Iteratively reweighted ℓ_1-penalized robust regression[*]

Xiaoou Pan

*Department of Mathematics, University of California, San Diego,
La Jolla, CA 92093, USA
e-mail:* xip024@ucsd.edu

Qiang Sun

*Department of Statistical Sciences, University of Toronto,
Toronto, ON M5S 3G3, Canada
e-mail:* qsun@utstat.toronto.edu

Wen-Xin Zhou

*Department of Mathematics, University of California, San Diego,
La Jolla, CA 92093, USA
e-mail:* wez243@ucsd.edu

Abstract: This paper investigates tradeoffs among optimization errors, statistical rates of convergence and the effect of heavy-tailed errors for high-dimensional robust regression with nonconvex regularization. When the additive errors in linear models have only bounded second moments, we show that iteratively reweighted ℓ_1-penalized adaptive Huber regression estimator satisfies exponential deviation bounds and oracle properties, including the oracle convergence rate and variable selection consistency, under a weak beta-min condition. Computationally, we need as many as $\mathcal{O}(\log s + \log \log d)$ iterations to reach such an oracle estimator, where s and d denote the sparsity and ambient dimension, respectively. Extension to a general class of robust loss functions is also considered. Numerical studies lend strong support to our methodology and theory.

Contents

1. Introduction

Suppose we observe independent and identically distributed (i.i.d.) data vectors $\{(y_i, \boldsymbol{x}_i) : 1 \leq i \leq n\}$ from $(y, \boldsymbol{x})$ that follows the linear model

$$y = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}^* + \varepsilon = \sum_{j=1}^{d} \beta_j^* x_j + \varepsilon, \tag{1.1}$$

where $\boldsymbol{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$ with $x_1 \equiv 1$ is the predictor, $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_d^*)^{\mathrm{T}}$ is the vector of regression coefficients with β_1^* denoting the intercept, and ε is an error term satisfying $\mathbb{E}(\varepsilon|\boldsymbol{x}) = 0$. This setting includes the location-scale model in which $\varepsilon = \sigma(\boldsymbol{x})e$, $\sigma(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$ is an unknown function, and e is independent of $\boldsymbol{x}$ and satisfies $\mathbb{E}(e) = 0$. We restrict attention to the high-dimensional regime in which the number of features d exceeds the sample size

n, and $\boldsymbol{\beta}^* \in \mathbb{R}^d$ is s-sparse with $s \ll n$. Of particular interest is the case where the error variable is *asymmetric* and *heavy-tailed* with only bounded second moment.

Since the invention of Lasso in the seminal work of [46], 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 $\widehat{\boldsymbol{\beta}}^{\text{Lasso}}$ solves the convex optimization problem $\min_{\boldsymbol{\beta} \in \mathbb{R}^d} (2n)^{-1} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^{\text{T}} \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_1$, where $\lambda > 0$ is the regularization parameter. The Lasso is an ℓ_1-penalized least squares method in nature: the quadratic loss is used as a goodness of fit measure and the ℓ_1-norm induces sparsity. To achieve better performance under different circumstances, several Lasso variants have been proposed and studied; see [3, 5, 14, 43, 51, 56, 57] to name a few. We refer to [8, 20] and [49] for comprehensive and systematic introductions 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 $\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \{(1/n) \sum_{i=1}^{n} \ell(y_i - \boldsymbol{x}_i^{\text{T}} \boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1\}$, 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 $\{\ell'(\varepsilon_i)\}_{i=1}^{n}$. When ℓ is the quadratic loss and ε is heavy-tailed, this vector is likely to have relatively many large coordinates. As a result, the combination of the rapid growth of ℓ with heavy-tailed sampling distribution inevitably leads to outliers, which will eventually be translated into spurious discoveries. Secondly, the ℓ_1-penalty introduces nonnegligible estimation bias [14, 56]. For correlated designs, the bias of the Lasso may offset true signals and creates spurious effects, leading to inconsistency in support recovery. Technically, this is expressed as the irrepresentable condition for the selection consistency of the Lasso [55]. Under restricted eigenvalue type conditions, the Lasso and its sorted variant Slope [1, 2] do achieve rate optimality for prediction and coefficient estimation. However, they do not benefit much from strong signals because the bias does not diminish as signal strengthens. Under the restricted isometry property (on the design) and Gaussian errors, [39] derived the lower bound for the minimax risk: $\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}^* \in \Omega(s,a)} \mathbb{E} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 \gtrsim \sigma^2 s/n$ when $a \gtrsim \sigma \sqrt{\log(ed/s)/n}$, where $\Omega(s, a) = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|\boldsymbol{\beta}\|_0 \leq s, \min_{j:\beta_j \neq 0} |\beta_j| \geq a\}$. For estimating such sparse vectors with sufficiently strong signals, Lasso can not achieve the oracle rate without a variant of the irrepresentable condition [37, 55], which is a condition on how strongly the important and unimportant variables can be correlated. This condition, however, is in general very restrictive; see [56] for counterexamples and numerical demonstrations. Recently, [23] provided precise descriptions of the lower and upper irrepresentable conditions, proven to be necessary and sufficient for the selection consistency of Lasso, and also pointed that the Lasso estimator cannot achieve selection consistency and $\sqrt{n}$-consistency simultaneously under general conditions.

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. When the regression error ε only has finite second moment, the Lasso still achieves the minimax rate $\sqrt{s\log(d)/n}$ (under ℓ_2-norm) but with worse deviations [28]. To reduce the ill-effects of outliers, a widely recognized strategy is to use a robust loss function that is globally Lipschitz continuous and locally quadratic. A prototypical example is the Huber loss [21]:

$$\ell_\tau(x) = \left\{ \begin{array}{ll} x^2/2 & \text{if } |x| \leq \tau, \\ \tau|x| - \tau^2/2 & \text{if } |x| > \tau, \end{array} \right. \tag{1.2}$$

where $\tau > 0$ is a robustification parameter that controls the tradeoff between the robustness and bias. Next, to alleviate the nonnegligible estimation bias introduced by ℓ_1-regularization, [14] introduced a family of folded-concave penalties, including the smoothly clipped absolute deviation (SCAD) penalty [14], minimax concave (MC+) penalty [52], and the capped ℓ_1-penalty [41, 54]. These ideas motivate the following nonconvex (folded concave) regularized M-estimator

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^d} \left\{ \widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) + \sum_{j=1}^d p_\lambda(\beta_j) \right\}, \tag{1.3}$$

where $\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) := (1/n)\sum_{i=1}^n \ell_\tau(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ is the empirical loss, $\tau > 0$ 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 [53] for a comprehensive survey of folded concave regularized methods.

In practice, it is inherently difficult to solve the nonconvex optimization problem (1.3) directly. 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 some local optimum that exists somewhere like a needle in a haystack. To mitigate the gap between statistical theory and algorithmic complexity, we apply an iteratively reweighted ℓ_1-penalized algorithm, which originates from [58], to adaptive Huber regression. This multi-step regularized robust regression procedure (provably) yields an estimator with desired oracle properties, and is computationally efficient because it only involves solving a sequence of (unconstrained) convex programs. Our theoretical analysis is based on and improves upon [16], who established the statistical and algorithmic theory for the iteratively reweighted ℓ_1-penalized least squares regression estimator. The aim of this paper is to explore a general class of robust loss functions, typified by the Huber loss, not merely for the purpose of generality but owing to a real downside of the quadratic 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

Nonasymptotic or finite-sample analysis of regularized regression methods beyond least squares, such as regularized empirical risk minimization (ERM) or M-estimation with a non-quadratic loss, can be divided into three categories depending on the form of the regularizer/penalty.

ℓ_1-REGULARIZATION: For high-dimensional sparse linear models, [38] and [15], respectively, proposed a robust version of Lasso based on geometric median and ℓ_1-penalized Huber's M-estimator. Both estimators achieve sub-Gaussian deviation bounds when the regression error only has finite variance. In such a heavy-tailed case, [28] showed that the Lasso still achieves the minimax rate under expectation but with much worse deviations. When the regression error only has finite $(1 + \delta)$-th absolute moment for some $\delta \in (0, 1)$, [44] established exponential deviation bounds for ℓ_1-penalized adaptive Huber regression estimator with a more delicate choice of the robustification parameter. For more general penalized M-estimators with a convex and Lipschitz continuous loss, [1] established both estimation bounds and sharp oracle inequalities. Their results do not require a local strong convexity on the loss function, thus also including the hinge loss and quantile regression loss. For nonconvex loss functions with a redescending derivative, typified by Tukey's bisquare loss, [36] proved the statistical consistency of the ℓ_1-penalized estimator subject to an ℓ_2-constraint to stationary points.

FOLDED CONCAVE REGULARIZATION: For folded concave penalized M-estimators subject to a convex side constraint, [31] and [32] were among the first to provide rigorous statistical and algorithmic theory for local optima. They quantified statistical accuracy by providing bounds on ℓ_1/ℓ_2- and prediction errors between stationary points and the population-level optimum. They also provided conditions under which the stationary point is unique, and proposed a composite gradient algorithm for provably solving the constrained optimization problem efficiently. In the context of generalized linear models with a sufficiently smooth link function and bounded covariates, [32] proved under the scaling $n \gtrsim s^3 \log(d)$ that the nonconvex regularized program subject to an ℓ_1-ball constraint has a unique stationary point given by the oracle estimator with high probability. For linear regression with symmetric heavy-tailed errors, [30] studied statistical consistency and asymptotic normality of nonconvex regularized robust M-estimators (also subject to an ℓ_1-ball constraint) with a locally strongly convex loss. For sub-Gaussian covariates, [30] proved the uniqueness of a stationary point which has ℓ_2- and ℓ_1-error bounds in the order of $\sqrt{s \log(d)/n}$ and $s\sqrt{\log(d)/n}$, respectively. Furthermore, under the scaling $n \gtrsim \max\{s^2, s \log(d)\}$ and the beta-min condition $\|\beta_{\mathcal{S}}^*\|_{\min} \gtrsim \lambda + \sqrt{\log(s)/n}$, this stationary point coincides with the oracle estimator.

In this paper, we address nonconvex regularized robust regression from a different angle. Motivated by the local linear approximation (LLA) algorithm proposed by [58], we apply an iteratively reweighted ℓ_1-penalized algorithm to adaptive Huber regression, which involves solving a sequence of (unconstrained)

convex programs. We simultaneously analyze the statistical property and algorithmic complexity of the solutions produced by such an iterative procedure. For sub-exponential covariates and asymmetric error with finite variance, we show that the multi-step penalized estimator, after $\mathcal{O}(\log s + \log(\log d))$ iterations, achieves exponential deviation bounds with ℓ_2- and ℓ_1-errors in the order of $\sqrt{s/n}$ and $s/\sqrt{n}$, respectively, under the scaling $n \gtrsim s \log(d)$ and the above beta-min condition. The strong oracle property can be obtained under slightly stronger moment condition and the scaling $n \gtrsim \max\{s^2, s \log(d)\}$.

ℓ_0-REGULARIZATION: Another popular class of sparse recovery algorithms is based on directly solving ℓ_0-constrained or ℓ_0-penalized empirical risk minimizations, which naturally produces sparse solutions. Such a formation is NP-hard, and believed to be intractable in practice. Despite its computational hardness, many practically useful algorithms have been proposed to solve ℓ_0-regularized ERM, while the statistical properties beyond least squares regression are much less studied. We refer to [4] and [19] for two comprehensive survey articles on ℓ_0-regularized regression methods.

The idea of having the robustification parameter grow with the sample size in order to achieve exponential deviations even when the sampling distribution only has finite variance dates back to [9] in the context of mean estimation. Therefore, the robustness considered in this paper is primarily about nonasymptotic exponential deviation of the estimator versus polynomial tail of the error distribution. The resulting procedure does sacrifice a fair amount of robustness to adversarial contamination of the data. To echo the message in [9], the motivation of this work is different from and should not be confused with the classical notion of robust statistics.

From a variable selection perspective, this paper focuses on oracle properties of multi-step penalized robust regression estimators when the signal is sufficiently strong. While allowing for heavy-tailed noise, the high-dimensional feature vector $\boldsymbol{x} \in \mathbb{R}^d$ is assumed to have either sub-exponential or sub-Gaussian tails. For more complex problems in which both the covariates and noise can be (i) heavy-tailed and/or (ii) adversarially contaminated, the estimator obtained by minimizing a robust loss function is still sensitive to outliers in the feature space. To achieve robustness in both feature and response spaces, recent years have witnessed a rapid development of the "median-of-means" (MOM) principle, which dates back to [40] and [22], and a variety of MOM-based procedures for regression and classification in both low- an high-dimensional settings [12, 13, 26, 27, 33, 35]. We refer to [34] for a recent survey. An interesting open problem is how to efficiently incorporate the MOM principle with nonconvex regularization or iteratively reweighted ℓ_1-regularization so as to achieve high degree of robustness and variable selection consistency simultaneously.

1.2. Notation

Let us summarize our notation. For every integer $k \geq 1$, we use $\mathbb{R}^k$ to denote the k-dimensional Euclidean space. The inner and Hadamard products of any two

vectors $\boldsymbol{u} = (u_1, \ldots, u_k)^{\mathrm{T}}, \boldsymbol{v} = (v_1, \ldots, v_k)^{\mathrm{T}} \in \mathbb{R}^k$ are defined by $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v} = \langle \boldsymbol{u}, \boldsymbol{v} \rangle = \sum_{i=1}^{k} u_i v_i$ and $\boldsymbol{u} \circ \boldsymbol{v} = (u_1 v_1, \ldots, u_k v_k)^{\mathrm{T}}$, respectively. We use $\|\cdot\|_p$ $(1 \le p \le \infty)$ to denote the ℓ_p-norm in $\mathbb{R}^k$: $\|\boldsymbol{u}\|_p = (\sum_{i=1}^{k} |u_i|^p)^{1/p}$ and $\|\boldsymbol{u}\|_\infty = \max_{1 \le i \le k} |u_i|$. Moreover, we write $\|\boldsymbol{u}\|_{\min} = \min_{1 \le i \le k} |u_i|$. For $k \ge 2$, $\mathbb{S}^{k-1} = \{\boldsymbol{u} \in \mathbb{R}^k : \|\boldsymbol{u}\|_2 = 1\}$ denotes the unit sphere in $\mathbb{R}^k$. For any function $f : \mathbb{R} \mapsto \mathbb{R}$ and vector $\boldsymbol{u} = (u_1, \ldots, u_k)^{\mathrm{T}} \in \mathbb{R}^k$, we write $f(\boldsymbol{u}) = (f(u_1), \ldots, f(u_k))^{\mathrm{T}} \in \mathbb{R}^k$.

For $k \ge 2$, I_k represents the identity/unit matrix of size k. For any $k \times k$ symmetric matrix $\Sigma \in \mathbb{R}^{k \times k}$, $\|\Sigma\|_2$ is the operator norm of Σ, and we use $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ to denote the minimal and maximal eigenvalues of Σ, respectively. For a positive semidefinite matrix $\Sigma \in \mathbb{R}^{k \times k}$, $\|\cdot\|_\Sigma$ denotes the norm linked to Σ given by $\|\boldsymbol{u}\|_\Sigma = \|\Sigma^{1/2}\boldsymbol{u}\|_2$, $\boldsymbol{u} \in \mathbb{R}^k$. For any two real numbers u and v, we write $u \vee v = \max(u, v)$ and $u \wedge v = \min(u, v)$. For any integer $d \ge 1$, we write $[d] = \{1, \ldots, d\}$. For any set $\mathcal{S}$, we use $|\mathcal{S}|$ to denote its cardinality, i.e., the number of elements in $\mathcal{S}$.

2. Regularized Huber M-estimation

We first revisit the ℓ_1-penalized Huber regression estimator in Section 2.1, and point out two different regimes for the robustification parameter τ. In Section 2.2, we propose a multi-step procedure, which is closely related to folded concave regularized Huber regression, for fitting high-dimensional sparse models with heavy-tailed noise. This multi-step penalized 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, $\mathcal{S} = \mathrm{supp}(\boldsymbol{\beta}^*) = \{1 \le j \le d : \beta_j^* \ne 0\} \subseteq [d]$ denotes the active set and $s = |\mathcal{S}|$ is the sparsity.

2.1. ℓ_1-penalized Huber regression

Given i.i.d. observations $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^{n}$ from the linear model (1.1), consider the ℓ_1-regularized Huber M-estimator, which we refer to as the *Huber-Lasso*,

$$\widehat{\boldsymbol{\beta}}^{\text{H-Lasso}} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\operatorname{argmin}} \{\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1\}, \tag{2.1}$$

where $\widehat{\mathcal{L}}_\tau(\cdot)$ is the empirical loss function defined in (1.3). Statistical properties of the penalized Huber M-estimator have been studied by [1, 15, 24, 30] under different assumptions. A less-noticed problem is the connection between the robustification parameter and the error distribution, which in turn quantifies the tradeoff between robustness and unbiasedness. Recent studies by [44] reveal that the use of Huber loss is particularly suited for heavy-tailed problems in both low and high dimensions. With a properly chosen robustification parameter, calibrated by the noise level, sample size and parametric dimension, the effects of the heavy-tailed noise can be removed or dampened.

Remark 2.1. In practice, it is natural to leave the intercept or a given subset of the parameters unpenalized in the penalized M-estimation framework. Denote

by $\mathcal{R}$ be a user-specified index set of unpenalized parameters, which contains at least index 1. A modified Huber-Lasso estimator is then defined as the solution to $\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \{\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}_{\mathcal{R}^c}\|_1\}$, where $\|\boldsymbol{\beta}_{\mathcal{R}^c}\|_1 = \sum_{j \in \mathcal{R}^c} |\beta_j|$. Similar theoretical analysis can be carried out with slight modifications, and thus will be omitted for ease of exposition.

We first impose the following assumptions on the data generating process. The (random) covariate vectors are assumed to be *sub-exponential/sub-gamma* [6], and we allow the regression errors to be heavy-tailed and asymmetric.

Condition 2.1. There exist some constant $\sigma_{\boldsymbol{x}}, c_0 \geq 1$ such that $\mathbb{P}(|\boldsymbol{u}^{\mathrm{T}} \boldsymbol{x}| \geq \sigma_{\boldsymbol{x}} t) \leq c_0 e^{-t}$ for all $\boldsymbol{u} \in \mathbb{S}^{d-1}$ and $t \geq 0$. For simplicity, we set $c_0 = 1$. Moreover, $\Sigma = \mathbb{E}(\boldsymbol{x} \boldsymbol{x}^{\mathrm{T}})$ is positive definite with $\rho_l = \lambda_{\min}(\Sigma) > 0$. The regression error ε satisfies $\mathbb{E}(\varepsilon|\boldsymbol{x}) = 0$ and $\mathbb{E}(\varepsilon^2|\boldsymbol{x}) \leq \sigma_2^2$ almost surely.

Theorem 2.1. Assume that Condition 2.1 holds for model (1.1). For every $t > 0$, any optimal solution $\widehat{\boldsymbol{\beta}}^{\text{H-Lasso}}$ to the convex program (2.1) with $\tau \asymp \sigma_2 \sqrt{n/(\log d + t)}$ and $\lambda \asymp \sigma_2 \sqrt{(\log d + t)/n}$ satisfies

$$\|\widehat{\boldsymbol{\beta}}^{\text{H-Lasso}} - \boldsymbol{\beta}^*\|_2 \leq c_1 \sigma_2 (\log d + t)^{1/2} \sqrt{\frac{s}{n}} \quad \text{and}$$
$$\|\widehat{\boldsymbol{\beta}}^{\text{H-Lasso}} - \boldsymbol{\beta}^*\|_1 \leq c_2 \sigma_2 (\log d + t)^{1/2} \frac{s}{\sqrt{n}} \tag{2.2}$$

with probability at least $1 - 3e^{-t}$ as long as $n \geq c_3(s \log d + t)$, where c_1–c_3 are constants that are independent of (n, d, s) and t.

Theorem 2.1 provides the error bounds for the one-step penalized estimator, and paves the way for our subsequent analysis for the multi-step procedure. Theorem 2.1 is a modified version of Theorem B.2 in [44] (when $\delta = 1$) with an explicit relation between deviation bound and confidence level under slightly relaxed moment condition on the design. When the (conditional) distribution of ε is symmetric, $\boldsymbol{\beta}^*$ can be identified as $\boldsymbol{\beta}^* \in \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^d} \mathbb{E}\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta})$. Then, with a fixed τ (e.g. $\tau \asymp \sigma_2$), Theorem 2.1 can also be obtained as a special case of Theorem 2.1 in [1] when the feature vector $\boldsymbol{x}$ is sub-Gaussian.

2.2. *Iteratively reweighted ℓ_1-penalized Huber regression*

For fitting sparse regression models, the Lasso-type estimators typically exhibit a suboptimal rate of convergence, as compared to the oracle rate achieved by nonconvex regularization methods, under a minimum signal strength condition [39, 53], also known as the beta-min condition [8, Section 7.4]. However, as noted previously, directly solving the nonconvex optimization problem (1.3) 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 folded concave penalties [58], we consider a multi-stage procedure that solves a sequence of convex programs

up to a prespecified optimization precision. This is an iteratively reweighted ℓ_1-penalized algorithm, which is similar in spirit to the iteratively reweighted basis-pursuit algorithms studied in [18]. Let $p_\lambda(\cdot)$ be a differentiable penalty function as in (1.3) and recall that $\widehat{\mathcal{L}}_\tau(\cdot)$ is the empirical loss function. Starting with an initial estimate $\widehat{\boldsymbol{\beta}}^{(0)} = (\widehat{\beta}_1^{(0)}, \ldots, \widehat{\beta}_d^{(0)})^{\mathrm{T}}$, consider a sequence of convex optimization programs $\{(\mathrm{P}_\ell)\}_{\ell \geq 1}$:

$$\min_{\boldsymbol{\beta}=(\beta_1,\ldots,\beta_d)^{\mathrm{T}}} \left\{ \widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) + \sum_{j=1}^d p_\lambda'(|\widehat{\beta}_j^{(\ell-1)}|)|\beta_j| \right\} \qquad (\mathrm{P}_\ell) \qquad (2.3)$$

for $\ell = 1, 2, \ldots$, where $\widehat{\boldsymbol{\beta}}^{(\ell)} = (\widehat{\beta}_1^{(\ell)}, \ldots, \widehat{\beta}_d^{(\ell)})^{\mathrm{T}}$ is the optimal solution to program (P_ℓ). Following [53], we assume the following conditions on the penalty function p_λ.

Condition 2.2. The penalty function p_λ 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 $t_1 \geq t_2 > 0$.

Prototypical examples of the penalty function $p(\cdot)$ in Condition 2.2 include the ℓ_1-function, the SCAD penalty [14], MC+ penalty [52], and capped-ℓ_1 function [54].

1. (SCAD) $p(t) = \int_0^{|t|} \min\{1, (1 - \frac{u-1}{a-1})_+\}\mathrm{d}u$ and $p'(t) = \mathrm{sign}(t)\min\{1, (1 - \frac{|t|-1}{a-1})_+\}$ for some $a > 2$. By a Bayesian argument, [14] suggested the choice of $a = 3.7$.
2. (MC+) $p(t) = \int_0^{|t|}(1 - u/a)_+\mathrm{d}u$ and $p'(t) = \mathrm{sign}(t)(1 - |t|/a)_+$ for some $a > 1$.
3. (Capped-ℓ_1) $p(t) = \min(a, |t|)$ and $p'(t) = I(|t| \leq a)$ for some $a > 0$.

For each $\ell \geq 1$, program (P_ℓ) corresponds to a weighted ℓ_1-penalized empirical Huber loss minimization of the form

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \{\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) + \|\boldsymbol{\lambda} \circ \boldsymbol{\beta}\|_1\}, \qquad (2.4)$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)^{\mathrm{T}}$ is a d-vector of regularization parameters with $\lambda_j \geq 0$. By convex optimization theory, any optimal solution $\widehat{\boldsymbol{\beta}}$ to the convex program (2.4) satisfies the first-order condition

$$\nabla \widehat{\mathcal{L}}_\tau(\widehat{\boldsymbol{\beta}}) + \boldsymbol{\lambda} \circ \boldsymbol{\xi} = \mathbf{0}_d \text{ for some } \boldsymbol{\xi} = (\xi_1, \ldots, \xi_d)^{\mathrm{T}} \in \partial\|\widehat{\boldsymbol{\beta}}\|_1 \subseteq [-1, 1]^d,$$

where $\nabla \widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) = (-1/n)\sum_{i=1}^n \ell_\tau'(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{x}_i$.

Definition 2.1. Following the terminology in [16], for a prespecified tolerance level $\epsilon > 0$, we say $\widetilde{\boldsymbol{\beta}}$ is an ϵ-optimal solution to (2.4) if $\omega_{\boldsymbol{\lambda}}(\widetilde{\boldsymbol{\beta}}) \leq \epsilon$, where

$$\omega_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) := \min_{\boldsymbol{\xi} \in \partial\|\boldsymbol{\beta}\|_1} \|\nabla\mathcal{L}_\tau(\boldsymbol{\beta}) + \boldsymbol{\lambda} \circ \boldsymbol{\xi}\|_\infty, \quad \boldsymbol{\beta} \in \mathbb{R}^d. \qquad (2.5)$$

In view of Definition 2.1, for a prespecified sequence of tolerance levels $\{\epsilon_\ell\}_{\ell \geq 1}$, we use $\widetilde{\boldsymbol{\beta}}^{(\ell)} = (\widetilde{\beta}_1^{(\ell)}, \ldots, \widetilde{\beta}_d^{(\ell)})^\mathsf{T}$ to denote an ϵ_ℓ-optimal solution to program (P_ℓ), that is,

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \{\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) + \|\boldsymbol{\lambda}^{(\ell-1)} \circ \boldsymbol{\beta}\|_1\},$$

where $\boldsymbol{\lambda}^{(\ell-1)} := p_\lambda'(|\widetilde{\boldsymbol{\beta}}^{(\ell-1)}|)$. For simplicity, we consider a trivial initial estimator $\widetilde{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$. Since $p_\lambda'(|\widetilde{\beta}_j^{(0)}|) = p_\lambda'(0) = \lambda$ for $j = 1, \ldots, d$, the program (P_1) coincides with that in (2.1). In Section 3, we will describe an iterative local adaptive majorize-minimization (I-LAMM) algorithm which produces an ϵ-optimal solution to (2.4) 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 is 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 ℓ_1-penalized Huber regression in (2.1). A tightening stage further refines this coarse contraction estimator consecutively, and eventually gives rise to an estimator that achieves the oracle rate $\sqrt{s/n}$ under a weak beta-min condition.

2.3. Deterministic analysis

To analyze the statistical properties of $\{\widetilde{\boldsymbol{\beta}}^{(\ell)}\}_{\ell \geq 1}$, we first define a "good" event regarding the restricted strong convexity (RSC) property of the empirical Huber loss over a local ℓ_1-cone.

Definition 2.2. For some $r, l, \kappa > 0$, define the event

$$\mathcal{E}_1(r, l, \kappa) = \left\{ \inf_{\boldsymbol{\beta} \in \boldsymbol{\beta}^* + \mathbb{B}(r) \cap \mathbb{C}(l)} \frac{\langle \nabla \widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) - \nabla \widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle}{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2} \geq \kappa \right\}, \qquad (2.6)$$

where $\mathbb{B}(r) = \mathbb{B}^d(r) = \{\boldsymbol{\delta} \in \mathbb{R}^d : \|\boldsymbol{\delta}\|_2 \leq r\}$ is an ℓ_2-ball and $\mathbb{C}(l) := \{\boldsymbol{\delta} \in \mathbb{R}^d : \|\boldsymbol{\delta}\|_1 \leq l\|\boldsymbol{\delta}\|_2\}$ is an ℓ_1-cone. Here $\boldsymbol{\beta} + \mathbb{B}(r) \cap \mathbb{C}(l) = \{\boldsymbol{\beta} + \boldsymbol{\delta} : \boldsymbol{\delta} \in \mathbb{B}(r) \cap \mathbb{C}(l)\}$.

Throughout the following, we assume that the penalty function $p_\lambda(\cdot)$ satisfies Condition 2.2. Moreover, define

$$\boldsymbol{w}^* = \nabla \widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}^*) - \nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*) \quad \text{and} \quad b_\tau^* = \|\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_2, \qquad (2.7)$$

where $\mathcal{L}_\tau(\boldsymbol{\beta}) = \mathbb{E}\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta})$ is the population loss. Here $\boldsymbol{w}^* \in \mathbb{R}^d$ is the centered gradient vector which corresponds to the stochastic error, and b_τ^* quantifies the (deterministic) approximation bias induced by the Huber loss. See Lemma A.1 in the Supplementary Material for an upper bound on the bias.

Remark 2.2. In this paper, we introduce the bias term b_τ^* into the results primarily because the error distribution, if not specified, can be asymmetric. This term is typically nullified in the literature due to two reasons. First, under

the symmetry assumption that ε (conditional on $\boldsymbol{x}$) is symmetric around zero, $\mathbb{E}\{\ell'_\tau(\varepsilon)\boldsymbol{x}\} = 0$ for any $\tau > 0$. Secondly, it is sometimes assumed that $\boldsymbol{x}_i$ and ε_i are independent, and both have zero means. Again, for any $\tau > 0$, it follows that $\mathbb{E}\{\ell'_\tau(\varepsilon)\boldsymbol{x}\} = \mathbb{E}\{\ell'_\tau(\varepsilon)\} \cdot \mathbb{E}(\boldsymbol{x}) = 0$. In these two scenarios, the bias term b^*_τ vanishes for any given τ.

Proposition 2.1. Let $\lambda, r, \kappa > 0$ satisfy

$$\lambda \geq s^{-1/2}b^*_\tau, \quad r > 2.5\kappa^{-1}s^{1/2}\lambda. \tag{2.8}$$

Then, conditioned on the event $\mathcal{E}_1(r, l, \kappa) \cap \{\lambda \geq 2(\|\boldsymbol{w}^*\|_\infty + \epsilon_1)\}$ with $l = 6s^{1/2}$, any ϵ_1-optimal solution $\widetilde{\boldsymbol{\beta}}^{(1)}$ of program (P$_1$) satisfies

$$\|\widetilde{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*\|_2 \leq 2.5\kappa^{-1}s^{1/2}\lambda. \tag{2.9}$$

Proposition 2.1 is deterministic in the sense that the error bound (2.9) holds conditioning on the event $\mathcal{E}_1(r, l, \kappa) \cap \{\lambda \geq 2(\|\boldsymbol{w}^*\|_\infty + \epsilon_1)\}$. Under Condition 2.1 (sub-exponential design and heavy-tailed error with finite variance), we will establish the delicate choices of λ, ϵ_1, r and sample size requirement in order that this event occurs with high probability. Specifically, we will show that

$$\|\widetilde{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*\|_2 \lesssim \sigma_2\sqrt{\frac{s\log(d)}{n}} \text{ with high probability as long as } n \gtrsim s\log(d).$$

Next, we investigate the statistical properties of $\{\widetilde{\boldsymbol{\beta}}^{(\ell)}\}_{\ell \geq 2}$ in the tightening stage. We impose a *minimum signal strength condition* on $\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_{\min} = \min_{j\in\mathcal{S}}|\beta^*_j|$, so that the error rate obtained in Proposition 2.1 is improvable [39, 53]. Recall that $s = |\mathcal{S}|$.

