Nonconvex Regularized Robust Regression with Oracle Properties in Polynomial Time

Xiaoou Pan∗, Qiang Sun† and Wen-Xin Zhou‡

Abstract

This paper investigates tradeoffs among optimization errors, statistical rates of convergence and the effect of heavy-tailed random errors for high-dimensional adaptive Huber regression with nonconvex regularization. When the additive errors in linear models have only bounded second moment, our results suggest that adaptive Huber regression with nonconvex regularization yields statistically optimal estimators that satisfy oracle properties as if the true underlying support set were known beforehand. Computationally, we need at most $O(\log s + \log \log d)$ convex relaxations to reach such oracle estimators, where $s$ and $d$ denote the sparsity and ambient dimension, respectively. Numerical studies lend strong support to our methodology and theory.

Keywords: Adaptive Huber regression, heavy-tailed noise, nonconvex regularization, optimization error, oracle property, statistical rate, tradeoff.

1 Introduction

Suppose we have collected independent and identically distributed (i.i.d.) copies \{\((y_i, x_i) : 1 \leq i \leq n\)\} of \((y, x)\) that follows the linear model

$$y = \beta_0^* + x^T \beta^* + \varepsilon,$$

where \(\beta^* = (\beta_1^*, \ldots, \beta_d^*)^T\) is a \(d\)-dimensional vector of coefficients, \(\beta_0^*\) is the intercept, \(\varepsilon = (\varepsilon, \ldots, \varepsilon)^T\) is a \(d\)-dimensional vector of error terms satisfying \(E(\varepsilon|x) = 0\). Our model includes the heteroscedastic case as a special case where \(\varepsilon = \sigma(x)e\), and \(\sigma(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}_+\) is a positive function and the random variable \(e\) is independent of \(x\) and satisfies \(E(e) = 0\). For simplicity, we use \(\theta^* = (\beta_0^*, \beta^*)^T \in \mathbb{R}^{d+1}\) to denote the vector

∗Department of Mathematics, University of California, San Diego, La Jolla, CA 92093, USA. E-mail: xip024@ucsd.edu.
†Department of Statistical Sciences, University of Toronto, Toronto, ON M5S 3G3, Canada. E-mail: qsun@utstat.toronto.edu.
‡Department of Mathematics, University of California, San Diego, La Jolla, CA 92093, USA. E-mail: wez243@ucsd.edu.
of unknown parameters, and write \( z_i = (1, x_i^\top)^\top \) for \( i = 1, \ldots, n \) and \( z = (1, x^\top)^\top \). Of particular interest is the case where the error term only has a bounded second moment.

We are interested in the high-dimensional regime, where the number of features \( d \) is larger than the sample size \( n \) and \( \beta^\star \) is sparse. Since the invention of Lasso two decades ago (Tibshirani, 1996; Santosa and Symes, 1986), a variety of variable selection methods have been developed for finding a small group of covariates that are associated with the response from a large pool. The Lasso estimator \( \hat{\theta}_{\text{Lasso}} \) solves the convex optimization problem

\[
\arg\min_{\theta = (\beta_0, \beta^\top)^\top \in \mathbb{R}^{d+1}} \left\{ \frac{1}{2n} \| y - Z\theta \|_2^2 + \lambda \| \beta \|_1 \right\},
\]

where \( \lambda > 0 \) is a tuning parameter and \( Z = (z_1, \ldots, z_n)^\top \). Essentially, the Lasso is an \( \ell_1 \)-regularized least squares method: the quadratic loss is used as a goodness of fit measure and the \( \ell_1 \)-norm induces sparsity. To achieve better performance under different circumstances, several Lasso variants have been proposed and studied; see, Fan and Li (2001), Zou and Hastie (2005), Zou (2006), Yuan and Lin (2006), Belloni, Chernozhukov and Wang (2011), Sun and Zhang (2012) and Bogdan et al. (2015), to name a few. We refer to Bühlmann and van de Geer (2011) and Hastie, Tibshirani and Wainwright (2015) for comprehensive reviews of high-dimensional statistical methods and theory.

As a general regression analysis method, the Lasso, along with many of its variants, has two potential downsides. First, the regularized least squares methods are sensitive to the tails of error distributions, even though various alternative penalties have been proposed to achieve better model selection performance. Consider a Lasso-type estimator that solves the penalized empirical risk minimization

\[
\min_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(y_i - z_i^\top \theta) + \lambda \| \beta \|_1 \right\},
\]

where \( \ell(x) : \mathbb{R} \mapsto [0, \infty) \) is a general loss function. The effects of the loss and noise on estimation error are coded in the vector \( \{\ell'(\varepsilon_i)\}_{i=1}^n \). If \( \ell \) is the quadratic loss, this vector is likely to have relatively many large coordinates when \( \varepsilon \) is heavy-tailed (Mendelson, 2018). As a result, the combination of the rapid growth of \( \ell \) with heavy-tailed sampling distribution inevitably leads to outliers, which will eventually be translated into spurious discoveries. Second, it has been recognized (Fan and Li, 2001; Zou, 2006) that convex penalties, typified the \( \ell_1 \)-norm, introduce nonnegligible estimation bias. Due to the bias of the \( \ell_1 \) penalty, the Lasso typically selects far larger model size since the visible bias in Lasso forces the cross-validation procedure to choose a smaller value of \( \lambda \).

In the presence of heavy-tailed noise, outliers occur more frequently and may have a significant impact on (regularized) empirical risk minimization when the loss grows quickly. To reduce the effects of outliers, a widely acknowledged strategy is to use a robust loss function that is globally Lipschitz continuous and locally quadratic. A prototypical example is the Huber loss (Huber, 1964):

\[
\ell_\tau(x) = \begin{cases} 
\frac{x^2}{2} & \text{if } |x| \leq \tau, \\
\tau|x| - \tau^2/2 & \text{if } |x| > \tau.
\end{cases}
\]
The Huber loss $\ell_\tau(\cdot)$ is parametrized by $\tau > 0$, which is referred to as the robustification parameter that controls the tradeoff between the robustness and bias. The second important issue is the choice of sparsity-inducing penalty. In order to eliminate the nonnegligible estimation bias introduced by convex regularization, Fan and Li (2001) introduced a family of folded-concave penalties, including the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), minimax concave penalty (MCP) (Zhang, 2010a), and the capped $\ell_1$-penalty (Zhang, 2010b). These ideas motivate us to propose the nonconvex regularized robust M-estimator of the form

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^{d+1}} \left\{ L_\tau(\theta) + \sum_{j=1}^{d} p_\lambda(\beta_j) \right\},$$

(4)

where $\theta = (\beta_0, \beta_1, \ldots, \beta_d)^T$, $L_\tau(\theta) := (1/n) \sum_{i=1}^{n} \ell_\tau(y_i - z_i^\top \theta) = (1/n) \sum_{i=1}^{n} \ell_\tau(y_i - \beta_0 - x_i^\top \beta)$ is the empirical loss function, $\tau > 0$ is a robustification parameter, and $p_\lambda : \mathbb{R} \mapsto [0, \infty)$ is a concave penalty function with a regularization parameter $\lambda > 0$. We refer to Zhang and Zhang (2012) for a survey of concave regularized methods.

In practice, it is inherently difficult to solve the nonconvex optimization problem (4) directly. Theoretically, statistical properties, such as the rate of convergence under various norms and oracle properties, are established for either the hypothetical global optimum that is unobtainable by any practical algorithm in polynomial time, or a local optimum that exists somewhere like a needle in a haystack. To close the gap between statistical theory and computational complexity, we propose a two-stage (contraction and tightening) regularized robust regression procedure, which yields a solution with desired oracle properties and is computationally efficient as it only involves solving a sequence of adaptive convex programs. Our work builds upon Fan et al. (2018), who studied the tradeoff between algorithmic complexity and statistical error when fitting high-dimensional models with a general but nonrobust loss function. The aim of this paper is to explore robust loss functions, not merely for the purpose of generality but owing to a real downside of the squared loss. Typified by the Huber loss, our general principle applies to a class of robust loss functions as will be discussed in Section 4. Software implementing the proposed procedure and reproducing our computational results is available at https://github.com/XiaoouPan/ILAMM.

1.1 Related literature

For classical linear models with heavy-tailed errors, Mendelson (2018) and Sun, Zhou and Fan (2017) studied the performance of empirical risk minimization (ERM) with a convex robust loss function; the former considered a class of sufficiently smooth convex loss functions and the latter focused on the Huber loss. From a non-asymptotic viewpoint, a well-chosen loss, calibrated to fit the noise level, sample size and dimensionality of the problem, alleviates the ill-effects of outliers. For nonconvex loss functions, typified by Tukey’s bisquare loss, Mei, Bai and Montanari (2018) investigated the statistical consistency of $\ell_2$-constrained M-estimators to stationary points.

In the high-dimensional regime that $d \gg n$, Minsker (2015) and Fan, Li and Wang (2017), respectively, proposed a median-of-means estimator based on Lasso and $\ell_1$-regularized Huber’s M-estimator. Both estimators achieve sub-Gaussian deviation bounds when the
regression error only has finite variance. Loh (2017) studied statistical consistency and asymptotic normality of nonconvex regularized robust $M$-estimators. Assuming a symmetric error distribution, the author proved that there exists a stationary point with $\ell_2$ and $\ell_1$ error bounds in the order of $s\log(d)/n$ and $s\sqrt{\log(d)/n}$, respectively; and then established uniqueness of the stationary point. Loh (2017) applied a projected gradient algorithm to solve an $\ell_1$-constrained optimization problem. In this paper, we simultaneously analyze the statistical property and algorithmic complexity of the solutions produced by our algorithm. Specifically, we show that at most $O(\log s + \log \log d)$ iterations are needed to deliver a statistically optimal estimator with $\ell_2$ and $\ell_1$ errors in the order of $s/n$ and $s/\sqrt{n}$, respectively.

1.2 Notation

Let us summarize our notation. For every integer $k \geq 1$, we use $\mathbb{R}^k$ to denote the the $k$-dimensional Euclidean space. The inner and Hadamard products of any two vectors $u = (u_1, \ldots, u_k)^T, v = (v_1, \ldots, v_k)^T \in \mathbb{R}^k$ are defined by $u^T v = \langle u, v \rangle = \sum_{i=1}^k u_i v_i$ and $u \circ v = (u_1 v_1, \ldots, u_k v_k)^T$, respectively. We use $\| \cdot \|_p$ ($1 \leq p \leq \infty$) to denote the $\ell_p$-norm in $\mathbb{R}^k$: $\|u\|_p = (\sum_{i=1}^k |u_i|^p)^{1/p}$ and $\|u\|_\infty = \max_{1 \leq i \leq k} |u_i|$. Moreover, we write $||u||_{\min} = \min_{1 \leq i \leq k} |u_i|$. For $k \geq 2$, $S^{k-1} = \{ u \in \mathbb{R}^k : \|u\|_2 = 1 \}$ denotes the unit sphere in $\mathbb{R}^k$. For any function $f : \mathbb{R} \mapsto \mathbb{R}$ and vector $u = (u_1, \ldots, u_k)^T \in \mathbb{R}^k$, we write $f(u) = (f(u_1), \ldots, f(u_k))^T \in \mathbb{R}^k$.

Throughout this paper, we use bold capital letters to represent matrices. For $k \geq 2$, $I_k$ represents the identity/unit matrix of size $k$. For any $k \times k$ symmetric matrix $A \in \mathbb{R}^{k \times k}$, $\| A \|_2$ is the operator norm of $A$, and we use $\lambda_{\min} A$ and $\lambda_{\max} A$ to denote the minimal and maximal eigenvalues of $A$, respectively. For a positive semidefinite matrix $A \in \mathbb{R}^{k \times k}$, $\| \cdot \|_A$ denotes the norm linked to $A$ given by $\|u\|_A = \|A^{1/2}u\|_2$, $u \in \mathbb{R}^k$. For any two real numbers $a$ and $b$, we write $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For two sequences of non-negative numbers $\{a_n\}_{n \geq 1}$ and $\{b_n\}_{n \geq 1}$, $a_n \lesssim b_n$ indicates that there exists a constant $C > 0$ independent of $n$ such that $a_n \leq C b_n$; $a_n \gtrsim b_n$ is equivalent to $b_n \lesssim a_n$; $a_n \asymp b_n$ is equivalent to $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For two numbers $C_1$ and $C_2$, we write $C_2 = C_2(C_1)$ if $C_2$ depends only on $C_1$. For any integer $d \geq 1$, we write $[d] = \{1, \ldots, d\}$. For any set $S$, we use $|S|$ to denote its cardinality, i.e. the number of elements in $S$.

2 Regularized Huber $M$-estimation

In this section, we first revisit the $\ell_1$-penalized Huber estimator in Section 2.1, and point out two interesting regimes for the robustification parameter $\tau$. In Section 2.2, we propose a two-stage procedure, which is closely related to nonconvex regularized Huber regression, for fitting high-dimensional sparse models with heavy-tailed noise. This two-stage robust regression method not only is computationally efficient, but also achieves optimal rate of convergence and oracle properties, as will be studied in Sections 2.3 and 2.4. Throughout, $S = \text{supp}(\beta^*) = \{ j : \beta^*_j \neq 0 \}$ denotes the active set and $s = |S|$ is the sparsity.
2.1 $\ell_1$-regularized Huber regression

Given i.i.d. observations $\{(y_i, x_i)\}_{i=1}^n$ from the linear model (1), consider the $\ell_1$-regularized Huber $M$-estimator, which we refer to as the Huber-Lasso,

$$\hat{\theta}^{H-\text{Lasso}} \in \arg\min_{\theta = (\beta_0, \beta^\top) \in \mathbb{R}^{d+1}} \{L_r(\theta) + \lambda \|\beta\|_1\},$$

(5)

where $L_r(\cdot)$ is the empirical loss function as in (4). Statistical properties of the penalized Huber $M$-estimator have been studied by Lambert-Lacroix and Zwald (2011), Fan, Li and Wang (2017) and Loh (2017) from different perspectives. A less-noticed problem, however, is the connection between the robustification parameter and the error distribution, which in turn quantifies the tradeoff between robustness and unbiasedness. In fact, recent studies (Sun, Zhou and Fan, 2017) reveal that the use of Huber loss is particularly suited for heavy-tailed problems in both low and high dimensions. With a well-chosen robustification parameter, calibrated by the noise level and the intrinsic dimensionality of the problem, the effects of the heavy-tailed noise can be removed or dampened.

For estimation purpose, $\hat{\theta}^{H-\text{Lasso}}$ is a regularized $M$-estimator of

$$\theta^*_r = (\beta^*_{0, r}, \beta^*_{r}^\top)^\top = \arg\min_{\theta = (\beta_0, \beta^\top) \in \mathbb{R}^{d+1}} E L_r(\theta).$$

(6)

The true parameter $\theta^* := (\beta^*_0, \beta^*_r^\top)^\top$, however, is identified as $\arg\min_{\beta_0, \beta} \sum_{i=1}^n E (y_i - \beta_0 - x_i^\top \beta)^2$. It is easy to see that, if the distribution of $\varepsilon$ is symmetric around zero, $\theta^*_r = \theta^*$ for any $\tau > 0$. This is not always true with general asymmetric noise variables. To have a better understanding of the role of $\tau$ in the Huber loss, in the next result we establish the connection between $\theta^*$ and $\theta^*_r$ in a fairly general context.

**Proposition 1.** Assume that $\varepsilon$ and $x$ are independent, and the function $\alpha \mapsto E\{\ell_r(\varepsilon - \alpha)\}$ has a unique minimizer, denoted by $\alpha_r = \arg\min_{\alpha \in \mathbb{R}} E\{\ell_r(\varepsilon - \alpha)\}$, satisfying $P(|\varepsilon - \alpha_r| \leq \tau) > 0$. Assume further that $E(\varepsilon \varepsilon^\top)$ is positive definite. Then

$$\beta^*_0, r = \beta^*_0 + \alpha_r \quad \text{and} \quad \beta^*_r = \beta^*.$$

(7)

For Huber regression in low dimensions (e.g. $d$ is fixed or $d$ slowly grows as a function of $n$), the calibration of $\tau$ is typically determined by the asymptotic 95% efficiency rule. While in the absence of symmetry assumption, Proposition 1 demonstrates that $\alpha_r$ also quantifies the approximation bias induced by the Huber loss, and its magnitude decays as $\tau$ increases (Wang et al., 2018). Therefore, throughout this paper we study the statistical properties of regularized Huber $M$-estimators under two scenarios: diverging $\tau$ and fixed $\tau$.

To begin with, we impose the following assumption on the data generating process. The covariate (random) vectors are assumed to be sub-Gaussian (Vershynin, 2012), while we allow the regression errors to be heavy-tailed and/or asymmetric.

**Condition 1.** There exists $\nu_\varepsilon > 0$ such that $P(|\langle u, z \rangle| \geq \nu_\varepsilon \|u\|_{\Sigma} \cdot t) \leq 2 e^{-t^2}$ for all $u \in \mathbb{R}^{d+1}$, where $\Sigma = (\sigma_{jk})_{0 \leq j, k \leq d} = E(z z^\top)$ is positive definite with $\sigma^2_{zz} := \max_{0 \leq j \leq d} \sigma_{jj}$. The regression error $\varepsilon$ satisfies $E(\varepsilon | x) = 0$ and $E(\varepsilon^2 | x) = \sigma^2_{\varepsilon}$ almost surely.

**Theorem 1.** Assume that Condition 1 holds.
(I) (Diverging $\tau$) Any optimal solution $\hat{\theta}^{\text{H-Lasso}}$ to the convex program (5) with $\tau = \sigma_z \sqrt{n/\log d}$ and $\lambda$ scaling as $\nu_z \sigma_z \sigma_\varepsilon \sqrt{\log(d)/n}$ satisfies
\[
\|\hat{\theta}^{\text{H-Lasso}} - \theta^*\|_2 \lesssim \Delta \lambda^{-1} s^{1/2} \lambda \quad \text{and} \quad \|\hat{\theta}^{\text{H-Lasso}} - \theta^*\|_1 \lesssim \Delta \lambda^{-1} s \lambda
\]
with probability at least $1 - 5d^{-1}$ as long as $n \geq c_1 s \log d$, where $c_1 > 0$ is a constant depending only on $(\nu_z, \sigma_z, \Delta \Sigma)$.

(II) (Fix $\tau$) Assume further that $\varepsilon$ and $x$ are independent. Let $\tau > 0$ be such that the function $\alpha \mapsto \mathbb{E}\{\ell_r(\varepsilon - \alpha)\}$ has a unique minimizer $\alpha_\tau = \text{argmin}_{\alpha \in \mathbb{R}} \mathbb{E}\{\ell_r(\varepsilon - \alpha)\}$ and that $\rho_\tau := \mathbb{P}(|\varepsilon - \alpha_\tau| \leq \tau/2) > 0$. Then any optimal solution $\hat{\theta}^{\text{H-Lasso}}$ with $\tau \asymp \sigma_z$ and $\lambda \asymp \nu_z \sigma_z \sigma_\varepsilon \sqrt{\log(d)/n}$ satisfies
\[
\|\hat{\theta}^{\text{H-Lasso}} - \theta^*_\tau\|_2 \lesssim \rho_\tau^{-1} \Delta \lambda^{-1} s^{1/2} \lambda \quad \text{and} \quad \|\hat{\theta}^{\text{H-Lasso}} - \theta^*_\tau\|_1 \lesssim \rho_\tau^{-1} \Delta \lambda^{-1} s \lambda
\]
with probability at least $1 - 5d^{-1}$ as long as $n \geq c_2 \rho_\tau^{-3} s \log d$, where $\theta^*_\tau = (\beta^*_0 + \alpha_\tau, \beta^*_\tau)^T$ and $c_2 > 0$ is a constant depending only on $(\nu_z, \sigma_z, \Delta \Sigma)$.

### 2.2 Two-stage procedure: tightening after contraction

For sparse linear regression, it is known that regularized $M$-estimators with convex penalties typically exhibit a suboptimal statistical rate of convergence, as compared to the oracle rate achieved by nonconvex regularization. However, as noted previously, directly solving the nonconvex optimization problem (4) is computationally challenging. Moreover, statistical properties are only established for the hypothetical global optimum (or some stationary point), which is typically unobtainable by any polynomial time algorithm.

Inspired by the local linear approximation to the folded concave penalty (Zou and Li, 2008), here we propose a two-stage procedure, contraction and tightening, that solves a sequence of convex programs up to a prespecified optimization precision. Let $p_\lambda(\cdot)$ be a differentiable penalty function as in (4) and recall that $L_\tau(\cdot)$ is the empirical loss function. Starting with an initial estimate $\hat{\theta}^{(0)} = (\beta_0^{(0)}, \beta_1^{(0)}, \ldots, \beta_d^{(0)})^T$, consider a sequence of convex optimization programs $\{(P_\ell)\}_{\ell \geq 1}$:
\[
\min_{\theta=(\beta_0, \beta_1, \ldots, \beta_d)^T} \left\{ \mathcal{L}_\tau(\theta) + \sum_{j=1}^d p_\lambda(|\beta_j^{(\ell-1)}|)|\beta_j| \right\} \quad (P_\ell) \tag{10}
\]
for $\ell = 1, 2, \ldots$, where $\hat{\theta}^{(\ell)} = (\beta_0^{(\ell)}, \beta_1^{(\ell)}, \ldots, \beta_d^{(\ell)})^T$ is the optimal solution to program $(P_\ell)$. Following Zhang and Zhang (2012), we assume the following conditions on the penalty function $p_\lambda$.

**Condition 2.** The penalty function $p_\lambda$ is of the form $p_\lambda(t) = \lambda^2 p(t/\lambda)$ for $t \in \mathbb{R}$, where $p : \mathbb{R} \mapsto [0, \infty)$ satisfies: (i) $p(t) = p(-t)$ for all $t$ and $p(0) = 0$; (ii) $p$ is nondecreasing on $[0, \infty)$; (iii) $p$ is differentiable almost everywhere on $(0, \infty)$ and $\lim_{t \to 0^+} p'(t) = 1$; (iv) $p'(t_1) \leq p'(t_2)$ for all $t_1 \geq t_2 \geq 0$.

For each $\ell \geq 1$, program $(P_\ell)$ corresponds to a weighted $\ell_1$-regularized empirical Huber loss minimization of the form
\[
\min_{\theta=(\beta_0, \beta^T)^T} \left\{ \mathcal{L}_\tau(\theta) + \|\lambda \circ \beta\|_1 \right\}, \tag{11}
\]
where $\lambda = (\lambda_1, \ldots, \lambda_d)^T$ be a $d$-vector of regularization parameters with $\lambda_j \geq 0$. By convex optimization theory, any optimal solution $\tilde{\theta} = (\tilde{\beta}_0, \tilde{\beta}^T)^T$ to the convex program (11) satisfies the first-order optimality condition

$$\partial_{\beta_0} \mathcal{L}_r(\tilde{\theta}) = 0, \quad \nabla_{\beta} \mathcal{L}_r(\tilde{\theta}) + \lambda \circ \xi = 0_d \quad \text{for some} \quad \xi \in \partial ||\tilde{\beta}||_1 \in [-1, 1]^d,$$

where we use the notations $\nabla \mathcal{L}_r(\theta) = (\partial_{\beta_0} \mathcal{L}_r(\theta), \nabla_{\beta} \mathcal{L}_r(\theta))^T \in \mathbb{R}^{d+1}$ with

$$\partial_{\beta_0} \mathcal{L}_r(\theta) = -\frac{1}{n} \sum_{i=1}^n \ell'_r(y_i - z_i^T \theta) \quad \text{and} \quad \nabla_{\beta} \mathcal{L}_r(\theta) = -\frac{1}{n} \sum_{i=1}^n \ell'_r(y_i - z_i^T \theta)x_i \in \mathbb{R}^d.$$

**Definition 1.** Following the terminology in Fan et al. (2018), for a prespecified tolerance level $\epsilon > 0$, we say $\tilde{\theta}$ is an $\epsilon$-optimal solution to (11) if $\omega_\lambda(\tilde{\theta}) \leq \epsilon$, where

$$\omega_\lambda(\theta) := \max \left\{ ||\partial_{\beta_0} \mathcal{L}_r(\theta)||_1, \min_{\xi \in \partial ||\beta||_1} \| \nabla_{\beta} \mathcal{L}_r(\theta) + \lambda \circ \xi \|_\infty \right\}, \quad \theta \in \mathbb{R}^{d+1}. \quad (12)$$

In view of Definition 1, for a prespecified sequence of tolerance levels $\{\epsilon_\ell\}_{\ell \geq 1}$, we use $\tilde{\theta}^{(\ell)} = (\tilde{\beta}_0^{(\ell)}, \tilde{\beta}_1^{(\ell)}, \ldots, \tilde{\beta}_d^{(\ell)})^T$ to denote an $\epsilon_\ell$-optimal solution to program $(P_\ell)$, that is,

$$\min_{\theta \in (\beta_0, \beta^T)^T} \{ \mathcal{L}_r(\theta) + \| \lambda^{(\ell-1)} \circ \beta ||_1 \},$$

where $\lambda^{(\ell-1)} := p_\lambda'(\tilde{\beta}^{(\ell-1)})$. For simplicity, we consider a trivial initial estimator $\tilde{\theta}^{(0)} = (\tilde{\beta}_0^{(0)}, \tilde{\beta}_1^{(0)}, \ldots, \tilde{\beta}_d^{(0)})^T = 0$. Since $p_\lambda'(||\beta_j^{(0)}||) = p_\lambda'(0) = \lambda$ for $j = 1, \ldots, d$, the program $(P_1)$ coincides with that in (5). In Section 3, we will describe an I-LAMM algorithm which produces an $\epsilon$-optimal solution to (11) after a few iterations.

The above procedure is sequential, and can be categorized into two stages: contraction ($\ell = 1$) and tightening ($\ell \geq 2$). As we will see in the next subsection, even starting with a trivial initial estimator that can be fairly remote from the true parameter, the contraction stage will produce a reasonably good estimator whose statistical error is of the order $\sqrt{\log(d) \cdot s/n}$. Essentially, the contraction stage is equivalent to the $\ell_1$-regularized Huber regression in (5). A tightening stage further refines this coarse contraction estimator consecutively, and eventually gives rise to an estimator that achieves the oracle rate $\sqrt{s/n}$.

### 2.3 Deterministic analysis

To analyze the statistical properties of $\{\tilde{\theta}^{(\ell)}\}_{\ell \geq 1}$, we first introduce a local curvature parameter, which is closely related to the restricted strong convexity property of the empirical Huber loss over a local $\ell_1$-cone.

