Adaptive Huber Regression[*]

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan

Abstract

Big data can easily be contaminated by outliers or contain variables with heavy-tailed distributions, which makes many conventional methods inadequate. To address this challenge, we propose the adaptive Huber regression for robust estimation and inference. The key observation is that the robustification parameter should adapt to the sample size, dimension and moments for optimal tradeoff between bias and robustness. Our theoretical framework deals with heavy-tailed distributions with bounded $(1+\delta)$-th moment for any $\delta > 0$. We establish a sharp phase transition for robust estimation of regression parameters in both low and high dimensions: when $\delta \geq 1$, the estimator admits a sub-Gaussian-type deviation bound without sub-Gaussian assumptions on the data, while only a slower rate is available in the regime $0 < \delta < 1$ and the transition is smooth and optimal. In addition, we extend the methodology to allow both heavy-tailed predictors and observation noise. Simulation studies lend further support to the theory. In a genetic study of cancer cell lines that exhibit heavy-tailedness, the proposed methods are shown to be more robust and predictive.

[*] Qiang Sun is Assistant Professor, Department of Statistical Sciences, University of Toronto, Toronto, ON M5S 3G3, Canada (E-mail: qsun@utstat.toronto.edu). Wen-Xin Zhou is Assistant Professor, Department of Mathematics, University of California, San Diego, La Jolla, CA 92093 (E-mail: wez243@ucsd.edu). Jianqing Fan is Honorary Professor, School of Data Science, Fudan University, Shanghai, China and Frederick L. Moore '18 Professor of Finance, Department of Operations Research and Financial Engineering, Princeton University, NJ 08544 (E-mail: jqfan@princeton.edu).

Keywords: Adaptive Huber regression, bias and robustness tradeoff, finite-sample inference, heavy-tailed data, nonasymptotic optimality, phase transition.

1 Introduction

Modern data acquisitions have facilitated the collection of massive and high dimensional data with complex structures. Along with holding great promises for discovering subtle population patterns that are less achievable with small-scale data, big data have introduced a series of new challenges to data analysis both computationally and statistically (Loh and Wainwright, 2015; Fan et al., 2018). During the last two decades, extensive progress has been made towards extracting useful information from massive data with high dimensional features and sub-Gaussian tails[1] (Tibshirani, 1996; Fan and Li, 2001; Efron et al., 2004; Bickel, Ritov and Tsybakov, 2009). We refer to the monographs, Bühlmann and van de Geer (2011) and Hastie, Tibshirani and Wainwright (2015), for a systematic coverage of contemporary statistical methods for high dimensional data.

The sub-Gaussian tails requirement, albeit being convenient for theoretical analysis, is not realistic in many practical applications since modern data are often collected with low quality. For example, a recent study on functional magnetic resonance imaging (fMRI) (Eklund, Nichols and Knutsson, 2016) shows that the principal cause of invalid fMRI inferences is that the data do not follow the assumed Gaussian shape, which speaks to the need of validating the statistical methods being used in the field of neuroimaging. In a microarray data example considered in Wang, Peng and Li (2015), it is observed that some gene expression levels have heavy tails as their kurtosises are much larger than 3, despite of the normalization methods used. In finance, the power-law nature of the distribution of returns has been validated as a stylized

[1] A random variable Z is said to have sub-Gaussian tails if there exists constants c_1 and c_2 such that $\mathbb{P}(|Z| > t) \le c_1 \exp(-c_2 t^2)$ for any $t \ge 0$.

fact (Cont, 2001). Fan et al. (2016) argued that heavy-tailed distribution is a stylized feature for high dimensional data and proposed a shrinkage principle to attenuate the influence of outliers. Standard statistical procedures that are based on the method of least squares often behave poorly in the presence of heavy-tailed data[2] (Catoni, 2012). It is therefore of ever-increasing interest to develop new statistical methods that are robust against heavy-tailed errors and other potential forms of contamination.

In this paper, we first revisit the robust regression that was initiated by Peter Huber in his seminal work Huber (1973). Asymptotic properties of the Huber estimator have been well studied in the literature. We refer to Huber (1973), Yohai and Maronna (1979), Portnoy (1985), Mammen (1989) and He and Shao (1996, 2000) for an unavoidably incomplete overview. However, in all of the aforementioned papers, the robustification parameter is suggested to be set as fixed according to the 95% asymptotic efficiency rule. Thus, this procedure can not estimate the model-generating parameters consistently when the sample distribution is asymmetric.

From a nonasymptotic perspective (rather than an asymptotic efficiency rule), we propose to use the Huber regression with an adaptive robustification parameter, which is referred to as the *adaptive Huber regression*, for robust estimation and inference. Our adaptive procedure achieves the nonasymptotic robustness in the sense that the resulting estimator admits exponential-type concentration bounds when only low-order moments exist. Moreover, the resulting estimator is also an asymptotically unbiased estimate for the parameters of interest. In particular, we do not impose symmetry and homoscedasticity conditions on error distributions, so that our problem is intrinsically different from median/quantile regression models, which are also of independent interest and serve as important robust techniques (Koenker, 2005).

We made several major contributions towards robust modeling in this paper. First and foremost, we establish nonasymptotic deviation bounds for adaptive Huber re-

[2] We say a random variable X has heavy tails if $\mathbb{P}(|X| > t)$ decays to zero polynomially in $1/t$ as $t \to \infty$.

gression when the error variables have only finite $(1+\delta)$-th moments. By providing a matching lower bound, we observe a sharp phase transition phenomenon, which is in line with that discovered by Devroye et al. (2016) for univariate mean estimation. Second, a similar phase transition for regularized adaptive Huber regression is established in high dimensions. By defining the effective dimension and effective sample size, we present nonasymptotic results under the two different regimes in a unified form. Last, by exploiting the localized analysis developed in Fan et al. (2018), we remove the artificial bounded parameter constraint imposed in previous works; see Loh and Wainwright (2015) and Fan, Li and Wang (2017). In the supplementary material, we present a nonasymptotic Bahadur representation for the adaptive Huber estimator when $\delta \geq 1$, which provides a theoretical foundation for robust finite-sample inference.

The rest of the paper proceeds as follows. The rest of this section is devoted to related literature. In Section 2, we revisit the Huber loss and robustification parameter, followed by the proposal of adaptive Huber regression in both low and high dimensions. We sharply characterize the nonasymptotic performance of the proposed estimators in Section 3. We describe the algorithm and implementation in Section 5. Section 6 is devoted to simulation studies and a real data application. In Section 4, we extend the methodology to allow possibly heavy-tailed covariates/predictors. All the proofs are collected in the supplemental material.

1.1 Related Literature

The terminology "robustness" used in this paper describes how stable the method performs with respect to the tail-behavior of the data, which can be either sub-Gaussian/sub-exponential or Pareto-like (Delaigle, Hall and Jin, 2011; Catoni, 2012; Devroye et al., 2016). This is different from the conventional perspective of robust statistics under Huber's ϵ-contamination model (Huber, 1964), for which a number of depth-based procedures have been developed since the groundbreaking work of John Tukey (Tukey, 1975). Significant contributions have also been made in Liu (1990), Liu, Parelius, and Singh (1999), Zuo and Serfling (2000), Mizera (2002) and Mizera and Müller (2004). We refer to Chen, Gao and Ren (2018) for the most recent result and a literature review concerning this problem.

Our main focus is on the conditional mean regression in the presence of heavy-tailed and asymmetric errors, which automatically distinguishes our method from quantile-based robust regressions (Koenker, 2005; Belloni and Chernozhukov, 2011; Wang, 2013; Fan, Fan and Barut, 2014; Zheng, Peng and He, 2015). In general, quantile regression is biased towards estimating the mean regression coefficient unless the error distributions are symmetric around zero. Another recent work that is related to ours is Alquier, Cottett and Lecué (2017). They studied a general class of regularized empirical risk minimization procedures with a particular focus on Lipschitz losses, which includes the quantile, hinge and logistic losses. Different from all these work, our goal is to estimate the mean regression coefficients robustly. The robustness is witnessed by a nonasymptotic analysis: the proposed estimators achieve sub-Gaussian deviation bounds when the regression errors have only finite second moments. Asymptotically, our proposed estimators are fully efficient: they achieve the same efficiency as the ordinary least squares estimators.

An important step towards estimation under heavy-tailedness has been made by Catoni (2012), whose focus is on estimating a univariate mean. Let X be a real-valued random variable with mean $\mu = \mathbb{E}(X)$ and variance $\sigma^2 = \mathrm{var}(X) > 0$, and assume that $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) from X. For any prespecified exception probability $t > 0$, Catoni constructs a robust mean estimator $\widehat{\mu}_{\mathrm{C}}(t)$ that deviates from the true mean μ logarithmically in $1/t$, that is,

$$\mathbb{P}\big[|\widehat{\mu}_{\mathrm{C}}(t) - \mu| \leq t\sigma/n^{1/2}\big] \geq 1 - 2\exp(-ct^2), \tag{1}$$

while the empirical mean deviates from the true mean only polynomially in $1/t^2$, namely subGaussian tails versus Cauchy tail in terms of t. Further, Devroye et al. (2016) developed adaptive sub-Gaussian estimators that are independent of the pre-specified exception probability. Beyond mean estimation, Brownlees, Joly and Lugosi (2015) extended Catoni's idea to study empirical risk minimization problems when the losses are unbounded. Generalizations of the univariate results to those for matrices, such as the covariance matrices, can be found in Catoni (2016), Minsker (2018), Giulini (2017) and Fan, Li and Wang (2017). Fan, Li and Wang (2017) modified Huber's procedure (Huber, 1973) to obtain a robust estimator, which is concentrated around the true mean with exponentially high probability in the sense of (1), and also proposed a robust procedure for sparse linear regression with asymmetric and heavy-tailed errors.

Notation: We fix some notations that will be used throughout this paper. For any vector $\boldsymbol{u} = (u_1, \ldots, u_d)^{\mathrm{T}} \in \mathbb{R}^d$ and $q \geq 1$, $\|\boldsymbol{u}\|_q = (\sum_{j=1}^d |u_j|^q)^{1/q}$ is the ℓ_q norm. For any vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, we write $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}^{\mathrm{T}} \boldsymbol{v}$. Moreover, we let $\|\boldsymbol{u}\|_0 = \sum_{j=1}^d 1(u_j \neq 0)$ denote the number of nonzero entries of $\boldsymbol{u}$, and set $\|\boldsymbol{u}\|_\infty = \max_{1 \leq j \leq d} |u_j|$. For two sequences of real numbers $\{a_n\}_{n \geq 1}$ and $\{b_n\}_{n \geq 1}$, $a_n \lesssim b_n$ denotes $a_n \leq C b_n$ for some constant $C > 0$ independent of n, $a_n \gtrsim b_n$ if $b_n \lesssim a_n$, and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For two scalars, we use $a \wedge b = \min\{a, b\}$ to denote the minimum of a and b. If $\mathbf{A}$ is an $m \times n$ matrix, we use $\|\mathbf{A}\|$ to denote its spectral norm, defined by $\|\mathbf{A}\| = \max_{\boldsymbol{u} \in \mathbb{S}^{n-1}} \|\mathbf{A}\boldsymbol{u}\|_2$, where $\mathbb{S}^{n-1} = \{\boldsymbol{u} \in \mathbb{R}^n : \|\boldsymbol{u}\|_2 = 1\}$ is the unit sphere in $\mathbb{R}^n$. For an $n \times n$ matrix $\mathbf{A}$, we use $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ to denote the maximum and minimum eigenvalues of $\mathbf{A}$, respectively. For two $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$, we write $\mathbf{A} \preceq \mathbf{B}$ if $\mathbf{B} - \mathbf{A}$ is positive semi-definite. For a function $f : \mathbb{R}^d \to \mathbb{R}$, we use $\nabla f \in \mathbb{R}^d$ to denote its gradient vector as long as it exists.

2 Methodology

We consider i.i.d. observations $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)$ that are generated from the following heteroscedastic regression model

$$y_i = \langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle + \varepsilon_i, \quad \text{with} \quad \mathbb{E}(\varepsilon_i | \boldsymbol{x}_i) = 0 \quad \text{and} \quad v_{i,\delta} = \mathbb{E}\big(|\varepsilon_i|^{1+\delta}\big) < \infty. \tag{2}$$

Assuming that the second moments are bounded ($\delta = 1$), the standard ordinary least squares (OLS) estimator, denoted by $\widehat{\boldsymbol{\beta}}^{\text{ols}}$, admits a suboptimal polynomial-type deviation bound, and thus does not concentrate around $\boldsymbol{\beta}^*$ tightly enough for large-scale simultaneous estimation and inference. The key observation that underpins this suboptimality of the OLS estimator is the sensitivity of quadratic loss to outliers (Huber, 1973; Catoni, 2012), while the Huber regression with a fixed tuning constant may lead to nonnegligible estimation bias. To overcome this drawback, we propose to employ the Huber loss with an adaptive robustification parameter to achieve robustness and (asymptotic) unbiasedness simultaneously. We begin with the definitions of the Huber loss and the corresponding robustification parameter.

Definition 1 (Huber Loss and Robustification Parameter)**.** The Huber loss $\ell_\tau(\cdot)$ (Huber, 1964) is defined as

$$\ell_\tau(x) = \begin{cases} x^2/2, & \text{if } |x| \le \tau, \\ \tau|x| - \tau^2/2, & \text{if } |x| > \tau, \end{cases}$$

where $\tau > 0$ is referred to as the robustification parameter that balances bias and robustness (Fan, Li and Wang, 2017).

The loss function $\ell_\tau(x)$ is quadratic for small values of x, and becomes linear when x exceeds τ in magnitude. The parameter τ therefore controls the blending of quadratic and ℓ_1 losses, which can be regarded as two extremes of the Huber loss with $\tau = \infty$ and $\tau \to 0$, respectively. Comparing with the least squares, outliers are down weighted in the Huber loss. We will use the name, *adaptive Huber loss*, to emphasize the fact that the parameter τ should adapt to the sample size, dimension and moments for a better tradeoff between bias and robustness. This distinguishes our framework from the classical setting. As $\tau \to \infty$ is needed to reduce the bias when the error distribution is asymmetric, this loss is also called the RA-quadratic (robust approximation to quadratic) loss in Fan, Li and Wang (2017).

