(\boldsymbol{\beta})$ being logistic loss, we further assume $\|\boldsymbol{\beta}\|_2 \leq R$ and $\|\boldsymbol{\beta}'\|_2 \leq R$, where R is specified in Definition 4.3. Then the surrogate loss function $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \mathcal{Q}_\lambda(\boldsymbol{\beta})$ satisfies the restricted strong convexity

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

and the restricted strong smoothness

$$\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}') \leq \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})^T(\boldsymbol{\beta}' - \boldsymbol{\beta}) + \frac{\rho_+ - \zeta_+}{2}\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2.$$

Here $\rho_- = \rho_-\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big)$ and $\rho_+ = \rho_+\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big)$ are the sparse eigenvalues specified in Assumption 4.4. As defined in regularity condition (a), $\zeta_-, \zeta_+ > 0$ are the concavity parameters of the nonconvex penalty, which satisfy (4.5) and (4.6).

Proof. See §C.4 in Appendix C for a detailed proof. □

A similar condition has been discussed by Negahban et al. (2012). The main difference is that, our constraint set is a sparse subspace while that of Negahban et al. (2012) is a cone.

Note that in Lemma 5.1, the strong convexity and smoothness of $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})$ rely on the sparsity of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$. Hence, we need to establish results regarding the sparsity of $\boldsymbol{\beta}_t^{(k)}$ throughout the whole iterative procedure. In the setting of logistic loss, we further need to provide an upper bound of $\|\boldsymbol{\beta}_t^{(k)}\|_2$. In the sequel, we provide several important lemmas regarding these required properties of $\boldsymbol{\beta}_t^{(k)}$. The first lemma provides a characterization of any sparse $\boldsymbol{\beta}$ with certain suboptimality.

Lemma 5.2. We assume that $\boldsymbol{\beta}$ satisfies

$$\|\boldsymbol{\beta}_{\overline{S^*}}\|_0 \leq \widetilde{s}, \quad \omega_\lambda(\boldsymbol{\beta}) \leq \lambda/2 \tag{5.1}$$

with $\lambda \geq \lambda_{\text{tgt}}$, where $\omega_\lambda(\boldsymbol{\beta})$ is the measure of suboptimality defined in (3.16). For logistic loss, we further assume $\|\boldsymbol{\beta}\|_2 \leq R$, where $R > 0$ is a constant specified in Definition 4.3. Under Assumption 4.1 and Assumption 4.4, $\boldsymbol{\beta}$ has the following statistical recovery property,

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq C\lambda\sqrt{s^*}, \quad \text{where} \quad C = \frac{21/8}{\rho_- - \zeta_-}.$$

Meanwhile, the objective function value evaluated at $\boldsymbol{\beta}$ satisfies

$$\phi_\lambda(\boldsymbol{\beta}) - \phi_\lambda(\boldsymbol{\beta}^*) \leq C'\lambda^2 s^*, \quad \text{where} \quad C' = \frac{21/2}{\rho_- - \zeta_-}.$$

Proof. See §C.5 of Appendix C for a detailed proof. □

Recall that in our approximate path following method, we use the approximate local solution $\widetilde{\boldsymbol{\beta}}_{t-1}$ obtained within the $(t-1)$-th path following stage to be the initialization of the t-th stage (Line 8 of Algorithm 1), i.e., $\boldsymbol{\beta}_t^{(0)} = \widetilde{\boldsymbol{\beta}}_{t-1}$. By setting $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}_{t-1} = \boldsymbol{\beta}_t^{(0)}$ and $\lambda = \lambda_t$ in Lemma 5.2, we can see that, if $\widetilde{\boldsymbol{\beta}}_{t-1}$ is sparse and $(\lambda_t/2)$-suboptimal, then the initial point $\boldsymbol{\beta}_t^{(0)}$ of the t-th stage has nice statistical recovery performance. However, it remains unclear whether the rest of $\boldsymbol{\beta}_t^{(k)}$'s $(k = 1, 2, \ldots)$ within the t-th stage also have similar recovery performance. To prove this, we first present Lemma 5.3, which shows that under the condition that $\boldsymbol{\beta}$ is sparse and $\phi_\lambda(\boldsymbol{\beta})$ is close to $\phi_\lambda(\boldsymbol{\beta}^*)$, $\boldsymbol{\beta}$ has desired statistical properties. After Lemma 5.3, we will explain that if $\boldsymbol{\beta}_t^{(0)}$ satisfies this condition, then all the $\boldsymbol{\beta}_t^{(k)}$'s $(k = 1, 2, \ldots)$ within the same path following stage also satisfy this condition and thus enjoys nice statistical properties.

Lemma 5.3. Suppose that, for $\lambda \geq \lambda_{\text{tgt}}$, $\boldsymbol{\beta}$ satisfies

$$\|\boldsymbol{\beta}_{\overline{S^*}}\|_0 \leq \widetilde{s}, \quad \phi_\lambda(\boldsymbol{\beta}) - \phi_\lambda(\boldsymbol{\beta}^*) \leq C\lambda^2 s^*, \quad \text{where} \quad C = \frac{21/2}{\rho_- - \zeta_-}.$$

For logistic loss, we further assume $\|\boldsymbol{\beta}\|_2 \leq R$, where R is a constant specified in Definition 4.3. Under Assumption 4.1 and Assumption 4.4, we have

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq C'\lambda\sqrt{s^*}, \quad \text{where} \quad C' = \frac{15/2}{\rho_- - \zeta_-}.$$

Proof. See §C.6 of Appendix C for a detailed proof. □

Let $\lambda = \lambda_t$ and $\boldsymbol{\beta} = \boldsymbol{\beta}_t^{(k)}$ in Lemma 5.3. It suggests that within the t-th path following stage, all $\boldsymbol{\beta}_t^{(k)}$'s ($k = 1, 2, \ldots$) have nice statistical recovery performance under three sufficient conditions: (i) Each $\boldsymbol{\beta}_t^{(k)}$ is sparse; (ii) The objective function value $\phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k)}\big)$ is sufficiently close to $\phi_{\lambda_t}(\boldsymbol{\beta}^*)$; (iii) For logistic loss, we further need $\|\boldsymbol{\beta}_t^{(k)}\|_2 \leq R$. For condition (ii), recall that if we set $\boldsymbol{\beta} = \boldsymbol{\beta}_t^{(0)}$ and $\lambda = \lambda_t$ in Lemma 5.2, then $\boldsymbol{\beta}_t^{(0)}$ being sparse and $(\lambda_t/2)$-suboptimal implies that $\phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(0)}\big)$ is sufficiently close to $\phi_{\lambda_t}(\boldsymbol{\beta}^*)$. Since the proximal-gradient method ensures the monotone decrease of $\big\{\phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k)}\big)\big\}_{k=0}^{\infty}$ within the t-th path following stage (see Lemma C.1 of Appendix C), we have that condition (ii) holds. Meanwhile, condition (iii) obviously holds because of the ℓ_2 constraint. To establish the statistical recovery performance of all the $\boldsymbol{\beta}_t^{(k)}$'s within the t-th stage, we further need to establish the sparsity of $\boldsymbol{\beta}_t^{(k)}$'s to make sure condition (i) holds. To prove this, we present Lemma 5.4, which states that if $\boldsymbol{\beta}$ is sparse, then a proximal-gradient update operation on $\boldsymbol{\beta}$ defined in (3.8) produces a sparse solution under certain conditions.

Lemma 5.4. Suppose that, for $\lambda \geq \lambda_{\mathrm{tgt}}$, $\boldsymbol{\beta}$ satisfies

$$\|\boldsymbol{\beta}_{\overline{S^*}}\|_0 \leq \widetilde{s}, \quad \phi_\lambda(\boldsymbol{\beta}) - \phi_\lambda(\boldsymbol{\beta}^*) \leq C\lambda^2 s^*, \quad \text{and} \quad L < 2(\rho_+ - \zeta_+), \quad \text{where } C = \frac{21/2}{\rho_- - \zeta_-}.$$

For logistic loss, we assume $\|\boldsymbol{\beta}\|_2 \leq R$, where R is specified in Definition 4.3. Under Assumption 4.1 and Assumption 4.4, the proximal-gradient update step defined in (3.8) produces a sparse solution, i.e.,

$$\big\|\big(\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}; R)\big)_{\overline{S^*}}\big\|_0 \leq \widetilde{s}.$$

Here we set $R = +\infty$ if the domain Ω in (3.8) is $\mathbb{R}^d$.

Proof. See §C.7 of Appendix C for a detailed proof. □

Consider $\boldsymbol{\beta} = \boldsymbol{\beta}_t^{(k-1)}$, $\lambda = \lambda_t$ and $L = L_t^{(k)}$, Lemma 5.4 states that, if $\boldsymbol{\beta}_t^{(k-1)}$ is sparse and the objective function value $\phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}\big)$ is close to $\phi_{\lambda_t}(\boldsymbol{\beta}^*)$, then $\boldsymbol{\beta}_t^{(k)} = \mathcal{T}_{L_t^{(k)},\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}; R\big)$ produced by the proximal-gradient update step (3.8) is also sparse. Within the t-th path following stage, if $\boldsymbol{\beta}_t^{(0)}$ is sparse, $\omega_{\lambda_t}\big(\boldsymbol{\beta}_t^{(0)}\big) \leq \lambda_t/2$, and for logistic loss $\big\|\boldsymbol{\beta}_t^{(0)}\big\|_2 \leq R$, then by Lemma 5.2 we have

$$\phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(0)}\big) - \phi_{\lambda_t}(\boldsymbol{\beta}^*) \leq \frac{21/2}{\rho_- - \zeta_-}\lambda_t^2 s^*.$$

Since $\big\{\phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k)}\big)\big\}_{k=0}^{\infty}$ decreases monotonically, we have

$$\phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k)}\big) - \phi_{\lambda_t}(\boldsymbol{\beta}^*) \leq \phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(0)}\big) - \phi_{\lambda_t}(\boldsymbol{\beta}^*) \leq \frac{21/2}{\rho_- - \zeta_-}\lambda_t^2 s^*, \quad \text{for } \ k = 1, 2, \ldots.$$

Assume that we have $L_t^{(k)} \leq 2(\rho_+ - \zeta_+)$ (which will be proved in Theorem 5.5). Applying Lemma 5.4 recursively, we obtain $\big\|\big(\boldsymbol{\beta}_t^{(k)}\big)_{\overline{S^*}}\big\|_0 \leq \widetilde{s}$ ($k = 1, 2, \ldots$). Meanwhile, we have $\big\|\boldsymbol{\beta}_t^{(k)}\big\|_2 \leq R$ due

to the ℓ_2 constraint. Then according to Lemma 5.3, all $\boldsymbol{\beta}_t^{(k)}$'s within the path-following t-th stage have nice recovery performance, i.e.,

$$\left\|\boldsymbol{\beta}_t^{(k)} - \boldsymbol{\beta}^*\right\|_2 \leq \frac{15/2}{\rho_- - \zeta_-}\lambda_t\sqrt{s^*}, \quad \text{for} \quad k = 1, 2, \ldots.$$

Furthermore, based on the sparsity of $\boldsymbol{\beta}_t^{(k)}$, we obtain the restricted strong convexity and smoothness of $\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta})$ by Lemma 5.1, which enable us to establish the geometric rate of convergence within the t-th path following stage. These results are presented in Theorem 5.5.

Theorem 5.5. We assume that, within the t-th path following stage, the proximal-gradient method in Algorithm 3 is initialized by $\boldsymbol{\beta}_t^{(0)}$ and $L_t^{(0)}$, which satisfy

$$\left\|\left(\boldsymbol{\beta}_t^{(0)}\right)_{\overline{S^*}}\right\|_0 \leq \widetilde{s}, \quad \omega_{\lambda_t}\left(\boldsymbol{\beta}_t^{(0)}\right) \leq \lambda_t/2, \quad \text{and} \quad L_t^{(0)} \leq 2(\rho_+ - \zeta_+).$$

For logistic loss we further assume $\left\|\boldsymbol{\beta}_t^{(0)}\right\|_2 \leq R$ with R specified in Definition 4.3. Then we have

$$\left\|\left(\boldsymbol{\beta}_t^{(k)}\right)_{\overline{S^*}}\right\|_0 \leq \widetilde{s}, \quad \left\|\boldsymbol{\beta}_t^{(k)} - \boldsymbol{\beta}^*\right\|_2 \leq \frac{15/2}{\rho_- - \zeta_-}\lambda_t\sqrt{s^*}, \quad \text{and } L_t^{(k)} \leq 2(\rho_+ - \zeta_+), \quad \text{for } k = 1, 2, \ldots. \tag{5.2}$$

Moreover, the iterative sequence $\left\{\boldsymbol{\beta}_t^{(k)}\right\}_{k=0}^{\infty}$ converges towards a unique exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$, which satisfies $\left\|\left(\widehat{\boldsymbol{\beta}}_{\lambda_t}\right)_{\overline{S^*}}\right\|_0 \leq \widetilde{s}$ and the exact optimality condition that $\omega_{\lambda_t}\left(\boldsymbol{\beta}_t^{(k)}\right) \leq 0$.

To achieve an approximate local solution $\widetilde{\boldsymbol{\beta}}_t$ such that $\omega_{\lambda_t}\left(\widetilde{\boldsymbol{\beta}}_t\right) \leq \lambda_t/4$, we need no more than $C' \log\left(4C\sqrt{s^*}\right)$ proximal-gradient iterations defined in Lines 5–9 of Algorithm 3. To achieve an approximate local solution $\widetilde{\boldsymbol{\beta}}_t$ such that $\omega_{\lambda_t}\left(\widetilde{\boldsymbol{\beta}}_t\right) \leq \epsilon_{\text{opt}}$, we need no more than $C' \log\left(C\lambda_t\sqrt{s^*}/\epsilon_{\text{opt}}\right)$ proximal-gradient iterations. Here

$$C = 2\sqrt{21} \cdot \sqrt{\kappa}(1+\kappa), \quad C' = 2 \Big/ \log\left(\frac{1}{1 - 1/(8\kappa)}\right),$$

where κ is the condition number defined in (4.4). In other words, within the t-th path following stage, the proximal-gradient method converges to $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ with a geometric rate of convergence.

Proof. See §C.8 of Appendix C for a detailed proof. □

To prove that the geometric rate of convergence and desired statistical recovery properties hold within all path following stages, i.e., $t = 0, \ldots, N$, we need to verify that the conditions of Theorem 5.5 hold at each stage. We prove by induction. Suppose the initialization of $(t-1)$-th path following stage satisfies

$$\left\|\left(\boldsymbol{\beta}_{t-1}^{(0)}\right)_{\overline{S^*}}\right\|_0 \leq \widetilde{s}, \quad \omega_{\lambda}\left(\boldsymbol{\beta}_{t-1}^{(0)}\right) \leq \lambda_t/2, \quad \text{and} \quad L_{t-1}^{(0)} \leq 2(\rho_+ - \zeta_+). \tag{5.3}$$

Applying Theorem 5.5, we obtain

$$\left\|\left(\boldsymbol{\beta}_{t-1}^{(k)}\right)_{\overline{S^*}}\right\|_0 \leq \widetilde{s}, \quad L_{t-1}^{(k)} \leq 2(\rho_+ - \zeta_+), \quad \text{for} \quad k = 1, 2, \ldots.$$

Consequently, the approximate solution $\widetilde{\boldsymbol{\beta}}_{t-1}$ produced by the $(t-1)$-th stage satisfies $\big\|\big(\widetilde{\boldsymbol{\beta}}_{t-1}\big)_{\overline{S^*}}\big\|_0 \leq \widetilde{s}$, while L_{t-1} satisfies $L_{t-1} \leq 2(\rho_+ - \zeta_+)$. Since we warm start the t-th path following stage with $\boldsymbol{\beta}_t^{(0)} = \widetilde{\boldsymbol{\beta}}_{t-1}$ and $L_t^{(0)} = L_{t-1}$ (Line 8 of Algorithm 1), we have

$$\big\|\big(\boldsymbol{\beta}_t^{(0)}\big)_{\overline{S^*}}\big\|_0 \leq \widetilde{s}, \quad L_t^{(0)} \leq 2(\rho_+ - \zeta_+). \tag{5.4}$$

Moreover, note that the stopping criterion of the proximal-gradient method ensures $\omega_{\lambda_{t-1}}\big(\widetilde{\boldsymbol{\beta}}_{t-1}\big) \leq \lambda_{t-1}/4$ (Line 9 of Algorithm 3). In Lemma C.8 of Appendix C we will prove this implies $\omega_{\lambda_t}\big(\widetilde{\boldsymbol{\beta}}_{t-1}\big) \leq \lambda_t/2$. Consequently, we have

$$\omega_{\lambda_t}\big(\boldsymbol{\beta}_t^{(0)}\big) \leq \lambda_t/2. \tag{5.5}$$

Therefore, we know that (5.3) implies (5.4) and (5.5). We will verify (5.4) and (5.5) hold for $t = 0$ in the proof of Theorem 4.5 in Appendix C.9. By induction, we have that (5.4) and (5.5) hold for $t = 0, \ldots, N$. As a consequence of Theorem 5.5, all path following stages have geometric rates of convergence along the solution path, which implies the global geometric rate of convergence. See Appendix C.9 for a detail proof. Meanwhile, every $\boldsymbol{\beta}_t^{(k)}$ possesses desired statistical properties, i.e.,

$$\big\|\boldsymbol{\beta}_t^{(k)} - \boldsymbol{\beta}^*\big\|_2 \leq \frac{15/2}{\rho_- - \zeta_-}\lambda_t\sqrt{s^*}, \quad \text{for} \;\; t = 1, \ldots, N \;\; \text{and} \;\; k = 0, 1, \ldots,$$

which further leads to the statistical rates of convergence of $\big\{\widetilde{\boldsymbol{\beta}}_t\big\}_{t=1}^N$ in Theorem 4.7, the more refined rates of convergence of $\big\{\widehat{\boldsymbol{\beta}}_{\lambda_t}\big\}_{t=1}^N$ in Theorem 4.8, and the support recovery results in Theorem 4.10. See §C.10–§C.12 of Appendix C for detailed proofs respectively.

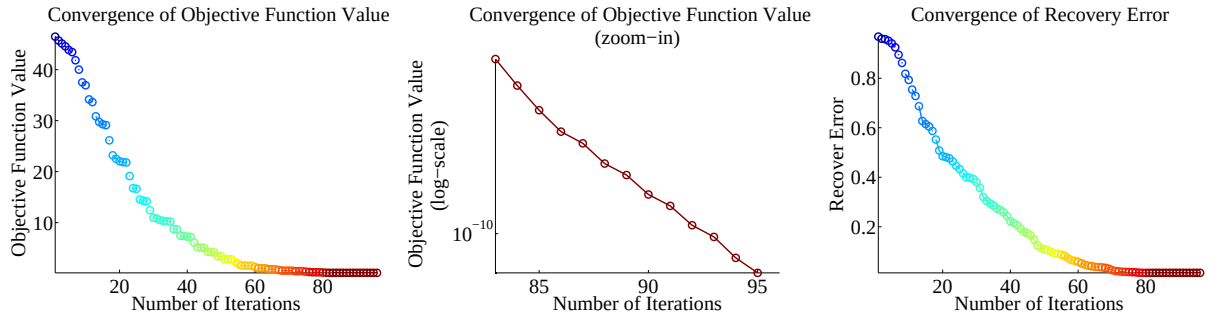
6 Numerical Results

We provide numerical results illustrating the computational efficiency and statistical accuracy of the proposed method. We consider two settings: (i) Semiparametric elliptical design regression with the MCP penalty; (ii) Logistic regression with the MCP penalty. In the first setting, both the loss and penalty functions are nonconvex, while in the second only the penalty function is nonconvex.

In the first experiment, we consider $\mathcal{L}(\boldsymbol{\beta})$ being semiparametric elliptical random design loss and $\mathcal{P}_\lambda(\boldsymbol{\beta})$ being the MCP penalty. The detailed settings are as follows:

- The design matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ contains $n = 500$ independent realizations of a random vector $\boldsymbol{X} \in \mathbb{R}^d$ with $d = 2500$, which follows a t-distribution with 5 degrees of freedom, zero mean and correlation matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}}^0$. We set the correlation matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}}^0$ to be $(\boldsymbol{\Sigma}_{\boldsymbol{X}}^0)_{i,j} = 0.8^{|i-j|}$ $(1 \leq i, j \leq d)$. Meanwhile, in the i-th data sample the response y_i follows a univariate t-distribution with 5 degrees of freedom, mean $\mathbf{x}_i^T\boldsymbol{\beta}^*$ and variance 0.01. Here $\mathbf{x}_i^T$ is the i-th row of the design matrix $\mathbf{X}$, and $\boldsymbol{\beta}^*$ is the true parameter vector specified as follows.

- For the true parameter vector $\boldsymbol{\beta}^* \in \mathbb{R}^d$, we set the first 100 coordinates of $\boldsymbol{\beta}^*$ to be independent realizations of a standard univariate Gaussian distribution (zero mean and unit variance), and the other coordinates to be zero, i.e., we set $s^* = |\mathrm{supp}(\boldsymbol{\beta}^*)| = 100$.



(a) (b) (c)

Figure 3: Semiparametric elliptical design regression with MCP: (a) Plot of the objective function value $\phi_\lambda\big(\boldsymbol{\beta}_t^{(k)}\big)$ along the entire regularization path; (b) Zoom-in plot of $\phi_\lambda\big(\boldsymbol{\beta}_N^{(k)}\big)$ (log-scale) within the N-th path following stage; (c) Plot of the recovery error $\big\|\boldsymbol{\beta}_t^{(k)} - \boldsymbol{\beta}^*\big\|_2$. Here we illustrate each path following stage ($t = 1, \ldots, N$) with a different color. Note that each point in the figure denotes $\boldsymbol{\beta}_t^{(k)}$, which corresponds to the k-th iteration of the proximal-gradient method (Algorithm 3) within the t-th path following stage.

- For the sequence of regularization parameters $\{\lambda_t\}_{t=0}^N$, we set $\lambda_{\mathrm{tgt}} = 0.05$ by cross-validation and $\lambda_0 = \|\nabla\mathcal{L}(\mathbf{0})\|_\infty = \big\|\widehat{\mathbf{K}}_{\boldsymbol{X},Y}\big\|_\infty$. Here $\widehat{\mathbf{K}}_{\boldsymbol{X},Y} \in \mathbb{R}^d$ is defined in (3.13). In our experiment, we fix the random seed to be "2" in MATLAB. In this setting, we observe $\lambda_0 = 2.8516$. We set $\eta = 0.9015$ so that the total number of regularization parameters is $N = \log(\lambda_{\mathrm{tgt}}/\lambda_0)/\log\eta = 39$.

- For the MCP penalty defined in (2.2), we set the tuning parameter to be $b = 1.1$. We set the optimization precision within the N-th path following stage to be $\epsilon_{\mathrm{opt}} = 10^{-6}$. Meanwhile, we set $L_{\min} = 10^{-6}$.

In Figure 3(a) we illustrate the convergence of the objective function value $\phi_\lambda\big(\boldsymbol{\beta}_t^{(k)}\big)$. In Figure 3(b) we zoom into the N-th path following stage and illustrate the geometric rate of convergence. In Figure 3(c) we illustrate the statistical recovery performance of the iterative sequence $\big\{\boldsymbol{\beta}_t^{(k)}\big\}_{t=1}^N$ ($k = 0, 1, \ldots$) attained by our path following method, i.e., $\big\|\boldsymbol{\beta}_t^{(k)} - \boldsymbol{\beta}^*\big\|_2$.

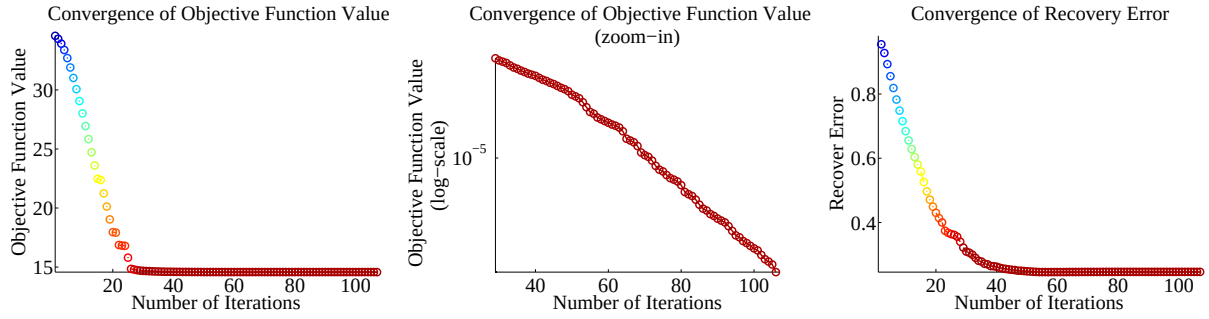
In the second experiment, we consider the setting where $\mathcal{L}(\boldsymbol{\beta})$ is logistic loss and $\mathcal{P}_\lambda(\boldsymbol{\beta})$ is the MCP penalty. The detailed settings are as follows:

- The design matrix $\mathbf{X}$ contains $n = 50$ independent realizations of a random vector $\boldsymbol{X} \in \mathbb{R}^d$ with $d = 100$, which follows a zero mean Gaussian distribution with covariance matrix $10 \cdot \mathbf{I}$. Here $\mathbf{I} \in \mathbb{R}^{d\times d}$ is the identity matrix. Corresponding to the i-th data sample, the response $y_i \in \{0, 1\}$ follows a Bernoulli distribution that satisfies $\mathbb{P}(Y = 0 \,|\, \boldsymbol{X} = \mathbf{x}_i) = \big(1 + \exp(\mathbf{x}_i^T\boldsymbol{\beta}^*)\big)^{-1}$. Here $\mathbf{x}_i^T$ is the i-th row of the design matrix $\mathbf{X}$, and $\boldsymbol{\beta}^*$ is the true parameter vector specified as follows. We set the radius R of the constraint set $\Omega = B_2(R)$ in (3.8) to be 10^3 (Line 3 of Algorithm 1).

- For the true parameter vector $\boldsymbol{\beta}^* \in \mathbb{R}^d$, we set the first 3 coordinates of $\boldsymbol{\beta}^*$ to be 20, and the other coordinates to be zero, i.e., we set $s^* = |\mathrm{supp}(\boldsymbol{\beta}^*)| = 3$.

- For the regularization parameters, we set $\lambda_{\mathrm{tgt}} = 0.12$ by cross-validation and $\lambda_0 = \|\nabla \mathcal{L}(\mathbf{0})\|_\infty$. In our experiment, we fix the random seed to be "2" in MATLAB. We observe that $\lambda_0 = 1.2$. Correspondingly, we set $\eta = 0.9035$ so that the total number of regularization parameters along the regularization path is $N = \log(\lambda_{\mathrm{tgt}}/\lambda_0)/\log\eta = 22$.
- For the MCP penalty defined in (2.2), we set the tuning parameter to be $b = 2$. We set the optimization precision within the N-th path following stage to be $\epsilon_{\mathrm{opt}} = 10^{-6}$. Meanwhile, we set $L_{\min} = 10^{-6}$.

Similar to Figure 3, in Figure 4 we illustrate the convergence of the objective function value, as well as the statistical recovery performance of the iterative sequence $\{\boldsymbol{\beta}_t^{(k)}\}_{t=1}^N$ $(k = 0, 1, \ldots)$ that is attained by our path following method.



(a) (b) (b)

Figure 4: Logistic regression with MCP: (a) Plot of the objective function value $\phi_\lambda(\boldsymbol{\beta}_t^{(k)})$ along the entire regularization path; (b) Zoom-in plot of $\phi_\lambda(\boldsymbol{\beta}_N^{(k)})$ (log-scale) within the N-th path following stage; (c) Plot of the recovery error $\|\boldsymbol{\beta}_t^{(k)} - \boldsymbol{\beta}^*\|_2$.

7 Conclusion

In this paper, we provided an integrated theory for penalized M-estimators with possibly nonconvex loss or penalty functions. These problems are motivated by generalized linear models with nonconvex penalties and semiparametric elliptical design regression, as well as a broad range of other applications. Since it is intractable to compute the global solutions of these problems due to the nonconvex formulation, we need to establish a theory that characterizes both the computational and statistical properties of the local solutions obtained by specific algorithms. For this purpose, we proposed an approximate regularization path following method which serves as a unified framework for solving a variety of high-dimensional sparse learning problems with nonconvexity. Computationally, our method enjoys a fast global geometric rate of convergence for calculating the entire regularization path; Statistically, all the approximate and exact local solutions along the regularization path attained by our method enjoy sharp statistical rate of convergence in both estimation and support recovery. In particular, we provide a sharp theoretical analysis that demonstrates the advantage of using nonconvex penalties. This paper demonstrates that under suitable condi-

tions, the entire regularization path of a broad class of nonconvex sparse learning problems can be efficiently obtained.

Acknowledgement

We sincerely thank Po-Ling Loh and Martin Wainwright for their helpful personal communications. Han Liu is supported by NSF Grant III-1116730 and NIH sub-awards from both Johns Hopkins University and Harvard University. Tong Zhang is supported by NSF IIS-1250985 and NSF DMS-1007527.