**Definition 2.** For $r > 0$, $a \geq 1$ and $\vartheta \in \mathbb{R}^{d+1}$, define

$$\kappa_r(a, r, \vartheta) = \inf \left\{ \frac{\langle \nabla \mathcal{L}_r(\theta) - \nabla \mathcal{L}_r(\vartheta), \theta - \vartheta \rangle}{\| \theta - \vartheta \|_2^2} : \theta \in \mathbb{B}_2(r, \vartheta), \theta \in \mathcal{C}(a, \vartheta) \right\},$$

where $\mathbb{B}_2(r, \vartheta) := \{ \theta \in \mathbb{R}^{d+1} : \| \theta - \vartheta \|_2 \leq r \}$ is the $\ell_2$-ball centered at $\vartheta$ with radius $r$, and $\mathcal{C}(a, \vartheta) := \{ \theta \in \mathbb{R}^{d+1} : \| \theta - \vartheta \|_1 \leq a \| \theta - \vartheta \|_2 \}$ is an $\ell_1$-cone.

Throughout the following, we assume that the penalty function $p_\lambda$ satisfies Condition 2.
Proposition 2. Let \((\lambda, \epsilon_1, r)\) satisfy
\[
\lambda \geq 2\{||\nabla L_r(\theta^*)||_\infty + \epsilon_1\} \quad \text{and} \quad r > \frac{1 + \sqrt{2}/2}{\kappa_\tau(4\sqrt{s + 1}, r, \theta^*)} s^{1/2}\lambda. \tag{13}
\]
Then any \(\epsilon_1\)-optimal solution \(\tilde{\theta}^{(1)}\) to (P1) satisfies
\[
||\tilde{\theta}^{(1)} - \theta^*||_2 \leq \frac{1 + \sqrt{2}/2}{\kappa_\tau(4\sqrt{s + 1}, r, \theta^*)} s^{1/2}\lambda. \tag{14}
\]

Proposition 2 is deterministic in the sense that the bound in (14) holds under the assumed scaling regardless of the sampling distribution. Indeed, both \(||\nabla L_r(\theta^*)||_\infty\) and \(\kappa_\tau(4\sqrt{s + 1}, r, \theta^*)\) are random and their magnitudes depend on the data generating process as well as \((n, d)\). In Section 2.4, we will show that under Condition 1,
\[
||\nabla L_r(\theta^*)||_\infty \lesssim \sqrt{\frac{\log d}{n}} \quad \text{and} \quad \kappa_\tau(4\sqrt{s + 1}, r, \theta^*) \text{ is bounded away from zero}
\]
with high probability as long as \(n \gtrsim s \log d\). Consequently, with \((\tau, \lambda, \epsilon_1, r)\) properly tuned, the contraction estimator \(\tilde{\theta}^{(1)}\) satisfies the bound
\[
||\tilde{\theta}^{(1)} - \theta^*||_2 \lesssim \sqrt{s \log d/n} \quad \text{with high probability as long as} \quad n \gtrsim s \log d.
\]

Next, we investigate the statistical properties of \(\{\tilde{\theta}^{(\ell)}\}_{\ell \geq 2}\) in the tightening stage. To refine the suboptimal rate obtained in Proposition 2, we further require a minimum signal strength condition on \(||\beta_S^\delta||_\min = \min_{j \in S} |\beta_j^\delta|\), where \(S = \text{supp}(\beta^*)\). For any subset \(E \subseteq [d]\) with cardinality \(|E| = k\), we write \(\nabla L_r(\theta^*)_E = (\partial_{\beta_k} L_r(\theta^*), (\nabla_{\beta L_r(\theta^*)})_E^\top) \in \mathbb{R}^{k+1}\).

Proposition 3. For a prespecified \(\delta \in (0, 1),\) assume there exists some constant \(\gamma_\delta > 0\) such that \(p'(\gamma_\delta) > 0, ||\beta_S^\delta||_\min \geq \gamma_\delta \lambda\) and \(\lambda \geq \max\{4, \frac{2}{p'(\gamma_\delta)}\}\{||\nabla L_r(\theta^*)||_\infty + \epsilon_\ell\} \) for all \(\ell \geq 1\). Moreover, assume there exists some \(r > \delta \gamma_\delta s^{1/2}\lambda\) such that
\[
\kappa_\tau(b\sqrt{2s}, r, \theta^*) \geq \frac{1 + \sqrt{2}/4}{\delta \gamma_\delta}, \tag{15}
\]
where \(b = 2 + \frac{2}{p'(\gamma_\delta)}\). Then any \(\epsilon_\ell\)-optimal solution \(\tilde{\theta}^{(\ell)}\) (\(\ell \geq 2\)) satisfies
\[
||\tilde{\theta}^{(\ell)} - \theta^*||_2 \leq \delta||\tilde{\theta}^{(\ell-1)} - \theta^*||_2 + \frac{p'_\lambda(||\beta_S^\delta| - \gamma_\delta \lambda)||_2 + ||\nabla L_r(\theta^*)_S||_2 + (s + 1)^{1/2}\epsilon_\ell}{\kappa_\tau(b\sqrt{2s}, r, \theta^*)}. \tag{16}
\]
Furthermore, it holds
\[
||\tilde{\theta}^{(\ell)} - \theta^*||_2 \leq \frac{1 + \sqrt{2}/4}{\kappa_\tau(4\sqrt{s + 1}, r, \theta^*)} \delta^{\ell-1}s^{1/2}\lambda + \frac{p'_\lambda(||\beta_S^\delta| - \gamma_\delta \lambda)||_2 + ||\nabla L_r(\theta^*)_S||_2 + (s + 1)^{1/2}\epsilon_\ell}{(1 - \delta)\kappa_\tau(b\sqrt{2s}, r, \theta^*)}. \tag{17}
\]
Proposition 3 unveils how the tightening stage improves the statistical rate: every tightening step shrinks the estimation error from the previous step by a δ-fraction. The second term on the right-hand side of (16) or (17) dominates the ℓ2-error, and up to a constant factor, consists of three components,

\[ \|p'_{\lambda}(|\beta^*_S| - \gamma \lambda)\|_2, \|\nabla \mathcal{L}_r(\theta^*)_S\|_2 \text{ and } (s + 1)^{1/2} \varepsilon \]

We identify \( \|p'_{\lambda}(|\beta^*_S| - \gamma \lambda)\|_2 \) as the shrinkage bias induced by the penalty function. This explains the limitation of the \( \ell_1 \)-regularized method in which \( p_\lambda(t) = \lambda |t| \) and \( p'_{\lambda}(t) = \lambda \text{sign}(t) \). Intuitively, choosing a proper penalty function \( p_\lambda(\cdot) \) with a decreasing derivative reduces the bias. The second term, \( \|\nabla \mathcal{L}_r(\theta^*)_S\|_2 \), reveals the oracle property. To see this, consider the oracle estimator defined as

\[ \hat{\theta}_{\text{oracle}} = \arg \min_{\theta \in (\beta_0, \beta^T)\mathbb{T}, \text{supp}(\beta) = S} \mathcal{L}_r(\theta) = \arg \min_{\theta \in (\beta_0, \beta^T)\mathbb{T}, \beta_{sc} = 0} \frac{1}{n} \sum_{i=1}^{n} \ell_r(y_i - \beta_0 - x_i^T \beta, S). \]  

(18)

Since \( s = |S| \ll n \), the finite sample theory for the Huber M-estimation in low dimensions (Wang et al., 2018) applies to \( \hat{\theta}_{\text{oracle}} \), indicating that with high probability,

\[ \|\hat{\theta}_{\text{oracle}} - \theta^*\|_2 \lesssim \|\nabla \mathcal{L}_r(\theta^*)_S\|_2. \]

According to Definition 1, the last term \( (s + 1)^{1/2} \varepsilon \) demonstrates the optimization error, which will be discussed in greater detail in Section 3.

The above results provide conditions under which the sequence of estimators \( \{\hat{\theta}^{(t)}\}_{t \geq 1} \) satisfy the contraction property and, meanwhile, fall in a local neighborhood of \( \theta^* \). However, our previous results for the \( \ell_1 \)-regularized method show that, in cases where the regression error is asymmetric and \( \tau \) scales as a constant, the statistical consistency of the resultant estimator is with respect to \( \theta^*_r \) defined in (6) rather than \( \theta^* \). With slight modifications, this fixed \( \tau \) scenario is also covered by the previous theory.

**Proposition 4.** The statements of Propositions 2 and 3 remain valid if \( \theta^* \) is replaced by \( \theta^*_r = (\beta^*_0, \beta^r_T)^T, S = \text{supp}(\beta^*_r) \) and \( s = |S| \).

Another important feature of the proposed procedure is that the resulting estimator satisfies the strong oracle property, as demonstrated by the following result. Let \( \{\hat{\theta}^{(t)}\}_{t \geq 1} \) be any optimal solutions to the convex programs \( \{(P_t)\}_{t \geq 1} \) in (10) with \( \hat{\theta}^{(0)} = 0 \). Similarly to Definition 2, we need the following quantity that quantifies the restricted strong convexity. For \( a \geq 1, r > 0, \vartheta \in \mathbb{R}^{d+1} \) and \( \mathcal{E} \subseteq \mathbb{R}^d \), define

\[ \bar{\kappa}_r(a, r, \vartheta, \mathcal{E}) = \inf \left\{ \frac{\langle \nabla \mathcal{L}_r(\theta) - \nabla \mathcal{L}_r(\theta'), \theta - \theta' \rangle}{\|\theta - \theta'\|_2^2} : \begin{array}{l} \theta, \theta' = (\beta^*_0, \beta^r)^T \in \mathbb{B}_2(r, \vartheta), \|\theta - \theta'\|_1 \leq a, \text{supp}(\beta') = \mathcal{E} \end{array} \right\} \]

**Proposition 5.** For a prespecified \( \delta \in (0, 1) \), assume there exist constants \( \gamma > \gamma_\delta > 0 \) such that \( p'(\gamma_\delta) > 0, p'(\gamma) = 0, \lambda \geq \max\{4, \frac{2}{p''(\gamma_\delta)}\}\|\nabla \mathcal{L}_r(\hat{\theta}_{\text{oracle}})\|_{\infty} \) and \( \|\beta^*_S\|_{\text{min}} \geq (\gamma_\delta + \gamma)\lambda \).

9
Moreover, assume that there exists \( \bar{r} \geq 1.25/(\delta \gamma_d) \) such that \( \tilde{r}_\tau(b\sqrt{s+1}, r, \theta^*, \mathcal{S}) \geq \bar{r} \) for some \( r > 1.25\bar{r}^{-1}s^{1/2} \lambda \) and \( b = 2 + \frac{1}{\rho'/(\gamma_d)} \),

\[
\|\hat{\theta}^{\text{oracle}} - \theta^*\|_2 \leq r \quad \text{and} \quad \|\hat{\theta}^{\text{oracle}} - \theta^*\|_\infty \leq \frac{\lambda}{5\delta \gamma_d}.
\]  

(19)

Then \( \hat{\theta}(\ell) = \hat{\theta}^{\text{oracle}} \) for all \( \ell \geq \log(s^{1/2}/\delta)/\log(1/\delta) \).

Similarly to Proposition 4, the strong oracle property also prevails if the target parameter is \( \theta^*_r \) instead of \( \theta^* \).

**Proposition 6.** The statement of Proposition 5 remains valid if \( \theta^* \) is replaced by \( \theta^*_r = (\beta^*_{0,r}, \beta^*_r)^\top \), \( \mathcal{S} = \text{supp}(\beta^*_r) \) and \( s = |\mathcal{S}| \).

The proofs of Propositions 2, 3 and 5 are provided in the Appendix. Since the proofs of Propositions 4 and 6 are almost identical, we omit the details to avoid repetition.

### 2.4 Random analysis

In this section, we complement the previous deterministic results with probabilistic bounds on three key random quantities, \( \kappa_r(a, r, \theta), \|\nabla L_r(\theta)\|_\infty \) and \( \|\nabla L_r(\theta)\|_2 \), for either \( \theta = \theta^* \) or \( \theta = \hat{\theta} \). Recall that \( z = (1, x)^\top \) and \( \Sigma = (\sigma_{jk})_{0 \leq j, k \leq d} = \mathbb{E}(zz^\top) \) is positive definite. Write \( \mathbb{B}_\Sigma(t, \theta) = \{\theta \in \mathbb{R}^{d+1} : \|\theta - \theta\|_\Sigma \leq t\} \) for \( t \geq 0 \) and \( \theta \in \mathbb{R}^{d+1} \). Moreover, given the true support set \( \mathcal{S} \subseteq [d] \) of \( \beta^* \), we define the following \((s + 1) \times (s + 1)\) principal submatrix of \( \Sigma \):

\[
\mathcal{S} = \mathbb{E}(zz^\top), \quad \text{where} \quad z_S = (1, x_S)^\top \in \mathbb{R}^{s+1}.
\]  

(20)

**Proposition 7.** Assume Condition 1 holds and let \( \nu = \sup_{u \in \mathbb{B}^d}\{\mathbb{E}(u, \Sigma^{-1/2}z)^4\}^{1/4} \).

(I) Let \( \tau, r_0 > 0 \) and \( a \geq 1 \) satisfy

\[
\tau \geq \max(4\sigma_z^2, 8\nu^2r_0) \quad \text{and} \quad n \geq c_0 \lambda^{-1} \left(\nu + \sigma_z \tau / r_0\right)^2a^2 \log d,
\]  

(21)

where \( c_0 > 0 \) is an absolute constant. Then with probability at least \( 1 - d^{-1} \),

\[
\frac{\langle \nabla L_r(\theta) - \nabla L_r(\theta^*), \theta - \theta^* \rangle}{\|\theta - \theta^*\|_\Sigma^2} \geq \frac{1}{2} \quad \text{for all} \quad \theta \in \mathbb{C}(a, \theta^*) \cap \mathbb{B}_\Sigma(r_0, \theta^*). \]  

(22)

(II) Assume further that the assumptions of Theorem 1, (II) are met. Let \( \tau, r_0 > 0 \) and \( a \geq 1 \) satisfy

\[
\tau \geq 8\rho_r^{-1/2} \nu^2r_0 \quad \text{and} \quad n \geq c_0 \rho_r^{-2} \lambda^{-1} (\nu + \sigma_z \tau / r_0)^2a^2 \log d.
\]  

(23)

Then with probability at least \( 1 - d^{-1} \),

\[
\frac{\langle \nabla L_r(\theta) - \nabla L_r(\theta^*_r), \theta - \theta^*_r \rangle}{\|\theta - \theta^*_r\|_\Sigma^2} \geq \frac{\rho_r}{2} \quad \text{for all} \quad \theta \in \mathbb{C}(a, \theta^*_r) \cap \mathbb{B}_\Sigma(r_0, \theta^*_r). \]  

(24)
A direct implication of Proposition 2 is that, provided inequality (22) or (24) holds with $r_0 = \frac{\lambda_0}{2} r$ for some $r > 0$, $\kappa_r(a, r, \theta^*)$ in Definition 2 satisfies $\kappa_r(a, r, \theta^*) \geq \frac{N}{2} / \lambda_0$ or $\kappa_r(a, r, \theta^*) \geq \rho_r \frac{N}{2} / \lambda_0$. The next proposition provides upper bounds on $\|\nabla L_\tau(\theta^*)\|_\infty$ and $\|\nabla L_\tau(\theta^*_\theta)\|_2$ for $\theta = \theta^*$ and $\theta = \theta^*_\tau$.

**Proposition 8.** Assume Condition 1 holds.

(I) With probability at least $1 - 3d^{-1}$,

$$\|\nabla L_\tau(\theta^*)\|_\infty \leq \sigma_\xi \left( 2\nu_\xi \sigma_\xi \sqrt{\frac{2\log d}{n}} + \nu_\xi \tau \frac{\log d}{n} + \frac{\sigma_\xi^2}{\tau} \right).$$

(25)

For any $t > 0$, it holds with probability at least $1 - e^{-t}$ that

$$\|S^{-1/2} \nabla L_\tau(\theta^*_\theta)\|_2 \leq 4\nu_\xi \sigma_\xi \left( 2\nu_\xi \sigma_\xi \sqrt{\frac{2s + 2 + t}{n}} + \nu_\xi \tau \frac{2s + 2 + t}{n} + \frac{\sigma_\xi^2}{\tau} \right),$$

(26)

where $S \in \mathbb{R}^{(s+1) \times (s+1)}$ is given in (20).

(II) Assume further that the assumptions of Theorem 1, (II) are met. Then with probability at least $1 - 3d^{-1}$,

$$\|\nabla L_\tau(\theta^*_\tau)\|_\infty \leq \nu_\xi \sigma_\xi \left( 2\nu_\xi \sigma_\xi \sqrt{\frac{2\log d}{n}} + \nu_\xi \tau \frac{\log d}{n} \right).$$

(27)

For any $t > 0$, it holds with probability at least $1 - e^{-t}$ that

$$\|S^{-1/2} \nabla L_\tau(\theta^*_\tau)\|_2 \leq \nu_\xi \left( 4\nu_\xi \sigma_\xi \sqrt{\frac{2s + 2 + t}{n}} + \nu_\xi \tau \frac{2s + 2 + t}{n} \right).$$

(28)

Together, Propositions 7 and 8 reveal the impact of the robustification parameter on the statistical performance of the resulting estimator. Their proofs closely follow those of Propositions 8 and 9 in Wang et al. (2018), and are omitted. As discussed in Section 2.3 above, the order of $\|\nabla L_\tau(\theta^*_\theta)\|_2$ determines the oracle statistical rate. In Theorem 2, we show that after only a small number of iterations, the proposed procedure leads to an estimator that achieves the oracle rate of convergence.

**Theorem 2.** Assume Conditions 1 and 2 hold. Moreover, for a prespecified $\delta \in (0, 1)$, assume there exists some $\gamma > \gamma_\delta := \frac{2 + \sqrt{2}}{2} \lambda_0^0 \lambda_0$ such that $p'(\gamma_\delta) > 0$, $p'(t) = 0$ for all $t \geq \gamma$ and $\|\theta_\theta\|_{\text{min}} \geq (\gamma_\delta + \gamma) \lambda$.

(I) (Diverging $\tau$) Let

$$\tau \asymp \sigma_\xi \sqrt{\frac{n}{s}}, \quad \lambda \asymp \sigma_\xi \sqrt{\frac{s}{n}} + \frac{\log d}{n}, \quad \max_{1 \leq t \leq T} \epsilon_t \leq \sqrt{\frac{1}{n}} \text{ with } \frac{T}{\log(\log d)} \leq \frac{\log(s \log d)}{\log(1/\delta)}. $$

Then with probability at least $1 - 4d^{-1} - e^{-s}$,

$$\|\theta(T) - \theta^*\|_2 \lesssim \sigma_\xi \sqrt{\frac{s}{n}} \quad \text{and} \quad \|\hat{\theta}(T) - \theta^*\|_1 \lesssim \sigma_\xi \frac{s}{\sqrt{n}},$$

(29)

as long as $n \gtrsim \max\{s \log d, s^{3/2}\}$. 

11
and then compute \( \theta_t \) and compute the value of such an algorithm is non-increasing in each step, since an algorithm first majorizes it by another function \( g \) to minimize a general function \( f \). To minimize a general function \( f \), assume there exists some \( \gamma > \gamma_0 \) and \( \gamma \) and suppose \( \gamma \approx \lambda \) and \( \lambda \approx \log(d) / \log(1/\delta) \). Then with probability at least 1 − 4d − 1 − e^s, that

\[
\theta^{(t)} - \theta^*_s \leq \sigma_x \frac{s}{\sqrt{n}} \quad \text{and} \quad \| \tilde{\theta}^{(t)} - \theta^*_s \|_2 \leq \sigma_x \sqrt{\frac{s}{n}}
\]

(II) Assume further that the conditions of Theorem 1, (II) hold. Then \( \bar{\theta}^{(T)} \) with

\[
\tau \geq \frac{\log d}{n}, \quad \max_{1 \leq t \leq T} \epsilon_t \leq \sqrt{\frac{1}{n}} \quad \text{and} \quad T \geq \frac{\log \log d}{\log(1/\delta)}
\]

satisfies, with probability at least 1 − 4d − 1 − e^s, that

\[
\| \hat{\theta}^{(T)} - \theta^*_s \|_2 \leq \sigma_x \sqrt{\frac{s}{n}} \quad \text{and} \quad \| \tilde{\theta}^{(T)} - \theta^*_s \|_1 \leq \sigma_x \frac{s}{\sqrt{n}}
\]

as long as \( n \geq s \log d \).

We refer to the conclusion of Theorem 2 as the weak oracle property in the sense that the proposed estimator shares the statistical rate of convergence with the oracle \( \hat{\theta}^{\text{oracle}} \) that knows \textit{a priori} the support \( S \) of \( \beta^* \). An even more intriguing result, as revealed by the following theorem, is that our estimator achieves the strong oracle property, namely, it coincides with the oracle with high probability.

**Theorem 3.** Assume Conditions 1 and 2 hold. Moreover, for a prespecified \( \delta \in (0,1) \), assume there exists some \( \gamma > \gamma_0 := 2.5/(\lambda \Sigma \delta) \) such that \( p'(\gamma_0) > 0, p'(t) = 0 \) for all \( t \geq \gamma \) and \( \| \beta^s \|_{\infty} \geq (\gamma_0 + \gamma) \lambda \).

(I) (Diverging \( \tau \)) Let \( (\tau, \lambda) \) satisfy \( \tau \asymp \sigma_x \sqrt{n(s + \log n)} \) and \( \lambda \asymp \sigma_x \sqrt{(s + \log d)/n} + \tau \log(d)/n \), and suppose \( n \geq \max\{s \log d, s^{3/2}\} \). Then with probability at least 1 − 4d − 1 − 2n − 1, \( \hat{\theta}^{(\ell)} = \hat{\theta}^{\text{oracle}} \) for all \( \ell \geq \log(s^{1/2}/\delta)/\log(1/\delta) \).

(II) (Fix \( \tau \)) Assume further that the conditions of Theorem 1, (II) hold. Let \( (\tau, \lambda) \) satisfy \( \tau \asymp \sigma_x \) and \( \lambda \asymp \sigma_x \sqrt{(s + \log d)/n} \), and suppose \( n \geq \max\{s \log d, s^2\} \). Then with probability at least 1 − 4d − 1 − 2n − 1, \( \hat{\theta}^{(\ell)} = \hat{\theta}^{\text{oracle}} \) for all \( \ell \geq \log(s^{1/2}/\delta)/\log(1/\delta) \).

### 3 Optimization algorithm

In this section, we use the local adaptive majorize-minimization (LAMM) principal (Fan et al., 2018) to derive an iterative algorithm for solving each subproblem \( (P_\ell) \) in (10):

\[
\min_{\theta = (\beta_0, \beta^T)^T} \left\{ L_\ell(\theta) + \| \lambda^{(\ell-1)} \circ \beta \|_1 \right\}.
\]

#### 3.1 LAMM algorithm

To minimize a general function \( f(\theta) \), at a given point \( \theta^{(k)} \), the majorize-minimization (MM) algorithm first majorizes it by another function \( g(\theta | \theta^{(k)}) \), which satisfies

\[
g(\theta | \theta^{(k)}) \geq f(\theta) \quad \text{and} \quad g(\theta^{(k)} | \theta^{(k)}) = f(\theta^{(k)}) \quad \text{for all } \theta,
\]

and then compute \( \theta^{(k+1)} = \arg\min_{\theta} g(\theta | \theta^{(k)}) \) (Lange, Hunter and Yang, 2000). The objective value of such an algorithm is non-increasing in each step, since

\[
f(\theta^{(k+1)}) \leq g(\theta^{(k+1)} | \theta^{(k)}) \leq g(\theta^{(k)} | \theta^{(k)}) \equiv f(\theta^{(k)}).
\]

(31)
Algorithm 1 LAMM algorithm at the $k$-th iteration of the $\ell$-th subproblem.

1: Algorithm: $\{\theta^{(\ell,k)}, \phi^{(\ell,k)}\} \leftarrow$ LAMM$(\lambda^{(\ell-1)}, \theta^{(\ell,k-1)}, \phi_0, \phi^{(\ell,k-1)})$
2: Input: $\lambda^{(\ell-1)}, \theta^{(\ell,k-1)}, \phi_0, \phi^{(\ell,k-1)}$
3: Initialize: $\phi^{(\ell,k)} \leftarrow \max\{\phi_0, \gamma_u^{-1}\phi^{(\ell,k-1)}\}$
4: Repeat
5: $\theta^{(\ell,k)} \leftarrow T_{\lambda^{(\ell-1)},\phi^{(\ell,k)}}(\theta^{(\ell,k-1)})$
6: If $F(\theta^{(\ell,k)}, \lambda^{(\ell-1)}) < \mathcal{L}_\tau(\theta^{(\ell,k-1)})$ then $\phi^{(\ell,k)} \leftarrow \gamma_u \phi^{(\ell,k)}$
7: Until $F(\theta^{(\ell,k)}, \lambda^{(\ell-1)}) \geq \mathcal{L}_\tau(\theta^{(\ell,k-1)})$
8: Return $\{\theta^{(\ell,k)}, \phi^{(\ell,k)}\}$

Fan et al. (2018) observed that the global majorization requirement is not necessary. It only requires the local properties

$$f(\theta^{(k+1)}) \leq g(\theta^{(k+1)}|\theta^{(k)}) \quad \text{and} \quad g(\theta^{(k)}|\theta^{(k)}) = f(\theta^{(k)})$$

(32)

for the inequalities in (31) to hold.

Using the above principal, it suffices to locally majorize the objective function $\mathcal{L}_\tau(\theta)$ in the above minimization problem. At the $k$-th step with working parameter vector $\theta^{(\ell,k-1)}$, we use an isotropic quadratic function, that is,

$$F(\theta; \phi, \theta^{(\ell,k-1)}) := \mathcal{L}_\tau(\theta^{(\ell,k-1)}) + \langle \nabla \mathcal{L}_\tau(\theta^{(\ell,k-1)}), \theta - \theta^{(\ell,k-1)} \rangle + \frac{\phi}{2}\|\theta - \theta^{(\ell,k-1)}\|_2^2,$$

(33)

to locally majorize $\mathcal{L}_\tau(\theta)$ such that

$$F(\theta^{(\ell,k)}; \phi^{(\ell,k)}, \theta^{(\ell,k-1)}) \geq \mathcal{L}_\tau(\theta^{(\ell,k)}),$$

(34)

where $\phi^{(\ell,k)}$ is a proper quadratic coefficient at the $k$-th update, and $\theta^{(\ell,k)}$ is the solution to

$$\min_{\theta} \{F(\theta; \phi^{(\ell,k)}, \theta^{(\ell,k-1)}) + \|\lambda^{(\ell-1)} \odot \beta\|_1\}.$$ 

It is easy to see that $\theta^{(\ell,k)} = (\beta_0^{(\ell,k)}, (\beta^{(\ell,k)})^\top)$ takes a simple explicit form

$$\begin{cases}
\beta_0^{(\ell,k)} = \beta_0^{(\ell,k-1)} - \partial_{\beta_0} \mathcal{L}_\tau(\theta^{(\ell,k-1)})/\phi^{(\ell,k)}, \\
\beta^{(\ell,k)} = S_{\text{soft}}(\beta^{(\ell,k-1)} - \nabla_{\beta} \mathcal{L}_\tau(\theta^{(\ell,k-1)})/\phi^{(\ell,k)}), \lambda^{(\ell-1)}/\phi^{(\ell,k)},
\end{cases}$$

(35)

where $S_{\text{soft}}(x, \lambda) := (\text{sign}(x_j) \max\{|x_j| - \lambda_j, 0\})$ is the soft-thresholding operator. For simplicity, we summarize and define the above update as $\theta^{(\ell,k)} = T_{\lambda^{(\ell-1)},\phi^{(\ell,k)}}(\theta^{(\ell,k-1)})$. Using this simple update formula of $\theta$, we iteratively search for the pair $(\phi^{(\ell,k)}, \theta^{(\ell,k)})$ to make the local majorization (34) hold. Starting with an initial quadratic coefficient $\phi = \phi_0$, say $10^{-4}$, we iteratively increase $\phi$ by a factor of $\gamma_u > 1$ and compute

$$\theta^{(\ell,k)} = T_{\lambda^{(\ell-1)},\phi^{(\ell,k)}}(\theta^{(\ell,k-1)}) \quad \text{with} \quad \phi^{(\ell,k)} = \gamma_u^{k-1}\phi_0,$$

until the local property (34) holds. This routine is summarized in Algorithm 1.
3.2 Complexity theory

To investigate the complexity theory of the proposed algorithm, we first impose the following regularity condition.