Define the empirical loss function $\mathcal{L}_\tau(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \ell_\tau(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle)$ for $\boldsymbol{\beta} \in \mathbb{R}^d$. The Huber estimator is defined through the following convex optimization problem:

$$\widehat{\boldsymbol{\beta}}_\tau = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \mathcal{L}_\tau(\boldsymbol{\beta}). \tag{3}$$

In low dimensions, under the condition that $v_\delta = n^{-1}\sum_{i=1}^n \mathbb{E}(|\varepsilon_i|^{1+\delta}) < \infty$ for some $\delta > 0$, we will prove that $\widehat{\boldsymbol{\beta}}_\tau$ with $\tau \asymp \min\{v_\delta^{1/(1+\delta)}, v_1^{1/2}\}\, n^{\max\{1/(1+\delta),1/2\}}$ (the first factor is kept in order to show its explicit dependence on the moment) achieves the tight upper bound $d^{1/2}\tau^{-(\delta\wedge 1)} \asymp d^{1/2}n^{-\min\{\delta/(1+\delta),1/2\}}$. The phase transition at $\delta = 1$ can be easily observed (see Figure 1). When higher moments exist ($\delta \geq 1$), robustification leads to a sub-Gaussian-type deviation inequality in the sense of (1).

In the high dimensional regime, we consider the following regularized adaptive Huber regression with a different choice of the robustification parameter:

$$\widehat{\boldsymbol{\beta}}_{\tau,\lambda} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \big\{\mathcal{L}_\tau(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1\big\}, \tag{4}$$

where $\tau \asymp \nu_\delta\{n/(\log d)\}^{\max\{1/(1+\delta),1/2\}}$ and $\lambda \asymp \nu_\delta\{(\log d)/n\}^{\min\{\delta/(1+\delta),1/2\}}$ with $\nu_\delta = \min\{v_\delta^{1/(1+\delta)}, v_1^{1/2}\}$. Let s be the size of the true support $\mathcal{S} = \mathrm{supp}(\boldsymbol{\beta}^*)$. We will show that the regularized Huber estimator achieves an upper bound that is of the order $s^{1/2}\{(\log d)/n\}^{\min\{\delta/(1+\delta),1/2\}}$ for estimating $\boldsymbol{\beta}^*$ in ℓ_2-error with high probability.

To unify the nonasymptotic upper bounds in the two different regimes, we define the *effective dimension*, d_{eff}, to be d in low dimensions and s in high dimensions.

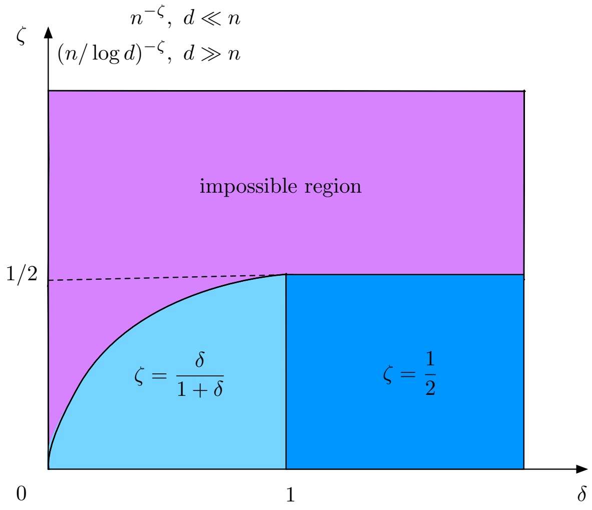

Figure 1: Phase transition in terms of ℓ_2-error for the adaptive Huber estimator. With fixed effective dimension, $\|\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*\|_2 \asymp n_{\text{eff}}^{-\delta/(1+\delta)}$, when $0 < \delta < 1$; $\|\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*\|_2 \asymp n_{\text{eff}}^{-1/2}$, when $\delta \geq 1$. Here n_{eff} is the effective sample size: $n_{\text{eff}} = n$ in low dimensions while $n_{\text{eff}} = n/\log d$ in high dimensions.

In other words, d_{eff} denotes the number of nonzero parameters of the problem. The *effective sample size*, n_{eff}, is defined as $n_{\text{eff}} = n$ and $n_{\text{eff}} = n/\log d$ in low and high dimensions, respectively. We will establish a phase transition: when $\delta \geq 1$, the proposed estimator enjoys a sub-Gaussian concentration, while it only achieves a slower concentration when $0 < \delta < 1$. Specifically, we show that, for any $\delta \in (0, \infty)$, the proposed estimators with $\tau \asymp \min\{v_\delta^{1/(1+\delta)}, v_1^{1/2}\}\, n_{\text{eff}}^{\max\{1/(1+\delta),1/2\}}$ achieve the following tight upper bound, up to logarithmic factors:

$$\big\|\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*\big\|_2 \lesssim d_{\text{eff}}^{1/2}\, n_{\text{eff}}^{-\min\{\delta/(1+\delta),1/2\}} \quad \text{with high probability.} \tag{5}$$

This finding is summarized in Figure 1.

3 Nonasymptotic Theory

3.1 Adaptive Huber Regression with Increasing Dimensions

We begin with the adaptive Huber regression in the low dimensional regime. First, we provide an upper bound for the estimation bias of Huber regression. We then establish the phase transition by establishing matching upper and lower bounds on the ℓ_2-error. The analysis is carried out under both fixed and random designs. The results under random designs are provided in the supplementary material. We start with the following regularity condition.

Condition 1. The empirical Gram matrix $\mathbf{S}_n := n^{-1}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$ is nonsingular. Moreover, there exist constants c_l and c_u such that $c_l \leq \lambda_{\min}(\mathbf{S}_n) \leq \lambda_{\max}(\mathbf{S}_n) \leq c_u$.

For any $\tau > 0$, $\widehat{\boldsymbol{\beta}}_\tau$ given in (3) is natural M-estimator of

$$\boldsymbol{\beta}_\tau^* := \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \mathbb{E}\{\mathcal{L}_\tau(\boldsymbol{\beta})\} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \mathbb{E}\{\ell_\tau(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle)\}, \tag{6}$$

where the expectation is taken over the regression errors. We call $\boldsymbol{\beta}_\tau^*$ the *Huber regression coefficient*, which is possibly different from the vector of true parameters $\boldsymbol{\beta}^*$. The estimation bias, measured by $\|\boldsymbol{\beta}_\tau^* - \boldsymbol{\beta}^*\|_2$, is a direct consequence of robustification and asymmetric error distributions. Heuristically, choosing a sufficiently large τ reduces bias at the cost of losing robustness (the extreme case of $\tau = \infty$ corresponds to the least squares estimator). Our first result shows how the magnitude of τ affects the bias $\|\boldsymbol{\beta}_\tau^* - \boldsymbol{\beta}^*\|_2$. Recall that $v_\delta = n^{-1}\sum_{i=1}^n v_{i,\delta}$ with $v_{i,\delta} = \mathbb{E}(|\varepsilon_i|^{1+\delta})$.

Proposition 1. Assume Condition 1 holds and that v_δ is finite for some $\delta > 0$. Then, the vector $\boldsymbol{\beta}_\tau^*$ of Huber regression coefficients satisfies

$$\|\boldsymbol{\beta}_\tau^* - \boldsymbol{\beta}^*\|_2 \leq 2c_l^{-1/2} v_\delta \tau^{-\delta} \tag{7}$$

provided $\tau \geq (4v_\delta \widetilde{M}^2)^{1/(1+\delta)}$ for $0 < \delta < 1$ or $\tau \geq (2v_1)^{1/2}\widetilde{M}$ for $\delta \geq 1$, where $\widetilde{M} = \max_{1\leq i\leq n} \|\mathbf{S}_n^{-1/2}\boldsymbol{x}_i\|_2$.

The total estimation error $\|\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*\|_2$ can therefore be decomposed into two parts

$$\underbrace{\|\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*\|_2}_{\text{Total error}} \leq \underbrace{\|\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*_\tau\|_2}_{\text{estimation error}} + \underbrace{\|\boldsymbol{\beta}^*_\tau - \boldsymbol{\beta}^*\|_2}_{\text{approximation bias}},$$

where the approximation bias is of order $\tau^{-\delta}$. A large τ reduces the bias but compromises the degree of robustness. Thus an optimal estimator is the one with τ diverging at a certain rate to achieve the optimal tradeoff between estimation error and approximation bias. Our next result presents nonasymptotic upper bounds on the ℓ_2-error with an exponential-type exception probability, when τ is properly tuned. Recall that $\nu_\delta = \min\{v_\delta^{1/(1+\delta)}, v_1^{1/2}\}$ for any $\delta > 0$.

Theorem 1 (Upper Bound)**.** Assume Condition 1 holds and $v_\delta < \infty$ for some $\delta > 0$. Let $L = \max_{1\leq i\leq n}\|\boldsymbol{x}_i\|_\infty$ and assume $n \geq C(L, c_l)d^2 t$ for some $C(L, c_l) > 0$ depending only on L and c_l. Then, for any $t > 0$ and $\tau_0 \geq \nu_\delta$, the estimator $\widehat{\boldsymbol{\beta}}_\tau$ with $\tau = \tau_0(n/t)^{\max\{1/(1+\delta),1/2\}}$ satisfies the bound

$$\|\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*\|_2 \leq 4c_l^{-1}L\tau_0\, d^{1/2}\left(\frac{t}{n}\right)^{\min\{\delta/(1+\delta),1/2\}} \tag{8}$$

with probability at least $1-(2d+1)e^{-t}$.

Remark 1. It is worth mentioning that the proposed robust estimator depends on the unknown parameter $v_\delta^{1/(1+\delta)}$. Adaptation to the unknown moment is indeed another important problem. In Section 6, we suggest a simple cross-validation scheme for choosing τ with desirable numerical performance. A general adaptive construction of τ can be obtained via Lepski's method (Lepski, 1991), which is more challenging due to unspecified constants. In the supplementary material, we discuss a variant of Lepski's method and establish its theoretical guarantee.

Remark 2. We do not assume $\mathbb{E}(|\varepsilon_i|^{1+\delta}|\boldsymbol{x}_i)$ to be a constant, and hence the proposed method accommodates heteroscedastic regression models. For example, ε_i can take the form of $\sigma(\boldsymbol{x}_i)v_i$, where $\sigma : \mathbb{R}^d \to (0, \infty)$ is a positive function, and v_i are random variables satisfying $\mathbb{E}(v_i) = 0$ and $\mathbb{E}(|v_i|^{1+\delta}) < \infty$.

Remark 3. We need the scaling condition to go roughly as $n \gtrsim d^2 t$ under fixed designs. With random designs, we show that the scaling condition can be relaxed to $n \gtrsim d + t$. Details are given in the supplementary material.

Theorem 1 indicates that, with only bounded $(1+\delta)$-th moment, the adaptive Huber estimator achieves the upper bound $d^{1/2}n^{-\min\{\delta/(1+\delta),1/2\}}$, up to a logarithmic factor, by setting $t = \log(nd)$. A natural question is whether the upper bound in (8) is optimal. To address this, we provide a matching lower bound up to a logarithmic factor. Let $\mathcal{P}_\delta^{v_\delta}$ be the class of all distributions on $\mathbb{R}$ whose $(1+\delta)$-th absolute central moment equals v_δ. Let $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{\mathrm{T}} = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^d) \in \mathbb{R}^{n\times d}$ be the design matrix and $\mathcal{U}_n = \{\boldsymbol{u} : \boldsymbol{u} \in \{-1, 1\}^n\}$.

Theorem 2 (Lower Bound)**.** Assume that the regression errors ε_i are i.i.d. from a distribution in $\mathcal{P}_\delta^{v_\delta}$ with $\delta > 0$. Suppose there exists a $\boldsymbol{u} \in \mathcal{U}_n$ such that $\|n^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{u}\|_{\min} \geq \alpha$ for some $\alpha > 0$. Then, for any $t \in [0, n/2]$ and any estimator $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(y_1, \ldots, y_n, t)$ possibly depending on t, we have

$$\sup_{\mathbb{P}\in\mathcal{P}_\delta^{v_\delta}} \mathbb{P}\left[\big\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\big\|_2 \geq \alpha c_u^{-1} \nu_\delta\, d^{1/2} \left(\frac{t}{n}\right)^{\min\{\delta/(1+\delta),1/2\}}\right] \geq \frac{e^{-2t}}{2},$$

where $c_u \geq \lambda_{\max}(\mathbf{S}_n)$.

Theorem 2 reveals that root-n consistency with exponential concentration is impossible when $\delta \in (0, 1)$. It widens the phenomenon observed in Theorem 3.1 in Devroye et al. (2016) for estimating a mean. In addition to the eigenvalue assumption, we need to assume that there exists a $\boldsymbol{u} \in \mathcal{U}_n \subseteq \mathbb{R}^n$ such that the minimum angle between $n^{-1}\boldsymbol{u}$ and $\boldsymbol{x}^j$ is non-vanishing. This assumption comes from the intuition that the linear subspace spanned by $\boldsymbol{x}^j$ is at most of rank d and thus cannot span the whole space $\mathbb{R}^n$. This assumption naturally holds in the univariate case where $\mathbf{X} = (1, \ldots, 1)^{\mathrm{T}}$ and we can take $\boldsymbol{u} = (1, \ldots, 1)^{\mathrm{T}}$ and $\alpha = 1$. More generally, $\|\mathbf{X}^{\mathrm{T}}\boldsymbol{u}/n\|_{\min} = \min\{|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}^1|/n, \ldots, |\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}^d|/n\}$. Taking $|\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}^1|/n$ for an example, since $\boldsymbol{u} \in \{-1, +1\}^n$, we can assume that each coordinate of $\boldsymbol{x}^1$ is positive. In this case, $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}^1/n = \sum_{i=1}^n |x_i^1|/n \geq \min_i |x_i^1|$, which is strictly positive with probability one, assuming $\boldsymbol{x}^1$ is drawn from a continuous distribution.