A Illustration of Regularity Condition (e) for Nonconvex Penalty

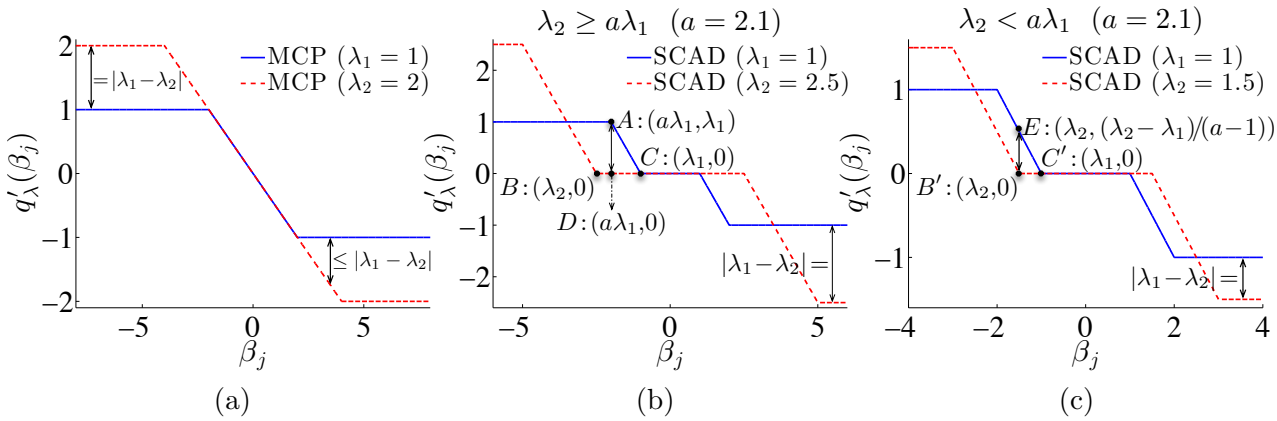


Figure 5: An illustration of regularity condition (e) for MCP and SCAD: (a) Plots of $q'_{\lambda_1}(\beta_j)$ and $q'_{\lambda_2}(\beta_j)$ for MCP with $\lambda_1 = 1$, $\lambda_2 = 2$ and $b = 2$; (b) Plots of $q'_{\lambda_1}(\beta_j)$ and $q'_{\lambda_2}(\beta_j)$ for SCAD with $\lambda_1 = 1$, $\lambda_2 = 2.5$ and $a = 2.1$; (c) Plots of $q'_{\lambda_1}(\beta_j)$ and $q'_{\lambda_2}(\beta_j)$ for SCAD with $\lambda_1 = 1$, $\lambda_2 = 1.5$ and $a = 2.1$. Subfigure (a) shows that regularity condition (e) holds for MCP. For SCAD, we consider two cases: $\lambda_2 \geq a\lambda_1$, as illustrated in (b); $\lambda_2 < a\lambda_1$ as illustrated in (c). In the first case, $|AD| = \lambda_1 \leq (a-1)\lambda_1 \leq |\lambda_1 - \lambda_2|$ since $a > 2$ and $\lambda_2 \geq a\lambda_1$. In the second case, $|B'E| = (\lambda_2 - \lambda_1)/(a-1) \leq |\lambda_1 - \lambda_2|$, because the slope of EC' is $(-1/(a-1))$ with $a > 2$.

B Derivation of Optimization Update Schemes

To simplify the notation, we denote $L_t^{(k)}$ by L, $\boldsymbol{\beta}_t^{(k-1)}$ by $\boldsymbol{\beta}'$, and λ_t by λ in the rest of this section.

Derivation of (3.10): If $\Omega = \mathbb{R}^d$, then we have

$$\begin{aligned}
\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}'; +\infty) &= \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}} \left\{\psi_{L,\lambda}(\boldsymbol{\beta};\boldsymbol{\beta}')\right\} \\
&= \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}} \left\{\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}') + \nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}')^T(\boldsymbol{\beta}-\boldsymbol{\beta}') + \frac{L}{2}\|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\right\} \\
&= \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}} \left\{\frac{1}{2}\Big\|\boldsymbol{\beta} - \underbrace{\Big(\boldsymbol{\beta}' - \frac{1}{L}\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}')\Big)}_{\bar{\boldsymbol{\beta}}}\Big\|_2^2 + \frac{\lambda}{L}\|\boldsymbol{\beta}\|_1\right\}. \qquad \text{(B.1)}
\end{aligned}$$

It is known that the minimizer of (B.1) can be obtained by soft-thresholding $\bar{\boldsymbol{\beta}}$ with the threshold of λ/L, i.e.,

$$\left(\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}';+\infty)\right)_j = \begin{cases} 0 & \text{if } |\bar{\beta}_j| \le \lambda/L, \\ \operatorname{sign}(\bar{\beta}_j)(|\bar{\beta}_j| - \lambda/L) & \text{if } |\bar{\beta}_j| > \lambda/L. \end{cases} \qquad \text{(B.2)}$$

Therefore we obtain the first update scheme (3.10) for $\Omega = \mathbb{R}^d$.

Derivation of (3.12): If $\Omega = B_2(R) = \left\{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_2^2 \le R^2\right\}$, by Lagrangian duality we can transform the original optimization problem with constraint into an unconstraint optimization problem. Hence, there exists a Lagrangian multiplier $\tau \ge 0$ such that

$$\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}'; R) = \underset{\boldsymbol{\beta}\in B_2(R)}{\operatorname{argmin}} \left\{\psi_{L,\lambda}(\boldsymbol{\beta};\boldsymbol{\beta}')\right\} = \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}} \left\{\psi_{L,\lambda}(\boldsymbol{\beta};\boldsymbol{\beta}') + \frac{\tau}{2}\|\boldsymbol{\beta}\|_2^2\right\}.$$

Consequently, based on (B.1) we have

$$\begin{aligned}
\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}'; R) &= \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}} \left\{\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}') + \nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}')^T(\boldsymbol{\beta}-\boldsymbol{\beta}') + \frac{L}{2}\|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 + \frac{\tau}{2}\|\boldsymbol{\beta}\|_2^2\right\} \\
&= \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}} \left\{\frac{L+\tau}{2}\|\boldsymbol{\beta}\|_2^2 - \Big(L\cdot\boldsymbol{\beta}' - \nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}')\Big)^T\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1\right\} \\
&= \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}} \left\{\frac{1}{2}\Big\|\boldsymbol{\beta} - \underbrace{\Big(\frac{L}{L+\tau}\boldsymbol{\beta}' - \frac{1}{L+\tau}\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}')\Big)}_{\frac{L}{L+\tau}\bar{\boldsymbol{\beta}}}\Big\|_2^2 + \frac{\lambda}{L+\tau}\|\boldsymbol{\beta}\|_1\right\}, \qquad \text{(B.3)}
\end{aligned}$$

where $\bar{\boldsymbol{\beta}} = \boldsymbol{\beta}' - \nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}')/L$. The minimizer of (B.3) can also be obtained by soft-thresholding, i.e.,

$$\left(\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}'; R)\right)_j = \begin{cases} 0 & \text{if } \dfrac{L}{L+\tau}|\bar{\beta}_j| \le \dfrac{\lambda}{L+\tau}, \\ \operatorname{sign}\Big(\dfrac{L}{L+\tau}\bar{\beta}_j\Big)\Big(\dfrac{L}{L+\tau}|\bar{\beta}_j| - \dfrac{\lambda}{L+\tau}\Big) & \text{if } \dfrac{L}{L+\tau}|\bar{\beta}_j| > \dfrac{\lambda}{L+\tau}. \end{cases} \qquad \text{(B.4)}$$

Comparing (B.4) with (B.2), we have

$$\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}';R)=\frac{L}{L+\tau}\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}';+\infty). \tag{B.5}$$

In other words, we can obtain the constraint solution $\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}';R)$ by first calculating the unconstraint solution $\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}';+\infty)$, and then rescaling it by a factor of $L/(L+\tau)$. Note that here the Lagrangian multiplier τ is unknown. We discuss the following two cases:

- If the constraint $\boldsymbol{\beta}\in B_2(R)$ is not active, then we have $\tau=0$ by complementary slackness, which implies
$$\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}';R)=\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}';+\infty).$$
Since the constraint is not active, we have
$$\|\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}';R)\|_2=\|\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}';+\infty)\|_2<R.$$
- If the constraint $\boldsymbol{\beta}\in B_2(R)$ is active, then we have $\tau\geq 0$ by complementary slackness. In this case, the minimizer $\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}';R)$ lies on the boundary of $B_2(R)$. By (B.5) we have
$$\|\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}';+\infty)\|_2=\frac{L+\tau}{L}\|\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}';R)\|_2=\frac{L+\tau}{L}R\geq R.$$
To obtain $\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}';R)$, we project $\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}';+\infty)$ onto $B_2(R)$, which can be achieved by
$$\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}';R)=\frac{R\cdot\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}';+\infty)}{\|\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}';+\infty)\|_2}.$$

Therefore we obtain the second update scheme (3.12) for $\Omega=B_2(R)$.

C Proof of Theoretical Results

To analyze the computational properties of our approximate regularization path following method, we first provide several useful lemmas about Nesterov's proximal-gradient method used within each stage of the path following method.

C.1 Preliminary Results about the Proximal-Gradient Method

Recall that the objective function can be formulated as $\phi_{\lambda_t}(\boldsymbol{\beta})=\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta})+\lambda_t\|\boldsymbol{\beta}\|_1$ where $\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta})=\mathcal{L}(\boldsymbol{\beta})+\mathcal{Q}_{\lambda_t}(\boldsymbol{\beta})$, while $\psi_{L_t^{(k)},\lambda_t}\big(\boldsymbol{\beta};\boldsymbol{\beta}_t^{(k-1)}\big)$ is the local quadratic approximation of $\phi_{\lambda_t}(\boldsymbol{\beta})$ at $\boldsymbol{\beta}_t^{(k-1)}$ defined in (3.7). The following lemma, which adapts from Nesterov (2007), characterizes the decrement of the objective function.

Lemma C.1. Under Assumption 4.4, we assume $\big\|\big(\boldsymbol{\beta}_t^{(k-1)}\big)_{\overline{S^*}}\big\|_0\leq\widetilde{s}$, where $\widetilde{s}$ is the positive integer specified in Assumption 4.4. For any $L_t^{(k)}>0$ and fixed $\lambda_t\in[\lambda_{\mathrm{tgt}},\lambda_0]$, we have

$$\phi_\lambda\big(\boldsymbol{\beta}_t^{(k)}\big)\leq\phi_\lambda\big(\boldsymbol{\beta}_t^{(k-1)}\big)-\frac{L_t^{(k)}}{2}\big\|\boldsymbol{\beta}_t^{(k)}-\boldsymbol{\beta}_t^{(k-1)}\big\|_2^2.$$

Recall that as defined in (3.16), $\omega_\lambda(\boldsymbol{\beta})$ describes the suboptimality of approximate solutions. The following lemma, which follows from Nesterov (2007), upper bounds $\omega_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k)}\big)$ with $\big\|\boldsymbol{\beta}_t^{(k)}-\boldsymbol{\beta}_t^{(k-1)}\big\|_2$.

Lemma C.2. Under the assumptions of Lemma C.1, then we have

$$\omega_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k)}\big) \le \big(L_t^{(k)} + \rho_+ - \zeta_-\big)\big\|\boldsymbol{\beta}_t^{(k)} - \boldsymbol{\beta}_t^{(k-1)}\big\|_2,$$

where $\rho_+ = \rho_+\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big)$ is the sparse eigenvalue specified in Assumption 4.4; As defined in regularity condition (a), $\zeta_+ > 0$ is the concavity parameter of the nonconvex penalty, which satisfies (4.6).

C.2 Upper Bounds of $\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty$

In this section, we provide upper bounds of $\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty$ to justify Assumption 4.1.

Lemma C.3. For least squares regression with sub-Gaussian noise and logistic regression, we assume that the columns of $\mathbf{X}$ are normalized in such a way that $\max_{j\in\{1,\ldots,d\}}\big\{\|\mathbf{X}_j\|_2\big\} \le \sqrt{n}$. Then we have

$$\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty \le C\sqrt{\frac{\log d}{n}} \tag{C.1}$$

with probability at least $1 - d^{-1}$, where C is a constant.

Proof. See Candés and Tao (2007); Zhang and Huang (2008); Zhang (2009); Bickel et al. (2009); Koltchinskii (2009a); van de Geer and Bühlmann (2009); Negahban et al. (2012); Wainwright (2009) for a detailed proof. □

Lemma C.4. For semiparametric elliptical design regression, we have, with probability at least $1 - (d+1)^{-5/2} - 2(d+1)^{-3}$,

$$\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty \le C\|\boldsymbol{\beta}^*\|_1\sqrt{\frac{\log d}{n}}, \tag{C.2}$$

where C is a constant.

Proof. See §D.3 of Appendix D for a detailed proof. □

C.3 Justification of Assumption 4.4

In this section, we show that Assumption 4.4 holds with high probability for semiparametric elliptical design loss and logistic loss.

First we provide two lemmas regarding the largest and smallest sparse eigenvalues of the Hessian matrix $\nabla^2\mathcal{L}(\boldsymbol{\beta})$ of semiparametric elliptical design loss and logistic loss. Then we will use them to justify Assumption 4.4.

Lemma C.5. Let n be the sample size, d be the dimension of $\boldsymbol{\beta}$, and $\boldsymbol{Z} \in \mathbb{R}^{d+1}$ be an elliptically distributed random vector defined in §2.2. The corresponding covariance matrix estimator $\widehat{\mathbf{K}}_{\boldsymbol{Z}} \in \mathbb{R}^{(d+1)\times(d+1)}$ is defined in (2.7), while its submatrix $\widehat{\mathbf{K}}_{\boldsymbol{X}} \in \mathbb{R}^{d\times d}$ is defined in (3.13). The Hessian matrix of semiparametric elliptical design loss is $\nabla^2\mathcal{L}(\boldsymbol{\beta}) = \widehat{\mathbf{K}}_{\boldsymbol{X}}$. Let s be a positive integer that indicates the sparsity level. Under suitable conditions (see Han and Liu (2013) for details), for a sufficiently large n, there exists an s such that $\rho_-\big(\nabla^2\mathcal{L}, s\big) > 0$ and $\rho_+\big(\nabla^2\mathcal{L}, s\big) < +\infty$, both with probability at least $1 - 2d^{-1} - 3d^{-2}$. Here $\rho_+\big(\nabla^2\mathcal{L}, s\big)$ and $\rho_-\big(\nabla^2\mathcal{L}, s\big)$ are defined in Definition 4.2.

Proof. See §D.2 in Appendix D for a detailed proof. □

In the following we provide a similar lemma for logistic loss. RIP-like conditions for logistic loss have been widely studied (van de Geer, 2008; Negahban et al., 2012; Loh and Wainwright, 2013). To simplify the analysis, we utilize a result from Loh and Wainwright (2013) to prove the following lemma.

Lemma C.6. Let n be the sample size, d be the dimension. Suppose $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \in \mathbb{R}^{n\times d}$ is a sub-Gaussian design matrix, where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent realizations of a sub-Gaussian random vector with zero mean, unit variance proxy and independent entries. For logistic loss, the Hessian matrix is defined in (4.2). Let s be a positive integer that indicates the sparsity level, and R be a positive constant. For a sufficiently large n, there exists an integer s such that $\rho_-\big(\nabla^2\mathcal{L}, s\big) > 0$ and $\rho_+\big(\nabla^2\mathcal{L}, s\big) < +\infty$, both with probability at least $1 - C\exp(-C'n)$, where $C, C' > 0$. Here $\rho_-\big(\nabla^2\mathcal{L}, s\big)$ and $\rho_+\big(\nabla^2\mathcal{L}, s\big)$ are defined in Definition 4.3.

Proof. For logistic loss, Loh and Wainwright (2013, Proposition 1) showed that, for $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^d$ such that $\|\boldsymbol{\beta}\|_2 \le R$ and $\|\boldsymbol{\beta}'\|_2 \le R$, we have

$$\mathcal{L}(\boldsymbol{\beta}') - \mathcal{L}(\boldsymbol{\beta}) - \nabla\mathcal{L}(\boldsymbol{\beta})^T(\boldsymbol{\beta}' - \boldsymbol{\beta}) \le C\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2^2 + \frac{2C}{3}\cdot\frac{\log d}{n}\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_1^2, \tag{C.3}$$

$$\mathcal{L}(\boldsymbol{\beta}') - \mathcal{L}(\boldsymbol{\beta}) - \nabla\mathcal{L}(\boldsymbol{\beta})^T(\boldsymbol{\beta}' - \boldsymbol{\beta}) \ge C'\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2^2 - C''\cdot\frac{\log d}{n}\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_1^2, \tag{C.4}$$

both with probability at least $1 - C'''\exp(-C''''n)$. All these constants are positive. By Taylor's theorem and the mean value theorem, we have

$$\mathcal{L}(\boldsymbol{\beta}') = \mathcal{L}(\boldsymbol{\beta}) + \nabla\mathcal{L}(\boldsymbol{\beta})^T(\boldsymbol{\beta}' - \boldsymbol{\beta}) + \frac{1}{2}(\boldsymbol{\beta}' - \boldsymbol{\beta})^T\nabla^2\mathcal{L}\big(\gamma\boldsymbol{\beta}' + (1-\gamma)\boldsymbol{\beta}\big)(\boldsymbol{\beta}' - \boldsymbol{\beta}),$$

where $\gamma \in [0, 1]$. Plugging this into the left-hand sides of (C.3) and (C.4), we obtain

$$\frac{1}{2}(\boldsymbol{\beta}' - \boldsymbol{\beta})^T\nabla^2\mathcal{L}\big(\gamma\boldsymbol{\beta}' + (1-\gamma)\boldsymbol{\beta}\big)(\boldsymbol{\beta}' - \boldsymbol{\beta}) \le C\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2 + \frac{2C}{3}\cdot\frac{\log d}{n}\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_1^2, \tag{C.5}$$

$$\frac{1}{2}(\boldsymbol{\beta}' - \boldsymbol{\beta})^T\nabla^2\mathcal{L}\big(\gamma\boldsymbol{\beta}' + (1-\gamma)\boldsymbol{\beta}\big)(\boldsymbol{\beta}' - \boldsymbol{\beta}) \ge C'\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2 - C''\cdot\frac{\log d}{n}\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_1^2. \tag{C.6}$$

Assume that $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ satisfy $\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_0 \le s$, which implies $\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_1 \le \sqrt{s}\cdot\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2$. Plugging this upper bound of $\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_1$ into the right-hand sides of (C.5) and (C.6), we have

$$\frac{1}{2}(\boldsymbol{\beta}' - \boldsymbol{\beta})^T\nabla^2\mathcal{L}\big(\gamma\boldsymbol{\beta}' + (1-\gamma)\boldsymbol{\beta}\big)(\boldsymbol{\beta}' - \boldsymbol{\beta}) \le \left(C + \frac{2C}{3}\cdot\frac{s\log d}{n}\right)\cdot\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2, \tag{C.7}$$

$$\frac{1}{2}(\boldsymbol{\beta}' - \boldsymbol{\beta})^T\nabla^2\mathcal{L}\big(\gamma\boldsymbol{\beta}' + (1-\gamma)\boldsymbol{\beta}\big)(\boldsymbol{\beta}' - \boldsymbol{\beta}) \ge \left(C' - C''\cdot\frac{s\log d}{n}\right)\cdot\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2. \tag{C.8}$$

In (C.7) and (C.8), taking $n \geq C''' \cdot s \log d/n$ with a sufficiently large $C''' > 0$, and dividing $\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2$ on both sides, we obtain

$$\frac{C'}{2} \leq \frac{1}{2} \cdot \frac{(\boldsymbol{\beta}' - \boldsymbol{\beta})^T}{\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2} \cdot \nabla^2 \mathcal{L}\big(\gamma \boldsymbol{\beta}' + (1-\gamma)\boldsymbol{\beta}\big) \cdot \frac{(\boldsymbol{\beta}' - \boldsymbol{\beta})}{\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2} \leq 2C. \tag{C.9}$$

Let $\boldsymbol{v} = (\boldsymbol{\beta}' - \boldsymbol{\beta})/\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2$. Obviously, $\boldsymbol{v}$ is an arbitrary vector that satisfies $\|\boldsymbol{v}\|_2 = 1$ and $\|\boldsymbol{v}\|_0 \leq s$. Taking $\boldsymbol{\beta}' \to \boldsymbol{\beta}$, we have $C' \leq \boldsymbol{v}^T \nabla^2 \mathcal{L}(\boldsymbol{\beta}) \boldsymbol{v} \leq 4C$ for any $\boldsymbol{\beta} \leq R$ and any $\boldsymbol{v}$ such that $\|\boldsymbol{v}\|_2 = 1$ and $\|\boldsymbol{v}\|_0 \leq s$. By Definition 4.3 of $\rho_-\big(\nabla^2 \mathcal{L}, s\big)$ and $\rho_+\big(\nabla^2 \mathcal{L}, s\big)$, we have $\rho_-\big(\nabla^2 \mathcal{L}, s\big) \geq C' > 0$ and $\rho_+\big(\nabla^2 \mathcal{L}, s\big) \leq 4C < +\infty$. Thus we conclude the proof. □

Equipped with Lemma C.5 and Lemma C.6, we are ready to justify Assumption 4.4 for semi-parametric elliptical design loss and logistic loss. Recall that $s^* = \|\boldsymbol{\beta}^*\|_0$, where $\boldsymbol{\beta}^*$ is the true parameter vector. We assume that Lemma C.5 or Lemma C.6 holds with $s = Cs^*$, $\rho_+\big(\nabla^2 \mathcal{L}, s\big) = C'$ and $\rho_-\big(\nabla^2 \mathcal{L}, s\big) = C''$, where C satisfies

$$C \geq 2\left(144 \cdot \left(\frac{2C'}{C''}\right)^2 + 250 \cdot \left(\frac{2C'}{C''}\right)\right) + 1. \tag{C.10}$$

Meanwhile, we set the concavity parameter of the nonconvex penalty to be $\zeta_+ = 0$ and $\zeta_- = C''/2$. Now we verify that there exists an integer $\widetilde{s} = (C-1)/2 \cdot s^*$, where C satisfies (C.10), that satisfies Assumption 4.4. Note that the condition number κ defined in (4.4) is

$$\kappa = \frac{\rho_+\big(\nabla^2 \mathcal{L}, s^* + 2\widetilde{s}\big) - \zeta_+}{\rho_-\big(\nabla^2 \mathcal{L}, s^* + 2\widetilde{s}\big) - \zeta_-} = \frac{\rho_+\big(\nabla^2 \mathcal{L}, Cs^*\big) - \zeta_+}{\rho_-\big(\nabla^2 \mathcal{L}, Cs^*\big) - \zeta_-} = \frac{\rho_+\big(\nabla^2 \mathcal{L}, s\big) - \zeta_+}{\rho_-\big(\nabla^2 \mathcal{L}, s\big) - \zeta_-} = \frac{C'}{C'' - C''/2} = \frac{2C'}{C''}.$$

Since $\widetilde{s} = (C-1)/2 \cdot s^*$ where C satisfies (C.10), we have

$$\widetilde{s} \geq \left(144 \cdot \left(\frac{2C'}{C''}\right)^2 + 250 \cdot \left(\frac{2C'}{C''}\right)\right) \cdot s^* = (144\kappa^2 + 250\kappa) \cdot s^*.$$

Hence we find an $\widetilde{s}$ that satisfies the requirements in Assumption 4.4.

C.4 Proof of Lemma 5.1

Proof. Recall that $\mathcal{Q}_\lambda(\boldsymbol{\beta})$ is the concave component of the nonconvex penalty $\mathcal{P}_\lambda(\boldsymbol{\beta})$, which implies $-\mathcal{Q}_\lambda(\boldsymbol{\beta})$ is convex. Meanwhile, recall that $\mathcal{Q}_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^d q_\lambda(\beta_j)$, where $q_\lambda(\beta_j)$ satisfies regularity condition (a). Hence we have

$$-\zeta_-(\beta_j' - \beta_j)^2 \leq \big(q_\lambda'(\beta_j') - q_\lambda'(\beta_j)\big)(\beta_j' - \beta_j) \leq -\zeta_+(\beta_j' - \beta_j)^2,$$

which implies the convex function $-\mathcal{Q}_\lambda(\boldsymbol{\beta})$ satisfies

$$\Big(\nabla\big(-\mathcal{Q}_\lambda(\boldsymbol{\beta}')\big) - \nabla\big(-\mathcal{Q}_\lambda(\boldsymbol{\beta})\big)\Big)^T(\boldsymbol{\beta}' - \boldsymbol{\beta}) \quad \leq \quad \zeta_-\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2, \tag{C.11}$$

$$\Big(\nabla\big(-\mathcal{Q}_\lambda(\boldsymbol{\beta}')\big) - \nabla\big(-\mathcal{Q}_\lambda(\boldsymbol{\beta})\big)\Big)^T(\boldsymbol{\beta}' - \boldsymbol{\beta}) \quad \geq \quad \zeta_+\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2. \tag{C.12}$$

According to Nesterov (2004, Theorem 2.1.5 & Theorem 2.1.9), (C.11) and (C.12) are equivalent definitions of strong smoothness and strong convexity respectively. In other words, $-\mathcal{Q}_\lambda(\boldsymbol{\beta})$ satisfies

$$-\mathcal{Q}_\lambda(\boldsymbol{\beta}') \leq -\mathcal{Q}_\lambda(\boldsymbol{\beta}) - \nabla\mathcal{Q}(\boldsymbol{\beta})^T(\boldsymbol{\beta}' - \boldsymbol{\beta}) + \frac{\zeta_-}{2}\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2, \tag{C.13}$$

$$-\mathcal{Q}_\lambda(\boldsymbol{\beta}') \geq -\mathcal{Q}_\lambda(\boldsymbol{\beta}) - \nabla\mathcal{Q}(\boldsymbol{\beta})^T(\boldsymbol{\beta}' - \boldsymbol{\beta}) + \frac{\zeta_+}{2}\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2. \tag{C.14}$$

For loss function $\mathcal{L}(\boldsymbol{\beta})$, by Taylor's theorem and the mean value theorem, we have

$$\mathcal{L}(\boldsymbol{\beta}') = \mathcal{L}(\boldsymbol{\beta}) + \nabla\mathcal{L}(\boldsymbol{\beta})^T(\boldsymbol{\beta}' - \boldsymbol{\beta}) + \frac{1}{2}(\boldsymbol{\beta}' - \boldsymbol{\beta})^T\nabla^2\mathcal{L}\big(\gamma\boldsymbol{\beta} + (1-\gamma)\boldsymbol{\beta}'\big)(\boldsymbol{\beta}' - \boldsymbol{\beta}), \tag{C.15}$$

where $\gamma \in [0,1]$. Note that we assume $\|(\boldsymbol{\beta}' - \boldsymbol{\beta})_{\overline{S^*}}\|_0 \leq 2\widetilde{s}$, which implies $\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_0 \leq s^* + 2\widetilde{s}$. For logistic loss, we assume $\|\boldsymbol{\beta}\|_2 \leq R$ and $\|\boldsymbol{\beta}'\|_2 \leq R$, which implies $\|\gamma\boldsymbol{\beta} + (1-\gamma)\boldsymbol{\beta}'\|_2 \leq R$ by the convexity of ℓ_2 norm. By Definition 4.2 and Definition 4.3, we have

$$\rho_-\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big) \leq \frac{(\boldsymbol{\beta}' - \boldsymbol{\beta})^T}{\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2}\nabla^2\mathcal{L}\big(\gamma\boldsymbol{\beta} + (1-\gamma)\boldsymbol{\beta}'\big)\frac{(\boldsymbol{\beta}' - \boldsymbol{\beta})}{\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2} \leq \rho_+\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big).$$

Plugging this into the right-hand side of (C.15), we have

$$\mathcal{L}(\boldsymbol{\beta}') \geq \mathcal{L}(\boldsymbol{\beta}) + \nabla\mathcal{L}(\boldsymbol{\beta})^T(\boldsymbol{\beta}' - \boldsymbol{\beta}) + \frac{\rho_-\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big)}{2}\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2, \tag{C.16}$$

$$\mathcal{L}(\boldsymbol{\beta}') \leq \mathcal{L}(\boldsymbol{\beta}) + \nabla\mathcal{L}(\boldsymbol{\beta})^T(\boldsymbol{\beta}' - \boldsymbol{\beta}) + \frac{\rho_+\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big)}{2}\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2. \tag{C.17}$$

Recall that $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \mathcal{Q}_\lambda(\boldsymbol{\beta})$. Subtracting (C.13) from (C.16), and (C.14) from (C.17), we obtain

$$\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}') \geq \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})^T(\boldsymbol{\beta}' - \boldsymbol{\beta}) + \frac{\rho_-\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big) - \zeta_-}{2}\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2$$

$$\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}') \leq \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})^T(\boldsymbol{\beta}' - \boldsymbol{\beta}) + \frac{\rho_+\big(\nabla^2\mathcal{L}, s^* + 2\widetilde{s}\big) - \zeta_+}{2}\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2.$$

Then we conclude the proof. □

C.5 Proof of Lemma 5.2

Proof. **Results for Statistical Recovery:** Since $\|\boldsymbol{\beta}_{\overline{S^*}}\|_0 \leq \widetilde{s}$ and $\|\boldsymbol{\beta}^*_{\overline{S^*}}\|_0 = 0$, we have $\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\overline{S^*}}\| \leq \widetilde{s}$. For logistic loss, we further have $\|\boldsymbol{\beta}\|_2 \leq R$ and $\|\boldsymbol{\beta}^*\|_2 \leq R$, where R is specified in Definition 4.3. Thus Lemma 5.1 gives

$$\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) \geq \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + (\boldsymbol{\beta}^* - \boldsymbol{\beta})^T\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \frac{\rho_- - \zeta_-}{2}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|_2^2, \tag{C.18}$$

$$\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) \geq \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) + \frac{\rho_- - \zeta_-}{2}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|_2^2. \tag{C.19}$$

Adding (C.18) and (C.19) and moving $(\boldsymbol{\beta}^* - \boldsymbol{\beta})^T\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})$ to the left-hand side, we obtain

$$(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) \geq (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) + (\rho_- - \zeta_-)\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|_2^2. \tag{C.20}$$

Let $\boldsymbol{\xi} \in \partial\|\boldsymbol{\beta}\|_1$ be the subgradient that attains the minimum in

$$\omega_\lambda(\boldsymbol{\beta}) = \min_{\boldsymbol{\xi}' \in \partial\|\boldsymbol{\beta}\|_1} \max_{\boldsymbol{\beta}' \in \Omega} \left\{ \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}')^T}{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_1} \big(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda \boldsymbol{\xi}'\big) \right\}.$$

Then we have

$$\omega_\lambda(\boldsymbol{\beta}) = \max_{\boldsymbol{\beta}' \in \Omega} \left\{ \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}')^T}{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_1} \big(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda \boldsymbol{\xi}\big) \right\}. \tag{C.21}$$

Adding $\lambda(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \boldsymbol{\xi}$ to the both sides of (C.20), we obtain

$$(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \big(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda \boldsymbol{\xi}\big) \geq (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) + (\rho_- - \zeta_-)\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|_2^2 + \lambda(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \boldsymbol{\xi}.$$

Since $\boldsymbol{\beta}^* \in \Omega$, by (C.21) we have

$$\frac{(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T}{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1} \big(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda \boldsymbol{\xi}\big) \leq \max_{\boldsymbol{\beta}' \in \Omega} \left\{ \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}')^T}{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_1} \big(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda \boldsymbol{\xi}\big) \right\} = \omega_\lambda(\boldsymbol{\beta}). \tag{C.22}$$

Recall that we assume $\omega_\lambda(\boldsymbol{\beta}) \leq \lambda/2$, we obtain

$$(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \big(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda \boldsymbol{\xi}\big) \leq \lambda/2 \cdot \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1. \tag{C.23}$$

Plugging (C.23) into the left-hand side of (C.20), we obtain

$$\lambda/2 \cdot \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \geq \underbrace{(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*)}_{\text{(i)}} + (\rho_- - \zeta_-)\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|_2^2 + \underbrace{\lambda(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \boldsymbol{\xi}}_{\text{(ii)}}. \tag{C.24}$$

Now we provide lower bounds of terms (i) and (ii) in (C.24) respectively.