Proposition 2.2. Assume there exists some $\gamma > 0$ such that $p'(\gamma) > 0$. Let

$$\lambda \geq s^{-1/2}b^*_\tau, \quad \kappa > \sqrt{5}/(2\gamma), \tag{2.10}$$

and choose $c > 0$ so that

$$0.5p'(\gamma)(c^2 + 1)^{1/2} + 2 = c\kappa\gamma. \tag{2.11}$$

Set $l = \{2 + \frac{2}{p'(\gamma)}\}(c^2 + 1)^{1/2}s^{1/2} + \frac{2}{p'(\gamma)}s^{1/2}$ and let $r > 0$ satisfy

$$r^{\text{crude}} := c\gamma s^{1/2}\lambda \leq r. \tag{2.12}$$

Under the minimum signal strength condition $\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_{\min} \geq \gamma\lambda$, and conditioned on event $\mathcal{E}_1(r, l, \kappa) \cap \{\lambda \geq \frac{2}{p'(\gamma)}(\|\boldsymbol{w}^*\|_\infty + \max_{\ell\geq 1}\epsilon_\ell)\}$, the ϵ_ℓ-optimal solutions $\widetilde{\boldsymbol{\beta}}^{(\ell)}$ ($\ell \geq 2$) satisfy

$$\|\widetilde{\boldsymbol{\beta}}^{(\ell)} - \boldsymbol{\beta}^*\|_2$$
$$\leq \delta \cdot \|\widetilde{\boldsymbol{\beta}}^{(\ell-1)} - \boldsymbol{\beta}^*\|_2 + \underbrace{\kappa^{-1}\{\|p'_\lambda(|\boldsymbol{\beta}^*_{\mathcal{S}}| - \gamma\lambda)\|_2 + \|\boldsymbol{w}^*_{\mathcal{S}}\|_2 + s^{1/2}\epsilon_\ell + b^*_\tau\}}_{=:r^{\text{ora}}}.$$
$$\tag{2.13}$$

where $\delta = \sqrt{5}/(2\gamma\kappa) \in (0,1)$. Furthermore, it holds

$$\|\widetilde{\boldsymbol{\beta}}^{(\ell)} - \boldsymbol{\beta}^*\|_2 \leq \delta^{\ell-1} r^{\mathrm{crude}} + (1-\delta)^{-1} r^{\mathrm{ora}} \quad \text{for any } \ell \geq 2. \qquad (2.14)$$

Proposition 2.2 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 (2.13) or (2.14) dominates the ℓ_2-error, and up to constant factors, consists of three components,

$$\underbrace{\|p_\lambda'(|\boldsymbol{\beta}_{\mathcal{S}}^*| - \gamma\lambda)\|_2}_{\text{shrinkage bias}}, \qquad \underbrace{\|\boldsymbol{w}_{\mathcal{S}}^*\|_2 + b_\tau^*}_{\text{oracle rate plus approx. bias}} \qquad \text{and} \qquad \underbrace{s^{1/2}\epsilon_\ell}_{\text{optimization error}} \qquad .$$

We identify $\|p_\lambda'(|\boldsymbol{\beta}_{\mathcal{S}}^*| - \gamma\lambda)\|_2$ as the shrinkage bias induced by the penalty function. This explains the limitation of the ℓ_1-penalty $p_\lambda(t) = \lambda|t|$ whose derivative $p_\lambda'(t) = \lambda\,\mathrm{sign}(t)$ ($t \neq 0$) does not vanish regardless of the signal strength. Intuitively, choosing a proper penalty function $p_\lambda(\cdot)$ with a descending derivative reduces the bias as signal strengthens. The second term, $\|\boldsymbol{w}_{\mathcal{S}}^*\|_2 + b_\tau^*$, reveals the oracle property. To see this, consider the oracle estimator defined as

$$\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} = \operatorname*{argmin}_{\boldsymbol{\beta}:\boldsymbol{\beta}_{\mathcal{S}^c}=\mathbf{0}} \widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) = \operatorname*{argmin}_{\boldsymbol{\beta}:\boldsymbol{\beta}_{\mathcal{S}^c}=\mathbf{0}} \frac{1}{n} \sum_{i=1}^n \ell_\tau(y_i - \boldsymbol{x}_{i,\mathcal{S}}^{\mathrm{T}} \boldsymbol{\beta}_{\mathcal{S}}). \qquad (2.15)$$

Since $s = |\mathcal{S}| \ll n$, the finite sample theory for Huber's M-estimation in low dimensions [44] applies to $\widehat{\boldsymbol{\beta}}^{\mathrm{ora}}$, indicating that with high probability,

$$\|\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*\|_2 \lesssim \|\boldsymbol{w}_{\mathcal{S}}^*\|_2 + b_\tau^*.$$

According to Definition 2.1, the last term $s^{1/2}\epsilon_\ell$ demonstrates the optimization error, which will be discussed in Section 3.

The above results provide conditions under which the sequence of estimators $\{\widehat{\boldsymbol{\beta}}^{(\ell)}\}_{\ell\geq 1}$ satisfies the contraction property and, meanwhile, falls in a local neighborhood of $\boldsymbol{\beta}^*$. Another important feature of the proposed procedure is that the resulting estimator satisfies the strong oracle property, as demonstrated by the following result. Let $\{\widehat{\boldsymbol{\beta}}^{(\ell)}\}_{\ell\geq 1}$ be any optimal solutions to the convex programs $\{(\mathrm{P}_\ell)\}_{\ell\geq 1}$ in (2.3) with $\widehat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$. Similarly to Definition 2.2, we define the following event in regard of the restricted strong convexity of the empirical Huber loss. For some $r, l, \kappa > 0$,

$$\mathcal{E}_2(r,l,\kappa) := \left\{ \inf_{(\boldsymbol{\beta}',\boldsymbol{\beta}'')\in\mathcal{C}(r,l)} \frac{\langle \nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}') - \nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}''), \boldsymbol{\beta}' - \boldsymbol{\beta}''\rangle}{\|\boldsymbol{\beta}' - \boldsymbol{\beta}''\|_2^2} \geq \kappa \right\}. \qquad (2.16)$$

where $\mathcal{C}(r,l) = \{(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) : \boldsymbol{\beta}_1 \in \boldsymbol{\beta}_2 + \mathbb{B}(r) \cap \mathbb{C}(l), \boldsymbol{\beta}_2 \in \boldsymbol{\beta}^* + \mathbb{B}_\Sigma(r), \mathrm{supp}(\boldsymbol{\beta}_2) \subseteq \mathcal{S}\}$ and $\mathbb{B}_\Sigma(r) = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|\boldsymbol{\beta}\|_\Sigma \leq r\}$. Moreover, define the "oracle" score $\boldsymbol{w}^{\mathrm{ora}} \in \mathbb{R}^d$ as

$$\boldsymbol{w}^{\mathrm{ora}} = \nabla\widehat{\mathcal{L}}_\tau(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}}), \qquad (2.17)$$

which satisfies $\boldsymbol{w}_{\mathcal{S}^c}^{\mathrm{ora}} = \mathbf{0}$.

Proposition 2.3. Suppose there exist constants $\gamma_1 > \gamma_0 > 0$ such that $p'(\gamma_0) \in (0, 1/2]$, $p'(\gamma_1) = 0$. For a prespecified $\delta \in (0, 1)$, let $\kappa \geq 1.25/(\delta\gamma_0)$ and choose $c_0 > 0$ so that

$$1 + 0.5p'(\gamma_0)(c_0^2 + 1)^{1/2} = c_0 \kappa \gamma_0. \tag{2.18}$$

Moreover, set $l = \{2 + \frac{2}{p'(\gamma_0)}\}(c_0^2 + 1)^{1/2}s^{1/2}$ and let $r \geq c_0\gamma_0 s^{1/2}\lambda$. Then, conditioned on the event

$$\left\{\|\boldsymbol{w}^{\mathrm{ora}}\|_\infty \leq \frac{p'(\gamma_0)}{2}\lambda\right\} \cap \left\{\|\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*\|_\infty \leq \frac{\lambda}{5\delta\kappa}\right\}$$
$$\cap \left\{\|\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*\|_\Sigma \leq r\right\} \cap \mathcal{E}_2(r, l, \kappa), \tag{2.19}$$

the *strong oracle property* holds under the *minimum signal strength condition* $\|\boldsymbol{\beta}_\mathcal{S}^*\|_{\min} \geq (\gamma_0 + \gamma_1)\lambda$: $\widehat{\boldsymbol{\beta}}^{(\ell)} = \widehat{\boldsymbol{\beta}}^{\mathrm{ora}}$ for all $\ell \geq \lceil \log(s^{1/2}/\delta)/\log(1/\delta) \rceil$.

The proofs of Propositions 2.1, 2.2 and 2.3 are provided in the Supplementary Material.

2.4. *Random analysis*

In this section, we complement the previous deterministic results with probabilistic bounds on the random events of interest. To be more specific, events $\mathcal{E}_1(r, l, \kappa)$ and $\mathcal{E}_2(r, l, \kappa)$ correspond to the RSC properties of $\widehat{\mathcal{L}}_\tau(\cdot)$. The order of the regularization parameter λ depends on $\|\boldsymbol{w}^*\|_\infty$, where $\boldsymbol{w}^* = \nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}^*) - \nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)$ is the centered score function evaluated at $\boldsymbol{\beta}^*$. The oracle convergence rate depends on the ℓ_2-norm of $\boldsymbol{w}_\mathcal{S}^* \in \mathbb{R}^s$, the subvector of $\boldsymbol{w}^*$ indexed by $\mathcal{S}$.

Under Condition 2.1, $\boldsymbol{x} = (x_1, \ldots, x_d)^{\mathrm{T}}$ is sub-exponential and $\Sigma = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}) = (\sigma_{jk})_{1 \leq j, k \leq d}$ is positive definite. Here we do not require the components of $\boldsymbol{x}$ to have zero means. Moreover, given the true active set $\mathcal{S} \subseteq [d]$ of $\boldsymbol{\beta}^*$, we define the following $s \times s$ principal submatrix of Σ:

$$\mathrm{S} = \mathbb{E}(\boldsymbol{x}_\mathcal{S}\boldsymbol{x}_\mathcal{S}^{\mathrm{T}}), \quad \text{where } \boldsymbol{x}_\mathcal{S} \in \mathbb{R}^s \text{ is the subvector of } \boldsymbol{x} \text{ indexed by } \mathcal{S}. \tag{2.20}$$

Throughout, "$\lesssim$" and "$\gtrsim$" stand for "$\leq$" and "$\geq$", respectively, up to constants that are independent of (n, d, s) but might depend on those in Condition 2.1. In particular, define

$$\rho_{\boldsymbol{x}} = \sup_{\boldsymbol{u} \in \mathbb{R}^d} \{\mathbb{E}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{u})^4\}^{1/4}/\{\mathbb{E}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{u})^2\}^{1/2}$$
$$= \sup_{\boldsymbol{u} \in \mathbb{S}^{d-1}} \{\mathbb{E}(\boldsymbol{u}^{\mathrm{T}}\Sigma^{-1/2}\boldsymbol{x})^4\}^{1/4} \geq 1, \tag{2.21}$$

which is a constant depending only on $\sigma_{\boldsymbol{x}}$.

Proposition 2.4. Assume Condition 2.1 holds, and let $\rho_u = \lambda_{\max}(\boldsymbol{\Sigma})$. Then, for any $t \geq 0$,

$$\inf_{\boldsymbol{\beta} \in \boldsymbol{\beta}^* + \mathbb{B}(r) \cap \mathbb{C}(l)} \frac{\langle \nabla \widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) - \nabla \widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle}{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2}$$

$$\geq \frac{3}{4}\rho_l - C_0\sigma_{\boldsymbol{x}}\frac{\tau l}{r}\sqrt{\frac{\log(2d)}{n}} - \rho_u\rho_{\boldsymbol{x}}^2\sqrt{\frac{2t}{n}} - \left(\frac{\tau}{r}\right)^2\frac{t}{3n} \qquad (2.22)$$

holds with probability at least $1 - e^{-t}$ as long as $\tau \geq \max\{C_1\sigma_2, C_2 r\}$ and $n \geq \log(2d)$, where C_0, C_1 are absolute constants and C_2 depends only on $\sigma_{\boldsymbol{x}}$.

The next proposition provides high probability bounds on $\|\boldsymbol{w}^*\|_\infty$ and $\|\boldsymbol{w}_{\mathcal{S}}^*\|_2$, where $\boldsymbol{w}^* = \nabla \widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}^*) - \nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)$.

Proposition 2.5. Assume Condition 2.1 holds. For any $t > 0$, the centered score $\boldsymbol{w}^* \in \mathbb{R}^d$ satisfies

$$\|\boldsymbol{w}^*\|_\infty \leq 2\sigma_{\boldsymbol{x}}\left\{\sigma_2\sqrt{\frac{\log(2d) + t}{n}} + \tau\frac{\log(2d) + t}{2n}\right\}. \qquad (2.23)$$

with probability at least $1 - e^{-t}$, and

$$\|\boldsymbol{w}_{\mathcal{S}}^*\|_2 \leq 3\sigma_{\boldsymbol{x}}\left(\sigma_2\sqrt{\frac{2s + t}{n}} + \tau\frac{2s + t}{2n}\right), \qquad (2.24)$$

with probability at least $1 - e^{-t}$.

Similarly to Theorem 2.1, Propositions 2.4 and 2.5 are also modified versions of Lemmas C.4 and C.6 in [44] under a weaker sub-exponential condition on the feature vector $\boldsymbol{x}$. Therefore in the proofs, we only provide the necessary steps that help improve upon the existing results. Together, Propositions 2.4 and 2.5 reveal the impact of the robustification parameter on the statistical properties of the resulting estimator. As discussed in Section 2.3 above, the order of $\|\boldsymbol{w}(\boldsymbol{\beta}^*)_{\mathcal{S}}\|_2$ determines the oracle rate of convergence. In Theorem 2.2, we show that after only a small number of iterations, the proposed procedure leads to an estimator that achieves the oracle rate of convergence. Recall from Section 2.2 that $\{\widetilde{\boldsymbol{\beta}}^{(\ell)}\}_{\ell=1,2,\dots}$ is a sequence of ϵ_ℓ-optimal solutions of the convex programs (2.3), initialized at $\widetilde{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$.

Theorem 2.2. Assume Conditions 2.1 and 2.2 hold, and there exist some $\gamma_1 > \gamma_0 > 0$ such that

$$\gamma_0 > \sqrt{5}/\rho_l, \quad p'(\gamma_0) > 0, \quad p'(t) = 0 \text{ for all } t \geq \gamma_1. \qquad (2.25)$$

Given $t \geq 0$, suppose the sample size satisfies $n \gtrsim s\log d + t$, and $\epsilon_\ell \leq \sqrt{1/n}$ for all $\ell \geq 1$. Moreover, suppose that we choose a regularization parameter $\lambda \asymp \sigma_2\sqrt{(\log d + t)/n}$, and let τ satisfy

$$\sigma_2 \lesssim \tau \lesssim \sigma_2\sqrt{\frac{n}{\log d + t}} \quad \text{and} \quad b_\tau^* = \|\mathbb{E}\{\ell_\tau'(\varepsilon)\boldsymbol{x}\}\|_2 \leq s^{1/2}\lambda. \qquad (2.26)$$

Then, under the minimum signal strength condition $\|\boldsymbol{\beta}_{\mathcal{S}}^*\|_{\min} \geq (\gamma_0 + \gamma_1)\lambda$, the multi-stage estimator $\widetilde{\boldsymbol{\beta}}^{(T)}$ with $T \gtrsim \lceil \frac{\log(\log d + t)}{\log(1/\delta)} \rceil$ satisfies the bounds

$$
\|\widetilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\|_2 \lesssim \sigma_2 \sqrt{\frac{s+t}{n}} + \tau \frac{s+t}{n} + b_\tau^* \quad \text{and}
$$

$$
\|\widetilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\|_1 \lesssim s^{1/2} \left(\sigma_2 \sqrt{\frac{s+t}{n}} + \tau \frac{s+t}{n} + b_\tau^* \right)
\tag{2.27}
$$

with probability at least $1 - 3e^{-t}$.

We refer to the conclusion of Theorem 2.2 as the *weak oracle property* in the sense that the proposed estimator achieves the same rate of convergence as the oracle $\widehat{\boldsymbol{\beta}}^{\mathrm{ora}}$ which knows *a priori* the support $\mathcal{S}$ of $\boldsymbol{\beta}^*$. We keep the two terms $\tau(s+t)/n$ and b_τ^* in the upper bounds of (2.27) to keep track the impact of τ on the estimator error: the former is part of the stochastic error and the latter characterizes the bias. Below are two cases that are of general interests.

1. (Symmetry) As discussed in Remark 2.2, if ε (conditional on $\boldsymbol{x}$) is symmetric around zero, then $b_\tau^* = 0$ for any $\tau > 0$. To certify (2.2), τ can be taken as a constant-multiple of σ_2, and the resulting error bounds become

$$
\|\widetilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\|_2 \lesssim \sigma_2 \sqrt{\frac{s+t}{n}} \quad \text{and} \quad \|\widetilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\|_1 \lesssim s^{1/2} \sigma_2 \sqrt{\frac{s+t}{n}}
$$

with probability at least $1 - 3e^{-t}$.

2. (Asymmetry) When the conditional distribution of ε_i is asymmetric, there will be a bias-robustness tradeoff. If ε only has bounded second moment, by Lemma A.1 in the Supplementary Material we have $b_\tau^* \lesssim \sigma_2^2 \tau^{-1}$ although $\tau b_\tau^* \to 0$ as $\tau \to \infty$. Then, the multi-step iterative estimator $\widetilde{\boldsymbol{\beta}}^{(T)}$ with $\tau \asymp \sigma_2 \sqrt{n/(s + \log d + t)}$ satisfies, under the scaling $n \gtrsim s \log d + t$, that

$$
\|\widetilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\|_2 \lesssim \sigma_2 \sqrt{\frac{s + \log d + t}{n}} \quad \text{and}
$$

$$
\|\widetilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\|_1 \lesssim s^{1/2} \sigma_2 \sqrt{\frac{s + \log d + t}{n}}
$$

with probability at least $1 - 3e^{-t}$.

A 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. Here we need slightly stronger moment conditions than those in Condition 2.1, that is, the random predictor $\boldsymbol{x}$ is *sub-Gaussian* and the noise variable ε satisfies an $L_{2+\eta}$-L_2 norm equivalence for some $\eta \in (0, 1]$.

Condition 2.3. There exists $\sigma_{\boldsymbol{x}} \geq 1$ such that $\mathbb{P}(|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}| \geq \sigma_{\boldsymbol{x}} t) \leq 2e^{-t^2/2}$ for all $\boldsymbol{u} \in \mathbb{S}^{d-1}$ and $t \geq 0$. Moreover, $\Sigma = (\sigma_{jk}) = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$ satisfies $\rho_l = \lambda_{\min}(\Sigma) > 0$ and

$$
\max_{j \in \mathcal{S}^c} \|\Sigma_{j\mathcal{S}}(\Sigma_{\mathcal{S}\mathcal{S}})^{-1}\|_1 \leq A_0
\tag{2.28}
$$

for some $A_0 > 0$. The random error ε satisfies $\mathbb{E}(\varepsilon|\boldsymbol{x}) = 0$ and $\mathbb{E}(\varepsilon^2|\boldsymbol{x}) \leq \sigma_2^2$, and $\{\mathbb{E}(|\varepsilon|^{2+\eta}|\boldsymbol{x})\}^{1/(2+\eta)} \leq a_\eta \{\mathbb{E}(\varepsilon^2|\boldsymbol{x})\}^{1/2}$ (almost surely) for some $\eta \in (0,1]$ and $a_\eta > 1$. Moreover, ε satisfies the *anti-concentration* property: there exists a constant $a_0 > 0$ such that

$$\mathbb{P}(a \leq \varepsilon \leq b|\boldsymbol{x}) \leq a_0(b - a) \quad \text{for all } a \leq b. \tag{2.29}$$

Theorem 2.3. Assume Conditions 2.2 and 2.3 hold, and there exist some $\gamma_1 > \gamma_0 > 2.5/\rho_l$ such that $p'(\gamma_0) > 0$, $p'(t) = 0$ for all $t \geq \gamma_1$. For any $t \geq 0$ and $q \geq \max(s, \log d)$, let $\lambda \asymp \sigma_2 \sqrt{(\log d + t)/n}$ and $\tau \asymp \sigma_2 \sqrt{n/(q + t)}$. Moreover, assume the sample size satisfies $n \gtrsim \max\{s \log d + t, (q + t)^{1+1/\eta}(\log d)^{-1/\eta}\}$, and the beta-min condition $\|\boldsymbol{\beta}_{\mathcal{S}}^*\|_{\min} \geq (\gamma_0 + \gamma_1)\lambda$. Then, with probability at least $1 - 8e^{-t}$, $\widehat{\boldsymbol{\beta}}^{(\ell)} = \widehat{\boldsymbol{\beta}}^{\text{ora}}$ provided $\ell \geq \lceil \log(s^{1/2}/\delta)/\log(1/\delta) \rceil$, where $\delta := 2.5/(\rho_l \gamma_0) \in (0,1)$.

Theorem 2.3 provides a useful complement to Theorem 2 in [30], and differs from it in two aspects. First, the latter studies the estimator obtained by solving the folded concave penalized optimization program in (1.3) subject to an ℓ_1-ball constraint in order to ensure the existence of local/global optima. Secondly, Theorem 2 in [30] establishes the strong oracle property for any stationary point $\widetilde{\boldsymbol{\beta}}$ of the program (1.3) (with an ℓ_1-ball constraint) that falls inside a local neighborhood of $\boldsymbol{\beta}^*$. In contrast, Theorem 2.3 concerns the strong oracle property of the proposed iteratively reweighted ℓ_1-penalized estimator obtained by solving a sequence of (unconstrained) convex programs (2.3).

Remark 2.3. A direct consequence of the strong oracle property is *variable selection consistency*, saying that

$$\mathbb{P}\{\operatorname{supp}(\widehat{\boldsymbol{\beta}}^{(\ell)}) = \mathcal{S}\} \to 1 \quad \text{as } n, d \to \infty.$$

In particular, assume Condition 2.3 holds with $\eta = 1$, implying that ε satisfies an L_3-L_2 norm equivalence. Then, Theorem 2.3 implies that the multi-step estimator $\widehat{\boldsymbol{\beta}}^{(\ell)}$ with $\lambda \asymp \sigma_2 \sqrt{(\log d)/n}$, $\tau \asymp \sigma_2 \sqrt{n/(s + \log d)}$ and $\ell \asymp \log s$ achieves variable selection consistency as $n, d \to \infty$ under the scaling $n \gtrsim \max(s \log d, s^2)$ and the necessary beta-min condition $\|\boldsymbol{\beta}_{\mathcal{S}}^*\|_{\min} \gtrsim \sigma_2 \sqrt{\log(d)/n}$ [39].

As previously noted, Lasso [46] achieves desirable risk properties, in terms of both estimation and prediction, under mild conditions, yet its variable selection consistency requires much stronger assumptions [23, 37, 48, 55]. In addition to sub-Gaussian errors, it requires a stronger beta-min condition—$\|\boldsymbol{\beta}_{\mathcal{S}}^*\|_{\min} \gtrsim \sigma_2 \sqrt{(s \log d)/n}$, and the *irrepresentable condition*

$$\max_{j \in \mathcal{S}^c} \|\Sigma_{j\mathcal{S}}(\Sigma_{\mathcal{S}\mathcal{S}})^{-1}\|_1 \leq a_0 < 1. \tag{2.30}$$

See, for example, Chapter 7 in [8] and Section 7.5 in [49]. We also refer to [23] for the precise lower and upper irrepresentable conditions, which are necessary and sufficient for the variable selection consistency of Lasso.

3. Optimization algorithm

In this section, we use the local adaptive majorize-minimization (LAMM) principal [16] to derive an iterative algorithm for solving each subproblem (P_ℓ) in (2.3):

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \{\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) + \|\boldsymbol{\lambda}^{(\ell-1)} \circ \boldsymbol{\beta}\|_1\}, \quad \ell = 1, 2, \ldots,$$

where $\boldsymbol{\lambda}^{(\ell-1)} = (\lambda_1^{(\ell-1)}, \ldots, \lambda_d^{(\ell-1)})^\mathrm{T} \in \mathbb{R}^d$ with $\lambda_j^{(\ell-1)} \geq 0$. Specifically, $\lambda_j^{(\ell-1)} = 0$ for some j means that the j-th coefficient is not penalized.

3.1. LAMM algorithm

To minimize a nonlinear function $f(\cdot)$ on $\mathbb{R}^d$, at a given point $\boldsymbol{\beta}^{(k)}$, the majorize-minimization (MM) algorithm first majorizes it by another function $g(\cdot|\boldsymbol{\beta}^{(k)})$, which satisfies

$$g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)}) \geq f(\boldsymbol{\beta}) \quad \text{and} \quad g(\boldsymbol{\beta}^{(k)}|\boldsymbol{\beta}^{(k)}) = f(\boldsymbol{\beta}^{(k)}) \quad \text{for any } \boldsymbol{\beta} \in \mathbb{R}^d,$$

and then compute $\boldsymbol{\beta}^{(k+1)} := \mathrm{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^d} g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)})$ [25]. The objective value of such an algorithm is non-increasing in each step, because

$$f(\boldsymbol{\beta}^{(k+1)}) \overset{\text{(i)}}{\leq} g(\boldsymbol{\beta}^{(k+1)} \,|\, \boldsymbol{\beta}^{(k)}) \overset{\text{(ii)}}{\leq} g(\boldsymbol{\beta}^{(k)} \,|\, \boldsymbol{\beta}^{(k)}) = f(\boldsymbol{\beta}^{(k)}), \tag{3.1}$$

where inequality (i) is due to the majorization property of $g(\cdot|\boldsymbol{\beta}^{(k)})$ and inequality (ii) follows from the definition $\boldsymbol{\beta}^{(k+1)}$. [16] observed that the global majorization requirement is not necessary. It only requires the local properties

$$f(\boldsymbol{\beta}^{(k+1)}) \leq g(\boldsymbol{\beta}^{(k+1)}|\boldsymbol{\beta}^{(k)}) \quad \text{and} \quad g(\boldsymbol{\beta}^{(k)}|\boldsymbol{\beta}^{(k)}) = f(\boldsymbol{\beta}^{(k)}) \tag{3.2}$$

for the inequalities in (3.1) to hold.

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

$$F(\boldsymbol{\beta}; \phi, \boldsymbol{\beta}^{(\ell,k-1)}) := \widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}^{(\ell,k-1)})$$
$$+ \langle \nabla \widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}^{(\ell,k-1)}), \boldsymbol{\beta} - \boldsymbol{\beta}^{(\ell,k-1)} \rangle + \frac{\phi}{2} \|\boldsymbol{\beta} - \boldsymbol{\beta}^{(\ell,k-1)}\|_2^2, \tag{3.3}$$

to locally majorize $\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta})$ such that

$$F(\boldsymbol{\beta}^{(\ell,k)}; \phi^{(\ell,k)}, \boldsymbol{\beta}^{(\ell,k-1)}) \geq \widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}^{(\ell,k)}), \tag{3.4}$$

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

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

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

$$\boldsymbol{\beta}^{(\ell,k)} = S_{\text{soft}}\big(\boldsymbol{\beta}^{(\ell,k-1)} - \nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}^{(\ell,k-1)})/\phi^{(\ell,k)}, \boldsymbol{\lambda}^{(\ell-1)}/\phi^{(\ell,k)}\big), \qquad (3.5)$$

where $S_{\text{soft}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) := (\text{sign}(\beta_j)\max\{|\beta_j| - \lambda_j, 0\})_{j=1,\ldots,d}$ is the soft-thresholding operator. For simplicity, we summarize and define the above update as $\boldsymbol{\beta}^{(\ell,k)} = T_{\boldsymbol{\lambda}^{(\ell-1)},\phi^{(\ell,k)}}(\boldsymbol{\beta}^{(\ell,k-1)})$. Using this simple update formula of $\boldsymbol{\beta}$, we iteratively search for the pair $(\phi^{(\ell,k)}, \boldsymbol{\beta}^{(\ell,k)})$ that ensures the local majorization (3.4). Starting with an initial quadratic coefficient $\phi = \phi_0$, say 10^{-4}, we iteratively increase ϕ by a factor of $\gamma_u > 1$ and compute

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

until the local property (3.4) holds. This routine is summarized in Algorithm 1.

Algorithm 1 LAMM algorithm at the k-th iteration of the ℓ-th subproblem.

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

3.2. Complexity theory

To investigate the complexity theory of the proposed algorithm, we first impose the following standard regularity conditions on the objective function.

Condition 3.1. $\nabla\widehat{\mathcal{L}}_\tau(\cdot)$ is L-Lipschitz continuous for some $L > 0$, that is, $\|\nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}_1) - \nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}_2)\|_\infty \leq L\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2$ for any $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^d$.

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

Theorem 3.1. Assume Condition 3.1 holds and the optimal solution $\widehat{\boldsymbol{\beta}}^{(1)}$ satisfies $\|\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*\|_2 \lesssim s^{1/2}\lambda$. Then, to attain an ϵ_c-optimal solution $\widetilde{\boldsymbol{\beta}}^{(1)}$, i.e. $\omega_{\boldsymbol{\lambda}^{(0)}}(\widetilde{\boldsymbol{\beta}}^{(1)}) \leq \epsilon_c$, in the contraction stage, we need as many as $C_1 L^2(1 + \gamma_u)^2(\|\boldsymbol{\beta}^*\|_2 + s^{1/2}\lambda)^2/\epsilon_c^2$ LAMM iterations in (3.5), where $C_1 > 0$ is a constant independent of (n, d, s).

The sublinear rate in the contraction stage is due to the lack of global strong convexity of the loss function in this stage, because we start with a naive initial value $\mathbf{0}$. Once we enter the contracting region where the estimator is relatively closer 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 of LAMM for solving each subproblem in the tightening stage. To this end, we describe a variant of the sparse eigenvalue condition.

Definition 3.1 (LSE—Localized Sparse Eigenvalue)**.** Given $r, \tau > 0$ and an integer $m \geq |\mathcal{S}|$, the localized sparse eigenvalues are defined as

$$\kappa_+(m, r, \tau) = \sup \left\{ \boldsymbol{\delta}^{\mathrm{T}} \nabla^2 \widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) \boldsymbol{\delta} : \boldsymbol{\delta} \in \mathbb{C}_0(m), \boldsymbol{\beta} \in \boldsymbol{\beta}^* + \mathbb{B}(r) \right\}$$

$$\text{and} \quad \kappa_-(m, r, \tau) = \inf \left\{ \boldsymbol{\delta}^{\mathrm{T}} \nabla^2 \widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) \boldsymbol{\delta} : \boldsymbol{\delta} \in \mathbb{C}_0(m), \boldsymbol{\beta} \in \boldsymbol{\beta}^* + \mathbb{B}(r) \right\},$$

where $\mathbb{C}_0(m) := \{ \boldsymbol{u} \in \mathbb{S}^{d-1} : \mathcal{S} \subseteq \mathrm{supp}(\boldsymbol{u}), |\mathrm{supp}(\boldsymbol{u})| \leq m \}$ denotes a sparse cone.

Condition 3.2. We say an LSE(C_0) condition holds for some $C_0 \geq 1$ if there exist an integer $s' \lesssim s$ and constants $\kappa^*, \kappa_*, C_1 > 0$ such that

$$0 < \kappa_* \leq \kappa_-(C_0 s + 2s', r, \tau) < \kappa_+(C_0 s + 2s', r, \tau) \leq \kappa^*$$
$$\text{and} \quad \kappa_+(s', r, \tau)/\kappa_-(C_0 s + 2s', r, \tau) \leq 1 + C_1 s'/s.$$

Note that if a vector $\boldsymbol{u} \in \mathbb{R}^d$ belongs to the sparse cone $\mathbb{C}_0(m)$ for some $m \geq 1$, by Cauchy-Schwarz inequality we have $\|\boldsymbol{u}\|_1 \leq m^{1/2}\|\boldsymbol{u}\|_2$. This implies that $\boldsymbol{u}$ also falls into the ℓ_1-cone $\mathbb{C}(m^{1/2})$ defined in (2.6). Proposition 2.4 will remain valid, possibly with different constants, if the ℓ_1-cone $\mathbb{C}(l)$ therein is replaced by a sparse cone. Note that Proposition 2.4 controls the minimum LSE. Similar results can be obtained to bound the maximum LSE from above.