Together, the upper and lower bounds show that the adaptive Huber estimator achieves near-optimal deviations. Moreover, it indicates that the Huber estimator with an adaptive τ exhibits a sharp phase transition: when $\delta \geq 1$, $\widehat{\boldsymbol{\beta}}_\tau$ converges to $\boldsymbol{\beta}^*$ at the parametric rate $n^{-1/2}$, while only a slower rate of order $n^{-\delta/(1+\delta)}$ is available when the second moment does not exist.

Remark 4. We provide a parallel analysis under random designs in the supplementary material. Beyond the nonasymptotic deviation bounds, we also prove a nonasymptotic Bahadur representation, which establishes a linear approximation of the nonlinear robust estimator. This result paves the way for future research on conducting statistical inference and constructing confidence sets under heavy-tailedness. Additionally, the proposed estimator achieves full efficiency: it is as efficient as the ordinary least squares estimator asymptotically, while the robustness is characterized via nonasymptotic performance.

3.2 Adaptive Huber Regression in High Dimensions

In this section, we study the regularized adaptive Huber estimator in high dimensions where d is allowed to grow with the sample size n exponentially. The analysis is carried out under fixed designs, and results for random designs are again provided in the supplementary material. We start with a modified version of the localized restricted eigenvalue introduced by Fan et al. (2018). Let $\mathbf{H}_\tau(\boldsymbol{\beta}) = \nabla^2 \mathcal{L}_\tau(\boldsymbol{\beta})$ denote the Hessian matrix. Recall that $\mathcal{S} = \operatorname{supp}(\boldsymbol{\beta}^*) \subseteq \{1, \ldots, d\}$ is the true support set with $|\mathcal{S}| = s$.

Definition 2 (Localized Restricted Eigenvalue, LRE)**.** The localized restricted eigenvalue of $\mathbf{H}_\tau$ is defined as

$$\kappa_+(m, \gamma, r) = \sup \Big\{ \langle \boldsymbol{u}, \mathbf{H}_\tau(\boldsymbol{\beta}) \boldsymbol{u} \rangle : (\boldsymbol{u}, \boldsymbol{\beta}) \in \mathcal{C}(m, \gamma, r) \Big\},$$
$$\kappa_-(m, \gamma, r) = \inf \Big\{ \langle \boldsymbol{u}, \mathbf{H}_\tau(\boldsymbol{\beta}) \boldsymbol{u} \rangle : (\boldsymbol{u}, \boldsymbol{\beta}) \in \mathcal{C}(m, \gamma, r) \Big\},$$

where $\mathcal{C}(m, \gamma, r) := \{ (\boldsymbol{u}, \boldsymbol{\beta}) \in \mathbb{S}^{d-1} \times \mathbb{R}^d : \forall J \subseteq \{1, \ldots, d\}$ satisfying $S \subseteq J, |J| \leq m, \|\boldsymbol{u}_{J^c}\|_1 \leq \gamma \|\boldsymbol{u}_J\|_1, \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \leq r \}$ is a local ℓ_1-cone.

The LRE is defined in a local neighborhood of $\boldsymbol{\beta}^*$ under ℓ_1-norm. This facilitates our proof, while Fan et al. (2018) use the ℓ_2-norm.

Condition 2. $\mathbf{H}_\tau$ satisfies the localized restricted eigenvalue condition $\mathrm{LRE}(k, \gamma, r)$, that is, $\kappa_l \leq \kappa_-(k, \gamma, r) \leq \kappa_+(k, \gamma, r) \leq \kappa_u$ for some constants $\kappa_u, \kappa_l > 0$.

The condition above is referred to as the LRE condition (Fan et al., 2018). It is a unified condition for studying generalized loss functions, whose Hessians may possibly depend on $\boldsymbol{\beta}$. For Huber loss, Condition 2 also involves the observation noise. The following definition concerns the restricted eigenvalues of $\mathbf{S}_n$ instead of $\mathbf{H}_\tau$.

Definition 3 (Restricted Eigenvalue, RE)**.** The restricted maximum and minimum eigenvalues of $\mathbf{S}_n$ are defined respectively as

$$\rho_+(m, \gamma) = \sup_{\boldsymbol{u}} \big\{ \langle \boldsymbol{u}, \mathbf{S}_n \boldsymbol{u} \rangle : \boldsymbol{u} \in \mathcal{C}(m, \gamma) \big\},$$
$$\rho_-(m, \gamma) = \inf_{\boldsymbol{u}} \big\{ \langle \boldsymbol{u}, \mathbf{S}_n \boldsymbol{u} \rangle : \boldsymbol{u} \in \mathcal{C}(m, \gamma) \big\},$$

where $\mathcal{C}(m, \gamma) := \{ \boldsymbol{u} \in \mathbb{S}^{d-1} : \forall J \subseteq \{1, \ldots, d\}$ satisfying $S \subseteq J, |J| \leq m, \|\boldsymbol{u}_{J^c}\|_1 \leq \gamma \|\boldsymbol{u}_J\|_1 \}$.

Condition 3. $\mathbf{S}_n$ satisfies the restricted eigenvalue condition $\mathrm{RE}(k,\gamma)$, that is, $\kappa_l \leq \rho_-(k,\gamma) \leq \rho_+(k,\gamma) \leq \kappa_u$ for some constants $\kappa_u, \kappa_l > 0$.

To make Condition 2 on $\mathbf{H}_\tau$ practically useful, in what follows, we show that Condition 3 implies Condition 2 with high probability. As before, we write $v_\delta = n^{-1}\sum_{i=1}^n v_{i,\delta}$ and $L = \max_{1\leq i\leq n}\|\boldsymbol{x}_i\|_\infty$.

Lemma 1. Condition 3 implies Condition 2 with high probability: if $0 < \kappa_l \leq \rho_-(k,\gamma) \leq \rho_+(k,\gamma) \leq \kappa_u < \infty$ for some $k \geq 1$ and $\gamma > 0$, then it holds with probability at least $1-e^{-t}$ that, $0 < \kappa_l/2 \leq \kappa_-(k,\gamma,r) \leq \kappa_+(k,\gamma,r) \leq \kappa_u < \infty$ provided $\tau \geq \max\{8Lr, c_1(L^2kv_\delta)^{1/(1+\delta)}\}$ and $n \geq c_2L^4k^2t$, where $c_1, c_2 > 0$ are constants depending only on (γ, κ_l).

With the above preparations in place, we are now ready to present the main results on the adaptive Huber estimator in high dimensions.

Theorem 3 (Upper Bound in High Dimensions)**.** Assume Condition 3 holds with $(k,\gamma) = (2s,3)$, $v_\delta < \infty$ for some $0 < \delta \leq 1$. For any $t > 0$ and $\tau_0 \geq \nu_\delta$, let $\tau = \tau_0(n/t)^{\max\{1/(1+\delta),1/2\}}$ and $\lambda \geq 4L\tau_0(t/n)^{\min\{\delta/(1+\delta),1/2\}}$. Then with probability at least $1-(2s+1)e^{-t}$, the ℓ_1-regularized Huber estimator $\widehat{\boldsymbol{\beta}}_{\tau,\lambda}$ defined in (4) satisfies

$$\left\|\widehat{\boldsymbol{\beta}}_{\tau,\lambda} - \boldsymbol{\beta}^*\right\|_2 \leq 3\kappa_l^{-1}s^{1/2}\lambda, \tag{9}$$

as long as $n \geq C(L,\kappa_l)s^2t$ for some $C(L,\kappa_l)$ depending only on (L,κ_l). In particular, with $t = (1+c)\log d$ for $c > 0$ we have

$$\left\|\widehat{\boldsymbol{\beta}}_{\tau,\lambda} - \boldsymbol{\beta}^*\right\|_2 \lesssim \kappa_l^{-1}L\tau_0\, s^{1/2}\left\{\frac{(1+c)\log d}{n}\right\}^{\min\{\delta/(1+\delta),1/2\}} \tag{10}$$

with probability at least $1-d^{-c}$.

The above result demonstrates that the regularized Huber estimator with an adaptive robustification parameter converges at the rate $s^{1/2}\{(\log d)/n\}^{\min\{\delta/(1+\delta),1/2\}}$ with overwhelming probability. Provided the observation noise has finite variance, the proposed estimator performs as well as the Lasso with sub-Gaussian errors. We advocate the adaptive Huber regression method since sub-Gaussian condition often fails in practice (Wang, Peng and Li, 2015; Eklund, Nichols and Knutsson, 2016).

Remark 5. As pointed out by a reviewer, if one pursues a sparsity-adaptive approach, such as the SLOPE (Bogdan et al., 2015; Bellec et al., 2018), the upper bound on ℓ_2-error can be improved from $\sqrt{s\log(d)/n}$ to $\sqrt{s\log(ed/s)/n}$. With heavy-tailed observation noise, it is interesting to investigate whether this sharper bound can be achieved by Huber-type regularized estimator. We leave this to future work as a significant amount of additional work is still needed. On the other hand, since $\log(ed/s) = 1 + \log d - \log s$ and $s \leq n$, $\log(ed/s)$ scales the same as $\log d$ so long as $\log d > a \log n$ for some $a > 1$.

Remark 6. Analogously to the low dimensional case, here we impose the sample size scaling $n \gtrsim s^2 \log d$ under fixed designs. In the supplementary material, we obtain minimax optimal ℓ_1-, ℓ_2- and prediction error bounds for $\widehat{\boldsymbol{\beta}}_{\tau,\lambda}$ with random designs under the scaling $n \gtrsim s \log d$.

Finally, we establish a matching lower bound for estimating $\boldsymbol{\beta}^*$. Recall the definition of $\mathcal{U}_n$ in Theorem 2.

Theorem 4 (Lower Bound in High Dimensions)**.** Assume that ε_i are independent from some distribution in $\mathcal{P}_\delta^{v_\delta}$. Suppose that Condition 3 holds with $k = 2s$ and $\gamma = 0$. Further assume that there exists a set $\mathcal{A}$ with $|\mathcal{A}| = s$ and $\mathbf{u} \in \mathcal{U}_n$ such that $\|\mathbf{X}_{\mathcal{A}}^{\mathrm{T}}\mathbf{u}/n\|_{\min} \geq \alpha$ for some $\alpha > 0$. Then, for any $A > 0$ and s-sparse estimator $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(y_1, \ldots, y_n, A)$ possibly depending on A, we have

$$\sup_{\mathbb{P} \in \mathcal{P}_\delta^{v_\delta}} \mathbb{P}\left[\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \geq \nu_\delta \frac{\alpha s^{1/2}}{\kappa_u} \left(\frac{A \log d}{2n} \right)^{\min\{\delta/(1+\delta), 1/2\}} \right] \geq 2^{-1} d^{-A},$$

as long as $n \geq 2(A \log d + \log 2)$.

Together, Theorems 3 and 4 show that the regularized adaptive Huber estimator achieves the optimal rate of convergence in ℓ_2-error. The proof, which is given in the supplementary material, involves constructing a sub-class of binomial distributions for the regression errors. Unifying the results in low and high dimensions, we arrive at the claim (5) and thus the phase transition in Figure 1.

4 Extension to Heavy-tailed Designs

In this section, we extend the idea of adaptive Huber regression described in Section 2 to the case where both the covariate vector $\boldsymbol{x}$ and the regression error ε exhibit heavy tails. We focus on the high dimensional regime $d \gg n$, where $\boldsymbol{\beta}^* \in \mathbb{R}^d$ is sparse with $s = \|\boldsymbol{\beta}^*\|_0 \ll n$. Observe that, for Huber regression, the linear part of the Huber loss penalizes the residuals, and therefore robustifies the quadratic loss in the sense that outliers in the response space (caused by heavy-tailed observation noise) are down weighted or removed. Since no robustification is imposed on the covariates, intuitively, the adaptive Huber estimator may not be robust against heavy-tailed covariates. In what follows, we modify the adaptive Huber regression to robustify both the covariates and regression errors.

To begin with, suppose we observe independent data $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ from $(y, \boldsymbol{x})$, which follows the linear model $y = \langle \boldsymbol{x}, \boldsymbol{\beta}^* \rangle + \varepsilon$. To robustify $\boldsymbol{x}_i$, we define truncated covariates $\boldsymbol{x}_i^{\varpi} = (\psi_{\varpi}(x_{i1}), \ldots, \psi_{\varpi}(x_{id}))^{\mathrm{T}}$, where $\psi_{\varpi}(x) := \min\{\max(-\varpi, x), \varpi\}$ and $\varpi > 0$ is a tuning parameter. Then we consider the modified adaptive Huber estimator (see Fan et al. (2016) for a general robustification principle)

$$\widehat{\boldsymbol{\beta}}_{\tau,\varpi,\lambda} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \big\{ \mathcal{L}_\tau^{\varpi}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1 \big\}, \tag{11}$$

where $\mathcal{L}_\tau^{\varpi}(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^n \ell_\tau(y_i - \langle \boldsymbol{x}_i^{\varpi}, \boldsymbol{\beta} \rangle)$ and $\lambda > 0$ is a regularization parameter.

Let $\mathcal{S}$ be the true support of $\boldsymbol{\beta}^*$ with sparsity $|\mathcal{S}| = s$, and denote by $\mathbf{H}_\tau^{\varpi}(\boldsymbol{\beta}) =$

$\nabla^2 \mathcal{L}_\tau^\varpi(\boldsymbol{\beta})$ the Hessian matrix of the modified Huber loss. To investigate the deviation property of $\widehat{\boldsymbol{\beta}}_{\tau,\varpi,\lambda}$, we impose the following mild moment assumptions.

Condition 4. (i) $\mathbb{E}(\varepsilon) = 0$, $\sigma^2 = \mathbb{E}(\varepsilon^2) > 0$ and $v_3 := \mathbb{E}(\varepsilon^4) < \infty$; (ii) The covariate vector $\boldsymbol{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$ is independent of ε and satisfies $M_4 := \max_{1 \le j \le d} \mathbb{E}(x_j^4) < \infty$.

We are now in place to state the main result of this section. Theorem 5 below demonstrates that the modified adaptive Huber estimator admits exponentially fast concentration when the convariates only have finite fourth moments, although at the cost of stronger scaling conditions.