**Condition 3.** $\nabla L_\tau(\theta)$ is locally $\rho_c$-Lipschitz continuous, that is,

$$\|\nabla L_\tau(\theta_1) - \nabla L_\tau(\theta_2)\|_2 \leq \rho_c \|\theta_1 - \theta_2\|_2,$$

where $\rho_c$ is the Lipschitz constant.

Our next theorem characterizes the computational complexity in the contraction stage. Recall that $\lambda(0) = (\lambda, \ldots, \lambda)^T \in \mathbb{R}^d$.

**Theorem 4.** Assume that Condition 3 holds and the optimal solution $\hat{\theta}^{(1)}$ satisfies $\|\hat{\theta}^{(1)} - \theta^*\|_2 \lesssim s^{1/2} \lambda$. Then, to attain an $\epsilon_c$-optimal solution $\tilde{\theta}^{(1)}$, i.e. $\omega_{\lambda(0)}(\tilde{\theta}^{(1)}) \leq \epsilon_c$, in the contraction stage, we need at most

$$C_1 \rho_c^2 (1 + \gamma_u)(\|\theta^*\|_2 + s^{1/2} \lambda)^2 / \epsilon_c^2$$

LAMM iterations in (35), where $C_1 > 0$ is a constant.

The sublinear rate in the contraction stage is due to the lack of strong convexity of the loss function in this stage, because we start with a naive initial value $0$. Once we enter the contracting region where the estimator is sparse and reasonably close to the underlying true parameter vector, the problem becomes strongly convex (at least with high probability). This endows the algorithm a linear rate of convergence. Our next theorem provides a formal statement on the geometric convergence rate for each subproblem in the tightening stage.

We need a form of sparse eigenvalue condition.

**Definition 3 (Localized Sparse Eigenvalue, LSE).** For any integer $m \geq |S|$ and $r, \tau > 0$, define the sparse cone

$$C(m, r, \tau) = \{ \tilde{u} = (u_0, u^T) \in \mathbb{R}^{d+1} : \|\tilde{u}\|_2 = 1, u_{\mathcal{J}^c} = 0 \text{ for some } \mathcal{J} \subseteq [d] \text{ s.t. } S \subseteq \mathcal{J} \text{ and } |\mathcal{J}| \leq m \}.$$  

The localized sparse eigenvalues are then defined as

$$\rho_+(m, r, \tau) = \sup \{ \tilde{u}^T \nabla^2 L_\tau(\theta) \tilde{u} : \tilde{u} \in \mathbb{C}(m, r, \tau), \theta \in \mathbb{B}_2(r, \theta^*) \}$$

and

$$\rho_- (m, r, \tau) = \inf \{ \tilde{u}^T \nabla^2 L_\tau(\theta) \tilde{u} : \tilde{u} \in \mathbb{C}(m, r, \tau), \theta \in \mathbb{B}_2(r, \theta^*) \}.$$  

**Condition 4.** We say that the LSE condition holds if there exist an integer $\tilde{s} \lesssim s$ and $r, \tau > 0$ such that

$$0 < \rho_s \leq \rho_-(2s + 2 + 2\tilde{s}, r, \tau) < \rho_+(2s + 2 + 2\tilde{s}, r, \tau) \leq \rho^* < +\infty$$

and

$$\frac{\rho_+(\tilde{s}, r, \tau)}{\rho_- (2s + 2 + 2\tilde{s}, r, \tau)} \leq 1 + C\tilde{s}/s,$$

where $\rho_s, \rho^*$ and $C$ are positive constants.
\textbf{Theorem 5.} Suppose that Condition 4 holds. To obtain an \( \epsilon_t \)-optimal solution \( \tilde{\theta}^{(t)} \), i.e. \( \omega_{\lambda(t-1)}(\tilde{\theta}^{(t)}) \leq \epsilon_t \), in the \( t \)-th subproblem for \( t \geq 2 \), we need at most \( C_1 \log(C_2 \lambda \sqrt{s}/\epsilon_t) \) LAMM iterations in (35), where \( C_1 \) and \( C_2 \) are positive constants.

Together, the above two results yield the following corollary, which characterizes the computational complexity of the entire algorithm.

\textbf{Corollary 1.} Suppose that Condition 4 holds. To achieve a sequence of approximate solutions \( \{\tilde{\theta}^{(t)}\}_{t=1}^T \) such that \( \omega_{\lambda(0)}(\tilde{\theta}^{(1)}) \leq \epsilon_c \leq \lambda \) and \( \omega_{\lambda(t-1)}(\tilde{\theta}^{(t)}) \leq \epsilon_t \leq \sqrt{T/n} \) for \( 2 \leq t \leq T \), the required number of LAMM iterations is at most \( C_1 \epsilon_{c}^{-2} + C_2(T-1) \log(\epsilon_{t}^{-1}) \), where \( C_1 \) and \( C_2 \) are positive constants.

\section{Extension to general robust losses}

In Sections 2 and 3, we have restricted our attention to the Huber loss. As a representative robust loss function, the Huber loss has the merit of being (i) globally \( \tau \)-Lipschitz continuous, and (ii) locally quadratic. A natural question arises that whether similar results, both statistical and computational, remain valid for more general loss functions that possess the above two features. In this section, we introduce a class of loss functions which, combined with nonconvex regularization, leads to statistically optimal and robust estimators.

\textbf{Condition 5} (Globally Lipschitz and locally quadratic loss functions). Consider a general loss function \( \ell_r(\cdot) \) that is of the form \( \ell_r(x) = \tau^2 \ell(x/\tau) \) for \( x \in \mathbb{R} \), where \( \tau : \mathbb{R} \mapsto [0, \infty) \) satisfies: (i) \( \ell'(0) = 0 \) and \( |\ell'(x)| \leq c_1 \) for all \( x \in \mathbb{R} \); (ii) \( \ell''(0) = 1 \) and \( \ell''(x) \geq c_2 \) for all \( |x| \leq c_3 \); and (iii) \( |\ell'(x) - x| \leq c_4 x^2 \) for all \( x \in \mathbb{R} \), where \( c_1 - c_4 \) are positive constants.

We first discuss the implications of the three properties in Condition 5. First, since \( \ell_r'(x) = \tau \ell'(x/\tau) \), it follows from property (i) that \( \sup_{x \in \mathbb{R}} |\ell_r'(x)| \leq c_1 \tau \). The boundedness of \( |\ell_r'| \) facilitates the use of Bernstein’s inequality on deriving upper bounds for the random quantities

\[ \|\nabla \mathcal{L}_r(\theta) - \mathbb{E} \nabla \mathcal{L}_r(\theta)\|_\infty \quad \text{and} \quad \|\nabla \mathcal{L}_r(\theta) - S - \mathbb{E} \nabla \mathcal{L}_r(\theta)S\|_2 \]

as in Proposition 8, where \( \mathcal{L}_r(\theta) = (1/n) \sum_{i=1}^n \ell_r(y_i - z_i^T \theta) \) and \( \theta = \theta^* \) or \( \theta^*_\tau \). Next, note that \( \ell_r'''(x) = \ell'''(x/\tau) \). Property (ii) indicates that \( \ell_r \) is strongly convex on \([-c_3 \tau, c_3 \tau]\), which turns out to be the key factor in establishing the restricted strong convexity condition on \( \mathcal{L}_r \). See Proposition 7, and Lemmas 3, 4 and 6. Lastly, property (iii) is particularly useful in the “diverging \( \tau \)” scenario. Even though it can be shown under property (i) that \( \mathcal{L}_r(\theta^*) \) is concentrated around its expected value \( \mathbb{E} \mathcal{L}_r(\theta^*) \) with high probability, in general \( \mathbb{E} \nabla \mathcal{L}_r(\theta^*) = -\mathbb{E} \{\ell_r'(\epsilon)z\} \) is nonzero. However, since \( \mathbb{E}(\epsilon|x) = 0 \), we have \( \mathbb{E}\{\ell_r'(\epsilon)|x\} = \mathbb{E}\{\ell_r'(\epsilon) - \epsilon|x\} = \tau \mathbb{E}\{\ell_r'(\epsilon/\tau) - \epsilon/\tau|x\} \). Together with property (iii), this implies

\[ |\mathbb{E}\{\ell_r'(\epsilon)|x\}| \leq c_4 \tau \mathbb{E}\{(\epsilon/\tau)^2|x\} = c_4 \sigma_\epsilon^2 \tau^{-1}. \]

The term “\( \sigma_\epsilon^2 \tau^{-1} \)”, which corresponds to the estimation bias, then arises on the right-hand of (25) and (26).

Below we list four examples of \( \ell \) that satisfy Condition 5.
1. (Huber loss): $\ell(x) = x^2/2 \cdot I(|x| \leq 1) + (|x| - 1/2) \cdot I(|x| > 1)$ with $\ell'(x) = xI(|x| \leq 1) + \text{sign}(x)I(|x| > 1)$ and $\ell''(x) = I(|x| \leq 1)$. Moreover, 

$$|\ell'(x) - x| = |x - \text{sign}(x)I(|x| > 1)| \leq x^2.$$ 

2. (Pseudo-Huber loss I): $\ell(x) = \sqrt{1 + x^2} - 1$, whose first and second derivatives are 

$$\ell'(x) = \frac{x}{\sqrt{1 + x^2}} \quad \text{and} \quad \ell''(x) = \frac{1}{(1 + x^2)^{3/2}},$$

respectively. It is easy to see that $\sup_{x \in \mathbb{R}} |\ell'(x)| \leq 1$ and $\ell''(x) \geq (1 + c^2)^{-3/2}$ for all $|x| \leq c$ and $c > 0$. Moreover, since $\ell'''(x) = -3(1 + x^2)^{-5/2}$ satisfies $|\ell'''(x)| < 0.9$ for all $x$, it follows from Taylor's theorem and Lagrange error bound that $|\ell'(x) - x| = |\ell'(x) - \ell'(0) - \ell''(0)x| \leq 0.45x^2$.

3. (Pseudo-Huber loss II): $\ell(x) = \log\{(e^x + e^{-x})/2\}$, whose first and second derivatives are, respectively,

$$\ell'(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad \text{and} \quad \ell''(x) = \frac{4}{(e^x + e^{-x})^2}.$$ 

It follows that $\sup_{x \in \mathbb{R}} |\ell'(x)| \leq 1$ and $\ell''(x) \geq 4(e^c + e^{-c})^{-2}$ for all $|x| \leq c$ and $c > 0$. Moreover, we calculate the third derivative $\ell'''(x) = -8(e^x - e^{-x})(e^x + e^{-x})^{-4}$ that satisfies $|\ell'''(x)| < 0.4$. Again, by Taylor’s theorem and Lagrange error bound, $|\ell'(x) - x| \leq 0.2x^2$.

4. (Smoothed Huber loss I): Note that the Huber loss is twice differentiable in $\mathbb{R}$, except at $\pm 1$. Modifying the Huber loss gives rise to the following function that is twice differentiable everywhere:

$$\ell(x) = \begin{cases} 
    x^2/2 - |x|^3/6 & \text{if } |x| \leq 1, \\
    |x|/2 - 1/6 & \text{if } |x| > 1,
\end{cases}$$

whose first and second derivatives are

$$\ell'(x) = \begin{cases} 
    x - \text{sign}(x) \cdot x^2/2 & \text{if } |x| \leq 1, \\
    \text{sign}(x)/2 & \text{if } |x| > 1,
\end{cases} \quad \ell''(x) = \begin{cases} 
    1 - |x| & \text{if } |x| \leq 1, \\
    0 & \text{if } |x| > 1.
\end{cases}$$

Direct calculations show that $\sup_{x \in \mathbb{R}} |\ell'(x)| \leq 1/2$ and $\ell''(x) \geq 1 - c$ for all $|x| \leq c$ and $0 < c < 1$. Since $\ell''$ is 1-Lipschitz continuous, we have $|\ell'(x) - x| \leq x^2/2$.

5. (Smoothed Huber loss II): Another smoothed version of the Huber loss function is

$$\ell(x) = \begin{cases} 
    x^2/2 - x^4/24 & \text{if } |x| \leq \sqrt{2}, \\
    (2\sqrt{2}/3)|x| - 1/2 & \text{if } |x| > \sqrt{2}.
\end{cases}$$

The derivative of this function, also known as the influence function, is used in Catoni and Giulini (2017) for mean vector estimation. We compute

$$\ell'(x) = \begin{cases} 
    x - x^3/6 & \text{if } |x| \leq \sqrt{2}, \\
    (2\sqrt{2}/3)\text{sign}(x) & \text{if } |x| > \sqrt{2},
\end{cases} \quad \ell''(x) = \begin{cases} 
    1 - x^2/2 & \text{if } |x| \leq \sqrt{2}, \\
    0 & \text{if } |x| > \sqrt{2}.
\end{cases}$$
It is easy to see that \( \sup_{x \in \mathbb{R}} |\ell'(x)| \leq 2\sqrt{2}/3 \) and \( \ell''(x) \geq 1 - c^2/2 \) for all \( |x| \leq c \) and \( 0 < c < \sqrt{2} \). Noting that \( \ell'' \) is \( \sqrt{2} \)-Lipschitz continuous, it holds \( |\ell'(x) - x| \leq x^2/\sqrt{2} \).

The loss functions discussed above, along with their derivatives up to order three, are all plotted in Figure 1 except for the Huber loss. Provided that the loss function \( \ell_\tau \) satisfies Condition 5, all the theoretical results in Sections 2.3 and 2.4 remain valid only with different constants. It is worth noticing that the four loss functions described in examples 2–5 and plotted in Figure 1 also have Lipschitz continuous second derivatives. In certain cases, this additional smoothness facilitates theoretical analysis and leads to more refined results.

Figure 1: Examples of robust loss functions and their derivatives.
Theorem 6. Assume Conditions 1 and 2 hold, and that ε and x are independent. For a prespecified δ ∈ (0, 1), assume there exists some γ ≥ γδ := 2.5/(2Σδ) such that p’(γδ) > 0, p’(t) = 0 for all t ≥ γ and ∥β∗∥min ≥ (γδ + γ)λ. Let ℓτ(x) = τ²ℓ(x/τ) be a loss function satisfying Condition 5 and moreover, ℓτ” is Lipschitz continuous. Let τ ≍ σε be such that the function α ↦ E{ℓ[(ε - α)/τ]} has a unique minimizer, denoted by ατ, and P(|ε - ατ| ≤ c₃τ/2) > 0 for c₃ > 0 given in Condition 5. Then, for the choice λ ≍ σε√log(d)/n, we have, with probability at least 1 - C(n⁻¹ + d⁻¹),

\[ \hat{\theta}^{(\ell)} = \hat{\theta}_{\text{oracle}} \quad \text{and} \quad \|\hat{\theta}^{(\ell)} - \theta^*_x\|_2 \lesssim \sigma_r \sqrt{\frac{s \log n}{n}} \]

for all ℓ ≥ log(s²/δ)/log(1/δ) as long as n ≥ max{s log d, s²}.

Theorem 6 refines Part (II) of Theorem 3 in respect to the order of λ, and therefore the minimum signal strength requirement: here λ is of the order σε√log(d)/n, while in Theorem 3, (II), λ scales as σε√(s + log d)/n.

5 Empirical study

In this section, we compare the empirical performance of the proposed two-stage robust regression approach with several benchmark methods, such as the Lasso (Tibshirani, 1996), the SCAD- and MCP-penalized least squares methods (Fan and Li, 2001; Zhang, 2010a). All the computational results presented below are reproducible using software available at https://github.com/XiaoouPan/ILAMM.

We generate data vectors \{(yᵢ, xᵢ)\}_{i=1}^{n} from two types of linear models:

1. (Homoscedastic model):
   \[ yᵢ = xᵢ^Tβ^* + \varepsilonᵢ \quad \text{with} \quad xᵢ \sim \mathcal{N}(0, I_d), \quad i = 1, \ldots, n; \]

2. (Heteroscedastic model):
   \[ yᵢ = xᵢ^Tβ^* + c^{-1}(xᵢ^Tβ^*)² \varepsilonᵢ \quad \text{with} \quad xᵢ \sim \mathcal{N}(0, I_d), \quad i = 1, \ldots, n, \]

where the constant c is chosen as c = \sqrt{3} ∥β^*∥² such that E{c^{-1}(xᵢ^Tβ^*)²}² = 1, and therefore the variance of the noise is the same as that of εᵢ.

In addition, we consider the following four error distributions:

1. Normal distribution \mathcal{N}(μ, σ²) with mean μ = 0 and standard deviation σ = 1.5;

2. Skewed generalized t distribution \mathcal{SG}(0, 5, 0.75, 2, 2.5) (Theodossiou, 1998) with mean μ = 0, variance σ² = q/(q - 2) = 5, q = 2.5, skewness parameter λ = 0.75 and shape parameter p = 2;

3. Lognormal distribution \mathcal{LN}(μ, σ²) with log location parameter μ = 0 and log shape parameter σ = 1.2;
4. Pareto distribution \( \text{Par}(x_m, \alpha) \) with scale parameter \( x_m = 2 \) and shape parameter \( \alpha = 2.2 \).

Except for the normal distribution, all the other three are skewed and heavy-tailed. To meet our model assumption, we subtract the mean from the lognormal and Pareto distributions.

In both homoscedastic and heteroscedastic models, the sample size \( n = 100 \), the ambient dimension \( d = 1000 \) and the sparsity parameter \( s = 6 \). The true vector of regression coefficients is \( \beta^* = (4, 3, 2, -2, -2, 2, 0, \ldots, 0)^T \), where the first 6 elements are non-zero and the rest are all equal to 0. We apply the proposed TAC (Tightening After Contraction) algorithm to compute all the estimators with tuning parameters \( \lambda \) and \( \tau \) chosen via three-fold cross-validation. To be more specific, we first choose a sequence of \( \lambda \) values the same way as in the \texttt{glmnet} algorithm (Friedman, Hastie and Tibshirani, 2010). Next, guided by its theoretically “optimal” choice, the candidate set for \( \tau \) is taken to be \( \{2^j \hat{\sigma}_{\text{MAD}} \sqrt{n/\log(nd)} : j = -2, -1, 0, 1, 2\} \), where \( \hat{\sigma}_{\text{MAD}} := \text{median}\{|\hat{R} - \text{median}(\hat{R})|\}/\Phi^{-1}(3/4) \) is the median absolute deviation (MAD) estimator using the residuals \( \hat{R} = (\hat{r}_1, \ldots, \hat{r}_n)^T \) obtained from the Lasso.

To highlight the tail robustness and oracle property of our algorithm, we consider the following four measurements to assess the empirical performance:

1. True positive, TP, which is the number of signal variables that are selected;
2. False positive, FP, which is the number of noise variables that are selected;
3. Relative error, \( \text{RE}_1 \) and \( \text{RE}_2 \), which is the relative error of an estimator \( \hat{\beta} \) with respect to the Lasso under \( \ell_1 \)- and \( \ell_2 \)-norms:

\[
\text{RE}_1 = \frac{||\hat{\beta} - \beta^*||_1}{||\hat{\beta}_{\text{Lasso}} - \beta^*||_1} \quad \text{and} \quad \text{RE}_2 = \frac{||\hat{\beta} - \beta^*||_2}{||\hat{\beta}_{\text{Lasso}} - \beta^*||_2}.
\]
<table>
<thead>
<tr>
<th>Error dist.</th>
<th>Lasso</th>
<th>SCAD</th>
<th>Huber-SCAD</th>
<th>MCP</th>
<th>Huber-MCP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>TP</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
</tr>
<tr>
<td></td>
<td>FP</td>
<td>24.44</td>
<td>3.11</td>
<td>2.19</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>RE₁</td>
<td>1.00</td>
<td>0.23</td>
<td>0.22</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>RE₂</td>
<td>1.00</td>
<td>0.32</td>
<td>0.33</td>
<td>0.30</td>
</tr>
<tr>
<td>Skewed t</td>
<td>TP</td>
<td>4.74</td>
<td>4.87</td>
<td>4.74</td>
<td>3.97</td>
</tr>
<tr>
<td></td>
<td>FP</td>
<td>20.78</td>
<td>18.49</td>
<td>11.48</td>
<td>4.28</td>
</tr>
<tr>
<td></td>
<td>RE₁</td>
<td>1.00</td>
<td>0.88</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>RE₂</td>
<td>1.00</td>
<td>0.91</td>
<td>0.86</td>
<td>0.94</td>
</tr>
<tr>
<td>Lognormal</td>
<td>TP</td>
<td>5.68</td>
<td>5.71</td>
<td>6.00</td>
<td>5.49</td>
</tr>
<tr>
<td></td>
<td>FP</td>
<td>29.70</td>
<td>16.75</td>
<td>3.80</td>
<td>4.32</td>
</tr>
<tr>
<td></td>
<td>RE₁</td>
<td>1.00</td>
<td>0.54</td>
<td>0.15</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>RE₂</td>
<td>1.00</td>
<td>0.62</td>
<td>0.23</td>
<td>0.60</td>
</tr>
<tr>
<td>Pareto</td>
<td>TP</td>
<td>5.64</td>
<td>5.67</td>
<td>6.00</td>
<td>5.44</td>
</tr>
<tr>
<td></td>
<td>FP</td>
<td>28.30</td>
<td>14.69</td>
<td>2.91</td>
<td>3.48</td>
</tr>
<tr>
<td></td>
<td>RE₁</td>
<td>1.00</td>
<td>0.51</td>
<td>0.14</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>RE₂</td>
<td>1.00</td>
<td>0.58</td>
<td>0.21</td>
<td>0.57</td>
</tr>
</tbody>
</table>

Table 1: Simulation results for the Lasso, SCAD, Huber-SCAD, MCP and Huber-MCP estimators under the homoscedastic model (38).

Tables 1 and 2 summarize the averages of each measurement, TP, FP, RE₁ and RE₂, over 200 replications under models (38) and (39). Here, Huber-SCAD and Huber-MCP signify the proposed two-stage algorithm using the SCAD and MCP penalties, respectively. When the noise distributions are heavy-tailed and/or skewed, we see that Huber-SCAD and Huber-MCP outperform SCAD and MCP, respectively, with fewer spurious discoveries (false positives) and smaller estimation errors. Under the homoscedastic normal model, Huber-SCAD and Huber-MCP perform similarly to their least squares counterparts; while under heteroscedasticity, the proposed algorithm exhibits a notable advantage over existing methods on selection consistency even though the error is normally distributed. In summary, these numerical studies validate our expectations that the proposed robust regression algorithm improves the Lasso as a general regression analysis method on two aspects: robustness against heavy-tailed (and even heteroscedastic) noise and selection consistency.
To further visualize the advantage of our robust regression methods over the existing ones (e.g. Lasso, SCAD and MCP), we draw the receiver operating characteristic (ROC) curve, which is the plot of true positive rate (TPR) against false positive rate (FPR) at various regularization parameters. Specifically, TPR and FPR are defined, respectively, as the ratio of true positive to $s$ and the ratio of false positive to $d - s$. We generate data vectors $\{(y_i, x_i)\}_{i=1}^n$ from both homoscedastic and heteroscedastic models (38) and (39) with sample size $n = 100$, dimension $d = 1000$ and sparsity $s = 10$. The true vector of regression coefficients is $\beta^* = (1.5, 1.5, \ldots, 1.5, 0, \ldots, 0)^T$, where the first 10 elements are non-zero with weaker signals than the previous experiment, and the rest are all equal to 0. We apply the proposed TAC algorithm to implement all the five methods, Lasso, SCAD, MCP, Huber-SCAD and Huber-MCP, with a sequence of $\lambda$ values chosen as before and $\tau$ as $\hat{\sigma}_{\text{MAD}} \sqrt{n/\log(nd)}$. For each combination of $\lambda$ and $\tau$, the empirical FPR and TPR are computed based on 200 simulations.

Figures 2 and 3 indicate evident advantage of Huber-SCAD and Huber-MCP over their least squares counterparts: the robust methods have a greater area under the curve (AUC) when the noise distribution is heavy-tailed and/or skewed in both homoscedastic and heteroscedastic models. Surprisingly, even in a normal model, the proposed methods still outperform the competitors by a visible margin.
Figure 2: Plots of ROC curves of the five methods under homoscedastic model (38) with errors generated from four distributions: normal, Student’s $t$, lognormal and Pareto.
Figure 3: Plots of ROC curves of the five methods under heteroscedastic model (39) with errors generated from four distributions: normal, Student’s $t$, lognormal and Pareto.

6 Discussion

In this paper, we have presented a general computational framework for fitting high-dimensional linear models with heavy-tailed noise. In such a case, a real side-effect of the quadratic loss has been recognized: the combination of its rapid growth with heavy-tailed sampling distributions inevitably leads to outliers (Fan, Li and Wang, 2017; Mendelson, 2018). To achieve robustness against such outliers, we discussed a general class of loss functions, typified by the Huber loss, which are globally Lipschitz continuous and locally
strongly convex. The use of nonconvex regularizers eliminates the bias introduced by convex penalty, such as the popular $\ell_1$-penalty, and therefore produce oracle solutions. The proposed tightening after contraction algorithm, which bypasses directly solving the non-convex optimization program, iteratively solves a sequence of adaptive convex programs with controlled algorithmic complexity. Statistically, our algorithm produces solutions that enjoy optimal rates of convergence and oracle properties. Moreover, our theoretical analysis provides key insights on calibrating the robustification parameters properly, which leads to the removal of outliers without sacrificing statistical efficiency.