Theorem 3.2. Assume LSE(C_0) condition holds for a sufficiently large $C_0 > 1$ and $\tau \gtrsim r \gtrsim \lambda \sqrt{s}$. To obtain an ϵ_{t}-optimal solution $\widetilde{\boldsymbol{\beta}}^{(\ell)}$, i.e. $\omega_{\boldsymbol{\lambda}^{(\ell-1)}}(\widetilde{\boldsymbol{\beta}}^{(\ell)}) \leq \epsilon_{\mathrm{t}}$, in the ℓ-th subproblem for $\ell \geq 2$, we need as many as $C_1 \log(C_2 s^{1/2} \lambda/\epsilon_{\mathrm{t}})$ LAMM iterations in (3.5), where C_1 and C_2 are positive constants.

We summarize the above two theorems in the following result, which characterizes the computational complexity of the whole algorithm.

Corollary 3.1. Assume that the conditions in Theorems 3.1 and 3.2 hold. To achieve a sequence of approximate solutions $\{\widetilde{\boldsymbol{\beta}}^{(\ell)}\}_{\ell=1}^T$ such that $\omega_{\boldsymbol{\lambda}^{(0)}}(\widetilde{\boldsymbol{\beta}}^{(1)}) \leq \epsilon_{\mathrm{c}} \lesssim \lambda$ and $\omega_{\boldsymbol{\lambda}^{(\ell-1)}}(\widetilde{\boldsymbol{\beta}}^{(\ell)}) \leq \epsilon_{\mathrm{t}} \lesssim \sqrt{1/n}$ for $2 \leq \ell \leq T$, the required number of LAMM iterations is of the order $C_1 \epsilon_{\mathrm{c}}^{-2} + C_2(T-1) \log(\epsilon_{\mathrm{t}}^{-1})$, where C_1 and C_2 are positive constants independent of (n, d, s).

4. Extension to general robust losses

Thus far, 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 τ-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 folded concave regularization, leads to statistically optimal estimators that are robust against heavy-tailed errors.

Condition 4.1 (Globally Lipschitz and locally quadratic loss functions). Consider a general loss function $\ell_\tau(\cdot)$ that is of the form $\ell_\tau(x) = \tau^2 \ell(x/\tau)$ for $x \in \mathbb{R}$, where $\ell : \mathbb{R} \mapsto [0, \infty)$ is convex and 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.

Note that Condition 4.1 excludes some important Lipschitz continuous functions, such as the check function for quantile regression and the hinge loss for classification, which do not have a local strong convexity. The recent works [1] and [12, 13] established optimal estimation and excess risk bounds for (regularized) empirical risk minimizers and MOM-type estimators based on general convex and Lipschitz loss functions even without a local quadratic behavior. Our work complements the existing results on ℓ_1-regularized ERM by showing oracle properties of nonconvex regularized methods under stronger signals. For this reason, we need an additional local strong convexity condition on the loss. It remains unclear whether the oracle rates or variable selection consistency can still be achieved without such a local curvature of the loss function.

We now discuss the implications of the three properties in Condition 4.1. First, since $\ell'_\tau(x) = \tau \ell'(x/\tau)$, it follows from property (i) that $\sup_{x \in \mathbb{R}} |\ell'_\tau(x)| \leq c_1 \tau$. The boundedness of $|\ell'_\tau|$ facilitates the use of Bernstein's inequality on deriving upper bounds for the random quantities $\|\boldsymbol{w}^*\|_\infty$ and $\|\boldsymbol{w}^*_\mathcal{S}\|_2$ as in Proposition 2.5, where $\boldsymbol{w}^* = \nabla \widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}^*) - \nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)$ with $\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) = (1/n) \sum_{i=1}^n \ell_\tau(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})$. Next, note that $\ell''_\tau(x) = \ell''(x/\tau)$. Property (ii) indicates that ℓ_τ 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 $\widehat{\mathcal{L}}_\tau$. See Proposition 2.4 and Lemma C.1. Lastly, property (iii) is particularly useful when the error distribution is asymmetric. Even though it can be shown under property (i) that $\nabla \widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}^*)$ is concentrated around its expected value $\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)$ with high probability, $\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*) = -\mathbb{E}\{\ell'_\tau(\varepsilon)\boldsymbol{x}\}$ is typically nonzero when the conditional distribution of ε is asymmetric. However, since $\mathbb{E}(\varepsilon|\boldsymbol{x}) = 0$, we have $\mathbb{E}\{\ell'_\tau(\varepsilon)|\boldsymbol{x}\} = \mathbb{E}\{\ell'_\tau(\varepsilon) - \varepsilon|\boldsymbol{x}\} = \tau \mathbb{E}\{\ell'(\varepsilon/\tau) - \varepsilon/\tau|\boldsymbol{x}\}$. Together with property (iii), this implies

$$|\mathbb{E}\{\ell'_\tau(\varepsilon)|\boldsymbol{x}\}| \leq c_4 \tau \mathbb{E}\{(\varepsilon/\tau)^2|\boldsymbol{x}\} = c_4 \sigma_2^2 \tau^{-1}.$$

We thus use $b^*_\tau = \|\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_2$ to quantify the bias; see Lemma A.1 and Theorem 2.2.

Below we list five examples of $\ell(\cdot)$ (including the Huber loss) that satisfy Condition 4.1.

1. (Huber loss): $\ell(x) = x^2/2 \cdot I(|x| \leq 1) + (|x| - 1/2) \cdot I(|x| > 1)$ with $\ell'(x) = xI(|x| \leq 1) + \mathrm{sign}(x)I(|x| > 1)$ and $\ell''(x) = I(|x| \leq 1)$. Moreover,

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

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 $\ell'''(x) = -3x(1 + x^2)^{-5/2}$ 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): The Huber loss is twice differentiable in $\mathbb{R}$, except at ± 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 ℓ'' 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 is used in [10] 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 ℓ'' 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 plotted in Figure 1 except for the Huber loss. Provided that the loss function $\ell_\tau(\cdot)$ satisfies Condition 4.1, 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 also have Lipschitz continuous second derivatives; see Figure 1. In fact, if the function ℓ satisfies $\ell'(0) = 0$, $\ell''(0) = 1$ and has L_2-Lipschitz second derivative, then property (iii) in Condition 4.1 holds with $c_2 = L_2/2$. The Lipschitz continuity of $\ell''(\cdot)$ also helps remove the anti-concentration condition (2.29) on ε.
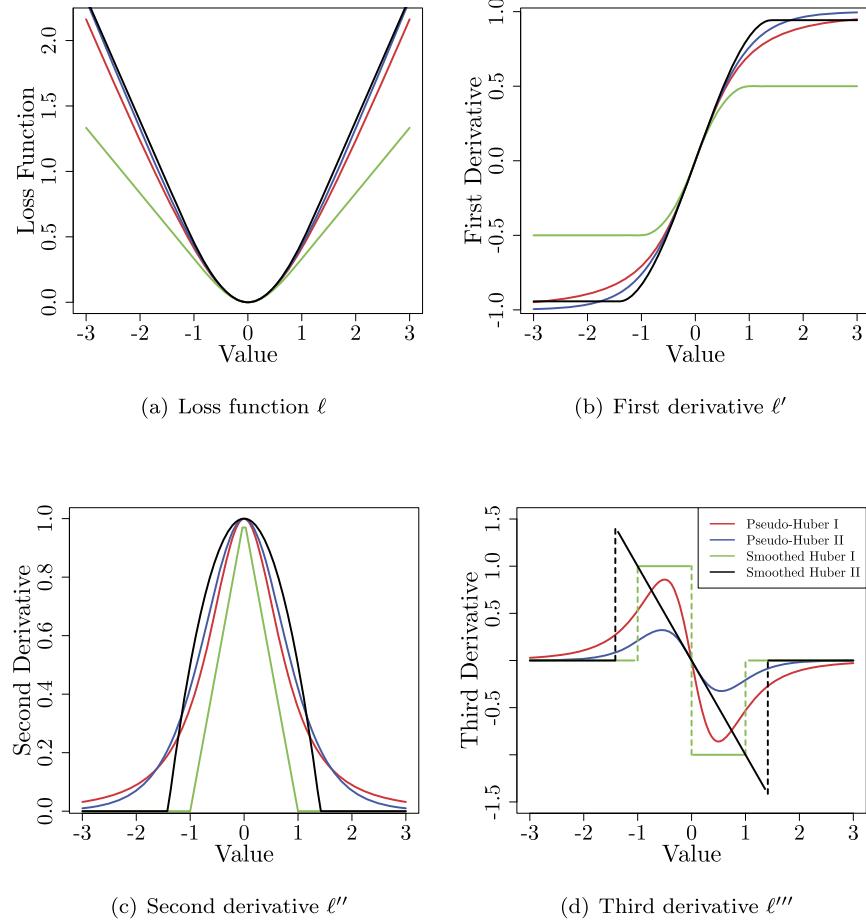

(a) Loss function ℓ

(b) First derivative ℓ'

(c) Second derivative ℓ''

(d) Third derivative ℓ'''

FIG 1. *Examples of robust loss functions and their derivatives.*

5. Numerical study

In this section, we compare the empirical performance of the proposed multi-step penalized robust regression estimator with several benchmark methods, such as the Lasso [46], the SCAD and MC+ penalized least squares [14, 52]. All the computational results presented below are reproducible using software available at https://github.com/XiaoouPan/ILAMM.

We generate data vectors $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ from two types of linear models:

1. (Homoscedastic model): $y_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^* + \varepsilon_i$ with $\boldsymbol{x}_i \sim \mathsf{N}(0, \mathrm{I}_d)$, $i = 1, \ldots, n$.
2. (Heteroscedastic model): $y_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^* + c^{-1}(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^*)^2 \varepsilon_i$ with $\boldsymbol{x}_i \sim \mathsf{N}(0, \mathrm{I}_d)$ for $i = 1, \ldots, n$, where the constant c is chosen as $c = \sqrt{3} \|\boldsymbol{\beta}^*\|_2^2$ such that $\mathbb{E}\{c^{-1}(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^*)^2\}^2 = 1$, and therefore the variance of the noise is the same as that of ε_i.

In addition, we consider the following four error distributions:

1. Normal distribution $\mathsf{N}(\mu, \sigma^2)$ with mean $\mu = 0$ and standard deviation $\sigma = 1.5$;
2. Skewed generalized t distribution $\mathsf{sgt}(0, 5, 0.75, 2, 2.5)$ [45] with mean $\mu = 0$, variance $\sigma^2 = q/(q-2) = 5$, $q = 2.5$, skewness parameter $\lambda = 0.75$ and shape parameter $p = 2$;
3. Lognormal distribution $\mathsf{LN}(\mu, \sigma^2)$ with log location parameter $\mu = 0$ and log shape parameter $\sigma = 1.2$;
4. Pareto distribution $\mathsf{Par}(x_m, \alpha)$ 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 $\boldsymbol{\beta}^* = (4, 3, 2, -2, -2, 2, 0, \ldots, 0)^{\mathrm{T}}$, 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 λ and τ chosen via three-fold cross-validation. To be more specific, we first choose a sequence of λ values the same way as in the **glmnet** algorithm [17]. Guided by its theoretically "optimal" magnitude, the candidate set for τ is taken to be $\{2^j \widehat{\sigma}^{\mathrm{MAD}} \sqrt{n/\log(nd)} : j = -2, -1, 0, 1, 2\}$, where $\widehat{\sigma}^{\mathrm{MAD}} := \mathrm{median}\{|\widehat{\boldsymbol{R}} - \mathrm{median}(\widehat{\boldsymbol{R}})|\}/\Phi^{-1}(3/4)$ is the median absolute deviation (MAD) estimator using the residuals $\widehat{\boldsymbol{R}} = (\widehat{r}_1, \ldots, \widehat{r}_n)^{\mathrm{T}}$ 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, RE_1 and RE_2, which is the relative error of an estimator $\widehat{\boldsymbol{\beta}}$ with respect to the Lasso under ℓ_1- and ℓ_2-norms:

$$RE_1 = \frac{||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1}{||\widehat{\boldsymbol{\beta}}^{\text{Lasso}} - \boldsymbol{\beta}^*||_1} \quad \text{and} \quad RE_2 = \frac{||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2}{||\widehat{\boldsymbol{\beta}}^{\text{Lasso}} - \boldsymbol{\beta}^*||_2}.$$

TABLE 1

Simulation results for the Lasso, SCAD, Huber-SCAD, MC+ and Huber-MC+ estimators under the homoscedastic model.

Error dist.		Lasso	SCAD	Huber-SCAD	MC+	Huber-MC+
Normal	TP	6.00(0)	6.00(0)	6.00(0)	6.00(0)	6.00(0)
	FP	24.44(14.25)	3.11(4.53)	2.19(3.87)	0.84(1.91)	0.53(1.27)
	RE_1	1.00	0.23(0.12)	0.22(0.11)	0.19(0.09)	0.19(0.09)
	RE_2	1.00	0.32(0.13)	0.33(0.13)	0.30(0.12)	0.30(0.12)
Skewed t	TP	4.74(1.37)	4.87(1.35)	4.74(1.39)	3.97(1.67)	3.97(1.62)
	FP	20.78(17.10)	18.49(9.65)	11.48(8.82)	4.28(4.24)	2.76(3.23)
	RE_1	1.00	0.88(0.22)	0.73(0.23)	0.73(0.21)	0.65(0.22)
	RE_2	1.00	0.91(0.17)	0.86(0.19)	0.94(0.20)	0.88(0.23)
Lognormal	TP	5.68(0.87)	5.71(0.84)	6.00(0.07)	5.49(1.14)	5.97(0.37)
	FP	29.70(16.66)	16.75(8.70)	3.80(4.52)	4.32(4.62)	0.91(1.95)
	RE_1	1.00	0.54(0.26)	0.15(0.12)	0.42(0.32)	0.13(0.11)
	RE_2	1.00	0.62(0.26)	0.23(0.14)	0.60(0.30)	0.22(0.14)
Pareto	TP	5.64(1.09)	5.67(1.01)	6.00(0)	5.44(1.35)	5.98(0.35)
	FP	28.30(16.21)	14.69(8.97)	2.91(4.34)	3.48(3.39)	0.71(1.71)
	RE_1	1.00	0.51(0.30)	0.14(0.08)	0.40(0.25)	0.13(0.17)
	RE_2	1.00	0.58(0.26)	0.21(0.11)	0.57(0.28)	0.22(0.22)

TABLE 2

Simulation results for the Lasso, SCAD, Huber-SCAD, MC+ and Huber-MC+ estimators under the heteroscedastic model.

Error dist.		Lasso	SCAD	Huber-SCAD	MC+	Huber-MC+
Normal	TP	6.00(0)	6.00(0)	5.96(0.40)	6.00(0)	5.98(0.28)
	FP	22.71(16.51)	3.29(5.76)	0.31(1.68)	0.88(2.03)	0.13(0.70)
	RE_1	1.00	0.28(0.17)	0.21(0.16)	0.24(0.14)	0.16(0.18)
	RE_2	1.00	0.38(0.19)	0.31(0.16)	0.36(0.18)	0.25(0.14)
Skewed t	TP	4.93(1.59)	5.04(1.53)	5.83(0.65)	4.58(1.76)	5.52(1.17)
	FP	22.99(18.62)	18.21(10.83)	2.71(4.29)	4.99(5.14)	0.92(2.42)
	RE_1	1.00	0.83(0.30)	0.26(0.26)	0.69(0.29)	0.27(0.27)
	RE_2	1.00	0.87(0.28)	0.34(0.26)	0.87(0.29)	0.35(0.28)
Lognormal	TP	5.74(0.96)	5.77(0.91)	6.00(0)	5.65(1.14)	6.00(0)
	FP	26.61(16.51)	11.28(9.50)	1.23(3.55)	2.62(3.45)	0.30(0.86)
	RE_1	1.00	0.45(0.28)	0.14(0.13)	0.35(0.23)	0.12(0.09)
	RE_2	1.00	0.53(0.26)	0.21(0.15)	0.50(0.26)	0.19(0.12)
Pareto	TP	5.67(1.19)	5.67(1.18)	5.97(0.42)	5.59(1.29)	5.95(0.55)
	FP	25.56(16.04)	10.13(10.31)	0.61(2.03)	2.80(4.06)	0.23(0.91)
	RE_1	1.00	0.46(0.29)	0.14(0.12)	0.39(0.28)	0.15(0.23)
	RE_2	1.00	0.55((0.29)	0.22(0.16)	0.54(0.31)	0.23(0.35)

Tables 1 and 2 summarize the averages of each measurement, TP, FP, RE_1, and RE_2 with their standard deviations in brackets, over 200 replications under both homoscedastic and heteroscadastic models. RE_1 and RE_2 for Lasso are defined to be one, so we omit their standard deviations. Here, Huber-SCAD and Huber-MC+ signify the proposed two-stage algorithm using the SCAD and MC+ penalties, respectively. When the noise distributions are heavy-tailed and/or skewed, we see that Huber-SCAD and Huber-MC+ outperform SCAD and MC+, respectively, with fewer spurious discoveries (false positives),

smaller estimation errors and less variability. Under the homoscedastic normal model, Huber-SCAD and Huber-MC+ 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 the multi-step penalized robust regression methods over the existing ones (e.g. Lasso, SCAD and MC+), 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 $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ from both homoscedastic and heteroscedastic models with sample size $n = 100$, dimension $d = 1000$ and sparsity $s = 10$. The true vector of regression coefficients is $\boldsymbol{\beta}^* = (1.5, 1.5, \ldots, 1.5, 0, \ldots, 0)^{\mathrm{T}}$, 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, MC+, Huber-SCAD and Huber-MC+, with a sequence of λ values chosen as before and τ as $\widehat{\sigma}^{\mathrm{MAD}}\sqrt{n/\log(nd)}$. For each combination of λ and τ, the empirical FPR and TPR are computed based on 200 simulations.

Figures 2 and 3 indicate evident advantage of Huber-SCAD and Huber-MC+ 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.

Acknowledgments

Appendix A: Preliminaries

Assume we observe independent data $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ from the linear model $y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \varepsilon_i$. Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)^{\mathrm{T}}$ be a d-vector of regularization parameters with $\lambda_j \geq 0$. Consider the optimization problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \{\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) + \|\boldsymbol{\lambda} \circ \boldsymbol{\beta}\|_1\}, \tag{A.1}$$
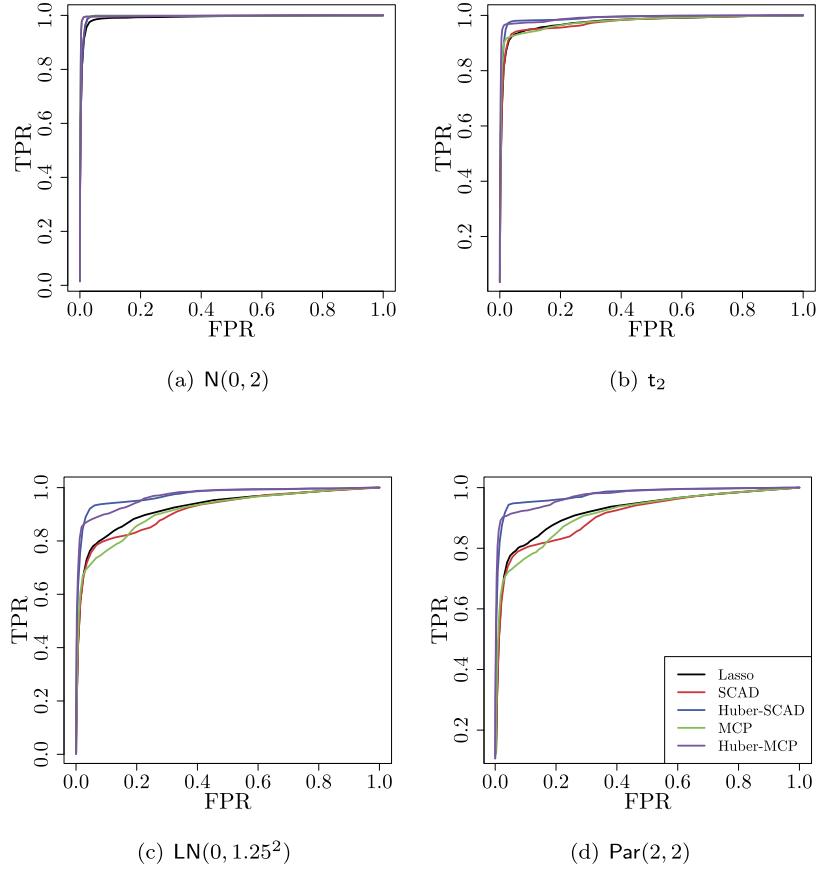
FIG 2. *Plots of ROC curves of the five methods under the homoscedastic model with errors generated from four distributions: normal, Student's t, lognormal and Pareto.*

where $\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n \ell_\tau(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ and $\boldsymbol{\lambda} \circ \boldsymbol{\beta} = (\lambda_1\beta_1, \ldots, \lambda_d\beta_d)^{\mathrm{T}}$. Moreover, define the population loss $\mathcal{L}_\tau(\boldsymbol{\beta}) = \mathbb{E}\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta})$.

The following result provides conditions under which an ϵ-optimal solution to the convex program (A.1) falls in an ℓ_1-cone. Recall that $\mathcal{S} = \mathrm{supp}(\boldsymbol{\beta}^*)$ and $\mathcal{S}^{\mathrm{c}} = [d] \setminus \mathcal{S}$. Moreover, define

$$\boldsymbol{w}(\boldsymbol{\beta}) = \nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) - \nabla\mathcal{L}_\tau(\boldsymbol{\beta}) \quad \text{and} \quad b(\boldsymbol{\beta}) = \|\nabla\mathcal{L}_\tau(\boldsymbol{\beta})\|_2,$$

which are, respectively, the centered score function and the approximation bias. We first characterize the magnitude of the bias $b_\tau^* := b(\boldsymbol{\beta}^*)$, as a function of τ.

Lemma A.1. Assume $\mu_1 = \sup_{\boldsymbol{u} \in \mathbb{S}^{d-1}} \mathbb{E}|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}| < \infty$, $\mathbb{E}(\varepsilon|\boldsymbol{x}) = 0$ and $\mathbb{E}(\varepsilon^2|\boldsymbol{x}) \le \sigma_2^2$ almost surely. Then $|b_\tau^*| \le \mu_1\sigma_2^2\tau^{-1}$ for any $\tau > 0$, and $\tau|b_\tau^*| \to 0$ as $\tau \to \infty$.

(a) $\mathsf{N}(0, 2)$

(b) t_2

(c) $\mathsf{LN}(0, 1.25^2)$

(d) $\mathsf{Par}(2, 2)$
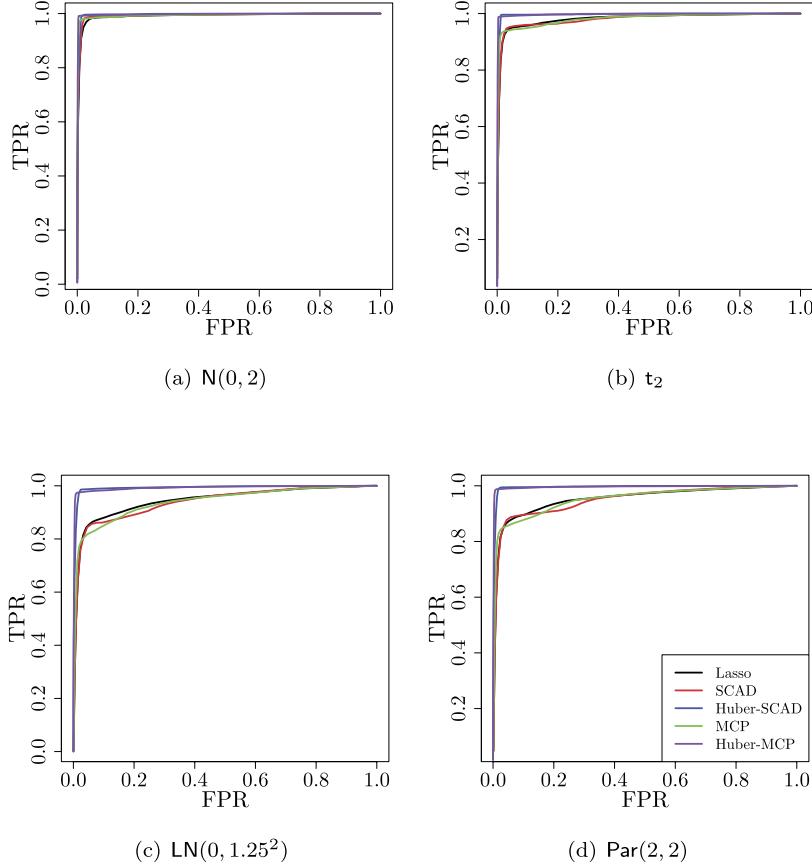
FIG 3. *Plots of ROC curves of the five methods under the heteroscedastic model with errors generated from four distributions: normal, Student's t, lognormal and Pareto.*

Proof of Lemma A.1. Note that $\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*) = \mathbb{E}\{\ell'_\tau(\varepsilon)\boldsymbol{x}\}$, where $\ell'_\tau(u) = uI(|u| \leq \tau) + \tau \operatorname{sign}(u)I(|u| > \tau)$. Recalling $\mathbb{E}(\varepsilon|\boldsymbol{x}) = 0$, it follows that

$$
\begin{aligned}
|\mathbb{E}\{\ell'_\tau(\varepsilon)|\boldsymbol{x}\}| &= |\mathbb{E}[\{\varepsilon - \tau \operatorname{sign}(\varepsilon)\}I\{|\varepsilon| > \tau\}|\boldsymbol{x}]| \\
&\leq \mathbb{E}[\{|\varepsilon| - \tau \operatorname{sign}(\varepsilon)\}I\{|\varepsilon| > \tau\}|\boldsymbol{x}] \\
&\leq \frac{\mathbb{E}\{(\varepsilon^2 - \tau^2)I(|\varepsilon| > \tau)|\boldsymbol{x}\}}{\tau} \leq \frac{\sigma_2^2 - \mathbb{E}\{\ell'_\tau(\varepsilon)^2|\boldsymbol{x}\}}{\tau} \quad \text{almost surely.}
\end{aligned}
$$

By the variational representation of the ℓ_2-norm, we have

$$
\|\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_2 = \sup_{\boldsymbol{u} \in \mathbb{S}^{d-1}} \mathbb{E}\{\ell'_\tau(\varepsilon)\boldsymbol{u}^\mathsf{T}\boldsymbol{x}\} \leq \sigma_2^2 \tau^{-1} \cdot \sup_{\boldsymbol{u} \in \mathbb{S}^{d-1}} \mathbb{E}|\boldsymbol{u}^\mathsf{T}\boldsymbol{x}| = \mu_1 \sigma_2^2 \tau^{-1},
$$

as claimed. The second claim follows from the fact that $\mathbb{E}\{\varepsilon^2 I(|\varepsilon| > \tau)|\boldsymbol{x}\} \to 0$ as $\tau \to \infty$. $\square$

Lemma A.2. Let $\mathcal{E}$ be a subset of $[d]$ that contains $\mathcal{S}$. For any $\boldsymbol{\beta} \in \mathbb{R}^d$ satisfying $\boldsymbol{\beta}_{\mathcal{E}^c} = \mathbf{0}$ and $\epsilon > 0$, provided $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)^{\mathrm{T}}$ satisfies $\|\boldsymbol{\lambda}_{\mathcal{E}^c}\|_{\min} > \|\boldsymbol{w}(\boldsymbol{\beta})\|_\infty + \epsilon$, any ϵ-optimal solution $\widetilde{\boldsymbol{\beta}}$ to (A.1) satisfies

$$\|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}^c}\|_1 \leq \frac{\{\|\boldsymbol{\lambda}\|_\infty + \|\boldsymbol{w}(\boldsymbol{\beta})\|_\infty + \epsilon\}\|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}}\|_1 + b(\boldsymbol{\beta})\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2}{\|\boldsymbol{\lambda}_{\mathcal{E}^c}\|_{\min} - \|\boldsymbol{w}(\boldsymbol{\beta})\|_\infty - \epsilon}.$$

Proof of Lemma A.2. For any $\boldsymbol{\xi} \in \partial\|\widetilde{\boldsymbol{\beta}}\|_1$, define $\boldsymbol{u}_{\boldsymbol{\xi}} = \nabla\widehat{\mathcal{L}}_\tau(\widetilde{\boldsymbol{\beta}}) + \boldsymbol{\lambda} \circ \boldsymbol{\xi} \in \mathbb{R}^d$. Note that

$$\begin{aligned}
\|\boldsymbol{u}_{\boldsymbol{\xi}}\|_\infty\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 &\geq \langle \boldsymbol{u}_{\boldsymbol{\xi}}, \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle \\
&= \underbrace{\langle \nabla\widehat{\mathcal{L}}_\tau(\widetilde{\boldsymbol{\beta}}) - \nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle}_{\geq 0} + \langle \nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) - \nabla\mathcal{L}_\tau(\boldsymbol{\beta}), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle \\
&\quad + \langle \nabla\mathcal{L}_\tau(\boldsymbol{\beta}), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle + \langle \boldsymbol{\lambda} \circ \boldsymbol{\xi}, \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle \\
&\geq -\|\boldsymbol{w}(\boldsymbol{\beta})\|_\infty\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 - b(\boldsymbol{\beta})\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 + \langle \boldsymbol{\lambda} \circ \boldsymbol{\xi}, \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle.
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
\langle \boldsymbol{\lambda} \circ \boldsymbol{\xi}, \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle &= \langle (\boldsymbol{\lambda} \circ \boldsymbol{\xi})_{\mathcal{E}^c}, (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}^c}\rangle + \langle (\boldsymbol{\lambda} \circ \boldsymbol{\xi})_{\mathcal{E}}, (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}}\rangle \\
&\geq \|\boldsymbol{\lambda}_{\mathcal{E}^c}\|_{\min}\|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}^c}\|_1 - \|\boldsymbol{\lambda}_{\mathcal{E}}\|_\infty\|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}}\|_1.
\end{aligned}$$

Together, the last two displays imply

$$\begin{aligned}
\|\boldsymbol{u}_{\boldsymbol{\xi}}\|_\infty\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 &\geq -\|\boldsymbol{w}(\boldsymbol{\beta})\|_\infty\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 - b(\boldsymbol{\beta})\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \\
&\quad + \|\boldsymbol{\lambda}_{\mathcal{E}^c}\|_{\min}\|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}^c}\|_1 - \|\boldsymbol{\lambda}_{\mathcal{E}}\|_\infty\|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}}\|_1.
\end{aligned}$$

Since the right-hand side of this inequality does not depend on $\boldsymbol{\xi}$, taking the infimum with respect to $\boldsymbol{\xi} \in \partial\|\widetilde{\boldsymbol{\beta}}\|_1$ on both sides to reach

$$\begin{aligned}
\omega_{\boldsymbol{\lambda}}(\widetilde{\boldsymbol{\beta}})\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 &\geq -\|\boldsymbol{w}(\boldsymbol{\beta})\|_\infty\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 - b(\boldsymbol{\beta})\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \\
&\quad + \|\boldsymbol{\lambda}_{\mathcal{E}^c}\|_{\min}\|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}^c}\|_1 - \|\boldsymbol{\lambda}_{\mathcal{E}}\|_\infty\|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}}\|_1.
\end{aligned}$$

By definition, $\widetilde{\boldsymbol{\beta}}$ is an ϵ-optimal solution so that $\omega_{\boldsymbol{\lambda}}(\widetilde{\boldsymbol{\beta}}) \leq \epsilon$. Putting together the pieces, we obtain

$$\begin{aligned}
\{\epsilon + \|\boldsymbol{w}(\boldsymbol{\beta})\|_\infty\}&\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 + b(\boldsymbol{\beta})\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \\
&\geq \|\boldsymbol{\lambda}_{\mathcal{E}^c}\|_{\min}\|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}^c}\|_1 - \|\boldsymbol{\lambda}_{\mathcal{E}}\|_\infty\|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}}\|_1.
\end{aligned}$$

Decompose $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$ as $\|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}}\|_1 + \|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}^c}\|_1$, the stated result follows immediately. $\square$

Lemma A.3. Consider some $\boldsymbol{\beta} \in \mathbb{R}^d$ satisfying $\boldsymbol{\beta}_{\mathcal{S}^c} = \mathbf{0}$, and let $\mathcal{E} \subseteq [d]$ be a subset that contains $\mathcal{S}$ and has cardinality $|\mathcal{E}| = k$. Assume that $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)^{\mathrm{T}}$ satisfies $\|\boldsymbol{\lambda}\|_\infty \leq \lambda$ and $\|\boldsymbol{\lambda}_{\mathcal{E}^c}\|_{\min} \geq \rho\lambda > 0$ for some $\rho \in (0, 1]$ and

$\lambda \geq s^{-1/2}b(\boldsymbol{\beta})$. Conditioned on event $\{\|\boldsymbol{w}(\boldsymbol{\beta})\|_\infty + \epsilon \leq 0.5\rho\lambda\}$, any ϵ-optimal solution $\widetilde{\boldsymbol{\beta}}$ to (A.1) satisfies $\widetilde{\boldsymbol{\beta}} \in \boldsymbol{\beta} + \mathbb{C}(l)$, where $l = (2 + \frac{2}{\rho})k^{1/2} + \frac{2}{\rho}s^{1/2}$. Moreover, let $r, \kappa > 0$ satisfy

$$r > \kappa^{-1}(0.5\rho k^{1/2} + 2s^{1/2})\lambda.$$

Then, conditioned on the event $\mathcal{E}_1(r, l, \kappa) \cap \{\|\boldsymbol{w}(\boldsymbol{\beta})\|_\infty + \epsilon \leq 0.5\rho\lambda\}$,

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \leq \kappa^{-1}\{\|\boldsymbol{\lambda}_{\mathcal{S}}\|_2 + \|\boldsymbol{w}(\boldsymbol{\beta})_{\mathcal{E}}\|_2 + k^{1/2}\epsilon\} + \kappa^{-1}b(\boldsymbol{\beta}) \qquad (A.2)$$

$$\leq \kappa^{-1}(0.5\rho k^{1/2} + 2s^{1/2})\lambda < r. \qquad (A.3)$$

Proof of Lemma A.3. For some $r > 0$ to be specified, define $\eta = \sup\{u \in [0, 1] : (1 - u)\boldsymbol{\beta} + u\widetilde{\boldsymbol{\beta}} \in \mathbb{B}(r)\}$, where $\mathbb{B}(r) = \{\boldsymbol{\delta} \in \mathbb{R}^d : \|\boldsymbol{\delta}\|_2 \leq r\}$. Note that $\eta = 1$ if $\widetilde{\boldsymbol{\beta}} \in \boldsymbol{\beta} + \mathbb{B}(r)$ and $\eta \in (0, 1)$ otherwise. Then, the intermediate estimate $\widetilde{\boldsymbol{\beta}}_\eta := \eta\widetilde{\boldsymbol{\beta}} + (1 - \eta)\boldsymbol{\beta}$ satisfies (i) $\widetilde{\boldsymbol{\beta}}_\eta \in \boldsymbol{\beta} + \mathbb{B}(r)$, (ii) $\widetilde{\boldsymbol{\beta}}_\eta$ lies on the boundary of $\boldsymbol{\beta} + \mathbb{B}(r)$ with $0 < \eta < 1$ if $\widetilde{\boldsymbol{\beta}} \notin \boldsymbol{\beta} + \mathbb{B}(r)$, and (iii) $\widetilde{\boldsymbol{\beta}}_\eta = \widetilde{\boldsymbol{\beta}}$ with $\eta = 1$ if $\widetilde{\boldsymbol{\beta}} \in \boldsymbol{\beta} + \mathbb{B}(r)$.