Theorem 5. Assume Condition 4 holds and let $\mathbf{H}_\tau^\varpi(\cdot)$ satisfy Condition 2 with $k = 2s$, $\gamma = 3$ and $r > 12\kappa_l^{-1}\lambda s$. Then, the modified adaptive Huber estimator $\widehat{\boldsymbol{\beta}}_{\tau,\varpi,\lambda}$ given in (11) satisfies, on the event $\mathcal{E}(\tau,\varpi,\lambda) = \big\{\|(\nabla \mathcal{L}_\tau^\varpi(\boldsymbol{\beta}^*))_{\mathcal{S}}\|_\infty \le \lambda/2\big\}$, that

$$\big\|\widehat{\boldsymbol{\beta}}_{\tau,\varpi,\lambda} - \boldsymbol{\beta}^*\big\|_2 \le 3\kappa_l^{-1} s^{1/2} \lambda.$$

For any $t > 0$, let the triplet (τ, ϖ, λ) satisfy

$$\lambda \ge 2M_4 \|\boldsymbol{\beta}^*\|_2\, s^{1/2} \varpi^{-2} + 8\big\{v_2 M_2^{1/2} + M_4 \|\boldsymbol{\beta}^*\|_2^3\, s^{3/2}\big\}\tau^{-2} \\ + 2\big(2\sigma^2 M_2 + 2M_4 \|\boldsymbol{\beta}^*\|_2^2\, s\big)^{1/2}\sqrt{\frac{t}{n}} + \varpi\tau\frac{t}{n}, \tag{12}$$

where $v_2 = \mathbb{E}(|\varepsilon|^3)$ and $M_2 = \max_{1 \le j \le d} \mathbb{E}(x_j^2)$. Then $\mathbb{P}\{\mathcal{E}(\tau,\varpi,\lambda)\} \ge 1 - 2se^{-t}$.

Remark 7. Assume that the quantities v_3, M_4 and $\|\boldsymbol{\beta}^*\|_2$ are all bounded. Taking $t \asymp \log d$ in (12), we see that $\widehat{\boldsymbol{\beta}}_{\tau,\varpi,\lambda}$ achieves a near-optimal convergence rate of order $s\sqrt{(\log d)/n}$ when the parameters (τ, ϖ, λ) scale as

$$\tau \asymp s^{1/2}\bigg(\frac{n}{\log d}\bigg)^{1/4}, \quad \varpi \asymp \bigg(\frac{n}{\log d}\bigg)^{1/4} \quad \text{and} \quad \lambda \asymp \sqrt{\frac{s \log d}{n}}.$$

We remark here that the theoretically optimal τ is different from that in the sub-Gaussian design case. See Theorem B.2 in the supplementary material.

5 Algorithm and Implementation

This section is devoted to computational algorithm and numerical implementation. We focus on the regularized adaptive Huber regression in (4), as (3) can be easily solved via the iteratively reweighted least squares method. To solve the convex optimization problem in (4), standard optimization algorithms, such as the cutting-plane or interior point method, are not scalable to large-scale problems.

In what follows, we describe a fast and easily implementable method using the local adaptive majorize-minimization (LAMM) principle (Fan et al., 2018). We say that a function $g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)})$ majorizes $f(\boldsymbol{\beta})$ at the point $\boldsymbol{\beta}^{(k)}$ if

$$g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)}) \geq f(\boldsymbol{\beta}) \quad \text{and} \quad g(\boldsymbol{\beta}^{(k)}|\boldsymbol{\beta}^{(k)}) = f(\boldsymbol{\beta}^{(k)}).$$

To minimize a general function $f(\boldsymbol{\beta})$, a majorize-minimization (MM) algorithm initializes at $\boldsymbol{\beta}^{(0)}$, and then iteratively computes $\boldsymbol{\beta}^{(k+1)} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)})$ for $k = 0, 1, \ldots$. The objective value of such an algorithm decreases in each step, since

$$f(\boldsymbol{\beta}^{(k+1)}) \overset{\text{major.}}{\leq} g(\boldsymbol{\beta}^{(k+1)} \,|\, \boldsymbol{\beta}^{(k)}) \overset{\text{min.}}{\leq} g(\boldsymbol{\beta}^{(k)} \,|\, \boldsymbol{\beta}^{(k)}) \overset{\text{init.}}{=} f(\boldsymbol{\beta}^{(k)}). \tag{13}$$

As pointed out by Fan et al. (2018), the majorization requirement only needs to hold locally at $\boldsymbol{\beta}^{(k+1)}$ when starting from $\boldsymbol{\beta}^{(k)}$. We therefore locally majorize $\mathcal{L}_\tau(\boldsymbol{\beta})$ in (4) at $\boldsymbol{\beta}^{(k)}$ by an isotropic quadratic function

$$g_k(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)}) = \mathcal{L}_\tau(\boldsymbol{\beta}^{(k)}) + \left\langle \nabla\mathcal{L}_\tau(\boldsymbol{\beta}^{(k)}),\, \boldsymbol{\beta} - \boldsymbol{\beta}^{(k)} \right\rangle + \frac{\phi_k}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(k)}\|_2^2,$$

where ϕ_k is a quadratic parameter such that $g_k(\boldsymbol{\beta}^{(k+1)}|\boldsymbol{\beta}^{(k)}) \geq \mathcal{L}_\tau(\boldsymbol{\beta}^{(k+1)})$. The

Algorithm 1 LAMM algorithm for regularized adaptive Huber regression.

```
**Algorithm**: $\{\boldsymbol{\beta}^{(k)}, \phi_k\}_{k=1}^{\infty} \leftarrow \mathrm{LAMM}(\lambda, \boldsymbol{\beta}^{(0)}, \phi_0, \epsilon\ )$
**Input**: $\lambda, \boldsymbol{\beta}^{(0)}, \phi_0, \epsilon$
**Initialize**: $\phi^{(\ell,k)} \leftarrow \max\{\phi_0, \gamma_u^{-1}\phi^{(\ell,k-1)}\}$
**for** $k = 0, 1, \ldots$ until $\|\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}\|_2 \le \epsilon$ **do**
    **Repeat**
        $\boldsymbol{\beta}^{(k+1)} \leftarrow T_{\lambda,\phi_k}(\boldsymbol{\beta}^{(k)})$
        **If** $g_k(\boldsymbol{\beta}^{(k+1)}|\boldsymbol{\beta}^{(k)}) < \mathcal{L}_\tau(\boldsymbol{\beta}^{(k+1)})$ **then** $\phi_k \leftarrow \gamma_u\phi_k$
    **Until** $g_k(\boldsymbol{\beta}^{(k+1)}|\boldsymbol{\beta}^{(k)}) \ge \mathcal{L}_\tau(\boldsymbol{\beta}^{(k+1)})$
    **Return** $\{\boldsymbol{\beta}^{(k+1)}, \phi_k\}$
**end for**
**Output**: $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(k+1)}$
```

isotropic form also allows a simple analytic solution to the subsequent majorized optimization problem:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^d}\left\{\left\langle\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^{(k)}), \boldsymbol{\beta} - \boldsymbol{\beta}^{(k)}\right\rangle + \frac{\phi_k}{2}\left\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(k)}\right\|_2^2 + \lambda\left\|\boldsymbol{\beta}\right\|_1\right\}. \tag{14}$$

It can be shown that (14) is minimized at

$$\boldsymbol{\beta}^{(k+1)} = T_{\lambda,\phi_k}(\boldsymbol{\beta}^{(k)}) = S\Big(\boldsymbol{\beta}^{(k)} - \phi_k^{-1}\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^{(k)}), \phi_k^{-1}\lambda\Big),$$

where $S(\mathbf{x}, \lambda)$ is the soft-thresholding operator defined by $S(\mathbf{x}, \lambda) = \mathrm{sign}(x_j)\max(|x_j| - \lambda, 0)$. The simplicity of this updating rule is due to the fact that (14) is an unconstrained optimization problem.

To find the smallest ϕ_k such that $g_k(\boldsymbol{\beta}^{(k+1)}|\boldsymbol{\beta}^{(k)}) \ge \mathcal{L}_\tau(\boldsymbol{\beta}^{(k+1)})$, the basic idea of LAMM is to start from a relatively small isotropic parameter $\phi_k = \phi_k^0$ and then successfully inflate ϕ_k by a factor $\gamma_u > 1$, say $\gamma_u = 2$. If the solution satisfies $g_k(\boldsymbol{\beta}^{(k+1)}|\boldsymbol{\beta}^{(k)}) \ge \mathcal{L}_\tau(\boldsymbol{\beta}^{(k+1)})$, we stop and obtain $\boldsymbol{\beta}^{(k+1)}$, which makes the target value non-increasing. We then continue with the iteration to produce next solution until the solution sequence $\{\boldsymbol{\beta}^{(k)}\}_{k=1}^{\infty}$ converges. A simple stopping criterion is $\|\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}\|_2 \le \epsilon$ for a sufficiently small ϵ, say 10^{-4}. We refer to Fan et al. (2018) for a detailed complexity analysis of the LAMM algorithm.

Table 1: Results for adaptive Huber regression (AHR) and ordinary least squares (OLS) when $n = 100$ and $d = 5$. The mean and standard deviation (std) of ℓ_2-error based on 100 simulations are reported.

Noise	AHR		OLS	
	mean	std	mean	std
Normal	0.566	0.189	0.567	0.191
Student's t	0.806	0.651	1.355	2.306
Log-normal	3.917	3.740	8.529	13.679

6 Numerical Studies

6.1 Tuning Parameter and Finite Sample Performance

For numerical studies and real data analysis, in the case where the actual order of moments is unspecified, we presume the variance is finite and therefore choose robustification and regularization parameters as follows:

$$\tau = c_\tau \times \widehat{\sigma} \left(\frac{n_{\text{eff}}}{t}\right)^{1/2} \quad \text{and} \quad \lambda = c_\lambda \times \widehat{\sigma} \left(\frac{n_{\text{eff}}}{t}\right)^{1/2},$$

where $\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^n (y_i - \bar{y})^2$ with $\bar{y} = n^{-1} \sum_{i=1}^n y_i$ serves as a crude preliminary estimate of σ^2, and the parameter t controls the confidence level. We set $t = \log n$ for simplicity except for the phase transition plot. The constant c_τ and c_λ are chosen via 3-fold cross-validation from a small set of constants, say $\{0.5, 1, 1.5\}$.

We generate data from the linear model

$$y_i = \langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle + \varepsilon_i, \quad i = 1, \ldots, n, \tag{15}$$

where ε_i are i.i.d. regression errors and $\boldsymbol{\beta}^* = (5, -2, 0, 0, 3, \underbrace{0, \ldots, 0}_{d-5})^{\mathrm{T}} \in \mathbb{R}^d$. Independent of ε_i, we generate $\boldsymbol{x}_i$ from standard multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. In this section, we set $(n, d) = (100, 5)$, and generate regression errors from three different distributions: the normal distribution $\mathcal{N}(0, 4)$, the t-distribution with degrees of freedom 1.5, and the log-normal distribution $\log \mathcal{N}(0,4)$. Both t and log-normal distributions are heavy-tailed, and produce outliers with high chance.

The results on ℓ_2-error for adaptive Huber regression and the least squares estimator, averaged over 100 simulations, are summarized in Table 1. In the case of normally distributed noise, the adaptive Huber estimator performs as well as the least squares. With heavy-tailed regression errors following Student's t or log-normal distribution, the adaptive Huber regression significantly outperforms the least squares. These empirical results reveal that adaptive Huber regression prevails across various scenarios: not only it provides more reliable estimators in the presence of heavy-tailed and/or asymmetric errors, but also loses almost no efficiency at the normal model.

6.2 Phase Transition

In this section, we validate the phase transition behavior of $\|\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*\|_2$ empirically. We generate continuous responses according to (15), where $\boldsymbol{\beta}^*$ and $\boldsymbol{x}_i$ are set the same way as before. We sample independent errors as $\varepsilon_i \sim t_{\mathrm{df}}$, Student's t-distribution with df degrees of freedom. Note that t_{df} has finite $(1+\delta)$-th moments provided $\delta < \mathrm{df} - 1$ and infinite df-th moment. Therefore, we take $\delta = \mathrm{df} - 1 - 0.05$ throughout.

In low dimensions, we take $(n, d) = (500, 5)$ and a sequence of degrees of freedoms (df's): $\mathrm{df} \in \{1.1, 1.2, \ldots, 3.0\}$; in high dimensions, we take $(n, d) = (500, 1000)$, with the same choice of df's. Tuning parameters (τ, λ) are calibrated similarly as before. Indicated by the main theorems, it holds

1. (Low dimension):

$$-\log\left(\|\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*\|_2\right) \asymp \frac{\delta}{1+\delta}\log(n) - \frac{1}{1+\delta}\log(v_\delta), \quad 0 < \delta \le 1,$$

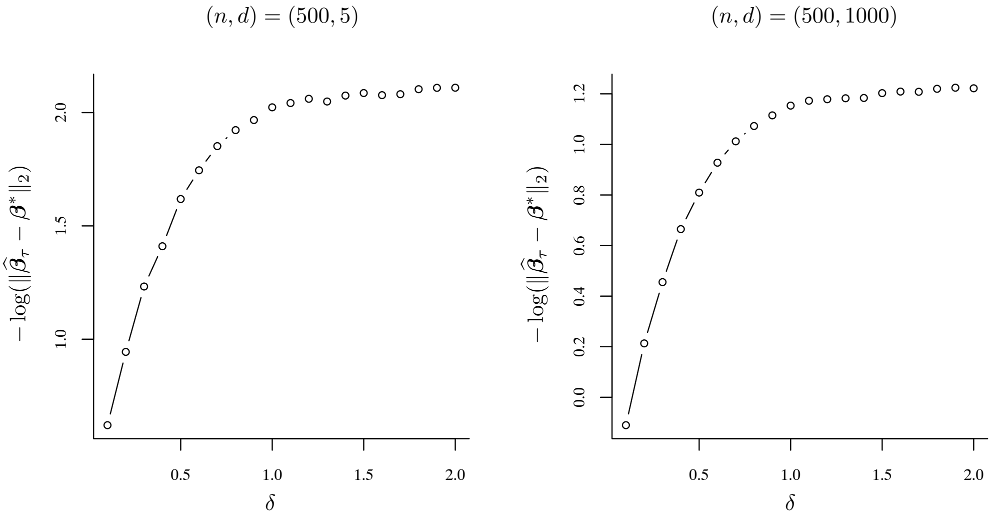

Figure 2: Negative log ℓ_2-error versus δ in low (left panel) and high (right panel) dimensions.