- **Bounding Term (i) in (C.24):** Recall that $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \mathcal{Q}_\lambda(\boldsymbol{\beta})$. We have

$$(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) = \underbrace{(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \nabla \mathcal{L}(\boldsymbol{\beta}^*)}_{\text{(i).a}} + \underbrace{(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \nabla \mathcal{Q}_\lambda(\boldsymbol{\beta}^*)}_{\text{(i).b}}. \tag{C.25}$$

  Separating the support of $\boldsymbol{\beta} - \boldsymbol{\beta}^*$ into S^* and $\overline{S^*}$, we obtain

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 = \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\overline{S^*}}\|_1 + \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_1.$$

  Then for term (i).a in (C.25), we have

$$\begin{aligned} (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \nabla \mathcal{L}(\boldsymbol{\beta}^*) &\geq -\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \|\nabla \mathcal{L}(\boldsymbol{\beta}^*)\|_\infty \\ &= -\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\overline{S^*}}\|_1 \|\nabla \mathcal{L}(\boldsymbol{\beta}^*)\|_\infty - \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_1 \|\nabla \mathcal{L}(\boldsymbol{\beta}^*)\|_\infty. \end{aligned} \tag{C.26}$$

  For term (i).b in (C.25), we have

$$(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \nabla \mathcal{Q}_\lambda(\boldsymbol{\beta}^*) = (\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}^T \big(\nabla \mathcal{Q}_\lambda(\boldsymbol{\beta}^*)\big)_{S^*} + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\overline{S^*}}^T \big(\nabla \mathcal{Q}_\lambda(\boldsymbol{\beta}^*)\big)_{\overline{S^*}}. \tag{C.27}$$

Note that $\mathcal{Q}_\lambda(\boldsymbol{\beta}^*)$ is separable. We have

$$(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}^T\big(\nabla\mathcal{Q}_\lambda(\boldsymbol{\beta}^*)\big)_{S^*}=\sum_{j\in S^*}(\beta_j-\beta_j^*)\cdot q'_\lambda(\beta_j^*)=(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}^T\nabla\mathcal{Q}_\lambda(\boldsymbol{\beta}^*), \tag{C.28}$$

$$(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}^T\big(\nabla\mathcal{Q}_\lambda(\boldsymbol{\beta}^*)\big)_{\overline{S^*}}=\sum_{j\in\overline{S^*}}(\beta_j-\beta_j^*)\cdot q'_\lambda(\beta_j^*)=\sum_{j\in\overline{S^*}}(\beta_j-\beta_j^*)\cdot q'_\lambda(0)=0, \tag{C.29}$$

where the second equation in (C.29) is because $\beta_j^*=0$ for $j\in\overline{S^*}$, and the third is by regularity condition (c) that $q'_\lambda(0)=0$. Plugging (C.28) and (C.29) into the right-hand side of (C.27), for term (i).b in (C.25) we obtain

$$(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\nabla\mathcal{Q}_\lambda(\boldsymbol{\beta}^*)=(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}^T\nabla\mathcal{Q}_\lambda(\boldsymbol{\beta}^*)\geq-\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1\|\nabla\mathcal{Q}_\lambda(\boldsymbol{\beta}^*)\|_\infty. \tag{C.30}$$

Plugging (C.26) and (C.30) into the right-hand side of (C.25), then for term (i) in (C.24) we obtain

$$\begin{aligned}&(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) \\ &\geq\ -\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty-\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty-\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1\|\nabla\mathcal{Q}_\lambda(\boldsymbol{\beta}^*)\|_\infty.\end{aligned} \tag{C.31}$$

- **Bounding Term (ii) in (C.24):** For term (ii) in (C.24), by separating the support of $\boldsymbol{\beta}-\boldsymbol{\beta}^*$ into S^* and $\overline{S^*}$ we have

$$\lambda(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\boldsymbol{\xi}=\lambda\underbrace{(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}^T\boldsymbol{\xi}_{S^*}}_{\text{(ii).a}}+\lambda\underbrace{(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}^T\boldsymbol{\xi}_{\overline{S^*}}}_{\text{(ii).b}}. \tag{C.32}$$

For term (ii).a in (C.32), since $\boldsymbol{\xi}\in\partial\|\boldsymbol{\beta}\|_1$, we have $\|\boldsymbol{\xi}_{S^*}\|_\infty\leq\|\boldsymbol{\xi}\|_\infty\leq1$, which implies

$$(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}^T\boldsymbol{\xi}_{S^*}\geq-\|\boldsymbol{\xi}_{S^*}\|_\infty\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1\geq-\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1. \tag{C.33}$$

For term (ii).b in (C.32), since $\boldsymbol{\beta}^*_{\overline{S^*}}=\mathbf{0}$, we have $(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}=\boldsymbol{\beta}_{\overline{S^*}}$. Recall that $\boldsymbol{\xi}\in\partial\|\boldsymbol{\beta}\|_1$. For $\beta_j\neq0$, since $\xi_j=\operatorname{sign}(\beta_j)$, we have $\beta_j\xi_j=|\beta_j|$. For $\beta_j=0$, we have $\beta_j\xi_j=|\beta_j|=0$. Therefore, we obtain

$$(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}^T\boldsymbol{\xi}_{\overline{S^*}}=\boldsymbol{\beta}_{\overline{S^*}}^T\boldsymbol{\xi}_{\overline{S^*}}=\sum_{j\in\overline{S^*}}\beta_j\xi_j=\sum_{j\in\overline{S^*}}|\beta_j|=\|\boldsymbol{\beta}_{\overline{S^*}}\|_1=\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1. \tag{C.34}$$

Plugging (C.33) and (C.34) into the right-hand side of (C.32), we obtain

$$\lambda(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\boldsymbol{\xi}\geq-\lambda\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1+\lambda\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1. \tag{C.35}$$

Plugging (C.31) and (C.35) into the right-hand side of (C.24), we obtain

$$\begin{aligned}&\lambda/2\cdot\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1 \\ &\geq\ \underbrace{-\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty-\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty-\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1\|\nabla\mathcal{Q}_\lambda(\boldsymbol{\beta}^*)\|_\infty}_{\text{(i) in (C.24)}} \\ &\quad+(\rho_--\zeta_-)\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_2^2\underbrace{-\lambda\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1+\lambda\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1}_{\text{(ii) in (C.24)}}.\end{aligned} \tag{C.36}$$

Again, we separate the left-hand side of (C.36) as $\lambda/2 \cdot \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 = \lambda/2 \cdot \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\overline{S^*}}\|_1 + \lambda/2 \cdot \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_1$. Rearranging the terms, we obtain

$$\begin{aligned}(\rho_- - \zeta_-)\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 + \underbrace{\big(\lambda/2 - \|\nabla \mathcal{L}(\boldsymbol{\beta}^*)\|_\infty\big)\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\overline{S^*}}\|_1}_{\text{(i)}} \\ \leq \big(3\lambda/2 + \underbrace{\|\nabla \mathcal{L}(\boldsymbol{\beta}^*)\|_\infty}_{\text{(ii)}} + \underbrace{\|\nabla \mathcal{Q}_\lambda(\boldsymbol{\beta}^*)\|_\infty}_{\text{(iii)}}\big)\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_1.\end{aligned} \tag{C.37}$$

For term (ii) in (C.37), by (4.1) in Assumption 4.1 and $\lambda \geq \lambda_{\text{tgt}}$ we have

$$\|\nabla \mathcal{L}(\boldsymbol{\beta}^*)\|_\infty \leq \lambda_{\text{tgt}}/8 \leq \lambda/8. \tag{C.38}$$

Meanwhile, (C.38) also implies that term (i) in (C.37) is positive. Recall that $\mathcal{Q}_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^d q_\lambda(\beta_j)$, where $q_\lambda(\beta_j)$ satisfies regularity condition (d). Hence for term (iii) in (C.37) we have

$$\|\nabla \mathcal{Q}_\lambda(\boldsymbol{\beta}^*)\|_\infty = \max_{1 \leq j \leq d} |q'_\lambda(\beta_j^*)| \leq \lambda. \tag{C.39}$$

In summary, from (C.37) we obtain

$$\begin{aligned}(\rho_- - \zeta_-)\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 &\leq \big(3\lambda/2 + \|\nabla \mathcal{L}(\boldsymbol{\beta}^*)\|_\infty + \|\nabla \mathcal{Q}_\lambda(\boldsymbol{\beta}^*)\|_\infty\big)\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_1 \\ &\leq (3\lambda/2 + \lambda/8 + \lambda)\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_1 \\ &\leq 21\lambda/8 \cdot \sqrt{s^*}\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_2 \\ &\leq 21\lambda/8 \cdot \sqrt{s^*}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2.\end{aligned} \tag{C.40}$$

According to (4.5), we have $\rho_- - \zeta_- > 0$. Therefore, (C.40) gives

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq \frac{21/8}{\rho_- - \zeta_-}\lambda\sqrt{s^*}, \tag{C.41}$$

which implies the first conclusion.

Results for the Objective Function Value: Note that on the right-hand side of (C.19), we have $\rho_- - \zeta_- > 0$, which gives

$$\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) \geq \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + (\boldsymbol{\beta}^* - \boldsymbol{\beta})^T \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}). \tag{C.42}$$

Meanwhile, since $\boldsymbol{\xi} \in \partial\|\boldsymbol{\beta}\|_1$, by the convexity of ℓ_1 norm we have

$$\lambda\|\boldsymbol{\beta}^*\|_1 \geq \lambda\|\boldsymbol{\beta}\|_1 + \lambda(\boldsymbol{\beta}^* - \boldsymbol{\beta})^T\boldsymbol{\xi}. \tag{C.43}$$

Recall that $\phi_\lambda(\boldsymbol{\beta}) = \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1$. Adding (C.42) and (C.43), we obtain

$$\phi_\lambda(\boldsymbol{\beta}^*) \geq \phi_\lambda(\boldsymbol{\beta}) + (\boldsymbol{\beta}^* - \boldsymbol{\beta})^T\big(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda\boldsymbol{\xi}\big), \tag{C.44}$$

which implies

$$\phi_\lambda(\boldsymbol{\beta}) - \phi_\lambda(\boldsymbol{\beta}^*) \leq (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T\big(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda\boldsymbol{\xi}\big) \leq \lambda/2 \cdot \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1.$$

Here the second inequality follows from (C.23), which is a direct consequence of the assumption that $\omega_\lambda(\boldsymbol{\beta}) \leq \lambda/2$. Separating the support of $\boldsymbol{\beta} - \boldsymbol{\beta}^*$ into S^* and $\overline{S^*}$, we obtain

$$\phi_\lambda(\boldsymbol{\beta}) - \phi_\lambda(\boldsymbol{\beta}^*) \leq \lambda/2 \cdot \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \leq \lambda/2 \cdot \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_1 + \lambda/2 \cdot \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\overline{S^*}}\|_1. \tag{C.45}$$

Now we provide an upper bound of $\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\overline{S^*}}\|_1$ on the right-hand side of (C.45). Note that, on the left-hand side of (C.37), we have $\rho_- - \zeta_- > 0$, which gives

$$\begin{aligned} &\big(\lambda/2 - \|\nabla \mathcal{L}(\boldsymbol{\beta}^*)\|_\infty\big)\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\overline{S^*}}\|_1 \\ &\quad \leq \quad \big(3\lambda/2 + \|\nabla \mathcal{L}(\boldsymbol{\beta}^*)\|_\infty + \|\nabla \mathcal{Q}_\lambda(\boldsymbol{\beta}^*)\|_\infty\big)\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_1. \end{aligned} \tag{C.46}$$

Note that in (C.47) we have $\|\nabla \mathcal{L}(\boldsymbol{\beta}^*)\|_\infty \leq \lambda/8$ by (C.38), and $\|\nabla \mathcal{Q}_\lambda(\boldsymbol{\beta}^*)\|_\infty \leq \lambda$ by (C.39). Hence we have

$$(\lambda/2 - \lambda/8)\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\overline{S^*}}\|_1 \leq (3\lambda/2 + \lambda/8 + \lambda)\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_1, \tag{C.47}$$

which implies $\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\overline{S^*}}\|_1 \leq 7\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_1$. Plugging this into the right-hand side of (C.45), we obtain

$$\phi_\lambda(\boldsymbol{\beta}) - \phi_\lambda(\boldsymbol{\beta}^*) \leq (\lambda/2 + 7\lambda/2)\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_1 \leq 4\lambda\sqrt{s^*}\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_2 \leq 4\lambda\sqrt{s^*}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2. \tag{C.48}$$

Plugging the upper bound of $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2$ in (C.41) into the right-hand side of (C.48), we obtain

$$\phi_\lambda(\boldsymbol{\beta}) - \phi_\lambda(\boldsymbol{\beta}^*) \leq \frac{21/2}{\rho_- - \zeta_-}\lambda^2 s^*.$$

Hence we reach the second conclusion. □

C.6 Proof of Lemma 5.3

Proof. Since $\|\boldsymbol{\beta}_{\overline{S^*}}\|_0 \leq \widetilde{s}$ and $\|\boldsymbol{\beta}^*_{\overline{S^*}}\|_0 = 0$, we have $\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\overline{S^*}}\|_0 \leq \widetilde{s}$. For logistic loss, we further have $\|\boldsymbol{\beta}\|_2 \leq R$ and $\|\boldsymbol{\beta}^*\|_2 \leq R$, where R is specified in Definition 4.3. Therefore, Lemma 5.1 gives

$$\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) + \frac{\rho_- - \zeta_-}{2}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|_2^2 \leq \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}). \tag{C.49}$$

Recall that $\phi_\lambda(\boldsymbol{\beta}) = \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1$. Hence, from our assumption that

$$\phi_\lambda(\boldsymbol{\beta}) - \phi_\lambda(\boldsymbol{\beta}^*) \leq \frac{21/2}{\rho_- - \zeta_-}\lambda^2 s^*$$

we obtain

$$\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) - \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) + \lambda(\|\boldsymbol{\beta}\|_1 - \|\boldsymbol{\beta}^*\|_1) \leq \frac{21/2}{\rho_- - \zeta_-}\lambda^2 s^*. \tag{C.50}$$

Plugging (C.49) into the left-hand side of (C.50), we have

$$(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) + \frac{\rho_- - \zeta_-}{2}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|_2^2 + \lambda(\|\boldsymbol{\beta}\|_1 - \|\boldsymbol{\beta}^*\|_1) \leq \frac{21/2}{\rho_- - \zeta_-}\lambda^2 s^*.$$

Moving $(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*)+\lambda(\|\boldsymbol{\beta}\|_1-\|\boldsymbol{\beta}^*\|_1)$ to its right-hand side, we obtain

$$\frac{\rho_- - \zeta_-}{2}\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_2^2 \le \frac{21/2}{\rho_- - \zeta_-}\lambda^2 s^* \underbrace{-(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*)}_{\text{(i)}} + \lambda\underbrace{\big(\|\boldsymbol{\beta}^*\|_1-\|\boldsymbol{\beta}\|_1\big)}_{\text{(ii)}}. \tag{C.51}$$

For term (i) in (C.51), following the same way we obtain the lower bound of term (i) in (C.24) (in the proof of Lemma 5.2), we can obtain the same result as in (C.31), which implies

$$\begin{aligned} &-(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) \\ &\le \|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty + \|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty + \|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1\|\nabla\mathcal{Q}_\lambda(\boldsymbol{\beta}^*)\|_\infty. \end{aligned} \tag{C.52}$$

For term (ii) in (C.51), separating the support of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ into S^* and $\overline{S^*}$ respectively, we obtain

$$\|\boldsymbol{\beta}^*\|_1-\|\boldsymbol{\beta}\|_1 = \|\boldsymbol{\beta}^*_{S^*}\|_1+\|\boldsymbol{\beta}^*_{\overline{S^*}}\|_1-\big(\|\boldsymbol{\beta}_{S^*}\|_1+\|\boldsymbol{\beta}_{\overline{S^*}}\|_1\big). \tag{C.53}$$

Note that $\boldsymbol{\beta}^*_{\overline{S^*}}=\mathbf{0}$, which gives $\boldsymbol{\beta}_{\overline{S^*}}=\boldsymbol{\beta}_{\overline{S^*}}-\boldsymbol{\beta}^*_{\overline{S^*}}=(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}$. Hence, from (C.53) we have

$$\|\boldsymbol{\beta}^*\|_1-\|\boldsymbol{\beta}\|_1 = \|\boldsymbol{\beta}^*_{S^*}\|_1-\|\boldsymbol{\beta}_{S^*}\|_1-\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1 \le \|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1-\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1, \tag{C.54}$$

where the inequality follows from the triangle inequality. Plugging (C.52) and (C.54) into the right-hand side of (C.51), we obtain

$$\begin{aligned} &\frac{\rho_- - \zeta_-}{2}\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_2^2 \\ &\le \underbrace{\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty + \|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty + \|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1\|\nabla\mathcal{Q}_\lambda(\boldsymbol{\beta}^*)\|_\infty}_{\text{(i) in (C.51)}} \\ &\quad + \lambda\underbrace{\big(\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1-\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1\big)}_{\text{(ii) in (C.51)}} + \frac{21/2}{\rho_- - \zeta_-}\lambda^2 s^*. \end{aligned} \tag{C.55}$$

Rearranging the terms in (C.55), we obtain

$$\begin{aligned} &\frac{\rho_- - \zeta_-}{2}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2 + \underbrace{\big(\lambda-\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty\big)\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1}_{\text{(i)}} \\ &\le \big(\lambda+\underbrace{\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty}_{\text{(ii)}}+\underbrace{\|\nabla\mathcal{Q}_\lambda(\boldsymbol{\beta}^*)\|_\infty}_{\text{(iii)}}\big)\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1 + \frac{21/2}{\rho_- - \zeta_-}\lambda^2 s^*. \end{aligned} \tag{C.56}$$

By (4.1) in Assumption 4.1 and $\lambda \ge \lambda_{\text{tgt}}$, for term (ii) in (C.56), we have

$$\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty \le \lambda_{\text{tgt}}/8 \le \lambda/8. \tag{C.57}$$

Moreover, (C.57) implies that term (i) in (C.56) is positive. For term (iii) in (C.56), since $\mathcal{Q}_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^d q_\lambda(\beta_j)$, where $q_\lambda(\beta_j)$ satisfies regularity condition (d), we have

$$\|\nabla\mathcal{Q}_\lambda(\boldsymbol{\beta}^*)\|_\infty \le \max_{1\le j\le d}|q'_\lambda(\beta^*_j)| \le \lambda. \tag{C.58}$$

Therefore, from (C.58) we obtain

$$\begin{aligned}\frac{\rho_- - \zeta_-}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 &\leq \big(\lambda + \|\nabla \mathcal{L}(\boldsymbol{\beta}^*)\|_\infty + \|\nabla \mathcal{Q}_\lambda(\boldsymbol{\beta}^*)\|_\infty\big)\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_1 + \frac{21/2}{\rho_- - \zeta_-}\lambda^2 s^* \\ &\leq (\lambda + \lambda/8 + \lambda)\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_1 + \frac{21/2}{\rho_- - \zeta_-}\lambda^2 s^* \\ &\leq 17/8 \cdot \lambda\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_1 + \frac{21/2}{\rho_- - \zeta_-}\lambda^2 s^*. \end{aligned} \tag{C.59}$$

To further provide an upper bound of the right-hand side of (C.59), we discuss two cases regarding the relationship between $\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_1$ and λs^*.

- If $7/(\rho_- - \zeta_-) \cdot \lambda s^* < \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_1$, then we have

  $$\frac{21/2}{\rho_- - \zeta_-}\lambda^2 s^* < 3/2 \cdot \lambda\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_1.$$

  Plugging this into the right-hand side of (C.59), we obtain

  $$\begin{aligned}\frac{\rho_- - \zeta_-}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 &\leq (17/8 \cdot \lambda + 3/2 \cdot \lambda)\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_1 \\ &\leq 29/8 \cdot \lambda\sqrt{s^*}\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_2 \\ &\leq 29/8 \cdot \lambda\sqrt{s^*}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2.\end{aligned}$$

  Dividing $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|_2$ on both sides, we have

  $$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq \frac{29/4}{\rho_- - \zeta_-}\lambda\sqrt{s^*}. \tag{C.60}$$

- If $\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_1 \leq 7/(\rho_- - \zeta_-) \cdot \lambda s^*$, then we have

  $$17/8 \cdot \lambda\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_1 < \frac{119/8}{\rho_- - \zeta_-}\lambda^2 s^*.$$

  Plugging this into the right-hand side of (C.59), we obtain

  $$\frac{\rho_- - \zeta_-}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \leq \frac{119/8}{\rho_- - \zeta_-}\lambda^2 s^* + \frac{21/2}{\rho_- - \zeta_-}\lambda^2 s^* = \frac{203/8}{\rho_- - \zeta_-}\lambda^2 s^*, \tag{C.61}$$

  which implies

  $$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq \frac{\sqrt{203}/2}{\rho_- - \zeta_-}\lambda\sqrt{s^*}. \tag{C.62}$$

Combining (C.60) and (C.62), since $\max\{29/4, \sqrt{203}/2\} \leq 15/2$, we obtain

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 < \frac{15/2}{\rho_- - \zeta_-}\lambda\sqrt{s^*}.$$

Hence we conclude the proof. □

C.7 Proof of Lemma 5.4

Proof. Recall that the proximal-gradient update step defined in (3.8) with $\Omega = \mathbb{R}^d$, i.e., $R = +\infty$, takes the form

$$\big(\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}; +\infty)\big)_j = \begin{cases} 0 & \text{if } |\bar{\beta}_j| \le \lambda/L, \\ \operatorname{sign}(\bar{\beta}_j)\big(|\bar{\beta}_j| - \lambda/L\big) & \text{if } |\bar{\beta}_j| > \lambda/L, \end{cases} \tag{C.63}$$

for $j = 1, \ldots, d$, where

$$\bar{\boldsymbol{\beta}} = \boldsymbol{\beta} - \frac{1}{L}\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}), \tag{C.64}$$

and $\bar{\beta}_j$ is the j-th dimension of $\bar{\boldsymbol{\beta}}$. Furthermore, if $\Omega = B_2(R)$ of radius $R \in (0, \infty)$, $\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}; R)$ can be obtained by projecting $\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}; +\infty)$ shown in (C.63) onto $B_2(R)$, i.e.,

$$\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}; R) = \begin{cases} \mathcal{T}_{L,\lambda}(\boldsymbol{\beta}; +\infty) & \text{if } \|\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}; +\infty)\|_2 < R, \\ \dfrac{R \cdot \mathcal{T}_{L,\lambda}(\boldsymbol{\beta}; +\infty)}{\|\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}; +\infty)\|_2} & \text{if } \|\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}; +\infty)\|_2 \ge R. \end{cases} \tag{C.65}$$

Note that $\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}; +\infty)$ and $\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}; R)$ have exactly the same sparsity pattern. Hence we focus on analyzing the sparsity pattern of $\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}; +\infty)$ in the following.

In fact, update scheme (C.63) defines a soft-thresholding operation on $\bar{\boldsymbol{\beta}}$ defined in (C.64), with the threshold value λ/L. To show $\big\|\big(\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}; +\infty)\big)_{\overline{S^*}}\big\|_0 \le \widetilde{s}$, we need to prove that, for $j \in \overline{S^*}$, the number of j's such that $|\bar{\beta}_j| > \lambda/L$ is no more than $\widetilde{s}$. To achieve this goal, we first reformulate $\bar{\boldsymbol{\beta}}$ as

$$\bar{\boldsymbol{\beta}} = \boldsymbol{\beta} - \frac{1}{L}\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) = \boldsymbol{\beta} - \frac{1}{L}\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) + \frac{1}{L}\big(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) - \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})\big). \tag{C.66}$$

Then it suffices to prove there exist integers $\widetilde{s}_1$, $\widetilde{s}_2$ and $\widetilde{s}_3$, which satisfy $\widetilde{s}_1 + \widetilde{s}_2 + \widetilde{s}_3 \le \widetilde{s}$, such that

$$\big|\big\{j \in \overline{S^*} : |\beta_j| \ge 1/4 \cdot \lambda/L\big\}\big| \;\le\; \widetilde{s}_1, \tag{C.67}$$

$$\Big|\Big\{j \in \overline{S^*} : \big|\big(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*)/L\big)_j\big| > 1/8 \cdot \lambda/L\Big\}\Big| \;\le\; \widetilde{s}_2, \tag{C.68}$$

$$\Big|\Big\{j \in \overline{S^*} : \big|\big(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})/L - \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*)/L\big)_j\big| \ge 5/8 \cdot \lambda/L\Big\}\Big| \;\le\; \widetilde{s}_3. \tag{C.69}$$

This is because, if (C.67)−(C.69) hold, then there are at most $\widetilde{s}_1 + \widetilde{s}_2 + \widetilde{s}_3 \le \widetilde{s}$ coordinates $j \in \overline{S^*}$ such that

$$|\beta_j| + \big|\big(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*)/L\big)_j\big| + \big|\big(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})/L - \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*)/L\big)_j\big| > \lambda/L.$$

Since by the triangular inequality (C.66) implies

$$|\bar{\beta}_j| \le |\beta_j| + \big|\big(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*)/L\big)_j\big| + \big|\big(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})/L - \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*)/L\big)_j\big|,$$

the number of coordinates $j \in \overline{S^*}$ such that $|\bar{\beta}_j| > \lambda/L$ is also upper bounded by $\widetilde{s}_1 + \widetilde{s}_2 + \widetilde{s}_3 \le \widetilde{s}$. In the following, we will prove (C.68)−(C.69) and specify the corresponding $\widetilde{s}_1$, $\widetilde{s}_2$ and $\widetilde{s}_3$.

Proof of (C.67): Note that for $j \in \overline{S^*}$, we have $\beta_j^* = 0$. Hence we have

$$\big|\big\{j \in \overline{S^*} : |\beta_j| \ge 1/4 \cdot \lambda/L\big\}\big| = \big|\big\{j \in \overline{S^*} : |\beta_j - \beta_j^*| \ge 1/4 \cdot \lambda/L\big\}\big|. \tag{C.70}$$

Meanwhile, note that

$$\begin{aligned}\frac{\lambda}{4L}\big|\big\{j\in\overline{S^*}:|\beta_j-\beta_j^*|\geq 1/4\cdot\lambda/L\big\}\big| &\leq \sum_{j\in\overline{S^*}}|\beta_j-\beta_j^*|\cdot\mathbb{1}\big(|\beta_j-\beta_j^*|\geq 1/4\cdot\lambda/L\big)\\ &\leq \sum_{j\in\overline{S^*}}|\beta_j-\beta_j^*|\\ &= \|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1. \end{aligned}\tag{C.71}$$

Plugging (C.71) into the right-hand side of (C.70), we obtain

$$\big|\big\{j\in\overline{S^*}:|\beta_j|\geq 1/4\cdot\lambda/L\big\}\big|\leq\frac{4L}{\lambda}\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1.\tag{C.72}$$

Now we provide an upper bound of $\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1$. Following the same way we derive (C.56) in the proof of Lemma 5.3, we can obtain

$$\begin{aligned}&\frac{\rho_- -\zeta_-}{2}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2+\big(\lambda-\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty\big)\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1 \\ &\leq \big(\lambda+\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty+\|\nabla\mathcal{Q}_\lambda(\boldsymbol{\beta}^*)\|_\infty\big)\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1+\frac{21/2}{\rho_- -\zeta_-}\lambda^2 s^*.\end{aligned}\tag{C.73}$$

According to (4.5), we have $\rho_- -\zeta_- >0$. Hence (C.73) implies

$$\begin{aligned}&\big(\lambda-\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty\big)\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1 \\ &\leq \big(\lambda+\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty+\|\nabla\mathcal{Q}_\lambda(\boldsymbol{\beta}^*)\|_\infty\big)\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1+\frac{21/2}{\rho_- -\zeta_-}\lambda^2 s^*.\end{aligned}\tag{C.74}$$

By (4.1) in Assumption 4.1 and $\lambda\geq\lambda_{\mathrm{tgt}}$, we have

$$\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty\leq\lambda_{\mathrm{tgt}}/8\leq\lambda/8.\tag{C.75}$$

Meanwhile, since $\mathcal{Q}_\lambda(\boldsymbol{\beta})=\sum_{j=1}^d q_\lambda(\beta_j)$ and $q_\lambda(\beta_j)$ satisfies regularity condition (d), we have

$$\|\nabla\mathcal{Q}_\lambda(\boldsymbol{\beta}^*)\|_\infty=\max_{1\leq j\leq d}|q_\lambda'(\beta_j^*)|\leq\lambda.\tag{C.76}$$

Plugging (C.75) and (C.76) into (C.74) and dividing λ on both sides, we obtain

$$7/8\cdot\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1\leq 17/8\cdot\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1+\frac{21/2}{\rho_- -\zeta_-}\lambda s^*.\tag{C.77}$$

Now we discuss two cases regarding the relationship between $\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1$ and λs^*.