By the convexity of Huber loss and Lemma F.2 in [16], we have

$$\langle \nabla\widehat{\mathcal{L}}_\tau(\widetilde{\boldsymbol{\beta}}_\eta) - \nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}), \widetilde{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta} \rangle \leq \eta\langle \nabla\widehat{\mathcal{L}}_\tau(\widetilde{\boldsymbol{\beta}}) - \nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle. \qquad (A.4)$$

First we bound the left-hand side of (A.4) from below. Conditioned on the stated event, Lemma A.2 indicates

$$\|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}^c}\|_1 \leq (1 + 2/\rho)\|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}}\|_1 + 2(\rho\lambda)^{-1}b(\boldsymbol{\beta})\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2,$$

from which it follows that $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \leq (2 + \frac{2}{\rho})k^{1/2}\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 + \frac{2}{\rho}\lambda^{-1}b(\boldsymbol{\beta})\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$. Provided that $\lambda \geq s^{-1/2}b(\boldsymbol{\beta})$, this implies $\widetilde{\boldsymbol{\beta}} \in \boldsymbol{\beta} + \mathbb{C}(l)$ with $l = (2 + \frac{2}{\rho})k^{1/2} + \frac{2}{\rho}s^{1/2}$. Since $\widetilde{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta} = \eta(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$, we have $\widetilde{\boldsymbol{\beta}}_\eta \in \boldsymbol{\beta} + \mathbb{B}(r) \cap \mathbb{C}(l)$ and conditioned on event $\mathcal{E}_1(r, l, \kappa)$,

$$\langle \nabla\widehat{\mathcal{L}}_\tau(\widetilde{\boldsymbol{\beta}}_\eta) - \nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}), \widetilde{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta} \rangle \geq \kappa\|\widetilde{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}\|_2^2. \qquad (A.5)$$

Next we upper bound the right-hand side of (A.4). For any $\boldsymbol{\xi} \in \partial\|\widetilde{\boldsymbol{\beta}}\|_1$, write

$$\langle \nabla\widehat{\mathcal{L}}_\tau(\widetilde{\boldsymbol{\beta}}) - \nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle$$
$$= \langle \boldsymbol{u}, \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle - \langle \boldsymbol{\lambda} \circ \boldsymbol{\xi}, \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle - \langle \nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle$$
$$:= \Pi_1 - \Pi_2 - \Pi_3, \qquad (A.6)$$

where $\boldsymbol{u} = \nabla\widehat{\mathcal{L}}_\tau(\widetilde{\boldsymbol{\beta}}) + \boldsymbol{\lambda} \circ \boldsymbol{\xi}$. For $\Pi_3 = \langle \nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle$, by the decomposition $\nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) = \boldsymbol{w}(\boldsymbol{\beta}) + \nabla\mathcal{L}_\tau(\boldsymbol{\beta})$ we have

$$|\Pi_3| \leq \|\boldsymbol{w}(\boldsymbol{\beta})_{\mathcal{E}}\|_2\|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}}\|_2$$
$$+ \|\boldsymbol{w}(\boldsymbol{\beta})_{\mathcal{E}^c}\|_\infty\|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}^c}\|_1 + b(\boldsymbol{\beta})\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2. \qquad (A.7)$$

Turning to Π_2, decompose $\boldsymbol{\lambda} \circ \boldsymbol{\xi}$ and $\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}$ according to $\mathcal{S} \cup (\mathcal{E} \setminus \mathcal{S}) \cup \mathcal{E}^{\mathrm{c}}$ to reach

$$\Pi_2 = \langle (\boldsymbol{\lambda} \circ \boldsymbol{\xi})_{\mathcal{S}}, (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{S}} \rangle + \langle (\boldsymbol{\lambda} \circ \boldsymbol{\xi})_{\mathcal{E} \setminus \mathcal{S}}, (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E} \setminus \mathcal{S}} \rangle + \langle (\boldsymbol{\lambda} \circ \boldsymbol{\xi})_{\mathcal{E}^{\mathrm{c}}}, (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}^{\mathrm{c}}} \rangle.$$

Since $\boldsymbol{\beta}_{\mathcal{E}^{\mathrm{c}}} = \mathbf{0}$ and $\boldsymbol{\xi} \in \partial \|\widetilde{\boldsymbol{\beta}}\|_1$, we have $\langle (\boldsymbol{\lambda} \circ \boldsymbol{\xi})_{\mathcal{E}^{\mathrm{c}}}, (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}^{\mathrm{c}}} \rangle = \langle \boldsymbol{\lambda}_{\mathcal{E}^{\mathrm{c}}}, |\widetilde{\boldsymbol{\beta}}_{\mathcal{E}^{\mathrm{c}}}| \rangle = \langle \boldsymbol{\lambda}_{\mathcal{E}^{\mathrm{c}}}, |(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}^{\mathrm{c}}}| \rangle$. Also, $\langle (\boldsymbol{\lambda} \circ \boldsymbol{\xi})_{\mathcal{E} \setminus \mathcal{S}}, (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E} \setminus \mathcal{S}} \rangle = \langle (\boldsymbol{\lambda} \circ \boldsymbol{\xi})_{\mathcal{E} \setminus \mathcal{S}}, \widetilde{\boldsymbol{\beta}}_{\mathcal{E} \setminus \mathcal{S}} \rangle \geq 0$. Therefore,

$$\Pi_2 \geq -\|\boldsymbol{\lambda}_{\mathcal{S}}\|_2 \|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{S}}\|_2 + \|\boldsymbol{\lambda}_{\mathcal{E}^{\mathrm{c}}}\|_{\min} \|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}^{\mathrm{c}}}\|_1. \tag{A.8}$$

Similarly, Π_1 satisfies the bound

$$|\Pi_1| \leq \|\boldsymbol{u}_{\mathcal{E}}\|_2 \|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}}\|_2 + \|\boldsymbol{u}\|_{\infty} \|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}^{\mathrm{c}}}\|_1. \tag{A.9}$$

Together, (A.6)–(A.9) yield

$$\langle \nabla \widehat{\mathcal{L}}_{\tau}(\widetilde{\boldsymbol{\beta}}) - \nabla \widehat{\mathcal{L}}_{\tau}(\boldsymbol{\beta}), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle$$
$$\leq -\{\|\boldsymbol{\lambda}_{\mathcal{E}^{\mathrm{c}}}\|_{\min} - \|\boldsymbol{w}(\boldsymbol{\beta})\|_{\infty}\} \|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}^{\mathrm{c}}}\|_1 + \|\boldsymbol{u}\|_{\infty} \|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}^{\mathrm{c}}}\|_1$$
$$+ \{\|\boldsymbol{w}(\boldsymbol{\beta})_{\mathcal{E}}\|_2 + \|\boldsymbol{u}_{\mathcal{E}}\|_2\} \|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}}\|_2 + \|\boldsymbol{\lambda}_{\mathcal{S}}\|_2 \|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{S}}\|_2 + b(\boldsymbol{\beta}) \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2.$$

Taking the infimum over $\boldsymbol{\xi} \in \partial \|\widetilde{\boldsymbol{\beta}}\|_1$ on both sides, it follows that

$$\langle \nabla \widehat{\mathcal{L}}_{\tau}(\widetilde{\boldsymbol{\beta}}) - \nabla \widehat{\mathcal{L}}_{\tau}(\boldsymbol{\beta}), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle$$
$$\leq -\{\|\boldsymbol{\lambda}_{\mathcal{E}^{\mathrm{c}}}\|_{\min} - \|\boldsymbol{w}(\boldsymbol{\beta})\|_{\infty} - \epsilon\} \|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}^{\mathrm{c}}}\|_1$$
$$+ \{\|\boldsymbol{w}(\boldsymbol{\beta})_{\mathcal{E}}\|_2 + k^{1/2} \epsilon\} \|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{E}}\|_2$$
$$+ \|\boldsymbol{\lambda}_{\mathcal{S}}\|_2 \|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{S}}\|_2 + b(\boldsymbol{\beta}) \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2. \tag{A.10}$$

It follows from (A.4), (A.5) and (A.10) that conditioned on $\mathcal{E}_1(r, l, \kappa) \cap \{\|\boldsymbol{w}(\boldsymbol{\beta})\|_{\infty} + \epsilon \leq 0.5 \rho \lambda\}$,

$$\kappa \|\widetilde{\boldsymbol{\beta}}_{\eta} - \boldsymbol{\beta}\|_2^2 \leq \{\|\boldsymbol{\lambda}_{\mathcal{S}}\|_2 + \|\boldsymbol{w}(\boldsymbol{\beta})_{\mathcal{E}}\|_2 + k^{1/2} \epsilon\} \|\widetilde{\boldsymbol{\beta}}_{\eta} - \boldsymbol{\beta}\|_2 + b(\boldsymbol{\beta}) \|\widetilde{\boldsymbol{\beta}}_{\eta} - \boldsymbol{\beta}\|_2, \tag{A.11}$$

On the same event, note that

$$\|\boldsymbol{w}(\boldsymbol{\beta})_{\mathcal{E}}\|_2 + k^{1/2} \epsilon \leq k^{1/2} \{\|\boldsymbol{w}(\boldsymbol{\beta})\|_{\infty} + \epsilon\} \leq 0.5 \rho k^{1/2} \lambda.$$

Moreover, recall that $\|\boldsymbol{\lambda}_{\mathcal{S}}\|_2 \leq s^{1/2} \lambda$ and $b(\boldsymbol{\beta}) \leq s^{1/2} \lambda$. Plugging these bounds into (A.11) yields

$$\|\widetilde{\boldsymbol{\beta}}_{\eta} - \boldsymbol{\beta}\|_2 \leq \kappa^{-1} \{(s^{1/2} + 0.5 \rho k^{1/2}) \lambda + b(\boldsymbol{\beta})\} \leq \kappa^{-1} (2 s^{1/2} + 0.5 \rho k^{1/2}) \lambda < r.$$

Hence, $\widetilde{\boldsymbol{\beta}}_{\eta}$ falls in the interior of $\boldsymbol{\beta} + \mathbb{B}(r)$. Via proof by contradiction, we must have $\eta = 1$ and $\widetilde{\boldsymbol{\beta}}_{\eta} = \widetilde{\boldsymbol{\beta}}$. Consequently, (A.2) and (A.3) follow, respectively, from (A.11) and the last display. $\qquad\square$

In Lemma A.3, we need λ to be sufficiently large in the sense that for some s-sparse vector $\boldsymbol{\beta}$,

$$\lambda \gtrsim \|\boldsymbol{w}(\boldsymbol{\beta})\|_\infty \quad \text{and} \quad \lambda \geq s^{-1/2}b(\boldsymbol{\beta}),$$

where $\boldsymbol{w}(\boldsymbol{\beta}) = \nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) - \nabla\mathcal{L}_\tau(\boldsymbol{\beta})$ is the centered gradient at $\boldsymbol{\beta}$ and $b(\boldsymbol{\beta}) = \|\nabla\mathcal{L}_\tau(\boldsymbol{\beta})\|_2$ quantifies the bias. To prove the weak oracle property (Proposition 2.2), we will take $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ and control the stochastic term $\|\boldsymbol{w}^*\|_\infty = \|\boldsymbol{w}(\boldsymbol{\beta}^*)\|_\infty$ and bias term $b(\boldsymbol{\beta}^*)$ separately. For some $\boldsymbol{\beta}$ which has vanishing or negligible bias, we will only focus on the stochastic term $\|\nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta})\|_\infty$, as described in the next lemma. Recall the event $\mathcal{E}_2(r, l, \kappa)$ defined in (2.16), and $\mathbb{B}_\Sigma(r) = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|\boldsymbol{\beta}\|_\Sigma \leq r\}$.

Lemma A.4. Consider some $\boldsymbol{\beta} \in \boldsymbol{\beta}^* + \mathbb{B}_\Sigma(r)$ $(r > 0)$ satisfying $\boldsymbol{\beta}_{\mathcal{S}^c} = \boldsymbol{0}$, and let $\mathcal{E} \subseteq [d]$ be a subset that contains $\mathcal{S}$ and has cardinality $|\mathcal{E}| = k$. Assume that $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)^{\mathrm{T}}$ satisfies $\|\boldsymbol{\lambda}\|_\infty \leq \lambda$ and $\|\boldsymbol{\lambda}_{\mathcal{E}^c}\|_{\min} \geq \rho\lambda > 0$ for some $\rho \in (0, 1]$. Conditioned on event $\{\|\nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta})\|_\infty \leq 0.5\rho\lambda\}$, any optimal solution $\widehat{\boldsymbol{\beta}}$ to (A.1) satisfies $\widehat{\boldsymbol{\beta}} \in \boldsymbol{\beta} + \mathbb{C}(l)$, where $l = (2 + \frac{2}{\rho})k^{1/2} + \frac{2}{\rho}s^{1/2}$. Moreover, let $r, \kappa > 0$ satisfy $r > \kappa^{-1}(0.5\rho k^{1/2} + 2s^{1/2})\lambda$. Then, conditioned on the event $\mathcal{E}_2(r, l, \kappa) \cap \{\|\nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta})\|_\infty \leq 0.5\rho\lambda\}$,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \leq \kappa^{-1}\{\|\boldsymbol{\lambda}_{\mathcal{S}}\|_2 + \|\nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta})_{\mathcal{E}}\|_2\}.$$

The proof of Lemma A.4 is based on the same arguments from the proof of Lemmas A.2 and A.3, and thus is omitted. To prove the strong oracle property (Proposition 2.3), we will apply Lemma A.4 with $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{\mathrm{ora}}$, the oracle estimator defined in (2.15).

Appendix B: Proofs of propositions

B.1. Proof of Proposition 2.1

With the initial estimate $\widetilde{\boldsymbol{\beta}}^{(0)} = \boldsymbol{0}_d$, we have $\boldsymbol{\lambda}^{(0)} = p'_\lambda(\boldsymbol{0}_d) = (\lambda, \ldots, \lambda)^{\mathrm{T}} \in \mathbb{R}^d$. Then (2.9) follows immediately from Lemma A.3 with $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, $\mathcal{E} = \mathcal{S}$ and $\rho = 1$. $\qquad\square$

B.2. Proof of Proposition 2.2

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, $\max_{j \in \mathcal{S}^c} |\widetilde{\beta}_j^{(\ell-1)}|$. Recall that $\boldsymbol{\lambda}^{(\ell-1)} = (\lambda_1^{(\ell-1)}, \ldots, \lambda_d^{(\ell-1)})^{\mathrm{T}} = (p'_\lambda(|\widetilde{\beta}_1^{(\ell-1)}|), \ldots, p'_\lambda(|\widetilde{\beta}_d^{(\ell-1)}|))^{\mathrm{T}}$ and $p_\lambda(t) = \lambda^2 p(t/\lambda)$ for $t \in \mathbb{R}$. Intuitively, the larger $|\widetilde{\beta}_j^{(\ell-1)}|$ 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 $\|\boldsymbol{\lambda}_{\mathcal{E}_\ell^c}^{(\ell-1)}\|_{\min}$.

Starting from $\widetilde{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$, we have $\boldsymbol{\lambda}^{(0)} = (\lambda, \ldots, \lambda)^{\mathrm{T}} \in \mathbb{R}^d$. Recall from (2.10) that $\lambda \geq s^{-1/2} b_\tau^*$, or equivalently, $\lambda^{-1} b_\tau^* \leq s^{1/2}$. Then, applying Lemma A.3 with $\mathcal{E} = \mathcal{S}$ and $l_0 = 6s^{1/2}$ we obtain that, conditioning on $\mathcal{E}_1(r, l_0, \kappa) \cap \{\lambda \geq \frac{2}{p'(\gamma)}(\|\boldsymbol{w}^*\|_\infty + \epsilon_1)\}$,

$$
\begin{aligned}
\|\widetilde{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*\|_2 &\leq \kappa^{-1}\big\{\big(\|\boldsymbol{\lambda}_{\mathcal{S}}^{(0)}\|_2 + \|\boldsymbol{w}_{\mathcal{S}}^*\|_2 + s^{1/2}\epsilon_1\big) + b_\tau^*\big\} \\
&\leq \kappa^{-1}\{1 + 0.5p'(\gamma)\}s^{1/2}\lambda + \kappa^{-1}b_\tau^* \\
&\leq \kappa^{-1}\{2 + 0.5p'(\gamma)\}s^{1/2}\lambda,
\end{aligned}
\tag{B.1}
$$

where the last inequality is due to (2.10). For $\ell \geq 1$, define the augmented set

$$
\mathcal{E}_\ell = \mathcal{S} \cup \{1 \leq j \leq d : \lambda_j^{(\ell-1)} < p'(\gamma)\lambda\},
\tag{B.2}
$$

which depends on the solution $\widetilde{\boldsymbol{\beta}}^{(\ell-1)}$ from the previous step. We claim that the above constructed sets satisfy

$$
|\mathcal{E}_\ell| < (c^2 + 1)s \quad \text{and} \quad \|\boldsymbol{\lambda}_{\mathcal{E}_\ell^c}^{(\ell-1)}\|_{\min} \geq p'(\gamma)\lambda,
\tag{B.3}
$$

where c is the constant determined by (2.11). If these were true, it follows from Lemma A.3 with $\rho = p'(\gamma)$, $k = (c^2 + 1)s$ and $l = (2 + \frac{2}{\rho})(c^2 + 1)^{1/2}s^{1/2} + \frac{2}{\rho}s^{1/2}$ that, conditioned on $\mathcal{E}_1(r, l, \kappa) \cap \{\lambda \geq \frac{2}{p'(\gamma)}(\|\boldsymbol{w}^*\|_\infty + \epsilon_1)\}$,

$$
\begin{aligned}
\|\widetilde{\boldsymbol{\beta}}^{(\ell)} - \boldsymbol{\beta}^*\|_2 &\leq \kappa^{-1}\big(\|\boldsymbol{\lambda}_{\mathcal{S}}^{(\ell-1)}\|_2 + \|\boldsymbol{w}_{\mathcal{E}_\ell}^*\|_2 + |\mathcal{E}_\ell|^{1/2}\epsilon_\ell\big) + \kappa^{-1}b_\tau^* \\
&< \kappa^{-1}\big\{1 + 0.5(c^2 + 1)^{1/2}p'(\gamma)\big\}s^{1/2}\lambda + \kappa^{-1}s^{1/2}\lambda \\
&= \kappa^{-1}\big\{2 + 0.5(c^2 + 1)^{1/2}p'(\gamma)\big\}s^{1/2}\lambda \\
&= \underbrace{c\gamma s^{1/2}\lambda}_{=:r^{\mathrm{crude}}} \leq r,
\end{aligned}
\tag{B.4}
$$
$$
\tag{B.5}
$$

where the last two steps follow from (2.11) and (2.12).

We prove the earlier claim (B.3) by the method of induction. For $\ell = 1$, we have $\boldsymbol{\lambda}^{(0)} = (\lambda, \ldots, \lambda)^{\mathrm{T}} \in \mathbb{R}^d$. Thus, (B.3) holds with $\mathcal{E}_1 = \mathcal{S}$. Next, assume (B.3) holds for some $\ell \geq 1$, from which (B.5) follows. To bound the cardinality of $\mathcal{E}_{\ell+1}$, note that for any $j \in \mathcal{E}_{\ell+1} \setminus \mathcal{S}$, $p'_\lambda(|\widetilde{\beta}_j^{(\ell)}|) = \lambda_j^{(\ell)} < p'(\gamma)\lambda = p'_\lambda(\gamma\lambda)$. This, together with the monotonicity of p'_λ on $\mathbb{R}_+$, implies $|\widetilde{\beta}_j^{(\ell)}| > \gamma\lambda$. Recalling that $\beta_j^* = 0$ for $j \in \mathcal{E}_{\ell+1} \setminus \mathcal{S}$, we obtain

$$
\begin{aligned}
|\mathcal{E}_{\ell+1} \setminus \mathcal{S}|^{1/2} &< \frac{1}{\gamma\lambda}\|\widetilde{\boldsymbol{\beta}}_{\mathcal{E}_{\ell+1}\setminus\mathcal{S}}^{(\ell)}\|_2 = \frac{1}{\gamma\lambda}\|(\widetilde{\boldsymbol{\beta}}^{(\ell)} - \boldsymbol{\beta}^*)_{\mathcal{E}_{\ell+1}\setminus\mathcal{S}}\|_2 \\
&\overset{\text{(i)}}{\leq} \frac{c\gamma s^{1/2}\lambda}{\gamma\lambda} = cs^{1/2},
\end{aligned}
\tag{B.6}
$$

where inequality (i) applies the bound (B.5). Hence, $|\mathcal{E}_{\ell+1}| \leq |\mathcal{S}| + |\mathcal{E}_{\ell+1} \setminus \mathcal{S}| < (c^2 + 1)s$. By (B.2) and the property $p'_\lambda(t) = \lambda p'(t/\lambda)$, we are guaranteed that

$$
\lambda_j^{(\ell)} \geq p'(\gamma)\lambda \geq 2(\|\boldsymbol{w}^*\|_\infty + \epsilon_{\ell+1}) \quad \text{for} \quad j \in \mathcal{E}_{\ell+1}^c.
$$

The two hypotheses in (B.3) then hold for $\ell + 1$, which completes the induction step. Consequently, the bounds (B.4) and (B.5) hold for any $\ell \geq 1$.

We have shown that under proper conditions, all the estimates $\widetilde{\boldsymbol{\beta}}^{(\ell)}$ fall in a local neighborhood of $\boldsymbol{\beta}^*$, i.e., $\|\widetilde{\boldsymbol{\beta}}^{(\ell)} - \boldsymbol{\beta}^*\|_\Sigma \leq r^{\mathrm{crude}} = c\gamma s^{1/2}\lambda$. To further refine this bound as signal strengthens, on the right-hand side of (B.4), we need to establish sharper bounds on

$$\|\boldsymbol{\lambda}_{\mathcal{S}}^{(\ell-1)}\|_2 = \sqrt{\sum_{j \in \mathcal{S}}\{\lambda_j^{(\ell-1)}\}^2} \quad \text{and} \quad \|\boldsymbol{w}_{\mathcal{E}_\ell}^*\|_2 + |\mathcal{E}_\ell|^{1/2}\epsilon_\ell,$$

and maintain the bias term b_τ^*, instead of replacing it with an upper bound $s^{1/2}\lambda$. For each $j \in [d]$, $\lambda_j^{(\ell-1)} = p_\lambda'(|\widetilde{\beta}_j^{(\ell-1)}|)$. If $|\widetilde{\beta}_j^{(\ell-1)} - \beta_j^*| \geq \gamma\lambda$, then $\lambda_j^{(\ell-1)} \leq \lambda \leq \gamma^{-1}|\widetilde{\beta}_j^{(\ell-1)} - \beta_j^*|$; otherwise if $|\widetilde{\beta}_j^{(\ell-1)} - \beta_j^*| \leq \gamma\lambda$, $\lambda_j^{(\ell-1)} \leq p_\lambda'(|\beta_j^*| - \gamma\lambda)$ due to monotonicity of p_λ'. Putting together the pieces, we conclude that

$$\|\boldsymbol{\lambda}_{\mathcal{S}}^{(\ell-1)}\|_2 \leq \|p_\lambda'(|\boldsymbol{\beta}_{\mathcal{S}}^*| - \gamma\lambda)\|_2 + \gamma^{-1}\|(\widetilde{\boldsymbol{\beta}}^{(\ell-1)} - \boldsymbol{\beta}^*)_{\mathcal{S}}\|_2.$$

For the remaining terms that involve $\mathcal{E}_\ell$, by the triangle inequality and (B.6) we obtain that

$$\begin{aligned}
&\|\boldsymbol{w}_{\mathcal{E}_\ell}^*\|_2 + |\mathcal{E}_\ell|^{1/2}\epsilon_\ell \\
&\leq \|\boldsymbol{w}_{\mathcal{S}}^*\|_2 + s^{1/2}\epsilon_\ell + |\mathcal{E}_\ell \setminus \mathcal{S}|^{1/2}\|\boldsymbol{w}^*\|_\infty + |\mathcal{E}_\ell \setminus \mathcal{S}|^{1/2}\epsilon_\ell \\
&< \|\boldsymbol{w}_{\mathcal{S}}^*\|_2 + s^{1/2}\epsilon_\ell + \frac{\|\boldsymbol{w}^*\|_\infty + \epsilon_\ell}{\gamma\lambda}\|(\widetilde{\boldsymbol{\beta}}^{(\ell-1)} - \boldsymbol{\beta}^*)_{\mathcal{E}_\ell \setminus \mathcal{S}}\|_2 \\
&\leq \|\boldsymbol{w}_{\mathcal{S}}^*\|_2 + s^{1/2}\epsilon_\ell + \frac{p'(\gamma)}{2\gamma}\|(\widetilde{\boldsymbol{\beta}}^{(\ell-1)} - \boldsymbol{\beta}^*)_{\mathcal{E}_\ell \setminus \mathcal{S}}\|_2.
\end{aligned}$$

Recall that $p'(\gamma) \leq 1$, and for any $a, b \geq 0$, $\sqrt{a} + \sqrt{b/4} \leq \sqrt{5(a+b)/4}$. Plugging the above refined bounds into (B.4) yields

$$\begin{aligned}
&\|\widetilde{\boldsymbol{\beta}}^{(\ell)} - \boldsymbol{\beta}^*\|_2 \\
&\leq \kappa^{-1}\big\{\|p_\lambda'(|\boldsymbol{\beta}_{\mathcal{S}}^*| - \gamma\lambda)\|_2 + \|\boldsymbol{w}_{\mathcal{S}}^*\|_2 + s^{1/2}\epsilon_\ell + b_\tau^*\big\} + \frac{\sqrt{5}}{2\gamma\kappa}\|(\widetilde{\boldsymbol{\beta}}^{(\ell-1)} - \boldsymbol{\beta}^*)_{\mathcal{E}_\ell}\|_2.
\end{aligned}$$

Taking $\delta = \sqrt{5}/(2\gamma\kappa) \in (0, 1)$, the contraction inequality (2.13) follows immediately. Finally, (2.14) is a direct consequence of (2.13) and (B.1). $\qquad\square$

B.3. Proof of Proposition 2.3

By construction, the oracle estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{ora}}$ is such that $\widehat{\boldsymbol{\beta}}_{\mathcal{S}^c}^{\mathrm{ora}} = \boldsymbol{0} \in \mathbb{R}^{d-s}$ and $\nabla\widehat{\mathcal{L}}_\tau(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}})_{\mathcal{S}} = \boldsymbol{0} \in \mathbb{R}^s$. With $\boldsymbol{w}^{\mathrm{ora}} = \nabla\widehat{\mathcal{L}}_\tau(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}})$, the proof strategy is similar to that in the proof of Proposition 2.2 with $\epsilon_\ell = 0$, because $\widehat{\boldsymbol{\beta}}^{(\ell)}$ are optimal solutions to (P_ℓ).