2. (High dimension):

$$-\log\left(\|\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*\|_2\right) \asymp \frac{\delta}{1+\delta}\log\left(\frac{n}{\log d}\right) - \frac{1}{1+\delta}\log(v_\delta), \quad 0 < \delta \le 1,$$

which are approximately $\log(n) \times \delta/(1+\delta)$ and $\log(n/\log d) \times \delta/(1+\delta)$, respectively, when n is sufficiently large.

Figure 2 displays the negative log ℓ_2-error versus δ in both low and high dimensions over 200 repetitions for each (n, d) combination. The empirically fitted curve closely resembles the theoretical curve displayed in Figure 1. These numerical results are in line with the theoretical findings, and empirically validate the phase transition of the adaptive Huber estimator.

We also compared the ℓ_2-error of the adaptive Huber estimator with that of the OLS estimator for t-distributed errors with varying degrees of freedoms. As shown in Figure 3, adaptive Huber exhibits a significant advantage especially when δ is small. The OLS slowly catches up as δ increases.

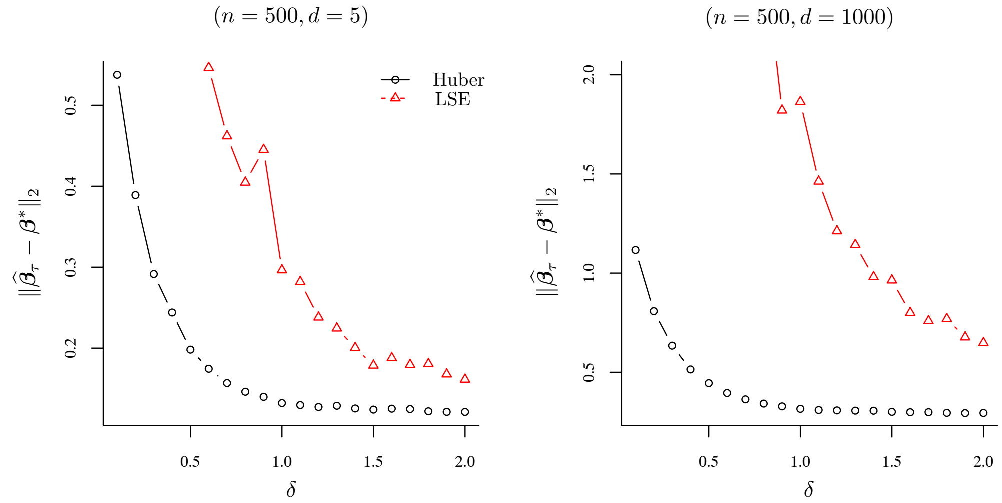

Figure 3: Comparison between the (regularized) adaptive Huber estimator and the (regularized) least squares estimator under ℓ_2-error.

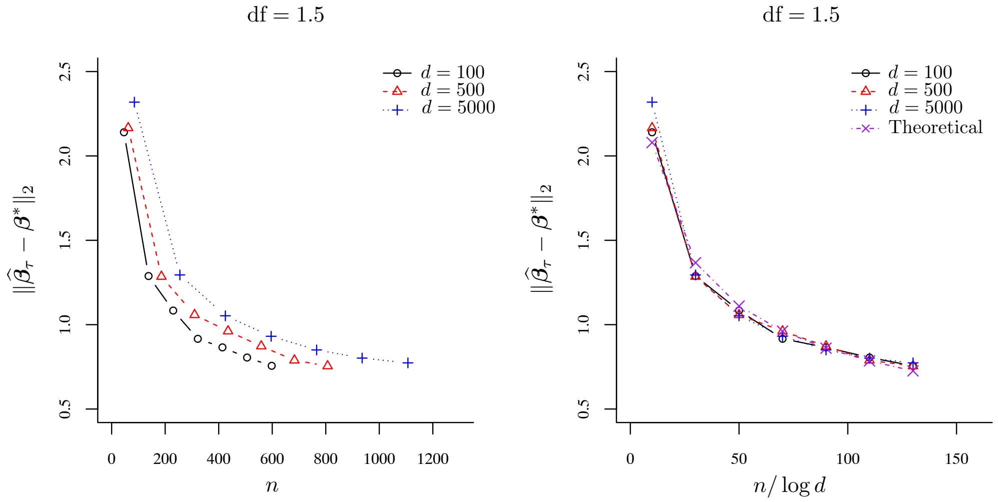

Figure 4: The ℓ_2-error versus sample size n (left panel) and the ℓ_2-error versus effective sample size $n_{\mathrm{eff}} = n/\log d$ (right panel).

6.3 Effective Sample Size

In this section, we verify the scaling behavior of $\|\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*\|_2$ with respect to the effective sample size. The data are generated in the same way as before except that the errors are drawn from $t_{1.5}$. As discussed in the previous subsection, we take $\delta = 0.45$ and then choose the robustification parameter as $\tau = c_\tau \widehat{v}_\delta (n/\log d)^{1/(1+\delta)}$, where $\widehat{v}_\delta$ is the $(1+\delta)$-th sample absolute central moment. For simplicity, we take $c_\tau = 0.5$ here since our goal is to demonstrate the scaling behavior as n grows, instead of to achieve the best finite-sample performance.

The left panel of Figure 4 plots the ℓ_2-error $\|\widehat{\boldsymbol{\beta}}_{\tau,\lambda} - \boldsymbol{\beta}^*\|_2$ versus sample size over 200 repetitions when the dimension $d \in \{100, 500, 5000\}$. In all three settings, the ℓ_2-error decays as the sample size grows. As expected, the curves shift to the right when the dimension increases. Theorem 3 provides a specific prediction about this scaling behavior: if we plot the ℓ_2-error versus effective sample size $(n/\log d)$, the curves should align roughly with the theoretical curve

$$\|\widehat{\boldsymbol{\beta}}_{\tau,\lambda} - \boldsymbol{\beta}^*\|_2 \asymp \left(\frac{n}{\log d}\right)^{-\delta/(1+\delta)}$$

for different values of d. This is validated empirically by the right panel of Figure 4. This near-perfect alignment in Figure 4 is also observed by Wainwright (2009) for Lasso with sub-Gaussian errors.

6.4 A Real Data Example: NCI-60 Cancer Cell Lines

We apply the proposed methodologies to the NCI-60, a panel of 60 diverse human cancel cell lines. The NCI-60 consists of data on 60 human cancer cell lines and can be downloaded from http://discover.nci.nih.gov/cellminer/. More details on data acquisition can be found in Shankavaram et al. (2007). Our aim is to investigate the effects of genes on protein expressions. The gene expression data were obtained with an Affymetrix HG-U133A/B chip, $\log_2$ transformed and normalized with the guanine dytosine robust multi-array analysis. We then combined the same gene expression variables measured by multiple different probes into one by taking their median, resulting in a set of $p = 17,924$ predictors. The protein expressions based on 162 antibodies were acquired via reverse-phase protein lysate arrays in their

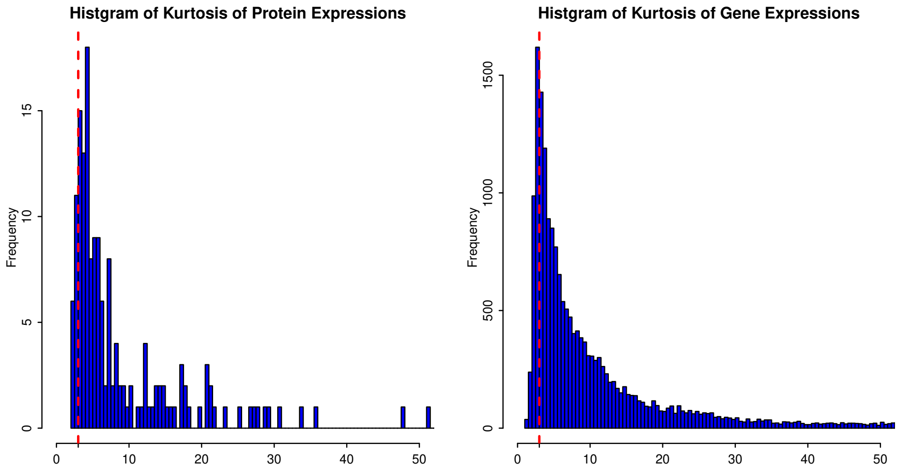

Figure 5: Histogram of kurtosises for the protein and gene expressions. The dashed red line at 3 is the kurtosis of a normal distribution.

original scale. One observation had to be removed since all values were missing in the gene expression data, reducing the number of observations to $n = 59$.

We first center all the protein and gene expression variables to have mean zero, and then plot the histograms of the kurtosises of all expressions in Figure 5. The left panel in the figure shows that, 145 out of 162 protein expressions have kurtosises larger than 3; and 49 larger than 9. In other words, more than 89.5% of the protein expression variables have tails heavier than the normal distribution, and about 30.2% are severely heavy-tailed with tails flatter than t_5, the t-distribution with 5 degrees of freedom. Similarly, about 36.5% of the gene expression variables, even after the $\log_2$-transformation, still exhibit empirical kurtosises larger than that of t_5. This suggests that, regardless of the normalization methods used, genomic data can still exhibit heavy-tailedness, which was also pointed out by Purdom and Holmes (2005).

We order the protein expression variables according to their scales, measured by the standard deviation. We show the results for the protein expressions based on the KRT19 antibody, the protein keratin 19, which constitutes the variable with the largest standard deviation, serving as one dependent variable. KRT19, a type I keratin, also known as Cyfra 21-1, is encoded by the *KRT19* gene. Due to its high sensitivity, the KRT19 antibody is the most used marker for the tumor cells disseminated in lymph nodes, peripheral blood, and bone marrow of breast cancer patients (Nakata et al., 2004). We denote the adaptive Huber regression as AHuber, and that with truncated covariates as TAHuber. We then compare AHuber and TAHuber with Lasso. Both regularization and robustification parameters are chosen by the ten-fold cross-validation.

To measure the predictive performance, we consider a robust prediction loss: the mean absolute error (MAE) defined as

$$\mathrm{MAE}\big(\widehat{\boldsymbol{\beta}}\big) = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \big| y_i^{\mathrm{test}} - \langle \boldsymbol{x}_i^{\mathrm{test}}, \widehat{\boldsymbol{\beta}} \rangle \big|,$$

where y_i^{test} and $\boldsymbol{x}_i^{\mathrm{test}}$, $i = 1, \ldots, n_{\mathrm{test}}$, denote the observations of the response and predictor variables in the test data, respectively. We report the MAE via the leave-one-out cross-validation. Table 2 reports the MAE, model size and selected genes for the considered methods. TAHuber clearly shows the smallest MAE, followed by AHuber and Lasso. The Lasso produces a fairly large model despite the small sample. Now it has been recognized that Lasso tends to select many noise variables along with the significant ones, especially when data exhibit heavy tails.

The Lasso selects a model with 42 genes but excludes the *KRT19* gene, which encodes the protein keratin 19. AHuber finds 11 genes including *KRT19*. TAHuber results in a model with 7 genes: *KRT19, MT1E, ARHGAP29, MALL, ANXA3, MAL2, BAMBI*. First, *KRT19* encodes the keratin 19 protein. It has been reported in Wu et al. (2008) that the *MT1E* expression is positively correlated with cancer cell migration and tumor stage, and the *MT1E* isoform was found to be present in estrogen receptor-negative breast cancer cell lines (Friedline et al., 1998). *ANXA3* is highly expressed in all colon cell lines and all breast-derived cell lines positive for the oestrogen receptor (Ross et al., 2000). A very recent study in Zhou et al. (2017) suggested that silencing the *ANXA3* expression by RNA interference inhibits the proliferation

Table 2: We report the mean absolute error (MAE) for protein expressions based on the KRT19 antibody from the NCI-60 cancer cell lines, computed from leave-one-out cross-validation. We also report the model size and selected genes for each method.

Method	MAE	Size	Selected Genes
Lasso	7.64	42	*FBLIM1, MT1E, EDN2, F3, FAM102B, S100A14, LAMB3, EPCAM, FN1, TM4SF1, UCHL1, NMU, ANXA3, PLAC8, SPP1, TGFBI, CD74, GPX3, EDN1, CPVL, NPTX2, TES, AKR1B10, CA2, TSPYL5, MAL2, GDA, BAMBI, CST6, ADAMTS15, DUSP6, BTG1, LGALS3, IFI27, MEIS2, TOX3, KRT23, BST2, SLPI, PLTP, XIST, NGFRAP1*
AHuber	6.74	11	*MT1E, ARHGAP29, CPCAM, VAMP8, MALL, ANXA3, MAL2, BAMBI, LGALS3, KRT19, TFF3*
TAHuber	5.76	7	*MT1E, ARHGAP29, MALL, ANXA3, MAL2, BAMBI, KRT19*

and invasion of breast cancer cells. Moreover, studies in Shangguan et al. (2012) and Kretzschmar (2000) showed that the *BAMBI* transduction significantly inhibited TGF-β/Smad signaling and expression of carcinoma-associated fibroblasts in human bone marrow mesenchymal stem cells (BM-MSCs), and disrupted the cytokine network mediating the interaction between MSCs and breast cancer cells. Consequently, the *BAMBI* transduction abolished protumor effects of BM-MSCs in vitro and in an orthotopic breast cancer xenograft model, and instead significantly inhibited growth and metastasis of coinoculated cancer. *MAL2* expressions were shown to be elevated at both RNA and protein levels in breast cancer (Shehata et al., 2008). It has also been shown that *MALL* is associated with various forms of cancer (Oh et al., 2005; Landi et al., 2014). However, the effect of *ARHGAP29* and *MALL* on breast cancer remains unclear and is worth further investigation.

Supplementary Materials

In the supplementary materials, we provide theoretical analysis under random designs, and proofs of all the theoretical results in this paper.

Acknowledgments

The authors thank the Editor, Associate Editor, and two anonymous referees for their valuable comments. This work is supported by a Connaught Award, NSERC Grant RGPIN-2018-06484, NSF Grants DMS-1662139, DMS-1712591, and DMS-1811376, NIH Grant 2R01-GM072611-14, and NSFC Grant 11690014.