- If $7/(\rho_- -\zeta_-)\cdot\lambda s^*<\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1$, then we have

  $$\frac{21/2}{\rho_- -\zeta_-}\lambda s^*\leq 3/2\cdot\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1.$$

  Plugging this into the right-hand side of (C.77), we obtain $\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1\leq 29/7\cdot\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1$, which implies

  $$\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1\leq 29/7\cdot\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1\leq 29/7\cdot\sqrt{s^*}\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_2\leq 29/7\cdot\sqrt{s^*}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2.\tag{C.78}$$

Plugging the upper bound of $\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2$ in Lemma 5.3 into the right-hand side of (C.78), we obtain

$$\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1 \leq 29/7\cdot\sqrt{s^*}\cdot\frac{15/2}{\rho_- - \zeta_-}\lambda\sqrt{s^*} = \frac{435/14}{\rho_- - \zeta_-}\lambda s^*. \tag{C.79}$$

- If $\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1 \leq 7/(\rho_- - \zeta_-)\cdot\lambda s^*$, then plugging this into the right-hand side of (C.77), we obtain

$$\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1 \leq 8/7\cdot\left(\frac{17/8\cdot 7}{\rho_- - \zeta_-}\lambda s^* + \frac{21/2}{\rho_- - \zeta_-}\lambda s^*\right) \leq \frac{29}{\rho_- - \zeta_-}\lambda s^*. \tag{C.80}$$

Combining (C.79) and (C.80), we obtain

$$\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1 \leq \frac{\max\{435/14, 29\}}{\rho_- - \zeta_-}\lambda s^* \leq \frac{435/14}{\rho_- - \zeta_-}\lambda s^*.$$

Plugging this into the right-hand side of (C.72), we obtain

$$\big|\big\{j\in\overline{S^*} : |\beta_j| \geq 1/4\cdot\lambda/L\big\}\big| \leq \frac{4L}{\lambda}\cdot\frac{435/14}{\rho_- - \zeta_-}\lambda s^* < \frac{125L}{\rho_- - \zeta_-}s^*.$$

Meanwhile, since we assume $L < 2(\rho_+ - \zeta_+)$, we have

$$\big|\big\{j\in\overline{S^*} : |\beta_j| \geq 1/4\cdot\lambda/L\big\}\big| < 250\cdot\frac{\rho_+ - \zeta_+}{\rho_- - \zeta_-}\cdot s^* = 250\kappa s^*,$$

where the last equality follows from the definition of the condition number κ in (4.4). Therefore we obtain (C.67) by setting $\widetilde{s}_1 = 250\kappa s^*$.

Proof of (C.68)**:** Recall that $\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \mathcal{Q}_\lambda(\boldsymbol{\beta})$. Hence we have

$$\big\|\big(\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*)\big)_{\overline{S^*}}\big\|_\infty \leq \big\|\big(\nabla\mathcal{L}(\boldsymbol{\beta}^*)\big)_{\overline{S^*}}\big\|_\infty + \big\|\big(\nabla\mathcal{Q}_\lambda(\boldsymbol{\beta}^*)\big)_{\overline{S^*}}\big\|_\infty. \tag{C.81}$$

By (4.1) in Assumption 4.1, we have

$$\big\|\big(\nabla\mathcal{L}(\boldsymbol{\beta}^*)\big)_{\overline{S^*}}\big\|_\infty \leq \|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty \leq \lambda/8. \tag{C.82}$$

Recall $\mathcal{Q}_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^d q_\lambda(\beta_j)$, where $q_\lambda(\beta_j)$ satisfies regularity condition (c) that $q'_\lambda(0) = 0$. Hence we have

$$\big\|\big(\nabla\mathcal{Q}_\lambda(\boldsymbol{\beta}^*)\big)_{\overline{S^*}}\big\|_\infty = \max_{j\in\overline{S^*}}\big|q'_\lambda(\beta_j^*)\big| = \max_{j\in\overline{S^*}}\big|q'_\lambda(0)\big| = 0, \tag{C.83}$$

where the second equation follows from the fact that $\beta_j^* = 0$ for $j\in\overline{S^*}$. Plugging (C.83) and (C.82) into the right-hand side of (C.81), we obtain $\big\|\big(\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*)\big)_{\overline{S^*}}\big\|_\infty = \max_{j\in\overline{S^*}}\big|\big(\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*)/L\big)_j\big| \leq \lambda/8$. Hence we have

$$\Big|\Big\{j\in\overline{S^*} : \big|\big(\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*)/L\big)_j\big| > 1/8\cdot\lambda/L\Big\}\Big| = 0.$$

Therefore, by setting $\widetilde{s}_2 = 0$, we obtain (C.68).

Proof of (C.69)**:** Consider an arbitrary subset S' such that

$$S' \subseteq \left\{ j : \left| \left(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) - \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) \right)_j \right| \geq 5/8 \cdot \lambda \right\}. \tag{C.84}$$

Let $s' = |S'|$. In the following we provide an upper bound of s'. Suppose $\boldsymbol{v} \in \mathbb{R}^d$ is chosen such that $v_j = \operatorname{sign}\{ \left(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) - \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) \right)_j \}$ for $j \in S'$, and $v_j = 0$ for $j \notin S'$. Hence we have

$$\begin{aligned} \boldsymbol{v}^T \left(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) - \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) \right) &= \sum_{j \in S'} v_j \left(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) - \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) \right)_j \\ &= \sum_{j \in S'} \left| \left(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) - \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) \right)_j \right| \\ &\geq 5/8 \cdot \lambda s'. \end{aligned} \tag{C.85}$$

Meanwhile, by Cauchy Schwarz inequality we have

$$\boldsymbol{v}^T \left(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) - \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) \right) \leq \|\boldsymbol{v}\|_2 \left\| \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) - \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) \right\|_2 \leq \sqrt{s'} \left\| \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) - \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) \right\|_2, \tag{C.86}$$

where the last inequality follows from the fact that $\|\boldsymbol{v}\|_2 \leq \sqrt{s'} \|\boldsymbol{v}\|_\infty = \sqrt{s'}$, because $\boldsymbol{v}$ is chosen such that $\|\boldsymbol{v}\|_0 = s'$. Combining (C.85) and (C.86), we have

$$5/8 \cdot \lambda s' \leq \boldsymbol{v}^T \left(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) - \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) \right) \leq \sqrt{s'} \left\| \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) - \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) \right\|_2. \tag{C.87}$$

Since $\|\boldsymbol{\beta}_{\overline{S^*}}\|_0 \leq \widetilde{s}$ and $\|\boldsymbol{\beta}^*_{\overline{S^*}}\|_0 = 0$, we have $\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\overline{S^*}}\| \leq \widetilde{s}$. In the setting of logistic loss, we further have $\|\boldsymbol{\beta}\|_2 \leq R$ and $\|\boldsymbol{\beta}^*\|_2 \leq R$, where R is specified in Definition 4.3. Therefore, Lemma 5.1 implies that $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})$ is restricted strongly smooth. Hence we have

$$\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) \leq \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) + \frac{\rho_+ - \zeta_+}{2} \|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|_2^2. \tag{C.88}$$

According to Nesterov (2004, Theorem 2.1.9), the strong smoothness of $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})$ is equivalent to the Lipschitz continuity of its gradient, i.e.,

$$\left\| \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) - \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) \right\|_2 \leq (\rho_+ - \zeta_+) \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2. \tag{C.89}$$

Plugging (C.89) into the right-hand side of (C.87), we obtain

$$5/8 \cdot \lambda s' \leq (\rho_+ - \zeta_+) \cdot \sqrt{s'} \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2. \tag{C.90}$$

Plugging the upper bound of $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2$ in Lemma 5.3 into the right-hand side of (C.90), we obtain

$$\sqrt{s'} \leq \frac{8}{5\lambda} \cdot (\rho_+ - \zeta_+) \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq \frac{8}{5\lambda} \cdot (\rho_+ - \zeta_+) \cdot \frac{15/2}{\rho_- - \zeta_-} \lambda \sqrt{s^*} = 12\kappa \sqrt{s^*}, \tag{C.91}$$

where the last equality follows from the definition of the condition number κ in (4.4). Hence we obtain $s' \leq 144\kappa^2 s^*$. Note that S' is defined as an arbitrary subset of $\left\{ j : \left| \left(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) - \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) \right)_j \right| \geq 5/8 \cdot \lambda \right\}$ and

$$\left\{ j \in \overline{S^*} : \left| \left(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) - \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) \right)_j \right| \geq 5/8 \cdot \lambda \right\} \subseteq \left\{ j : \left| \left(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) - \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*) \right)_j \right| \geq 5/8 \cdot \lambda \right\}.$$

Hence we have

$$\left|\left\{j \in \overline{S^*} : \left|\left(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})/L - \nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}^*)/L\right)_j\right| \geq 5/8 \cdot \lambda/L\right\}\right| \leq 144\kappa^2 s^*.$$

Therefore, by setting $\widetilde{s}_3 = 144\kappa^2 s^*$, we obtain (C.69).

In summary, we prove that (C.68)−(C.69) hold with $\widetilde{s}_1 = 250\kappa s^*$, $\widetilde{s}_2 = 0$ and $\widetilde{s}_2 = 144\kappa^2 s^*$. In Assumption 4.4, we assume $\widetilde{s} \geq 144\kappa^2 + 250\kappa$, which implies $\widetilde{s}_1 + \widetilde{s}_2 + \widetilde{s}_3 \leq \widetilde{s}$. Therefore we have $\left\|\left(\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}; +\infty)\right)_{\overline{S^*}}\right\|_0 < \widetilde{s}$. Since $\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}; R)$ has the same sparsity pattern as $\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}; +\infty)$, we also have that $\left\|\left(\mathcal{T}_{L,\lambda}(\boldsymbol{\beta}; R)\right)_{\overline{S^*}}\right\|_0 < \widetilde{s}$ for $R \in (0, +\infty)$. Hence we conclude the proof. □

C.8 Proof of Theorem 5.5

We first provide a useful lemma. It states that if $\boldsymbol{\beta}$ is ϵ-suboptimal with respect to the regularization parameter λ and sufficiently sparse, then for $\lambda' \leq \lambda$ the objective function value $\phi_{\lambda'}(\boldsymbol{\beta})$ is close to $\phi_{\lambda'}\big(\widehat{\boldsymbol{\beta}}_{\lambda'}\big)$. Here $\widehat{\boldsymbol{\beta}}_{\lambda'}$ is the exact local solution corresponding to λ'.

Lemma C.7. Let $\lambda \geq \lambda_{\text{tgt}}$ and $\lambda' \in [\lambda_{\text{tgt}}, \lambda]$. Suppose $\|\boldsymbol{\beta}_{\overline{S^*}}\|_0 \leq \widetilde{s}$ and $\omega_\lambda(\boldsymbol{\beta}) \leq \epsilon$. Let $\widehat{\boldsymbol{\beta}}_{\lambda'}$ be the exact local solution corresponding to λ', which satisfies the exact optimality condition in (3.14) and $\left\|\left(\widehat{\boldsymbol{\beta}}_{\lambda'}\right)_{\overline{S^*}}\right\|_0 \leq \widetilde{s}$. For logistic loss, we further assume $\max\left\{\|\boldsymbol{\beta}\|_2, \left\|\widehat{\boldsymbol{\beta}}_{\lambda'}\right\|_2\right\} \leq R$, where R is specified in Definition 4.3. Under Assumption 4.1 and Assumption 4.4, we have

$$\phi_{\lambda'}(\boldsymbol{\beta}) - \phi_{\lambda'}\big(\widehat{\boldsymbol{\beta}}_{\lambda'}\big) \leq C\big(\epsilon + 2(\lambda - \lambda')\big) \cdot (\lambda' + \lambda)s^*, \quad \text{where} \quad C = \frac{21}{\rho_- - \zeta_-}.$$

Proof. Since $\|\boldsymbol{\beta}_{\overline{S^*}}\|_0 \leq \widetilde{s}$ and $\left\|\left(\widehat{\boldsymbol{\beta}}_{\lambda'}\right)_{\overline{S^*}}\right\|_0 \leq \widetilde{s}$, we have $\left\|\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\lambda'}\right)_{\overline{S^*}}\right\| \leq 2\widetilde{s}$. In the setting of logistic loss, we further have $\|\boldsymbol{\beta}\|_2 \leq R$ and $\left\|\widehat{\boldsymbol{\beta}}_{\lambda'}\right\|_2 \leq R$. Therefore, Lemma 5.1 gives

$$\widetilde{\mathcal{L}}_{\lambda'}\big(\widehat{\boldsymbol{\beta}}_{\lambda'}\big) \geq \widetilde{\mathcal{L}}_{\lambda'}(\boldsymbol{\beta}) + \big(\widehat{\boldsymbol{\beta}}_{\lambda'} - \boldsymbol{\beta}\big)^T \nabla \widetilde{\mathcal{L}}_{\lambda'}(\boldsymbol{\beta}) + \frac{\rho_- - \zeta_-}{2}\left\|\widehat{\boldsymbol{\beta}}_{\lambda'} - \boldsymbol{\beta}\right\|_2^2 \geq \widetilde{\mathcal{L}}_{\lambda'}(\boldsymbol{\beta}) + \big(\widehat{\boldsymbol{\beta}}_{\lambda'} - \boldsymbol{\beta}\big)^T \nabla \widetilde{\mathcal{L}}_{\lambda'}(\boldsymbol{\beta}), \tag{C.92}$$

where the second inequality is because $\rho_- - \zeta_- > 0$, which follows from (4.5).

Let $\boldsymbol{\xi} \in \partial\|\boldsymbol{\beta}\|_1$ be the subgradient that attains the minimum in

$$\omega_\lambda(\boldsymbol{\beta}) = \min_{\boldsymbol{\xi}' \in \partial\|\boldsymbol{\beta}\|_1} \max_{\boldsymbol{\beta}' \in \Omega} \left\{\frac{(\boldsymbol{\beta} - \boldsymbol{\beta}')^T}{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_1}\big(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda\boldsymbol{\xi}'\big)\right\}, \tag{C.93}$$

where $\Omega = B_2(R)$ in the setting of logistic loss and $\Omega = \mathbb{R}^d$ in other settings. Since $\boldsymbol{\xi}$ is a minimizer, we have

$$\omega_\lambda(\boldsymbol{\beta}) = \max_{\boldsymbol{\beta}' \in \Omega} \left\{\frac{(\boldsymbol{\beta} - \boldsymbol{\beta}')^T}{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_1}\big(\nabla \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda\boldsymbol{\xi}\big)\right\}. \tag{C.94}$$

By the convexity of ℓ_1 norm, we also have

$$\lambda'\left\|\widehat{\boldsymbol{\beta}}_{\lambda'}\right\|_1 \geq \lambda'\|\boldsymbol{\beta}\|_1 + \lambda'\boldsymbol{\xi}^T\big(\widehat{\boldsymbol{\beta}}_{\lambda'} - \boldsymbol{\beta}\big). \tag{C.95}$$

Recall that the objective function $\phi_\lambda(\boldsymbol{\beta})$ is defined as $\phi_\lambda(\boldsymbol{\beta}) = \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1$. Adding (C.92) and (C.95), we obtain

$$\phi_{\lambda'}\big(\widehat{\boldsymbol{\beta}}_{\lambda'}\big) \geq \phi_{\lambda'}(\boldsymbol{\beta}) + \big(\nabla \widetilde{\mathcal{L}}_{\lambda'}(\boldsymbol{\beta}) + \lambda'\boldsymbol{\xi}\big)^T\big(\widehat{\boldsymbol{\beta}}_{\lambda'} - \boldsymbol{\beta}\big).$$

Hence we have

$$\begin{aligned}\phi_{\lambda'}(\boldsymbol{\beta})-\phi_{\lambda'}(\widehat{\boldsymbol{\beta}}_{\lambda'}) &\le \big(\nabla\widetilde{\mathcal{L}}_{\lambda'}(\boldsymbol{\beta})+\lambda'\boldsymbol{\xi}\big)^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\lambda'})\\ &= \Big(\underbrace{\big(\nabla\mathcal{L}(\boldsymbol{\beta})+\nabla\mathcal{Q}_\lambda(\boldsymbol{\beta})\big)}_{\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})}+\lambda\boldsymbol{\xi}\big)+\big(\nabla\mathcal{Q}_{\lambda'}(\boldsymbol{\beta})-\nabla\mathcal{Q}_\lambda(\boldsymbol{\beta})\big)+\big(\lambda'\boldsymbol{\xi}-\lambda\boldsymbol{\xi}\big)\Big)^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\lambda'})\\ &\le \underbrace{\big(\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})+\lambda\boldsymbol{\xi}\big)^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\lambda'})}_{\text{(i)}}+\underbrace{\big\|\nabla\mathcal{Q}_{\lambda'}(\boldsymbol{\beta})-\nabla\mathcal{Q}_\lambda(\boldsymbol{\beta})\big\|_\infty}_{\text{(ii)}}\underbrace{\big\|\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\lambda'}\big\|_1}_{\text{(iv)}} \qquad \text{(C.96)}\\ &\quad+\underbrace{\|\lambda'\boldsymbol{\xi}-\lambda\boldsymbol{\xi}\|_\infty}_{\text{(iii)}}\underbrace{\big\|\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\lambda'}\big\|_1}_{\text{(iv)}}.\end{aligned}$$

Now we provide upper bounds of terms (i)−(iv) correspondingly.

Bounding Term (i) in (C.96): According to (C.94), we have

$$\frac{(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\lambda'})^T}{\big\|\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\lambda'}\big\|_1}\big(\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})+\lambda\boldsymbol{\xi}\big)\le\max_{\boldsymbol{\beta}'\in\Omega}\left\{\frac{(\boldsymbol{\beta}-\boldsymbol{\beta}')^T}{\|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_1}\big(\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})+\lambda\boldsymbol{\xi}\big)\right\}=\omega_\lambda(\boldsymbol{\beta})\le\epsilon,$$

where the last inequality is our assumption. Therefore we obtain

$$\big(\nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})+\lambda\boldsymbol{\xi}\big)^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\lambda'})\le\epsilon\cdot\big\|\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\lambda'}\big\|_1. \qquad \text{(C.97)}$$

We will provide an upper bound of $\big\|\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\lambda'}\big\|_1$ when we handle term (iv).

Bounding Term (ii) in (C.96): Recall that $\mathcal{Q}_\lambda(\boldsymbol{\beta})=\sum_{i=1}^d q_\lambda(\beta_j)$. We have

$$\big\|\nabla\mathcal{Q}_{\lambda'}(\boldsymbol{\beta})-\nabla\mathcal{Q}_\lambda(\boldsymbol{\beta})\big\|_\infty=\max_{1\le j\le d}\big|q_{\lambda'}(\beta_j)-q_\lambda(\beta_j)\big|\le\max_{1\le j\le d}|\lambda'-\lambda|=\lambda-\lambda', \qquad \text{(C.98)}$$

where the inequality follows from regularity condition (e), and the last equality is because $\lambda\ge\lambda'$.

Bounding Term (iii) in (C.96): Since $\boldsymbol{\xi}\in\partial\|\boldsymbol{\beta}\|_1$, we have $\|\boldsymbol{\xi}\|_\infty\le 1$. Then we obtain

$$\|\lambda'\boldsymbol{\xi}-\lambda\boldsymbol{\xi}\|_\infty=|\lambda'-\lambda|\|\boldsymbol{\xi}\|_\infty\le|\lambda-\lambda'|=\lambda-\lambda'. \qquad \text{(C.99)}$$

Bounding Term (iv) in (C.96): Note that

$$\big\|\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{\lambda'}\big\|_1\le\underbrace{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1}_{\text{(iv).a}}+\underbrace{\big\|\widehat{\boldsymbol{\beta}}_{\lambda'}-\boldsymbol{\beta}^*\big\|_1}_{\text{(iv).b}}. \qquad \text{(C.100)}$$

For term (iv).a, since $\boldsymbol{\beta}$ satisfies $\|\boldsymbol{\beta}_{\overline{S^*}}\|_0\le\widetilde{s}$, $\omega_\lambda(\boldsymbol{\beta})\le\lambda/2$, and $\|\boldsymbol{\beta}\|_2\le R$ for logistic loss, we have that $\boldsymbol{\beta}$ satisfies the assumptions of Lemma 5.2. Following the same way we obtain (C.47) in the proof of Lemma 5.2, we can get

$$(\lambda/2-\lambda/8)\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1\le(3\lambda/2+\lambda/8+\lambda)\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1,$$

which implies $\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1\le 7\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1$. Hence we obtain

$$\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1\le\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1+\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1\le 8\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_1\le 8\sqrt{s^*}\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_2\le 8\sqrt{s^*}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2.$$

Plugging in the upper bound of $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2$ in Lemma 5.2, we obtain

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \leq \frac{21}{\rho_- - \zeta_-}\lambda s^*. \tag{C.101}$$

Meanwhile, for term (iv).b, note that we assume $\widehat{\boldsymbol{\beta}}_{\lambda'}$ satisfies $\big\|(\widehat{\boldsymbol{\beta}}_{\lambda'})_{\overline{S^*}}\big\|_0 \leq \widetilde{s}$ and $\big\|\widehat{\boldsymbol{\beta}}_{\lambda'}\big\|_2 \leq R$ for logistic loss. Moreover, since $\widehat{\boldsymbol{\beta}}_{\lambda'}$ is an exact local solution, it satisfies the exact optimality condition $\omega\big(\widehat{\boldsymbol{\beta}}_{\lambda'}\big) \leq 0$, which implies $\omega\big(\widehat{\boldsymbol{\beta}}_{\lambda'}\big) < \lambda'/2$. Hence $\widehat{\boldsymbol{\beta}}_{\lambda'}$ also satisfies the conditions of Lemma 5.2. Similar to (C.101), we have

$$\big\|\widehat{\boldsymbol{\beta}}_{\lambda'} - \boldsymbol{\beta}^*\big\|_1 \leq \frac{21}{\rho_- - \zeta_-}\lambda' s^*. \tag{C.102}$$

Plugging (C.102) and (C.101) into (C.100), for term (iv) in (C.96), we obtain

$$\big\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\lambda'}\big\|_1 \leq \frac{21}{\rho_- - \zeta_-}(\lambda' + \lambda)s^*. \tag{C.103}$$

Plugging (C.97)−(C.99) and (C.103) into the right-hand side of (C.96), we obtain

$$\begin{aligned}
&\phi_{\lambda'}(\boldsymbol{\beta}) - \phi_{\lambda'}\big(\widehat{\boldsymbol{\beta}}_{\lambda'}\big) \\
&\leq \underbrace{\epsilon \cdot \frac{21}{\rho_- - \zeta_-}(\lambda' + \lambda)s^*}_{\text{(i) in (C.96)}} + \Big(\underbrace{(\lambda - \lambda')}_{\text{(ii) in (C.96)}} + \underbrace{(\lambda - \lambda')}_{\text{(iii) in (C.96)}} \Big) \cdot \underbrace{\frac{21}{\rho_- - \zeta_-}(\lambda' + \lambda)s^*}_{\text{(iv) in (C.96)}} \\
&\leq \frac{21}{\rho_- - \zeta_-}\big(\epsilon + 2(\lambda - \lambda')\big) \cdot (\lambda' + \lambda)s^*,
\end{aligned}$$

where the upper bound of term (i) in (C.96) is obtained by plugging (C.103) into the right-hand side of (C.97). Hence we conclude the proof. □

Now we are ready to prove Theorem 5.5.

Proof. **Sparsity of $\{\boldsymbol{\beta}_t^{(k)}\}_{k=0}^{\infty}$ within the t-th Stage:** In the following, we provide results concerning the sparsity of the sequence $\{\boldsymbol{\beta}_t^{(k)}\}_{k=0}^{\infty}$ within the t-th path following stage. In the following we prove this by induction. Note that the initialization satisfies

$$\big\|(\boldsymbol{\beta}_t^{(0)})_{\overline{S^*}}\big\|_0 \leq \widetilde{s}, \quad \omega_{\lambda_t}\big(\boldsymbol{\beta}_t^{(0)}\big) \leq \lambda_t/2, \quad \text{and} \quad L_t^{(0)} \leq 2(\rho_+ - \zeta_+). \tag{C.104}$$

By Lemma 5.2 we have

$$\phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(0)}\big) - \phi_{\lambda_t}(\boldsymbol{\beta}^*) \leq \frac{21/2}{\rho_- - \zeta_-}\lambda_t^2 s^*, \tag{C.105}$$

Suppose that, at the $(k-1)$-th iteration of the proximal-gradient method (Lines 5−9 of Algorithm 3), we have

$$\big\|(\boldsymbol{\beta}_t^{(k-1)})_{\overline{S^*}}\big\|_0 \leq \widetilde{s}, \quad L_t^{(k-1)} \leq 2(\rho_+ - \zeta_+), \quad \text{and } \phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}\big) - \phi_{\lambda_t}(\boldsymbol{\beta}^*) \leq \frac{21/2}{\rho_- - \zeta_-}\lambda_t^2 s^*, \tag{C.106}$$

Then according to Lemma 5.4, we have that $\boldsymbol{\beta}_t^{(k)} = \mathcal{T}_{L_t^{(k)},\lambda_t}(\boldsymbol{\beta}_t^{(k-1)}; R)$ satisfies

$$\big\|(\boldsymbol{\beta}_t^{(k)})_{\overline{S^*}}\big\|_0 \le \widetilde{s}. \tag{C.107}$$

Note that, in the setting of logistic loss, we always have $\big\|\boldsymbol{\beta}_t^{(k)}\big\|_2 \le R$ for $k = 0, 1, \ldots$ because of the ℓ_2 constraint $\Omega = B_2(R)$. Since $\big\|(\boldsymbol{\beta}_t^{(k-1)})_{\overline{S^*}}\big\|_0 \le \widetilde{s}$ and $\big\|(\boldsymbol{\beta}_t^{(k)})_{\overline{S^*}}\big\|_0 \le \widetilde{s}$ imply $\big\|(\boldsymbol{\beta}_t^{(k-1)} - \boldsymbol{\beta}_t^{(k)})_{\overline{S^*}}\big\| \le 2\widetilde{s}$, by Lemma 5.1 we have

$$\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k)}) \ge \widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)}) + \nabla\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)})^T(\boldsymbol{\beta}_t^{(k)} - \boldsymbol{\beta}_t^{(k-1)}) + \frac{\rho_- - \zeta_-}{2}\big\|\boldsymbol{\beta}_t^{(k)} - \boldsymbol{\beta}_t^{(k-1)}\big\|_2^2, \tag{C.108}$$

$$\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k)}) \le \widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)}) + \nabla\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)})^T(\boldsymbol{\beta}_t^{(k)} - \boldsymbol{\beta}_t^{(k-1)}) + \frac{\rho_+ - \zeta_+}{2}\big\|\boldsymbol{\beta}_t^{(k)} - \boldsymbol{\beta}_t^{(k-1)}\big\|_2^2. \tag{C.109}$$

Now we prove that (C.109) guarantees the line-search method in Algorithm 2 produces $L_t^{(k)} \le 2(\rho_+ - \zeta_+)$. We prove by contradiction. We assume that, when the line-search method stops, it outputs $L_t^{(k)} > 2(\rho_+ - \zeta_+)$. Recall that we double $L_t^{(k)}$ at each line-search iteration (Line 6 of Algorithm 2). Then at the line-search iteration right before the line-search method stops, we have $L_t^{(k)\prime} = L_t^{(k)}/2 > (\rho_+ - \zeta_+)$. Remind that the objective function $\phi_\lambda(\boldsymbol{\beta}) = \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1$. Adding $\lambda_t\big\|\boldsymbol{\beta}_t^{(k)}\big\|_1$ to the both sides of (C.109), we obtain

$$\begin{aligned}
\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k)}) &= \widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k)}) + \lambda_t\big\|\boldsymbol{\beta}_t^{(k)}\big\|_1 \\
&\le \widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)}) + \nabla\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)})^T(\boldsymbol{\beta}_t^{(k)} - \boldsymbol{\beta}_t^{(k-1)}) + \frac{\rho_+ - \zeta_+}{2}\big\|\boldsymbol{\beta}_t^{(k)} - \boldsymbol{\beta}_t^{(k-1)}\big\|_2^2 + \lambda_t\big\|\boldsymbol{\beta}_t^{(k)}\big\|_1 \\
&\le \widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)}) + \nabla\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)})^T(\boldsymbol{\beta}_t^{(k)} - \boldsymbol{\beta}_t^{(k-1)}) + \frac{L_t^{(k)\prime}}{2}\big\|\boldsymbol{\beta}_t^{(k)} - \boldsymbol{\beta}_t^{(k-1)}\big\|_2^2 + \lambda_t\big\|\boldsymbol{\beta}_t^{(k)}\big\|_1 \\
&= \psi_{L_t^{(k)\prime},\lambda_t}(\boldsymbol{\beta}_t^{(k)}; \boldsymbol{\beta}_t^{(k-1)}),
\end{aligned}$$

where the last equality follows from (3.7). The stopping criterion of Algorithm 2 (Line 7) implies that the line-search method should have already stopped and output $L_t^{(k)\prime} = L_t^{(k)}/2$, which contradicts our assumption that the line-search method outputs $L_t^{(k)}$. Therefore we have

$$L_t^{(k)} \le 2(\rho_+ - \zeta_+). \tag{C.110}$$

Moreover, according to (C.108) and (C.109), Lemma C.1 holds, i.e.,

$$\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k)}) \le \phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)}) - \frac{L_t^{(k)}}{2}\big\|\boldsymbol{\beta}_t^{(k)} - \boldsymbol{\beta}_t^{(k-1)}\big\|_2^2, \tag{C.111}$$

which implies

$$\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k)}) - \phi_{\lambda_t}(\boldsymbol{\beta}^*) \le \phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)}) - \frac{L_t^{(k)}}{2}\big\|\boldsymbol{\beta}_t^{(k)} - \boldsymbol{\beta}_t^{(k-1)}\big\|_2^2 - \phi_{\lambda_t}(\boldsymbol{\beta}^*) \le \frac{21/2}{\rho_- - \zeta_-}\lambda_t^2 s^*. \tag{C.112}$$