References


Lemma 1. Let \(L \in \mathbb{R}^{d \times d}\) satisfy (\ref{eq:lemma1}). \(\frac{1}{2} \| x \|^2 \) is an \(\ell_1\)-vector of regularization parameters with \(\lambda_j \geq 0\). Consider the optimization problem

\[
\min_{\beta \in \mathbb{R}^{d+1}} \{ L_\tau(\beta) + \| \lambda \circ \beta \|_1 \},
\]

where \(L_\tau(\beta) = (1/n) \sum_{i=1}^n \ell_i(y_i - z_i^\top \beta)\) and \(\lambda \circ \beta = (\lambda_1 \beta_1, \ldots, \lambda_d \beta_d)^\top\).

The following result provides conditions under which an \(\epsilon\)-optimal solution to the convex program (\ref{eq:lemma1}) falls in an \(\ell_1\)-cone. Recall that \(S = \text{supp} (\beta^*)\).

Lemma 1. Let \(\mathcal{E}\) be a subset of \([d]\) satisfying \(S \subseteq \mathcal{E}\). For any \(\beta = (\beta_0, \beta^\top)^\top \in \mathbb{R}^{d+1}\) satisfying \(\beta_{\mathcal{E}^c} = 0\) and \(\epsilon > 0\), provided \(\lambda = (\lambda_1, \ldots, \lambda_d)^\top\) satisfies \(\| \lambda_{\mathcal{E}^c} \|_{\min} > \| \nabla L_\tau(\beta) \|_\infty + \epsilon\), any \(\epsilon\)-optimal solution \(\bar{\beta} = (\bar{\beta}_0, \bar{\beta}^\top)^\top\) to (\ref{eq:lemma1}) satisfies

\[
\| (\bar{\beta} - \beta)_{\mathcal{E}^c} \|_1 \leq \frac{(\| \lambda \|_{\infty} + \| \nabla L_\tau(\beta) \|_{\infty} + \epsilon) \| (\bar{\beta} - \beta)_{\mathcal{E}^c} \|_1 + \| \nabla L_\tau(\beta) \|_{\infty} + \epsilon) |\bar{\beta}_0 - \beta_0|}{\| \lambda_{\mathcal{E}^c} \|_{\min} - \| \nabla L_\tau(\beta) \|_\infty - \epsilon}.
\]

Proof of Lemma 1. For any \(\xi \in \partial \| \bar{\beta} \|_1\), define \(u = u_\xi = \nabla L_\tau(\bar{\beta}) + (\lambda \circ \xi)^\top \in \mathbb{R}^{d+1}\). Note that

\[
\| u \|_\infty \bar{\theta} - \theta \|_1 \geq \langle u, \bar{\theta} - \theta \rangle
\]

\[
= \langle \nabla L_\tau(\bar{\theta}) - \nabla L_\tau(\theta), \bar{\theta} - \theta \rangle + \langle \nabla L_\tau(\theta), \bar{\theta} - \theta \rangle + \langle \lambda \circ \xi, \bar{\beta} - \beta \rangle
\]

\[
\geq \langle \nabla L_\tau(\theta), \bar{\theta} - \theta \rangle + \langle \lambda \circ \xi, \bar{\beta} - \beta \rangle
\]

\[
\geq -\| \nabla L_\tau(\theta) \|_\infty \| \bar{\theta} - \theta \|_1 + \langle \lambda \circ \xi, \bar{\beta} - \beta \rangle.
\]
Moreover, we have
\[
\langle \lambda \circ \xi, \tilde{\beta} - \beta \rangle = \langle (\lambda \circ \xi)_{\xi^c}, (\tilde{\beta} - \beta)_{\xi^c} \rangle + \langle (\lambda \circ \xi)_{\xi}, (\tilde{\beta} - \beta)_{\xi} \rangle \\
g \geq \|\lambda_{\xi^c}\|_{\min} \| (\tilde{\beta} - \beta)_{\xi^c} \|_1 - \|\lambda_{\xi}\|_{\infty} \| (\tilde{\beta} - \beta)_{\xi} \|_1.
\]
Together, the last two displays imply
\[
\|u\|_{\infty} \|\tilde{\theta} - \theta\|_1 \geq -\|\nabla L_\tau(\theta)\|_{\infty} \|\tilde{\theta} - \theta\|_1 + \|\lambda_{\xi^c}\|_{\min} \| (\tilde{\beta} - \beta)_{\xi^c} \|_1 - \|\lambda_{\xi}\|_{\infty} \| (\tilde{\beta} - \beta)_{\xi} \|_1.
\]
Since the right-hand side of this inequality does not depend on \(\xi\), taking the infimum with respect to \(\xi \in \partial \|\tilde{\beta}\|_1\) on both sides to reach
\[
\omega_{\lambda_{\xi}}(\tilde{\theta}) \|\tilde{\theta} - \theta\|_1 \geq -\|\nabla L_\tau(\theta)\|_{\infty} \|\tilde{\theta} - \theta\|_1 + \|\lambda_{\xi^c}\|_{\min} \| (\tilde{\beta} - \beta)_{\xi^c} \|_1 - \|\lambda_{\xi}\|_{\infty} \| (\tilde{\beta} - \beta)_{\xi} \|_1.
\]
By definition, \(\tilde{\theta}\) is an \(\epsilon\)-optimal solution so that \(\omega_{\lambda_{\xi}}(\tilde{\theta}) \leq \epsilon\). Putting together the pieces we obtain
\[
\{\epsilon + \|\nabla L_\tau(\theta)\|_{\infty}\} \|\tilde{\theta} - \theta\|_1 \geq \|\lambda_{\xi^c}\|_{\min} \| (\tilde{\beta} - \beta)_{\xi^c} \|_1 - \|\lambda_{\xi}\|_{\infty} \| (\tilde{\beta} - \beta)_{\xi} \|_1.
\]
Decompose \(\|\tilde{\theta} - \theta\|_1\) as \(\| (\tilde{\beta} - \beta)_{\xi^c} \|_1 + \| (\tilde{\beta} - \beta)_{\xi} \|_1 + |\tilde{\beta}_0 - \beta_0|\), the stated result follows immediately.

**Lemma 2.** Let \(E \subseteq [d]\) be such that \(S \subseteq E\) and \(k = |E|\). For \(\theta = (\beta_0, \beta^T)\) \(\in \mathbb{R}^{d+1}\) satisfying \(\beta_{\xi^c} = 0\), assume that \(\lambda = (\lambda_1, \ldots, \lambda_d)^T\) satisfies \(\|\lambda\|_{\infty} \leq \lambda\) and \(\|\lambda_{\xi^c}\|_{\min} \geq a\lambda \geq 2\|\nabla L_\tau(\theta)\|_{\infty} + \epsilon\) for some \(a \in (0, 1]\). Then any \(\epsilon\)-optimal solution \(\tilde{\theta}\) to (40) satisfies \(\tilde{\theta} \in \mathbb{C}(b\sqrt{k} + 1, \theta)\) where \(b = 2 + 2/a\). Furthermore, provided \(r > \kappa_\tau(b\sqrt{k} + 1, r, \theta)^{-1} \{s^{1/2} + (k + 1)^{1/2}a/2\}\) in Definition 2, we have
\[
\|\tilde{\theta} - \theta\|_2 \leq \frac{\|\nabla L_\tau(\theta)\|_2 + (k + 1)^{1/2} \epsilon}{\kappa_\tau(b\sqrt{k} + 1, r, \theta)} \tag{41}
\]
and
\[
\|\tilde{\theta} - \theta\|_2 \leq \frac{s^{1/2} + a(k + 1)^{1/2}a/2}{\kappa_\tau(b\sqrt{k} + 1, r, \theta)} \tag{42}
\]
where \(\nabla L_\tau(\theta)_{\xi} = (\partial_{\beta_\xi} L_\tau(\theta), \nabla L_\tau(\theta)_{\xi}^T)^T\).

**Proof of Lemma 2.** Define the \(\ell_2\)-ball in \(\mathbb{R}^{d+1}\) centered at \(\theta = (\beta_0, \beta^T)^T\): \(\mathbb{B}_2(t, \theta) = \{ \theta \in \mathbb{R}^{d+1} : \|\theta - \theta\|_2 \leq t\}\) for \(t > 0\). For some \(r > 0\) to be specified, let \(\tilde{\theta}_\eta = \eta\tilde{\theta} + (1 - \eta)\theta\) \((0 < \eta \leq 1)\) be an intermediate estimator such that (i) \(\tilde{\theta}_\eta \in \mathbb{B}_2(r, \theta)\), (ii) \(\tilde{\theta}_\eta\) lies on the boundary of \(\mathbb{B}_2(r, \theta)\) with \(0 < \eta < 1\) if \(\tilde{\theta} \notin \mathbb{B}_2(r, \theta)\), and (iii) \(\tilde{\theta}_\eta = \tilde{\theta}\) with \(\eta = 1\) if \(\tilde{\theta} \in \mathbb{B}_2(r, \theta)\).

Since the Huber loss is convex, from Lemma F.2 in Fan et al. (2018) we obtain that
\[
\langle \nabla L_\tau(\tilde{\theta}_\eta) - \nabla L_\tau(\tilde{\theta}), \tilde{\theta}_\eta - \tilde{\theta}\rangle \leq \eta\langle \nabla L_\tau(\tilde{\theta}) - \nabla L_\tau(\tilde{\theta}), \tilde{\theta} - \tilde{\theta}\rangle. \tag{43}
\]
First we bound the left-hand side of (43) from below. Under the assumed scaling, Lemma 1 indicates
\[
\| (\tilde{\beta} - \beta)_{\xi^c} \|_1 \leq (1 + 2/a)\| (\tilde{\beta} - \beta)_{\xi} \|_1 + |\tilde{\beta}_0 - \beta_0|.
\]
As a consequence, \( \| \tilde{\theta} - \theta \|_1 \leq (2 + 2/a) \|(\tilde{\beta} - \beta)\|_1 + 2|\tilde{\beta}_0 - \beta_0| \leq b(k + 1)^{1/2}\| \tilde{\theta} - \theta \|_2 \). According to Definition 2, this implies \( \tilde{\theta} \in \mathbb{C}(b\sqrt{k + 1}, \theta) \). Since \( \tilde{\theta}_T - \theta = \eta(\tilde{\theta} - \theta) \), we have \( \tilde{\theta}_T \in \mathbb{C}(b\sqrt{k + 1}, \theta) \cap \mathbb{B}_2(r, \theta) \) and

\[
\langle \nabla L_r(\tilde{\theta}_T) - \nabla L_r(\tilde{\theta}), \tilde{\theta}_T - \theta \rangle \geq \kappa_r(b\sqrt{k + 1}, r, \theta)\| \tilde{\theta}_T - \theta \|^2_2.
\] (44)

Next we upper bound the right-hand side of (43). For any \( \xi \in \partial\| \beta \|_1 \), write

\[
\langle \nabla L_r(\tilde{\theta}) - \nabla L_r(\theta), \tilde{\theta} - \theta \rangle
= \langle u, \tilde{\theta} - \theta \rangle - \langle \lambda \circ \xi, \tilde{\beta} - \beta \rangle - \langle \nabla L_r(\theta), \tilde{\theta} - \theta \rangle
:= \Pi_1 - \Pi_2 - \Pi_3,
\] (45)

where \( u = \nabla L_r(\tilde{\theta}) + (0, (\lambda \circ \xi)')^T \). For \( \Pi_3 = \langle \nabla L_r(\theta), \tilde{\theta} - \theta \rangle \), we write \( \tilde{\mathbf{v}} \in R_\xi(\tilde{v}_0, \mathbf{v}_T) \) for any \( \tilde{v} = (v_0, \mathbf{v}_T)^T \) for \( \tilde{v}_0 = \mathbf{v}_T \in \mathbb{R}^{d+1} \). Then it can be easily shown that

\[
\| \Pi_3 \| \leq \| \nabla L_r(\theta) \|_{\mathcal{E}} \| (\tilde{\beta} - \beta) \|_{\mathcal{E}} \| (\tilde{\beta} - \beta) \|_{\mathcal{E}} \|_1.
\] (46)

Tuning to \( \Pi_2 \), decompose \( \lambda \circ \xi \) and \( \tilde{\beta} - \beta \) according to \( \mathcal{S} \cup (\mathcal{E} \setminus \mathcal{S}) \cup \mathcal{E} \) to reach

\[
\Pi_2 = \langle (\lambda \circ \xi)\mathcal{S}, (\tilde{\beta} - \beta)\mathcal{S} \rangle + \langle (\lambda \circ \xi)\mathcal{E} \setminus \mathcal{S}, (\tilde{\beta} - \beta)\mathcal{E} \setminus \mathcal{S} \rangle + \langle (\lambda \circ \xi)\mathcal{E}, (\tilde{\beta} - \beta)\mathcal{E} \rangle.
\]

Since \( \beta_{\mathcal{E} \setminus \mathcal{S}} = 0 \) and \( \xi \in \partial\| \beta \|_1 \), we have \( \langle (\lambda \circ \xi)\mathcal{E}, (\tilde{\beta} - \beta)\mathcal{E} \rangle = \langle \lambda_{\mathcal{E} \setminus \mathcal{S}}, (\tilde{\beta} - \beta)\mathcal{E} \rangle = \langle \lambda_{\mathcal{E} \setminus \mathcal{S}}, (\tilde{\beta} - \beta)\mathcal{E} \rangle \rangle \geq 0 \). Also, \( \langle (\lambda \circ \xi)\mathcal{S} \setminus \mathcal{S}, (\tilde{\beta} - \beta)\mathcal{S} \setminus \mathcal{S} \rangle = \langle (\lambda \circ \xi)\mathcal{S}, (\tilde{\beta} - \beta)\mathcal{S} \rangle \leq 0 \). Therefore,

\[
\Pi_2 \geq -\| \lambda_{\mathcal{S}} \|_{\mathcal{S}} \| (\tilde{\beta} - \beta) \|_{\mathcal{S}} \| (\tilde{\beta} - \beta) \|_{\mathcal{S}} \|_1.
\] (47)

Similarly, \( \Pi_1 \) satisfies the bound

\[
\| \Pi_1 \| \leq \| u \|_{\mathcal{E}} \| (\tilde{\beta} - \beta) \|_{\mathcal{E}} \|_2 + \| u \|_{\mathcal{E}} \| (\tilde{\beta} - \beta) \|_{\mathcal{E}} \|_1.
\] (48)

Together, (45)–(48) yield

\[
\langle \nabla L_r(\tilde{\theta}) - \nabla L_r(\theta), \tilde{\theta} - \theta \rangle
\leq -\| \lambda_{\mathcal{E}} \|_{\mathcal{S}} \| (\tilde{\beta} - \beta) \|_{\mathcal{S}} \| (\tilde{\beta} - \beta) \|_{\mathcal{S}} \|_1
\]

\[
+ \| \nabla L_r(\theta) \|_{\mathcal{E}} \| (\tilde{\beta} - \beta) \|_{\mathcal{E}} \|_2 + \| \lambda_{\mathcal{S}} \|_{\mathcal{S}} \| (\tilde{\beta} - \beta) \|_{\mathcal{S}} \|_2.
\]

Taking the infimum over \( \xi \in \partial\| \beta \|_1 \) on both sides, it follows that

\[
\langle \nabla L_r(\tilde{\theta}) - \nabla L_r(\theta), \tilde{\theta} - \theta \rangle
\leq -\| \lambda_{\mathcal{E}} \|_{\mathcal{S}} \| (\tilde{\beta} - \beta) \|_{\mathcal{S}} \| (\tilde{\beta} - \beta) \|_{\mathcal{S}} \|_1
\]

\[
+ \| \nabla L_r(\theta) \|_{\mathcal{E}} \| (\tilde{\beta} - \beta) \|_{\mathcal{E}} \|_2 + (k + 1)^{1/2} \| (\tilde{\theta}_T - \theta) \|_{\mathcal{E}} \| (\tilde{\beta} - \beta) \|_{\mathcal{S}} \|_2.
\] (49)

Recall that \( \| \lambda_{\mathcal{E}} \|_{\mathcal{S}} \geq 2 \| \| \nabla L_r(\theta) \|_{\mathcal{S}} + \epsilon \) \), it follows from (43), (44) and (49) that

\[
\kappa_r(b\sqrt{k + 1}, r, \theta)\| \tilde{\theta}_T - \theta \|^2_2 \leq \| \nabla L_r(\theta) \|_{\mathcal{E}} \| (\tilde{\beta} - \beta) \|_{\mathcal{E}} \|_2 + (k + 1)^{1/2} \| \lambda_{\mathcal{S}} \|_{\mathcal{S}} \| (\tilde{\theta}_T - \theta) \|_2.
\] (50)

Under the assumed scaling, we have \( \| \nabla L_r(\theta) \|_{\mathcal{E}} \|_2 + (k + 1)^{1/2} \leq (k + 1)^{1/2} a\lambda / 2 \) and \( \| \lambda_{\mathcal{S}} \|_{\mathcal{S}} \leq s^{1/2} \lambda \). Substituting these bounds into (50) implies

\[
\| \tilde{\theta}_T - \theta \|_2 \leq \frac{s^{1/2} + (k + 1)^{1/2} a / 2}{\kappa_r(b\sqrt{k + 1}, r, \theta)} \lambda < r.
\]

Therefore, \( \tilde{\theta}_T \) falls in the interior of \( \mathbb{B}_2(r, \theta) \), enforcing \( \eta = 1 \) and \( \tilde{\theta}_T = \tilde{\theta} \). Consequently, (41) and (42) follow directly from (50) and the last display, respectively. \( \square \)
B Proofs of main results

B.1 Proof of Theorem 1

Part (I) is a direct consequence of Theorem 5 in Wang et al. (2018). In what follows, we provide a self-contained proof of (9). Write \( \hat{\theta} = \hat{\theta}^{\text{H-Lasso}} \) for simplicity. For \( t > 0 \), define the \( \ell_2 \)-ball \( \mathbb{B}_2(t, \theta^*_\tau) = \{ \theta \in \mathbb{R}^{d+1} : \| \theta - \theta^*_\tau \|_2 \leq t \} \) centered at \( \theta^*_\tau \) with radius \( t \), where \( \| \cdot \|_2 \) denotes the \( \ell_2 \)-norm defined by \( \|u\|_2 = \|\Sigma^{1/2}u\|_2 \) for \( u \in \mathbb{R}^{d+1} \). For some \( r > 0 \) to be specified, if \( \hat{\theta} \notin \mathbb{B}_2(r, \theta^*_\tau) \), there exists \( \eta \in (0, 1) \) such that \( \hat{\theta} := \theta^*_\tau + \eta(\hat{\theta} - \theta^*_\tau) \in \partial \mathbb{B}_2(r, \theta^*_\tau) \); otherwise if \( \hat{\theta} \in \mathbb{B}_2(r, \theta^*_\tau) \), we simply take \( \eta = 1 \) so that \( \hat{\theta} = \theta^*_\tau \). Applying Lemma 2 in Appendix C of Sun, Zhou and Fan (2017) to \( L_{\tau}(\theta) = (1/n) \sum_{i=1}^n \ell_{\tau}(y_i - \zeta_i^\top \theta) \) gives

\[
\langle \nabla L_{\tau}(\hat{\theta}) - \nabla \ell_{\tau}^*(\theta^*_\tau), \hat{\theta} - \theta^*_\tau \rangle \leq \eta \langle \nabla L_{\tau}(\hat{\theta}) - \nabla \ell_{\tau}^*(\theta^*_\tau), \hat{\theta} - \theta^*_\tau \rangle.
\] (51)

In the following, we deal with the left-hand and right-hand sides of (51) separately, starting with latter.

Write \( (\hat{v}_0, \hat{\nu})^\top = \hat{\theta} - \theta^*_\tau \). From Proposition 1 we see that \( \beta^*_\tau = \beta^* \) and \( \text{supp}(\beta^*_\tau) = S \). By the convexity of Huber loss, \( \hat{\theta} \) satisfies the Karush-Kuhn-Tucker (KKT) conditions, which say that \( \nabla L_{\tau}(\hat{\theta}) + \lambda \hat{z} = 0 \), where \( \hat{z} = (0, \hat{\nu}^\top)^\top \) and \( \hat{u} \in \partial \|\hat{\beta}\|_1 \). Then, under the scaling \( \lambda \geq 2\|\nabla L_{\tau}(\theta^*_\tau)\|_\infty \), it holds

\[
0 \leq \langle \nabla L_{\tau}(\hat{\theta}) - \nabla \ell_{\tau}^*(\theta^*_\tau), \hat{\theta} - \theta^*_\tau \rangle \\
= -\langle \nabla L_{\tau}(\theta^*_\tau), \hat{\theta} - \theta^*_\tau \rangle - \lambda(\hat{u}, \hat{\beta} - \beta^*) \\
\leq \lambda(\|\beta^*_\tau\|_1 - \|\hat{v} + \beta^*\|_1) + \frac{\lambda}{2}\|\hat{\theta} - \theta^*_\tau\|_1 \\
\leq \lambda(\|\beta^*_S\|_1 + \|\hat{v}_S\|_1 - \|\beta^*_S\|_1 - \|\hat{v}_S\|_1) + \frac{\lambda}{2}(\|\hat{v}_S\|_1 + \|\hat{v}_S\|_1 + \|\hat{v}_0\|) \\
= \frac{\lambda}{2}(3\|\hat{v}_S\|_1 - \|\hat{v}_S\|_1) + \frac{\lambda}{2}\|\hat{v}_0\| \leq \frac{3}{2}(s + 1)^{1/2}\lambda\|\hat{\theta} - \theta^*_\tau\|_2.
\] (52)

This in turn implies that provided \( \lambda \geq 2\|\nabla L_{\tau}(\theta^*_\tau)\|_\infty \), \( \hat{\theta} \) falls in the \( \ell_1 \)-cone \( C := \{ \theta \in \mathbb{R}^{d+1} : \|\nu_S\|_1 + \|v_0\|_1 + \nu_0 \} \) for \( (\nu, v^\top)^\top = \theta - \theta^*_\tau \).

Next we bound the left-hand side of (51) from below. Since \( \hat{\theta} - \theta^*_\tau = \eta(\hat{\theta} - \theta^*_\tau) \), we also have \( \hat{\theta} \in C \) as long as \( \lambda \geq 2\|\nabla L_{\tau}(\theta^*_\tau)\|_\infty \). The following proposition indicates that under certain constraints, the empirical Huber loss satisfies the restricted strong convexity condition over \( \mathbb{B}_2(r, \theta^*_\tau) \cap C \) with high probability.

Lemma 3. Write \( \nu^4 = \sup_{u \in S^d} \mathbb{E}(u, \Sigma^{-1/2}z)^4 \) and let \( r > 0 \) satisfy

\[
r \leq \frac{1}{8} \rho_{\tau}^{1/2} r^{-2}\tau \quad \text{and} \quad n \geq c_0 \Delta^{-1} \rho_{\tau}^{-2} (\nu_s \sigma \tau / r)^2 s \log d,
\] (53)

where \( c_0 > 0 \) is an absolute constant. Then with probability at least \( 1 - d^{-1} \),

\[
\langle \nabla L_{\tau}(\theta) - \nabla \ell_{\tau}^*(\theta^*_\tau), \theta - \theta^*_\tau \rangle \geq \frac{r_r}{2} \|\theta - \theta^*_\tau\|_\Sigma^2
\] (54)

holds uniformly over \( \theta \in \mathbb{B}_2(r, \theta^*_\tau) \cap C \).
In view of (52) and (54), we take \( r = \rho_r^{1/2} \nu^{-2} \tau/8 \) such that with probability at least \( 1 - d^{-1} \),
\[
\frac{\rho_r}{2} \Delta \Sigma \| \tilde{\theta} - \theta_r^* \|_\Sigma \leq \frac{\rho_r}{2} \| \tilde{\theta} - \theta_r^* \|_\Sigma \leq \frac{3}{2} (s+1)^{1/2} \lambda \| \tilde{\theta} - \theta_r^* \|_2.
\]
This, together with the fact that \( \tilde{\theta} \in \mathbb{C} \), implies
\[
\| \tilde{\theta} - \theta_r^* \|_\Sigma \leq 3 \Delta \Sigma \rho_r^{-1} (s+1)^{1/2} \lambda \quad \text{and} \quad \| \tilde{\theta} - \theta_r^* \|_1 \leq 12 \Delta \Sigma \rho_r^{-1} (s+1) \lambda \quad (55)
\]
provided \( \lambda \geq 2 \| \nabla \mathcal{L}_r(\theta_r^*) \|_\infty \) and (53) is met. In addition, if \( \tau > 24 \Delta \Sigma \rho_r^{-3/2} \nu^2 (s+1)^{1/2} \lambda \), then \( \| \tilde{\theta} - \theta_r^* \|_\Sigma < r \) with probability at least \( 1 - d^{-1} \). As a consequence, \( \tilde{\theta} \) and \( \tilde{\theta} \) must coincide and the error bounds in (55) hold for \( \tilde{\theta} \).

It remains to show that the constraint \( \lambda \geq 2 \| \nabla \mathcal{L}_r(\theta_r^*) \|_\infty \) is fulfilled with high probability. Recalling \( \theta_r^* = \arg \min_\theta \mathbb{E} \mathcal{L}_r(\theta) \), then by the convexity of \( \theta \mapsto \mathbb{E} \mathcal{L}_r(\theta) \) and Proposition 1, we have
\[
0 = \nabla \mathbb{E} \mathcal{L}_r(\theta_r^*) = \nabla \mathcal{L}_r(\theta_r^*) = -\frac{1}{n} \sum_{i=1}^n \mathbb{E} \{ \ell'_r(\epsilon_i - \alpha_r) z_i \}.
\]
This means that \( \nabla \mathcal{L}_r(\theta_r^*) = - (1/n) \sum_{i=1}^n \ell'_r(\epsilon_i - \alpha_r) z_i \) is a sum of independent mean zero random vectors. Further, by the definition of \( \alpha_r, \mathbb{E} \ell'_r(\epsilon - \alpha_r) = 0 \) and
\[
\mathbb{E} \{ \ell'_r(\epsilon - \alpha_r) \}^2 \leq 2 \mathbb{E} \ell_r(\epsilon - \alpha_r) \leq 2 \mathbb{E} \ell_r(\epsilon) \leq \sigma_r^2.
\]
Following the proof of Proposition C.2 in Wang et al. (2018), it can be similarly shown that for any \( x > 0 \),
\[
\| \nabla \mathcal{L}_r(\theta_r^*) \|_\infty \leq \nu_x \sigma_x \left( 2 \sigma_x \sqrt{\frac{x}{n}} + \frac{\tau x}{2n} \right) \quad (57)
\]
with probability at least \( 1 - 2(d+1)e^{-x} \). Taking \( x = 2 \log d \) in (57) we see that provided
\[
\lambda \geq \nu_x \sigma_x \left( 2 \sigma_x \sqrt{\frac{2 \log d}{n}} + \tau \log d \right),
\]
the event \( \{ \lambda \geq 2 \| \nabla \mathcal{L}_r(\theta_r^*) \|_\infty \} \) occurs with probability at least \( 1 - 2(d^{-1} + d^{-2}) \).