References

Alquier, P., Cottet, V. and Lecué, G. (2017). Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions. Preprint. Available at arXiv:1702.01402.

Belloni, A. and Chernozhukov, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, **39** 82–130.

Bellec, P. C., Lecué, G. and Tsybakov, A. B. (2018). Slope meets Lasso: Improved oracle bounds and optimality. *The Annals of Statistics*, **46** 3603–3642.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, **37** 1705–1732.

Bogdan, M., van den Berg, E., Sabatti, C., Su, W. and Candès, E. J. (2015). SLOPE–Adaptive variable selection via convex optimization. *The Annals of Applied Statistics*, **9** 1103–1140.

Brownlees, C., Joly, E. and Lugosi, G. (2015). Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, **43** 2507–2536.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer, Heidelberg.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Annales de I'Institut Henri Poincaré - Probabilités et Statistiques*, **48** 1148–1185.

Catoni, O. (2016). PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. Preprint. Available at arXiv:1603.05229.

Chen, M., Gao, C. and Ren, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *The Annals of Statistics*, **46**, 1932–1960.

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, **1**, 223–236.

Delaigle, A., Hall, P. and Jin, J. (2011). Robustness and accuracy of methods for high dimensional data analysis based on Student's t-statistic. *Journal of the Royal Statistical Society,* Series B, **73** 283–301.

Devroye, L., Lerasle, M., Lugosi, G. and Oliveira, R. I. (2016). Sub-Gaussian mean estimators. *The Annals of Statistics*, **44** 2695–2725.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, **32** 407–499.

Eklund, A., Nichols, T. and Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, **113** 7900–7905.

Fan, J., Fan, Y. and Barut, E. (2014). Adaptive robust variable selection. *The Annals of Statistics*, **42** 324–351.

Fan, J., Li, Q. and Wang, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society,* Series B, **79** 247–265.

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

Fan, J., Liu, H., Sun, Q. and Zhang, T. (2018). I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *The Annals of Statistics*, **96** 1348–1360.

Fan, J., Wang, W. and Zhu, Z. (2016). A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. Available at arXiv:1603.08315.

Friedline, J. A., Garrett, S. H., Somji, S., Todd, J. H. and Sens, D. A. (1998). Differential expression of the MT-1E gene in estrogen-receptor-positive and-negative human breast cancer cell lines. *The American Journal of Pathology*, **152** 23–27.

Giulini, I. (2017). Robust PCA and pairs of projections in a Hilbert space. *Electronic Journal of Statistics*, **11** 3903–3926.

Hastie, T., Tibshirani, R. and Wainwright, M. J. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations.* CRC Press.

He, X. and Shao, Q.-M. (1996). A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *The Annals of Statistics*, **24** 2608–2630.

He, X. and Shao, Q.-M. (2000). On parameters of increasing dimensions. *Journal of Multivariate Analysis*, **73** 120–135.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, **35** 73–101.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, **1** 799–821.

Koenker, R. (2005). *Quantile Regression.* Cambridge University Press, New York.

Kretzschmar, M. (2000) Transforming growth factor-β and breast cancer: transforming growth factor-β/Smad signaling defects and cancer. *Breast Cancer Research*, **2** 107–115.

Landi, A., Vermeire, J., Iannucci, V., Vanderstraeten, H., Naessens, E., Bentahir, M. and Verhasselt, B. (2014). Genome-wide shRNA screening identifies host factors involved in early endocytic events for HIV-1-induced CD4 down-regulation. *Retrovirology*, **11** 118–129.

Lepski, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *IEEE Transactions on Information Theory*, **36** 682–697.

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics*, **18** 405–414.

Liu, R. Y., Parelius, J. M. and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference, (with discussion and a rejoinder by Liu and Singh). *The Annals of Statistics*, **27** 783–858.

Loh, P. and Wainwright, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, **16** 559–616.

Mammen, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *The Annals of Statistics*, **17** 382–400.

Minsker, S. (2018). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics*, **46** 2871–2903.

Mizera, I. (2002). On depth and deep points: A calculus. *The Annals of Statistics*, **30** 1681–1736.

Mizera, I. and Müller, C. H. (2004). Location-scale depth. *Journal of the American Statistical Association*, **99** 949–966.

Nakata, B., Takashima, T., Ogawa, Y., Ishikawa, T. and Hirakawa, K. (2004). Serum CYFRA 21-1 (cytokeratin-19 fragments) is a useful tumour marker for detecting disease relapse and assessing treatment efficacy in breast cancer. *British Journal of Cancer*, **91** 873–878.

Oh, J. H., Yang, J. O., Hahn, Y., Kim, M. R., Byun, S. S., Jeon, Y. J., Kim, J. M., Song, K. S., Noh, S. M., Kim, S. and Yoo, H. S. (2005). Transcriptome analysis of human gastric cancer. *Mammalian Genome*, **16** 942–954.

Portnoy, S. (1985). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large; II. Normal approximation. *The Annals of Statistics*, **13** 1403–1417.

Purdom, E. and Holmes, S. P. (2005). Error distribution for gene expression data. *Statistical Applications in Genetics and Molecular Biology*, **4**: 16.

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, W., Jeffrey, S. S., Van de Rijn, M., Pergamenschikov, A., Lee, J. C. F., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D. and Brown, P. O. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, **24**, 227–235.

Shangguan, L., Ti, X., Krause, U., Hai, B., Zhao, Y., Yang, Z. and Liu, F. (2012). Inhibition of TGF-β/Smad signaling by BAMBI blocks differentiation of human mesenchymal stem cells to carcinoma-associated fibroblasts and abolishes their protumor effects. *Stem Sells*, **30** 2810–2819.

Shankavaram, U. T., Reinhold, W. C., Nishizuka, S., Major, S., Morita, D., Chary, K. K., Reimers, M. A., Scherf, U. Kahn, A., Dolginow, D., Cossman, J., Kaldjian, E. P., Scudiero, D. A., Petricoin, E., Liotta, L., Lee, J. K. and Weinstein, J. N. (2007). Transcript and protein expression profiles of the NCI-60 cancer cell panel: An integromic microarray study. *Molecular Cancer Therapeutics*, **40** 2877–2909.

Shehata, M., Bièche, I., Boutros, R., Weidenhofer, J., Fanayan, S., Spalding, L., Zeps, N., Byth, K., Bright, R. K., Lidereau, R. and Byrne, J. A. (2008). Nonredundant functions for tumor protein D52-like proteins support specific targeting of TPD52. *Clinical Cancer Research*, **14** 5050–5060.

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

Tukey, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*, **2** 523–531.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, **55** 2183–2202.

Wang, L. (2013). The L_1 penalized LAD estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, **120** 135–151.

Wang, L., Peng, B. and Li, R. (2015). A high-dimensional nonparametric multivariate test for mean vector. *Journal of the American Statistical Association*, **110** 1658–1669.

Wang, L., Wu, Y. and Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, **107** 214–222.

Wu, Y., Siadaty, M. S., Berens, M. E., Hampton, G. M. and Theodorescu, D. (2008). Overlapping gene expression profiles of cell migration and tumor invasion in human bladder cancer identify metallothionein E1 and nicotinamide N-methyltransferase as novel regulators of cell migration. *Oncogene*, **27** 6679–6689.

Yohai, V. J. and Maronna, R. A. (1979). Asymptotic behavior of M-estimators for the linear model. *The Annals of Statistics*, **7** 258–268.

Zheng, Q., Peng, L. and He, X. (2015). Globally adaptive quantile regression with ultra-high dimensional data. *The Annals of Statistics*, **43** 2225–2258.

Zhou, T., Li, Y., Yang, L., Liu, L., Ju, Y. and Li, C. (2017). Silencing of ANXA3 expression by RNA interference inhibits the proliferation and invasion of breast cancer cells. *Oncology Reports*, **37** 388-398.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *The Annals of Statistics*, **28** 461–482.

Supplementary material

A A Lepski-type method

Adapting the unknown robustification parameter depends on the value of the variance provided it exists. Through Lepski's renowned adaptation method (Lepski, 1991), this can be done without actually knowing the variance in advance. Assume that $v_1 = n^{-1}\sum_{i=1}^n \mathbb{E}(\varepsilon_i^2) < \infty$ and let $\sigma_{\max}, \sigma_{\min} > 0$ be such that $\sigma_{\min} \le v_1^{1/2} \le \sigma_{\max}$. Here, parameters $\sigma_{\max}$ and $\sigma_{\min}$ serve as crude preliminary upper and lower bounds for $v_1^{1/2}$, respectively.

For a prespecified $a > 1$, let $\sigma_j = \sigma_{\min} a^j$ and define the set

$$\mathcal{J} = \mathcal{J}_a = \big\{ j = 0, 1, 2, \ldots : \sigma_{\min} \le \sigma_j < a\sigma_{\max} \big\}$$

with its cardinality satisfying $\mathrm{card}(\mathcal{J}) \le 1 + \log_a(\sigma_{\max}/\sigma_{\min})$. For every predetermined $t > 0$, compute a collection of Huber estimators $\{\widehat{\boldsymbol\beta}_{\tau_j}\}_{j\in\mathcal{J}}$, where $\tau_j = \sigma_j (n/t)^{1/2}$ for $j \in \mathcal{J}$. Set

$$\widehat{j} = \min\left\{ j \in \mathcal{J} : \big\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol\beta}_{\tau_k} - \widehat{\boldsymbol\beta}_{\tau_j})\big\|_2 \le 8\widetilde{L}\sigma_j\, d^{1/2}\sqrt{\frac{t}{n}} \ \text{ for all } k > j, k \in \mathcal{J} \right\},$$

where $\widetilde{L} := \max_{1\le i\le n} \|\mathbf{S}_n^{-1/2}\boldsymbol{x}_i\|_\infty$ assuming $\mathbf{S}_n = n^{-1}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm T}$ is positive definite. The final data-driven estimator is then defined as $\widehat{\boldsymbol\beta} = \widehat{\boldsymbol\beta}_{\tau_{\widehat{j}}}$.

Theorem A.1. *For any $t > 0$, the data-dependent estimator $\widehat{\boldsymbol\beta}$ satisfies the bound*

$$\big\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol\beta} - \boldsymbol\beta^*)\big\|_2 \le 12a\widetilde{L}v_1^{1/2}d^{1/2}\sqrt{\frac{t}{n}} \tag{A.1}$$

with probability at least $1 - (2d+1)\log_a(a\sigma_{\max}/\sigma_{\min})e^{-t}$, provided the sample size satisfies $n \ge 8\max(4\widetilde{L}^2 d, \widetilde{L}^4 d^2)t$.

Lepski-type construction relies on preliminary crude upper and lower bounds for $v_1^{1/2}$, which are usually unknown in advance. In practice, one can take $\sigma_{\min} = \widehat{\sigma}/K$ and $\sigma_{\max} = K\widehat{\sigma}$ for some $K > 1$, where $\widehat{\sigma}^2 := (n-d)^{-1}\sum_{i=1}^n (y_i - \langle \boldsymbol{x}_i, \widehat{\boldsymbol\beta}^{\mathrm{ols}}\rangle)^2$ and $\widehat{\boldsymbol\beta}^{\mathrm{ols}}$ is the least squares estimator. Moreover, one may choose $a = 1.5$ and $t = \log n$ or $\log(nd)$. However, the effectiveness of this method depends on how sharp the constants are in the theoretical bounds. We note that all constants in Theorems 1 and A.1 are explicit, although they might not be sharp. Finding sharp constants remains open. Since the current content already consists of long and technical arguments, we will not pursue this particular goal in this paper.

Proof of Theorem A.1. Following the proof of Theorem 1 which is given in Appendix C, it can be similarly proved that, for any $\tau = \tau_0(n/t)^{1/2}$ with $\tau_0 \ge v_1^{1/2}$,

$$\big\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol\beta}_\tau - \boldsymbol\beta^*)\big\|_2 \le 4\widetilde{L}\tau_0\, d^{1/2}\sqrt{\frac{t}{n}} \tag{A.2}$$

with probability at least $1-(2d+1)e^{-t}$ as long as $n \ge 8\max(4\widetilde{L}^2 d, \widetilde{L}^4 d^2)t$.

Let $j^* = \min\{j \in \mathcal{J} : \sigma_j \ge v_1^{1/2}\}$ and note that $v_1^{1/2} \le \sigma_{j^*} \le a v_1^{1/2}$. By the definition of $\widehat{j}$,

$$\begin{aligned}\{\widehat{j} > j^*\} &\subseteq \bigcup_{j\in\mathcal{J}:j>j^*}\left\{\big\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol\beta}_{\tau_j} - \widehat{\boldsymbol\beta}_{\tau_{j^*}})\big\|_2 > 8\widetilde{L}\sigma_j\, d^{1/2}\sqrt{\frac{t}{n}}\right\}\\ &\subseteq \bigcup_{j\in\mathcal{J}:j\ge j^*}\left\{\big\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol\beta}_{\tau_j} - \boldsymbol\beta^*)\big\|_2 > 4\widetilde{L}\sigma_j\, d^{1/2}\sqrt{\frac{t}{n}}\right\}.\end{aligned}$$

Define the event

$$\mathcal{E} = \bigcap_{j\in\mathcal{J}:j\ge j^*}\left\{\big\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol\beta}_{\tau_j} - \boldsymbol\beta^*)\big\|_2 \le 4\widetilde{L}\sigma_j\, d^{1/2}\sqrt{\frac{t}{n}}\right\}$$

such that $\mathcal{E} \subseteq \{\widehat{j} \le j^*\}$. From (A.2) we see that for each $j \ge j^*$,

$$\big\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol\beta}_{\tau_j} - \boldsymbol\beta^*)\big\|_2 \le 4\widetilde{L}\sigma_j\, d^{1/2}\sqrt{\frac{t}{n}}$$

with probability at least $1-(2d+1)e^{-t}$ under the prescribed sample size scaling. By the union bound, we obtain that

$$\begin{aligned}\mathbb{P}(\mathcal{E}^{\rm c}) &\le \sum_{j\in\mathcal{J}:j\ge j^*}\mathbb{P}\left\{\big\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol\beta}_{\tau_j} - \boldsymbol\beta^*)\big\|_2 > 4\widetilde{L}\sigma_j\, d^{1/2}\sqrt{\frac{t}{n}}\right\}\\ &\le (2d+1)|\mathcal{J}|e^{-t} \le (2d+1)\{1+\log_a(\sigma_{\max}/\sigma_{\min})\}e^{-t}.\end{aligned}$$

On the event $\mathcal{E}$, $\widehat{j} \le j^*$ and thus

$$\begin{aligned}\big\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol\beta} - \boldsymbol\beta^*)\big\|_2 &\le \big\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol\beta}_{\tau_{\widehat{j}}} - \widehat{\boldsymbol\beta}_{\tau_{j^*}})\big\|_2 + \big\|\mathbf{S}_n^{1/2}(\widehat{\boldsymbol\beta}_{\tau_{j^*}} - \boldsymbol\beta^*)\big\|_2\\ &\le 8\widetilde{L}\sigma_{j^*}\, d^{1/2}\sqrt{\frac{t}{n}} + 4\widetilde{L}\sigma_{j^*}\, d^{1/2}\sqrt{\frac{t}{n}} \le 12a\widetilde{L}v_1^{1/2}d^{1/2}\sqrt{\frac{t}{n}}.\end{aligned}$$

Together, the last two displays yield (A.1). □

B Random Design Analysis

In this section, we derive counterparts of the results in Section 3 under random designs. First we impose the following moment conditions on the covariates and regression errors.