According to (C.107) and (C.110)−(C.112), now we have

$$\big\|(\boldsymbol{\beta}_t^{(k)})_{\overline{S^*}}\big\|_0 \le \widetilde{s}, \quad L_t^{(k)} \le 2(\rho_+ - \zeta_+), \quad \text{and} \quad \phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k)}) - \phi_{\lambda_t}(\boldsymbol{\beta}^*) \le \frac{21/2}{\rho_- - \zeta_-}\lambda_t^2 s^*. \tag{C.113}$$

Combining (C.104), (C.106) and (C.113), by induction we prove that (C.113) holds for all $k = 0, 1, \ldots$ within the t-th path following stage. Furthermore, by Lemma 5.3 we know that all $\boldsymbol{\beta}_t^{(k)}$'s have nice statistical recovery properties, i.e.,

$$\big\|\boldsymbol{\beta}_t^{(k)} - \boldsymbol{\beta}^*\big\|_2 \leq \frac{15/2}{\rho_- - \zeta_-}\lambda_t\sqrt{s^*}, \quad \text{for} \quad k = 0, 1, \ldots.$$

Convergence to a Unique Local Solution: In the following, we prove that, within the t-th path following stage, the limit point of the sequence $\{\boldsymbol{\beta}_t^{(k)}\}_{k=0}^{\infty}$ generated by Algorithm 3 is a unique exact local solution. Since $\big\|(\boldsymbol{\beta}_t^{(0)})_{\overline{S^*}}\big\| \leq \widetilde{s}$, according to the restricted strong convexity of $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta})$ in Lemma 5.1, the sub-level set

$$\Big\{\boldsymbol{\beta} : \phi_{\lambda_t}(\boldsymbol{\beta}) \leq \phi_{\lambda_t}(\boldsymbol{\beta}_t^{(0)}),\ \big\|(\boldsymbol{\beta}_t^{(0)} - \boldsymbol{\beta})_{\overline{S^*}}\big\| \leq 2\widetilde{s}\Big\}$$

is bounded. From (C.111) and (C.113) we have

$$\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k)}) \leq \phi_{\lambda_t}(\boldsymbol{\beta}_t^{(0)}) \quad \text{and} \quad \big\|(\boldsymbol{\beta}_t^{(k)})_{\overline{S^*}}\big\|_0 \leq \widetilde{s}, \quad \text{for} \quad k = 1, 2, \ldots.$$

Thus $\{\boldsymbol{\beta}_t^{(k)}\}_{k=0}^{\infty}$ is bounded, which implies $\{\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k)})\}_{k=0}^{\infty}$ is also bounded. Meanwhile, (C.111) implies that $\{\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k)})\}_{k=0}^{\infty}$ decreases monotonically. By the Bolzano-Weierstrass theorem, the limit point of $\{\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k)})\}_{k=0}^{\infty}$ is unique, which implies

$$\lim_{k\to\infty}\Big\{\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k)}) - \phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)})\Big\} = 0.$$

Consequently, by (C.111) we have that, for any limit point of $\{\boldsymbol{\beta}^{(k)}\}_{k=0}^{\infty}$,

$$\lim_{k\to\infty}\Big\{\big\|\boldsymbol{\beta}_t^{(k)} - \boldsymbol{\beta}_t^{(k-1)}\big\|_2\Big\} \leq \frac{2}{L_t^{(k)}} \cdot \lim_{k\to\infty}\Big\{\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k)}) - \phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)})\Big\} = 0.$$

Moreover, Lemma C.2 implies

$$\lim_{k\to\infty}\Big\{\omega_{\lambda_t}(\boldsymbol{\beta}_t^{(k)})\Big\} \leq \Big(L_t^{(k)} + (\rho_+ - \zeta_+)\Big) \cdot \lim_{k\to\infty}\Big\{\big\|\boldsymbol{\beta}_t^{(k)} - \boldsymbol{\beta}_t^{(k-1)}\big\|_2\Big\} = 0.$$

In other words, $\{\boldsymbol{\beta}_t^{(k)}\}_{k=0}^{\infty}$ has a convergent subsequence such that $\lim_{k\to\infty}\{\omega_{\lambda_t}(\boldsymbol{\beta}_t^{(k)})\} \leq 0$. Furthermore, it implies that such a convergent subsequence of $\{\boldsymbol{\beta}_t^{(k)}\}_{k=0}^{\infty}$ converges towards an exact local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ that satisfies the exact optimal condition in (3.14). By (C.113) we have $\big\|(\boldsymbol{\beta}_t^{(k)})_{\overline{S^*}}\big\|_0 \leq \widetilde{s}$ $(k = 1, 2, \ldots)$, which implies $\big\|(\widehat{\boldsymbol{\beta}}_{\lambda_t})_{\overline{S^*}}\big\|_0 \leq \widetilde{s}$.

Now we prove the uniqueness of this exact local solution by contradiction. Let $\boldsymbol{\xi} \in \partial\big\|\widehat{\boldsymbol{\beta}}_{\lambda_t}\big\|_1$ be the subgradient that attains the minimum in

$$\omega_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) = \min_{\boldsymbol{\xi}' \in \partial\|\widehat{\boldsymbol{\beta}}_{\lambda_t}\|_1} \max_{\boldsymbol{\beta}' \in \Omega}\left\{\frac{(\widehat{\boldsymbol{\beta}}_{\lambda_t} - \boldsymbol{\beta}')^T}{\big\|\widehat{\boldsymbol{\beta}}_{\lambda_t} - \boldsymbol{\beta}'\big\|_1}\Big(\nabla\widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) + \lambda_t\boldsymbol{\xi}'\Big)\right\}. \tag{C.114}$$

Since $\omega_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) \leq 0$, we have

$$\max_{\boldsymbol{\beta}' \in \Omega}\left\{\frac{(\widehat{\boldsymbol{\beta}}_{\lambda_t} - \boldsymbol{\beta}')^T}{\big\|\widehat{\boldsymbol{\beta}}_{\lambda_t} - \boldsymbol{\beta}'\big\|_1}\Big(\nabla\widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) + \lambda_t\boldsymbol{\xi}\Big)\right\} \leq 0. \tag{C.115}$$

We assume there exists another local solution $\widehat{\boldsymbol{\beta}}'_{\lambda_t}$, which is the limit point of another convergent subsequence of $\{\boldsymbol{\beta}_t^{(k)}\}_{k=0}^{\infty}$. Since $\|(\widehat{\boldsymbol{\beta}}'_{\lambda_t})_{\overline{S^*}}\|_0 \le \widetilde{s}$, we have $\|(\widehat{\boldsymbol{\beta}}'_{\lambda_t} - \widehat{\boldsymbol{\beta}}_{\lambda_t})_{\overline{S^*}}\| \le 2\widetilde{s}$. In the setting of logistic loss, we have $\|\widehat{\boldsymbol{\beta}}'_{\lambda_t}\|_2 \le R$ and $\|\widehat{\boldsymbol{\beta}}_{\lambda_t}\|_2 \le R$ by the ℓ_2 constraint. Hence Lemma 5.1 implies

$$\widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}'_{\lambda_t}) \ge \widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) + (\widehat{\boldsymbol{\beta}}'_{\lambda_t} - \widehat{\boldsymbol{\beta}}_{\lambda_t})^T \nabla\widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) + \frac{\rho_- - \zeta_-}{2}\|\widehat{\boldsymbol{\beta}}'_{\lambda_t} - \widehat{\boldsymbol{\beta}}_{\lambda_t}\|_2^2. \tag{C.116}$$

Meanwhile, the convexity of ℓ_1 norm implies

$$\lambda_t\|\widehat{\boldsymbol{\beta}}'_{\lambda_t}\|_1 \ge \lambda_t\|\widehat{\boldsymbol{\beta}}_{\lambda_t}\|_1 + \lambda_t(\widehat{\boldsymbol{\beta}}'_{\lambda_t} - \widehat{\boldsymbol{\beta}}_{\lambda_t})^T\boldsymbol{\xi}. \tag{C.117}$$

Recall that the objective function $\phi_\lambda(\boldsymbol{\beta}) = \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1$. Adding (C.116) and (C.117), we obtain

$$\phi_{\lambda_t}(\widehat{\boldsymbol{\beta}}'_{\lambda_t}) - \phi_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) \ge \underbrace{\left(\nabla\widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) + \lambda_t\boldsymbol{\xi}\right)^T(\widehat{\boldsymbol{\beta}}'_{\lambda_t} - \widehat{\boldsymbol{\beta}}_{\lambda_t})}_{\text{(i)}} + \frac{\rho_- - \zeta_-}{2}\|\widehat{\boldsymbol{\beta}}'_{\lambda_t} - \widehat{\boldsymbol{\beta}}_{\lambda_t}\|_2^2. \tag{C.118}$$

Since (C.115) implies

$$\frac{(\widehat{\boldsymbol{\beta}}_{\lambda_t} - \widehat{\boldsymbol{\beta}}'_{\lambda_t})^T}{\|\widehat{\boldsymbol{\beta}}_{\lambda_t} - \widehat{\boldsymbol{\beta}}'_{\lambda_t}\|_1}\left(\nabla\widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) + \lambda_t\boldsymbol{\xi}\right) \le \max_{\boldsymbol{\beta}'\in\Omega}\left\{\frac{(\widehat{\boldsymbol{\beta}}_{\lambda_t} - \boldsymbol{\beta}')^T}{\|\widehat{\boldsymbol{\beta}}_{\lambda_t} - \boldsymbol{\beta}'\|_1}\left(\nabla\widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) + \lambda_t\boldsymbol{\xi}\right)\right\} \le 0,$$

term (i) in (C.118) is nonnegative. Hence we obtain

$$\phi_{\lambda_t}(\widehat{\boldsymbol{\beta}}'_{\lambda_t}) - \phi_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) \ge \frac{\rho_- - \zeta_-}{2}\|\widehat{\boldsymbol{\beta}}'_{\lambda_t} - \widehat{\boldsymbol{\beta}}_{\lambda_t}\|_2^2. \tag{C.119}$$

Because we already know that the limit point of $\{\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k)})\}_{k=0}^{\infty}$ is unique, which implies $\phi_{\lambda_t}(\widehat{\boldsymbol{\beta}}'_{\lambda_t}) - \phi_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) = 0$. Then we obtain $\|\widehat{\boldsymbol{\beta}}'_{\lambda_t} - \widehat{\boldsymbol{\beta}}_{\lambda_t}\|_2^2 = 0$, which contradicts our assumption that $\widehat{\boldsymbol{\beta}}'_{\lambda_t} \ne \widehat{\boldsymbol{\beta}}_{\lambda_t}$. In other words, we prove that the sequence $\{\boldsymbol{\beta}_t^{(k)}\}_{k=0}^{\infty}$ converges to a unique local solution $\widehat{\boldsymbol{\beta}}_{\lambda_t}$.

Geometric Rate of Convergence of Algorithm 3: Now we establish the geometric rate of convergence of Algorithm 3. According to the stopping criterion of Algorithm 2, we have

$$\begin{aligned}
&\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k)}) \le \psi_{L_t^{(k)},\lambda_t}(\boldsymbol{\beta}_t^{(k)}; \boldsymbol{\beta}_t^{(k-1)}) \\
&= \min_{\boldsymbol{\beta}}\left\{\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)}) + \nabla\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)})^T(\boldsymbol{\beta} - \boldsymbol{\beta}_t^{(k-1)}) + \frac{L_t^{(k)}}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_t^{(k-1)}\|_2^2 + \lambda_t\|\boldsymbol{\beta}\|_1\right\} \\
&\le \min_{\substack{\boldsymbol{\beta}=\alpha\widehat{\boldsymbol{\beta}}_{\lambda_t}+(1-\alpha)\boldsymbol{\beta}_t^{(k-1)} \\ \alpha\in[0,1]}}\left\{\underbrace{\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)}) + \nabla\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)})^T(\boldsymbol{\beta} - \boldsymbol{\beta}_t^{(k-1)})}_{\text{(i)}} + \frac{L_t^{(k)}}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_t^{(k-1)}\|_2^2 + \lambda_t\|\boldsymbol{\beta}\|_1\right\}.
\end{aligned} \tag{C.120}$$

For term (i), since $\|(\boldsymbol{\beta}_t^{(k-1)})_{\overline{S^*}}\|_0 \le \widetilde{s}$, $\|(\widehat{\boldsymbol{\beta}}_{\lambda_t})_{\overline{S^*}}\|_0 \le \widetilde{s}$ and $\boldsymbol{\beta} = \alpha\widehat{\boldsymbol{\beta}}_{\lambda_t} + (1-\alpha)\boldsymbol{\beta}_t^{(k-1)}$ with $\alpha \in [0,1]$, we have $\|(\boldsymbol{\beta} - \boldsymbol{\beta}_t^{(k-1)})_{\overline{S^*}}\|_0 \le 2\widetilde{s}$. For logistic loss, since $\|\boldsymbol{\beta}_t^{(k-1)}\|_2 \le R$ and $\|\widehat{\boldsymbol{\beta}}_{\lambda_t}\|_2 \le R$, we have $\|\boldsymbol{\beta}\|_2 \le R$, since the ℓ_2 ball $B_2(R)$ is a convex set. Applying Lemma 5.1, we have

$$\begin{aligned}
\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}) &\ge \widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)}) + \nabla\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)})^T(\boldsymbol{\beta} - \boldsymbol{\beta}_t^{(k-1)}) + \frac{\rho_- - \zeta_-}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_t^{(k-1)}\|_2^2 \\
&\ge \widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)}) + \nabla\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)})^T(\boldsymbol{\beta} - \boldsymbol{\beta}_t^{(k-1)}),
\end{aligned} \tag{C.121}$$

where the second inequality follows from (4.5). Plugging (C.121) into (C.120), we obtain

$$\phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k)}\big) \leq \min_{\substack{\boldsymbol{\beta}=\alpha\widehat{\boldsymbol{\beta}}_{\lambda_t}+(1-\alpha)\boldsymbol{\beta}_t^{(k-1)} \\ \alpha\in[0,1]}} \bigg\{\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}) + \frac{L_t^{(k)}}{2}\big\|\boldsymbol{\beta}-\boldsymbol{\beta}_t^{(k-1)}\big\|_2^2 + \lambda_t\|\boldsymbol{\beta}\|_1\bigg\}. \tag{C.122}$$

Since $\big\|\big(\boldsymbol{\beta}_t^{(k-1)}\big)_{\overline{S^*}}\big\|_0 \leq \widetilde{s}$ and $\big\|\big(\widehat{\boldsymbol{\beta}}_{\lambda_t}\big)_{\overline{S^*}}\big\|_0 \leq \widetilde{s}$ imply $\big\|\big(\widehat{\boldsymbol{\beta}}_{\lambda_t}-\boldsymbol{\beta}_t^{(k-1)}\big)_{\overline{S^*}}\big\|_0 \leq 2\widetilde{s}$, Lemma 5.1 implies that the strong convexity of $\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta})$ holds for $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ and $\boldsymbol{\beta}_t^{(k-1)}$. Hence we have

$$\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}) = \widetilde{\mathcal{L}}_{\lambda_t}\big(\alpha\widehat{\boldsymbol{\beta}}_{\lambda_t}+(1-\alpha)\boldsymbol{\beta}^{(k-1)}\big) \leq \alpha\widetilde{\mathcal{L}}_{\lambda_t}\big(\widehat{\boldsymbol{\beta}}_{\lambda_t}\big) + (1-\alpha)\widetilde{\mathcal{L}}_{\lambda_t}\big(\boldsymbol{\beta}^{(k-1)}\big). \tag{C.123}$$

Meanwhile, by the convexity of ℓ_1 norm we have

$$\lambda_t\|\boldsymbol{\beta}\|_1 = \lambda_t\big\|\alpha\widehat{\boldsymbol{\beta}}_{\lambda_t}+(1-\alpha)\boldsymbol{\beta}^{(k-1)}\big\|_1 \leq \alpha\lambda_t\big\|\widehat{\boldsymbol{\beta}}_{\lambda_t}\big\|_1 + (1-\alpha)\big\|\boldsymbol{\beta}^{(k-1)}\big\|_1. \tag{C.124}$$

Plugging (C.123) and (C.124) into the right-hand side of (C.122), we obtain

$$\begin{aligned}
\phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k)}\big) \leq \min_{\alpha\in[0,1]} \bigg\{ & \alpha\Big(\widetilde{\mathcal{L}}_{\lambda_t}\big(\widehat{\boldsymbol{\beta}}_{\lambda_t}\big) + \lambda_t\big\|\widehat{\boldsymbol{\beta}}_{\lambda_t}\big\|_1\Big) + (1-\alpha)\Big(\widetilde{\mathcal{L}}_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}\big) + \lambda_t\big\|\boldsymbol{\beta}_t^{(k-1)}\big\|_1\Big) \\
& + \frac{L_t^{(k)}}{2}\big\|\alpha\widehat{\boldsymbol{\beta}}_{\lambda_t}+(1-\alpha)\boldsymbol{\beta}_t^{(k-1)}-\boldsymbol{\beta}_t^{(k-1)}\big\|_2^2\bigg\} \\
= \min_{\alpha\in[0,1]} \bigg\{ & \alpha\phi_{\lambda_t}\big(\widehat{\boldsymbol{\beta}}_{\lambda_t}\big) + (1-\alpha)\phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}\big) + \frac{L_t^{(k)}}{2}\big\|\alpha\widehat{\boldsymbol{\beta}}_{\lambda_t}+(1-\alpha)\boldsymbol{\beta}_t^{(k-1)}-\boldsymbol{\beta}_t^{(k-1)}\big\|_2^2\bigg\} \\
\leq \min_{\alpha\in[0,1]} \bigg\{ & \phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}\big) - \alpha\Big(\phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}\big) - \phi_{\lambda_t}\big(\widehat{\boldsymbol{\beta}}_{\lambda_t}\big)\Big) + \frac{\alpha^2 L_t^{(k)}}{2}\underbrace{\big\|\boldsymbol{\beta}_t^{(k-1)}-\widehat{\boldsymbol{\beta}}_{\lambda_t}\big\|_2^2}_{\text{(i)}}\bigg\}.
\end{aligned} \tag{C.125}$$

For term (i), similar to (C.119), applying the exact optimality condition of $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ and the restricted strong convexity of $\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta})$, we obtain

$$\phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}\big) - \phi_{\lambda_t}\big(\widehat{\boldsymbol{\beta}}_{\lambda_t}\big) \geq \frac{\rho_- - \zeta_-}{2}\big\|\boldsymbol{\beta}_t^{(k-1)}-\widehat{\boldsymbol{\beta}}_{\lambda_t}\big\|_2^2.$$

Plugging this into the right-hand side of (C.125), we obtain

$$\begin{aligned}
\phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k)}\big) \leq \min_{\alpha\in[0,1]} \bigg\{ & \phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}\big) - \alpha\Big(\phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}\big) - \phi_{\lambda_t}\big(\widehat{\boldsymbol{\beta}}_{\lambda_t}\big)\Big) \\
& + \frac{\alpha^2 L_t^{(k)}}{2}\cdot\frac{2}{\rho_- - \zeta_-}\Big(\phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}\big) - \phi_{\lambda_t}\big(\widehat{\boldsymbol{\beta}}_{\lambda_t}\big)\Big)\bigg\}.
\end{aligned} \tag{C.126}$$

The right-hand side of (C.126) attains the minimum when $\alpha = (\rho_- - \zeta_-)/\big(2L_t^{(k)}\big)$. Plugging this value of α, we obtain

$$\phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k)}\big) \leq \phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}\big) - \frac{\rho_- - \zeta_-}{4L_t^{(k)}}\Big(\phi_{\lambda_t}\big(\boldsymbol{\beta}_t^{(k-1)}\big) - \phi_{\lambda_t}\big(\widehat{\boldsymbol{\beta}}_{\lambda_t}\big)\Big),$$

which implies

$$\begin{aligned}\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k)}) - \phi_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) &\leq \Big(\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)}) - \phi_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t})\Big) - \frac{\rho_- - \zeta_-}{4L_t^{(k)}}\Big(\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)}) - \phi_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t})\Big) \\ &= \Big(1 - \frac{\rho_- - \zeta_-}{4L_t^{(k)}}\Big)\Big(\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)}) - \phi_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t})\Big). \qquad \text{(C.127)}\end{aligned}$$

Recall that in (C.113) we have $L_t^{(k)} \leq 2(\rho_+ - \zeta_+)$ $(k = 0, 1, \ldots)$. Plugging in this into the right-hand side of (C.127), we obtain

$$\begin{aligned}\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k)}) - \phi_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) &\leq \Big(1 - \frac{1}{8}\cdot\underbrace{\frac{\rho_- - \zeta_-}{\rho_+ - \zeta_+}}_{1/\kappa}\Big)\Big(\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k-1)}) - \phi_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t})\Big) \\ &= \Big(1 - \frac{1}{8\kappa}\Big)^2\Big(\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k-2)}) - \phi_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t})\Big) \\ &\vdots \\ &= \Big(1 - \frac{1}{8\kappa}\Big)^k\Big(\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(0)}) - \phi_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t})\Big). \qquad \text{(C.128)}\end{aligned}$$

Here κ is the condition number defined in (4.4). Now we are ready to characterize the total number of proximal-gradient steps required to obtain an approximate solution $\widetilde{\boldsymbol{\beta}}_t = \boldsymbol{\beta}_t^{(k+1)}$ that satisfies

$$\omega_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_t) \leq \lambda_t/4 \quad (t = 1, \ldots, N-1), \quad \text{or} \quad \omega_{\lambda_t}(\widetilde{\boldsymbol{\beta}}) \leq \epsilon_{\text{opt}} \quad (t = N). \qquad \text{(C.129)}$$

From Lemma C.2, we have

$$\omega_{\lambda_t}(\boldsymbol{\beta}_t^{(k+1)}) \leq \Big(L_t^{(k+1)} + (\rho_+ - \zeta_+)\Big)\big\|\boldsymbol{\beta}_t^{(k+1)} - \boldsymbol{\beta}_t^{(k)}\big\|_2 = L_t^{(k+1)}\left(1 + \frac{\rho_+ - \zeta_+}{L_t^{(k+1)}}\right)\big\|\boldsymbol{\beta}_t^{(k+1)} - \boldsymbol{\beta}_t^{(k)}\big\|_2. \qquad \text{(C.130)}$$

Note that the stopping criterion of the line-search method (Line 7 of Algorithm 2) implies $L_t^{(k+1)} \geq \rho_- - \zeta_-$. Otherwise, we assume that $L_t^{(k+1)} < \rho_- - \zeta_-$. Since $\big\|(\boldsymbol{\beta}_t^{(k+1)})_{\overline{S^*}}\big\|_0 \leq \widetilde{s}$ and $\big\|(\boldsymbol{\beta}_t^{(k)})_{\overline{S^*}}\big\|_0 \leq \widetilde{s}$ imply $\big\|(\boldsymbol{\beta}_t^{(k)} - \boldsymbol{\beta}_t^{(k+1)})_{\overline{S^*}}\big\|_0 \leq 2\widetilde{s}$, by Lemma 5.1 we have

$$\begin{aligned}&\psi_{L_t^{(k+1)},\lambda_t}(\boldsymbol{\beta}_t^{(k+1)}; \boldsymbol{\beta}_t^{(k)}) \\ &= \widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k)}) + \nabla\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k)})^T(\boldsymbol{\beta}_t^{(k+1)} - \boldsymbol{\beta}_t^{(k)}) + \frac{L^{(k+1)}}{2}\big\|\boldsymbol{\beta}_t^{(k+1)} - \boldsymbol{\beta}_t^{(k)}\big\|_2^2 + \lambda_t\big\|\boldsymbol{\beta}_t^{(k+1)}\big\|_1 \\ &< \widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k)}) + \nabla\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k)})^T(\boldsymbol{\beta}_t^{(k+1)} - \boldsymbol{\beta}_t^{(k)}) + \frac{\rho_- - \zeta_-}{2}\big\|\boldsymbol{\beta}_t^{(k+1)} - \boldsymbol{\beta}_t^{(k)}\big\|_2^2 + \lambda_t\big\|\boldsymbol{\beta}_t^{(k+1)}\big\|_1 \\ &\leq \widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}_t^{(k+1)}) + \lambda_t\big\|\boldsymbol{\beta}_t^{(k+1)}\big\|_1 \\ &= \phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k+1)}),\end{aligned}$$

where the first equality follows from the definition in (3.7), the first inequality is because we assume $L_t^{(k+1)} < \rho_- - \zeta_-$, the second inequality follows from the restricted strong convexity by Lemma 5.1.

However, this contradicts the stopping criterion $\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k+1)}) \leq \psi_{L_t^{(k+1)},\lambda_t}(\boldsymbol{\beta}_t^{(k+1)}; \boldsymbol{\beta}_t^{(k)})$. Therefore we have proved $L_t^{(k+1)} \geq \rho_- - \zeta_-$. From (C.130) we have

$$\omega_{\lambda_t}(\boldsymbol{\beta}_t^{(k+1)}) \leq L_t^{(k+1)}\left(1 + \frac{\rho_+ - \zeta_+}{\rho_- - \zeta_-}\right)\left\|\boldsymbol{\beta}_t^{(k+1)} - \boldsymbol{\beta}_t^{(k)}\right\|_2 = L_t^{(k+1)}(1+\kappa)\left\|\boldsymbol{\beta}_t^{(k+1)} - \boldsymbol{\beta}_t^{(k)}\right\|_2. \quad \text{(C.131)}$$

Moreover, by Lemma C.1 we have

$$\frac{L_t^{(k+1)}}{2}\left\|\boldsymbol{\beta}_t^{(k+1)} - \boldsymbol{\beta}_t^{(k)}\right\|_2^2 \leq \phi_\lambda(\boldsymbol{\beta}_t^{(k)}) - \phi_\lambda(\boldsymbol{\beta}_t^{(k+1)}).$$

Plugging this into the right-hand side of (C.131), we obtain

$$\begin{aligned}\omega_{\lambda_t}(\boldsymbol{\beta}_t^{(k+1)}) &\leq (1+\kappa)L_t^{(k+1)}\left\|\boldsymbol{\beta}_t^{(k+1)} - \boldsymbol{\beta}_t^{(k)}\right\|_2 \\ &\leq (1+\kappa)\sqrt{2L_t^{(k+1)}\left(\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k)}) - \phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k+1)})\right)}.\end{aligned}$$

According to (C.111), the sequence $\{\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k)})\}_{k=0}^{\infty}$ decreases monotonically. Therefore, we have $\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k+1)}) \geq \phi_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t})$, which implies

$$\omega_{\lambda_t}(\boldsymbol{\beta}_t^{(k+1)}) \leq (1+\kappa)\sqrt{2L_t^{(k+1)}\left(\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k)}) - \phi_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t})\right)}. \quad \text{(C.132)}$$

Now we provide an upper bound of the right-hand side of (C.132). Recall that in (C.113) we have $L_t^{(k)} \leq 2(\rho_+ - \zeta_+)$ $(k = 0, 1, \ldots)$, and in (C.128) we have $\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(k)}) - \phi_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) \leq (1 - 1/(8\kappa))^k\left(\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(0)}) - \phi_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t})\right)$. Note that we assume $\left\|(\boldsymbol{\beta}_t^{(0)})_{\overline{S^*}}\right\|_0 \leq \widetilde{s}$ and $\omega_{\lambda_t}(\boldsymbol{\beta}_t^{(0)}) \leq \lambda_t/2$. In Lemma C.7, we set $\lambda' = \lambda = \lambda_t$ and $\epsilon = \lambda_t/2$, then we have

$$\phi_{\lambda_t}(\boldsymbol{\beta}_t^{(0)}) - \phi_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) \leq \frac{21}{\rho_- - \zeta_-}{\lambda_t}^2 s^*.$$

Plugging these into the right-hand side of (C.132), we obtain

$$\omega_{\lambda_t}(\boldsymbol{\beta}_t^{(k+1)}) \leq (1+\kappa)\sqrt{4(\rho_+ - \zeta_+)\cdot\left(1 - \frac{1}{8\kappa}\right)^k \frac{21}{\rho_- - \zeta_-}{\lambda_t}^2 s^*} = (1+\kappa)\sqrt{84\kappa\left(1 - \frac{1}{8\kappa}\right)^k}\cdot\lambda_t\sqrt{s^*}.$$

Therefore, for $t = 1, \ldots, N-1$, to ensure that $\boldsymbol{\beta}_t^{(k+1)}$ satisfies $\omega_{\lambda_t}(\boldsymbol{\beta}_t^{(k+1)}) \leq \lambda_t/4$, it suffices to make k satisfy

$$(1+\kappa)\sqrt{84\kappa\left(1 - \frac{1}{8\kappa}\right)^k}\cdot\lambda_t\sqrt{s^*} \leq \lambda_t/4,$$

which implies

$$k \geq 2\log\left(8\sqrt{21}\cdot\sqrt{\kappa}(1+\kappa)\cdot\sqrt{s^*}\right)\Big/\log\left(1 - \frac{1}{8\kappa}\right).$$

Similarly, for $t = N$, to ensure that $\boldsymbol{\beta}_t^{(k+1)}$ satisfies $\omega_{\lambda_t}(\boldsymbol{\beta}^{(k+1)}) \leq \epsilon_{\text{opt}}$, k should satisfy

$$k \geq 2\log\left(2\sqrt{21}\cdot\sqrt{\kappa}(1+\kappa)\cdot\sqrt{s^*}\lambda_t/\epsilon_{\text{opt}}\right)\Big/\log\left(1 - \frac{1}{8\kappa}\right).$$

Therefore we conclude the proof of Theorem 5.5. □

C.9 Proof of Theorem 4.5

Before we lay out the proof, we present a useful lemma. It ensures that the approximate solution $\widetilde{\boldsymbol{\beta}}_{t-1}$, which is obtained from the $(t-1)$-th path following stage, is $(\lambda_t/2)$-suboptimal with respect to regularization parameter λ_t, i.e., $\omega_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_{t-1}) \leq \lambda_t/2$.