Putting together the pieces to obtain the stated result (9). \( \square \)

B.1.1 Proof of Lemma 3

Following the proof of Proposition B.1 in Wang et al. (2018), we have
\[
\langle \nabla \mathcal{L}_r(\theta) - \nabla \mathcal{L}_r(\theta_r^*), \theta - \theta_r^* \rangle \geq - \frac{1}{n} \sum_{i=1}^n \{ \ell'_r(y_i - z_i^* \theta_r^*) - \ell'_r(y_i - z_i^* \theta) \} z_i^* (\theta - \theta_r^*)_i \mathcal{E}_{i,r},
\]
where \( \mathcal{E}_{i,r} = \{ |\epsilon_i - \alpha_r| \leq \tau/2 \} \cap \{ |z_i^* (\theta - \theta_r^*)_i \|_\Sigma \leq \tau/(2r) \} \). On the event \( \mathcal{E}_{i,r} \),
\[
|y_i - z_i^* \theta_r^*| = |\epsilon_i + \beta_0^r - \beta_0^r r| = |\epsilon_i - \alpha_r| \leq \tau \quad \text{and} \quad |y_i - z_i^* \theta| \leq \tau \quad \text{for all} \quad \theta \in B_\Sigma(r, \theta_r^*). \]
Let
\( g(\theta) \) and \( \Delta_r \) be as in (B.19) and (B.20) in Wang et al. (2018) with the exception that \( \theta^* \) being replaced by \( \theta^* r \). For \( \theta \in B_2(r, \theta^* r) \), it holds

\[
\frac{\langle \nabla L_r(\theta) - \nabla L_r(\theta^* r), \theta - \theta^* \rangle}{\| \theta - \theta^* \|^2_{\Sigma}} \geq \frac{\mathbb{E} g(\theta)}{\| \theta - \theta^* \|^2_{\Sigma}} - \Delta_r.
\]

By the definition of \( \rho_r \) and Markov’s inequality, for every \( \theta \in B_2(r, \theta^* r) \) we have

\[
\mathbb{E} g(\theta) \geq (\rho_r - 16\nu^4 r^2 \tau^{-2})\| \theta - \theta^* \|^2_{\Sigma} \geq \frac{3}{4} \rho_r \| \theta - \theta^* \|^2_{\Sigma}
\]

as long as \( r \leq \rho_r^{1/2}/(8\nu^2) \). Moreover, it follows from Lemma C.1 in Wang et al. (2018) that for any \( x > 0 \),

\[
\Delta_r \leq \mathbb{E} \Delta_r + \{(\mathbb{E} \Delta_r)^{1/2} (2r) + \sqrt{\nu^2} \} \sqrt{\frac{x}{n}} + \frac{\tau^2 x}{48\tau^2 n}
\]

with probability at least \( 1 - e^{-x} \) and \( \mathbb{E} \Delta_r \leq \sqrt{2\pi} \{8\Lambda_{\Sigma}^{-1/2} (\tau/r) (s + 1)^{1/2} \mathbb{E} \| G_n \|_\infty + n^{-1/2} \} \), where \( G_n = (1/n) \sum_{i=1}^n g_i z_i \) and \( g_1, \ldots, g_n \) are i.i.d. standard normal random variables. It can be shown that \( \mathbb{E} \| G_n \|_\infty \lesssim \nu_2 \sigma_2 \{ \sqrt{\log(d)/n} + \log(d)/n \} \); see (89) for details. Taking \( x = \log d \) we obtain that with probability at least \( 1 - d^{-1} \), \( \Delta_r \leq \rho_r/4 \) provided \( n \gtrsim \lambda_\Sigma^{-1} \rho_r^{-2} (\nu_2 \sigma_2 \tau/r)^2 s \log d \). Combining this with (58) and (59) proves (54).

\[ \Box \]

### B.2 Proof of Proposition 2

With the initial estimate \( \bar{\theta}^{(0)} = 0_{d+1} \), we have \( \lambda^{(0)} = p_\lambda(0_d) = (\lambda, \ldots, \lambda)^\top \in \mathbb{R}^d \). Then (14) follows immediately from Lemma 2 with \( \theta = \theta^* \), \( \mathcal{E} = \mathcal{S} \) and \( a = 1 \).

\[ \Box \]

### B.3 Proof of Proposition 3

In order to improve the statistical rate at step \( \ell \geq 1 \), we need to control the magnitude of the spurious discoveries from the last step, that is, \( \max_{j \in \mathcal{S}} |\bar{\beta}_j^{(\ell-1)}| \). Recall that \( \lambda^{(\ell-1)} = (\lambda_1^{(\ell-1)}, \ldots, \lambda_d^{(\ell-1)})^\top = (p_\lambda((\bar{\beta}_1^{(\ell-1)})^\top), \ldots, p_\lambda((\bar{\beta}_d^{(\ell-1)})^\top))^\top \) and \( p_\lambda(t) = \lambda^2 p(t/\lambda) \) for \( t \in \mathbb{R} \). Intuitively, the larger \( |\bar{\beta}_j^{(\ell-1)}| \) is, the smaller \( \lambda_j^{(\ell-1)} \) is. Motivated by this observation, we construct an augmented set \( \mathcal{E}_\ell \) of \( \mathcal{S} \) in each step and control the magnitude of \( \| \lambda_{\mathcal{E}_\ell^{(\ell-1)}} \|_\min \).

With \( \bar{\theta}^{(0)} = 0 \), we have \( \lambda^{(0)} = (\lambda, \ldots, \lambda)^\top \in \mathbb{R}^d \). Under the scaling \( \lambda \geq 4 \{ \| \nabla L_r(\theta^*) \|_\infty + \epsilon_1 \} \), it follows from Lemma 2 with \( \mathcal{E} = \mathcal{S} \) that

\[
\| \tilde{\theta}^{(1)} - \theta^* \|_2 \leq \| \lambda_{\mathcal{S}}^{(0)} \|_2 + \| \nabla L_r(\theta^*) \|_2 \mathcal{S} + (|\mathcal{S}| + 1)^{1/2} \epsilon_1 \]

\[
\leq \frac{s^{1/2} + (s + 1)^{1/2} / 4}{\kappa_r (4\sqrt{s + 1}, r, \theta^*)} \lambda \leq \frac{1 + \sqrt{2} / 4}{\kappa_r (4\sqrt{s + 1}, r, \theta^*)} s^{1/2} \lambda.
\]

For \( \ell \geq 1 \), define

\[
\mathcal{E}_\ell = \mathcal{S} \cup \{ 1 \leq j \leq d : \lambda_j^{(\ell-1)} < p'(\gamma_0) \lambda \}.
\]
Note that the subset $E_{\ell}$ depends only on $\beta^{(l-1)}$. Provided
\[ |E_{\ell}| \leq 2s - 1 \quad \text{and} \quad \|\lambda_{S_{\ell}}^{(l-1)}\|_{\min} \geq p'(\gamma_{\delta})\lambda, \quad (62) \]
under the scaling $\lambda \geq \max\{4, \frac{2}{p'(\gamma_{\delta})}\}\|\nabla \mathcal{L}_r(\theta^*)\|_{\infty} + \epsilon_{\ell}$, it follows from Lemma 2 with $a = p'(\gamma_{\delta})$ and $b = 2 + \frac{2}{p'(\gamma_{\delta})}$ that
\[ \|\beta \theta^{(l)} - \theta^*\|_{2} \leq \frac{\|\lambda_{S_{\ell}}^{(l-1)}\|_{2} + \|\nabla \mathcal{L}_r(\theta^*)\|_{\infty} + (|E_{\ell}| + 1)^{1/2}\epsilon_{\ell}}{\kappa_{r}(b\sqrt{2s}, r, \theta^*)} \](63)
\[ \leq \frac{s^{1/2} + (2s)^{1/2}/4}{\kappa_{r}(b\sqrt{2s}, r, \theta^*)} \]
\[ \leq \frac{1 + \sqrt{2}/4}{\kappa_{r}(b\sqrt{2s}, r, \theta^*)} s^{1/2} \lambda \leq \delta s^{1/2} \gamma_{\delta}\lambda < r. \quad (64) \]

Claim (62) can be proved by induction as follows. For $\ell = 1$, we have $\lambda^{(0)} = (\lambda, \ldots, \lambda) \in \mathbb{R}^{d}$. Thus, (62) holds with $E_{1} = S$. Next, assume (62) holds for some $\ell \geq 1$, from which (64) follows. To bound the cardinality of $E_{\ell+1}$, note that for any $j \in E_{\ell+1} \setminus S$, $p'_{\lambda}(\beta^{(l)}_{j}) = \lambda^{(l)}_{j} < p'(\gamma_{\delta})\lambda = p'_{\lambda}(\gamma_{\delta}\lambda)$. This, together with the monotonicity of $p'_{\lambda}$ on $\mathbb{R}_{+}$, implies $|\beta^{(l)}_{j}| > \gamma_{\delta}\lambda$. Recalling that $\beta^{(l)}_{j} = 0$ for $j \in E_{\ell+1} \setminus S$, we obtain
\[ |E_{\ell+1} \setminus S|^{1/2} = \frac{\|\beta^{(l)}_{E_{\ell+1} \setminus S}\|_{2}}{\gamma_{\delta}\lambda} = \frac{\|(\beta^{(l)} - \theta^*)_{E_{\ell+1} \setminus S}\|_{2}}{\gamma_{\delta}\lambda} \leq \delta s^{1/2}. \quad (65) \]

Hence, $|E_{\ell+1}| \leq |S| + |E_{\ell+1} \setminus S| < (1 + \delta s)s < 2s$. In other words, $|E_{\ell+1}| \leq 2s - 1$. By (61) and the property $p'_{\lambda}(t) = \lambda p'(t/\lambda)$, we have $\lambda^{(l)}_{j} \geq p'(\gamma_{\delta})\lambda \geq 2\|\nabla \mathcal{L}_r(\theta^*)\|_{\infty} + \epsilon_{\ell+1}$ for $j \in E_{\ell}$. The two hypotheses in (62) then hold for $\ell + 1$, which completes the induction step. Consequently, (63) and (64) hold for any $\ell \geq 1$.

We have shown that under some scaling conditions, all the estimates $\beta^{(l)}$ fall in an $\ell_{2}$-ball centered at $\theta^*$ with radius of order $s^{1/2}\lambda$. To further refine this bound, in view of (63), it suffices to establish a sharper bound on $\|\lambda_{S}^{(l-1)}\|_{2} = \sqrt{\sum_{j \in S}(|\lambda_{j}^{(l-1)}|^{2})}$. For each $j \in [d]$, $\lambda_{j}^{(l-1)} = p'_{\lambda}(\beta^{(l-1)}_{j})$. If $|\beta^{(l-1)}_{j} - \beta^{*}_{j}| \geq \gamma_{\delta}\lambda$, then $\lambda_{j}^{(l-1)} \leq \gamma_{\delta}^{-1} |\beta^{(l-1)}_{j} - \beta^{*}_{j}|$; otherwise if $|\beta^{(l-1)}_{j} - \beta^{*}_{j}| \leq \gamma_{\delta}\lambda$, $\lambda_{j}^{(l-1)} \leq p'_{\lambda}(\beta^{*}_{j} - \gamma_{\delta}\lambda)$ due to monotonicity of $p'_{\lambda}$. Putting the pieces together, we conclude that
\[ \|\lambda_{S}^{(l-1)}\|_{2} \leq \|p'_{\lambda}(|\beta^{*}_{j} - \gamma_{\delta}\lambda)|\|_{2} + \gamma_{\delta}^{-1} \|[\beta^{(l-1)} - \beta^{*}]_{S}\|_{2}. \]

Plugging this into (63) gives
\[ \|\beta \theta^{(l)} - \theta^*\|_{2} \leq \frac{\|p'_{\lambda}(\beta^{*}_{j} - \gamma_{\delta}\lambda)\|_{2} + \|\nabla \mathcal{L}_r(\theta^*)\|_{\infty} + (|E_{\ell}| + 1)^{1/2}\epsilon_{\ell}}{\kappa_{r}(b\sqrt{2s}, r, \theta^*)} + \frac{1}{\kappa_{r}(b\sqrt{2s}, r, \theta^*)} \|[\beta^{(l-1)} - \beta^{*}]_{S}\|_{2}. \quad (66) \]
For the remaining terms that depend on $E_\ell \supseteq S$, by the triangle inequality and (65) we obtain that
\[
\|\nabla L_\tau(\theta^*)_{E_\ell}\|_2 + (|E_\ell| + 1)^{1/2}\epsilon_\ell
\]
\[
\leq \|\nabla L_\tau(\theta^*)_{S}\|_2 + (s + 1)^{1/2}\epsilon_\ell + |E_\ell \setminus S|^{1/2}\|\nabla L_\tau(\theta^*)\|_{\infty} + |E_\ell \setminus S|^{1/2}\epsilon_\ell
\]
\[
\leq \|\nabla L_\tau(\theta^*)_{S}\|_2 + (s + 1)^{1/2}\epsilon_\ell + \|\nabla L_\tau(\theta^*)\|_{\infty} + \epsilon_\ell \|\bar{\beta}^{(\ell-1)} - \beta^*\|_{E_\ell \setminus S} 2
\]
\[
\leq \|\nabla L_\tau(\theta^*)_{S}\|_2 + (s + 1)^{1/2}\epsilon_\ell + \frac{p'(\gamma_\delta)}{2}\|\bar{\beta}^{(\ell-1)} - \beta^*\|_{E_\ell \setminus S} 2.
\]

This, together with (15) and (66), proves the contraction property (16). Finally, (17) is a direct consequence of (16) and (60).

B.4 Proof of Proposition 5

Observe that the oracle estimator $\hat{\theta}_{\text{oracle}} = (\hat{\theta}_{\text{oracle}}^{0}, (\hat{\beta}_{\text{oracle}})^{\top})$ shares the active set with $\theta^*$, i.e. $\text{supp}(\hat{\beta}_{\text{oracle}}) = S$, and satisfies $\nabla L_\tau(\hat{\theta}_{\text{oracle}})_{S} = 0$. The proof strategy is similar to that in the proof of Proposition 3 with $\epsilon_\ell = 0$ since $\hat{\theta}^{(\ell)}$ are optimal solutions to $(P_\ell)$.

Under the constraints $\lambda \geq 4\|\nabla L_\tau(\hat{\theta}_{\text{oracle}})\|_{\infty}$ and $\|\hat{\theta}_{\text{oracle}} - \theta^*\|_{2} \leq r$, and following the proof of Proposition 3 with $\kappa_\tau$ replaced by $\bar{\kappa}_\tau$, we obtain that
\[
\|\hat{\theta}^{(\ell)} - \hat{\theta}_{\text{oracle}}\|_{2} \leq \frac{\|\lambda_{S}^{(\ell-1)}\|_{2} + \|\nabla L_\tau(\hat{\theta}_{\text{oracle}})_{E_\ell}\|_{2}}{\bar{\kappa}_\tau(b\sqrt{s} + 1, r, \theta^*, S)} \leq 1.25\bar{\kappa}_\tau^{-1}s^{1/2}\lambda \leq \frac{\delta \gamma_\delta s^{1/2}\lambda}{r},
\]
where, in view of (61) and (62), $E_\ell = S \cup \{1 \leq j \leq d : \lambda_{j}^{(\ell-1)} < p'(\gamma_\delta)\lambda\}$ is such that $|E_\ell| \leq 2s - 1$, and hence $|E_\ell \setminus S| \leq s - 1$. Moreover, define the subsets
\[
S_\ell = \{1 \leq j \leq d : |\hat{\beta}_j^{(\ell)} - \beta_j^*| \geq \gamma_\delta \lambda\}, \quad \ell = 0, 1, 2, \ldots.
\]
Starting with $\hat{\theta}^{(0)} = 0$, it holds under the minimum signal strength condition that $S_0 = S$.

For $\|\lambda_{S}^{(\ell-1)}\|_{2}$, note that if $j \in S \cap S_{\ell-1}$, $\lambda_{j}^{(\ell-1)} \leq p'_\lambda(\beta_j^* - \gamma_\delta \lambda)$ due to monotonicity; otherwise if $j \in S \cap S_{\ell-1}$, $\lambda_{j}^{(\ell-1)} \leq \lambda$. Therefore,
\[
\|\lambda_{S}^{(\ell-1)}\|_{2} \leq \|p'_\lambda(\beta_j^* - \gamma_\delta \lambda)\|_{2} + \lambda|S \cap S_{\ell-1}|^{1/2}.
\]
Since $\|\beta_j^*\|_{\text{min}} \geq \gamma_\delta \lambda + \gamma \lambda$ and $p'_\lambda(t) = 0$ for all $t \geq \gamma \lambda$, $\|p'_\lambda(\beta_j^* - \gamma_\delta \lambda)\|_{2}$ vanishes. Turning to $\|\nabla L_\tau(\hat{\theta}_{\text{oracle}})_{E_\ell}\|_{2}$, recall that $\nabla L_\tau(\hat{\theta}_{\text{oracle}})_{S} = 0$, we have
\[
\|\nabla L_\tau(\hat{\theta}_{\text{oracle}})_{E_\ell}\|_{2} = \|\nabla \hat{\theta}_{\tau}(\hat{\theta}_{\text{oracle}})_{E_\ell \setminus S}\|_{2} \leq \|\nabla L_\tau(\hat{\theta}_{\text{oracle}})\|_{\infty}|E_\ell \setminus S|^{1/2}.
\]
For each $j \in E_\ell \setminus S$, $\beta_j^* = 0$ and $\lambda_j^{(\ell-1)} = p'_\lambda(\beta_j^{(\ell-1)}) < p'(\gamma_\delta)\lambda = p'_\lambda(\gamma_\delta \lambda)$. Hence, $|\hat{\beta}_j^{(\ell-1)} - \beta_j| > \gamma_\delta \lambda$ so that $j \in S_{\ell-1} \setminus S$. As a consequence, $E_\ell \setminus S \subseteq S_{\ell-1} \setminus S$, which in turn implies
\[
\|\nabla L_\tau(\hat{\theta}_{\text{oracle}})_{E_\ell}\|_{2} \leq \|\nabla L_\tau(\hat{\theta}_{\text{oracle}})\|_{\infty}|S_{\ell-1} \setminus S|^{1/2}.
\]
Substituting the above estimates into (67) yields
\[
\| \hat{\theta}^{(t)} - \hat{\theta}_{\text{oracle}} \|_2 \leq \frac{|S \cap S_{t-1}|^{1/2} + |S_{t-1} \setminus S|^1/2}{\lambda} \\
\leq \frac{\sqrt{17}}{4} |S_{t-1}|^{1/2} \lambda \rho \delta.
\] (68)

Next we bound \( |S_{\ell}| (\ell \geq 1) \), the cardinality of \( S_{\ell} \). By (19), it holds for any \( j \in S_{\ell} \) that
\[
|\hat{\beta}_j^{(t)} - \hat{\beta}_{j,\text{oracle}}| \geq \gamma_{\delta} \lambda - \| \theta_{\text{oracle}} - \theta^* \|_{\infty} > \frac{\sqrt{17}}{4} \lambda \rho \delta.
\]
Together with (68), this implies
\[
|S_{\ell}|^{1/2} < \frac{\| \hat{\theta}^{(t)} - \hat{\theta}_{\text{oracle}} \|_2}{\frac{\sqrt{17}}{4} \lambda \rho \delta} \leq \delta |S_{t-1}|^{1/2}, \quad \ell \geq 1.
\]
Recall that \( S_0 = S \), we have \( |S_{\ell}|^{1/2} < \delta s^{1/2} \) for any \( \ell \geq 1 \). As long as \( \ell \geq T := \log(s^{1/2})/\log(1/\delta) \), \( |S_{\ell}| < 1 \) and thus \( S_{\ell} = \emptyset \). Consequently, it follows from (68) that \( \hat{\theta}^{(t)} = \theta_{\text{oracle}} \) for all \( \ell \geq T + 1 \). This completes the proof.

**B.5 Proof of Theorem 2**

We only prove (29) since (30) can be shown via the same argument with slight modifications. Let \( b = 2 + 2/p' \gamma_{\delta} \), \( r_0 = \tau/(8\sqrt{2}) \) and \( r = \bar{\Sigma}^{-1/2} r_0 \). Applying Proposition 7 with \( a = b\sqrt{2} \), we obtain that with probability at least \( 1 - d^{-1}, \kappa_{\tau}(a,r,\theta^*) \geq \bar{\Sigma}/2 = (1 + \sqrt{2}/4)/(\delta \gamma_{\delta}) \) as long as \( \tau \geq 4\sigma_{\epsilon} \) and \( n \geq \lambda^{-1}b^2\nu_{\bar{\Sigma}}^2 \sigma_{\epsilon}^2 s \log d \). Therefore, provided \( \tau > 8\bar{\Sigma}^{-1/2} \nu_{\bar{\Sigma}}^2 \sigma_{\epsilon}^2 s \log d \), (15) is fulfilled with probability at least \( 1 - d^{-1} \) under the assumed scaling. Next, with \( \tau \propto \sigma_{\epsilon} \sqrt{n/s} \), it follows from (25) that with probability at least \( 1 - 3d^{-1} \),
\[
\| \nabla L_{\tau}(\theta^*) \|_{\infty} \lesssim \nu_{\epsilon} \sigma_{\epsilon} \left( \sigma_{\epsilon} \sqrt{n} + \frac{\tau \log d}{n} \right).
\]
Letting \( \lambda \) scale with \( \nu_{\epsilon} \sigma_{\epsilon} \{ \sigma_{\epsilon} \sqrt{n} + \tau \log (d/n) \} \) to ensure that with probability at least \( 1 - 3d^{-1}, \lambda \geq \max\{ 4, 2/p'(\gamma_{\delta}) \} \{ \| \nabla L_{\tau}(\theta^*) \|_{\infty} + \epsilon_{\ell} \} \) for all \( \ell \geq 1 \). Putting together the pieces, we conclude that with probability at least \( 1 - 4d^{-1} \), the assumptions of Proposition 3 are fulfilled, under which it holds for any \( \ell \geq 1 \) that
\[
\| \hat{\theta}^{(t)} - \theta^* \|_1 \lesssim s^{1/2} \| \hat{\theta}^{(t)} - \theta^* \|_2 \quad \text{and} \quad \| \hat{\theta}^{(t)} - \theta^* \|_2 \lesssim \lambda^{-1} \{ \delta^{\ell-1}s^{1/2} \lambda + \| \nabla L_{\tau}(\theta^*) \|_2 + \sqrt{s/n} \}.
\]
Combined with (26) and the choice of \( (\lambda, \tau) \), this proves (29).

**B.6 Proof of Theorem 3**

*Proof of Part (I).* The proof is based primarily on Proposition 5, combined with complementary probabilistic analysis. We start with establishing the required statistical properties of the oracle estimator \( \hat{\theta}_{\text{oracle}} \) defined in (18). Since the oracle \( \hat{\theta}_{\text{oracle}} \) knows the actual support set \( S \), it is essentially an unpenalized Huber estimator computed from \( \{(y_i, z_i^S)\}_{i=1}^n \) satisfying \( y_i = z_i^S \theta_S^* + \epsilon_i \), where \( z_i^S = (1, x_i^S) \) and \( \theta_S^* = (\beta_0^S, \beta_S^T) \). Let \( S \) be as in (20),
and take $\tau \asymp \sigma_\varepsilon \sqrt{n/(s + \log n)}$. Then, it follows from Theorem B.1 in Sun, Zhou and Fan (2017) after slight modifications that, provided $n \gtrsim \nu_2^2(s + \log n)$,

$$
\| \Sigma^{1/2}(\hat{\theta}_{\text{oracle}} - \theta^*) \|_2 = \| \Sigma^{1/2}(\hat{\theta}_{\text{oracle}} - \theta^*)_S \|_2 \lesssim \sigma_\varepsilon \sqrt{\frac{s + \log n}{n}}
$$

(69)

holds with probability at least $1 - n^{-1}$. For the error under $\ell_\infty$-norm, it is obvious that

$$
\| \hat{\theta}_{\text{oracle}} - \theta^* \|_\infty = \| (\hat{\theta}_{\text{oracle}} - \theta^*)_S \|_\infty 
$$

$$
\leq \| (\hat{\theta}_{\text{oracle}} - \theta^*)_S \|_2 \leq 2^{1/2} \| \Sigma^{1/2}(\hat{\theta}_{\text{oracle}} - \theta^*)_S \|_2.
$$

(70)

Let $b = 2 + 2/p'(\gamma_\delta)$ be as in Proposition 5. We then need the following two technical lemmas, which provide lower and upper bounds for $\tilde{\kappa}_\tau(b\sqrt{s + 1}, r, \theta^*, S)$ and $\| \nabla L_{\tau}(\hat{\theta}_{\text{oracle}}) \|_\infty$, respectively.

**Lemma 4.** Write $\nu^4 = \sup_{u \in S^n} \mathbb{E}(u, \Sigma^{-1/2}z)^4$ and let $a \geq 1$, $r_0 > 0$ satisfy

$$
\tau \geq \max\{8\sqrt{2}\sigma_\varepsilon, 32\nu^2 r_0\} \text{ and } n \geq c_0 \max\{(\tau/r_0)^2s, \Sigma^{-1}(\nu_2 \sigma_\varepsilon r/r_0)^2 a^2 \log d\},
$$

(71)

where $c_0 > 0$ is an absolute constant. Then with probability at least $1 - d^{-1}$,

$$
\langle \nabla L_{\tau}(\theta_1) - \nabla L_{\tau}(\theta_2), \theta_1 - \theta_2 \rangle \geq \frac{1}{2} \| \theta_1 - \theta_2 \|_S^2
$$

(72)

holds uniformly over $(\theta_1, \theta_2) \in C(a, r_0) := \{(\theta, \theta') : \theta, \theta' = (\beta_0', \beta'\tau) \in \mathbb{B}_r(0, \theta^*), \| \theta - \theta' \|_1 \leq a \| \theta - \theta' \|_2, \text{supp}(\beta') = S\}$.

**Lemma 5.** With $\tau \asymp \sigma_\varepsilon \sqrt{n/(s + \log n)}$, it holds with probability at least $1 - 3d^{-1} - 2n^{-1}$ that

$$
\| \nabla L_{\tau}(\hat{\theta}_{\text{oracle}}) \|_\infty \lesssim \nu_2 \sigma_\varepsilon r \sqrt{\frac{\log d}{n}} + \lambda \Sigma^{-1/2} \sigma_\varepsilon \sqrt{\frac{s + \log n}{n}}
$$

(73)

as long as $n \gtrsim \nu_2^2(s + \log n)$.