Recall that $l = \{2 + \frac{2}{p'(\gamma_0)}\}(c_0^2 + 1)^{1/2}s^{1/2}$ with $c_0 > 0$ determined by (2.18). Conditioned on the event

$$\{\|\boldsymbol{w}^{\mathrm{ora}}\|_\infty \leq 0.5p'(\gamma_0)\lambda\} \cap \{\|\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*\|_\Sigma \leq r\} \cap \mathcal{E}_2(r, l, \kappa),$$

following the proof of Proposition 2.2 and applying Lemma A.4 with $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{\mathrm{ora}}$ and $\mathcal{E} = \mathcal{E}_\ell$, it can be similarly shown that

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}}^{(\ell)} - \widehat{\boldsymbol{\beta}}^{\mathrm{ora}}\|_2 &\leq \kappa^{-1}\big(\|\boldsymbol{\lambda}_{\mathcal{S}}^{(\ell-1)}\|_2 + \|\boldsymbol{w}_{\mathcal{E}_\ell}^{\mathrm{ora}}\|_2\big) \\
&\leq \kappa^{-1}\big(s^{1/2}\lambda + |\mathcal{E}_\ell|^{1/2}\|\boldsymbol{w}^{\mathrm{ora}}\|_\infty\big) \\
&< \kappa^{-1}\big\{1 + 0.5p'(\gamma_0)(c_0^2 + 1)^{1/2}\big\}s^{1/2}\lambda = c_0\gamma_0 s^{1/2}\lambda \leq r, \quad \text{(B.7)}
\end{aligned}$$

where similarly to (B.2) and (B.3), $\mathcal{E}_\ell = \mathcal{S} \cup \{1 \leq j \leq d : \lambda_j^{(\ell-1)} < p'(\gamma_0)\lambda\}$ is such that $|\mathcal{E}_\ell \setminus \mathcal{S}| < c_0^2 s$ and thus $|\mathcal{E}_\ell| < (c_0^2 + 1)s$. In this case, the approximation bias is hidden in $\|\boldsymbol{w}^{\mathrm{ora}}\|_\infty$. Moreover, define a sequence of subsets

$$\mathcal{S}_\ell = \{1 \leq j \leq d : |\widehat{\beta}_j^{(\ell)} - \beta_j^*| \geq \gamma_0\lambda\}, \quad \ell = 0, 1, 2, \ldots.$$

Starting with $\widehat{\boldsymbol{\beta}}^{(0)} = \boldsymbol{0}$, it holds under the minimum signal strength condition that $\mathcal{S}_0 = \mathcal{S}$.

To obtain a refined upper bound on $\|\boldsymbol{\lambda}_{\mathcal{S}}^{(\ell-1)}\|_2$, note that if $j \in \mathcal{S} \cap \mathcal{S}_{\ell-1}^{\mathrm{c}}$, $\lambda_j^{(\ell-1)} = p'_\lambda(|\widehat{\beta}_j^{(\ell-1)}|) \leq p'_\lambda(|\beta_j^*| - \gamma_0\lambda)$ due to monotonicity; otherwise if $j \in \mathcal{S} \cap \mathcal{S}_{\ell-1}$, $\lambda_j^{(\ell-1)} \leq \lambda$. Therefore,

$$\|\boldsymbol{\lambda}_{\mathcal{S}}^{(\ell-1)}\|_2 \leq \|p'_\lambda(|\boldsymbol{\beta}_{\mathcal{S}}^*| - \gamma_0\lambda)\|_2 + \lambda|\mathcal{S} \cap \mathcal{S}_{\ell-1}|^{1/2}.$$

Since $\|\boldsymbol{\beta}_{\mathcal{S}}^*\|_{\min} \geq (\gamma_0 + \gamma_1)\lambda$ and $p'_\lambda(t) = 0$ for all $t \geq \gamma_1\lambda$, $\|p'_\lambda(|\boldsymbol{\beta}_{\mathcal{S}}^*| - \gamma_0\lambda)\|_2$ vanishes. Turning to $\|\boldsymbol{w}_{\mathcal{E}_\ell}^{\mathrm{ora}}\|_2$, by the first-order condition of minimizing $\boldsymbol{\beta}_{\mathcal{S}} \mapsto (1/n)\sum_{i=1}^n \widehat{\mathcal{L}}_\tau(y_i - \boldsymbol{x}_{i,\mathcal{S}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{S}})$, we have $\boldsymbol{w}_{\mathcal{S}}^{\mathrm{ora}} = \boldsymbol{0}$ and hence

$$\|\boldsymbol{w}_{\mathcal{E}_\ell}^{\mathrm{ora}}\|_2 = \|\boldsymbol{w}_{\mathcal{E}_\ell \setminus \mathcal{S}}^{\mathrm{ora}}\|_2 \leq \|\boldsymbol{w}^{\mathrm{ora}}\|_\infty |\mathcal{E}_\ell \setminus \mathcal{S}|^{1/2}.$$

For each $j \in \mathcal{E}_\ell \setminus \mathcal{S}$, $\beta_j^* = 0$ and $\lambda_j^{(\ell-1)} = p'_\lambda(|\widehat{\beta}_j^{(\ell-1)}|) < p'(\gamma_0)\lambda = p'_\lambda(\gamma_0\lambda)$. Hence, $|\widehat{\beta}_j^{(\ell-1)} - \beta_j^*| = |\widehat{\beta}_j^{(\ell-1)}| > \gamma_0\lambda$ so that $j \in \mathcal{S}_{\ell-1} \setminus \mathcal{S}$. Therefore, $\mathcal{E}_\ell \setminus \mathcal{S} \subseteq \mathcal{S}_{\ell-1} \setminus \mathcal{S}$. Combined with the earlier bound, we arrive at

$$\|\boldsymbol{w}_{\mathcal{E}_\ell}^{\mathrm{ora}}\|_2 \leq \|\boldsymbol{w}^{\mathrm{ora}}\|_\infty |\mathcal{S}_{\ell-1} \setminus \mathcal{S}|^{1/2}.$$

Since $p'(\gamma_0) \leq 1/2$, substituting the above estimates into (B.7) yields

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}}^{(\ell)} - \widehat{\boldsymbol{\beta}}^{\mathrm{ora}}\|_2 &\leq \frac{|\mathcal{S} \cap \mathcal{S}_{\ell-1}|^{1/2} + |\mathcal{S}_{\ell-1} \setminus \mathcal{S}|^{1/2}/4}{\kappa}\lambda \\
&\leq \frac{\sqrt{17}}{4}\frac{\lambda}{\kappa}|\mathcal{S}_{\ell-1}|^{1/2}. \quad \text{(B.8)}
\end{aligned}$$

Next we bound $|\mathcal{S}_\ell|$ ($\ell \geq 1$), the cardinality of $\mathcal{S}_\ell$. By (2.19), it holds for any $j \in \mathcal{S}_\ell$ that

$$|\widehat{\beta}_j^{(\ell)} - \widehat{\beta}_j^{\mathrm{ora}}| \geq \gamma_0 \lambda - \|\widehat{\beta}^{\mathrm{ora}} - \beta^*\|_\infty \geq 1.25 \frac{\lambda}{\delta\kappa} - 0.2 \frac{\lambda}{\delta\kappa} = 1.05 \frac{\lambda}{\delta\kappa}.$$

In conjunction with (B.8), this implies

$$|\mathcal{S}_\ell|^{1/2} \leq \frac{\|\widehat{\beta}^{(\ell)} - \widehat{\beta}^{\mathrm{ora}}\|_2}{1.05\lambda/(\kappa\delta)} \leq \frac{\frac{\sqrt{17}}{4}(\lambda/\kappa)|\mathcal{S}_{\ell-1}|^{1/2}}{1.05\lambda/(\kappa\delta)} < \delta|\mathcal{S}_{\ell-1}|^{1/2}, \quad \ell \geq 1. \quad \text{(B.9)}$$

Recall that $\mathcal{S}_0 = \mathcal{S}$, we have $|\mathcal{S}_\ell|^{1/2} < \delta^\ell s^{1/2}$ for any $\ell \geq 1$. As long as $\ell \geq T := \lceil \log(s^{1/2})/\log(1/\delta) \rceil$, we are guaranteed that $|\mathcal{S}_\ell| < 1$, i.e. $\mathcal{S}_\ell = \emptyset$. Consequently, it follows from (B.8) that $\widehat{\beta}^{(\ell)} = \widehat{\beta}^{\mathrm{ora}}$ for all $\ell \geq T+1$. This completes the proof. $\qquad\square$

B.4. Proof of Proposition 2.4

The proof is based on similar arguments that were used in the proof of Lemmas C.3 and C.4 in [44]. We only present the necessary steps in order to slightly relax the sub-Gaussian condition on $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})^{\mathrm{T}}$.

For any $\boldsymbol{\beta} \in \boldsymbol{\beta}^* + \mathbb{B}(r) \cap \mathbb{C}(l)$, write $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$. Following the proof of Lemma C.3 in [44], it can be shown under Condition 2.1 that

$$\langle \nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) - \nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle$$
$$\geq \left\{ 1 - (2\sigma_2/\tau)^2 \right\} \|\boldsymbol{\delta}\|_\Sigma^2 - \mathbb{E}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\delta})^2 I\{|\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\delta}|/\|\boldsymbol{\delta}\|_2 \geq \tau/(4r)\} - \Delta(r,l) \cdot \|\boldsymbol{\delta}\|_2^2$$
$$\geq \left\{ 1 - (2\sigma_2/\tau)^2 - \rho_{\boldsymbol{x}}^2 e^{-\tau/(8\sigma_{\boldsymbol{x}} r)} \right\} \cdot \|\boldsymbol{\delta}\|_\Sigma^2 - \Delta(r,l) \cdot \|\boldsymbol{\delta}\|_2^2, \quad \text{(B.10)}$$

where the second inequality uses the bound

$$\mathbb{E}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\delta})^2 I\{|\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\delta}|/\|\boldsymbol{\delta}\|_2 \geq \tau/(4r)\} \leq \left\{ \mathbb{E}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\delta})^4 \right\}^{1/2} \mathbb{P}\left(|\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\delta}|/\|\boldsymbol{\delta}\|_2 \geq \tau/(4r) \right)^{1/2}$$
$$\leq \rho_{\boldsymbol{x}}^2 e^{-\tau/(8\sigma_{\boldsymbol{x}} r)} \|\boldsymbol{\delta}\|_\Sigma^2$$

and $\Delta(r,l) = \sup_{\boldsymbol{\delta} \in \mathbb{B}(r) \cap \mathbb{C}(l)} (1/n) \sum_{i=1}^n |f_{\boldsymbol{\delta}}(\boldsymbol{x}_i, \varepsilon_i) - \mathbb{E}f_{\boldsymbol{\delta}}(\boldsymbol{x}_i, \varepsilon_i)|$ with

$$f_{\boldsymbol{\delta}}(\boldsymbol{x}_i, \varepsilon_i) := I(|\varepsilon_i| \leq \tau/2) \cdot \varphi_{\tau\|\boldsymbol{\delta}\|_2/(2r)}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta})/\|\boldsymbol{\delta}\|_2^2$$

and $\varphi_c(u) := u^2 I(|u| \leq c/2) + (|u| - c)^2 I(c/2 < |u| \leq c)$ for $u \in \mathbb{R}$ and $c \geq 0$. We thus let $\tau \geq \max\{4\sqrt{2}\sigma_2, 8\log(8\rho_{\boldsymbol{x}}^2)\sigma_{\boldsymbol{x}} r\}$ so that $(2\sigma_2/\tau)^2 \leq 1/8$ and $\rho_{\boldsymbol{x}}^2 e^{-\tau/(8\sigma_{\boldsymbol{x}} r)} \leq 1/8$. It then follows from (B.10) that

$$\langle \nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) - \nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle \geq \frac{3}{4} \cdot \|\boldsymbol{\delta}\|_\Sigma^2 - \Delta(r,l) \cdot \|\boldsymbol{\delta}\|_2^2 \quad \text{(B.11)}$$

holds uniformly over $\boldsymbol{\beta} \in \boldsymbol{\beta}^* + \mathbb{B}(r) \cap \mathbb{C}(l)$.

It remains to control the supremum $\Delta(r,l)$. Note that $\varphi_c(\cdot)$ is c-Lipschitz continuous, and satisfies $\varphi_{bc}(bu) = b^2\varphi_c(u)$ for $b, c > 0$ and $u \in \mathbb{R}$. Thus, the above $f_{\boldsymbol{\delta}}(\boldsymbol{x}_i, \varepsilon_i)$ can be simplified as

$$f_{\boldsymbol{\delta}}(\boldsymbol{x}_i, \varepsilon_i) = I(|\varepsilon_i| \leq \tau/2) \cdot \varphi_{\tau/(2r)}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2).$$

Moreover, since $0 \leq \varphi_c(u) \leq \min\{(c/2)^2, u^2\}$, we have $0 \leq f_{\boldsymbol{\delta}}(\boldsymbol{x}_i, \varepsilon_i) \leq \tau^2/(4r)^2$ and by (2.21), $\mathbb{E}f_{\boldsymbol{\delta}}^2(\boldsymbol{x}_i, \varepsilon_i) \leq \mathbb{E}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2)^4 \leq \rho_{\boldsymbol{x}}^4\{\mathbb{E}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2)^2\}^2 \leq (\rho_u\rho_{\boldsymbol{x}}^2)^2$ for all $\boldsymbol{\delta} \in \mathbb{R}^d$. Then, applying Bousquet's version of Talagrand's inequality (see, e.g. Theorem 7.3 in [7]), we obtain that for every $t \geq 0$,

$$\Delta(r,l) \leq \mathbb{E}\Delta(r,l) + \sqrt{\mathbb{E}\Delta(r,l)(\tau/2r)^2 t/n + 2(\rho_u\rho_{\boldsymbol{x}}^2)^2 t/n} + (\tau/4r)^2 t/(3n)$$

$$\leq 1.25\mathbb{E}\Delta(r,l) + \rho_u\rho_{\boldsymbol{x}}^2\sqrt{2t/n} + (\tau/r)^2 t/(3n) \tag{B.12}$$

with probability at least $1 - e^{-t}$. For $\mathbb{E}\Delta(r,l)$, using Rademacher symmetrization gives

$$\mathbb{E}\Delta(r,l) \leq 2\mathbb{E}\left\{\sup_{\boldsymbol{\delta}\in\mathbb{B}(r)\cap\mathbb{C}(l)}\left|\frac{1}{n}\sum_{i=1}^{n}e_i f_{\boldsymbol{\delta}}(\boldsymbol{x}_i, \varepsilon_i)\right|\right\},$$

where $e_1, \ldots, e_n$ are independent Rademacher random variables. Since $\varphi_c(\cdot)$ is c-Lipshitz, $f_{\boldsymbol{\delta}}(\boldsymbol{x}_i, \varepsilon_i)$ is a $(\tau/2r)$-Lipschitz function in $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2$, i.e., for any sample $(\boldsymbol{x}_i, \varepsilon_i)$ and parameters $\boldsymbol{\delta}, \boldsymbol{\delta}' \in \mathbb{R}^d$,

$$|f_{\boldsymbol{\delta}}(\boldsymbol{x}_i, \varepsilon_i) - f_{\boldsymbol{\delta}'}(\boldsymbol{x}_i, \varepsilon_i)| \leq \frac{\tau}{2r}|\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2 - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta}'/\|\boldsymbol{\delta}'\|_2|.$$

Moreover, observe that $f_{\boldsymbol{\delta}}(\boldsymbol{x}_i, \varepsilon_i) = 0$ for any $\boldsymbol{\delta}$ such that $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2 = 0$, and $I(|\varepsilon_i| \leq \tau/2) \in \{0, 1\}$. Then, applying Talagrand's contraction principle (see, e.g. Theorem 4.4, Theorem 4.12 and (4.20) in [29]) yields

$$\mathbb{E}\Delta(r,l) \leq 2\mathbb{E}\left\{\sup_{\boldsymbol{\delta}\in\mathbb{B}(r)\cap\mathbb{C}(l)}\left|\frac{1}{n}\sum_{i=1}^{n}e_i f_{\boldsymbol{\delta}}(\boldsymbol{x}_i, \varepsilon_i)\right|\right\}$$

$$\leq \frac{2\tau}{r}\mathbb{E}\left\{\sup_{\boldsymbol{\delta}\in\mathbb{B}(r)\cap\mathbb{C}(l)}\left|\frac{1}{n}\sum_{i=1}^{n}e_i \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2\right|\right\} \leq \frac{2\tau l}{r}\cdot\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}e_i \boldsymbol{x}_i\right\|_{\infty}, \tag{B.13}$$

where the last inequality uses the cone constraint that $\|\boldsymbol{\delta}\|_1 \leq l\|\boldsymbol{\delta}\|_2$. Next, we apply a maximal inequality for sub-exponential random variables to bound the last term on the right-hand side of (B.13). For $j = 1, \ldots, d$, define partial sums $S_j = \sum_{i=1}^{n}e_i x_{ij}$, of which each summand satisfies $\mathbb{E}(e_i x_{ij}) = 0$ and $\mathbb{E}(e_i x_{ij})^2 = \sigma_{jj}$. More over, for $k = 2, 4, \ldots$, $\mathbb{E}|e_i|^k = 1$ and

$$\mathbb{E}|e_i x_{ij}|^k \leq \sigma_{\boldsymbol{x}}^k \cdot k \int_0^{\infty} t^{k-1}\mathbb{P}(|x_{ij}| \geq \sigma_{\boldsymbol{x}}t)\, \mathrm{d}t \leq \sigma_{\boldsymbol{x}}^k \cdot k \int_0^{\infty} t^{k-1}e^{-t}\, \mathrm{d}t$$

$$= k!\sigma_{\boldsymbol{x}}^k = \frac{k!}{2}\underbrace{2\sigma_{\boldsymbol{x}}^2}_{\geq \sigma_{jj}}\sigma_{\boldsymbol{x}}^{k-2}.$$

By the symmetry of Rademacher random variables, and Bernstein's inequality (see, e.g. Theorems 2.10 in [6]), we obtain that

$$\log \mathbb{E}e^{\lambda S_j} = \log \mathbb{E}e^{-\lambda S_j} \le \psi(\lambda) := \frac{\nu \lambda^2}{2(1 - c\lambda)} \quad \text{for all } \lambda \in (0, 1/c),$$

where $\nu = 2n\sigma_{\boldsymbol{x}}^2$ and $c = \sigma_{\boldsymbol{x}}$. Following the proof of Theorems 2.5 in [6], it can be shown that

$$\mathbb{E} \max_{1 \le j \le d} |S_j| \le \inf_{\lambda \in (0, 1/c)} \left\{ \frac{\log(2d) + \psi(\lambda)}{\lambda} \right\} = \sqrt{2v \log(2d)} + c \log(2d).$$

Re-arranging terms and using (B.13), we find that

$$\mathbb{E}\Delta(r, l) \le 2\sigma_{\boldsymbol{x}}(\tau l/r) \left\{ 2\sqrt{\frac{\log(2d)}{n}} + \frac{\log(2d)}{n} \right\}.$$

Combining this with (B.11) and (B.12) yields the bound (2.22). $\quad\square$

B.5. Proof of Proposition 2.5

Write $S_j = (1/n) \sum_{i=1}^{n} (\xi_i x_{ij} - \mathbb{E}\xi_i x_{ij})$ with $\xi_i := \ell'_\tau(\varepsilon_i)$ for $j = 1, \ldots, d$, so that $\|\boldsymbol{w}^*\|_\infty = \max_{1 \le j \le d} |S_j|$. Note that $\ell'_\tau(u) = uI(|u| \le \tau) + \tau \operatorname{sign}(u) I(|u| > \tau)$, we have $\mathbb{E}(\xi_i^2 | \boldsymbol{x}_i) \le \sigma_2^2$ and $|\xi_i| \le \tau$. It follows that $\mathbb{E}(\xi_i x_{ij})^2 \le \sigma_2^2 \sigma_{jj}$ and under Condition 2.1,

$$\mathbb{E}|\xi_i x_{ij}|^k \le \tau^{k-2}\sigma_2^2 \cdot \mathbb{E}|x_{ij}|^k \le \tau^{k-2}\sigma_2^2 \cdot k!\sigma_{\boldsymbol{x}}^k \le \frac{k!}{2} \cdot 2\sigma_2^2 \sigma_{\boldsymbol{x}}^2 \cdot (\tau\sigma_{\boldsymbol{x}})^{k-2},$$

for $k = 3, 4, \ldots$. Bernstein's inequality, in conjunction with the union bound, implies that for any $x \ge 0$,

$$\max_{1 \le j \le d} |S_j| \le \sigma_{\boldsymbol{x}} \left(2\sigma_2 \sqrt{\frac{2x}{n}} + \tau \frac{x}{n} \right)$$

with probability at least $1 - 2de^{-x}$. Taking $x = \log(2d) + t$ proves (2.23). Next we use a standard covering argument to prove (2.24). For any $\epsilon \in (0, 1)$, there exists an ϵ-net $\mathcal{N}_\epsilon$ of the unit sphere in $\mathbb{R}^s$ with cardinality $|\mathcal{N}_\epsilon| \le (1 + 2/\epsilon)^s$ such that

$$\|\boldsymbol{w}_{\mathcal{S}}^*\|_2 \le \frac{1}{1 - \epsilon} \max_{\boldsymbol{u} \in \mathcal{N}_\epsilon} \frac{1}{n} \sum_{i=1}^{n} \left(\xi_i \boldsymbol{u}^{\mathrm{T}} \boldsymbol{x}_{i,\mathcal{S}} - \mathbb{E}\xi_i \boldsymbol{u}^{\mathrm{T}} \boldsymbol{x}_{i,\mathcal{S}} \right). \tag{B.14}$$

For every $\boldsymbol{u} \in \mathcal{N}_\epsilon$, Bernstein's condition holds: $\mathbb{E}(\xi_i \boldsymbol{u}^{\mathrm{T}} \boldsymbol{x}_{i,\mathcal{S}})^2 \le 2\sigma_2^2 \sigma_{\boldsymbol{x}}^2$ and for $k = 3, 4, \ldots$,

$$\mathbb{E}|\xi_i \boldsymbol{u}^{\mathrm{T}} \boldsymbol{x}_{i,\mathcal{S}}|^k \le \frac{k!}{2} \cdot 2\sigma_2^2 \sigma_{\boldsymbol{x}}^2 \cdot (\tau\sigma_{\boldsymbol{x}})^{k-2}.$$

Again, applying (one-sided) Bernstein's inequality yields, for any $x > 0$,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\xi_i\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}} - \mathbb{E}\xi_i\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}}\right) \leq \sigma_{\boldsymbol{x}}\left(2\sigma_2\sqrt{\frac{x}{n}} + \tau\frac{x}{n}\right)$$

with probability at least $1 - e^{-x}$. Consequently, from the union bound and (B.14), we have

$$\|\boldsymbol{w}(\boldsymbol{\beta})_{\mathcal{S}}\|_2 \leq \frac{\sigma_{\boldsymbol{x}}}{1-\epsilon}\left(2\sigma_2\sqrt{\frac{x}{n}} + \tau\frac{x}{n}\right)$$

with probability at least $1 - e^{\log(1+2/\epsilon)s - x}$. Taking $\epsilon = 1/3$ and $x = 2s + t$ proves (2.24). □

Appendix C: Proofs of theorems

C.1. Proof of Theorem 2.1

We will apply Propositions 2.1, 2.4 and 2.5 to prove Theorem 2.1. To begin with, let $\lambda, r, \kappa > 0$ satisfy (2.8), that is, $\lambda \geq s^{-1/2}b_\tau^*$ and $r > 2.5\kappa^{-1}s^{1/2}\lambda$. Applying Proposition 2.1 with $\epsilon_1 = 0$ yields that, conditioned on the event $\mathcal{E}_1(r, 6s^{1/2}, \kappa) \cap \{\lambda \geq 2\|\boldsymbol{w}^*\|_\infty\}$, the Huber-Lasso estimator defined in (2.1) satisfies

$$\|\widehat{\boldsymbol{\beta}}^{\text{H-Lasso}} - \boldsymbol{\beta}^*\|_2 \leq \kappa^{-1}\left(1.5s^{1/2}\lambda + b_\tau^*\right) \leq 2.5\kappa^{-1}s^{1/2}\lambda. \qquad (\text{C.1})$$

It remains to control the above event of interest. Taking $l = 6s^{1/2}$ and $\kappa = \rho_l/2$, it follows from Proposition 2.4 that $\mathbb{P}\{\mathcal{E}_1(r, l, \kappa)\} \geq 1 - e^{-t}$ as long as $\tau \gtrsim \max(\sigma_2, r)$ and $n \gtrsim (\tau/r)^2(s\log d + t)$. Furthermore, Proposition 2.5 ensures that event $\{\lambda \geq 2\|\boldsymbol{w}^*\|_\infty\}$ occurs with probability at least $1 - e^{-t}$ if the regularization parameter satisfies

$$\lambda \geq \sigma_{\boldsymbol{x}}\left(4\sigma_2\sqrt{\frac{\log(2d) + t}{n}} + 2\tau\frac{\log(2d) + t}{n}\right).$$

Finally, we take $\tau \asymp \sigma_2\sqrt{n/(\log d + t)}$ and $r \asymp \tau$. By Lemma A.1 and Condition 2.1, $b_\tau^* \leq \sigma_{\boldsymbol{x}}\sigma_2^2\tau^{-1} \asymp \sigma_2\sqrt{(\log d + t)/n}$. Putting together the pieces, we conclude that under the scaling $n \gtrsim s\log d + t$, and if λ has magnitude of the order within the range

$$\sigma_2\sqrt{\frac{\log d + t}{n}} \lesssim \lambda \lesssim \sigma_2 s^{-1/2}\sqrt{\frac{n}{\log d + t}},$$

the event $\mathcal{E}_1(r, 6s^{1/2}, \rho_l/2) \cap \{\lambda \geq 2\|\boldsymbol{w}^*\|_\infty\}$ occurs with probability at least $1 - 2e^{-t}$. Combined with (C.1), this proves the claimed bound. □

C.2. Proof of Theorem 2.2

We will apply Propositions 2.2, 2.4 and 2.5 to prove (2.27). The key is to control the random events from Proposition 2.2, which relies on a delicate combination of all the parameters. Similarly to the proof of Theorem 2.1, we take $\kappa = \rho_l/2$, and let $\lambda \geq s^{-1/2}b_\tau^*$, where $b_\tau^* = \|\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_2 \leq \sigma_{\boldsymbol{x}}\sigma_2^2\tau^{-1}$ due to Condition 2.1. Given $\gamma_1 > \gamma_0 > 0$ satisfying (2.25), define $\delta = p'(\gamma_0)/(\rho_l\gamma_0) \in (0,1)$, and let $c > 0$ be the constant determined by $0.5p'(\gamma_0)(c^2+1)^{1/2} + 4 = c\rho_l\gamma_0$, which verifies (2.11) with $\kappa = \rho_l/2$. Moreover, set $l = \{2 + \frac{2}{p'(\gamma_0)}\}(c^2+1)^{1/2}s^{1/2} + \frac{2}{p'(\gamma_0)}s^{1/2}$, and let

$$r \geq c\gamma_0 s^{1/2}\lambda =: r^{\text{crude}}.$$

Consequently, we apply Proposition 2.2 to conclude that, conditioned on event $\mathcal{E}_1(r, l, \rho_l/2) \cap \{\|\boldsymbol{w}^*\|_\infty + n^{-1/2} \leq 0.5p'(\gamma_0)\lambda\}$, the ϵ_ℓ-optimal solutions $\widetilde{\boldsymbol{\beta}}^{(\ell)}$ satisfy

$$\|\widetilde{\boldsymbol{\beta}}^{(\ell)} - \boldsymbol{\beta}^*\|_2 \leq \delta \cdot \|\widetilde{\boldsymbol{\beta}}^{(\ell-1)} - \boldsymbol{\beta}^*\|_2$$
$$+ 2\rho_l^{-1}\big\{\|p'_\lambda(|\boldsymbol{\beta}_{\mathcal{S}}^*| - \gamma_0\lambda)\|_2 + \|\boldsymbol{w}_{\mathcal{S}}^*\|_2 + (s/n)^{1/2} + b_\tau^*\big\}$$

for all $\ell \geq 2$. Under the minimum signal strength condition $\|\boldsymbol{\beta}_{\mathcal{S}}^*\|_{\min} \geq (\gamma_0+\gamma_1)\lambda$, and due to the fact that $p'_\lambda(t) = 0$ for all $t \geq \gamma_1\lambda$, the deterministic term $\|p'_\lambda(|\boldsymbol{\beta}_{\mathcal{S}}^*| - \gamma_0\lambda)\|_2$ vanishes, thus implying

$$\|\widetilde{\boldsymbol{\beta}}^{(\ell)} - \boldsymbol{\beta}^*\|_2 \leq \delta^{\ell-1}r^{\text{crude}} + 2(1-\delta)^{-1}\rho_l^{-1}\big\{\|\boldsymbol{w}_{\mathcal{S}}^*\|_2 + (s/n)^{1/2} + b_\tau^*\big\} \quad \text{(C.2)}$$

for all $\ell \geq 2$.

Next, we control the event $\mathcal{E}_1(r, l, \rho_l/2) \cap \{\|\boldsymbol{w}^*\|_\infty + n^{-1/2} \leq 0.5p'(\gamma_0)\lambda\}$ and the oracle error term $\|\boldsymbol{w}_{\mathcal{S}}^*\|_2$. Given $t \geq 0$, it follows from Proposition 2.4 with the above $l \asymp s^{1/2}$ and $\kappa = \rho_l/2$ that event $\mathcal{E}_1(r, l, \kappa)$ occurs with probability at least $1 - e^{-t}$ as long as $\tau \gtrsim \max(\sigma_2, r)$ and $n \gtrsim (\tau/r)^2(s\log d + t)$. We therefore take $r \asymp \tau$ throughout the proof. Turning to the gradient vector $\boldsymbol{w}^* \in \mathbb{R}^d$, applying Proposition 2.5 yields that with probability at least $1 - 2e^{-t}$,

$$\|\boldsymbol{w}^*\|_\infty \lesssim \sigma_2\sqrt{\frac{\log d + t}{n}} + \tau\frac{\log d + t}{n} \quad \text{and} \quad \|\boldsymbol{w}_{\mathcal{S}}^*\|_2 \lesssim \sigma_2\sqrt{\frac{s+t}{n}} + \tau\frac{s+t}{n}.$$

Based on the above analysis, we choose the regularization parameter $\lambda = C\sigma_2\sqrt{(\log d + t)/n}$ for a sufficiently large C. Under the scaling $n \gtrsim s\log d + t$, and if τ has magnitude of the order within the range of σ_2 to $\sigma_2\sqrt{n/(\log d + t)}$, it follows from (C.2) that with probability at least $1 - 3e^{-t}$,

$$\|\widetilde{\boldsymbol{\beta}}^{(\ell)} - \boldsymbol{\beta}^*\|_2 \lesssim \delta^{\ell-1}s^{1/2}\lambda + \frac{1}{1-\delta}\left(\sigma_2\sqrt{\frac{s+t}{n}} + \tau\frac{s+t}{n} + b_\tau^*\right) \quad \text{for all } \ell \geq 1.$$

This leads to the claimed bound by letting $\ell \gtrsim \lceil \log(\log d + t)/\log(1/\delta)\rceil$. $\quad\square$

C.3. Proof of Theorem 2.3

The proof is based primarily on Proposition 2.3, combined with complementary probabilistic analysis. For $t \geq 0$ and a prescribed $q \geq \max(s, \log d)$, set $\tau = \sigma_2 \sqrt{n/(q+t)}$. In order to apply the high-level result in Proposition 2.3, we need the following two technical lemmas to control the events in (2.19). The former controls the event $\mathcal{E}_2(r, l, \kappa)$ defined in (2.16) under proper sample size requirement, and the latter provides upper bounds on the ℓ_∞-error terms $\|\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*\|_\infty$ and $\|\boldsymbol{w}^{\mathrm{ora}}\|_\infty = \|\nabla \widehat{\mathcal{L}}_\tau(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}})\|_\infty$.