Condition 1. In linear model (2), the covariate vectors $\boldsymbol{x}_i \in \mathbb{R}^d$ are i.i.d. from a sub-Gaussian random vector $\boldsymbol{x}$, i.e. $\mathbb{P}(|\langle \boldsymbol{u}, \widetilde{\boldsymbol{x}} \rangle| \geq y) \leq 2\exp(-y^2\|\boldsymbol{u}\|_2^2/A_0^2)$ for all $y \in \mathbb{R}$ and $\boldsymbol{u} \in \mathbb{R}^d$, where $\widetilde{\boldsymbol{x}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$ with $\boldsymbol{\Sigma} = (\sigma_{jk})_{1\leq j,k\leq d} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$ being positive definite and $A_0 > 0$ is a constant. The regression errors ε_i are independent and satisfy $\mathbb{E}(\varepsilon_i|\boldsymbol{x}_i) = 0$ and $v_{i,\delta} = \mathbb{E}(|\varepsilon_i|^{1+\delta}|\boldsymbol{x}_i) < \infty$ almost surely for some $\delta > 0$.

Throughout this section, for simplicity, we assume the independent regression errors ε_i in model (2) are homoscedastic in the sense that $v_{i,\delta}$ does not depend on $\boldsymbol{x}_i$. The conditional heteroscedastic model can be allowed with slight modifications as before. With this setup, we write

$$v_\delta = \frac{1}{n}\sum_{i=1}^n v_{i,\delta} \quad \text{and} \quad \nu_\delta = \min\{v_\delta^{1/(1+\delta)}, v_1^{1/2}\}, \quad \delta > 0.$$

Assuming the $d \times d$ matrix $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$ is positive definite, we use $\|\cdot\|_{\boldsymbol{\Sigma},2}$ to denote the rescaled ℓ_2-norm on $\mathbb{R}^d$:

$$\|\boldsymbol{u}\|_{\boldsymbol{\Sigma},2} = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2, \quad \boldsymbol{u} \in \mathbb{R}^d.$$

Moreover, we use ψ_τ to denote the derivative of Huber loss, that is,

$$\psi_\tau(x) = \ell'_\tau(x) = \mathrm{sign}(x)\min(|x|, \tau), \quad x \in \mathbb{R}. \tag{B.1}$$

B.1 Huber regression in low dimensions

In the low dimensional regime "$d \ll n$", we consider the Huber estimator

$$\widehat{\boldsymbol{\beta}}_\tau = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \mathcal{L}_\tau(\boldsymbol{\beta}),$$

where $\mathcal{L}_\tau(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \ell_\tau(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle)$ is the empirical Huber loss function and $\tau > 0$ is the robustification parameter. Under Condition 1, the following theorem provides (i) exponential-type concentration inequalities for $\widehat{\boldsymbol{\beta}}_\tau$ when τ is properly calibrated, and (ii) a nonasymptotic Bahadur representation result under the finite variance condition on regression errors, i.e. $\delta = 1$.

Theorem B.1. Suppose Condition 1 holds.

(I) For any $t > 0$ and $\tau_0 \geq \nu_\delta$, the estimator $\widehat{\boldsymbol{\beta}}_\tau$ with $\tau = \tau_0\{n/(d+t)\}^{\max\{1/(1+\delta),1/2\}}$ satisfies

$$\mathbb{P}\left\{ \|\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2} \geq C_1\tau_0\left(\frac{d+t}{n}\right)^{\min\{\delta/(1+\delta),1/2\}} \right\} \leq 2e^{-t} \tag{B.2}$$

as long as $n \geq C_2(d+t)$, where $C_1, C_2 > 0$ depend only on A_0.

(II) Assume that $v_1 < \infty$. For any $t > 0$ and $\tau_0 \geq v_1^{1/2}$, the estimator $\widehat{\boldsymbol{\beta}}_\tau$ with $\tau = \tau_0\sqrt{n/(d+t)}$ satisfies

$$\mathbb{P}\left\{ \left\| \boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*) - \frac{1}{n}\sum_{i=1}^n \psi_\tau(\varepsilon_i)\widetilde{\boldsymbol{x}}_i \right\|_2 \geq C_3\tau_0\frac{d+t}{n} \right\} \leq 3e^{-t} \tag{B.3}$$

provided $n \geq C_2(d+t)$, where $C_3 > 0$ depends only on A_0.

With random designs, the first part of Theorem B.1 provides concentration inequalities for the ℓ_2-error under finite $(1+\delta)$-th moment conditions with $\delta > 0$; when the second moments are finite, the second part gives a finite-sample approximation of $\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*$ by a sum of independent random vectors. The remainder of such an approximation exhibits sub-exponential tails. Unlike the least squares estimator, the adaptive Huber estimator does not admit an explicit closed-form representation, which causes the main difficulty for analyzing its asymptotic and nonasymptotic properties. Theorem B.1 reveals that, up to a higher-order remainder, the distributional property of $\widehat{\boldsymbol{\beta}}_\tau$ mainly depends on a linear stochastic term that is much easier to deal with.

Regarding the truncated random variable $\psi_\tau(\varepsilon_i)$, the following result shows that the differences between the first two moments of $\psi_\tau(\varepsilon_i)$ and ε_i depend on both τ and the moments of ε_i. The higher moment ε_i has, the faster these differences decay as a function of τ. We summarize this observation in the following proposition. We drop i for ease of presentation.

Proposition B.1. Assume that $\mathbb{E}(\varepsilon) = 0$, $\sigma^2 = \mathbb{E}(\varepsilon^2) > 0$ and $\mathbb{E}(|\varepsilon|^{2+\kappa}) < \infty$ from some $\kappa \geq 0$. Then we have

$$|\mathbb{E}\psi_\tau(\varepsilon)| \leq \min\left\{\tau^{-1}\sigma^2, \tau^{-1-\kappa}\mathbb{E}\big(|\varepsilon|^{2+\kappa}\big)\right\}.$$

Moreover, if $\kappa > 0$,

$$\sigma^2 - 2\kappa^{-1}\tau^{-\kappa}\mathbb{E}\big(|\varepsilon|^{2+\kappa}\big) \leq \mathbb{E}\{\psi_\tau^2(\varepsilon)\} \leq \sigma^2.$$

Proposition B.1, along with Theorem B.1, shows that the adaptive Huber estimator achieves nonasymptotic robustness against heavy-tailed errors, while enjoying high efficiency when τ diverges to ∞. In particular, taking $t = \log n$, we see that under the scaling $n \gtrsim d$, the robust estimator $\widehat{\boldsymbol{\beta}}_\tau$ with $\tau \asymp \sqrt{n/(d+\log n)}$ satisfies

$$\bigg\| \widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^* - \frac{1}{n}\sum_{i=1}^n \psi_\tau(\varepsilon_i)\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_i \bigg\|_2 = O\bigg(\frac{d+\log n}{n}\bigg)$$

with probability at least $1 - O(n^{-1})$. From an asymptotic point of view, this implies that if the dimension d, as a function of n, satisfies

$$d = o(n) \ \text{ as } n \to \infty,$$

then for any deterministic vector $\boldsymbol{a} \in \mathbb{R}^d$, the distribution of $\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^* \rangle$ is close to that of $n^{-1}\sum_{i=1}^n \psi_\tau(\varepsilon_i)\langle \boldsymbol{a}, \boldsymbol{\Sigma}^{-1}\boldsymbol{x}_i\rangle$. If $\varepsilon_1, \ldots, \varepsilon_n$ are independent from ε with variance σ^2 and $\mathbb{E}(|\varepsilon|^{2+\kappa}) < \infty$ for some $\kappa > 0$, taking $\tau \asymp \sqrt{n/(d+\log n)}$ in Proposition B.1 implies that $n^{-1/2}\sum_{i=1}^n \psi_\tau(\varepsilon_i)\langle \boldsymbol{a}, \boldsymbol{\Sigma}^{-1}\boldsymbol{x}_i\rangle$ follows a normal distribution with mean zero and variance $\sigma^2 \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}\|_2^2$ asymptotically.

B.2 Huber regression in high dimensions

In the high dimensional setting where $d \gg n$ and $s = \|\boldsymbol{\beta}^*\|_0 \ll n$, we investigate the ℓ_1-regularized Huber estimator

$$\widehat{\boldsymbol{\beta}}_{\tau,\lambda} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \big\{ \mathcal{L}_\tau(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1 \big\} \tag{B.4}$$

under Condition 1, where τ and λ represent, respectively, the robustification and regularization parameters.

Theorem B.2. Assume Condition 1 holds and that the unknown $\boldsymbol{\beta}^*$ is sparse with $s = \|\boldsymbol{\beta}^*\|_0$. Then any optimal solution $\widehat{\boldsymbol{\beta}}_{\tau,\lambda}$ to the convex program (B.4) with

$$\tau = \tau_0 \bigg(\frac{n}{\log d}\bigg)^{\max\{1/(1+\delta),1/2\}} \qquad (\tau_0 \geq \nu_\delta) \tag{B.5}$$

and λ scaling as $A_0 \sigma_{\max} \tau_0 \{(\log d)/n\}^{\min\{\delta/(1+\delta),1/2\}}$ satisfies the bounds

$$\begin{aligned} \big\|\widehat{\boldsymbol{\beta}}_{\tau,\lambda} - \boldsymbol{\beta}^*\big\|_{\boldsymbol{\Sigma},2} &\lesssim \kappa_l^{-1/2} A_0 \sigma_{\max}\tau_0\, s^{1/2} \bigg(\frac{\log d}{n}\bigg)^{\min\{\delta/(1+\delta),1/2\}} \\ \text{and} \quad \big\|\widehat{\boldsymbol{\beta}}_{\tau,\lambda} - \boldsymbol{\beta}^*\big\|_1 &\lesssim \kappa_l^{-1} A_0 \sigma_{\max}\tau_0\, s \bigg(\frac{\log d}{n}\bigg)^{\min\{\delta/(1+\delta),1/2\}} \end{aligned} \tag{B.6}$$

with probability at least $1 - 3d^{-1}$ as long as $n \geq C\kappa_l^{-1}\sigma_{\max}^2 s \log d$, where $C > 0$ is a constant only depending on A_0, $\sigma_{\max} = \max_{1\leq j\leq d}\sigma_{jj}^{1/2}$ and $\kappa_l = \lambda_{\min}(\boldsymbol{\Sigma})$.

Provided the distribution of ε_i has finite variance, i.e. $\delta = 1$, Theorem B.2 asserts that the ℓ_1-regularized Huber regression with properly tuned (τ, λ) gives rise to statistically consistent estimators with ℓ_1- and ℓ_2-errors scaling as $s\sqrt{(\log d)/n}$ and $\sqrt{s(\log d)/n}$, respectively, under the sample size scaling $n \gtrsim s\log d$. These rates are the minimax rates enjoyed by the standard Lasso with Gaussian/sub-Gaussian errors (Bickel, Ritov and Tsybakov, 2009; Wainwright, 2009).

The results of Theorem B.2 are useful complements to those in Theorem 4 under fixed designs. Taking $t = \log d$ therein, we see that the ℓ_2-error bound in (10) almost coincides with that in (B.6) up to constant factors. The sample size scaling under random designs is optimal and better than the scaling under fixed designs: the former is of order $O(s\log d)$, while the latter is of order $O(s^2\log d)$. Technically, the sample size scaling is required to ensure the restricted strong convexity of Huber loss in a neighborhood of $\boldsymbol{\beta}^*$; see Lemma 1 in the main text and Lemma C.4 below. Since most existing works on analyzing high dimensional M-estimators beyond the least squares have focused on random designs (see, e.g. Belloni and Chernozhukov (2011), Negahban et al. (2012) and the references therein), it is not clear what the optimal sample size scaling is under fixed designs, although it is possible that the additional s factor in Theorem 4 is purely an artifact of the proof technique. We refer to van de Geer (2008) for a study of generalized linear models in high dimensions. To achieve the oracle rate for the excess risk, the sparsity s is required to be of order $O(\sqrt{n/\log n})$, or equivalently, the required sample size scales as $s^2\log n$.

We complete this section by a prediction error bound for $\widehat{\boldsymbol{\beta}}_{\tau,\lambda}$, which is a direct consequence of Theorem B.2.

Corollary B.1. Under the conditions of Theorem B.2, it holds

$$\frac{1}{\sqrt{n}}\big\|\mathbf{X}(\widehat{\boldsymbol{\beta}}_{\tau,\lambda}-\boldsymbol{\beta}^*)\big\|_2 \lesssim \kappa_l^{-1/2} A_0\sigma_{\max}\tau_0\, s^{1/2}\bigg(\frac{\log d}{n}\bigg)^{\min\{\delta/(1+\delta),1/2\}} \tag{B.7}$$

with probability at least $1-5d^{-1}$, where $\mathbf{X} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)^{\mathrm{T}}$ is the $n\times d$ design matrix.