According to Lemma 4, we take $r_0 = \tau/(32\nu^2)$, $r = \Sigma^{-1/2} r_0$ and $a = b\sqrt{s + 1}$. Then, under the sample size requirement $n \gtrsim \lambda \Sigma^{-1/2} \nu_2^2 \sigma_\varepsilon^2 \log d$, $\tilde{\kappa}_\tau(b\sqrt{s + 1}, r, \theta^*, S) \geq \lambda \Sigma/2$ with probability at least $1 - d^{-1}$. With the above choice of $r$, it follows from (69) that $\| \hat{\theta}_{\text{oracle}} - \theta^* \|_2 \leq r$ holds with probability at least $1 - n^{-1}$ as long as $n \gtrsim (\lambda \Sigma/\Delta \Sigma )^{1/2} \nu_2^2 s$. Taking $\tilde{\kappa} = \Delta \Sigma / 2$ in Proposition 5 and recall the inequality $\lambda \Sigma^{1/2} \| \hat{\theta}_{\text{oracle}} - \theta^* \|_\infty \leq \| \Sigma^{1/2}(\hat{\theta}_{\text{oracle}} - \theta^*)_S \|_2$, we see that if $\lambda$ satisfies

$$
\max\{4, \frac{2}{p'/(\gamma_\delta)}\} \| \nabla L_{\tau}(\hat{\theta}_{\text{oracle}}) \|_\infty \sqrt{2.5 \delta \lambda \Sigma^{1/2} \| (\hat{\theta}_{\text{oracle}} - \theta^*)_S \|_S} \leq \lambda < \frac{\tilde{\kappa} r}{1.25 s^{1/2}} = \frac{\Delta \Sigma^{1/2} \nu_2^2 s^{1/2}}{80 \lambda \Sigma^{1/2} \nu_2^2 s^{1/2}}.
$$

(74)

then $\hat{\theta}_\ell = \hat{\theta}_{\text{oracle}}$ for all $\ell \gtrsim \log(s^{1/2}/\delta)/\log(1/\delta)$. In fact, (69), (70) and Lemma 5 together imply that, with $\lambda \asymp \sigma_\varepsilon \sqrt{(s + \log n)/n} + \tau \log(d)/n$, $(74)$ holds with probability at least $1 - 3d^{-1} - 2n^{-1}$ as long as $n \gtrsim (\lambda \Sigma/\Delta \Sigma) \nu_2^2 s^{3/2}$. Putting together the pieces completes the
proof of Part (I).

**Proof of Part (II).** The proof strategy is similar in spirit to the previous one. Applying Theorem 4 in Wang et al. (2018) to $\hat{\beta}_S^{\text{oracle}} = (\hat{\beta}_0^{\text{oracle}}, (\hat{\beta}_S^{\text{oracle}})')^T$ with $\tau \asymp \sigma_z$ yields that under the scaling $n \gtrsim s + \log n$,

$$
\| (\hat{\theta}^{\text{oracle}} - \theta^*_r)_{S} \|_{\infty} = \| S^{1/2} (\hat{\theta}^{\text{oracle}} - \theta^*_r)_{S} \|_2 \lesssim \sigma_z \sqrt{s + \log n}
$$

with probability at least $1 - n^{-1}$.

Regarding the quantity $\bar{\kappa}(b\sqrt{s + 1}, r, \theta^*_r, S)$, the following lemma is a direct counterpart of Lemma 4.

**Lemma 6.** Write $\nu^4 = \sup_{u \in S^d} \mathbb{E}(u, \Sigma^{-1/2} z)^4$ and let $a \geq 1, r_0 > 0$ satisfy

$$
\tau \geq 16\sqrt{5} \rho^2 \nu^2 r_0 \text{ and } n \geq c_0 \rho^2 \max\{ (\tau/r_0)^2 s, \frac{\Delta \Sigma}{\sigma_z^2} (\nu_2 \sigma_z \tau/r_0)^2 a^2 \log d \},
$$

where $c_0 > 0$ is an absolute constant. Then with probability at least $1 - d^{-1}$,

$$
(\nabla L_{\tau}(\theta_1) - \nabla L_{\tau}(\theta_2), \theta_1 - \theta_2) \geq \frac{\rho^2}{2} \| \theta_1 - \theta_2 \|_2^2
$$

holds uniformly over $(\theta_1, \theta_2) \in \mathcal{C}(a, r_0)$.

As a consequence, we take $r_0 = \rho^{-1/2} \tau/(16\sqrt{5} \nu^2), r = \lambda^{-1/2} r_0$ and $a = b\sqrt{s + 1}$ such that under the scaling $n \gtrsim \rho^{-2} \Delta \Sigma^{-1} b^2 \nu^2 \sigma_z^2 s \log d$, $\bar{\kappa}(b\sqrt{s + 1}, r, \theta^*_r, S) \gtrsim \rho \Delta \Sigma/2$ with probability at least $1 - d^{-1}$. With the above choice of $r$ and $\tau \asymp \sigma_z$, $\| \theta^{\text{oracle}} - \theta^*_r \|_2 \leq r$ with probability at least $1 - n^{-1}$ provided $n \gtrsim \rho^{-1} (\Delta \Sigma/\Delta \Sigma) \nu^4 s$. From Proposition 5 with $\bar{\kappa} = \rho \Delta \Sigma/2$, we see that if $\lambda$ satisfies

$$
\max\left\{ 4, \frac{2}{\rho^2 \tau^2 (73) } \right\} \| \nabla L_{\tau}(\hat{\theta}^{\text{oracle}}) \|_{\infty} \sqrt{2.5 \rho^2 \Delta \Sigma^{-1/2}} \| (\hat{\theta}^{\text{oracle}} - \theta^*_r)_{S} \|_2
\leq \lambda \leq \frac{\bar{\kappa} r}{1.25 s^{1/2}} = \frac{\rho^2 \Delta \Sigma^{-1/2} \tau}{40 \sqrt{5} \nu^2 \Delta \Sigma^{1/2} s^{1/2}},
$$

(75)

then $\hat{\theta}(\ell) = \hat{\theta}^{\text{oracle}}$ for all $\ell \geq \log(s^{1/2}/\delta)/\log(1/\delta)$. It remains to bound $\| \nabla L_{\tau}(\hat{\theta}^{\text{oracle}}) \|_{\infty}$. Following the same arguments to those in the proof of Lemma 5, it can be shown that with probability at least $1 - n^{-1}$,

$$
\| \nabla L_{\tau}(\hat{\theta}^{\text{oracle}}) \|_{\infty} - \| \nabla L_{\tau}(\theta^*_r) \|_{\infty} \lesssim \lambda^{-1/2} \| (\hat{\theta}^{\text{oracle}} - \theta^*_r)_{S} \|_2
$$

as long as $n \gtrsim \nu^4 (s + \log n)$. Finally, in view of (27), we take $\lambda \asymp \sigma_z \sqrt{(s + \log d)/n}$ so that (75) holds with probability at least $1 - 3d^{-1} 2n^{-1}$, provided $n \gtrsim \max\{ s^2, s \log d \}$. Combining this with the lower bound on $\bar{\kappa}(b\sqrt{s + 1}, r, \theta^*_r, S)$ proves the stated result. \hfill \Box

**B.6.1 Proof of Lemma 4**

The proof is based on an argument similar to that given in the proof of Proposition 2 in Loh (2017). Since the magnitude of the robustification parameter $\tau$ plays a critical role in
subsequent analysis, we provide details of the proof that highlight the choice of \( \tau \) and its connection with the sample size scaling.

To begin with, note that

\[
\mathcal{T}(\theta_1, \theta_2) := \langle \nabla \mathcal{L}_\tau(\theta_1) - \nabla \mathcal{L}_\tau(\theta_2), \theta_1 - \theta_2 \rangle \\
= \frac{1}{n} \sum_{i=1}^{n} \{ \ell''_\tau(y_i - z^*_i \theta_2) - \ell''_\tau(y_i - z^*_i \theta_1) \} z^*_i (\theta_1 - \theta_2) \\
\geq \frac{1}{n} \sum_{i=1}^{n} \{ \ell''_\tau(y_i - z^*_i \theta_2) - \ell''_\tau(y_i - z^*_i \theta_1) \} z^*_i (\theta_1 - \theta_2) I_{\mathcal{E}_i},
\]

where \( I_{\mathcal{E}_i} \) is the indicator function of the event

\[
\mathcal{E}_i := \{ |\varepsilon_i| \leq \tau/4 \} \cap \{ |\langle z_i, \theta_2 - \theta^* \rangle| \leq \tau/2 \} \cap \{ |\langle z_i, \theta_1 - \theta_2 \rangle| \leq \tau/4 \},
\]

on which \( |y_i - z^*_i \theta_2| \leq |\varepsilon_i| + |z^*_i (\theta_2 - \theta^*)| \leq \tau/2 \) and \( |y_i - z^*_i \theta_1| \leq |z^*_i (\theta_1 - \theta_2)| + |z^*_i (\theta_2 - \theta^*)| + |\varepsilon_i| \leq \tau \) for all \( \theta_1, \theta_2 \in \mathbb{B}_2(r, \theta^*) \). For any \( R > 0 \), define two continuous functions

\[
\varphi_R(u) = \begin{cases} 
u^2 & \text{if } |u| \leq \frac{R}{2}, \\ (u-R)^2 & \text{if } \frac{R}{2} \leq u \leq R, \\ (u+R)^2 & \text{if } -R \leq u \leq -\frac{R}{2}, \\ 0 & \text{if } |u| > R, \end{cases} \quad \text{and } \phi_R(u) = \begin{cases} 1 & \text{if } |u| \leq \frac{R}{2}, \\ 2 - \frac{2}{R}u & \text{if } R/2 \leq u \leq R, \\ 2 + \frac{2}{R}u & \text{if } -R \leq u \leq -R/2, \\ 0 & \text{if } |u| > R, \end{cases}
\]

which are smoothed versions of \( u^2 I(|u| \leq R) \) and \( I(|u| \leq R) \), respectively. Recall that \( \ell''_\tau(u) = 1 \) for \( |u| \leq \tau \), then it follows from (76) that

\[
\mathcal{T}(\theta_1, \theta_2) \geq g(\theta_1, \theta_2) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\mathcal{E}_i} \varphi_{\tau|\theta_1-\theta_2|\Sigma/(4\tau)}(\langle z_i, \theta_1 - \theta_2 \rangle) \phi_{\tau/4}(\langle z_i, \theta_2 - \theta^* \rangle) I(|\varepsilon_i| \leq \tau/4) \\
= \mathbb{E}g(\theta_1, \theta_2) + g(\theta_1, \theta_2) - \mathbb{E}g(\theta_1, \theta_2).
\]

In what follows, we deal with \( \mathbb{E}g(\theta_1, \theta_2) \) and \( g(\theta_1, \theta_2) - \mathbb{E}g(\theta_1, \theta_2) \), separately.

Using the properties that

\[
u^2 I(|u| \leq R/2) \leq \varphi_R(u) \leq \nu^2 I(|u| \leq R) \quad \text{and } \phi_R(u) \geq I(|u| \leq R/2),
\]

we have

\[
\mathbb{E}g(\theta_1, \theta_2) \geq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\varphi_{\tau|\theta_1-\theta_2|\Sigma/(4\tau)}(\langle z_i, \theta_1 - \theta_2 \rangle) I(|\langle z_i, \theta_2 - \theta^* \rangle| \leq \tau/8) I(|\varepsilon_i| \leq \tau/4) \\
\geq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(\langle z_i, \theta_1 - \theta_2 \rangle)^2 I(|\langle z_i, \theta_1 - \theta_2 \rangle| \leq \tau \|\theta_1 - \theta_2\|_{\Sigma/(8\tau)}) \\
- \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(\langle z_i, \theta_1 - \theta_2 \rangle)^2 I(|\langle z_i, \theta_2 - \theta^* \rangle| > \tau/8) \\
- \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(\langle z_i, \theta_1 - \theta_2 \rangle)^2 I(|\varepsilon_i| > \tau/4).
\]
Write $\delta = \theta_1 - \theta_2$ for $\theta_1, \theta_2 \in S(r, \theta^*)$. By Markov’s inequality,

$$E(z_i, \delta)^2 I(|\{z_i, \delta\}| > \tau \|\delta\|_\Sigma/(8r)) \leq (8r/\tau)^2 E(z_i, \delta)^4/\|\delta\|^2_\Sigma \leq (8\nu^2r/\tau)^2 \|\delta\|^2_\Sigma,$$

$$E(z_i, \delta)^2 I(|\langle z_i, \theta_2 - \theta^*\rangle| > \tau/8) \leq (8/\tau)^2 E(z_i, \delta)^2 |\langle z_i, \theta_2 - \theta^*\rangle|^2 \leq (8\nu^2r/\tau)^2 \|\delta\|^2_\Sigma$$

and $E(z_i, \delta)^2 I(|\varepsilon_i| > \tau/4) \leq (4/\tau)^2 \sigma_2^2 \|\delta\|^2_\Sigma$. Substituting these into (79) yields

$$E g(\theta_1, \theta_2) \geq \{1 - 2(8\nu^2r/\tau)^2 - (4/\tau)^2 \sigma_2^2\} \|\delta\|^2_\Sigma.$$ 

Provided $\tau \geq \max\{8\sqrt{2}\sigma_x, 32\nu^2r\}$, this further implies

$$E g(\theta_1, \theta_2) \geq \frac{3}{4} \|\theta_1 - \theta_2\|^2_\Sigma \tag{80}$$

uniformly over $\theta_1, \theta_2 \in S(r, \theta^*)$.

To bound $g(\theta_1, \theta_2) - E g(\theta_1, \theta_2)$ uniformly over $(\theta_1, \theta_2) \in C(a, r)$, we define

$$\Delta = \sup_{(\theta_1, \theta_2) \in C(a, r)} \frac{|g(\theta_1, \theta_2) - E g(\theta_1, \theta_2)|}{\|\theta_1 - \theta_2\|^2_\Sigma}.$$ 

For $g(\theta_1, \theta_2)$ given in (78), write $g(\theta_1, \theta_2) = (1/n) \sum_{i=1}^{n} g_i(\theta_1, \theta_2)$. Since $\varphi_R(u) \leq R^2/4$ and $0 \leq \phi_R(u) \leq 1$ for all $u \in \mathbb{R}$, $0 \leq g_i(\theta_1, \theta_2) \leq (\tau/8r)^2 \|\theta_1 - \theta_2\|^2_\Sigma$. By Bousquet’s concentration theorem (Bousquet, 2003), for any $x > 0$,

$$\Delta \leq E\Delta + \sqrt{\frac{2x}{n}} \sqrt{\sigma_n^2 + 2(\tau/8r)^2 E\Delta} + (\tau/8r)^2 \frac{x}{3n} \tag{81}$$

with probability at least $1 - e^{-x}$, where $\sigma_n^2 := \sup_{(\theta_1, \theta_2) \in C(a, r)} E g(\theta_1, \theta_2)^2/\|\theta_1 - \theta_2\|^4_\Sigma$ is such that $\sigma_n^2 \leq \nu^4$. It suffices to bound the expected value $E\Delta$. Apply the symmetrization inequality for empirical processes and the connection between Gaussian complexity and Rademacher complexity, we obtain that

$$E\Delta \leq \sqrt{2\pi} E\left\{ \sup_{(\theta_1, \theta_2) \in C(a, r)} |G_{\theta_1, \theta_2}| \right\}, \tag{82}$$

where, with $\delta = \theta_1 - \theta_2$,

$$G_{\theta_1, \theta_2} := \frac{1}{n} \sum_{i=1}^{n} \frac{g_i}{\|\delta\|^2_\Sigma} \varphi_R|\delta|/\|\delta\|^2_\Sigma((z_i, \delta)) \phi_R/4((z_i, \theta_2 - \theta^*)) I(|\varepsilon_i| \leq \tau/4)$$

and $g_i$ are i.i.d. standard normal random variables that are independent of the data. Let $E^*$ be the conditional expectation given $\{(y_i, x_i)\}_{i=1}^{n}$. Apply the properties of Gaussian processes (Ledoux and Talagrand, 1991), for any $(\theta_1^*, \theta_2^*) \in C(a, r)$ fixed, it holds

$$E^*\left\{ \sup_{(\theta_1, \theta_2) \in C(a, r)} |G_{\theta_1, \theta_2}| \right\} \leq E^*|G_{\theta_1^*, \theta_2^*}| + 2E^*\left\{ \sup_{(\theta_1, \theta_2) \in C(a, r)} G_{\theta_1, \theta_2} \right\}.$$ 

Taking expectations with respect to $\{(y_i, x_i)\}_{i=1}^{n}$ on both sides, we also have

$$E\left\{ \sup_{(\theta_1, \theta_2) \in C(a, r)} |G_{\theta_1, \theta_2}| \right\} \leq E|G_{\theta_1^*, \theta_2^*}| + 2E\left\{ \sup_{(\theta_1, \theta_2) \in C(a, r)} \|G_{\theta_1, \theta_2}\| \right\}. \tag{83}$$
To bound \( \mathbb{E}[G_{\theta,\theta}^*] \), we take \( \theta_1^* = (\beta_0 + r, \beta^* \tau)^T \) and \( \theta_2^* = \theta^* = (\beta_0^*, \beta^* \tau)^T \) such that

\[
G_{\theta_1,\theta_2}^* = \frac{\varphi_{\tau/4}(r)}{r^2n} \sum_{i=1}^{n} g_i I(|\epsilon_i| \leq \tau/4).
\]

Provided \( \tau \geq 4r, \varphi_{\tau/4}(r) \leq r^2 \) and hence \( \mathbb{E}[G_{\theta_1,\theta_2}^*] \leq n^{-1/2} \).

Next, we apply the Gaussian comparison theorem to bound the expectation of the (conditional) Gaussian supremum \( \mathbb{E}^* \{ \sup_{(\theta_1,\theta_2) \in C(a,r)} G_{\theta_1,\theta_2} \} \), from which an upper bound for \( \mathbb{E}\{ \sup_{(\theta_1,\theta_2) \in C(a,r)} G_{\theta_1,\theta_2} \} \) follows immediately. Let \( \text{var}^* \) denote the conditional variance.

Given \( \{(y_i, x_i)\}_{i=1}^{n} \) and set \( \eta_i = f(|\epsilon_i| \leq \tau/4) \). For \( (\theta_1, \theta_2), (\theta_1', \theta_2') \in C(a,r) \), write \( \delta = \theta_1 - \theta_2 \), \( \delta' = \theta_1' - \theta_2' \), and note that

\[
G_{\theta_1,\theta_2} - G_{\theta'_1,\theta'_2} = G_{\theta_1,\theta_2} - G_{\theta'_1,\theta_2} + G_{\theta'_1,\theta_2} - G_{\theta'_1,\theta'_2} = \frac{1}{n} \sum_{i=1}^{n} \frac{g_i \varphi_{\tau/4}(\|\delta_i\|/\|\delta\|_{\Sigma}/(4r)) (\langle z_i, \delta \rangle)}{\delta_{i}^{2} \Sigma} \{ \varphi_{\tau/4}(\langle z_i, \theta_2 - \theta' \rangle) - \varphi_{\tau/4}(\langle z_i, \theta_2 - \theta' \rangle) \}
\]

Applying the Lipschitz property of \( \varphi_R \), i.e. \( |\varphi_R(u) - \varphi_R(v)| \leq 2|u - v|/R \), and recall that \( \varphi_R(u) \leq R^2/4 \), we have

\[
\text{var}^*(G_{\theta_1,\theta_2} - G_{\theta'_1,\theta'_2}) \leq \frac{1}{n^2} \sum_{i=1}^{n} \frac{\varphi_{\tau/4}(\|\delta_i\|/\|\delta\|_{\Sigma}/(4r)) (\langle z_i, \delta \rangle)^2}{\|\delta\|^{2}_{\Sigma}} (\frac{\varphi_{\tau/4}(\|\delta\|/\|\delta_{i}^{*}\|_{\Sigma}/(4r)) (\langle z_i, \delta \rangle)^2}{\|\delta_{i}^{*}\|^{2}_{\Sigma}} - \frac{\varphi_{\tau/4}(\|\delta\|/\|\delta\|_{\Sigma}/(4r)) (\langle z_i, \delta \rangle)^2}{\|\delta\|^{2}_{\Sigma}})
\]

Moreover, from the homogeneity and Lipschitz properties of \( \varphi_R \), i.e.

\[
\varphi_{cR}(cu) = c^2 \varphi_R(u) \quad \text{and} \quad |\varphi_R(u) - \varphi_R(v)| \leq R|u - v|, \quad u, v \in \mathbb{R}, c > 0,
\]

we deduce that

\[
\text{var}^*(G_{\theta_1,\theta_2} - G_{\theta'_1,\theta'_2}) \leq \frac{\tau^2}{(4r)^2 n^2} \sum_{i=1}^{n} \frac{\langle z_i, \delta \rangle}{\|\delta\|^{2}_{\Sigma}} (\frac{\langle z_i, \delta \rangle}{\|\delta\|^{2}_{\Sigma}} - \frac{\langle z_i, \delta \rangle}{\|\delta\|^{2}_{\Sigma}})^2.
\]

In view of (84), (85) and the inequality \( \text{var}(G_{\theta_1,\theta_2} - G_{\theta'_1,\theta'_2}) \leq 2\text{var}^*(G_{\theta_1,\theta_2} - G_{\theta'_1,\theta'_2}) + 2\text{var}^*(G_{\theta'_1,\theta'_2} - G_{\theta'_1,\theta'_2}) \), we define a second conditional Gaussian process indexed by
Together, \((82), (83), (86)\) and \((87)\) deliver the bound
\[
\mathbb{E}_\theta \left\{ \sup_{(\theta_1, \theta_2) \in \mathcal{C}(a, r)} \mathcal{G}_{\theta_1, \theta_2} \right\} \leq 2 \mathbb{E}_\theta \left\{ \sup_{(\theta_1, \theta_2) \in \mathcal{C}(a, r)} Z_{\theta_1, \theta_2} \right\},
\]
which also holds if \(\mathbb{E}_*\) is replaced by \(\mathbb{E}\). To bound the supremum of \(Z_{\theta_1, \theta_2}\), using the inequality \(\|\theta_2 - \theta_1\|_1 \leq a \|\theta_2 - \theta_1\|_2 \leq \frac{\Delta a}{2} \|\theta_2 - \theta_1\|_\Sigma\), we obtain
\[
\mathbb{E}\left\{ \sup_{(\theta_1, \theta_2) \in \mathcal{C}(a, r)} Z_{\theta_1, \theta_2} \right\} \leq \sqrt{2} \frac{2r}{8r^2} \mathbb{E} \left\{ \frac{1}{n} \sum_{i=1}^n g_i' S^{-1/2} z_i \right\} + \frac{\sqrt{2}a^2}{4} \frac{\rho}{\Delta a \Sigma r} \mathbb{E} \left\{ \frac{1}{n} \sum_{i=1}^n g_i'' z_i \right\}_\infty.
\]
Together, \((82), (83), (86)\) and \((87)\) deliver the bound
\[
\mathbb{E} \Delta \leq \sqrt{\frac{2}{n}} \left\{ \sqrt{\frac{2}{n}} + \frac{\tau}{r} \sqrt{\frac{s + 1}{n}} + \frac{2a\tau}{\Delta a \Sigma r} \mathbb{E} \left( \max_{0 \leq j \leq d} \left\{ \frac{1}{n} \sum_{i=1}^n g_i z_{ij} \right\} \right) \right\},
\]
where \(z_i = (z_{i0}, z_{i1}, \ldots, z_{id})^\top\). Finally, we bound the right-hand side of \((88)\). By the sub-Gaussianity of \(z\), Lemma 5.5 in Vershynin (2012) and the Legendre duplication formula, i.e. \(\Gamma(s)\Gamma(s + 1/2) = 2^{1-2s} \sqrt{\pi} \Gamma(2s)\), we calculate that \(\mathbb{E}(g_i z_{ij})^2 = \sigma_{jj}\) and for \(m \geq 3\),
\[
\mathbb{E}|g_i z_{ij}|^m = \mathbb{E}|g_i|^m \cdot \mathbb{E}|z_{ij}|^m \leq 2^{m/2} \Gamma(m + 1) \frac{m!}{\sqrt{\pi}} \cdot \nu_z^m \sigma_{jj}^{m/2} m \Gamma(m/2)
\leq \nu_z^m \sigma_{jj}^{m/2} 2^{m/2} 2^{1-m} \Gamma(m) = 2 \nu_z^m \sigma_{jj}^{m/2} (m - 1)! \left\{ \frac{m - 1}{2} \right\}^{m/2 - 1} \cdot \nu_z \left( \sigma_{jj} / 2 \right)^{m/2 - 2/2}.
\]
Hence, it follows from Lemma 14.12 in Bühlmann and van de Geer (2011) that
\[
\mathbb{E} \left( \max_{0 \leq j \leq d} \left\{ \frac{1}{n} \sum_{i=1}^n g_i z_{ij} \right\} \right) \leq \nu_z \sigma \left\{ \sqrt{\frac{2 \log(d + 2)}{n}} + \frac{\log(d + 2)}{n} \right\},
\]
Combining \((81), (88)\) and \((89)\), we determine that with probability at least \(1 - d^{-1}, \Delta \leq 1/4\) as long as \(n \geq \max\{\tau/r, s\} \Delta^{-1} \nu_z \sigma \sigma \sigma \tau/r a^2 \log d\). This, together with \((78)\) and \((80)\), proves the stated result \((72)\). \(\square\)
B.6.2 Proof of Lemma 5

By the generalized mean value theorem for vector-valued functions, we have

$$\nabla L_r(\theta_{\text{oracle}}) = \nabla L_r(\theta^*) + \left( \int_0^1 \nabla^2 L_r((1-t)\theta^* + t\theta_{\text{oracle}}) dt, \theta_{\text{oracle}} - \theta^* \right).$$

Note that \(\nabla^2 L_r(\theta) = (1/n) \sum_{i=1}^n I(y_i - z_i^T \theta) \leq \tau\) for \(\theta \in \mathbb{R}^{d+1}\). Then, by the triangle inequality and the Cauchy-Schwarz inequality,

$$\|\nabla L_r(\theta_{\text{oracle}})\|_\infty \leq \|\nabla L_r(\theta^*)\|_\infty + \left\| \int_0^1 \nabla^2 L_r((1-t)\theta^* + t\theta_{\text{oracle}}) dt, \theta_{\text{oracle}} - \theta^* \right\|_\infty,$$

$$\leq \|\nabla L_r(\theta^*)\|_\infty + \left\| \int_0^1 \nabla^2 L_r((1-t)\theta^* + t\theta_{\text{oracle}}) dt, \theta_{\text{oracle}} - \theta^* \right\|_2,$$

$$\leq \|\nabla L_r(\theta^*)\|_\infty + \|S^{1/2}(\theta_{\text{oracle}} - \theta^*)\|_2 \cdot \frac{1}{n} \left\| \sum_{i=1}^n z_i S_i^T z_i^\top \right\|_2,$$

where \(z_i = S^{-1/2} z_i S\) are i.i.d. isotropic sub-Gaussian random vectors. Recall that \(\theta_{\text{oracle}}\) satisfies the bound \(\|S^{1/2}(\theta_{\text{oracle}} - \theta^*)\|_2 \lesssim \sigma \sqrt{(s + \log n)/n}\) with probability at least \(1 - n^{-1}\). For \(\|/(1/n) \sum_{i=1}^n z_i z_i^\top\|_2\), it follows from Theorem 5.39 in Vershynin (2012) that

$$\left\| \frac{1}{n} \sum_{i=1}^n z_i z_i^\top - I_{s+1} \right\|_2 \lesssim \nu^2 \sqrt{\frac{s + \log n}{n}},$$

with probability at least \(1 - n^{-1}\). Combine these bounds with (25) proves (73). \(\square\)

B.6.3 Proof of Lemma 6

This proof is almost identical to that of Lemma 4, except that the event \(E_i\) in (77) and \(g(\theta_1, \theta_2)\) in (78) should now be replaced by

$$E_i = \{|\varepsilon_i - \alpha_r| \leq \tau/2\} \cap \{|(z_i, \theta_2 - \theta^*)| \leq \tau/4\} \cap \left\{ \frac{|(z_i, \theta_1 - \theta_2)|}{\|\theta_1 - \theta_2\|_{\Sigma}} \leq \frac{\tau}{8r} \right\},$$

and

$$g(\theta_1, \theta_2) := \frac{1}{n} \sum_{i=1}^n \varphi_{\rho \|\theta_1 - \theta_2\|_{\Sigma}/(8r)}(\langle z_i, \theta_1 - \theta_2 \rangle) \phi_{\tau/4}(\langle z_i, \theta_2 - \theta^* \rangle) I(|\varepsilon_i - \alpha_r| \leq \tau/2),$$

respectively. Similarly to (79), we have

$$\mathbb{E} g(\theta_1, \theta_2) \geq \frac{1}{n} \sum_{i=1}^n \mathbb{E} \phi_{\rho \|\theta_1 - \theta_2\|_{\Sigma}/(8r)}(\langle z_i, \theta_1 - \theta_2 \rangle) I(|(z_i, \theta_2 - \theta^*)| \leq \tau/8) I(|\varepsilon_i - \alpha_r| \leq \tau/2)$$

$$\quad \geq \frac{1}{n} \sum_{i=1}^n \mathbb{E} (z_i, \theta_1 - \theta_2)^2 I(|\varepsilon_i - \alpha_r| \leq \tau/2)$$

$$\quad - \frac{1}{n} \sum_{i=1}^n \mathbb{E} (z_i, \theta_1 - \theta_2)^2 I(|(z_i, \theta_2 - \theta^*)| > \tau/8)$$

$$\quad - \frac{1}{n} \sum_{i=1}^n \mathbb{E} (z_i, \theta_1 - \theta_2)^2 I\{|(z_i, \theta_1 - \theta_2)| > \tau \|\theta_1 - \theta_2\|_{\Sigma}/(16r)\}$$

$$\geq \{\rho \tau - (16\nu^2 r/\tau)^2 - (8\nu^2 r/\tau)^2\} \|\theta_1 - \theta_2\|^2_{\Sigma} \geq \frac{3}{4} \rho \tau \|\theta_1 - \theta_2\|^2_{\Sigma},$$

41
where the last inequality holds if \( \tau \geq 16\sqrt{5} \rho_r^{-1/2} \nu^2 r \). Keep all other statements the same, we then get the desired result.

\[ \square \]

### B.7 Proof of Theorem 6

The proof is almost identical to that of Theorem 3, except that (70) and (73) therein are now replaced, respectively, by (90) and (91) in the following lemma.