Lemma C.1. Under the conditions of the theorem, let $\tau, r, l > 0$ satisfy

$$\tau \geq \max(C_0 \sigma_2, C_1 r) \quad \text{and} \quad n \gtrsim (\tau/r)^2 (s + l^2 \log d + t), \qquad \text{(C.3)}$$

where C_0 is an absolute constant and C_1 depends only on $\sigma_{\boldsymbol{x}}$. Then, with probability at least $1 - e^{-t}$,

$$\langle \nabla \widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}_1) - \nabla \widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \rangle \geq \frac{\rho_l}{2} \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 \qquad \text{(C.4)}$$

holds uniformly over $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \in \mathcal{C}(r, l)$, where $\mathcal{C}(r, l)$ is defined in (2.16).

The next lemma provides statistical properties of the oracle estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{ora}}$ defined in (2.15) with $\tau = \sigma_2 \sqrt{n/(q+t)}$. Since the oracle $\widehat{\boldsymbol{\beta}}^{\mathrm{ora}}$ has access to the true active set $\mathcal{S}$, it is essentially an unpenalized Huber estimator based on $\{(y_i, \boldsymbol{x}_{i, \mathcal{S}})\}_{i=1}^n$.

Lemma C.2. Under the sample size scaling $n \gtrsim q + t$, the following bounds

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*\|_\Sigma &\lesssim \sigma_2 \sqrt{\frac{s+t}{n}}, \\
\|\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*\|_\infty &\lesssim \sigma_2 \sqrt{\frac{\log s + t}{n}} + \sigma_2 \left(\frac{q+t}{n}\right)^{(1+\eta)/2},
\end{aligned} \qquad \text{(C.5)}$$

and

$$\|\nabla \widehat{\mathcal{L}}_\tau(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}})\|_\infty \lesssim \sigma_2 \sqrt{\frac{\log s + t}{n}} + \sigma_2 \left(\frac{q+t}{n}\right)^{(1+\eta)/2} \qquad \text{(C.6)}$$

hold with probability at least $1 - 7e^{-t}$.

Compared to Propositions 2.4 and 2.5, the proofs of Lemmas C.1 and C.2, which are placed in the following two subsections, require a more delicate analysis of the local behavior of the gradient process $\{\nabla \widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}), \boldsymbol{\beta} \in \mathbb{R}^d\}$ around both the underlying vector $\boldsymbol{\beta}^*$ and the oracle estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{ora}}$, with the latter being random itself.

With the above preparations, we are ready to prove the result. In Proposition 2.3, we set $\kappa = \rho_l/2$, $\delta = 2.5/(\rho_l\gamma_0) \in (0,1)$ and $l = \{2 + \frac{2}{p'(\gamma_0)}\}(c_0^2 + 1)^{1/2}s^{1/2}$, where c_0 is determined by (2.18). Taking $r = \tau/C_1$ in Lemma C.1, we obtain that event $\mathcal{E}_2(r,l,\kappa)$ happens with probability at least $1 - e^{-t}$ as long as $n \gtrsim \sigma_2$ and $n \gtrsim s\log d + t$. Next, let $\lambda = C\sigma_2\sqrt{(\log d + t)/n}$ for a sufficient large constant C. Then, it follows from Lemma C.2 that the event $\{\|\boldsymbol{w}^{\mathrm{ora}}\|_\infty \leq 0.5p'(\gamma_0)\lambda\} \cap \{\|\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*\|_\infty \leq \lambda/(5\delta\kappa) = \gamma_0\lambda/6.25\} \cap \{\|\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*\|_\Sigma \leq r\}$ occurs with probability at least $1 - 7e^{-t}$ as long as $n \gtrsim (q + t)^{1+1/\eta}(\log d)^{-1/\eta}$. Finally, the strong oracle property is a direct consequence of Proposition 2.3. □

C.3.1. *Proof of Lemma C.1*

By the convexity of the loss function, for any $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^d$,

$$D(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) := \langle \nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}_1) - \nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \rangle$$

$$= \frac{1}{n}\sum_{i=1}^n \{\ell'_\tau(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_2) - \ell'_\tau(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1)\}\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)$$

$$\geq \frac{1}{n}\sum_{i=1}^n \{\ell'_\tau(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_2) - \ell'_\tau(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1)\}\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)I_{\mathcal{E}_i}, \quad \text{(C.7)}$$

where $I_{\mathcal{E}_i}$ is the indicator function of the event $\mathcal{E}_i := \{|\varepsilon_i| \leq \tau/4\} \cap \{|\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_2 - \boldsymbol{\beta}^*)| \leq \tau/4\} \cap \{|\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)| \leq \tau\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2/(2r)\}$ on which $|y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_2| \leq |\varepsilon_i| + |\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_2 - \boldsymbol{\beta}^*)| \leq \tau/2$ and $|y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1| \leq |\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)| + |\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_2 - \boldsymbol{\beta}^*)| + |\varepsilon_i| \leq \tau$ for all $\boldsymbol{\beta}_1 \in \boldsymbol{\beta}_2 + \mathbb{B}(r)$. Similarly to the proof of Proposition 2 in [30], for any $R > 0$, define Lipschitz continuous functions

$$\varphi_R(u) = u^2 I(|u| \leq R/2) + (|u| - R)^2 I(R/2 < |u| \leq R),$$
$$\text{and} \quad \phi_R(u) = I(|u| \leq R/2) + \{2 - (2u/R)\operatorname{sign}(u)\}I(R/2 < |u| \leq R),$$

which are smoothed versions of $u \mapsto u^2 I(|u| \leq R)$ and $u \mapsto I(|u| \leq R)$, respectively. Moreover, $\varphi_R(u) \leq u^2 I(|u| \leq R)$ and $\phi_R(u) \leq I(|u| \leq R)$. By (C.7) and the fact that $\ell''_\tau(u) = 1$ for $|u| \leq \tau$,

$$D(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$$
$$\geq \frac{1}{n}\sum_{i=1}^n \{\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\}^2 I_{|\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_2)|/\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|_2 \leq \frac{\tau}{2r}} I_{|\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_2-\boldsymbol{\beta}^*)| \leq \frac{\tau}{4}} I_{|\varepsilon_i| \leq \frac{\tau}{4}}$$

$$\geq D_0(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) := \frac{1}{n}\sum_{i=1}^n \varphi_{\frac{\tau}{2r}\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|_2}(\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2))\phi_{\frac{\tau}{4}}(\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_2 - \boldsymbol{\beta}^*))I_{|\varepsilon_i| \leq \frac{\tau}{4}}$$

$$= \mathbb{E}D_0(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) + D_0(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) - \mathbb{E}D_0(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2). \quad \text{(C.8)}$$

In what follows, we deal with $\mathbb{E}D_0(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and $D_0(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) - \mathbb{E}D_0(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, separately.

Noting that $\varphi_R(u) \geq u^2 I(|u| \leq R/2)$ and $\phi_R(u) \geq I(|u| \leq R/2)$, we have

$$\mathbb{E}D_0(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \geq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\varphi_{\frac{\tau}{2r}\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|_2}(\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2))I_{|\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_2-\boldsymbol{\beta}^*)| \leq \frac{\tau}{8}}I_{|\varepsilon_i| \leq \frac{\tau}{4}}$$

$$\geq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\{(\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2))\}^2 I_{|\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_2)| \leq \frac{\tau}{4r}\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|_2}$$

$$- \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\{(\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2))\}^2 I_{|\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_2-\boldsymbol{\beta}^*)| > \frac{\tau}{8}} - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\{\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\}^2 I_{|\varepsilon_i| > \frac{\tau}{4}}.$$
$$(C.9)$$

Write $\boldsymbol{\delta} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2$ for $\boldsymbol{\beta}_1 \in \boldsymbol{\beta}_2 + \mathbb{B}(r)$ and $\boldsymbol{\beta}_2 \in \boldsymbol{\beta}^* + \mathbb{B}_\Sigma(r)$. By Hölder's inequality and (2.21),

$$\mathbb{E}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta})^2 I_{|\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta}| > \frac{\tau}{4r}\|\boldsymbol{\delta}\|_2} \leq 2\rho_{\boldsymbol{x}}^2 e^{-(\tau/8\sigma_{\boldsymbol{x}} r)^2}\|\boldsymbol{\delta}\|_\Sigma^2,$$

$$\mathbb{E}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta})^2 I_{|\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_2-\boldsymbol{\beta}^*)| > \frac{\tau}{8}} \leq (8/\tau)^2 \mathbb{E}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta})^2 \{\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_2 - \boldsymbol{\beta}^*)\}^2 \leq (8\rho_{\boldsymbol{x}}^2 r/\tau)^2 \|\boldsymbol{\delta}\|_\Sigma^2$$

and $\mathbb{E}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta})^2 I_{|\varepsilon_i| > \frac{\tau}{4}} \leq (4\sigma_2/\tau)^2 \|\boldsymbol{\delta}\|_\Sigma^2$. Substituting these into (C.9) yields

$$\mathbb{E}D_0(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \geq \left\{1 - (4\sigma_2/\tau)^2 - 2\rho_{\boldsymbol{x}}^2 e^{-(\tau/8\sigma_{\boldsymbol{x}} r)^2} - (8\rho_{\boldsymbol{x}}^2 r/\tau)^2\right\}\|\boldsymbol{\delta}\|_\Sigma^2.$$

Let $\tau \geq 8\max(2\sigma_2, \{\log(16\rho_{\boldsymbol{x}}^2)\}^{1/2}\sigma_{\boldsymbol{x}} r, 4\rho_{\boldsymbol{x}}^2 r)$, so that

$$(4\sigma_2/\tau)^2 \leq 1/16, \quad 2\rho_{\boldsymbol{x}}^2 e^{-(\tau/8\sigma_{\boldsymbol{x}} r)^2} \leq 1/8 \quad \text{and} \quad (8\rho_{\boldsymbol{x}}^2 r/\tau)^2 \leq 1/16.$$

It thus follows that

$$\mathbb{E}D_0(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \geq \frac{3}{4}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_\Sigma^2 \geq \frac{3}{4}\rho_l\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 \qquad (C.10)$$

uniformly over $\boldsymbol{\beta}_1 \in \boldsymbol{\beta}_2 + \mathbb{B}(r)$ and $\boldsymbol{\beta}_2 \in \boldsymbol{\beta}^* + \mathbb{B}_\Sigma(r)$.

Next, we will establish a high probability bound for the supremum

$$\Delta(r, l) := \sup_{(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \in \mathcal{C}(r,l)} \frac{|D_0(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) - \mathbb{E}D_0(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)|}{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2}.$$

Note that $\varphi_{cR}(cu) = c^2 \varphi_R(u)$ for any $c > 0$ and $u \in \mathbb{R}$. For each pair $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, we write $\boldsymbol{\delta} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2$ and define

$$f_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2}(\boldsymbol{x}_i, \varepsilon_i) = \varphi_{\frac{\tau}{2r}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2) \cdot \phi_{\frac{\tau}{4}}(\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_2 - \boldsymbol{\beta}^*)) \cdot I_{|\varepsilon_i| \leq \frac{\tau}{4}},$$

so that $\Delta(r, l) = \sup_{(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \in \mathcal{C}(r,l)} |(1/n)\sum_{i=1}^{n} f_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2}(\boldsymbol{x}_i, \varepsilon_i) - \mathbb{E}f_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2}(\boldsymbol{x}_i, \varepsilon_i)|$. Since $0 \leq \varphi_R(u) \leq \min\{(R/2)^2, u^2\}$ and $0 \leq \phi_R(u) \leq 1$ for all $u \in \mathbb{R}$, we have

$$0 \leq f_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2}(\boldsymbol{x}_i, \varepsilon_i) \leq (\tau/4r)^2 \quad \text{and} \quad \mathbb{E}f_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2}^2(\boldsymbol{x}_i, \varepsilon_i) \leq \rho_u^2 \rho_{\boldsymbol{x}}^4.$$

By Bousquet's version of Talagrand's inequality [7] and (B.12), for any $t \geq 0$,

$$\Delta(r,l) \leq 1.25\mathbb{E}\Delta(r,l) + \rho_u \rho_{\boldsymbol{x}}^2 \sqrt{2t/n} + (\tau/r)^2 t/(3n) \qquad \text{(C.11)}$$

holds with probability at least $1 - e^{-t}$.

It suffices to bound the expected value $\mathbb{E}\Delta(r,l)$. Applying the symmetrization inequality for empirical processes and the connection between Gaussian complexity and Rademacher complexity (see, e.g. Lemma 4.5 in [29]), we obtain that

$$\mathbb{E}\Delta(r,l) \leq 2 \cdot \sqrt{\frac{\pi}{2}} \cdot \mathbb{E}\left\{ \sup_{(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) \in \mathcal{C}(r,l)} |\mathbb{G}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}| \right\}, \qquad \text{(C.12)}$$

where

$$\mathbb{G}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2} := \frac{1}{n} \sum_{i=1}^{n} g_i \varphi_{\frac{\tau}{2r}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2) \phi_{\frac{\tau}{4}}(\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_2 - \boldsymbol{\beta}^*)) I_{|\varepsilon_i| \leq \frac{\tau}{4}}$$

with $\boldsymbol{\delta} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2$ and g_i's are independent standard normal random variables that are independent of the observations. In particular, $\mathbb{G}_{\boldsymbol{\beta}^*,\boldsymbol{\beta}^*}$ is defined as zero. Let $\mathbb{E}^*$ be the conditional expectation given $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$. By symmetry,

$$\mathbb{E}^*\left\{ \sup_{(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) \in \mathcal{C}(r,l)} |\mathbb{G}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}| \right\} \leq \underbrace{\mathbb{E}^*|\mathbb{G}_{\boldsymbol{\beta}^*,\boldsymbol{\beta}^*}|}_{=0} + 2\mathbb{E}^*\left\{ \sup_{(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) \in \mathcal{C}(r,l)} \mathbb{G}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2} \right\}. \quad \text{(C.13)}$$

Next, we apply the Gaussian comparison theorem to bound the remaining term $\mathbb{E}^*\{\sup_{(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) \in \mathcal{C}(r,l)} \mathbb{G}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}\}$, from which a bound on $\mathbb{E}\{\sup_{(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) \in \mathcal{C}(r,l)} \mathbb{G}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}\}$ follows immediately. For another pair $(\boldsymbol{\beta}_1',\boldsymbol{\beta}_2') \in \mathcal{C}(r,l)$, write $\boldsymbol{\delta}' = \boldsymbol{\beta}_1' - \boldsymbol{\beta}_2'$, and note that

$$\mathbb{G}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2} - \mathbb{G}_{\boldsymbol{\beta}_1',\boldsymbol{\beta}_2'}$$
$$= \mathbb{G}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2} - \mathbb{G}_{\boldsymbol{\beta}_2'+\boldsymbol{\delta},\boldsymbol{\beta}_2'} + \mathbb{G}_{\boldsymbol{\beta}_2'+\boldsymbol{\delta},\boldsymbol{\beta}_2'} - \mathbb{G}_{\boldsymbol{\beta}_1',\boldsymbol{\beta}_2'}$$
$$= \frac{1}{n} \sum_{i=1}^{n} g_i \varphi_{\frac{\tau}{2r}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2)\{\phi_{\frac{\tau}{4}}(\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_2 - \boldsymbol{\beta}^*)) - \phi_{\frac{\tau}{4}}(\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_2' - \boldsymbol{\beta}^*))\} I_{|\varepsilon_i| \leq \frac{\tau}{4}}$$
$$+ \frac{1}{n} \sum_{i=1}^{n} g_i \phi_{\frac{\tau}{4}}(\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_2' - \boldsymbol{\beta}^*))\{\varphi_{\frac{\tau}{2r}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2) - \varphi_{\frac{\tau}{2r}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta}'/\|\boldsymbol{\delta}'\|_2)\} I_{|\varepsilon_i| \leq \frac{\tau}{4}}.$$

By the Lipschitz properties of ϕ_R and φ_R, i.e., $|\phi_R(u) - \phi_R(v)| \leq \frac{2}{R}|u - v|$ and $|\varphi_R(u) - \varphi_R(v)| \leq R|u - v|$, and recall that $\varphi_R(u) \leq (R/2)^2$, we have

$$\mathbb{E}^*(\mathbb{G}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2} - \mathbb{G}_{\boldsymbol{\beta}_2'+\boldsymbol{\delta},\boldsymbol{\beta}_2'})^2 \leq \frac{1}{n^2} \sum_{i=1}^{n} \left(\frac{\tau}{4r}\right)^4 \left(\frac{8}{\tau}\right)^2 \{\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_2')\}^2$$

$$= \left(\frac{\tau}{2r^2}\right)^2 \frac{1}{n^2} \sum_{i=1}^{n} \{\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_2')\}^2 \qquad \text{(C.14)}$$

and

$$\mathbb{E}^*(\mathbb{G}_{\boldsymbol{\beta}_2'+\boldsymbol{\delta},\boldsymbol{\beta}_2'} - \mathbb{G}_{\boldsymbol{\beta}_1',\boldsymbol{\beta}_2'})^2$$

$$\leq \frac{1}{n^2} \sum_{i=1}^n \big\{ \varphi_{\frac{\tau}{2r}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2) - \varphi_{\frac{\tau}{2r}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta}'/\|\boldsymbol{\delta}'\|_2) \big\}^2$$

$$\leq \left(\frac{\tau}{2r}\right)^2 \frac{1}{n^2} \sum_{i=1}^n \big(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2 - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta}'/\|\boldsymbol{\delta}'\|_2\big)^2. \qquad (\mathrm{C.15})$$

Motivated by (C.14), (C.15) and the inequality that

$$\mathbb{E}^*(\mathbb{G}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2} - \mathbb{G}_{\boldsymbol{\beta}_1',\boldsymbol{\beta}_2'})^2 \leq 2\mathbb{E}^*(\mathbb{G}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2} - \mathbb{G}_{\boldsymbol{\beta}_2'+\boldsymbol{\delta},\boldsymbol{\beta}_2'})^2 + 2\mathbb{E}^*(\mathbb{G}_{\boldsymbol{\beta}_2'+\boldsymbol{\delta},\boldsymbol{\beta}_2'} - \mathbb{G}_{\boldsymbol{\beta}_1',\boldsymbol{\beta}_2'})^2,$$

we define another (conditional) Gaussian process $\{\mathbb{Z}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}, (\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) \in \mathcal{C}(r,l)\}$ as

$$\mathbb{Z}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2} = \frac{\sqrt{2}\tau}{2r^2} \cdot \frac{1}{n} \sum_{i=1}^n g_i' \boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_2 - \boldsymbol{\beta}^*) + \frac{\sqrt{2}\tau}{2r} \cdot \frac{1}{n} \sum_{i=1}^n g_i'' \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2$$

$$= \frac{\sqrt{2}\tau}{2r^2} \cdot \frac{1}{n} \sum_{i=1}^n g_i' \boldsymbol{x}_{i,\mathcal{S}}^{\mathrm{T}}(\boldsymbol{\beta}_2 - \boldsymbol{\beta}^*)_{\mathcal{S}} + \frac{\sqrt{2}\tau}{2r} \cdot \frac{1}{n} \sum_{i=1}^n g_i'' \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2,$$

where $g_1', g_1'', \ldots, g_n', g_n''$ are independent standard normal random variables that are independent of all the other variables. We have established that $\mathbb{E}^*(\mathbb{G}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2} - \mathbb{G}_{\boldsymbol{\beta}_1',\boldsymbol{\beta}_2'})^2 \leq \mathbb{E}^*(\mathbb{Z}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2} - \mathbb{Z}_{\boldsymbol{\beta}_1',\boldsymbol{\beta}_2'})^2$. Then, applying Sudakov-Fernique's Gaussian comparison inequality (see, e.g. Theorem 7.2.11 in [47]) yields

$$\mathbb{E}^* \left\{ \sup_{(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)\in\mathcal{C}(r,l)} \mathbb{G}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2} \right\} \leq \mathbb{E}^* \left\{ \sup_{(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)\in\mathcal{C}(r,l)} \mathbb{Z}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2} \right\}, \qquad (\mathrm{C.16})$$

which remains valid if $\mathbb{E}^*$ is replaced by $\mathbb{E}$. For the supremum of $\mathbb{Z}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}$, it is easy to see that

$$\mathbb{E} \left\{ \sup_{(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)\in\mathcal{C}(r,l)} \mathbb{Z}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2} \right\}$$

$$\leq \frac{\sqrt{2}\tau}{2r} \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n g_i' (\Sigma_{\mathcal{SS}})^{-1/2} \boldsymbol{x}_{i,\mathcal{S}} \right\|_2 + \frac{\sqrt{2}\tau l}{2r} \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n g_i'' \boldsymbol{x}_i \right\|_\infty$$

$$\leq \frac{\sqrt{2}\tau}{2r} \sqrt{\frac{s}{n}} + \frac{\sqrt{2}\tau l}{2r} \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n g_i'' \boldsymbol{x}_i \right\|_\infty. \qquad (\mathrm{C.17})$$

Together, (C.12), (C.13), (C.16) and (C.17) deliver the bound

$$\mathbb{E}\Delta(r,l) \leq 2\sqrt{\pi} \frac{\tau}{r} \left\{ \sqrt{\frac{s}{n}} + l\, \mathbb{E} \left(\max_{1\leq j\leq d} \left| \frac{1}{n} \sum_{i=1}^n g_i x_{ij} \right| \right) \right\}. \qquad (\mathrm{C.18})$$

Finally we bound the maximum under expectation on the right-hand side of (C.18). Write $S_j = \sum_{i=1}^{n} g_i x_{ij}$ for $j = 1, \ldots, d$. Under Condition 2.3, for each $1 \leq j \leq d$ and $m \geq 2$ we have

$$\mathbb{E}|x_j|^m = \sigma_{\boldsymbol{x}}^m m \int_0^\infty t^{m-1} \mathbb{P}(|x_j| \geq \sigma_{\boldsymbol{x}} t) \, dt$$
$$\leq 2\sigma_{\boldsymbol{x}}^m m \int_0^\infty t^{m-1} e^{-t^2/2} \, dt = 2^{m/2} \sigma_{\boldsymbol{x}}^m m \Gamma(m/2).$$

Let $g \sim \mathsf{N}(0,1)$ be independent of $\boldsymbol{x}$. Using the Legendre duplication formula, i.e., $\Gamma(s)\Gamma(s + 1/2) = 2^{1-2s}\sqrt{\pi}\,\Gamma(2s)$, and some algebra, we get

$$\mathbb{E}|gx_j|^m \leq 2^{m/2}\frac{\Gamma(\frac{m+1}{2})}{\sqrt{\pi}} \cdot 2^{m/2}\sigma_{\boldsymbol{x}}^m m \Gamma(m/2) = 2\sigma_{\boldsymbol{x}}^m m! = \frac{m!}{2} \underbrace{4\sigma_{\boldsymbol{x}}^2}_{\geq \mathbb{E}(g_j x_j)^2} \sigma_{\boldsymbol{x}}^{m-2}.$$

Hence, using Bernstein's inequality and the symmetry of normal distribution yields

$$\log \mathbb{E}e^{\lambda S_j} = \log \mathbb{E}e^{-\lambda S_j} \leq \frac{4\sigma_{\boldsymbol{x}}^2 n \lambda^2}{2(1 - \sigma_{\boldsymbol{x}}\lambda)}$$

for all $\lambda \in (0, 1/\sigma_{\boldsymbol{x}})$. Combined with Theorem 2.5 in [6], this implies

$$\mathbb{E}\left(\max_{1 \leq j \leq d} \left| \frac{1}{n} \sum_{i=1}^{n} g_i x_{ij} \right| \right) = \mathbb{E} \max_{1 \leq j \leq d} |S_j/n|$$
$$\leq \sigma_{\boldsymbol{x}} \left\{ 2\sqrt{\frac{2\log(2d)}{n}} + \frac{\log(2d)}{n} \right\}. \tag{C.19}$$

Combining (C.18), (C.19) with the concentration inequality (C.11), we determine that with probability at least $1 - e^{-t}$, $\Delta(r, l) \leq \rho_l/4$ as long as $n \gtrsim (\tau/r)^2(s + l^2 \log d + t)$. This, together with (C.8) and (C.10), proves the claim (C.4). $\qquad\square$

C.3.2. *Proof of Lemma C.2*

To begin with, consider the decomposition

$$\nabla\widehat{\mathcal{L}}_\tau(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}}) = \boldsymbol{w}(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}}) - \boldsymbol{w}(\boldsymbol{\beta}^*) + \nabla\mathcal{L}_\tau(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}}) + \boldsymbol{w}^*,$$

where $\boldsymbol{w}(\boldsymbol{\beta}) = \nabla\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta}) - \nabla\mathcal{L}_\tau(\boldsymbol{\beta})$, $\mathcal{L}_\tau(\boldsymbol{\beta}) = \mathbb{E}\widehat{\mathcal{L}}_\tau(\boldsymbol{\beta})$ and $\boldsymbol{w}^* = \boldsymbol{w}(\boldsymbol{\beta}^*)$. In the following, we control the ℓ_∞-norms of the three terms, $\boldsymbol{w}(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}}) - \boldsymbol{w}(\boldsymbol{\beta}^*)$, $\nabla\mathcal{L}_\tau(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}})$ and $\boldsymbol{w}^*$, separately. Throughout the proof, we take $\tau = \sigma_2\sqrt{n/(q+t)}$ for some $q \geq \max(s, \log d)$ and $t \geq 0$.

Applying Proposition 2.5 to the centered gradient $\boldsymbol{w}^* = ((\boldsymbol{w}_{\mathcal{S}}^*)^{\mathrm{T}}, (\boldsymbol{w}_{\mathcal{S}^c}^*))^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^d$ with slight modifications, we obtain that with probability at least $1 - 2e^{-t}$,

$$\|\boldsymbol{w}_{\mathcal{S}}^*\|_\infty \lesssim \sigma_2 \sqrt{\frac{\log s + t}{n}} \quad \text{and} \quad \|\boldsymbol{w}_{\mathcal{S}^c}^*\|_\infty \lesssim \sigma_2 \sqrt{\frac{\log(d-s) + t}{n}}, \qquad (\text{C.20})$$

thus implying $\|\boldsymbol{w}^*\|_\infty \lesssim \sigma_2 \sqrt{(\log d + t)/n}$ with the same probability.

Recall that $\widehat{\boldsymbol{\beta}}^{\mathrm{ora}}$ and $\boldsymbol{\beta}^*$ have the same support $\mathcal{S} \subseteq [d]$. Define the oracle local neighborhood $\Theta^*(r) = \{\boldsymbol{\beta} \in \boldsymbol{\beta}^* + \mathbb{B}_\Sigma(r) : \boldsymbol{\beta}_{\mathcal{S}^c} = \boldsymbol{0}\}$. Then, conditioned on the event $\{\|\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*\|_\Sigma \leq r\}$,

$$\|\boldsymbol{w}(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}}) - \boldsymbol{w}(\boldsymbol{\beta}^*)\|_\infty \leq \sup_{\boldsymbol{\beta} \in \Theta^*(r)} \|\boldsymbol{w}(\boldsymbol{\beta}) - \boldsymbol{w}(\boldsymbol{\beta}^*)\|_\infty. \qquad (\text{C.21})$$

We thus focus on the supremum on the right-hand side of (C.21). For every s-sparse vector $\boldsymbol{\beta} \in \Theta^*(r)$, we write $\boldsymbol{\delta} = (\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\mathcal{S}} \in \mathbb{R}^s$. For $j = 1, \ldots, d$, let $\boldsymbol{e}_j \in \mathbb{R}^d$ be the coordinate vector that has 1 on its j-th coordinate and 0 elsewhere, and define $\Delta_j^0(\boldsymbol{\delta}) = \langle \boldsymbol{w}(\boldsymbol{\beta}) - \boldsymbol{w}(\boldsymbol{\beta}^*), \boldsymbol{e}_j \rangle = (1/n) \sum_{i=1}^n (\eta_{ij} - \mathbb{E}\eta_{ij})$ for $\boldsymbol{\delta} \in \mathbb{R}^s$, where $\eta_{ij} = x_{ij}\{\ell'_\tau(\varepsilon_i - \boldsymbol{x}_{i,\mathcal{S}}^{\mathrm{T}}\boldsymbol{\delta}) - \ell'_\tau(\varepsilon_i)\}$. Consequently, we have

$$\sup_{\boldsymbol{\beta} \in \Theta^*(r)} \|\boldsymbol{w}(\boldsymbol{\beta}) - \boldsymbol{w}(\boldsymbol{\beta}^*)\|_\infty \leq \max_{1 \leq j \leq d} \sup_{\|\boldsymbol{\delta}\|_{\mathrm{S}} \leq r} \Delta_j^0(\boldsymbol{\delta}) \bigvee \max_{1 \leq j \leq d} \sup_{\|\boldsymbol{\delta}\|_{\mathrm{S}} \leq r} -\Delta_j^0(\boldsymbol{\delta}),$$
$$(\text{C.22})$$

where $\mathrm{S} = \Sigma_{\mathcal{S}\mathcal{S}} \in \mathbb{R}^{s \times s}$. In order to bound the local fluctuation $\sup_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_{\mathrm{S}} \leq r} \Delta_j^0(\boldsymbol{\delta})$, we need to control the moment generating function of $\Delta_j^0(\boldsymbol{\delta})$ for each $\boldsymbol{\delta} \in \mathbb{R}^s$. By the Lipschitz continuity of $\ell'_\tau(\cdot)$, $|\mathbb{E}(\eta_{ij})| \leq \mathbb{E}|x_{ij}\boldsymbol{x}_{i,\mathcal{S}}^{\mathrm{T}}\boldsymbol{\delta}| \leq \sigma_{jj}^{1/2}\|\boldsymbol{\delta}\|_{\mathrm{S}}$, $\mathbb{E}(\eta_{ij}^2|\boldsymbol{x}_i) \leq x_{ij}^2(\boldsymbol{x}_{i,\mathcal{S}}^{\mathrm{T}}\boldsymbol{\delta})^2$ and

$$\mathbb{E}\{(\eta_{ij} - \mathbb{E}\eta_{ij})^2|\boldsymbol{x}_i\} \leq 2\mathbb{E}(\eta_{ij}^2|\boldsymbol{x}_i) + 2(\mathbb{E}\eta_{ij})^2 \leq 2x_{ij}^2(\boldsymbol{x}_{i,\mathcal{S}}^{\mathrm{T}}\boldsymbol{\delta})^2 + 2\sigma_{jj}\|\boldsymbol{\delta}\|_{\mathrm{S}}^2.$$

The above moment inequalities, combined with the elementary inequality $|e^u - 1 - u| \leq (u^2/2)e^{u \vee 0}$, imply that for any $\lambda \in \mathbb{R}$ and $\lambda^* = \lambda/(\sigma_{jj}^{1/2}\|\boldsymbol{\delta}\|_{\mathrm{S}})$,

$$\mathbb{E}e^{\lambda\sqrt{n}\Delta_j^0(\boldsymbol{\delta})/(\sigma_{jj}^{1/2}\|\boldsymbol{\delta}\|_{\mathrm{S}})} = \prod_{i=1}^n \mathbb{E}e^{\frac{\lambda^*}{\sqrt{n}}(\eta_{ij} - \mathbb{E}\eta_{ij})}$$

$$\leq \prod_{i=1}^n \mathbb{E}\left\{1 + \frac{\lambda^{*2}}{2n}(\eta_{ij} - \mathbb{E}\eta_{ij})^2 e^{\frac{|\lambda^*|}{\sqrt{n}}|\eta_{ij} - \mathbb{E}\eta_{ij}|}\right\}$$

$$\leq \prod_{i=1}^n \left\{1 + \frac{\lambda^{*2}e^{|\lambda|/\sqrt{n}}}{2n}\mathbb{E}(\eta_{ij} - \mathbb{E}\eta_{ij})^2 e^{\frac{|\lambda^*|}{\sqrt{n}}|x_{ij}\boldsymbol{x}_{i,\mathcal{S}}^{\mathrm{T}}\boldsymbol{\delta}|}\right\}$$

$$\leq \prod_{i=1}^n \left\{1 + \frac{\lambda^2}{n}e^{|\lambda|/\sqrt{n}}\mathbb{E}e^{\frac{|\lambda^*|}{\sqrt{n}}|x_{ij}\boldsymbol{x}_{i,\mathcal{S}}^{\mathrm{T}}\boldsymbol{\delta}|} + \frac{\lambda^{*2}}{n}e^{|\lambda|/\sqrt{n}}\mathbb{E}x_{ij}^2(\boldsymbol{x}_{i,\mathcal{S}}^{\mathrm{T}}\boldsymbol{\delta})^2 e^{\frac{|\lambda^*|}{\sqrt{n}}|x_{ij}\boldsymbol{x}_{i,\mathcal{S}}^{\mathrm{T}}\boldsymbol{\delta}|}\right\}.$$
$$(\text{C.23})$$