Lemma C.8. Let $\widetilde{\boldsymbol{\beta}}_{t-1}$ $(t = 1, \ldots, N)$ be the approximate solution obtained from the $(t-1)$-th path following stage (Line 8 of Algorithm 1). If $\omega_{\lambda_{t-1}}(\widetilde{\boldsymbol{\beta}}_{t-1}) \leq \lambda_{t-1}/4$. Under Assumption 4.1 and Assumption 4.4, we have

$$\omega_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_{t-1}) \leq \lambda_t/2,$$

where $\lambda_t = \eta\lambda_{t-1}$ with $\eta \in [0.9, 1)$.

Proof. Consider regularization parameter λ_{t-1}. Let $\boldsymbol{\xi} \in \partial\|\widetilde{\boldsymbol{\beta}}_{t-1}\|_1$ be the subgradient that attains the minimum in

$$\omega_{\lambda_{t-1}}(\widetilde{\boldsymbol{\beta}}_{t-1}) = \min_{\boldsymbol{\xi}' \in \partial\|\widetilde{\boldsymbol{\beta}}_{t-1}\|_1} \max_{\boldsymbol{\beta}' \in \Omega} \left\{ \frac{(\widetilde{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}')^T}{\|\widetilde{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}'\|_1} \left(\nabla \widetilde{\mathcal{L}}_{\lambda_{t-1}}(\widetilde{\boldsymbol{\beta}}_{t-1}) + \lambda_{t-1}\boldsymbol{\xi}' \right) \right\}, \tag{C.133}$$

which implies

$$\omega_{\lambda_{t-1}}(\widetilde{\boldsymbol{\beta}}_{t-1}) = \max_{\boldsymbol{\beta}' \in \Omega} \left\{ \frac{(\widetilde{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}')^T}{\|\widetilde{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}'\|_1} \left(\nabla \widetilde{\mathcal{L}}_{\lambda_{t-1}}(\widetilde{\boldsymbol{\beta}}_{t-1}) + \lambda_{t-1}\boldsymbol{\xi} \right) \right\}. \tag{C.134}$$

Now we consider regularization parameter λ_t. We have

$$\begin{aligned} \omega_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_{t-1}) &= \min_{\boldsymbol{\xi}' \in \partial\|\widetilde{\boldsymbol{\beta}}_{t-1}\|_1} \max_{\boldsymbol{\beta}' \in \Omega} \left\{ \frac{(\widetilde{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}')^T}{\|\widetilde{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}'\|_1} \left(\nabla \widetilde{\mathcal{L}}_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_{t-1}) + \lambda_t\boldsymbol{\xi}' \right) \right\} \\ &\leq \max_{\boldsymbol{\beta}' \in \Omega} \left\{ \frac{(\widetilde{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}')^T}{\|\widetilde{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}'\|_1} \left(\nabla \widetilde{\mathcal{L}}_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_{t-1}) + \lambda_t\boldsymbol{\xi} \right) \right\}, \end{aligned} \tag{C.135}$$

where $\boldsymbol{\xi}$ is defined as the minimizer of (C.133). Recall that $\nabla \widetilde{\mathcal{L}}_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_{t-1}) = \nabla \mathcal{L}(\widetilde{\boldsymbol{\beta}}_{t-1}) + \nabla \mathcal{Q}_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_{t-1})$. We have

$$\begin{aligned} \nabla \widetilde{\mathcal{L}}_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_{t-1}) + \lambda_t\boldsymbol{\xi} &= \left(\nabla \mathcal{L}(\widetilde{\boldsymbol{\beta}}_{t-1}) + \nabla \mathcal{Q}_{\lambda_{t-1}}(\widetilde{\boldsymbol{\beta}}_{t-1}) + \lambda_t\boldsymbol{\xi} \right) + (\lambda_{t-1}\boldsymbol{\xi} - \lambda_t\boldsymbol{\xi}) \\ &\quad + \left(\nabla \mathcal{Q}_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_{t-1}) - \nabla \mathcal{Q}_{\lambda_{t-1}}(\widetilde{\boldsymbol{\beta}}_{t-1}) \right). \end{aligned}$$

Plugging this into the right-hand side of (C.135), we obtain

$$\begin{aligned} \omega_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_{t-1}) \leq & \underbrace{\max_{\boldsymbol{\beta}' \in \Omega} \left\{ \frac{(\widetilde{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}')^T}{\|\widetilde{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}'\|_1} \left(\nabla \widetilde{\mathcal{L}}_{\lambda_{t-1}}(\widetilde{\boldsymbol{\beta}}_{t-1}) + \lambda_{t-1}\boldsymbol{\xi} \right) \right\}}_{\text{(i)}} + \underbrace{\max_{\boldsymbol{\beta}' \in \Omega} \left\{ \frac{(\widetilde{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}')^T}{\|\widetilde{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}'\|_1} (\lambda_{t-1}\boldsymbol{\xi} - \lambda_t\boldsymbol{\xi}) \right\}}_{\text{(ii)}} \\ & + \underbrace{\max_{\boldsymbol{\beta}' \in \Omega} \left\{ \frac{(\widetilde{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}')^T}{\|\widetilde{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}'\|_1} \left(\nabla \mathcal{Q}_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_{t-1}) - \nabla \mathcal{Q}_{\lambda_{t-1}}(\widetilde{\boldsymbol{\beta}}_{t-1}) \right) \right\}}_{\text{(iii)}}. \end{aligned} \tag{C.136}$$

According to (C.134), term (i) in (C.136) is equal to $\omega_{\lambda_{t-1}}(\widetilde{\boldsymbol{\beta}}_{t-1})$, which is upper bounded by $\lambda_{t-1}/4$ by our assumption. For term (ii) in (C.136), we have

$$\begin{aligned}\max_{\boldsymbol{\beta}'\in\Omega}\left\{\frac{(\widetilde{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta}')^T}{\|\widetilde{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta}'\|_1}(\lambda_{t-1}\boldsymbol{\xi}-\lambda_t\boldsymbol{\xi})\right\} &\leq \max_{\boldsymbol{\beta}'\in\mathbb{R}^d}\left\{\frac{(\widetilde{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta}')^T}{\|\widetilde{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta}'\|_1}(\lambda_{t-1}\boldsymbol{\xi}-\lambda_t\boldsymbol{\xi})\right\}\\ &= \|\lambda_{t-1}\boldsymbol{\xi}-\lambda_t\boldsymbol{\xi}\|_\infty\\ &\leq \lambda_{t-1}-\lambda_t,\end{aligned}$$

where first inequality is due to the duality between ℓ_1 and ℓ_∞ norm, while the second inequality is due to the fact that $\lambda_{t-1}>\lambda_t$ and $\|\boldsymbol{\xi}\|_\infty\leq 1$, which follows from $\boldsymbol{\xi}\in\partial\|\widetilde{\boldsymbol{\beta}}_{t-1}\|_1$. Similarly, for term (iii) we have

$$\begin{aligned}\max_{\boldsymbol{\beta}'\in\Omega}\left\{\frac{(\widetilde{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta}')^T}{\|\widetilde{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta}'\|_1}\left(\nabla\mathcal{Q}_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_{t-1})-\nabla\mathcal{Q}_{\lambda_{t-1}}(\widetilde{\boldsymbol{\beta}}_{t-1})\right)\right\} &\leq \|\nabla\mathcal{Q}_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_{t-1})-\nabla\mathcal{Q}_{\lambda_{t-1}}(\widetilde{\boldsymbol{\beta}}_{t-1})\|_\infty\\ &= \max_{1\leq j\leq d}\left|q'_{\lambda_t}((\widetilde{\beta}_{t-1})_j)-q'_{\lambda_{t-1}}((\widetilde{\beta}_{t-1})_j)\right|\\ &\leq \lambda_{t-1}-\lambda_t,\end{aligned}$$

where the second inequality follows from regularity condition (e). Hence, from (C.136) we obtain

$$\omega_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_{t-1})\leq \underbrace{\lambda_{t-1}/4}_{\text{(i) in (C.136)}}+\underbrace{\lambda_{t-1}-\lambda_t}_{\text{(ii) in (C.136)}}+\underbrace{\lambda_{t-1}-\lambda_t}_{\text{(iii) in(C.136)}}\leq\big(1/(4\eta)+1/\eta-1+1/\eta-1\big)\lambda_t\leq\lambda_t/2,$$

where the last inequality is obtained by plugging in $\eta\in[0.9,1)$. Hence we conclude the proof. □

Now we are ready to prove Theorem 4.5.

Proof. **Geometric Rate of Convergence within Each Stage:** The stopping criterion of Algorithm 3 (Line 9) implies

$$\omega_{\lambda_{t-1}}(\widetilde{\boldsymbol{\beta}}_{t-1})\leq\lambda_{t-1}/4,\quad\text{for}\quad t=1,\ldots,N.$$

By Lemma C.8 we have

$$\omega_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_{t-1})\leq\lambda_t/2,\quad\text{for}\quad t=1,\ldots,N.\tag{C.137}$$

Recall that we initialize the t-th stage with $\widetilde{\boldsymbol{\beta}}_{t-1}=\boldsymbol{\beta}_t^0$ and $L_{t-1}=L_t^{(0)}$ (Line 8 of Algorithm 1). By Theorem 5.5, as long as $\|(\widetilde{\boldsymbol{\beta}}_{t-1})_{\overline{S^*}}\|_0\leq\widetilde{s}$ and $L_{(t-1)}\leq 2(\rho_+-\zeta_+)$, we have

$$\|(\boldsymbol{\beta}_t^{(k)})_{\overline{S^*}}\|_0\leq\widetilde{s},\quad L_t^{(k)}\leq 2(\rho_+-\zeta_+),\quad\text{for}\quad k=1,2,\ldots,$$

which implies $\|(\widetilde{\boldsymbol{\beta}}_t)_{\overline{S^*}}\|_0\leq\widetilde{s}$ and $L_t\leq 2(\rho_+-\zeta_+)$. Recall that we initialize the entire path following procedure with $\widetilde{\boldsymbol{\beta}}_0=\mathbf{0}$ and $L_0=L_{\min}\leq 2(\rho_+-\zeta_+)$ (Line 4 of Algorithm 1). By induction we obtain

$$\|(\widetilde{\boldsymbol{\beta}}_t)_{\overline{S^*}}\|_0\leq\widetilde{s},\quad L_t\leq 2(\rho_+-\zeta_+),\quad\text{for}\quad t=1,\ldots,N.$$

By setting $\lambda = \lambda_t$ and $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}_t$ $(t = 1, \ldots, N)$ in Theorem 5.5, we obtain that, within the t-th stage $(t = 1, \ldots, N-1)$, the total number of proximal-gradient iterations is no more than

$$2\log\big(8\sqrt{21}\cdot\sqrt{\kappa}(1+\kappa)\cdot\sqrt{s^*}\big)\Big/\log\left(\frac{1}{1-1/(8\kappa)}\right),$$

while within the N-th stage, the total number of proximal-gradient steps is no more that

$$2\log\big(2\sqrt{21}\cdot\sqrt{\kappa}(1+\kappa)\cdot\sqrt{s^*}\lambda_{\mathrm{tgt}}/\epsilon_{\mathrm{opt}}\big)\Big/\log\left(\frac{1}{1-1/(8\kappa)}\right).$$

Hence we obtain the first conclusion.

Geometric Rate of Convergence over the Full Path: Now we prove the second statement about the total number of proximal-gradient steps along the entire solution path. The total number of path following stages is

$$N = \log(\lambda_{\mathrm{tgt}}/\lambda_0)/\log\eta.$$

Together with the first result, we have that the total number of proximal-gradient steps is no more than

$$(N-1)C'\log\big(4C\sqrt{s^*}\big) + C'\log\big(C\sqrt{s^*}\lambda_{\mathrm{tgt}}/\epsilon_{\mathrm{opt}}\big).$$

where

$$C = 2\sqrt{21}\cdot\sqrt{\kappa}(1+\kappa), \quad C' = 2\Big/\log\left(\frac{1}{1-1/(8\kappa)}\right).$$

Geometric Rate of Convergence of the Objective Function Values: Now we prove the third statement concerning the objective function value. For $t = 1, \ldots, N-1$, by (C.137) we have $\omega_{\lambda_{t+1}}(\widetilde{\boldsymbol{\beta}}_t) \le \lambda_{t+1}/2$. Setting $\lambda' = \lambda_{\mathrm{tgt}}$, $\lambda = \lambda_{t+1}$, $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}_t$ and $\epsilon = \lambda_{t+1}/2$ in Lemma C.7, we obtain

$$\phi_{\lambda_{\mathrm{tgt}}}(\widetilde{\boldsymbol{\beta}}_t) - \phi_{\lambda_{\mathrm{tgt}}}(\widehat{\boldsymbol{\beta}}_{\lambda_{\mathrm{tgt}}}) \le \frac{21}{\rho_- - \zeta_-}\big(\lambda_{t+1}/2 + 2(\lambda_{t+1} - \lambda_{\mathrm{tgt}})\big)\cdot(\lambda_{\mathrm{tgt}} + \lambda_{t+1})s^*.$$

Since $\lambda_{\mathrm{tgt}} \le \lambda_{t+1}$, we have

$$\phi_{\lambda_{\mathrm{tgt}}}(\widetilde{\boldsymbol{\beta}}_t) - \phi_{\lambda_{\mathrm{tgt}}}(\widehat{\boldsymbol{\beta}}_{\lambda_{\mathrm{tgt}}}) \le \frac{21}{\rho_- - \zeta_-}\big(\lambda_{t+1}/2 + 2\lambda_{t+1}\big)\cdot 2\lambda_{t+1}s^* = \frac{105\cdot\lambda_{t+1}^2 s^*}{\rho_- - \zeta_-}.$$

Since $\lambda_{t+1} = \eta^{t+1}\lambda_0$, we obtain

$$\phi_{\lambda_{\mathrm{tgt}}}(\widetilde{\boldsymbol{\beta}}_t) - \phi_{\lambda_{\mathrm{tgt}}}(\widehat{\boldsymbol{\beta}}_{\lambda_{\mathrm{tgt}}}) \le \eta^{2(t+1)}\frac{105\cdot\lambda_0^2 s^*}{\rho_- - \zeta_-}, \quad \text{for } t = 1, \ldots, N-1.$$

Similarly, for $t = N$, we have $\omega_{\lambda_{\mathrm{tgt}}}(\widetilde{\boldsymbol{\beta}}_N) \le \epsilon_{\mathrm{opt}}$. By setting $\lambda = \lambda' = \lambda_{\mathrm{tgt}}$ and $\epsilon = \epsilon_{\mathrm{opt}}$ in Lemma C.7, we have

$$\phi_{\lambda_{\mathrm{tgt}}}(\widetilde{\boldsymbol{\beta}}_t) - \phi_{\lambda_{\mathrm{tgt}}}(\widehat{\boldsymbol{\beta}}_{\lambda_{\mathrm{tgt}}}) \le \frac{21\cdot\lambda_{\mathrm{tgt}}s^*}{\rho_- - \zeta_-}\epsilon_{\mathrm{opt}}.$$

Therefore we conclude the proof of Theorem 4.5. □

C.10 Proof of Theorem 4.7

Proof. Recall that $\widetilde{\boldsymbol{\beta}}_t$ is the approximate local solution obtained from the t-th path following stage (Lines 8 and 12 of Algorithm 1). Hence it satisfies the stopping criterion of the proximal-gradient method (Line 9 of Algorithm 3), i.e., for $t = 1, \ldots, N-1$ we have $\omega_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_t) \leq \lambda_t/4 < \lambda_t/2$, while for $t = N$ we have $\omega_{\lambda_t}(\widetilde{\boldsymbol{\beta}}_t) \leq \epsilon_{\text{opt}} \ll \lambda_{\text{tgt}}/4 < \lambda_t/2$. Meanwhile, by (5.2) in Theorem 5.5, $\widetilde{\boldsymbol{\beta}}_t$ satisfies $\|(\widetilde{\boldsymbol{\beta}}_t)_{\overline{S^*}}\|_0 \leq \widetilde{s}$. For logistic loss, we further have $\|\widetilde{\boldsymbol{\beta}}_t\|_2 \leq R$ due to the ℓ_2 constraint. Therefore Lemma 5.2 gives

$$\|\widetilde{\boldsymbol{\beta}}_t - \boldsymbol{\beta}^*\|_2 \leq \frac{21/8}{\rho_- - \zeta_-}\lambda_t\sqrt{s^*}, \quad \text{for} \quad t = 1, \ldots, N,$$

which concludes the proof. □

C.11 Proof of Theorem 4.8

Proof. We denote the subgradients by $\boldsymbol{\xi}^* \in \partial\|\boldsymbol{\beta}^*\|_1$ and $\widehat{\boldsymbol{\xi}} \in \partial\|\widehat{\boldsymbol{\beta}}_{\lambda_t}\|_1$. In particular, we set $\widehat{\boldsymbol{\xi}}$ to be the subgradient that attains the minimum in

$$\omega_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) = \min_{\boldsymbol{\xi}' \in \partial\|\widehat{\boldsymbol{\beta}}_{\lambda_t}\|_1} \max_{\boldsymbol{\beta}' \in \Omega} \left\{ \frac{(\widehat{\boldsymbol{\beta}}_{\lambda_t} - \boldsymbol{\beta}')^T}{\|\widehat{\boldsymbol{\beta}}_{\lambda_t} - \boldsymbol{\beta}'\|_1} \left(\nabla \widetilde{\mathcal{L}}_\lambda(\widehat{\boldsymbol{\beta}}_{\lambda_t}) + \lambda \boldsymbol{\xi}' \right) \right\}.$$

Recall that $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ satisfies the exact optimality condition that $\omega_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) \leq 0$, hence we have

$$\max_{\boldsymbol{\beta}' \in \Omega} \left\{ (\widehat{\boldsymbol{\beta}}_{\lambda_t} - \boldsymbol{\beta}')^T \left(\nabla \widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) + \lambda_t \widehat{\boldsymbol{\xi}} \right) \right\} \leq 0. \tag{C.138}$$

By Theorem 5.5 we have $\|(\widehat{\boldsymbol{\beta}}_{\lambda_t})_{\overline{S^*}}\|_0 \leq \widetilde{s}$. Since $\|(\widehat{\boldsymbol{\beta}}_{\lambda_t} - \boldsymbol{\beta}^*)_{\overline{S^*}}\|_0 \leq \widetilde{s}$, according to Lemma 5.1 the restricted convexity holds for $\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta})$ at $\boldsymbol{\beta}_t$ and $\boldsymbol{\beta}^*$, i.e.,

$$\widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) \geq \widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}^*) + \nabla \widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}^*)^T(\widehat{\boldsymbol{\beta}}_{\lambda_t} - \boldsymbol{\beta}^*) + \frac{\rho_- - \zeta_-}{2}\|\widehat{\boldsymbol{\beta}}_{\lambda_t} - \boldsymbol{\beta}^*\|_2^2, \tag{C.139}$$

$$\widetilde{\mathcal{L}}_{\lambda_t}(\boldsymbol{\beta}^*) \geq \widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) + \nabla \widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t})^T(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{\lambda_t}) + \frac{\rho_- - \zeta_-}{2}\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{\lambda_t}\|_2^2. \tag{C.140}$$

Meanwhile, by the convexity of ℓ_1 norm, we have

$$\lambda_t\|\widehat{\boldsymbol{\beta}}_{\lambda_t}\|_1 \geq \lambda_t\|\boldsymbol{\beta}^*\|_1 + \lambda_t(\widehat{\boldsymbol{\beta}}_{\lambda_t} - \boldsymbol{\beta}^*)^T\boldsymbol{\xi}^*, \tag{C.141}$$

$$\lambda_t\|\boldsymbol{\beta}^*\|_1 \geq \lambda_t\|\widehat{\boldsymbol{\beta}}_{\lambda_t}\|_1 + \lambda_t(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{\lambda_t})^T\widehat{\boldsymbol{\xi}}. \tag{C.142}$$

Recall that $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \mathcal{Q}_\lambda(\boldsymbol{\beta})$. Adding (C.139)−(C.142), we obtain

$$\begin{aligned} 0 \geq\ & \underbrace{\left(\nabla\mathcal{L}(\boldsymbol{\beta}^*) + \nabla\mathcal{Q}_{\lambda_t}(\boldsymbol{\beta}^*) + \lambda_t\boldsymbol{\xi}^*\right)^T(\widehat{\boldsymbol{\beta}}_{\lambda_t} - \boldsymbol{\beta}^*)}_{\text{(i)}} + \underbrace{\left(\nabla\widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) + \lambda_t\widehat{\boldsymbol{\xi}}\right)^T(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{\lambda_t})}_{\text{(ii)}} \\ & + (\rho_- - \zeta_-)\|\widehat{\boldsymbol{\beta}}_{\lambda_t} - \boldsymbol{\beta}^*\|_2^2. \end{aligned} \tag{C.143}$$

According to (C.138) we have

$$\Big(\nabla\widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t})+\lambda_t\widehat{\boldsymbol{\xi}}\Big)^T(\widehat{\boldsymbol{\beta}}_{\lambda_t}-\boldsymbol{\beta}^*)\le\max_{\boldsymbol{\beta}'\in\Omega}\Big\{(\widehat{\boldsymbol{\beta}}_{\lambda_t}-\boldsymbol{\beta}')^T\Big(\nabla\widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t})+\lambda\widehat{\boldsymbol{\xi}}\Big)\Big\}\le 0,$$

which implies that term (ii) in (C.143) is nonnegative. Moving term (i) in (C.143) to its left-hand side, we obtain

$$\begin{aligned}(\rho_- -\zeta_-)\big\|\widehat{\boldsymbol{\beta}}_{\lambda_t}-\boldsymbol{\beta}^*\big\|_2^2 &\le \big(\nabla\mathcal{L}(\boldsymbol{\beta}^*)+\nabla\mathcal{Q}_{\lambda_t}(\boldsymbol{\beta}^*)+\lambda_t\boldsymbol{\xi}^*\big)^T(\widehat{\boldsymbol{\beta}}_{\lambda_t}-\boldsymbol{\beta}^*)\\ &\le \min_{\boldsymbol{\xi}^*\in\partial\|\boldsymbol{\beta}^*\|_1}\Big\{\sum_{j=1}^d\Big|\big(\nabla\mathcal{L}(\boldsymbol{\beta}^*)+\nabla\mathcal{Q}_{\lambda_t}(\boldsymbol{\beta}^*)+\lambda_t\boldsymbol{\xi}^*\big)_j\Big|\cdot\Big|\big(\boldsymbol{\beta}^*-\widehat{\boldsymbol{\beta}}_{\lambda_t}\big)_j\Big|\Big\}.\end{aligned}\tag{C.144}$$

In the following, we decompose the summation in (C.144) into three parts: $j\in\overline{S^*}$, $j\in S_1^*$ and $j\in S_2^*$, where $S_1^*=\{j:|\beta_j|\ge\nu_t\}$ and $S_2^*=\{j:|\beta_j|<\nu_t\}$. Here $\nu_t>0$ is defined in (4.16).

- For $j\in\overline{S^*}$, according to regularity condition (c), we have

  $$\big(\nabla\mathcal{Q}_{\lambda_t}(\boldsymbol{\beta}^*)\big)_j=q'_{\lambda_t}(\beta_j^*)=q'_{\lambda_t}(0)=0,\quad\text{for}\quad j\in\overline{S^*}.$$

  By (4.1) in Assumption 4.1, we have

  $$\max_{j\in\overline{S^*}}\Big|\big(\nabla\mathcal{L}(\boldsymbol{\beta}^*)\big)_j\Big|\le\max_{1\le j\le d}\Big|\big(\nabla\mathcal{L}(\boldsymbol{\beta}^*)\big)_j\Big|=\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty\le\lambda_{\text{tgt}}/8\le\lambda_t/8<\lambda_t.$$

  Hence we have

  $$\max_{j\in\overline{S^*}}\Big|\big(\nabla\mathcal{L}(\boldsymbol{\beta}^*)+\mathcal{Q}_{\lambda_t}(\boldsymbol{\beta}^*)\big)_j\Big|\le\lambda_t.$$

  Meanwhile, since $\boldsymbol{\xi}^*\in\partial\|\boldsymbol{\beta}^*\|_1$, we have $\lambda_t\boldsymbol{\xi}_j^*\in[-\lambda_t,\lambda_t]$. Therefore, for any $j\in\overline{S^*}$, we can always find a ξ_j^* such that

  $$\Big|\big(\nabla\mathcal{L}(\boldsymbol{\beta}^*)+\nabla\mathcal{Q}_{\lambda_t}(\boldsymbol{\beta}^*)\big)_j+\lambda_t\xi_j^*\Big|=0,$$

  which implies

  $$\min_{\boldsymbol{\xi}^*\in\partial\|\boldsymbol{\beta}^*\|_1}\Big\{\Big|\big(\nabla\mathcal{L}(\boldsymbol{\beta}^*)+\nabla\mathcal{Q}_{\lambda_t}(\boldsymbol{\beta}^*)+\lambda_t\boldsymbol{\xi}^*\big)_j\Big|\Big\}=0,\quad\text{for}\quad j\in\overline{S^*}.$$

  Thus we obtain

  $$\min_{\boldsymbol{\xi}^*\in\partial\|\boldsymbol{\beta}^*\|_1}\Big\{\sum_{j\in\overline{S^*}}\Big|\big(\nabla\mathcal{L}(\boldsymbol{\beta}^*)+\nabla\mathcal{Q}_{\lambda_t}(\boldsymbol{\beta}^*)+\lambda_t\boldsymbol{\xi}^*\big)_j\Big|\cdot\Big|\big(\boldsymbol{\beta}^*-\widehat{\boldsymbol{\beta}}_{\lambda_t}\big)_j\Big|\Big\}=0.\tag{C.145}$$

- For $j\in S_1^*\subseteq S^*$, we have $|\beta_j^*|\ge\nu_t$. Recall that $\mathcal{P}_\lambda(\boldsymbol{\beta})=\mathcal{Q}_\lambda(\boldsymbol{\beta})+\lambda\|\boldsymbol{\beta}\|_1$. By our assumption on $\mathcal{P}_{\lambda_t}(\boldsymbol{\beta})$ in (4.16), we have

  $$\big(\nabla\mathcal{Q}_{\lambda_t}(\boldsymbol{\beta}^*)+\lambda_t\boldsymbol{\xi}^*\big)_j=p'_{\lambda_t}(\beta_j^*)=0,\quad\text{for}\quad j\in S_1^*,$$

which implies

$$\begin{aligned}
&\min_{\boldsymbol{\xi}^* \in \partial \|\boldsymbol{\beta}^*\|_1} \Big\{ \sum_{j \in S_1^*} \Big| \big(\nabla \mathcal{L}(\boldsymbol{\beta}^*) + \nabla \mathcal{Q}_{\lambda_t}(\boldsymbol{\beta}^*) + \lambda_t \boldsymbol{\xi}^*\big)_j \Big| \cdot \Big| \big(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{\lambda_t}\big)_j \Big| \Big\} \\
&= \sum_{j \in S_1^*} \Big| \big(\nabla \mathcal{L}(\boldsymbol{\beta}^*)\big)_j \Big| \cdot \Big| \big(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{\lambda_t}\big)_j \Big| \\
&\leq \big\| \big(\nabla \mathcal{L}(\boldsymbol{\beta}^*)\big)_{S_1^*} \big\|_2 \cdot \big\| \boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{\lambda_t} \big\|_2. \qquad \text{(C.146)}
\end{aligned}$$

- For $j \in S_2^* \subseteq S^*$, we have $|\beta_j^*| < \nu_t$. By (4.1) in Assumption 4.1, we have

$$\max_{j \in S_2^*} \Big| \big(\nabla \mathcal{L}(\boldsymbol{\beta}^*)\big)_j \Big| \leq \max_{1 \leq j \leq d} \Big| \big(\nabla \mathcal{L}(\boldsymbol{\beta}^*)\big)_j \Big| = \|\nabla \mathcal{L}(\boldsymbol{\beta}^*)\|_\infty \leq \lambda_t/8 \leq \lambda_t/8.$$

Meanwhile we have

$$\max_{j \in S_2^*} \Big| \big(\nabla \mathcal{Q}_{\lambda_t}(\boldsymbol{\beta}^*)\big)_j \Big| = \max_{j \in S_2^*} \big| q'_{\lambda_t}(\beta_j^*) \big| \leq \max_{1 \leq j \leq d} \big| q'_{\lambda_t}(\beta_j^*) \big| \leq \lambda_t,$$

where the last inequality follows from regularity condition (d). Also, since $\boldsymbol{\xi}^* \in \partial \|\boldsymbol{\beta}^*\|_1$, we have $|\xi_j^*| \leq 1$. Therefore we obtain that, for $j \in S_2^*$,

$$\Big| \big(\nabla \mathcal{L}(\boldsymbol{\beta}^*) + \nabla \mathcal{Q}_{\lambda_t}(\boldsymbol{\beta}^*) + \lambda_t \boldsymbol{\xi}^*\big)_j \Big| \leq \max_{j \in S_2^*} \Big| \big(\nabla \mathcal{L}(\boldsymbol{\beta}^*)\big)_j \Big| + \max_{j \in S_2^*} \Big| \big(\nabla \mathcal{Q}_{\lambda_t}(\boldsymbol{\beta}^*)\big)_j \Big| + \lambda_t \leq 3\lambda_t.$$

which implies

$$\begin{aligned}
&\min_{\boldsymbol{\xi}^* \in \partial \|\boldsymbol{\beta}^*\|_1} \Big\{ \sum_{j \in S_2^*} \Big| \big(\nabla \mathcal{L}(\boldsymbol{\beta}^*) + \nabla \mathcal{Q}_{\lambda_t}(\boldsymbol{\beta}^*) + \lambda_t \boldsymbol{\xi}^*\big)_j \Big| \cdot \Big| \big(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{\lambda_t}\big)_j \Big| \Big\} \leq 3\lambda_t \sum_{j \in S_2^*} \Big| \big(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{\lambda_t}\big)_j \Big| \\
&= 3\lambda_t \big\| \big(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{\lambda_t}\big)_{\overline{S_2^*}} \big\|_1 \leq 3\lambda_t \sqrt{s^*} \big\| \big(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{\lambda_t}\big)_{\overline{S_2^*}} \big\|_2 \leq 3\lambda_t \sqrt{s_2^*} \big\| \boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{\lambda_t} \big\|_2. \qquad \text{(C.147)}
\end{aligned}$$

Plugging (C.145)–(C.147) into the right-hand side of (C.144), we obtain

$$\big\| \widehat{\boldsymbol{\beta}}_{\lambda_t} - \boldsymbol{\beta}^* \big\|_2 \leq \frac{1}{\rho_- - \zeta_-} \Big(\big\| \big(\nabla \mathcal{L}(\boldsymbol{\beta}^*)\big)_{S_1^*} \big\|_2 + 3\lambda_t \sqrt{s_2^*} \Big),$$

which concludes the proof of Theorem 4.8. □

C.12 Proof of Lemma 4.9 and Theorem 4.10

First we prove Lemma 4.9, which states that the oracle estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{O}}$ is uniquely defined and has some nice statistical recovery property.