**Lemma 7.** There exist absolute constants \( C_1, C_2 > 0 \) such that the oracle estimator \( \hat{\theta}^{\text{oracle}} \) with \( \tau \asymp \sigma_\varepsilon \) satisfies

\[
\| \hat{\theta}^{\text{oracle}} - \theta^*_\tau \|_\infty \lesssim \Delta \sigma_\varepsilon \left( \frac{s + \log n}{n} + \sqrt{\frac{\log n}{n}} \right)
\]

(90)

with probability at least \( 1 - C_1 n^{-1} \), and

\[
\| \nabla L(\hat{\theta}^{\text{oracle}}) \|_\infty \lesssim \sigma_\varepsilon \left\{ \sqrt{\frac{\log d}{n}} + \frac{s + \log n}{n} \sqrt{\log n} \right\}
\]

(91)

with probability at least \( 1 - C_2(n^{-1} + d^{-1}) \).

Keep all other statements the same to reach the desired result.

\[ \square \]

#### B.7.1 Proof of Lemma 7

Since both \( \hat{\theta}^{\text{oracle}} \) and \( \theta^*_\tau \) are sparse, throughout the proof we view them as \( \mathbb{R}^{s+1} \)-valued vectors and regard \( L_\tau \) as a function defined on \( \mathbb{R}^{s+1} \).

We first prove (90). Applying Theorem B.1 in Sun, Zhou and Fan (2017) after slight adjustment, we obtain that the oracle \( \hat{\theta}^{\text{oracle}} \) with \( \tau \asymp \sigma_\varepsilon \) satisfies

\[
\| \hat{\theta}^{\text{oracle}} - \theta^*_\tau \|_S \lesssim \sigma_\varepsilon \sqrt{\frac{s + \log n}{n}}
\]

(92)

and

\[
\| S_{\tau}^{1/2} (\hat{\theta}^{\text{oracle}} - \theta^*_\tau) - S_{\tau}^{-1/2} \frac{1}{n} \sum_{i=1}^{n} \xi_i z_i S \|_2 \lesssim \sigma_\varepsilon \frac{s + \log d}{n}
\]

(93)

with probability at least \( 1 - n^{-1} \) as long as \( n \gtrsim s + \log n \), where \( S_{\tau} = \mathbb{E}\{ \ell''((\varepsilon - \alpha_\tau)/\tau) z_i S z_i^\top \} = \varrho_{\tau} S \) and \( \varrho_{\tau} = \mathbb{E}(\ell''((\varepsilon - \alpha_\tau)/\tau)) \geq c_2 \mathbb{P}(|\varepsilon - \alpha_\tau| \leq c_3 r) > 0 \). Here the constants \( c_2, c_3 \) are from Condition 5. Define independent random vectors \( w_i = S_{\tau}^{-1} z_i S \) such that \( \mathbb{E}(w_i w_i^\top) = \varrho_{\tau}^{-2} S^{-1} \). By the triangle inequality,

\[
\| \hat{\theta}^{\text{oracle}} - \theta^*_\tau \|_\infty \lesssim \frac{1}{\sqrt{\varrho_{\tau} \Delta \sigma_\varepsilon}} \left\| S_{\tau}^{1/2} (\hat{\theta}^{\text{oracle}} - \theta^*_\tau) - S_{\tau}^{-1/2} \frac{1}{n} \sum_{i=1}^{n} \xi_i w_i \right\|_2 + \left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i w_i \right\|_\infty.
\]

(94)
To bound the second term \( \|(1/n) \sum_{i=1}^{n} \zeta_i w_i \|_{\infty} \), let \( \| \cdot \|_{\psi_2} \) denote the sub-Gaussian norm for random variables/vectors (Vershynin, 2012), and note that for any \( a \in \mathbb{R}^{s+1} \),
\[
\| (a, w_i) \|_{\psi_2} = \vartheta_{\tau}^{-1} \| (S^{-1/2} a, z_{iS} z_{iS}^T) \|_{\psi_2} \leq \vartheta_{\tau}^{-1} \| S^{-1/2} a \|_2 \| S^{-1/2} z_{iS} \|_{\psi_2} \leq \vartheta_{\tau}^{-1} \nu \| S^{-1/2} a \|_2.
\]
Taking the supremum over \( a \in \mathbb{S}^s \), this implies \( \| w_i \|_{\psi_2} \leq \vartheta_{\tau}^{-1} \Delta S^{-1/2} \nu \). Following the same argument that leads to (27), it can be shown that with probability at least \( 1 - n^{-1} \), \( \|(1/n) \sum_{i=1}^{n} \zeta_i w_i \|_{\infty} \leq \vartheta_{\tau}^{-1} \Delta S^{-1/2} \nu \sigma \sqrt{\log(n)/n} \). Combining this with (93) and (94) proves (90).

Next we bound \( \| \nabla L_{\tau}(\hat{\theta}_{\text{oracle}}) \|_{\infty} \). By the mean value theorem for vector-valued functions, \( \nabla L_{\tau}(\hat{\theta}_{\text{oracle}}) \) can be written as
\[
\nabla L_{\tau}(\hat{\theta}_{\text{oracle}}) = \nabla L_{\tau}(\theta^*_\tau) + \int_0^1 (\nabla^2 L_{\tau}(\hat{\theta}_t), \hat{\theta}_{\text{oracle}} - \theta^*_\tau) dt,
\]
where \( \hat{\theta}_t = (1 - t)\theta^*_\tau + t \hat{\theta}_{\text{oracle}} \) and \( \nabla^2 L_{\tau}(\hat{\theta}_t) = (1/n) \sum_{i=1}^{n} \ell''(y_i - z_i^T \hat{\theta}_t)/\tau) z_{iS} z_{iS}^T \). Based on this expansion, we use the following inequality to bound \( \| \nabla L_{\tau}(\hat{\theta}_{\text{oracle}}) \|_{\infty} \) from above:
\[
\| \nabla L_{\tau}(\hat{\theta}_{\text{oracle}}) \|_{\infty} \leq \| \nabla L_{\tau}(\theta^*_\tau) \|_{\infty} + \left\| \int_0^1 (\nabla^2 L_{\tau}(\hat{\theta}_t) - S_\tau, \hat{\theta}_{\text{oracle}} - \theta^*_\tau) dt \right\|_{\infty} + \| S_\tau \|_{\infty} \| \hat{\theta}_{\text{oracle}} - \theta^*_\tau \|_{\infty}.
\]

In what to follow, we deal with the three terms on the right-hand side of (95) separately. By (27), it holds with probability at least \( 1 - d^{-1} \) that \( \| \nabla L_{\tau}(\theta^*_\tau) \|_{\infty} \leq \sigma \sqrt{\log(d)/n} \). Turn to the second term, decompose \( \nabla^2 L_{\tau}(\theta^*_\tau) - S_\tau \) as \( \nabla^2 L_{\tau}(\theta^*_\tau) = \nabla^2 L_{\tau}(\theta^*_\tau) + \nabla^2 L_{\tau}(\hat{\theta}_{\text{oracle}}) - S_\tau \). Under the condition that \( \ell'' \) is \( L_2 \)-Lipschitz, for any \( 0 \leq t \leq 1 \) it holds
\[
\| (\nabla^2 L_{\tau}(\hat{\theta}_t) - S_\tau, \hat{\theta}_{\text{oracle}} - \theta^*_\tau) \|_{\infty} \leq \frac{L_2^2}{\tau n} \sum_{i=1}^{n} \| z_{iS} \|_{\infty} (z_{iS}, \hat{\theta}_{\text{oracle}} - \theta^*_\tau)^2 \leq \frac{L_2^2}{\tau} \max_{1 \leq i \leq n} \| z_{iS} \|_{\infty} \| S^{-1/2} S S^{-1/2} \|_2 \| \hat{\theta}_{\text{oracle}} - \theta^*_\tau \|_{S},
\]
where \( \bar{S} := (1/n) \sum_{i=1}^{n} z_{iS} z_{iS}^T \). For every \( 1 \leq i \leq n \) and \( 1 \leq j \leq s \), \( z_{ij} \) is a sub-Gaussian random variable satisfying \( \mathbb{P}(|z_{ij}| \geq \nu \sigma_{ij}^{1/2} t) \leq 2e^{-t^2} \) for all \( t \in \mathbb{R} \). By the union bound, this implies
\[
\mathbb{P} \left( \max_{1 \leq i \leq n} \| z_{iS} \|_{\infty} \geq \nu \sigma_{ij}^{1/2} t \right) \leq 2sn \times e^{-t^2}.
\]
Provide \( n \geq s \), taking \( t = 2\sqrt{\log n} \) we obtain that with probability at least \( 1 - n^{-1} \),
\[
\max_{1 \leq i \leq n} \| z_{iS} \|_{\infty} \leq 2\nu \sigma \sqrt{n \log n}.
\]
For \( \nabla^2 L_{\tau}(\theta^*_\tau) - S_\tau \), since \( \ell'' \) is bounded from above, similarly to Theorem 5.39 in Vershynin (2012) it can be shown that
\[
\| S^{-1/2} (\nabla^2 L_{\tau}(\theta^*_\tau) - S_\tau) S^{-1/2} \|_2 \sqrt{\| S^{-1/2} \bar{S} S^{-1/2} - I_{s+1} \|_2} \lesssim \sqrt{s + \log n}/n
\]
(98)
with probability at least $1 - 2n^{-1}$. Together, (96), (97) and (98) imply

$$\left\| \int_0^1 \langle \nabla^2 \ell_r(\tilde{\theta}_t) - S_r, \hat{\theta}^{\text{oracle}}_r - \theta^*_r \rangle dt \right\|_\infty$$

$$\leq \left( L_2 / \tau \right) \max_{1 \leq i \leq n} \| z_i s \|_\infty \| S^{-1/2} SS^{-1/2} \|_2 \| \hat{\theta}^{\text{oracle}}_r - \theta^*_r \|_S$$

$$+ \lambda_S^{1/2} \| S^{-1/2} (\nabla^2 \ell_r(\theta^*_r) - S_r) S^{-1/2} \|_2 \| \hat{\theta}^{\text{oracle}}_r - \theta^*_r \|_S$$

$$\lesssim \left( L_2 / \tau \right) \sqrt{\log n} \| \hat{\theta}^{\text{oracle}}_r - \theta^*_r \|_S + \lambda_S^{1/2} \sqrt{\left( s + \log n \right) \left( \log n / n \right) \| \hat{\theta}^{\text{oracle}}_r - \theta^*_r \|_S}$$

with probability at least $1 - 3n^{-1}$ as long as $n \gtrsim s + \log n$.

It remains to bound $\| S_r(\hat{\theta}^{\text{oracle}}_r - \theta^*_r) \|_\infty$ in (95). Similar to (94), we have

$$\| S_r(\hat{\theta}^{\text{oracle}}_r - \theta^*_r) \|_\infty$$

$$\leq \| S_r(\hat{\theta}^{\text{oracle}}_r - \theta^*_r) - \frac{1}{n} \sum_{i=1}^n \xi_i z_i s \|_\infty + \frac{1}{n} \sum_{i=1}^n \xi_i z_i s \|_\infty$$

$$\leq \| S_r(\hat{\theta}^{\text{oracle}}_r - \theta^*_r) - \frac{1}{n} \sum_{i=1}^n \xi_i z_i s \|_2 + \frac{1}{n} \sum_{i=1}^n \xi_i z_i s \|_\infty$$

$$\leq (\rho, \lambda_S)^{1/2} \left\| S_r^{1/2} (\hat{\theta}^{\text{oracle}}_r - \theta^*_r) S_r - S_r^{-1/2} \frac{1}{n} \sum_{i=1}^n \xi_i z_i s \right\|_2 + \left\| \frac{1}{n} \sum_{i=1}^n \xi_i z_i s \right\|_\infty.$$
where $\zeta(k) = (0, \xi(k)^T)^T$ and $\tilde{\lambda} = (0, \lambda^T)^T$. For any $u$ such that $\|u\|_1 = 1$, we have

$$
\langle \nabla L(\theta(k)) + \tilde{\lambda} \odot \zeta(k), u \rangle = \langle \nabla L(\theta(k)), u \rangle - \langle \nabla L(\theta(k-1)) + \phi(k)(\theta(k) - \theta(k-1)), u \rangle
$$

$$
= \langle \nabla L(\theta(k)) - \nabla L(\theta(k-1)), u \rangle - \langle \phi(k)(\theta(k) - \theta(k-1)), u \rangle
$$

$$
\leq \|\nabla L(\theta(k)) - \nabla L(\theta(k-1))\|_\infty + \phi(k)\|\theta(k) - \theta(k-1)\|_\infty
$$

$$
\leq (\phi(k) + \rho_c)\|\theta(k) - \theta(k-1)\|_2,
$$

where the last inequality is due to the Lipschitz continuity of $\nabla L$. Taking the supremum over all $u$ satisfying $\|u\|_1 \leq 1$, we obtain

$$
\omega_\lambda(\theta(k)) \leq (\phi(k) + \rho_c)\|\theta(k) - \theta(k-1)\|_2.
$$

It remains to show that $\phi(k) \leq \gamma_u \rho_c$ for any $k$. This is guaranteed by the iterative LAMM algorithm. Otherwise, if $\phi(k) > \gamma_u \rho_c$, then $\phi' \equiv \phi(k)/\gamma_u > \rho_c$ is the quadratic parameter in the previous iteration for searching $\phi$ such that

$$
F(\tilde{\theta}(k); \phi', \theta(k-1)) < L(\tilde{\theta}(k))
$$

where $\tilde{\theta}(k)$ is the new updated parameter vector under the quadratic coefficient $\phi'$. On the other hand, it follows from the definition of $F$ and the Lipschitz continuity of $\nabla L$ that

$$
F(\tilde{\theta}(k); \phi', \theta(k-1)) + \lambda\|\theta\|_1 = L(\theta(k-1)) + \langle \nabla L(\theta(k-1)), \tilde{\theta}(k) - \theta(k-1) \rangle + \frac{\phi'}{2} \|\tilde{\theta}(k) - \theta(k-1)\|_2
$$

$$
> L(\theta(k-1)) + \langle \nabla L(\theta(k-1)), \tilde{\theta}(k) - \theta(k-1) \rangle + \frac{\rho_c}{2} \|\tilde{\theta}(k) - \theta(k-1)\|_2
$$

$$
\geq L(\tilde{\theta}(k)).
$$

This leads to a contradiction, indicating that $\phi(k) \leq \gamma_u \rho_c$. \hfill \Box

The second lemma is a modified version of Lemma E.4 in Fan et al. (2018). We reproduce its proof for completeness. Let $\Psi(\theta, \lambda) = L(\theta) + \|\lambda \circ \beta\|_1$ with $\lambda = \lambda(0)$.

**Lemma 9.** For any $\theta \in \mathbb{R}^{d+1}$, we have

$$
\Psi(\theta, \lambda) - \Psi(\theta(k), \lambda) \geq \frac{\phi(k)}{2} \{\|\theta - \theta(k)\|_2^2 - \|\theta - \theta(k-1)\|_2^2\}.
$$

**Proof of Lemma 9.** Because $F(\theta; \phi(k), \theta(k-1))$ majorizes $L(\theta)$ at $\theta(k)$, we have

$$
\Psi(\theta, \lambda) - \Psi(\theta(k), \lambda) \geq \Psi(\theta(k), \lambda) - \{F(\theta(k); \phi(k), \theta(k-1)) + \|\lambda \circ \beta(k)\|_1\}.
$$

By the convexity of $L(\theta)$ and $\|\lambda \circ \beta\|_1$,

$$
L(\theta) \geq L(\phi(k)) + \langle \nabla L(\theta(k-1)), \theta - \theta(k-1) \rangle + \frac{\phi'}{2} \|\theta(k) - \theta(k-1)\|_2^2
$$

and $\|\lambda \circ \beta\|_1 \geq \|\lambda \circ \beta(k)\|_1 + \langle \lambda \circ \xi(k), \beta - \beta(k) \rangle$

for any $\xi(k) \in \partial \|\beta(k)\|_1$. Together, (100) and (101) imply

$$
\Psi(\theta, \lambda) \geq L(\theta(k-1)) + \langle \nabla L(\theta(k-1)), \theta - \theta(k-1) \rangle + \frac{\phi(k)}{2} \|\theta(k) - \theta(k-1)\|_2^2
$$

$$
+ \|\lambda \circ \beta(k)\|_1 + \langle \lambda \circ \xi(k), \beta - \beta(k) \rangle.
$$
Plugging the expression of $F(\theta^{(k)}; \phi^{(k)}, \theta^{(k-1)})$ in (33) and (102) into (99), we obtain

$$
\Psi(\theta, \lambda) - \Psi(\theta^{(k)}, \lambda) \geq -\frac{\phi^{(k)}}{2} \|\theta^{(k)} - \theta^{(k-1)}\|_2^2 + \langle \nabla \mathcal{L}(\theta^{(k-1)}), \theta - \theta^{(k)} \rangle + \langle \lambda \circ \xi^{(k)}, \beta - \beta^{(k)} \rangle.
$$

(103)

By the first-order optimality condition, there exists some $\xi \in \partial \|\beta^{(k)}\|_1$ such that

$$
\nabla \mathcal{L}(\theta^{(k-1)}) + \phi^{(k)}(\theta^{(k)} - \theta^{(k-1)}) + \lambda \circ \xi^{(k)} = 0,
$$

where $\xi^{(k)} = (0, \xi^T)^T$. Substituting this into (103), and by direct calculations, we arrive at the conclusion.

Recall that $\Psi(\theta, \lambda) = \mathcal{L}(\theta) + \|\lambda \circ \beta\|_1$ and $\hat{\theta}^{(1)} \in \min_{\theta} \Psi(\theta, \lambda)$ denotes the optimal solution at the contraction stage.

**Lemma 10.** For any $k \geq 1$, we have

$$
\Psi(\theta^{(k)}, \lambda) - \Psi(\hat{\theta}^{(1)}, \lambda) \leq \frac{\max_{1 \leq j \leq k} \phi^{(j)}}{2k} \|\theta^{(0)} - \hat{\theta}^{(1)}\|_2^2.
$$

Proof of Lemma 10. For simplicity, we write $\hat{\theta} = \hat{\theta}^{(1)}$, and define $\phi_{\max} = \max_{1 \leq j \leq k} \phi^{(j)}$ and $\phi_{\min} = \min_{1 \leq j \leq k} \phi^{(j)} > 0$. Taking $\theta = \hat{\theta}$ in Lemma 9 gives

$$
0 \geq \Psi(\hat{\theta}, \lambda) - \Psi(\theta^{(j)}, \lambda) \geq \frac{\phi^{(j)}}{2} \{\|\hat{\theta} - \theta^{(j)}\|_2^2 - \|\hat{\theta} - \theta^{(j-1)}\|_2^2\}
$$

for all $j \geq 1$. Summing over $j$ from 1 to $k$ yields

$$
\sum_{j=1}^{k} \frac{2}{\phi^{(j)}} \{\Psi(\hat{\theta}, \lambda) - \Psi(\theta^{(j)}, \lambda)\} \geq \sum_{j=1}^{k} \{\|\theta^{(j)} - \hat{\theta}\|_2^2 - \|\theta^{(j-1)} - \hat{\theta}\|_2^2\},
$$

which further implies

$$
\frac{2}{\phi_{\max}} \left\{ k\Psi(\hat{\theta}, \lambda) - \sum_{j=1}^{k} \Psi(\theta^{(j)}, \lambda) \right\} \geq \|\theta^{(k)} - \hat{\theta}\|_2^2 - \|\theta^{(0)} - \hat{\theta}\|_2^2\}
$$

(104)

Again, by Lemma 9 with $\theta = \theta^{(j-1)}$ and $k = j$,

$$
\Psi(\theta^{(j-1)}, \lambda) - \Psi(\theta^{(j)}, \lambda) \geq \frac{\phi^{(j)}}{2} \|\theta^{(j)} - \theta^{(j-1)}\|_2^2 \geq \frac{\phi_{\min}}{2} \|\theta^{(j)} - \theta^{(j-1)}\|_2^2.
$$

Multiplying both sides of the above inequality by $j - 1$ and summing over $j$, we obtain

$$
\frac{2}{\phi_{\min}} \sum_{j=1}^{k} ((j-1)\Psi(\theta^{(j-1)}, \lambda) - j\Psi(\theta^{(j)}, \lambda) + \Psi(\theta^{(j)}, \lambda)) \geq \sum_{j=1}^{k} (j-1)\|\theta^{(j)} - \theta^{(j-1)}\|_2^2,
$$

or equivalently,

$$
\frac{2}{\phi_{\min}} \left\{ -k\Psi(\theta^{(k)}, \lambda) + \sum_{j=1}^{k} \Psi(\theta^{(j)}, \lambda) \right\} \geq \sum_{j=1}^{k} (j-1)\|\theta^{(j)} - \theta^{(j-1)}\|_2^2.
$$

(105)
Putting (104) and (105) together to reach

$$\frac{2k}{\phi_{\min}} \{ \Psi(\hat{\theta}, \lambda) - \Psi(\theta^{(k)}, \lambda) \}$$

$$\geq \frac{\phi_{\max}}{\phi_{\min}} \| \theta^{(k)} - \hat{\theta} \|_2^2 + \sum_{j=1}^{k} (j - 1) \| \theta^{(j)} - \theta^{(j-1)} \|_2^2 - \frac{\phi_{\max}}{\phi_{\min}} \| \theta^{(0)} - \hat{\theta} \|_2^2,$$

from which it follows immediately that

$$\frac{2k}{\phi_{\max}} \{ \Psi(\theta^{(k)}, \lambda) - \Psi(\hat{\theta}, \lambda) \} \leq \| \theta^{(0)} - \hat{\theta} \|_2^2.$$

This completes the proof. \hfill \square

**B.8.2 Proof of the theorem**

Recall that \( \theta^{(k)} = \theta^{(1,k)} \) and \( \phi^{(k)} = \phi^{(1,k)} \). By Lemma 8 and its proof,

$$\omega_{\lambda}(\theta^{(k)}) \leq (\phi^{(k)} + \rho_c) \| \theta^{(k)} - \theta^{(k-1)} \|_2 \leq \rho_c(1 + \gamma_u) \| \theta^{(k)} - \theta^{(k-1)} \|_2.$$

Next, taking \( \theta = \theta^{(k-1)} \) in Lemma 9 yields

$$\Psi(\theta^{(k-1)}, \lambda) - \Psi(\theta^{(k)}, \lambda) \geq \frac{\phi^{(k)}}{2} \| \theta^{(k-1)} - \theta^{(k)} \|_2^2.$$

Together, the last two displays lead to a bound for the suboptimality measure

$$\omega_{\lambda}(\theta^{(k)}) \leq \rho_c(1 + \gamma_u) \left[ \frac{2}{\phi^{(k)}} \{ \Psi(\theta^{(k-1)}, \lambda) - \Psi(\theta^{(k)}, \lambda) \} \right]^{1/2} \leq \rho_c(1 + \gamma_u) \sqrt{\frac{\phi^{(k)}}{\phi^{(k-1)}}} \| \hat{\theta} \|_2.$$

Recall that \( \{ \Psi(\theta^{(k)}, \lambda) \}_{k=0}^{\infty} \) is a non-increasing sequence, i.e.

$$\Psi(\hat{\theta}^{(1)}, \lambda) \leq \cdots \leq \Psi(\theta^{(k)}, \lambda) \leq \cdots \leq \Psi(\theta^{(0)}, \lambda).$$

Then it follows from (106) and Lemma 10 that

$$\omega_{\lambda}(\theta^{(k)}) \leq \rho_c(1 + \gamma_u) \left[ \frac{2}{\phi^{(k)}} \{ \Psi(\theta^{(k-1)}, \lambda) - \Psi(\hat{\theta}, \lambda) \} \right]^{1/2}$$

$$\leq \frac{\rho_c(1 + \gamma_u)}{\sqrt{k-1}} \sqrt{\max_{1 \leq j \leq k-1} \frac{\phi^{(j)}}{\phi^{(k)}}} \| \hat{\theta} \|_2,$$

where we used the fact that \( \theta^{(0)} = 0 \). By the triangle inequality,

$$\omega_{\lambda}(\theta^{(k)}) \leq \frac{\rho_c(1 + \gamma_u)}{\sqrt{k}} (\| \theta^{*} \|_2 + \| \hat{\theta} - \theta^{*} \|_2).$$

Therefore, in the contraction stage, we need \( k \geq \{ \rho_c(1 + \gamma_u) (\| \theta^{*} \|_2 + \| \hat{\theta} - \theta^{*} \|_2) / \epsilon_c \}^2 \) to ensure \( \omega_{\lambda}(\theta^{(k)}) \leq \epsilon_c \). This proves the stated result. \hfill \square
B.9 Proof of Theorem 5

For convenience, we omit the index $\ell$, and use $\hat{\theta}, \theta^{(k)}, \lambda$ and $\mathcal{E}$ to denote $\hat{\theta}^{(\ell)}, \theta^{(\ell,k)}, \lambda^{(\ell-1)}$ and $\mathcal{E}_\ell$, respectively. Moreover, write $\mathcal{L}(\theta) = \mathcal{L}_s(\theta)$, and define $\Psi(\theta, \lambda) = \mathcal{L}(\theta) + \|\lambda \circ \beta\|_1 = \mathcal{L}(\theta) + \|\lambda^{(\ell-1)} \circ \beta\|_1$ so that $\hat{\theta} \in \min_\theta \Psi(\theta, \lambda)$.