Applying Hölder's inequality to the exponential moments on the right-hand side of (C.23), we have

$$
\mathbb{E}x_{ij}^2(\boldsymbol{x}_{i,\mathcal{S}}^{\mathrm{T}}\boldsymbol{\delta})^2 e^{\frac{|\lambda^*|}{\sqrt{n}}|x_{ij}\boldsymbol{x}_{i,\mathcal{S}}^{\mathrm{T}}\boldsymbol{\delta}|}
$$

$$
\leq \sigma_{jj}\|\boldsymbol{\delta}\|_{\mathrm{S}}^2 \cdot \left\{ \mathbb{E}(x_{ij}/\sigma_{jj}^{1/2})^4 e^{\frac{|\lambda|}{\sqrt{n}}x_{ij}^2/\sigma_{jj}} \right\}^{1/2}
$$

$$
\times \left(\mathbb{E}(\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{x}_{i,\mathcal{S}}/\|\boldsymbol{\delta}\|_{\mathrm{S}})^4 e^{\frac{|\lambda|}{\sqrt{n}}(\boldsymbol{x}_{i,\mathcal{S}}^{\mathrm{T}}\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_{\mathrm{S}})^2} \right)^{1/2}
$$

and

$$
\mathbb{E}e^{\frac{|\lambda^*|}{\sqrt{n}}|x_{ij}\boldsymbol{x}_{i,\mathcal{S}}^{\mathrm{T}}\boldsymbol{\delta}|} \leq \left(\mathbb{E}e^{\frac{|\lambda|}{\sqrt{n}}x_{ij}^2/\sigma_{jj}} \right)^{1/2} \cdot \left(\mathbb{E}e^{\frac{|\lambda|}{\sqrt{n}}(\boldsymbol{x}_{i,\mathcal{S}}^{\mathrm{T}}\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_{\mathrm{S}})^2} \right)^{1/2}.
$$

Substituting these bounds into the earlier inequality (C.23), we find that for any $|\lambda| \leq \sqrt{n}/C_1$,

$$
\mathbb{E}e^{\lambda\sqrt{n}\Delta_j^0(\boldsymbol{\delta})/(\sigma_{jj}^{1/2}\|\boldsymbol{\delta}\|_{\mathrm{S}})} \leq e^{C_2^2\lambda^2/2},
$$

where $C_1, C_2 > 0$ depend only on υ_1 in Condition 2.3. A similar argument can be used to establish the same bound for each pair $(\boldsymbol{\delta}, \boldsymbol{\delta}')$, that is,

$$
\mathbb{E}e^{\lambda\sqrt{n}\{\Delta_j^0(\boldsymbol{\delta})-\Delta_j^0(\boldsymbol{\delta}')\}/(\sigma_{jj}^{1/2}\|\boldsymbol{\delta}-\boldsymbol{\delta}'\|_{\mathrm{S}})} \leq e^{C_2^2\lambda^2/2} \quad \text{for all } |\lambda| \leq \sqrt{n}/C_1.
$$

The above inequality certifies condition $(\mathcal{E}d)$ in [42] (see Section 2 in the supplement), so that Corollary 2.2 therein applies to the process $\{\Delta_j^0(\boldsymbol{\delta})\}_{\boldsymbol{\delta}\in\mathbb{R}^s:\|\boldsymbol{\delta}\|_{\mathrm{S}}\leq r}$: with probability at least $1 - e^{-x}$,

$$
\sup_{\boldsymbol{\beta}\in\Theta^*(r)} \langle \boldsymbol{w}(\boldsymbol{\beta}) - \boldsymbol{w}(\boldsymbol{\beta}^*), \boldsymbol{e}_j \rangle = \sup_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_{\mathrm{S}}\leq r} \Delta_j^0(\boldsymbol{\delta}) \lesssim r\sqrt{\frac{s+x}{n}}
$$

as long as $n \gtrsim s + x$. Combined with (C.22) and the union bound, we find that

$$
\sup_{\boldsymbol{\beta}\in\Theta^*(r)} \|\boldsymbol{w}(\boldsymbol{\beta}) - \boldsymbol{w}(\boldsymbol{\beta}^*)\|_\infty \lesssim r\sqrt{\frac{s+x}{n}}
$$

with probability at least $1 - 2de^{-x}$ provided $n \gtrsim s + x$. Taking $x = \log(2d) + t$, it follows from (C.21) that conditioned on $\{\|\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*\|_\Sigma \leq r\}$,

$$
\|\boldsymbol{w}(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}}) - \boldsymbol{w}(\boldsymbol{\beta}^*)\|_\infty \lesssim r\sqrt{\frac{s+\log d+t}{n}} \tag{C.24}
$$

holds with probability at least $1 - e^{-t}$ as long as $n \gtrsim s + \log d + t$.

Tuning to $\|\nabla\mathcal{L}_\tau(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}})\|_\infty$, again, we control this term conditioned on the same event above. Following the proof of Lemma A.1, it can be similarly shown that

$$
\|\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_\infty \lesssim (a_\eta\sigma_2)^{2+\eta}\tau^{-1-\eta}. \tag{C.25}
$$

For any $\boldsymbol{\beta} \in \Theta^*(r)$, write $\boldsymbol{\delta} = (\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\mathcal{S}} \in \mathbb{R}^s$, and note that

$$\nabla \mathcal{L}_\tau(\boldsymbol{\beta}) - \nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*) = \mathbb{E}\big\{\ell'_\tau(\varepsilon) - \ell'_\tau(\varepsilon - \boldsymbol{x}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{\delta})\big\}\boldsymbol{x}$$

$$= \mathbb{E}\int_{-\boldsymbol{x}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{\delta}}^0 \ell''_\tau(\varepsilon + u)\,\mathrm{d}u \cdot \boldsymbol{x}$$

$$= \mathbb{E}\ell''_\tau(\varepsilon)\boldsymbol{x}\boldsymbol{x}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{\delta} + \mathbb{E}\int_{-\boldsymbol{x}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{\delta}}^0 \big\{\ell''_\tau(\varepsilon + u) - \ell''_\tau(\varepsilon)\big\}\,\mathrm{d}u \cdot \boldsymbol{x}.$$

Let $\mathbb{E}_{\boldsymbol{x}}$ and $\mathbb{P}_{\boldsymbol{x}}$ be the conditional expectation and probability given $\boldsymbol{x}$, respectively. By the anti-concentration property (2.29) of the distribution of ε given $\boldsymbol{x}$, we see that for any $u \in \mathbb{R}$,

$$|\mathbb{E}_{\boldsymbol{x}}\{\ell''_\tau(\varepsilon + u) - \ell''_\tau(\varepsilon)\}| = |\mathbb{P}_{\boldsymbol{x}}(|\varepsilon + u| \leq \tau) - \mathbb{P}_{\boldsymbol{x}}(|\varepsilon| \leq \tau)| \leq a_0|u|.$$

Together, the last two displays imply

$$\|\nabla \mathcal{L}_\tau(\boldsymbol{\beta}) - \nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*) + \mathrm{H}_{\cdot\mathcal{S}}\boldsymbol{\delta}\|_\infty$$
$$\leq \frac{a_0}{2}\max_{1 \leq j \leq d}\mathbb{E}|x_j|(\boldsymbol{x}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{\delta})^2 \leq \frac{a_0}{2}\max_{1 \leq j \leq d}\sigma_{jj}^{1/2}\rho_{\boldsymbol{x}}^2\|\boldsymbol{\delta}\|_{\mathrm{S}}^2, \tag{C.26}$$

where $\mathrm{H}_{\cdot\mathcal{S}} := \mathbb{E}\{\ell''_\tau(\varepsilon)\boldsymbol{x}\boldsymbol{x}_{\mathcal{S}}^{\mathrm{T}}\} \in \mathbb{R}^{d \times s}$ is the submatrix of $\mathrm{H} = \nabla^2 \mathcal{L}_\tau(\boldsymbol{\beta}^*) = \mathbb{E}\{\ell''_\tau(\varepsilon)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$. For the linear term $\mathrm{H}_{\cdot\mathcal{S}}\boldsymbol{\delta}$, write $\Sigma_{\cdot\mathcal{S}} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}_{\mathcal{S}}^{\mathrm{T}})$ and note that

$$\|(\mathrm{H}_{\cdot\mathcal{S}} - \Sigma_{\cdot\mathcal{S}})\boldsymbol{\delta}\|_\infty \leq \max_{1 \leq j \leq d}\mathbb{E}\big\{\mathbb{P}_{\boldsymbol{x}}(|\varepsilon| \geq \tau) \cdot |x_j\boldsymbol{x}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{\delta}|\big\}$$

$$\leq \max_{1 \leq j \leq d}\sigma_{jj}^{1/2}\sigma_2^2\|\boldsymbol{\delta}\|_{\mathrm{S}}\tau^{-2}. \tag{C.27}$$

Together, (C.25), (C.26) and (C.27) imply that conditioned on $\{\|\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*\|_\Sigma \leq r\}$,

$$\|\nabla \mathcal{L}_\tau(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}}) + \Sigma_{\cdot\mathcal{S}}(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*)_{\mathcal{S}}\|_\infty \lesssim r^2 + \sigma_2^{2+\eta}\tau^{-1-\eta} + \sigma_2^2\tau^{-2}r. \tag{C.28}$$

Next we consider the oracle estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{ora}}$ with $\tau = \sigma_2\sqrt{n/(q+t)}$. Following an argument similar to that used to prove Theorem 2.1 in [11], it can be shown that with probability at least $1 - 3e^{-t}$,

$$\|\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*\|_\Sigma = \|(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*)_{\mathcal{S}}\|_{\mathrm{S}} \lesssim \sigma_2\sqrt{\frac{s+t}{n}} \tag{C.29}$$

and

$$\left\|\mathrm{S}^{1/2}(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*)_{\mathcal{S}} - \mathrm{S}^{-1/2}\frac{1}{n}\sum_{i=1}^n \ell'_\tau(\varepsilon_i)\boldsymbol{x}_{i,\mathcal{S}}\right\|_2 \lesssim \sigma_2\frac{s+t}{n}, \tag{C.30}$$

where $\mathrm{S} = \mathbb{E}(\boldsymbol{x}_{\mathcal{S}}\boldsymbol{x}_{\mathcal{S}}^{\mathrm{T}}) = \Sigma_{\mathcal{S}\mathcal{S}}$. Note that the linear term $(1/n)\sum_{i=1}^n \ell'_\tau(\varepsilon_i)\boldsymbol{x}_{i,\mathcal{S}}$ in the Bahadur representation bound (C.30) can be written as

$$\frac{1}{n}\sum_{i=1}^n \ell'_\tau(\varepsilon_i)\boldsymbol{x}_{i,\mathcal{S}} = -\boldsymbol{w}_{\mathcal{S}}^* - \nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)_{\mathcal{S}}.$$

It follows that

$$
\begin{aligned}
\|\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*\|_\infty &= \|(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*)_{\mathcal{S}}\|_\infty \\
&\leq \|(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*)_{\mathcal{S}} + \mathrm{S}^{-1}\boldsymbol{w}_{\mathcal{S}}^* + \mathrm{S}^{-1}\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)_{\mathcal{S}}\|_\infty + \|\mathrm{S}^{-1}\boldsymbol{w}_{\mathcal{S}}^* + \mathrm{S}^{-1}\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)_{\mathcal{S}}\|_\infty \\
&\leq \|(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*)_{\mathcal{S}} + \mathrm{S}^{-1}\boldsymbol{w}_{\mathcal{S}}^* + \mathrm{S}^{-1}\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)_{\mathcal{S}}\|_2 + \|\mathrm{S}^{-1}\boldsymbol{w}_{\mathcal{S}}^*\|_\infty + \|\mathrm{S}^{-1}\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)_{\mathcal{S}}\|_2 \\
&\leq \|\mathrm{S}^{-1}\|_2^{1/2}\big\{\|\mathrm{S}^{1/2}(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*)_{\mathcal{S}} + \mathrm{S}^{-1/2}\boldsymbol{w}_{\mathcal{S}}^* \\
&\qquad\qquad + \mathrm{S}^{-1/2}\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)_{\mathcal{S}}\|_2 + \|\mathrm{S}^{-1/2}\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)_{\mathcal{S}}\|_2\big\} + \|\mathrm{S}^{-1}\boldsymbol{w}_{\mathcal{S}}^*\|_\infty.
\end{aligned}
$$

Similarly to Lemma A.1, we obtain that $\|\mathrm{S}^{-1/2}\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)_{\mathcal{S}}\|_2 \leq (a_\eta\sigma_2)^{2+\eta}\tau^{-1-\eta}$. For $\|\mathrm{S}^{-1}\boldsymbol{w}_{\mathcal{S}}^*\|_\infty$, following the proof of Proposition 2.5, it can be similarly shown that with probability at least $1 - e^{-t}$,

$$
\|\mathrm{S}^{-1}\boldsymbol{w}_{\mathcal{S}}^*\|_\infty \lesssim \sigma_2\sqrt{\frac{\log(2s) + t}{n}}.
$$

Putting together the pieces, we conclude that the ℓ_2-error bound (C.29) and the ℓ_∞-error bound

$$
\|\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*\|_\infty \lesssim \sigma_2\sqrt{\frac{\log s + t}{n}} + \sigma_2\left(\frac{q + t}{n}\right)^{(1+\eta)/2}
$$

hold with probability $1 - 4e^{-t}$ as long as $n \gtrsim q + t$. Combined with (C.29), this proves (C.5).

Finally, it remains to deal with $\|\Sigma_{\cdot\mathcal{S}}(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*)_{\mathcal{S}}\|_\infty$. Under condition (2.28),

$$
\|\Sigma_{\cdot\mathcal{S}}\boldsymbol{\delta}\|_\infty \leq \max_{j\in\mathcal{S}^{\mathrm{c}}}\|\Sigma_{j\mathcal{S}}(\Sigma_{\mathcal{S}\mathcal{S}})^{-1}\|_1 \cdot \|\Sigma_{\mathcal{S}\mathcal{S}}\boldsymbol{\delta}\|_\infty \leq A_0 \cdot \|\mathrm{S}\boldsymbol{\delta}\|_\infty \quad \text{for any } \boldsymbol{\delta} \in \mathbb{R}^s.
$$

Using the previous bounds (C.20), (C.25) and (C.30), we obtain that

$$
\begin{aligned}
&\|\mathrm{S}(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*)_{\mathcal{S}}\|_\infty \\
&\leq \left\|\mathrm{S}(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*)_{\mathcal{S}} - \frac{1}{n}\sum_{i=1}^n \ell'_\tau(\varepsilon_i)\boldsymbol{x}_{i,\mathcal{S}}\right\|_\infty + \left\|\frac{1}{n}\sum_{i=1}^n \ell'_\tau(\varepsilon_i)\boldsymbol{x}_{i,\mathcal{S}}\right\|_\infty \\
&\leq \left\|\mathrm{S}(\widehat{\boldsymbol{\beta}}^{\mathrm{ora}} - \boldsymbol{\beta}^*)_{\mathcal{S}} - \frac{1}{n}\sum_{i=1}^n \ell'_\tau(\varepsilon_i)\boldsymbol{x}_{i,\mathcal{S}}\right\|_2 + \|\boldsymbol{w}_{\mathcal{S}}^*\|_\infty + \|\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)_{\mathcal{S}}\|_\infty \\
&\lesssim \sigma_2\frac{s + t}{n} + \sigma_2\sqrt{\frac{\log s + t}{n}} + a_\eta^{2+\eta}\sigma_2\left(\frac{q + t}{n}\right)^{(1+\eta)/2}.
\end{aligned}
$$

Combining this bound with (C.24), (C.28) and (C.29) yields the claim (C.6). $\qquad\square$

C.4. *Proof of Theorem 3.1*

For simplicity, we write $\boldsymbol{\beta}^{(k)} = \boldsymbol{\beta}^{(1,k)}$, $\phi^{(k)} = \phi^{(1,k)}$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(0)}$ throughout this section.

C.4.1. Technical lemmas

We first present three technical lemmas, which are the key ingredients to the proof. The first lemma provides an alternative to the stopping rule.

Lemma C.3. $\omega_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{(k)}) \leq L(1 + \gamma_u)\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}\|_2$.

Proof of Lemma C.3. For simplicity, we write $\mathcal{L}(\cdot) = \widehat{\mathcal{L}}_\tau(\cdot)$ as the loss function of interest. Since $\boldsymbol{\beta}^{(k)}$ is the exact solution at the k-th iteration when $\ell = 1$, the first-order optimality condition holds: there exists some $\boldsymbol{\xi}^{(k)} \in \partial\|\boldsymbol{\beta}^{(k)}\|_1$ such that

$$\nabla\mathcal{L}(\boldsymbol{\beta}^{(k-1)}) + \phi^{(k)}(\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}) + \boldsymbol{\lambda} \circ \boldsymbol{\xi}^{(k)} = \mathbf{0}_d.$$

For any $\boldsymbol{u} \in \mathbb{R}^d$ such that $\|\boldsymbol{u}\|_1 = 1$, we have

$$
\begin{aligned}
&\langle \nabla\mathcal{L}(\boldsymbol{\beta}^{(k)}) + \boldsymbol{\lambda} \circ \boldsymbol{\xi}^{(k)}, \boldsymbol{u} \rangle \\
&= \langle \nabla\mathcal{L}(\boldsymbol{\beta}^{(k)}), \boldsymbol{u} \rangle - \langle \nabla\mathcal{L}(\boldsymbol{\beta}^{(k-1)}) + \phi^{(k)}(\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}), \boldsymbol{u} \rangle \\
&= \langle \nabla\mathcal{L}(\boldsymbol{\beta}^{(k)}) - \nabla\mathcal{L}(\boldsymbol{\beta}^{(k-1)}), \boldsymbol{u} \rangle - \langle \phi^{(k)}(\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}), \boldsymbol{u} \rangle \\
&\leq \|\nabla\mathcal{L}(\boldsymbol{\beta}^{(k)}) - \nabla\mathcal{L}(\boldsymbol{\beta}^{(k-1)})\|_\infty + \phi^{(k)}\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}\|_\infty \\
&\leq (\phi^{(k)} + L)\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}\|_2,
\end{aligned}
$$

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

$$\omega_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{(k)}) \leq (\phi^{(k)} + L)\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}\|_2.$$

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

$$F(\widetilde{\boldsymbol{\beta}}^{(k)}; \phi', \boldsymbol{\beta}^{(k-1)}) < \mathcal{L}(\widetilde{\boldsymbol{\beta}}^{(k)}),$$

where $\widetilde{\boldsymbol{\beta}}^{(k)}$ is the new updated parameter vector under the quadratic coefficient ϕ'. On the other hand, it follows from the definition of F and the Lipschitz continuity of $\nabla\mathcal{L}$ that

$$
\begin{aligned}
&F(\widetilde{\boldsymbol{\beta}}^{(k)}; \phi', \boldsymbol{\beta}^{(k-1)}) + \lambda\|\boldsymbol{\beta}\|_1 \\
&= \mathcal{L}(\boldsymbol{\beta}^{(k-1)}) + \langle \nabla\mathcal{L}(\boldsymbol{\beta}^{(k-1)}), \widetilde{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k-1)} \rangle + \frac{\phi'}{2}\|\widetilde{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k-1)}\|_2^2 \\
&> \mathcal{L}(\boldsymbol{\beta}^{(k-1)}) + \langle \nabla\mathcal{L}(\boldsymbol{\beta}^{(k-1)}), \widetilde{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k-1)} \rangle + \frac{L}{2}\|\widetilde{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k-1)}\|_2^2 \\
&\geq \mathcal{L}(\widetilde{\boldsymbol{\beta}}^{(k)}).
\end{aligned}
$$

This leads to a contradiction, indicating that $\phi^{(k)} \leq L\gamma_u$. $\qquad\square$

The second lemma is a modified version of Lemma E.4 in [16]. We reproduce its proof here for completeness. Let $\Psi(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \mathcal{L}(\boldsymbol{\beta}) + \|\boldsymbol{\lambda} \circ \boldsymbol{\beta}\|_1$ with $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(0)}$.

Lemma C.4. For any $\boldsymbol{\beta} \in \mathbb{R}^d$, we have

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

Proof of Lemma C.4. Since $F(\boldsymbol{\beta}; \phi^{(k)}, \boldsymbol{\beta}^{(k-1)})$ majorizes $\mathcal{L}(\boldsymbol{\beta})$ at $\boldsymbol{\beta}^{(k)}$, we have

$$\Psi(\boldsymbol{\beta}, \boldsymbol{\lambda}) - \Psi(\boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}) \geq \Psi(\boldsymbol{\beta}, \boldsymbol{\lambda}) - \{F(\boldsymbol{\beta}^{(k)}; \phi^{(k)}, \boldsymbol{\beta}^{(k-1)}) + \|\boldsymbol{\lambda} \circ \boldsymbol{\beta}^{(k)}\|_1\}. \tag{C.31}$$

By the convexity of $\mathcal{L}(\cdot)$ and $\boldsymbol{\beta} \mapsto \|\boldsymbol{\lambda} \circ \boldsymbol{\beta}\|_1$,

$$\mathcal{L}(\boldsymbol{\beta}) \geq \mathcal{L}(\boldsymbol{\beta}^{(k-1)}) + \langle \nabla \mathcal{L}(\boldsymbol{\beta}^{(k-1)}), \boldsymbol{\beta} - \boldsymbol{\beta}^{(k-1)} \rangle \quad \text{and}$$
$$\|\boldsymbol{\lambda} \circ \boldsymbol{\beta}\|_1 \geq \|\boldsymbol{\lambda} \circ \boldsymbol{\beta}^{(k)}\|_1 + \langle \boldsymbol{\lambda} \circ \boldsymbol{\xi}^{(k)}, \boldsymbol{\beta} - \boldsymbol{\beta}^{(k)} \rangle$$

for any $\boldsymbol{\xi}^{(k)} \in \partial \|\boldsymbol{\beta}^{(k)}\|_1$. This further implies

$$\Psi(\boldsymbol{\beta}, \boldsymbol{\lambda}) \geq \mathcal{L}(\boldsymbol{\beta}^{(k-1)}) + \langle \nabla \mathcal{L}(\boldsymbol{\beta}^{(k-1)}), \boldsymbol{\beta} - \boldsymbol{\beta}^{(k-1)} \rangle$$
$$+ \|\boldsymbol{\lambda} \circ \boldsymbol{\beta}^{(k)}\|_1 + \langle \boldsymbol{\lambda} \circ \boldsymbol{\xi}^{(k)}, \boldsymbol{\beta} - \boldsymbol{\beta}^{(k)} \rangle. \tag{C.32}$$

Plugging the expression of $F(\boldsymbol{\beta}^{(k)}; \phi^{(k)}, \boldsymbol{\beta}^{(k-1)})$ in (3.3) and (C.32) into (C.31), we obtain

$$\Psi(\boldsymbol{\beta}, \boldsymbol{\lambda}) - \Psi(\boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}) \geq -\frac{\phi^{(k)}}{2} \|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}\|_2^2$$
$$+ \langle \nabla \mathcal{L}(\boldsymbol{\beta}^{(k-1)}), \boldsymbol{\beta} - \boldsymbol{\beta}^{(k)} \rangle + \langle \boldsymbol{\lambda} \circ \boldsymbol{\xi}^{(k)}, \boldsymbol{\beta} - \boldsymbol{\beta}^{(k)} \rangle. \tag{C.33}$$

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

$$\nabla \mathcal{L}(\boldsymbol{\beta}^{(k-1)}) + \phi^{(k)}(\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}) + \boldsymbol{\lambda} \circ \boldsymbol{\xi}^{(k)} = \mathbf{0}.$$

Substituting this into (C.33) proves the claimed bound. $\square$

Recall that $\Psi(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \mathcal{L}(\boldsymbol{\beta}) + \|\boldsymbol{\lambda} \circ \boldsymbol{\beta}\|_1$ and $\widehat{\boldsymbol{\beta}}^{(1)} \in \min_{\boldsymbol{\beta} \in \mathbb{R}^d} \Psi(\boldsymbol{\beta}, \boldsymbol{\lambda})$ denotes the optimal solution in the contraction stage.

Lemma C.5. For any $k \geq 1$, we have

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

Proof of Lemma C.5. For simplicity, we write $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{(1)}$, 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 $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ in Lemma C.4 gives

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

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

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

which further implies

$$\frac{2}{\phi_{\max}} \left\{ k\Psi(\widehat{\boldsymbol{\beta}}, \boldsymbol{\lambda}) - \sum_{j=1}^{k} \Psi(\boldsymbol{\beta}^{(j)}, \boldsymbol{\lambda}) \right\} \geq \|\boldsymbol{\beta}^{(k)} - \widehat{\boldsymbol{\beta}}\|_2^2 - \|\boldsymbol{\beta}^{(0)} - \widehat{\boldsymbol{\beta}}\|_2^2. \qquad \text{(C.34)}$$

Again, by Lemma C.4 with $\boldsymbol{\beta} = \boldsymbol{\beta}^{(j-1)}$ and $k = j$,

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

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

$$\frac{2}{\phi_{\min}} \sum_{j=1}^{k} \{(j-1)\Psi(\boldsymbol{\beta}^{(j-1)}, \boldsymbol{\lambda}) - j\Psi(\boldsymbol{\beta}^{(j)}, \boldsymbol{\lambda}) + \Psi(\boldsymbol{\beta}^{(j)}, \boldsymbol{\lambda})\}$$

$$\geq \sum_{j=1}^{k} (j-1)\|\boldsymbol{\beta}^{(j)} - \boldsymbol{\beta}^{(j-1)}\|_2^2,$$

or equivalently,

$$\frac{2}{\phi_{\min}} \left\{ -k\Psi(\boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}) + \sum_{j=1}^{k} \Psi(\boldsymbol{\beta}^{(j)}, \boldsymbol{\lambda}) \right\} \geq \sum_{j=1}^{k} (j-1)\|\boldsymbol{\beta}^{(j)} - \boldsymbol{\beta}^{(j-1)}\|_2^2. \qquad \text{(C.35)}$$

Together, (C.34) and (C.35) imply

$$\frac{2k}{\phi_{\min}} \{\Psi(\widehat{\boldsymbol{\beta}}, \boldsymbol{\lambda}) - \Psi(\boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda})\}$$

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

from which it follows immediately that

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

This completes the proof. □

C.4.2. Proof of the theorem

Recall that $\boldsymbol{\beta}^{(k)} = \boldsymbol{\beta}^{(1,k)}$ and $\phi^{(k)} = \phi^{(1,k)}$. By Lemma C.3 and its proof,

$$\omega_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{(k)}) \leq (\phi^{(k)} + L)\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}\|_2 \leq L(1 + \gamma_u)\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}\|_2.$$

Next, taking $\boldsymbol{\beta} = \boldsymbol{\beta}^{(k-1)}$ in Lemma C.4 yields

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

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

$$\omega_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{(k)}) \leq L(1 + \gamma_u)\left[\frac{2}{\phi^{(k)}}\left\{\Psi(\boldsymbol{\beta}^{(k-1)}, \boldsymbol{\lambda}) - \Psi(\boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda})\right\}\right]^{1/2}. \qquad \text{(C.36)}$$

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

$$\Psi(\widehat{\boldsymbol{\beta}}^{(1)}, \boldsymbol{\lambda}) \leq \cdots \leq \Psi(\boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}) \leq \cdots \leq \Psi(\boldsymbol{\beta}^{(0)}, \boldsymbol{\lambda}).$$

Then, it follows from (C.36) and Lemma C.5 that

$$\begin{aligned}
\omega_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{(k)}) &\leq L(1 + \gamma_u)\left[\frac{2}{\phi^{(k)}}\{\Psi(\boldsymbol{\beta}^{(k-1)}, \boldsymbol{\lambda}) - \Psi(\widehat{\boldsymbol{\beta}}, \boldsymbol{\lambda})\}\right]^{1/2} \\
&\leq \frac{L(1 + \gamma_u)}{\sqrt{k-1}}\sqrt{\frac{\max_{1 \leq j \leq k-1}\phi^{(j)}}{\phi^{(k)}}}\|\widehat{\boldsymbol{\beta}}\|_2,
\end{aligned}$$

where we used the fact that $\boldsymbol{\beta}^{(0)} = \mathbf{0}$. By the triangle inequality,

$$\omega_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{(k)}) \lesssim \frac{L(1 + \gamma_u)}{\sqrt{k}}\left(\|\boldsymbol{\beta}^*\|_2 + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2\right).$$

Therefore, in the contraction stage, we need $k \gtrsim \{L(1 + \gamma_u)(\|\boldsymbol{\beta}^*\|_2 + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2)/\epsilon_c\}^2$ to ensure $\omega_{\boldsymbol{\lambda}^{(0)}}(\boldsymbol{\beta}^{(k)}) \leq \epsilon_c$. This proves the stated result. $\qquad \square$

C.5. Proof of Theorem 3.2

For convenience, we omit the index ℓ, and use $\widehat{\boldsymbol{\beta}}$, $\boldsymbol{\beta}^{(k)}$, $\boldsymbol{\lambda}$ and $\mathcal{E}$ to denote $\widehat{\boldsymbol{\beta}}^{(\ell)}$, $\boldsymbol{\beta}^{(\ell,k)}$, $\boldsymbol{\lambda}^{(\ell-1)}$ and $\mathcal{E}_{\ell}$, respectively, where $\mathcal{E}_{\ell}$ is the subset defined in (B.2) satisfying $\mathcal{S} \subseteq \mathcal{E}_{\ell}$ and $|\mathcal{E}_{\ell}| \leq C_0 s$ for some constant $C_0 > 1$. Moreover, write $\mathcal{L}(\cdot) = \widehat{\mathcal{L}}_{\tau}(\cdot)$, and define $\Psi(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \mathcal{L}(\boldsymbol{\beta}) + \|\boldsymbol{\lambda} \circ \boldsymbol{\beta}\|_1 = \mathcal{L}(\boldsymbol{\beta}) + \|\boldsymbol{\lambda}^{(\ell-1)} \circ \boldsymbol{\beta}\|_1$, so that $\widehat{\boldsymbol{\beta}} \in \min_{\boldsymbol{\beta}} \Psi(\boldsymbol{\beta}, \boldsymbol{\lambda})$.

C.5.1. Technical lemmas

We first provide several technical lemmas along with the proofs.

Lemma C.6. For any k-sparse ($k \geq 1$) vectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \boldsymbol{\beta}^* + \mathbb{B}(r)$, we have

$$\frac{1}{2}\kappa_-(2k, r, \tau)\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 \leq D_{\mathcal{L}}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \leq \frac{1}{2}\kappa_+(2k, r, \tau)\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2.$$

where $D_{\mathcal{L}}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) := \mathcal{L}(\boldsymbol{\beta}_1) - \mathcal{L}(\boldsymbol{\beta}_2) - \langle \nabla \mathcal{L}(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \rangle$ is the Bregman divergence.