Proof. To prove that the global minimizer of (4.19) is unique even for nonconvex loss functions, in the following we show that $\mathcal{L}(\boldsymbol{\beta})$ is actually strongly convex on the sparse set $\{\boldsymbol{\beta} : \mathrm{supp}\,(\boldsymbol{\beta}) \subseteq S^*\}$. Assume that $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ satisfy $\mathrm{supp}(\boldsymbol{\beta}) \subseteq S^*$ and $\mathrm{supp}(\boldsymbol{\beta}') \subseteq S^*$. By Taylor's theorem and the mean value theorem, we have

$$\mathcal{L}(\boldsymbol{\beta}') = \mathcal{L}(\boldsymbol{\beta}) + \nabla \mathcal{L}(\boldsymbol{\beta})^T (\boldsymbol{\beta}' - \boldsymbol{\beta}) + \frac{1}{2} (\boldsymbol{\beta}' - \boldsymbol{\beta})^T \nabla^2 \mathcal{L}\big(\gamma \boldsymbol{\beta}' + (1-\gamma)\boldsymbol{\beta}\big)(\boldsymbol{\beta}' - \boldsymbol{\beta}), \qquad \text{(C.148)}$$

where $\gamma \in [0,1]$. Note that we have $\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_0 = s^* < s^* + 2\widetilde{s}$. By Definition 4.2 and Definition 4.3, we have

$$\frac{(\boldsymbol{\beta}' - \boldsymbol{\beta})^T}{\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2} \nabla^2 \mathcal{L}\big(\gamma\boldsymbol{\beta} + (1-\gamma)\boldsymbol{\beta}'\big) \frac{(\boldsymbol{\beta}' - \boldsymbol{\beta})}{\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2} \geq \rho_-\big(\nabla^2 \mathcal{L}, s^* + 2\widetilde{s}\big).$$

Plugging this into the right-hand side of (C.148), we obtain

$$\mathcal{L}(\boldsymbol{\beta}') \geq \mathcal{L}(\boldsymbol{\beta}) + \nabla \mathcal{L}(\boldsymbol{\beta})^T (\boldsymbol{\beta}' - \boldsymbol{\beta}) + \frac{\rho_-}{2} \|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2, \tag{C.149}$$

where $\rho_- = \rho_-\big(\nabla^2 \mathcal{L}, s^* + 2\widetilde{s}\big)$ is a positive constant according to Assumption 4.4. Note that (C.149) holds for any $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ such that $\mathrm{supp}(\boldsymbol{\beta}) \subseteq S^*$ and $\mathrm{supp}(\boldsymbol{\beta}') \subseteq S^*$. Therefore, $\mathcal{L}(\boldsymbol{\beta})$ is strongly convex on this sparse set, which implies the minimizer of (4.19) is unique.

Now we prove the statistical recovery property of the oracle estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{O}}$ in the setting where $\mathcal{L}(\boldsymbol{\beta})$ is least squares loss. Let $\widehat{\boldsymbol{\beta}}'_{\mathrm{O}}, \boldsymbol{\beta}^{*\prime} \in \mathbb{R}^{s^*}$ be the restrictions of $\widehat{\boldsymbol{\beta}}_{\mathrm{O}}, \boldsymbol{\beta}^* \in \mathbb{R}^d$ to S^* respectively, and $\mathbf{X}_{S^*} \in \mathbb{R}^{n \times s^*}$ be a new matrix containing the columns of $\mathbf{X}$, i.e., $\mathbf{X}_j$, that satisfy $j \in S^*$. Since $\widehat{\boldsymbol{\beta}}'_{\mathrm{O}}$ is the solution to the ordinary least squares problem

$$\widehat{\boldsymbol{\beta}}'_{\mathrm{O}} = \underset{\boldsymbol{\beta}' \in \mathbb{R}^{s^*}}{\mathrm{argmin}} \frac{1}{2n} \|\mathbf{X}_{S^*} \boldsymbol{\beta}' - \mathbf{y}\|_2^2,$$

it has the closed-form expression of

$$\widehat{\boldsymbol{\beta}}'_{\mathrm{O}} = (\mathbf{X}_{S^*}^T \mathbf{X}_{S^*})^{-1} \mathbf{X}_{S^*}^T \mathbf{y}.$$

Here we still need to prove that $\mathbf{X}_{S^*}^T \mathbf{X}_{S^*} \in \mathbb{R}^{s^* \times s^*}$ is invertible. Note that the smallest eigenvalue of $\mathbf{X}_{S^*}^T \mathbf{X}_{S^*}$ is defined as

$$\Lambda_{\min}\big(\mathbf{X}_{S^*}^T \mathbf{X}_{S^*}\big) = \inf\left\{ \boldsymbol{v}^T \mathbf{X}_{S^*}^T \mathbf{X}_{S^*} \boldsymbol{v} : \|\boldsymbol{v}\|_2 = 1,\ \boldsymbol{v} \in \mathbb{R}^{s^*} \right\},$$

which satisfies

$$\begin{aligned}
\Lambda_{\min}\big(\mathbf{X}_{S^*}^T \mathbf{X}_{S^*}\big) &= \inf\left\{ \boldsymbol{v}^T \mathbf{X}^T \mathbf{X} \boldsymbol{v} : \|\boldsymbol{v}\|_2 = 1,\ \boldsymbol{v} \in \mathbb{R}^d,\ \mathrm{supp}(\boldsymbol{v}) = S^* \right\} \\
&\geq \inf\left\{ \boldsymbol{v}^T \mathbf{X}^T \mathbf{X} \boldsymbol{v} : \|\boldsymbol{v}\|_2 = 1,\ \boldsymbol{v} \in \mathbb{R}^d,\ \|\boldsymbol{v}\|_0 \leq s^* \right\} \\
&\geq \inf\left\{ \boldsymbol{v}^T \mathbf{X}^T \mathbf{X} \boldsymbol{v} : \|\boldsymbol{v}\|_2 = 1,\ \boldsymbol{v} \in \mathbb{R}^d,\ \|\boldsymbol{v}\|_0 \leq s^* + 2\widetilde{s} \right\} \\
&= n\rho_-\big(\nabla^2 \mathcal{L}, s^* + 2\widetilde{s}\big) \\
&> 0.
\end{aligned} \tag{C.150}$$

Here the first and second inequality are due to $\{\boldsymbol{v} : \mathrm{supp}(\boldsymbol{v}) = S^*\} \subseteq \{\boldsymbol{v} : \|\boldsymbol{v}\|_0 \leq s^*\} \subseteq \{\boldsymbol{v} : \|\boldsymbol{v}\|_0 \leq s^* + 2\widetilde{s}\}$, while the second equality follows from Definition 4.2, because in the setting of least squares loss $\nabla^2 \mathcal{L}(\boldsymbol{\beta}) = \mathbf{X}^T \mathbf{X}/n$, and the last inequality follows from Assumption 4.4. Therefore the smallest eigenvalue of $\mathbf{X}_{S^*}^T \mathbf{X}_{S^*}$ is positive, which implies that $\mathbf{X}_{S^*}^T \mathbf{X}_{S^*}$ is invertible.

By our assumption on $(Y|\boldsymbol{X} = \mathbf{x}_i)$, we have $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon} = \mathbf{X}_{S^*} \boldsymbol{\beta}^{*\prime} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is a zero mean sub-Gaussian random vector with independent entries and variance proxy σ^2. Therefore, we have

$$\widehat{\boldsymbol{\beta}}'_{\mathrm{O}} - \boldsymbol{\beta}^{*\prime} = (\mathbf{X}_{S^*}^T \mathbf{X}_{S^*})^{-1} \mathbf{X}_{S^*}^T \mathbf{y} - \boldsymbol{\beta}^{*\prime} = (\mathbf{X}_{S^*}^T \mathbf{X}_{S^*})^{-1} \mathbf{X}_{S^*}^T (\mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}) - \boldsymbol{\beta}^{*\prime} = (\mathbf{X}_{S^*}^T \mathbf{X}_{S^*})^{-1} \mathbf{X}_{S^*}^T \boldsymbol{\epsilon}.$$

Now we provide an upper bound of $\|\widehat{\boldsymbol{\beta}}'_{\mathrm{O}} - \boldsymbol{\beta}^{*\prime}\|_\infty$. Note that the j-th entry of $(\mathbf{X}_{S^*}^T\mathbf{X}_{S^*})^{-1}\mathbf{X}_{S^*}^T\boldsymbol{\epsilon} \in \mathbb{R}^{s^*}$ can be denoted as $\boldsymbol{e}_j(\mathbf{X}_{S^*}^T\mathbf{X}_{S^*})^{-1}\mathbf{X}_{S^*}^T\boldsymbol{\epsilon}$. Here $\boldsymbol{e}_j \in \mathbb{R}^{s^*}$ denotes a vector that is all-zero expect an "1" in its j-th coordinate. For any j, $\boldsymbol{e}_j(\mathbf{X}_{S^*}^T\mathbf{X}_{S^*})^{-1}\mathbf{X}_{S^*}^T\boldsymbol{\epsilon}$ is a sub-Gaussian random variable with variance proxy $\|\boldsymbol{e}_j(\mathbf{X}_{S^*}^T\mathbf{X}_{S^*})^{-1}\mathbf{X}_{S^*}^T\|_2^2\sigma^2$. Therefore we have

$$\mathbb{P}\big(|\boldsymbol{e}_j(\mathbf{X}_{S^*}^T\mathbf{X}_{S^*})^{-1}\mathbf{X}_{S^*}^T\boldsymbol{\epsilon}| > t\big) \le 2\exp\Big(-t^2/\big(\|\boldsymbol{e}_j(\mathbf{X}_{S^*}^T\mathbf{X}_{S^*})^{-1}\mathbf{X}_{S^*}^T\|_2^2\sigma^2\big)\Big),$$

which implies

$$\mathbb{P}\left(\max_{j\in\{1,\ldots,s^*\}}|\boldsymbol{e}_j(\mathbf{X}_{S^*}^T\mathbf{X}_{S^*})^{-1}\mathbf{X}_{S^*}^T\boldsymbol{\epsilon}| > t\right) \le 2s^*\exp\left(-t^2\Big/\Big(\max_{j\in\{1,\ldots,s^*\}}\|\boldsymbol{e}_j(\mathbf{X}_{S^*}^T\mathbf{X}_{S^*})^{-1}\mathbf{X}_{S^*}^T\|_2^2\sigma^2\Big)\right).$$

Taking $t = C\max_{j\in\{1,\ldots,s^*\}}\|\boldsymbol{e}_j(\mathbf{X}_{S^*}^T\mathbf{X}_{S^*})^{-1}\mathbf{X}_{S^*}^T\|_2\sigma\cdot\sqrt{2\log s^*}$ with $C>0$, we have that

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}'_{\mathrm{O}} - \boldsymbol{\beta}^{*\prime}\|_\infty = \|(\mathbf{X}_{S^*}^T\mathbf{X}_{S^*})^{-1}\mathbf{X}_{S^*}^T\boldsymbol{\epsilon}\|_\infty &= \max_{j\in\{1,\ldots,s^*\}}|\boldsymbol{e}_j(\mathbf{X}_{S^*}^T\mathbf{X}_{S^*})^{-1}\mathbf{X}_{S^*}^T\boldsymbol{\epsilon}| \\ &\le C\max_{j\in\{1,\ldots,s^*\}}\|\boldsymbol{e}_j(\mathbf{X}_{S^*}^T\mathbf{X}_{S^*})^{-1}\mathbf{X}_{S^*}^T\|_2\sigma\cdot\sqrt{2\log s^*} \quad \text{(C.151)}\end{aligned}$$

holds with probability at least $1-2\exp(-C^2)/s^*$. In other words, there exists a constant $C>0$ sufficiently large such that (C.151) holds with high probability. Note that, for any $j\in\{1,\ldots,d\}$

$$\begin{aligned}\|\boldsymbol{e}_j(\mathbf{X}_{S^*}^T\mathbf{X}_{S^*})^{-1}\mathbf{X}_{S^*}^T\|_2^2 = \boldsymbol{e}_j(\mathbf{X}_{S^*}^T\mathbf{X}_{S^*})^{-1}\mathbf{X}_{S^*}^T\mathbf{X}_{S^*}(\mathbf{X}_{S^*}^T\mathbf{X}_{S^*})^{-1}\boldsymbol{e}_j^T &= \boldsymbol{e}_j(\mathbf{X}_{S^*}^T\mathbf{X}_{S^*})^{-1}\boldsymbol{e}_j^T \\ &\le \Lambda_{\max}\big((\mathbf{X}_{S^*}^T\mathbf{X}_{S^*})^{-1}\big) \\ &= 1/\Lambda_{\min}\big(\mathbf{X}_{S^*}^T\mathbf{X}_{S^*}\big) \\ &\le 1/(n\rho_-),\end{aligned}$$

where the last inequality follows from (C.150). Plugging this into (C.151), we obtain

$$\|\widehat{\boldsymbol{\beta}}'_{\mathrm{O}} - \boldsymbol{\beta}^{*\prime}\|_\infty \le C\sigma\sqrt{2/\rho_-}\cdot\sqrt{\frac{\log s^*}{n}}.$$

Recall that $\widehat{\boldsymbol{\beta}}'_{\mathrm{O}}$ and $\boldsymbol{\beta}^{*\prime}$ are the restrictions of $\widehat{\boldsymbol{\beta}}_{\mathrm{O}}$ and $\boldsymbol{\beta}^*$ to S^*, and $\mathrm{supp}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}) \subseteq S^*$. Therefore we obtain

$$\|\widehat{\boldsymbol{\beta}}_{\mathrm{O}} - \boldsymbol{\beta}^*\|_\infty \le C\sigma\sqrt{2/\rho_-}\cdot\sqrt{\frac{\log s^*}{n}},$$

which concludes the proof. □

Now we prove Theorem 4.10.

Proof. Let $\widehat{\boldsymbol{\xi}} \in \partial\|\widehat{\boldsymbol{\beta}}_{\lambda_t}\|_1$. We set $\widehat{\boldsymbol{\xi}}$ to be the subgradient that attains the minimum in

$$\omega_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) = \min_{\boldsymbol{\xi}'\in\partial\|\widehat{\boldsymbol{\beta}}_{\lambda_t}\|_1}\max_{\boldsymbol{\beta}'\in\Omega}\left\{\frac{(\widehat{\boldsymbol{\beta}}_{\lambda_t}-\boldsymbol{\beta}')^T}{\|\widehat{\boldsymbol{\beta}}_{\lambda_t}-\boldsymbol{\beta}'\|_1}\Big(\nabla\widetilde{\mathcal{L}}_\lambda(\widehat{\boldsymbol{\beta}}_{\lambda_t}) + \lambda\boldsymbol{\xi}'\Big)\right\}.$$

Since $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ satisfies the exact optimality condition that $\omega_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) \le 0$, we have

$$\max_{\boldsymbol{\beta}'\in\Omega}\Big\{(\widehat{\boldsymbol{\beta}}_{\lambda_t}-\boldsymbol{\beta}')^T\Big(\nabla\widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) + \lambda_t\widehat{\boldsymbol{\xi}}\Big)\Big\} \le 0. \quad \text{(C.152)}$$

Now we prove there exists some $\boldsymbol{\xi}_{\mathrm{O}} \in \partial\|\widehat{\boldsymbol{\beta}}_{\mathrm{O}}\|_1$ such that $\widehat{\boldsymbol{\beta}}_{\mathrm{O}}$ satisfies the exact optimality condition

$$\max_{\boldsymbol{\beta}' \in \Omega}\Big\{(\widehat{\boldsymbol{\beta}}_{\mathrm{O}} - \boldsymbol{\beta}')^T\Big(\nabla\widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}) + \lambda_t\boldsymbol{\xi}_{\mathrm{O}}\Big)\Big\} \leq 0. \tag{C.153}$$

Recall that $\widetilde{\mathcal{L}}_{\lambda}(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \mathcal{Q}_{\lambda}(\boldsymbol{\beta})$. In (C.153), we have

$$\begin{aligned}
&(\widehat{\boldsymbol{\beta}}_{\mathrm{O}} - \boldsymbol{\beta}')^T\Big(\nabla\widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}) + \lambda_t\boldsymbol{\xi}_{\mathrm{O}}\Big) \\
&= (\widehat{\boldsymbol{\beta}}_{\mathrm{O}} - \boldsymbol{\beta}')^T\Big(\nabla\mathcal{L}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}) + \nabla\mathcal{Q}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}) + \lambda_t\boldsymbol{\xi}_{\mathrm{O}}\Big) \\
&= \underbrace{\sum_{1\leq j\leq d}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}} - \boldsymbol{\beta}')_j\Big(\nabla\mathcal{Q}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}) + \lambda_t\boldsymbol{\xi}_{\mathrm{O}}\Big)_j}_{\text{(i)}} + \underbrace{(\widehat{\boldsymbol{\beta}}_{\mathrm{O}} - \boldsymbol{\beta}')^T\nabla\mathcal{L}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}})}_{\text{(ii)}}.
\end{aligned} \tag{C.154}$$

For term (i) in (C.154), we decompose the summation into two parts: $j \in S^*$ and $j \in \overline{S^*}$.

- For $j \in \overline{S^*}$, since $\big(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}\big)_j = 0$, by regularity condition (c) we have

  $$\Big(\nabla\mathcal{Q}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}})\Big)_j = 0.$$

  Note that $\boldsymbol{\xi}_{\mathrm{O}} \in \partial\|\widehat{\boldsymbol{\beta}}_{\mathrm{O}}\|_1$. By setting $(\boldsymbol{\xi}_{\mathrm{O}})_j = 0$ for $j \in \overline{S^*}$, we obtain

  $$\sum_{j\in\overline{S^*}}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}} - \boldsymbol{\beta}')_j\Big(\nabla\mathcal{Q}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}) + \lambda_t\boldsymbol{\xi}_{\mathrm{O}}\Big)_j = 0.$$

- For $j \in S^*$, by assumption we have $\big|\big(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}\big)_j\big| > \nu_t$. Recall that $\mathcal{P}_{\lambda}(\boldsymbol{\beta}) = \mathcal{Q}_{\lambda}(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1$. Thus we have

  $$\Big(\nabla\mathcal{Q}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}) + \lambda_t\boldsymbol{\xi}_{\mathrm{O}}\Big)_j = \Big(\nabla\mathcal{P}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}})\Big)_j = p'_{\lambda_t}\Big(\big(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}\big)_j\Big) = 0,$$

  where the second equality follows from our assumption in (4.16).

Therefore, as long as $\boldsymbol{\xi}_{\mathrm{O}} \in \partial\|\widehat{\boldsymbol{\beta}}_{\mathrm{O}}\|_1$ satisfies $(\boldsymbol{\xi}_{\mathrm{O}})_j = 0$ for $j \in \overline{S^*}$, term (i) is always zero for any $\boldsymbol{\beta}'$.

For term (ii) in (C.154), note that $\widehat{\boldsymbol{\beta}}_{\mathrm{O}}$ is the global solution to the minimization problem in (4.19). Hence $\widehat{\boldsymbol{\beta}}_{\mathrm{O}}$ satisfies the exact optimality condition

$$\max_{\boldsymbol{\beta}' \in \Omega}\Big\{(\widehat{\boldsymbol{\beta}}_{\mathrm{O}} - \boldsymbol{\beta}')^T\nabla\mathcal{L}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}})\Big\} \leq 0.$$

Therefore, taking maximum over $\boldsymbol{\beta}' \in \Omega$ on both sides of (C.154), we obtain (C.153).

Now we are ready to prove that $\widehat{\boldsymbol{\beta}}_{\lambda_t} = \widehat{\boldsymbol{\beta}}_{\mathrm{O}}$. Note that the oracle estimator satisfies $\mathrm{supp}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}) \subseteq S^*$. Meanwhile, by Theorem 5.5 we have $\big\|\big(\widehat{\boldsymbol{\beta}}_{\lambda_t}\big)_{\overline{S^*}}\big\|_0 \leq \widetilde{s}$. Hence we have $\big\|\big(\widehat{\boldsymbol{\beta}}_{\lambda_t} - \widehat{\boldsymbol{\beta}}_{\mathrm{O}}\big)_{\overline{S^*}}\big\|_0 \leq \widetilde{s}$. Therefore Lemma 5.1 yields

$$\widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) \geq \widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}) + \nabla\widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}})^T(\widehat{\boldsymbol{\beta}}_{\lambda_t} - \widehat{\boldsymbol{\beta}}_{\mathrm{O}}) + \frac{\rho_- - \zeta_-}{2}\|\widehat{\boldsymbol{\beta}}_{\lambda_t} - \widehat{\boldsymbol{\beta}}_{\mathrm{O}}\|_2^2, \tag{C.155}$$

$$\widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}) \geq \widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) + \nabla\widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t})^T(\widehat{\boldsymbol{\beta}}_{\mathrm{O}} - \widehat{\boldsymbol{\beta}}_{\lambda_t}) + \frac{\rho_- - \zeta_-}{2}\|\widehat{\boldsymbol{\beta}}_{\mathrm{O}} - \widehat{\boldsymbol{\beta}}_{\lambda_t}\|_2^2. \tag{C.156}$$

Meanwhile, by the convexity of ℓ_1 norm, we have

$$\lambda_t\|\widehat{\boldsymbol{\beta}}_{\lambda_t}\|_1 \geq \lambda_t\|\widehat{\boldsymbol{\beta}}_{\mathrm{O}}\|_1 + \lambda_t(\widehat{\boldsymbol{\beta}}_{\lambda_t} - \widehat{\boldsymbol{\beta}}_{\mathrm{O}})^T\boldsymbol{\xi}_{\mathrm{O}}, \tag{C.157}$$

$$\lambda_t\|\widehat{\boldsymbol{\beta}}_{\mathrm{O}}\|_1 \geq \lambda_t\|\widehat{\boldsymbol{\beta}}_{\lambda_t}\|_1 + \lambda_t(\widehat{\boldsymbol{\beta}}_{\mathrm{O}} - \widehat{\boldsymbol{\beta}}_{\lambda_t})^T\widehat{\boldsymbol{\xi}}. \tag{C.158}$$

Adding (C.155)−(C.158), we obtain

$$0 \geq \underbrace{\Big(\nabla\widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) + \lambda_t\widehat{\boldsymbol{\xi}}\Big)^T(\widehat{\boldsymbol{\beta}}_{\mathrm{O}} - \widehat{\boldsymbol{\beta}}_{\lambda_t})}_{\text{(i)}} + \underbrace{\Big(\nabla\widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\mathrm{O}}) + \lambda_t\boldsymbol{\xi}_{\mathrm{O}}\Big)^T(\widehat{\boldsymbol{\beta}}_{\lambda_t} - \widehat{\boldsymbol{\beta}}_{\mathrm{O}})}_{\text{(ii)}} + (\rho_- - \zeta_-)\|\widehat{\boldsymbol{\beta}}_{\lambda_t} - \widehat{\boldsymbol{\beta}}_{\mathrm{O}}\|_2^2.$$

According to (C.152), we have

$$(\widehat{\boldsymbol{\beta}}_{\lambda_t} - \widehat{\boldsymbol{\beta}}_{\mathrm{O}})^T\Big(\nabla\widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) + \lambda_t\widehat{\boldsymbol{\xi}}\Big) \leq \max_{\boldsymbol{\beta}'\in\Omega}\Big\{(\widehat{\boldsymbol{\beta}}_{\lambda_t} - \boldsymbol{\beta}')^T\Big(\nabla\widetilde{\mathcal{L}}_{\lambda_t}(\widehat{\boldsymbol{\beta}}_{\lambda_t}) + \lambda_t\widehat{\boldsymbol{\xi}}\Big)\Big\} \leq 0,$$

which implies term (i) is nonnegative. Similarly, according to (C.153), term (ii) is also nonnegative. Hence we have $(\rho_- - \zeta_-)\|\widehat{\boldsymbol{\beta}}_{\lambda_t} - \widehat{\boldsymbol{\beta}}_{\mathrm{O}}\|_2^2 \leq 0$. By (4.5) we have $\rho_- - \zeta_- > 0$, which implies $\widehat{\boldsymbol{\beta}}_{\lambda_t} = \widehat{\boldsymbol{\beta}}_{\mathrm{O}}$. Thus we conclude that $\widehat{\boldsymbol{\beta}}_{\lambda_t}$ is the oracle estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{O}}$, which exactly recovers the support of $\boldsymbol{\beta}^*$. □

D Theoretical Results about Semiparametric Elliptical Design Regression

In this section, we first introduce the Catoni's M-estimator of standard deviation, then we provide the detailed proofs of some necessary results regarding semiparametric elliptical design regression[1].

D.1 Catoni's M-Estimator of Standard Deviation

Catoni (2012) proposed a novel method to estimate the mean and standard deviation of heavy-tail distributions. Let $\boldsymbol{Z} = (Z_1, \ldots, Z_{d+1})$ be the elliptically distributed random vector defined in §2.2. We consider the estimator of the marginal mean $\mathbb{E}(Z_j)$ $(j = 1, \ldots, d+1)$. Let $h : \mathbb{R} \to \mathbb{R}$ be a continuous strictly increasing function satisfying

$$-\log(1 - x + x^2/2) \leq h(x) \leq \log(1 + x + x^2/2).$$

For instance, we choose $h(\cdot)$ to be

$$h(x) = \begin{cases} \log(1 + x + x^2/2), & \text{if } x \geq 0, \\ -\log(1 - x + x^2/2), & \text{otherwise.} \end{cases}$$

Let $\delta \in (0, 1)$ be such that $n \geq 2\log(1/\delta)$. We introduce

$$a_\delta = \sqrt{2\log(1/\delta)\Big/\Big(nv + \frac{2nv\log(1/\delta)}{n - 2\log(1/\delta)}\Big)}, \tag{D.1}$$

[1] §D.1, Lemma D.1 and Corollary D.2 come from an unpublished internal technical report. We provide them here for completeness.

where v is an upper bound of $\mathrm{Var}(Z_j)$ for all j. Catoni's estimator of $\mathbb{E}(Z_j)$ is defined as $\widehat{\mu}_j = \widehat{\mu}_j(n, \delta)$ such that

$$\sum_{i=1}^{n} h\big(\alpha_\delta(z_{i,j} - \widehat{\mu}_j)\big) = 0, \quad j = 1, \ldots, d+1, \tag{D.2}$$

where $z_{i,j}$ is the i-th $(i = 1, \ldots, n)$ realizations of Z_j. As $h(\cdot)$ is differentiable everywhere, we can solve (D.2) with Newton's method efficiently. Similarly we can estimate $\mathbb{E}(Z_j^2)$ with $\widehat{m}_j$ defined in a similar way. Then we obtain an estimator of the marginal standard deviation σ_j

$$\widehat{\sigma}_j = \sqrt{\widehat{m}_j - \widehat{\mu}_j^2}, \quad j = 1, \ldots, d+1. \tag{D.3}$$

D.2 Proof of Lemma C.5

To establish results concerning the smallest sparse eigenvalue for $\widehat{\mathbf{K}}_{\boldsymbol{X}}$, we need to prove several concentration results. The next lemma and proposition provide the concentration inequality for Catoni's estimator of marginal standard deviation, which is defined in (D.3). We first consider the estimator of variance in the following lemma.