B.9.1 Technical lemmas

We first provide several technical lemmas along with their proofs. Recall the sparse cone $\mathbb{C}(m, r, \tau)$ given in (37).

Lemma 11. For any $\theta_1, \theta_2 \in \mathbb{C}(m/2, r, \tau) \cap \mathbb{B}_2(r, \theta^*)$, we have

$$\frac{1}{2} \rho_-(m, r, \tau)\|\theta_1 - \theta_2\|^2 \leq D_\mathcal{L}(\theta_1, \theta_2) \leq \frac{1}{2} \rho_+(m, r, \tau)\|\theta_1 - \theta_2\|^2,$$

where $D_\mathcal{L}(\theta_1, \theta_2) := \mathcal{L}(\theta_1) - \mathcal{L}(\theta_2) - \langle \nabla \mathcal{L}(\theta_2), \theta_1 - \theta_2 \rangle$.

Proof of Lemma 11. By a second-order Taylor series expansion, there exists some $\gamma \in [0, 1]$ such that $\hat{\theta} = \gamma \theta_1 + (1 - \gamma)\theta_2 \in \mathbb{B}_2(r, \theta^*)$ and $D_\mathcal{L}(\theta_1, \theta_2) = (1/2)(\theta_1 - \theta_2)^T \nabla^2 \mathcal{L}(\theta)(\theta_1 - \theta_2)$.

The stated bounds then follow directly from Definition 3.

The next lemma converts the bound on $\Psi(\theta, \lambda) - \Psi(\theta^*, \lambda)$ to that on $\|\theta - \theta^*\|_2$. Recall that for any subset $\mathcal{E} \subseteq [d]$, we write $\theta_{\mathcal{E}} = (\theta_0, \theta_\mathcal{E})^T$ and $\theta_{\mathcal{E}^c} = \beta_{\mathcal{E}^c}$.

Lemma 12. Assume Condition 4 holds. Let $\mathcal{E} \subseteq [d]$ be a subset satisfying $\mathcal{S} \subseteq \mathcal{E}$ and $|\mathcal{E}| \leq 2s$. Assume further that $\lambda \geq \max\{4\|\nabla \mathcal{L}(\theta^*)\|_\infty, \|\lambda\|_\infty\}$ and $\|\lambda_{\mathcal{E}^c}\|_{\min} \geq \lambda/2$. Then, for any $\theta \in \mathbb{B}_2(r, \theta^*)$ satisfying $\|\theta_{\mathcal{E}^c}\|_0 \leq \tilde{s}$ and $\Psi(\theta, \lambda) - \Psi(\theta^*, \lambda) \leq C\lambda^2 s$, we have

$$\|\theta - \theta^*\|_2 \leq C_1 \lambda \sqrt{\tilde{s}} \quad \text{and} \quad \|\theta - \theta^*\|_1 \leq C_2 \lambda s,$$

where $C_1, C_2 > 0$ depend only on $C$ and localized sparse eigenvalues.

Proof of Lemma 12. We omit the arguments in $\rho_-(m, r, \tau)$ and $\rho_+(m, r, \tau)$ whenever there is no ambiguity. Using Lemma 11, we immediately obtain

$$\mathcal{L}(\theta^*) + \langle \nabla \mathcal{L}(\theta^*), \theta - \theta^* \rangle + \frac{\rho}{2} \|\theta - \theta^*\|^2 \leq \mathcal{L}(\theta).$$

Because $\Psi(\theta) - \Psi(\theta^*) \leq C\lambda^2 s$, or equivalently,

$$\mathcal{L}(\theta) - \mathcal{L}(\theta^*) + (\|\lambda \circ \beta\|_1 - \|\lambda \circ \beta^*\|_1) \leq C\lambda^2 s,$$

it holds

$$\frac{\rho}{2} \|\theta - \theta^*\|^2 \leq C\lambda^2 s - \langle \nabla \mathcal{L}(\theta^*), \theta - \theta^* \rangle + (\|\lambda \circ \beta^*\|_1 - \|\lambda \circ \beta\|_1).$$

After some simple algebra, it can be derived that

$$\Pi \leq \|(\theta - \theta^*)_{\mathcal{E}^c}\|_1 \sum_{i=1}^{\|\mathcal{L}(\theta^*)\|_\infty,} \|(\theta - \theta^*)_{\mathcal{E}^c}\|_1 \sum_{i=1}^{\|\mathcal{L}(\theta^*)\|_\infty,} \|(\beta - \beta^*)_{\mathcal{E}^c}\|_1.$$
Combining the above bounds
\[
\frac{\rho}{2} \|\theta - \theta^*\|_2^2 + \{\lambda/2 - \|\nabla L(\theta^*)\|_\infty\}\|\theta - \theta^*\|_\infty \leq \{\lambda + \|\nabla L(\theta^*)\|_\infty\}\|\theta - \theta^*\|_1 + C\lambda^2 s.
\]
which further implies
\[
\frac{\rho}{2} \|\theta - \theta^*\|_2^2 \leq \frac{5\lambda}{4}\|\theta - \theta^*\|_1 + C\lambda^2 s.
\]
To further bound the right-hand side of the above inequality, we discuss two cases regarding the magnitude of \(\|\theta - \theta^*\|_1\) comparing with \(\lambda s\):

- If \(5\lambda \|\theta - \theta^*\|_1 / 4 \leq C\lambda^2 s\), we have
  \[
  \frac{\rho}{2} \|\theta - \theta^*\|_2^2 \leq 2C\lambda^2 s, \quad \text{and thus } \|\theta - \theta^*\|_2 \leq 2\sqrt{\frac{C}{\rho} \lambda \sqrt{s}}. \quad (108)
  \]
- If \(5\lambda \|\theta - \theta^*\|_1 / 4 > C\lambda^2 s\), we have
  \[
  \frac{\rho}{2} \|\theta - \theta^*\|_2^2 \leq \frac{5\lambda}{2}\|\theta - \theta^*\|_1 \leq \frac{5}{2}\lambda (2s + 1)^{1/2}\|\theta - \theta^*\|_2,
  \]
which further yields
\[
\|\theta - \theta^*\|_2 \leq \frac{5\sqrt{3}}{\rho} \lambda \sqrt{s}. \quad (109)
\]
Combining (108) and (109), we obtain
\[
\|\theta - \theta^*\|_2 \leq \max\left\{2\sqrt{\frac{C}{\rho} \lambda \sqrt{s}}, \frac{5\sqrt{3}}{\rho} \lambda \sqrt{s}\right\} \lambda \sqrt{s} \lesssim \lambda \sqrt{s}.
\]
Since \(\theta - \theta^*\) is at most \((s + 1 + \tilde{s})\)-sparse, \(\|\theta - \theta^*\|_1 \leq (s + 1 + \tilde{s})^{1/2}\|\theta - \theta^*\|_2\). The stated results then follow immediately.

**Lemma 13.** Assume Condition 4 holds and \(4\{\|\nabla L(\theta^*)\|_\infty + \epsilon_0 \vee \epsilon_1\} \leq \lambda \lesssim r / \sqrt{s}\). For any \(\ell \geq 2\), the solution sequence \(\{\theta^{(\ell,k)}\}_{k \geq 0}\) satisfies
\[
\|\theta^{(\ell,k)}\|_\infty \leq \tilde{s}, \quad \|\theta^{(\ell,k)} - \theta^*\|_2 \leq C_1 \lambda \sqrt{s} \quad \text{and} \quad \|\theta^{(\ell,k)} - \theta^*\|_1 \leq C_2 \lambda s, \quad (110)
\]
where \(C_1, C_2 > 0\) are constants depending only on the localized sparse eigenvalues.

**Proof of Lemma 15.** We prove the theorem by the method of induction on \((\ell, k)\). Throughout, \(C\) denotes a constant independent of \((s, d, n)\) and may take different values at each appearance. For the 1st subproblem, directly applying Proposition 4.1 and Lemma 5.4 in Fan et al. (2018) we obtain that \(\|\tilde{\theta}^{(1)} - \theta\|_2 \leq C\rho^* \lambda \sqrt{s} < r\), \(\|\tilde{\theta}^{(1)} - \theta\|_1 \leq C\rho^* \lambda s\) and \(\tilde{\theta}^{(1)}\) is \((s + 1 + \tilde{s})\)-sparse, where \(\tilde{s} \leq Cs\). It follows that \(\theta^{(2,0)} = \tilde{\theta}^{(1)}\) falls in a localized sparse set.
To apply the method of induction, first we assume that for any $k$, $\theta^{(2,k)}$ falls in a localized sparse set such that (110) holds. We then use Lemma E.13 in Fan et al. (2018) to show that $\theta^{(2,k+1)}$ also falls in a localized sparse set. To this end, we need to verify two conditions. The first one, $\|\lambda^{e,f}_{\ell}\|_{\min} \geq \lambda/2$ is guaranteed by Claim (62) in the proof of Proposition 3, when taking $a = 1/2$ therein. For the second condition, it suffices to show

$$\Psi(\theta^{(2,k)}, \lambda^{(1)}) - \Psi(\theta^{*}, \lambda^{(1)}) \leq (1 + \zeta)\rho_s^{-1}\lambda^2 s,$$

where $\zeta = \rho_s/\rho_s$. Using the mean value theorem, there exists some convex combination of $\theta^{(2,k)}$ and $\theta^{*}$, say $\tilde{\theta}$, such that

$$\Psi(\theta^{(2,k)}, \lambda^{(1)}) - \Psi(\theta^{*}, \lambda^{(1)}) = \mathcal{L}(\theta^{(2,k)}) - \mathcal{L}(\theta^{*}) + \{\|\lambda^{(1)} \circ \theta^{(2,k)}\|_1 - \|\lambda^{(1)} \circ \theta^{*}\|_1\}$$

$$\leq (\nabla \mathcal{L}(\theta^{*}), \theta^{(2,k)} - \theta^{*}) + \frac{1}{2}(\theta^{(2,k)} - \theta^{*})^\top \nabla^2 \mathcal{L}(\tilde{\theta})(\theta^{(2,k)} - \theta^{*})$$

$$+ \|\lambda^{(1)} \circ (\theta^{(2,k)} - \theta^{*})\|_1$$

$$\leq ||\nabla \mathcal{L}(\theta^{*})||_{\infty} ||\theta^{(2,k)} - \theta^{*}||_1 + \frac{1}{2}\rho_s^{*}||\theta^{(2,k)} - \theta^{*}||_2^2 + \lambda \|\theta^{(2,k)} - \theta^{*}\|_1$$

$$\leq C \frac{1}{4}\rho_s^{-1}\lambda^2 s + C^2\rho_s^{-2}\lambda^2 s + C\rho_s^{-1}\lambda^2 s \leq (1 + \zeta)\rho_s^{-1}\lambda^2 s.$$

With above preparations, it follows from Lemma E.13 in Fan et al. (2018) with slight modification that $\|\theta^{(2,k+1)}\|_0 \leq s + 1 + \tilde{s}$.

Next, we show that $\|\theta^{(2,k+1)} - \theta^{*}\|_2 \leq \rho_s^{-1}\lambda\sqrt{s}$. Again, by Lemma 9,

$$\Psi(\theta^{(2,k+1)}, \lambda^{(1)}) - \Psi(\theta^{(2,k)}, \lambda^{(1)}) \leq -\frac{c(2,k+1)}{2}\|\theta^{(2,k+1)} - \theta^{(2,k)}\|_2.$$

This implies that $\{\Psi(\theta^{(2,k)}, \lambda^{(1)}) - \Psi(\theta^{*}, \lambda^{(1)})\}_{k \geq 1}$ is a non-increasing sequence. By induction, it follows that

$$\Psi(\theta^{(2,k+1)}, \lambda^{(1)}) - \Psi(\theta^{*}, \lambda^{(1)}) \leq \Psi(\theta^{(2,k)}, \lambda^{(1)}) - \Psi(\theta^{*}, \lambda^{(1)}) \leq (1 + \zeta)\rho_s^{-1}\lambda^2 s.$$

Combining this with Lemma 12 gives the desired bounds on $\|\theta^{(2,k+1)} - \theta^{*}\|_2$ and $\|\theta^{(2,k+1)} - \theta^{*}\|_1$.

Finally, by an argument similar to that in the proof of Lemma 5.4 in Fan et al. (2018), we can derive the stated results for all $\ell \geq 3$. \hfill $\Box$

For $\epsilon > 0$, let $\tilde{\theta} = (\tilde{\beta}_0, \tilde{\beta}_i)^\top$ be an $\epsilon$-optimal solution to the program $\min_{\theta} \{\mathcal{L}_\ell(\theta) + \|\lambda \circ \theta\|_1\}$. The following lemma provides conditions under which $\tilde{\theta}$ falls in an $\ell_1$-cone.

**Lemma 14.** Let $E \subseteq [d]$ be a subset satisfying $S \subseteq E$. Moreover, assume $\lambda \geq \|\lambda\|_{\infty} \vee 4\{\|\nabla \mathcal{L}(\theta^*)\|_{\infty} + \epsilon\}$ and $\|\lambda_{E^c}\|_{\min} \geq \lambda/2$. Then we have

$$\|\tilde{\theta}^* - \mathcal{E}\|_1 \leq \frac{\|\lambda\|_{\infty} + \|\nabla \mathcal{L}(\theta^*)\|_{\infty} + \epsilon}{\|\lambda_{E^c}\|_{\min} - \|\nabla \mathcal{L}(\theta^*)\|_{\infty} - \epsilon} \|\mathcal{E} - \theta^*\|_1 \leq 5\|\tilde{\theta} - \theta^*\|_E.$$

**Proof of Lemma 14.** For any $\zeta = (0, \xi)^\top$ with $\xi \in \partial\|\beta\|_1$, let $u = \nabla \mathcal{L}(\tilde{\theta}) + \lambda \circ \zeta$ where $\bar{\lambda} = (0, \lambda^T)^\top$. By the convexity of $\mathcal{L}$, $\langle \nabla \mathcal{L}(\tilde{\theta}) - \nabla \mathcal{L}(\theta^*), \tilde{\theta} - \theta^* \rangle \geq 0$. This, together with the inequality $\langle \nabla \mathcal{L}(\tilde{\theta}) + \lambda \circ \zeta, \tilde{\theta} - \theta^* \rangle \leq \|u\|_{\infty} \|\theta - \theta^*\|_1$, implies

$$0 \leq \|u\|_{\infty} \|\tilde{\theta} - \theta^*\|_1 - \langle \nabla \mathcal{L}(\theta^*), \tilde{\theta} - \theta^* \rangle - (\lambda \circ \xi, \tilde{\beta} - \beta^*) \geq (1 + \zeta)\rho_s^{-1}\lambda^2 s.$$ (111)
For I and II, note that $I \geq -\|\nabla \mathcal{L}(\theta^*)\|_{\infty} \|\hat{\theta} - \theta\|_1$, and

\[
II = \langle \lambda \circ \xi, \hat{\beta} - \beta^* \rangle = \langle \lambda \circ \xi, (\hat{\beta} - \beta^*)_{E^c} \rangle + \langle \lambda \circ \xi, (\hat{\beta} - \beta^*)_{E} \rangle \\
\geq \|\lambda_{E^c}\|_{\min} \| (\hat{\beta} - \beta^*)_{E^c} \|_1 - \|\lambda_{E}\|_{\infty} \| (\hat{\beta} - \beta^*)_{E} \|_1.
\]

Substituting the above bounds into (111) and taking the infimum over $\xi \in \partial\|\hat{\beta}\|_1$ yields

\[
0 \leq -\|\lambda_{E^c}\|_{\min} - \{\|\nabla \mathcal{L}(\theta^*)\|_{\infty} + \omega_{\lambda}(\hat{\beta})\} \|\hat{\theta} - \theta^*\|_{E^c} \|_1 \\
+ \{\|\lambda_{E}\|_{\infty} + \|\nabla \mathcal{L}(\theta^*)\|_{\infty} + \omega_{\lambda}(\hat{\theta})\} \|\hat{\theta} - \theta^*\|_{E} \|_1,
\]

or equivalently

\[
\| (\hat{\theta} - \theta^*)_{E^c} \|_1 \leq \frac{\|\lambda\|_{\infty} + \|\nabla \mathcal{L}(\theta^*)\|_{\infty} + \omega_{\lambda}(\hat{\beta})}{\|\lambda_{E^c}\|_{\min} - \{\|\nabla \mathcal{L}(\theta^*)\|_{\infty} + \omega_{\lambda}(\hat{\theta})\}} \|\hat{\theta} - \theta^*\|_{E} \|_1.
\]

This leads to the stated result. \hfill \Box

B.9.2 Proof of the theorem

Restricting our attention to the $\ell$-th subproblem, we write $\phi^{(k)} = \phi^{(\ell, k)}$ for simplicity. Define the subset $\mathbb{L} = \{\alpha \hat{\theta} + (1 - \alpha) \theta^{(k-1)} : 0 \leq \alpha \leq 1\}$. Due to local majorization, we have

\[
\Psi(\theta^{(k)}, \lambda, L) \\
\leq \min_{\theta \in \mathbb{L}} \left\{ \mathcal{L}(\theta^{(k-1)}) + \langle \nabla \mathcal{L}(\theta^{(k-1)}), \theta - \theta^{(k-1)} \rangle + \frac{\phi^{(k)}}{2} \|\theta - \theta^{(k-1)}\|_2^2 + \|\lambda \circ \beta\|_1 \right\} \\
\leq \min_{\theta \in \mathbb{L}} \left\{ \mathcal{L}(\theta) + \frac{\phi^{(k)}}{2} \|\theta - \theta^{(k-1)}\|_2^2 + \|\lambda \circ \beta\|_1 \right\},
\]

where we used the convexity of $\mathcal{L}(\theta)$ in the second inequality. Since $\Psi(\theta, \lambda) = \mathcal{L}(\theta) + \|\lambda \circ \beta\|_1$ is minimized at $\hat{\theta}$, by convexity we have

\[
\Psi(\theta^{(k)}, \lambda) \leq \min_{\theta \in \mathbb{L}} \left\{ \Psi(\theta, \lambda) + \frac{\phi^{(k)}}{2} \|\theta - \theta^{(k-1)}\|_2^2 \right\} \\
\leq \min_{0 \leq \alpha \leq 1} \left\{ \alpha \Psi(\hat{\theta}, \lambda) + (1 - \alpha) \Psi(\theta^{(k-1)}, \lambda) + \frac{\alpha^2 \phi^{(k)}}{2} \|\theta^{(k-1)} - \hat{\theta}\|_2^2 \right\} \\
= \min_{0 \leq \alpha \leq 1} \left\{ \Psi(\theta^{(k-1)}, \lambda) - \alpha \{\Psi(\theta^{(k-1)}, \lambda) - \Psi(\hat{\theta}, \lambda)\} + \frac{\alpha^2 \phi^{(k)}}{2} \|\theta^{(k-1)} - \hat{\theta}\|_2^2 \right\}.
\]

(112)

Next, we bound the right-hand side of (112). By Lemma 13,

\[
\| (\theta^{(k-1)})_{S^c} \|_0 \leq \tilde{s}, \| \theta^{(k-1)} - \theta^* \|_2 \lesssim \lambda \sqrt{\tilde{s}} \leq r \quad \text{and} \quad \| \theta^{(k-1)} - \theta^* \|_2 \lesssim \lambda s.
\]

Similarly, it can be shown the the optimum $\hat{\theta}$ satisfies the same properties. Hence,

\[
\theta^{(k)}, \hat{\theta} \in C(s + \tilde{s} + 1, r, \tau) \cap \mathbb{B}_2(r, \theta^*).
\]
By the first-order optimality condition, there exists some \( \hat{\xi} \in \partial\|\hat{\theta}\|_1 \) such that \( \nabla \mathcal{L}(\hat{\theta}) + \hat{\lambda} \circ \hat{\xi} = 0 \), where \( \hat{\lambda} = (0, \lambda^*) \), \( \hat{\xi} = (0, \hat{\xi})^\top \in \mathbb{R}^{d+1} \). Moreover, define \( D_{\mathcal{L}}(\theta_1, \theta_2) = \mathcal{L}(\theta_1) - \mathcal{L}(\theta_2) - \langle \nabla \mathcal{L}(\theta_2), \theta_1 - \theta_2 \rangle \). Using Definition 3, Lemma 11, and the convexity of \( \mathcal{L} \) and \( \ell_1 \)-norm, \( \Psi(\theta^{(k-1)}, \lambda) - \Psi(\hat{\theta}, \lambda) \) can be bounded as
\[
\Psi(\theta^{(k-1)}, \lambda) - \Psi(\hat{\theta}, \lambda) \geq (\nabla \mathcal{L}(\hat{\theta}) + \hat{\lambda} \circ \hat{\xi}, \theta^{(k-1)} - \hat{\theta}) + D_{\mathcal{L}}(\theta^{(k-1)}, \hat{\theta}) \geq \frac{\rho_-}{2} \|\theta^{(k-1)} - \hat{\theta}\|^2_2,
\]
where \( \rho_- = \rho_-(2s + 2\bar{s} + 2, r, \tau) \). Plugging this bound into (112) yields
\[
\Psi(\theta^{(k)}, \lambda) \leq \min_{0 \leq \alpha \leq 1} \left[ \Psi(\theta^{(k-1)}, \lambda) - \alpha \{ \Psi(\theta^{(k-1)}, \lambda) - \Psi(\hat{\theta}, \lambda) \} + \frac{\alpha^2 \phi(k)}{\rho_-} \{ \Psi(\theta^{(k-1)}, \lambda) - \Psi(\hat{\theta}, \lambda) \} \right]
\leq \Psi(\theta^{(k-1)}, \lambda) - \frac{\rho_-}{4\phi(k)} \{ \Psi(\theta^{(k-1)}, \lambda) - \Psi(\hat{\theta}, \lambda) \}.
\]
Following the proof of Lemma 8, it can be similarly shown that \( \phi(k) \leq \gamma_u \rho^* \) under Condition 4. Consequently,
\[
\Psi(\theta^{(k)}, \lambda) - \Psi(\hat{\theta}, \lambda) \leq \left( 1 - \frac{1}{4\gamma_u \zeta} \right)^k \{ \Psi(\theta^{(0)}, \lambda) - \Psi(\hat{\theta}, \lambda) \},
\]
where \( \zeta = \rho^*/\rho_* \).

By an argument similar to that in the proof of Lemma 8, we can show that, for \( \ell \geq 2 \),
\[
\omega_{\lambda^{(\ell-1)}}(\theta^{(\ell,k)}) \leq \rho^*(1 + \gamma_u)\|\theta^{(\ell,k)} - \theta^{(\ell,k-1)}\|_2.
\]
Further, using Lemma 9 to bound \( \|\theta^{(\ell,k)} - \theta^{(\ell,k-1)}\|_2 \) from above and noting that \( \phi(k) \geq \rho_* \), we obtain
\[
\omega_{\lambda^{(\ell-1)}}(\theta^{(\ell,k)}) \leq (1 + \gamma_u)\rho^* \sqrt{2/\rho_*} \{ \Psi(\theta^{(\ell,k-1)}, \lambda^{(\ell-1)}) - \Psi(\theta^{(\ell,k)}, \lambda^{(\ell-1)}) \}
\leq (1 + \gamma_u)\sqrt{2\zeta \rho^*} \{ \Psi(\theta^{(\ell,k-1)}, \lambda^{(\ell-1)}) - \Psi(\hat{\theta}^{(\ell)}, \lambda^{(\ell-1)}) \}
\leq (1 + \gamma_u)\sqrt{2\zeta \rho^*} \left( 1 - \frac{1}{4\gamma_u \zeta} \right)^{k-1} \{ \Psi(\theta^{(\ell,0)}, \lambda^{(\ell-1)}) - \Psi(\hat{\theta}^{(\ell)}, \lambda^{(\ell-1)}) \}
\leq C(1 + \gamma_u)\sqrt{\frac{\zeta \rho^*}{\phi^{(\ell,0)}}} \left( 1 - \frac{1}{4\gamma_u \zeta} \right)^{k-1} \lambda^2 s \leq C(1 + \gamma_u)\zeta \sqrt{\left( 1 - \frac{1}{4\gamma_u \zeta} \right)^{k-1}} \lambda^2 s,
\]
where we used Lemmas 9 and 13 in the last step.

To make the right-hand side of the above inequality smaller than \( \epsilon_t \), we need \( k \) to be sufficiently large that \( k \geq C_1 \log(C_2 \lambda \sqrt{s}/\epsilon_t) \), where \( C_1, C_2 > 0 \) are constants depending only on localized sparse eigenvalues and \( \gamma_u \). This completes the proof. \( \square \)