Proof of Lemma C.6. By a second-order Taylor series expansion, there exists some $\gamma \in [0, 1]$ such that $\widetilde{\boldsymbol{\beta}} = \gamma\boldsymbol{\beta}_1 + (1 - \gamma)\boldsymbol{\beta}_2 \in \boldsymbol{\beta}^* + \mathbb{B}(r)$ and $D_{\mathcal{L}}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = (1/2)(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)^{\mathrm{T}}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}})(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)$. The stated bounds then follow directly from Definition 3.1. $\square$

The next lemma converts the bound on $\Psi(\boldsymbol{\beta}, \boldsymbol{\lambda}) - \Psi(\boldsymbol{\beta}^*, \boldsymbol{\lambda})$ to that on $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2$. Recall that for any subset $\mathcal{E} \subseteq [d]$, we write $\boldsymbol{\beta}_{\mathcal{E}}$ as a subvector of $\boldsymbol{\beta}$ indexed by $\mathcal{E}$.

Lemma C.7. Assume LSE(1) condition holds. Let $\mathcal{E} \subseteq [d]$ be a subset satisfying $\mathcal{S} \subseteq \mathcal{E}$ and $|\mathcal{E}| \leq C_0 s$ for some $C_0 \geq 1$. Assume further that $\lambda \geq \max\{4\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty, \|\boldsymbol{\lambda}\|_\infty\}$ and $\|\boldsymbol{\lambda}_{\mathcal{E}^c}\|_{\min} \geq \lambda/2$. Then, for any $\boldsymbol{\beta} \in \boldsymbol{\beta}^* + \mathbb{B}(r)$ satisfying $\|\boldsymbol{\beta}_{\mathcal{S}^c}\|_0 \leq s'$ and $\Psi(\boldsymbol{\beta}, \boldsymbol{\lambda}) - \Psi(\boldsymbol{\beta}^*, \boldsymbol{\lambda}) \leq Cs\lambda^2$, we have

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq C_1 s^{1/2}\lambda \quad \text{and} \quad \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \leq C_2\sqrt{s(s + s')}\lambda,$$

where $C_1, C_2 > 0$ depend only on C_0, C and localized sparse eigenvalues.

Proof of Lemma C.7. We omit the arguments in $\kappa_-(s + s', r, \tau)$ and $\kappa_+(s + s', r, \tau)$ whenever there is no ambiguity. For any $\boldsymbol{\beta} \in \boldsymbol{\beta}^* + \mathbb{B}(r)$ satisfying $\|\boldsymbol{\beta}_{\mathcal{S}^c}\|_0 \leq s'$, note that $\|\boldsymbol{\beta}\|_0 \leq s + s'$ and $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_0 \leq s + s'$. Using Lemma C.6 yields

$$\mathcal{L}(\boldsymbol{\beta}^*) + \langle \nabla\mathcal{L}(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle + \frac{\kappa_-}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \leq \mathcal{L}(\boldsymbol{\beta}).$$

Since $\Psi(\boldsymbol{\beta}) - \Psi(\boldsymbol{\beta}^*) \leq Cs\lambda^2$, or equivalently,

$$\mathcal{L}(\boldsymbol{\beta}) - \mathcal{L}(\boldsymbol{\beta}^*) + (\|\boldsymbol{\lambda} \circ \boldsymbol{\beta}\|_1 - \|\boldsymbol{\lambda} \circ \boldsymbol{\beta}^*\|_1) \leq Cs\lambda^2, \tag{C.37}$$

it follows

$$\frac{\kappa_-}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \leq Cs\lambda^2 - \underbrace{\langle \nabla\mathcal{L}(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle}_{\mathrm{I}} + \underbrace{(\|\boldsymbol{\lambda} \circ \boldsymbol{\beta}^*\|_1 - \|\boldsymbol{\lambda} \circ \boldsymbol{\beta}\|_1)}_{\mathrm{II}}.$$

After some simple algebra, it can be derived that

$$|\mathrm{I}| \leq \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\mathcal{E}^c}\|_1\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty + \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\mathcal{E}}\|_1\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty,$$
$$\mathrm{II} \leq \lambda\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\mathcal{E}}\|_1 - (\lambda/2)\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\mathcal{E}^c}\|_1.$$

Combining the above bounds gives

$$\frac{\kappa_-}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 + \{\lambda/2 - \|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty\}\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\mathcal{E}^c}\|_1$$
$$\leq \{\lambda + \|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty\}\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\mathcal{E}}\|_1 + Cs\lambda^2,$$

which further implies

$$\frac{\kappa_-}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \leq \frac{5\lambda}{4}\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\mathcal{E}}\|_1 + Cs\lambda^2.$$

To bound the right-hand side of the above inequality, we discuss two cases regarding the magnitude of $\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\mathcal{E}}\|_1$ as compared to $s\lambda$:

- If $5\lambda\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\mathcal{E}}\|_1/4 \leq Cs\lambda^2$, we have

$$\frac{\kappa_-}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \leq 2Cs\lambda^2, \text{ and hence}$$
$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq 2(C/\kappa_-)^{1/2}s^{1/2}\lambda. \tag{C.38}$$

- If $5\lambda\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\mathcal{E}}\|_1/4 > Cs\lambda^2$, we have

$$\frac{\kappa_-}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \leq \frac{5}{2}\lambda\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\mathcal{E}}\|_1 \leq \frac{5}{2}\lambda(C_0 s)^{1/2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2,$$

  thus implying

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq 5C_0^{1/2}\kappa_-^{-1}s^{1/2}\lambda. \tag{C.39}$$

Combining (C.38) and (C.39), we obtain

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq \max\left\{2(C/\kappa_-)^{1/2}, 5C_0^{1/2}\kappa_-^{-1}\right\}s^{1/2}\lambda \asymp s^{1/2}\lambda.$$

Since $\boldsymbol{\beta} - \boldsymbol{\beta}^*$ is at most $(s + s')$-sparse, $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \leq (s + s')^{1/2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2$. The stated results then follow immediately. $\square$

Recall that $\mathcal{E}_\ell$ is the subset defined in (B.2) satisfying $\mathcal{S} \subseteq \mathcal{E}_\ell$ and $|\mathcal{E}_\ell| \leq C_0 s$ for some $C_0 > 1$.

Lemma C.8. Assume LSE(C_0) condition holds and $4\{\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty + \epsilon_{\rm c} \vee \epsilon_{\rm t}\} \leq \lambda \lesssim s^{-1/2}r$. For any $\ell \geq 2$, the solution sequence $\{\boldsymbol{\beta}^{(\ell,k)}\}_{k \geq 0}$ satisfies

$$\|\boldsymbol{\beta}^{(\ell,k)}_{\mathcal{E}_\ell^c}\|_0 \leq s', \quad \|\boldsymbol{\beta}^{(\ell,k)} - \boldsymbol{\beta}^*\|_2 \leq C_1 s^{1/2}\lambda \quad \text{and}$$
$$\|\boldsymbol{\beta}^{(\ell,k)} - \boldsymbol{\beta}^*\|_1 \leq C_2 s\lambda, \tag{C.40}$$

where $C_1, C_2 > 0$ are constants depending only on the localized sparse eigenvalues.

Proof of Lemma C.8. We prove the theorem by the method of induction on (ℓ, k). Throughout, C denotes a constant independent of (n, d, s) and may take

different values at each appearance. For the first subproblem, directly applying Proposition 4.1 and Lemma 5.4 in [16] we obtain that $\|\widetilde{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}\|_2 \leq C\kappa_*^{-1}s^{1/2}\lambda < r$, $\|\widetilde{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}\|_1 \leq C\kappa_*^{-1}s\lambda$ and $\widetilde{\boldsymbol{\beta}}^{(1)}$ is $(s+s')$-sparse, where $s' \leq Cs$. It follows that $\boldsymbol{\beta}^{(2,0)} = \widetilde{\boldsymbol{\beta}}^{(1)}$ falls in a localized sparse set.

To apply the method of induction, first we assume that for any k, $\boldsymbol{\beta}^{(2,k)}$ falls in a localized sparse set such that (C.40) holds. We then use Lemma E.13 in [16] to show that $\boldsymbol{\beta}^{(2,k+1)}$ also falls in a localized sparse set. To this end, we need to verify two conditions. The first one, $\|\boldsymbol{\lambda}_{\mathcal{E}_\ell^c}^{(\ell)}\|_{\min} \geq \lambda/2$ is guaranteed by Claim (B.3) in the proof of Proposition 2.2, when γ is such that $p'(\gamma) = 1/2$ and $|\mathcal{E}_l| \leq C_0 s$ for some $C_0 > 1$. For the second condition, it suffices to show

$$\Psi(\boldsymbol{\beta}^{(2,k)}, \boldsymbol{\lambda}^{(1)}) - \Psi(\boldsymbol{\beta}^*, \boldsymbol{\lambda}^{(1)}) \lesssim (1+\zeta)\kappa_*^{-1}s\lambda^2,$$

where $\zeta = \kappa^*/\kappa_*$. Using the mean value theorem, there exists some convex combination of $\boldsymbol{\beta}^{(2,k)}$ and $\boldsymbol{\beta}^*$, say $\widetilde{\boldsymbol{\beta}}$, such that

$$
\begin{aligned}
&\Psi(\boldsymbol{\beta}^{(2,k)}, \boldsymbol{\lambda}^{(1)}) - \Psi(\boldsymbol{\beta}^*, \boldsymbol{\lambda}^{(1)}) \\
&= \mathcal{L}(\boldsymbol{\beta}^{(2,k)}) - \mathcal{L}(\boldsymbol{\beta}^*) + \{\|\boldsymbol{\lambda}^{(1)} \circ \boldsymbol{\beta}^{(2,k)}\|_1 - \|\boldsymbol{\lambda}^{(1)} \circ \boldsymbol{\beta}^*\|_1\} \\
&\leq \langle \nabla\mathcal{L}(\boldsymbol{\beta}^*), \boldsymbol{\beta}^{(2,k)} - \boldsymbol{\beta}^* \rangle + \frac{1}{2}(\boldsymbol{\beta}^{(2,k)} - \boldsymbol{\beta}^*)^{\mathrm{T}} \nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}})(\boldsymbol{\beta}^{(2,k)} - \boldsymbol{\beta}^*) \\
&\quad + \|\boldsymbol{\lambda}^{(1)} \circ (\boldsymbol{\beta}^{(2,k)} - \boldsymbol{\beta}^*)\|_1 \\
&\leq \|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty \|\boldsymbol{\beta}^{(2,k)} - \boldsymbol{\beta}^*\|_1 + \frac{1}{2}\kappa^* \|\boldsymbol{\beta}^{(2,k)} - \boldsymbol{\beta}^*\|_2^2 + \lambda\|\boldsymbol{\beta}^{(2,k)} - \boldsymbol{\beta}^*\|_1 \\
&\leq \frac{C}{4}\kappa_*^{-1}s\lambda^2 + \frac{C^2}{2}\kappa^*\kappa_*^{-2}s\lambda^2 + C\kappa_*^{-1}s\lambda^2 \lesssim (1+\zeta)\kappa_*^{-1}s\lambda^2.
\end{aligned}
$$

With above preparations, it follows from Lemma E.13 in [16] with slight modification that $\|\boldsymbol{\beta}_{\mathcal{E}_\ell^c}^{(2,k+1)}\|_0 \leq s'$ and hence $\|\boldsymbol{\beta}^{(2,k+1)}\|_0 \leq C_0 s + s'$.

Next, we show that $\|\boldsymbol{\beta}^{(2,k+1)} - \boldsymbol{\beta}^*\|_2 \lesssim \kappa_*^{-1}s^{1/2}\lambda$. Again, by Lemma C.4,

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

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

$$\Psi(\boldsymbol{\beta}^{(2,k+1)}, \boldsymbol{\lambda}^{(1)}) - \Psi(\boldsymbol{\beta}^*, \boldsymbol{\lambda}^{(1)}) \leq \Psi(\boldsymbol{\beta}^{(2,k)}, \boldsymbol{\lambda}^{(1)}) - \Psi(\boldsymbol{\beta}^*, \boldsymbol{\lambda}^{(1)}) \lesssim (1+\zeta)\kappa_*^{-1}s\lambda^2.$$

Combining this with Lemma C.7 gives the desired bounds on $\|\boldsymbol{\beta}^{(2,k+1)} - \boldsymbol{\beta}^*\|_2$ and $\|\boldsymbol{\beta}^{(2,k+1)} - \boldsymbol{\beta}^*\|_1$.

Finally, by an argument similar to that in the proof of Lemma 5.4 in [16], we can derive the stated results for all $\ell \geq 3$. □

For $\epsilon > 0$, let $\widetilde{\boldsymbol{\beta}}$ be an ϵ-optimal solution to the program $\min_{\boldsymbol{\beta}}\{\mathcal{L}(\boldsymbol{\beta}) + \|\boldsymbol{\lambda} \circ \boldsymbol{\beta}\|_1\}$. The following lemma provides conditions under which $\widetilde{\boldsymbol{\beta}}$ falls in an ℓ_1-cone.

Lemma C.9. Let $\mathcal{E} \subseteq [d]$ be a subset satisfying $\mathcal{S} \subseteq \mathcal{E}$, and assume $\lambda \geq \|\boldsymbol{\lambda}\|_\infty \vee 4\{\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty + \epsilon\}$ and $\|\boldsymbol{\lambda}_{\mathcal{E}^c}\|_{\min} \geq \lambda/2$. Then, any ϵ-optimal solution $\widetilde{\boldsymbol{\beta}}$ satisfies the cone constraint

$$\|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{E}^c}\|_1 \leq \frac{\|\boldsymbol{\lambda}\|_\infty + \|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty + \epsilon}{\|\boldsymbol{\lambda}_{\mathcal{E}^c}\|_{\min} - \|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty - \epsilon}\|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{E}}\|_1 \leq 5\|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{E}}\|_1.$$

Proof of Lemma C.9. For any $\boldsymbol{\xi} \in \partial\|\widetilde{\boldsymbol{\beta}}\|_1$, let $\boldsymbol{u} = \nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}) + \boldsymbol{\lambda}\circ\boldsymbol{\xi}$. By the convexity of $\mathcal{L}(\cdot)$, $\langle\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}) - \nabla\mathcal{L}(\boldsymbol{\beta}^*), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle \geq 0$. This, together with the inequality $\langle\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}) + \boldsymbol{\lambda}\circ\boldsymbol{\xi}, \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle \leq \|\boldsymbol{u}\|_\infty\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$, implies

$$0 \leq \|\boldsymbol{u}\|_\infty\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 - \underbrace{\langle\nabla\mathcal{L}(\boldsymbol{\beta}^*), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle}_{\text{I}} - \underbrace{\langle\boldsymbol{\lambda}\circ\boldsymbol{\xi}, \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle}_{\text{II}}. \tag{C.41}$$

For I and II, note that $\text{I} \geq -\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$, and

$$\text{II} = \langle\boldsymbol{\lambda}\circ\boldsymbol{\xi}, \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle = \langle(\boldsymbol{\lambda}\circ\boldsymbol{\xi})_{\mathcal{E}^c}, (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{E}^c}\rangle + \langle(\boldsymbol{\lambda}\circ\boldsymbol{\xi})_{\mathcal{E}}, (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{E}}\rangle$$
$$\geq \|\boldsymbol{\lambda}_{\mathcal{E}^c}\|_{\min}\|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{E}^c}\|_1 - \|\boldsymbol{\lambda}_{\mathcal{E}}\|_\infty\|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{E}}\|_1.$$

Substituting the above bounds into (C.41) and taking the infimum over $\boldsymbol{\xi} \in \partial\|\widetilde{\boldsymbol{\beta}}\|_1$ yields

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

or equivalently,

$$\|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{E}^c}\|_1 \leq \frac{\|\boldsymbol{\lambda}\|_\infty + \|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty + \omega_{\boldsymbol{\lambda}}(\widetilde{\boldsymbol{\beta}})}{\|\boldsymbol{\lambda}_{\mathcal{E}^c}\|_{\min} - \{\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty + \omega_{\boldsymbol{\lambda}}(\widetilde{\boldsymbol{\beta}})\}}\|(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{E}}\|_1.$$

This proves the stated result. $\qquad\square$

C.5.2. *Proof of the theorem*

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

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

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

$$
\begin{aligned}
\Psi(\boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}) &\leq \min_{\boldsymbol{\beta} \in \mathbb{L}} \left\{ \Psi(\boldsymbol{\beta}, \boldsymbol{\lambda}) + \frac{\phi^{(k)}}{2} \|\boldsymbol{\beta} - \boldsymbol{\beta}^{(k-1)}\|_2^2 \right\} \\
&\leq \min_{0 \leq \alpha \leq 1} \left\{ \alpha \Psi(\widehat{\boldsymbol{\beta}}, \boldsymbol{\lambda}) + (1 - \alpha)\Psi(\boldsymbol{\beta}^{(k-1)}, \boldsymbol{\lambda}) + \frac{\alpha^2 \phi^{(k)}}{2} \|\boldsymbol{\beta}^{(k-1)} - \widehat{\boldsymbol{\beta}}\|_2^2 \right\} \\
&= \min_{0 \leq \alpha \leq 1} \left\{ \Psi(\boldsymbol{\beta}^{(k-1)}, \boldsymbol{\lambda}) - \alpha\{\Psi(\boldsymbol{\beta}^{(k-1)}, \boldsymbol{\lambda}) - \Psi(\widehat{\boldsymbol{\beta}}, \boldsymbol{\lambda})\} \right. \\
&\qquad\qquad \left. + \frac{\alpha^2 \phi^{(k)}}{2} \|\boldsymbol{\beta}^{(k-1)} - \widehat{\boldsymbol{\beta}}\|_2^2 \right\}.
\end{aligned}
\tag{C.42}
$$

Next, we bound the right-hand side of (C.42). By Lemma C.8,

$$
\|(\boldsymbol{\beta}^{(k-1)})_{\mathcal{E}_\ell^c}\|_0 \leq s', \quad \|\boldsymbol{\beta}^{(k-1)} - \boldsymbol{\beta}^*\|_2 \lesssim s^{1/2}\lambda \lesssim r \quad \text{and} \quad \|\boldsymbol{\beta}^{(k-1)} - \boldsymbol{\beta}^*\|_2 \lesssim s\lambda.
$$

Similarly, it can be shown the the optimum $\widehat{\boldsymbol{\beta}}$ satisfies the same properties. Hence,

$$
\boldsymbol{\beta}^{(k)}, \widehat{\boldsymbol{\beta}} \in \boldsymbol{\beta}^* + \mathbb{B}(r) \quad \text{and} \quad \|\boldsymbol{\beta}^{(k)} - \widehat{\boldsymbol{\beta}}\|_0 \leq |\mathcal{E}_\ell| + 2s' \leq C_0 s + 2s'.
$$

By the first-order optimality condition, there exists some $\widehat{\boldsymbol{\xi}} \in \partial\|\widehat{\boldsymbol{\beta}}\|_1$ such that $\nabla\mathcal{L}(\widehat{\boldsymbol{\beta}}) + \boldsymbol{\lambda} \circ \widehat{\boldsymbol{\xi}} = \mathbf{0}$. Moreover, define $D_{\mathcal{L}}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \mathcal{L}(\boldsymbol{\beta}_1) - \mathcal{L}(\boldsymbol{\beta}_2) - \langle\nabla\mathcal{L}(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\rangle$. Using Definition 3.1, Lemma C.6, and the convexity of $\mathcal{L}(\cdot)$ and ℓ_1-norm, we obtain that

$$
\begin{aligned}
&\Psi(\boldsymbol{\beta}^{(k-1)}, \boldsymbol{\lambda}) - \Psi(\widehat{\boldsymbol{\beta}}, \boldsymbol{\lambda}) \\
&\geq \langle\nabla\mathcal{L}(\widehat{\boldsymbol{\beta}}) + \boldsymbol{\lambda} \circ \widehat{\boldsymbol{\xi}}, \boldsymbol{\beta}^{(k-1)} - \widehat{\boldsymbol{\beta}}\rangle + D_{\mathcal{L}}(\boldsymbol{\beta}^{(k-1)}, \widehat{\boldsymbol{\beta}}) \geq \frac{\kappa_-}{2} \|\boldsymbol{\beta}^{(k-1)} - \widehat{\boldsymbol{\beta}}\|_2^2,
\end{aligned}
$$

where $\kappa_- = \kappa_-(C_0 s + 2s', r, \tau)$. Plugging this bound into (C.42) yields

$$
\begin{aligned}
&\Psi(\boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}) \\
&\leq \min_{0 \leq \alpha \leq 1} \left[\Psi(\boldsymbol{\beta}^{(k-1)}, \boldsymbol{\lambda}) - \alpha\{\Psi(\boldsymbol{\beta}^{(k-1)}, \boldsymbol{\lambda}) - \Psi(\widehat{\boldsymbol{\beta}}, \boldsymbol{\lambda})\} \right. \\
&\qquad\qquad \left. + \frac{\alpha^2 \phi^{(k)}}{\kappa_-}\{\Psi(\boldsymbol{\beta}^{(k-1)}, \boldsymbol{\lambda}) - \Psi(\widehat{\boldsymbol{\beta}}, \boldsymbol{\lambda})\} \right] \\
&\leq \Psi(\boldsymbol{\beta}^{(k-1)}, \boldsymbol{\lambda}) - \frac{\kappa_-}{4\phi^{(k)}}\{\Psi(\boldsymbol{\beta}^{(k-1)}, \boldsymbol{\lambda}) - \Psi(\widehat{\boldsymbol{\beta}}, \boldsymbol{\lambda})\}.
\end{aligned}
$$

Following the proof of Lemma C.3, it can be similarly shown that $\phi^{(k)} \leq \gamma_u \kappa^*$ under Condition 3.2. Consequently,

$$
\Psi(\boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}) - \Psi(\widehat{\boldsymbol{\beta}}, \boldsymbol{\lambda}) \leq \left(1 - \frac{1}{4\gamma_u \zeta}\right)^k \{\Psi(\boldsymbol{\beta}^{(0)}, \boldsymbol{\lambda}) - \Psi(\widehat{\boldsymbol{\beta}}, \boldsymbol{\lambda})\},
$$

where $\zeta = \kappa^*/\kappa_*$.

By an argument similar to that in the proof of Lemma C.3, we can show that, for $\ell \geq 2$,

$$\omega_{\boldsymbol{\lambda}^{\ell-1}}(\boldsymbol{\beta}^{(\ell,k)}) \leq \kappa^*(1 + \gamma_u)\|\boldsymbol{\beta}^{(\ell,k)} - \boldsymbol{\beta}^{(\ell,k-1)}\|_2.$$

Further, using Lemma C.4 to bound $\|\boldsymbol{\beta}^{(\ell,k)} - \boldsymbol{\beta}^{(\ell,k-1)}\|_2$ from above and noting that $\phi^{(k)} \geq \kappa_*$, we obtain

$$\omega_{\boldsymbol{\lambda}^{(\ell-1)}}(\boldsymbol{\beta}^{(\ell,k)})$$
$$\leq (1 + \gamma_u)\kappa^*\sqrt{(2/\kappa_*)\{\Psi(\boldsymbol{\beta}^{(\ell,k-1)}, \boldsymbol{\lambda}^{(\ell-1)}) - \Psi(\boldsymbol{\beta}^{(\ell,k)}, \boldsymbol{\lambda}^{(\ell-1)})\}}$$
$$\leq (1 + \gamma_u)\sqrt{2\zeta\kappa^*\{\Psi(\boldsymbol{\beta}^{(\ell,k-1)}, \boldsymbol{\lambda}^{(\ell-1)}) - \Psi(\widehat{\boldsymbol{\beta}}^{(\ell)}, \boldsymbol{\lambda}^{(\ell-1)})\}}$$
$$\leq (1 + \gamma_u)\sqrt{2\zeta\kappa^*\left(1 - \frac{1}{4\gamma_u\zeta}\right)^{k-1}\{\Psi(\boldsymbol{\beta}^{(\ell,0)}, \boldsymbol{\lambda}^{(\ell-1)}) - \Psi(\widehat{\boldsymbol{\beta}}^{(\ell)}, \boldsymbol{\lambda}^{(\ell-1)})\}}$$
$$\leq C(1 + \gamma_u)\sqrt{\frac{\zeta\kappa^*}{\phi^{(\ell,0)}}\left(1 - \frac{1}{4\gamma_u\zeta}\right)^{k-1}s\lambda^2} \leq C(1 + \gamma_u)\zeta\sqrt{\left(1 - \frac{1}{4\gamma_u\zeta}\right)^{k-1}s\lambda^2},$$

where the last step applies Lemmas C.4 and C.8.

To make the right-hand side of the above inequality smaller than ϵ_t, we need k to be sufficiently large that $k \geq C_1 \log(C_2 s^{1/2}\lambda/\epsilon_t)$, where $C_1, C_2 > 0$ are constants depending only on localized sparse eigenvalues and γ_u. This completes the proof. □

References

[1] ALQUIER, P., COTTET, V. and LECUÉ, G. (2019). Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions. *Ann. Statist.* **47** 2117–2144. MR3953446

[2] BELLEC, P. C., LECUÉ, G. and TSYBAKOV, A. B. (2018). Slope meets Lasso: Improved oracle bounds and optimality. *Ann. Statist.* **46** 3603–3642. MR3852663

[3] BELLONI, A., CHERNOZHUKOV. V. and WANG, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806. MR2860324

[4] BERTSIMAS, D., PAUPHILET, J. and VAN PARYS, B. (2020). Sparse regression: Scalable algorithms and empirical performance. *Statist. Sci.* **35** 555–578. MR4175381

[5] BOGDAN, M., VAN DEN BERG, E., SABATTI, C., SU, W. and CANDÉS, E. J. (2015). SLOPE–Adaptive variable selection via convex optimization. *Ann. Appl. Stat.* **9** 1103–1140. MR3418717

[6] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford Univ. Press, Oxford. MR3185193

[7] BOUSQUET, O. (2003). Concentration inequalities for sub-additive functions using the entropy method. *In Stochastic Inequalities and Applications. Progress in Probability* **56** 213–247. Birkhäuser, Basel. MR2073435

[8] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer, Heidelberg. MR2807761

[9] CATONI, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185. MR3052407

[10] CATONI, O. and GIULINI, I. (2017). Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint* arXiv:1712.02747.

[11] CHEN, X. and ZHOU, W.-X. (2020). Robust inference via multiplier bootstrap. *Ann. Statist.* **48** 1665–1691. MR4124339

[12] CHINOT, G., LECUÉ, G. and LERASLE, M. (2020a). Robust statistical learning with Lipschitz and convex loss functions. *Probab. Theory Relat. Fields* **45** 866–896. MR4087486

[13] CHINOT, G., LECUÉ, G. and LERASLE, M. (2020b). Robust high dimensional learning for Lipschitz and convex losses. *J. Mach. Learn. Res.* **21**(233): 1–47. MR4209519

[14] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[15] FAN, J., LI, Q. and WANG, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 247–265. MR3597972

[16] FAN, J., LIU, H., SUN, Q. and ZHANG, T. (2018). I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *Ann. Statist.* **46** 814–841. MR3782385

[17] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22. MR1082147

[18] GAÏFFAS, S. and LECUÉ, G. (2011). Weighted algorithms for compressed sensing and matrix completion. *arXiv preprint* arXiv:1107.1638.

[19] HASTIE, T., TIBSHIRANI, R. and TIBSHIRANI, R. (2020). Best subset, forward stepwise or Lasso? Analysis and recommendations based on extensive comparisons. *Statist. Sci.* **35** 579–592. MR4175382

[20] HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations.* CRC Press, Boca Raton. MR3616141

[21] HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101. MR0161415

[22] JERRUM, M. R., VALIANT, L. G. and VAZIRANI, V. V. (1986). Random generation of combinatorial structures from a uniform distribution. *Theor. Comput. Sci.* **43** 169–188. MR0855970

[23] LAHIRI, S. N. (2021). Necessary and sufficient conditions for variable se-

lection consistency of the LASSO in high dimensions. *Ann. Statist.* **49** 820–844. MR4255109

[24] LAMBERT-LACROIX, S. and ZWALD, L. (2011). Robust regression through the Huber's criterion and adaptive lasso penalty. *Electron. J. Statist.* **5** 1015–1053. MR2836768

[25] LANGE, K., HUNTER, D. R. and YANG, I. (2000). Optimization transfer using surrogate objective functions. *J. Comput. Graph. Stat.* **9** 1–20. MR1819865

[26] LECUÉ, G. and LERASLE, M. (2018). Learning from MOM's principles: Le Cam's approach. *Stochast. Process. Appl.* **129** 4385–4410. MR4013866

[27] LECUÉ, G. and LERASLE, M. (2020). Robust machine learning by median-of-means: Theory and practice. *Ann. Statist.* **48** 906–931. MR4102681

[28] LECUÉ, G. and MENDELSON, S. (2018). Regularization and the small-ball method I: Sparse recovery. *Ann. Statist.* **46** 611–641. MR3782379

[29] LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, Berlin. MR1102015

[30] LOH, P. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Ann. Statist.* **45** 866–896. MR3650403

[31] LOH, P.-L. and WAINWRIGHT, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.* **16** 559–616. MR3335800

[32] LOH, P.-L. and WAINWRIGHT, M. J. (2017). Support recovery without incoherence: A case for nonconvex regularization. *Ann. Statist.* **45** 2455–2482. MR3737898

[33] LUGOSI, G. and MENDELSON, S. (2019a). Regularization, sparse recovery, and median-of-means tournaments. *Bernoulli* **25** 2075–2106. MR3961241

[34] LUGOSI, G. and MENDELSON, S. (2019b). Mean estimation and regression under heavy-tailed distributions: A survey. *Found. Comput. Math.* **19** 1145–1190. MR4017683

[35] LUGOSI, G. and MENDELSON, S. (2020). Risk minimization by median-of-means tournaments. *J. Eur. Math. Soc.* **22** 925–965. MR4055993

[36] MEI, S., BAI, Y. and MONTANARI, A. (2018). The landscape of empirical risk for nonconvex losses. *Ann. Statist.* **46** 2747–2774. MR3851754

[37] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[38] MINSKER, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** 2308–2335. MR3378468

[39] NDAOUD, M. (2019). Interplay of minimax estimation and minimax support recovery under sparsity. In *Proc. Mach. Learn. Res.* **98** 647–668. MR3932863

[40] NEMIROVSKY, A. S. and YUDIN, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization.* Wiley, New York. MR0702836

[41] SHEN, X., PAN, W. and ZHU, Y. (2012). Likelihood-based selection and sharp parameter estimation. *J. Amer. Statist. Assoc.* **107** 223–232. MR2949354

[42] SPOKOINY, V. (2012). Parametric estimation. Finite sample theory. *Ann. Statist.* **40** 2877–2909. MR3097963

[43] SUN, T. and ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166

[44] SUN, Q., ZHOU, W.-X. and FAN, J. (2020). Adaptive Huber regression. *J. Amer. Statist. Assoc.* **115** 254–265. MR4078461

[45] THEODOSSIOU, P. (1998). Financial data and the skewed generalized *t* distribution. *Manag. Sci.* **44** 1650–1661.

[46] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58** 267–288. MR1379242

[47] VERSHYNIN, R. (2018). *High-Dimensional Probability.* Cambridge University Press, Cambridge. MR3837109

[48] WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory* **55** 2183–2202. MR2729873

[49] WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics.* Cambridge University Press, Cambridge. MR3967104

[50] WANG, L., ZHENG, C., ZHOU, W. and ZHOU, W.-X. (2021). A new principle for tuning-free Huber regression. *Statistica Sinica,* to appear.

[51] YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574

[52] ZHANG, C.-H. (2010a). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

[53] ZHANG, C.-H. and ZHANG, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.* **27** 576–593. MR3025135

[54] ZHANG, T. (2010b). Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.* **11** 1081–1107. MR2629825

[55] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

[56] ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

[57] ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327

[58] ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443