Lemma D.1. Let $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ be a random vector and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be n independent realizations of $\boldsymbol{X}$ with $\mathrm{Var}(X_j) = v_j$ and $\mathbb{E}\big(X_j^4\big) \le M$, for $j = 1, \ldots, d$. We assume that

$$\max_{1 \le j \le d}\big\{|\mathbb{E}(X_j)|\big\} \le \mu_{\max}, \quad v_{\max} = \max_{1 \le j \le d}\big\{v_j\big\}.$$

For the estimator $\widehat{v}_j = \widehat{m}_j - \widehat{\mu}_j^2$ with $\widehat{m}_j$ and $\widehat{\mu}_j$ defined in (D.2), if $n > 5 \log d$, we have, with probability at least $1 - 2d^{-3}$,

$$\max_{1 \le j \le d}\Big\{\big|v_j - \widehat{v}_j\big|\Big\} \le C\sqrt{\frac{\log d}{n}},$$

where C is a constant.

Proof. For $j \in \{1, \ldots, d\}$, we use $\widehat{m}_j$ to estimate $\mathbb{E}\big(X_j^2\big)$. Catoni (2012) showed that

$$\mathbb{P}\Big(\big|\widehat{m}_j - \mathbb{E}\big(X_j^2\big)\big| > t\Big) \le \exp\Big(-\frac{nt^2}{M}\Big).$$

Taking a union bound, we have

$$\mathbb{P}\Big(\max_{1 \le j \le d}\Big\{\big|\widehat{m}_j - \mathbb{E}\big(X_j^2\big)\big|\Big\} > t\Big) \le d\exp\Big(-\frac{nt^2}{M}\Big),$$

or equivalently, with probability at least $1 - d^{-3}$,

$$\max_{1 \le j \le d}\Big\{\big|\widehat{m}_j - \mathbb{E}\big(X_j^2\big)\big|\Big\} \le 2\sqrt{M}\sqrt{\frac{\log d}{n}}. \tag{D.4}$$

Meanwhile, we use $\widehat{\mu}_j$ to estimate $\mathbb{E}(X_j)$. By similar arguments as above, we have

$$\max_{1 \le j \le d}\Big\{\big|\widehat{\mu}_j - \mathbb{E}(X_j)\big|\Big\} \le 2\sqrt{v_{\max}}\sqrt{\frac{\log d}{n}} \tag{D.5}$$

with probability at least $1 - d^{-3}$.

Note that

$$\max_{1\le j\le d}\Big\{\Big|\widehat{\mu}_j^2 - \big(\mathbb{E}(X_j)\big)^2\Big|\Big\} \le \max_{1\le j\le d}\Big\{\big|\widehat{\mu}_j - \mathbb{E}(X_j)\big|\Big\}\cdot \max_{1\le j\le d}\Big\{\big|\widehat{\mu}_j + \mathbb{E}(X_j)\big|\Big\}.$$

Since we assume that $\max_{1\le j\le d}\{\mathbb{E}(X_j)\} \le \mu_{\max}$, we have

$$\max_{1\le j\le d}\Big\{\Big|\widehat{\mu}_j^2 - \big(\mathbb{E}(X_j)\big)^2\Big|\Big\} \le \left(4\mu_{\max} + 4\sqrt{v_{\max}}\sqrt{\frac{\log d}{n}}\right)\cdot\sqrt{v_{\max}}\sqrt{\frac{\log d}{n}} \tag{D.6}$$

with probability at least $1 - d^{-3}$. Since $\log d/n < 1$, from (D.6) we have,

$$\max_{1\le j\le d}\Big\{\Big|\widehat{\mu}_j^2 - \big(\mathbb{E}(X_j)\big)^2\Big|\Big\} \le (4\mu_{\max} + 4\sqrt{v_{\max}})\cdot\sqrt{v_{\max}}\sqrt{\frac{\log d}{n}}. \tag{D.7}$$

Combining (D.4) and (D.7), we have, with probability at least $1 - 2d^{-3}$,

$$\max_{1\le j\le d}\Big\{\big|\widehat{m}_j - \widehat{\mu}_j^2 - \mathrm{Var}(X_j)\big|\Big\} \le C\sqrt{\frac{\log d}{n}},$$

where $C = 2\sqrt{M} + \big(4\mu_{\max} + 4\sqrt{v_{\max}}\big)\sqrt{v_{\max}}$. □

We use $\widehat{\sigma}_j = \sqrt{\widehat{v}_j}$ to estimate $\sigma_j = \sqrt{v_j}$. Using Lemma D.1, we derive a concentration inequality for $\widehat{\sigma}_j$ in the following corollary.

Corollary D.2. Let $\sigma_j = \sqrt{v_j}$ and $\widehat{\sigma}_j = \sqrt{\widehat{v}_j}$ for $j = 1, ..., d$. By assuming $\sigma_j \ge \sigma_{\min} > 0$ for all $j = 1, ..., d$, we have, with probability at least $1 - 2d^{-3}$,

$$\max_{1\le j\le d}\Big\{|\sigma_j - \widehat{\sigma}_j|\Big\} \le C\sqrt{\frac{\log d}{n}},$$

where C is a constant.

Proof. By Lemma D.1, we have, with probability at least $1 - 2d^{-3}$,

$$\max_{1\le j\le d}\Big\{|v_j - \widehat{v}_j|\Big\} \le C\sqrt{\frac{\log d}{n}}.$$

Since $|v_j - \widehat{v}_j| = |\sigma_j - \widehat{\sigma}_j|\cdot|\sigma_j + \widehat{\sigma}_j|$, it follows that

$$\max_{1\le j\le d}\Big\{|\sigma_j - \widehat{\sigma}_j|\Big\} \le \frac{C}{\min_{1\le j\le d}\Big\{|\sigma_j + \widehat{\sigma}_j|\Big\}}\sqrt{\frac{\log d}{n}} \le \frac{C}{\sigma_{\min}}\sqrt{\frac{\log d}{n}}.$$

As we assume that $\sigma_j > \sigma_{\min}$ for all j, we conclude the proof. □

Before we establish the sparse eigenvalue condition for $\widehat{\mathbf{K}}_{\boldsymbol{X}}$, we provide a concentration result of $\widehat{\mathbf{R}}_{\boldsymbol{X}}$ in the following lemma.

Lemma D.3 (Han and Liu (2013))**.** Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be n realizations of a random vector $\boldsymbol{X} \sim \mathrm{EC}_d(0, \boldsymbol{\Sigma}_{\boldsymbol{X}}, \Xi)$ as in Definition 2.1. We assume that the smallest eigenvalue of the generalized correlation matrix $\boldsymbol{\Sigma}^0_{\boldsymbol{X}}$ is strictly positive. Under the sign sub-Gaussian condition (See Han and Liu (2013) for more details), the correlation matrix estimator $\widehat{\mathbf{R}}_{\boldsymbol{X}}$ defined in (2.6) satisfies that, with probability at least $1 - 2d^{-1} - d^{-2}$,

$$\sup_{\|\boldsymbol{v}\|_0 \le s} \left\{ \frac{\left| \boldsymbol{v}^T (\widehat{\mathbf{R}}_{\boldsymbol{X}} - \boldsymbol{\Sigma}^0_{\boldsymbol{X}}) \boldsymbol{v} \right|}{\|\boldsymbol{v}\|_2^2} \right\} \le C \sqrt{\frac{s \log d}{n}}$$

for $s \in \{1, \ldots, d\}$ and a sufficiently large n.

We now prove Lemma C.5.

Proof. Let $\mathbf{D} = \mathrm{diag}(\sigma_1, \ldots, \sigma_d)$ and $\widehat{\mathbf{D}} = \mathrm{diag}(\widehat{\sigma}_1, \ldots, \widehat{\sigma}_d)$. First we consider the smallest sparse eigenvalue, which satisfies

$$\begin{aligned} \rho_-(\nabla^2 \mathcal{L}, s) &= \inf_{\|\boldsymbol{v}\|_0 \le s} \left\{ \frac{\boldsymbol{v}^T \widehat{\mathbf{K}}_{\boldsymbol{X}} \boldsymbol{v}}{\|\boldsymbol{v}\|_2^2} \right\} \\ &= \inf_{\|\boldsymbol{v}\|_0 \le s} \left\{ \frac{(\widehat{\mathbf{D}}\boldsymbol{v})^T \widehat{\mathbf{R}}_{\boldsymbol{X}} (\widehat{\mathbf{D}}\boldsymbol{v})}{\|\widehat{\mathbf{D}}\boldsymbol{v}\|_2^2} \cdot \frac{\|\widehat{\mathbf{D}}\boldsymbol{v}\|_2^2}{\|\boldsymbol{v}\|_2^2} \right\} \\ &\ge \inf_{\|\boldsymbol{v}\|_0 \le s} \left\{ \frac{\boldsymbol{v}^T \widehat{\mathbf{R}}_{\boldsymbol{X}} \boldsymbol{v}}{\|\boldsymbol{v}\|_2^2} \right\} \cdot \min_{1 \le j \le d} \{\widehat{\sigma}_j\}. \end{aligned} \tag{D.8}$$

The first term on the right-hand side of (D.8) is the smallest sparse eigenvalue of $\widehat{\mathbf{R}}_{\boldsymbol{X}}$. Since we have from Lemma D.3 that, with probability at least $1 - 2d^{-1} - d^{-2}$,

$$\sup_{\|\boldsymbol{v}\|_0 \le s} \left\{ \frac{\left| \boldsymbol{v}^T (\widehat{\mathbf{R}}_{\boldsymbol{X}} - \boldsymbol{\Sigma}^0_{\boldsymbol{X}}) \boldsymbol{v} \right|}{\|\boldsymbol{v}\|_2^2} \right\} \le C \sqrt{\frac{s \log d}{n}}.$$

Then for a sufficiently large n, we have

$$\boldsymbol{v}^T (\boldsymbol{\Sigma}^0_{\boldsymbol{X}} - \widehat{\mathbf{R}}_{\boldsymbol{X}}) \boldsymbol{v} \le C \sqrt{\frac{s \log d}{n}} \le \frac{1}{2} \Lambda_{\min}(\boldsymbol{\Sigma}^0_{\boldsymbol{X}}), \quad \text{for} \quad \|\boldsymbol{v}\|_0 \le s.$$

Here $\Lambda_{\min}(\boldsymbol{\Sigma}^0_{\boldsymbol{X}})$ denotes the smallest eigenvalue of $\boldsymbol{\Sigma}^0_{\boldsymbol{X}}$, which is strictly positive by assumption. Then we obtain

$$\frac{1}{2} \Lambda_{\min}(\boldsymbol{\Sigma}^0_{\boldsymbol{X}}) \le \boldsymbol{v}^T \boldsymbol{\Sigma}^0_{\boldsymbol{X}} \boldsymbol{v} - \frac{1}{2} \Lambda_{\min}(\boldsymbol{\Sigma}^0_{\boldsymbol{X}}) \le \boldsymbol{v}^T \widehat{\mathbf{R}}_{\boldsymbol{X}} \boldsymbol{v}, \quad \text{for} \quad \|\boldsymbol{v}\|_0 \le s.$$

Taking infimum over both sides, we get

$$\inf_{\|\boldsymbol{v}\|_0 \le s} \left\{ \frac{\boldsymbol{v}^T \widehat{\mathbf{R}}_{\boldsymbol{X}} \boldsymbol{v}}{\|\boldsymbol{v}\|_2^2} \right\} \ge \frac{1}{2} \Lambda_{\min}(\boldsymbol{\Sigma}^0_{\boldsymbol{X}}) > 0. \tag{D.9}$$

We now consider $\min_{1\le j\le d}\{\widehat{\sigma}_j\}$ in (D.8). In Corollary D.2 we prove that, with probability at least $1-2d^{-3}$,

$$|\sigma_j-\widehat{\sigma}_j|\le C'\sqrt{\frac{\log d}{n}},\quad \text{for}\ \ 1\le j\le d,$$

where C' is a constant. For a sufficiently large n, we have

$$\widehat{\sigma}_j\ge\frac{1}{2}\sigma_j>0,\quad \text{for}\ \ 1\le j\le d$$

with the same probability. Taking minimum over both sides, we get

$$\min_{1\le j\le d}\{\widehat{\sigma}_j\}\ge\frac{1}{2}\min_{1\le j\le d}\{\sigma_j\}>0 \tag{D.10}$$

with probability at least $1-2d^{-2}$. Plugging (D.9) and (D.10) into the right-hand side of (D.8), we reach the conclusion that $\rho_-\big(\nabla^2\mathcal{L},s\big)>0$.

Now we consider the largest sparse eigenvalue, which satisfies

$$\begin{aligned}\rho_+\big(\nabla^2\mathcal{L},s\big) &= \sup_{\|\boldsymbol{v}\|_0\le s}\left\{\frac{\boldsymbol{v}^T\widehat{\mathbf{K}}_{\boldsymbol{X}}\boldsymbol{v}}{\|\boldsymbol{v}\|_2^2}\right\}\\ &= \sup_{\|\boldsymbol{v}\|_0\le s}\left\{\frac{(\widehat{\mathbf{D}}\boldsymbol{v})^T\widehat{\mathbf{R}}_{\boldsymbol{X}}(\widehat{\mathbf{D}}\boldsymbol{v})}{\|\widehat{\mathbf{D}}\boldsymbol{v}\|_2^2}\cdot\frac{\|\widehat{\mathbf{D}}\boldsymbol{v}\|_2^2}{\|\boldsymbol{v}\|_2^2}\right\}\\ &\ge \sup_{\|\boldsymbol{v}\|_0\le s}\left\{\frac{\boldsymbol{v}^T\widehat{\mathbf{R}}_{\boldsymbol{X}}\boldsymbol{v}}{\|\boldsymbol{v}\|_2^2}\right\}\cdot\min_{1\le j\le d}\{\widehat{\sigma}_j\}.\end{aligned} \tag{D.11}$$

The first term on the right-hand side of (D.11) is the largest sparse eigenvalue of $\widehat{\mathbf{R}}_{\boldsymbol{X}}$. Since we have from Lemma D.3 that, with probability at least $1-2d^{-1}-d^{-2}$,

$$\sup_{\|\boldsymbol{v}\|_0\le s}\left\{\frac{\big|\boldsymbol{v}^T(\widehat{\mathbf{R}}_{\boldsymbol{X}}-\boldsymbol{\Sigma}^0_{\boldsymbol{X}})\boldsymbol{v}\big|}{\|\boldsymbol{v}\|_2^2}\right\}\le C\sqrt{\frac{s\log d}{n}}.$$

Then for a sufficiently large n, we have

$$\boldsymbol{v}^T(\widehat{\mathbf{R}}_{\boldsymbol{X}}-\boldsymbol{\Sigma}^0_{\boldsymbol{X}})\boldsymbol{v}\le C\sqrt{\frac{s\log d}{n}}\le\frac{1}{2}\Lambda_{\max}(\boldsymbol{\Sigma}^0_{\boldsymbol{X}}),\quad \text{for}\ \ \|\boldsymbol{v}\|_0\le s.$$

Here $\Lambda_{\max}(\boldsymbol{\Sigma}^0_{\boldsymbol{X}})$ denotes the largest eigenvalue of $\boldsymbol{\Sigma}^0_{\boldsymbol{X}}<+\infty$. Then we obtain

$$\boldsymbol{v}^T\widehat{\mathbf{R}}_{\boldsymbol{X}}\boldsymbol{v}\le\boldsymbol{v}^T\boldsymbol{\Sigma}^0_{\boldsymbol{X}}\boldsymbol{v}+\frac{1}{2}\Lambda_{\max}(\boldsymbol{\Sigma}^0_{\boldsymbol{X}})\le\frac{3}{2}\Lambda_{\max}(\boldsymbol{\Sigma}^0_{\boldsymbol{X}}),\quad \text{for}\ \ \|\boldsymbol{v}\|_0\le s.$$

Taking supremum over both sides, we get

$$\sup_{\|\boldsymbol{v}\|_0\le s}\left\{\frac{\boldsymbol{v}^T\widehat{\mathbf{R}}_{\boldsymbol{X}}\boldsymbol{v}}{\|\boldsymbol{v}\|_2^2}\right\}\le\frac{1}{2}\Lambda_{\max}(\boldsymbol{\Sigma}^0_{\boldsymbol{X}})<+\infty. \tag{D.12}$$

We now consider $\max_{1\le j\le d}\{\widehat{\sigma}_j\}$ in (D.11). In Corollary D.2 we prove that, with probability at least $1-2d^{-3}$,

$$|\sigma_j - \widehat{\sigma}_j| \le C'\sqrt{\frac{\log d}{n}}, \quad \text{for } 1 \le j \le d,$$

where C' is a constant. For a sufficiently large n, we have

$$\widehat{\sigma}_j \le \frac{3}{2}\sigma_j < +\infty, \quad \text{for } 1 \le j \le d$$

with the same probability. Taking minimum over both sides, we get

$$\max_{1\le j\le d}\{\widehat{\sigma}_j\} \le \frac{3}{2}\max_{1\le j\le d}\{\sigma_j\} < +\infty \tag{D.13}$$

with probability at least $1-2d^{-2}$. Plugging (D.12) and (D.13) into the right-hand side of (D.11), we reach the conclusion that $\rho_+\big(\nabla^2\mathcal{L}, s\big) < +\infty$. □

D.3 Proof of Lemma C.4

Proof. For semiparametric elliptical design regression, we have

$$\nabla\mathcal{L}(\boldsymbol{\beta}^*) = \widehat{\mathbf{K}}_{\boldsymbol{X},Y} - \widehat{\mathbf{K}}_{\boldsymbol{X}}\boldsymbol{\beta}^* = \widehat{\mathbf{K}}_{\boldsymbol{X},Y} - \boldsymbol{\Sigma}_{\boldsymbol{X},Y} + \boldsymbol{\Sigma}_{\boldsymbol{X},Y} - \widehat{\mathbf{K}}_{\boldsymbol{X}}\boldsymbol{\beta}^*,$$

where $\widehat{\mathbf{K}}_{\boldsymbol{X}} \in \mathbb{R}^{d\times d}$ and $\widehat{\mathbf{K}}_{\boldsymbol{X},Y} \in \mathbb{R}^{d\times 1}$ are the submatrices of $\widehat{\mathbf{K}}_{\boldsymbol{Z}} \in \mathbb{R}^{(d+1)\times(d+1)}$ defined in (3.13). Since $\mathbb{E}(Y|\boldsymbol{X}=\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}^*$, we have

$$\boldsymbol{\Sigma}_{\boldsymbol{X},Y} = \mathbb{E}(\boldsymbol{X}Y) = \mathbb{E}(\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{\beta}^*) = \boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{\beta}^*.$$

Hence we have

$$\begin{aligned}\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty &= \big\|\widehat{\mathbf{K}}_{\boldsymbol{X},Y} - \boldsymbol{\Sigma}_{\boldsymbol{X},Y} + \boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{\beta}^* - \widehat{\mathbf{K}}_{\boldsymbol{X}}\boldsymbol{\beta}^*\big\|_\infty \\ &\le \big\|\widehat{\mathbf{K}}_{\boldsymbol{X},Y} - \boldsymbol{\Sigma}_{\boldsymbol{X},Y}\big\|_\infty + \big\|\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{\beta}^* - \widehat{\mathbf{K}}_{\boldsymbol{X}}\boldsymbol{\beta}^*\big\|_\infty.\end{aligned}$$

Before we upper bound the two terms on the right-hand side, we establish a concentration inequality for $\widehat{\mathbf{K}}_{\boldsymbol{Z}}$. Let $\mathbf{D}_{\boldsymbol{Z}} = \mathrm{diag}(\sigma_1,\ldots,\sigma_{d+1})$ and $\widehat{\mathbf{D}}_{\boldsymbol{Z}} = \mathrm{diag}(\widehat{\sigma}_1,\ldots,\widehat{\sigma}_{d+1})$, where $\sigma_1,\ldots,\sigma_{d+1}$ are the marginal standard deviations of $\boldsymbol{Z} \in \mathbb{R}^{(d+1)} = (Y,\boldsymbol{X})^T$ while $\widehat{\sigma}_1,\ldots,\widehat{\sigma}_{d+1}$ are the corresponding Catoni's estimators defined in (D.3). We have

$$\boldsymbol{\Sigma}_{\boldsymbol{Z}} = \mathbf{D}_{\boldsymbol{Z}}\boldsymbol{\Sigma}^0_{\boldsymbol{Z}}\mathbf{D}_{\boldsymbol{Z}}, \quad \widehat{\mathbf{K}}_{\boldsymbol{Z}} = \widehat{\mathbf{D}}_{\boldsymbol{Z}}\widehat{\mathbf{R}}_{\boldsymbol{Z}}\widehat{\mathbf{D}}_{\boldsymbol{Z}},$$

where $\widehat{\mathbf{R}}_{\boldsymbol{Z}}$ is the rank-based estimator of the generalized correlation matrix $\boldsymbol{\Sigma}^0_{\boldsymbol{Z}}$ defined in (2.6). Han and Liu (2012) proved that, with probability at least at least $1-(d+1)^{-5/2}$,

$$\big\|\widehat{\mathbf{R}}_{\boldsymbol{Z}} - \boldsymbol{\Sigma}^0_{\boldsymbol{Z}}\big\|_{\max} \le C\sqrt{\frac{\log(d+1)}{n}},$$

where $\|\mathbf{M}\|_{\max} = \max_{1\le i,j\le d}\{|M_{i,j}|\}$ for $\mathbf{M}\in\mathbb{R}^{d\times d}$. We have

$$\begin{aligned}
&\big\|\widehat{\mathbf{D}}_{\boldsymbol{Z}}\widehat{\mathbf{R}}_{\boldsymbol{Z}}\widehat{\mathbf{D}}_{\boldsymbol{Z}} - \mathbf{D}_{\boldsymbol{Z}}\boldsymbol{\Sigma}^0_{\boldsymbol{Z}}\mathbf{D}_{\boldsymbol{Z}}\big\|_{\max} \qquad (D.14)\\
&= \big\|\mathbf{D}_{\boldsymbol{Z}}(\widehat{\mathbf{R}}_{\boldsymbol{Z}} - \boldsymbol{\Sigma}^0_{\boldsymbol{Z}})\mathbf{D}_{\boldsymbol{Z}} + (\widehat{\mathbf{D}}_{\boldsymbol{Z}} - \mathbf{D}_{\boldsymbol{Z}})\widehat{\mathbf{R}}_{\boldsymbol{Z}}\mathbf{D}_{\boldsymbol{Z}} + \widehat{\mathbf{D}}_{\boldsymbol{Z}}\widehat{\mathbf{R}}_{\boldsymbol{Z}}(\widehat{\mathbf{D}}_{\boldsymbol{Z}} - \mathbf{D}_{\boldsymbol{Z}})\big\|_{\max}\\
&\le \big\|\mathbf{D}_{\boldsymbol{Z}}(\widehat{\mathbf{R}}_{\boldsymbol{Z}} - \boldsymbol{\Sigma}^0_{\boldsymbol{Z}})\mathbf{D}_{\boldsymbol{Z}}\big\|_{\max} + \big\|(\widehat{\mathbf{D}}_{\boldsymbol{Z}} - \mathbf{D}_{\boldsymbol{Z}})\widehat{\mathbf{R}}_{\boldsymbol{Z}}\mathbf{D}_{\boldsymbol{Z}}\big\|_{\max} + \big\|\widehat{\mathbf{D}}_{\boldsymbol{Z}}\widehat{\mathbf{R}}_{\boldsymbol{Z}}(\widehat{\mathbf{D}}_{\boldsymbol{Z}} - \mathbf{D}_{\boldsymbol{Z}})\big\|_{\max}\\
&\le \|\mathbf{D}_{\boldsymbol{Z}}\|^2_{\max}\big\|\widehat{\mathbf{R}}_{\boldsymbol{Z}} - \boldsymbol{\Sigma}^0_{\boldsymbol{Z}}\big\|^2_{\max} + \|\mathbf{D}_{\boldsymbol{Z}}\|_{\max}\big\|\widehat{\mathbf{D}}_{\boldsymbol{Z}} - \mathbf{D}_{\boldsymbol{Z}}\big\|_{\max} + \big\|\widehat{\mathbf{D}}_{\boldsymbol{Z}}\big\|_{\max}\big\|\widehat{\mathbf{D}}_{\boldsymbol{Z}} - \mathbf{D}_{\boldsymbol{Z}}\big\|_{\max}.
\end{aligned}$$

Following similar arguments in Corollary D.2, we have

$$\big\|\widehat{\mathbf{D}}_{\boldsymbol{Z}} - \mathbf{D}_{\boldsymbol{Z}}\big\|_{\max} \le C\sqrt{\frac{\log(d+1)}{n}}, \quad \big\|\widehat{\mathbf{D}}_{\boldsymbol{Z}}\big\|_{\max} \le \big\|\mathbf{D}_{\boldsymbol{Z}}\big\|_{\max} + C\sqrt{\frac{\log(d+1)}{n}}$$

with probability at least $1-2(d+1)^{-3}$. We assume that σ_j $(1\le j\le d+1)$ is upper bounded, from (D.14) we have, with probability at least $1-(d+1)^{-5/2}-2(d+1)^{-3}$,

$$\big\|\boldsymbol{\Sigma}_{\boldsymbol{Z}} - \widehat{\mathbf{K}}_{\boldsymbol{Z}}\big\|_{\max} \le C\sqrt{\frac{\log(d+1)}{n}},$$

which implies that with the same probability,

$$\begin{aligned}
&\big\|\widehat{\mathbf{K}}_{\boldsymbol{X},Y} - \boldsymbol{\Sigma}_{\boldsymbol{X},Y}\big\|_{\infty} \le C\sqrt{\frac{\log(d+1)}{n}},\\
&\big\|\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{\beta}^* - \widehat{\mathbf{K}}_{\boldsymbol{X}}\boldsymbol{\beta}^*\big\|_{\infty} \le \|\boldsymbol{\beta}^*\|_1\big\|\boldsymbol{\Sigma}_{\boldsymbol{X}} - \widehat{\mathbf{K}}_{\boldsymbol{X}}\big\|_{\max} \le C\|\boldsymbol{\beta}^*\|_1\sqrt{\frac{\log(d+1)}{n}}.
\end{aligned}$$

Then we reach the conclusion. □

References

Agarwal, A., Negahban, S. and Wainwright, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics* **40** 2452–2482.

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

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

Bunea, F., Tsybakov, A. and Wegkamp, M. (2007). Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics* **1** 169–194.

Candés, E. and Tao, T. (2007). The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics* 2313–2351.

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

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **48** 1148–1185.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *The Annals of statistics* **32** 407–499.

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. (2012). Strong oracle optimality of folded concave penalized estimation. *arXiv preprint arXiv:1210.5992* .

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

Garrigues, P. and Ghaoui, L. (2008). An homotopy algorithm for the lasso with online observations. In *Neural Information Processing Systems (NIPS)*, vol. 21. Citeseer.

Gärtner, B., Jaggi, M. and Maria, C. (2012). An exponential lower bound on the complexity of regularization paths. *Journal of Computational Geometry* **3** 168–195.

Hale, E. T., Yin, W. and Zhang, Y. (2008). Fixed-point continuation for ℓ_1-minimization: Methodology and convergence. *SIAM Journal on Optimization* **19** 1107–1130.

Han, F. and Liu, H. (2012). Transelliptical component analysis. In *Advances in Neural Information Processing Systems 25*.

Han, F. and Liu, H. (2013). Optimal rates of convergence of transelliptical component analysis. Tech. rep., Department of Operation Research and Financial Engineering, Princeton University.

Hastie, T., Rosset, S., Tibshirani, R. and Zhu, J. (2005). The entire regularization path for the support vector machine. *Journal of Machine Learning Research* **5** 1391.

Hunter, D. R. and Li, R. (2005). Variable selection using mm algorithms. *Annals of statistics* **33** 1617.

Koltchinskii, V. (2009a). The dantzig selector and sparsity oracle inequalities. *Bernoulli* **15** 799–828.

Koltchinskii, V. (2009b). Sparsity in penalized empirical risk minimization. *Ann. Inst. H. Poincaré Probab. Statist* **45** 7–57.

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

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics* **40** 1637–1664.

Loh, P.-L. and Wainwright, M. J. (2013). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *arXiv preprint arXiv:1305.2436* .

Luo, Z.-Q. and Tseng, P. (1992). On the linear convergence of descent methods for convex essentially smooth minimization. *SIAM Journal on Control and Optimization* **30** 408–425.

Mairal, J. and Yu, B. (2012). Complexity analysis of the lasso regularization path. *arXiv preprint arXiv:1205.0079* .

Mazumder, R., Friedman, J. H. and Hastie, T. (2011). Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association* **106**.

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

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

Nesterov, Y. (2007). Gradient methods for minimizing composite functions. *preprint* .

Osborne, M. R., Presnell, B. and Turlach, B. A. (2000). On the lasso and its dual. *Journal of Computational and Graphical statistics* **9** 319–337.

Park, M. Y. and Hastie, T. (2007). ℓ_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69** 659–677.

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

Raskutti, G., Wainwright, M. J. and Yu, B. (2010). Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research* **99** 2241–2259.

Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *The Annals of Statistics* 1012–1030.

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

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

van de Geer, S. (2000). *Empirical processes in M-estimation*, vol. 45. Cambridge university press.

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

van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics* **3** 1360–1392.

Wainwright, M. (2009). Sharp thresholds for high dimensional and noisy sparsity recovery using $\ell_1$constrained quadratic programming. *IEEE Transactions on Information Theory* **55** 2183–2201.

Wen, Z., Yin, W., Goldfarb, D. and Zhang, Y. (2010). A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization, and continuation. *SIAM Journal on Scientific Computing* **32** 1832–1857.

Wright, S., Nowak, R. and Figueiredo, M. (2009). Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing* **57** 2479–2493.

Xiao, L. and Zhang, T. (2012). A proximal-gradient homotopy method for the sparse least-squares problem. *arXiv preprint arXiv:1203.3002* .

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. *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 ℓ_1 regularization. *The Annals of Statistics* **37** 2109–2144.

Zhang, T. (2010b). Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research* **11** 1087–1107.

Zhang, T. (2012). Multistage convex relaxation for feature selection. *Bernoulli* To appear.

Zhao, P. and Yu, B. (2007). Stagewise lasso. *The Journal of Machine Learning Research* **8** 2701–2726.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics* **36** 1509.