C Proofs of Main Theorems

Throughout the proofs, we use $\psi_\tau = \ell_\tau'$ as in definition (B.1) and let $\|\cdot\|_{\boldsymbol{\Sigma},2}$ be the rescaled ℓ_2-norm on $\mathbb{R}^d$ given by $\|\boldsymbol{u}\|_{\boldsymbol{\Sigma},2} = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2$ for $\boldsymbol{u}\in\mathbb{R}^d$.

C.1 Auxiliary Lemmas

First we collect several auxiliary lemmas. Our first lemma concerns the localized analysis that can be utilized to remove the parameter constraint in previous works. It is established in Fan et al. (2018) and we reproduce it here for completeness.

Lemma C.1. Let $D_{\mathcal{L}}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) = \mathcal{L}(\boldsymbol{\beta}_1)-\mathcal{L}(\boldsymbol{\beta}_2)-\langle\nabla\mathcal{L}(\boldsymbol{\beta}_2),\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\rangle$ and $D^s_{\mathcal{L}}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) = D_{\mathcal{L}}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)+D_{\mathcal{L}}(\boldsymbol{\beta}_2,\boldsymbol{\beta}_1)$. For $\boldsymbol{\beta}_\eta = \boldsymbol{\beta}^*+\eta(\boldsymbol{\beta}-\boldsymbol{\beta}^*)$ with $\eta\in(0,1]$ and any convex loss functions $\mathcal{L}$, we have

$$D^s_{\mathcal{L}}(\boldsymbol{\beta}_\eta, \boldsymbol{\beta}^*) \leq \eta D^s_{\mathcal{L}}(\boldsymbol{\beta}, \boldsymbol{\beta}^*).$$

Proof of Lemma C.1. Let $Q(\eta) = D_{\mathcal{L}}(\boldsymbol{\beta}_\eta, \boldsymbol{\beta}^*) = \mathcal{L}(\boldsymbol{\beta}_\eta) - \mathcal{L}(\boldsymbol{\beta}^*) - \langle \nabla \mathcal{L}(\boldsymbol{\beta}^*), \boldsymbol{\beta}_\eta - \boldsymbol{\beta}^* \rangle$. Noting that the derivative of $\mathcal{L}(\boldsymbol{\beta}_\eta)$ with respect to η is $\frac{d}{d\eta}\mathcal{L}(\boldsymbol{\beta}_\eta) = \langle \nabla \mathcal{L}(\boldsymbol{\beta}_\eta), \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle$, we have

$$Q'(\eta) = \langle \nabla \mathcal{L}(\boldsymbol{\beta}_\eta) - \nabla \mathcal{L}(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle.$$

Then, the symmetric Bregman divergence $D^s_{\mathcal{L}}(\boldsymbol{\beta}_\eta - \boldsymbol{\beta}^*)$ can be written as

$$D^s_{\mathcal{L}}(\boldsymbol{\beta}_\eta, \boldsymbol{\beta}^*) = \langle \nabla \mathcal{L}(\boldsymbol{\beta}_\eta) - \nabla \mathcal{L}(\boldsymbol{\beta}^*), \eta(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \rangle = \eta Q'(\eta), \quad 0 < \eta \leq 1.$$

Taking $\eta = 1$ in the above equation, we have $Q'(1) = D^s_{\mathcal{L}}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$ as a special case. If $Q(\eta)$ is convex, then $Q'(\eta)$ is non-decreasing and thus

$$D^s_{\mathcal{L}}(\boldsymbol{\beta}_\eta, \boldsymbol{\beta}^*) = \eta Q'(\eta) \leq \eta Q'(1) = \eta D^s_{\mathcal{L}}(\boldsymbol{\beta}, \boldsymbol{\beta}^*).$$

It remains to show the convexity of $\eta \in [0,1] \mapsto Q(\eta)$; or equivalently, the convexity of $\mathcal{L}(\boldsymbol{\beta}_\eta)$ and $\langle \nabla \mathcal{L}(\boldsymbol{\beta}^*), \boldsymbol{\beta}^* - \boldsymbol{\beta}_\eta \rangle$, respectively. First, note that $\boldsymbol{\beta}_\eta$, as a function of η, is linear in η, that is, $\boldsymbol{\beta}_{\alpha_1\eta_1 + \alpha_2\eta_2} = \alpha_1 \boldsymbol{\beta}_{\eta_1} + \alpha_2 \boldsymbol{\beta}_{\eta_2}$ for all $\eta_1, \eta_2 \in [0,1]$ and $\alpha_1, \alpha_2 \geq 0$ satisfying $\alpha_1 + \alpha_2 = 1$. Then, the convexity of $\eta \mapsto \mathcal{L}(\boldsymbol{\beta}_\eta)$ follows from this linearity and the convexity of the Huber loss. The convexity of the second term follows directly from the bi-linearity of the inner product. □

The following two lemmas provide restricted strong convexity properties for the Huber loss in a local vicinity of the true parameter under both fixed and random designs.

Lemma C.2. Assume that Condition 1 holds and that $v_\delta = n^{-1}\sum_{i=1}^n \mathbb{E}(|\varepsilon_i|^{1+\delta}) < \infty$ for some $0 < \delta \leq 1$. Then for any $t, r > 0$, the Hessian matrix $\nabla^2 \mathcal{L}_\tau(\boldsymbol{\beta})$ with $\tau > 2Mr$ satisfies that, with probability greater than $1 - e^{-t}$,

$$\begin{aligned} \min_{\boldsymbol{\beta} \in \mathbb{R}^d : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq r} & \lambda_{\min}\big(\nabla^2 \mathcal{L}_\tau(\boldsymbol{\beta})\big) \\ & \geq \big\{1 - (2Mr/\tau)^2\big\} c_l - M^2\big\{(2/\tau)^{1+\delta} v_\delta + (2n)^{-1/2} t^{1/2}\big\}, \end{aligned} \tag{C.1}$$

where $M = \max_{1 \leq i \leq n} \|\boldsymbol{x}_i\|_2$.

Proof of Lemma C.2. To begin with, note that

$$\mathbf{H}_n(\boldsymbol{\beta}) = \nabla^2 \mathcal{L}_\tau(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} 1\big(|y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}| \leq \tau\big),$$

where $\mathbf{S}_n$ is given in Condition 1. For each $\boldsymbol{\beta} \in \mathbb{R}^d$, define its centered and rescaled version $\boldsymbol{\beta}_0 = \boldsymbol{\beta} - \boldsymbol{\beta}^*$ such that $y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle = \varepsilon_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_0 \rangle$. Using the inequality that

$$1\big(|y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle| > \tau\big) \le 1\big(|\varepsilon_i| > \tau/2\big) + 1\big(|\langle \boldsymbol{x}_i, \boldsymbol{\beta}_0 \rangle| > \tau/2\big),$$

we have, for any $\boldsymbol{u} \in \mathbb{S}^{d-1}$ and $\boldsymbol{\beta} \in \mathbb{R}^d$ satisfying $\|\boldsymbol{\beta}_0\|_2 \le r$,

$$\begin{aligned} &\langle \boldsymbol{u}, \mathbf{H}_n(\boldsymbol{\beta}) \boldsymbol{u} \rangle \\ &\ge \|\mathbf{S}_n^{1/2} \boldsymbol{u}\|_2^2 - \frac{1}{n} \sum_{i=1}^n \langle \boldsymbol{x}_i, \boldsymbol{u} \rangle^2 1\big(|\varepsilon_i| > \tau/2\big) - \frac{1}{n} \sum_{i=1}^n \langle \boldsymbol{x}_i, \boldsymbol{u} \rangle^2 1\big(|\langle \boldsymbol{x}_i, \boldsymbol{\beta}_0 \rangle| > \tau/2\big) \\ &\ge \|\mathbf{S}_n^{1/2} \boldsymbol{u}\|_2^2 - \max_{1 \le i \le n} \|\boldsymbol{x}_i\|_2^2 \left\{ \frac{1}{n} \sum_{i=1}^n 1\big(|\varepsilon_i| > \tau/2\big) + \frac{4}{\tau^2} \|\boldsymbol{\beta}_0\|_2^2 \, \|\mathbf{S}_n^{1/2} \boldsymbol{u}\|_2^2 \right\} \\ &\ge c_l \big\{1 - (2Mr/\tau)^2\big\} - \frac{M^2}{n} \sum_{i=1}^n 1\big(|\varepsilon_i| > \tau/2\big), \end{aligned}$$

provided that $\tau > 2Mr$. For any $z \ge 0$, it follows from Hoeffding's inequality that, with probability at least $1 - e^{-2nz^2}$,

$$\frac{1}{n} \sum_{i=1}^n 1\big(|\varepsilon_i| > \tau/2\big) \le \frac{1}{n} \sum_{i=1}^n \mathbb{P}\big(|\varepsilon_i| > \tau/2\big) + z.$$

This, together with the inequality $\mathbb{P}(|\varepsilon_i| > \tau/2) \le (2/\tau)^{1+\delta} v_{i,\delta}$ and Condition 1, implies that, with probability at least $1 - e^{-2nz^2}$,

$$\big\langle \boldsymbol{u}, \mathbf{H}_n(\boldsymbol{\beta}) \boldsymbol{u} \big\rangle \ge \big\{1 - (2Mr/\tau)^2\big\} c_l - M^2 \big\{(2/\tau)^{1+\delta} v_\delta + z\big\}.$$

This proves (C.1) immediately by taking $z = \sqrt{t/(2n)}$. □

Lemma C.3. Assume $v_\delta < \infty$ for some $0 < \delta \le 1$ and $(\mathbb{E}\langle \boldsymbol{u}, \widetilde{\boldsymbol{x}} \rangle^4)^{1/4} \le A_1 \|\boldsymbol{u}\|_2$ for all $\boldsymbol{u} \in \mathbb{R}^d$ and some constant $A_1 > 0$. Moreover, let $\tau, r > 0$ satisfy

$$\tau \ge 2 \max\big\{(4v_\delta)^{1/(1+\delta)}, 4A_1^2 r\big\} \quad \text{and} \quad n \gtrsim (\tau/r)^2 (d + t). \tag{C.2}$$

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

$$\big\langle \nabla \mathcal{L}_\tau(\boldsymbol{\beta}) - \nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^* \big\rangle \ge \frac{1}{4} \big\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\big\|_{\boldsymbol{\Sigma},2}^2 \tag{C.3}$$

uniformly over $\boldsymbol{\beta} \in \Theta_0(r) = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2} \le r\}$.

Proof of Lemma C.3. To begin with, note that

$$\begin{aligned}\mathcal{T}(\boldsymbol{\beta}) &:= \langle \nabla \mathcal{L}_\tau(\boldsymbol{\beta}) - \nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle \\ &= \frac{1}{n}\sum_{i=1}^n \{\psi_\tau(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle) - \psi_\tau(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle)\} \langle \boldsymbol{x}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle \\ &\geq \frac{1}{n}\sum_{i=1}^n \{\psi_\tau(\varepsilon_i) - \psi_\tau(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle)\} \langle \boldsymbol{x}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle 1\{\mathcal{E}_i\}, \end{aligned} \tag{C.4}$$

where $1\{\mathcal{E}_i\}$ denotes the indication function of the event

$$\mathcal{E}_i = \big\{|\varepsilon_i| \leq \tau/2\big\} \cap \big\{|\langle \boldsymbol{x}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle| \leq \tau \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2}/(2r)\big\}.$$

On $\mathcal{E}_i$, it holds $|y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle| \leq |\varepsilon_i| + |\langle \boldsymbol{x}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle| \leq \tau/2 + \tau/2 = \tau$ for all $\boldsymbol{\beta} \in \Theta_0(r)$. Since $\psi'_\tau(x) = 1$ for $|x| \leq \tau$, the right-hand of (C.4) can be bounded from below by

$$\frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{x}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle^2 1\big\{|\langle \boldsymbol{x}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle| \leq \tau \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2}/(2r)\big\} 1\big\{|\varepsilon_i| \leq \tau/2\big\}. \tag{C.5}$$

To bound the right-hand of (C.5), the main difficulty is that the indicator function is non-smooth. To deal with this issue, we define the following "smoothed" functions: for any $R > 0$, write

$$\phi_R(x) = \begin{cases} x^2 & \text{if } |x| \leq R/2, \\ (x-R)^2 & \text{if } R/2 < x \leq R, \\ (x+R)^2 & \text{if } -R \leq x \leq -R/2, \\ 0 & \text{if } |x| > R, \end{cases} \quad \text{and} \quad \varphi_R(y) = 1(|y| \leq R).$$

It is easy to see that the function ϕ_R is R-Lipschitz and satisfies

$$x^2 1(|x| \leq R/2) \leq \phi_R(x) \leq x^2 1(|x| \leq R). \tag{C.6}$$

Together, (C.4), (C.5) and (C.6) imply

$$\mathcal{T}(\boldsymbol{\beta}) \geq g(\boldsymbol{\beta}) := \frac{1}{n}\sum_{i=1}^n \phi_{\tau\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2}/(2r)}(\langle \boldsymbol{x}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle) \varphi_{\tau/2}(\varepsilon_i). \tag{C.7}$$

For $r > 0$, define $\Delta(r) = \sup_{\boldsymbol{\beta} \in \Theta_0(r)} |g(\boldsymbol{\beta}) - \mathbb{E} g(\boldsymbol{\beta})| / \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2}^2$, such that

$$\frac{\mathcal{T}(\boldsymbol{\beta})}{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2}^2} \geq \frac{\mathbb{E} g(\boldsymbol{\beta})}{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2}^2} - \Delta(r) \tag{C.8}$$

for all $\boldsymbol{\beta} \in \Theta_0(r)$. In the following, we establish lower and upper bounds for $\mathbb{E} g(\boldsymbol{\beta})$

and $\Delta(r)$, respectively, starting with the former.

For $\boldsymbol{\beta} \in \mathbb{R}^d$, write $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$. By (C.7) and Markov's inequality,

$$\begin{aligned}\mathbb{E}g(\boldsymbol{\beta}) &\geq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle^2 - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle^2 1\big\{|\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle| \geq \tau\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}/(4r)\big\} \\ &\quad - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle^2 1(|\varepsilon_i| > \tau/2) \\ &\geq \boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\delta} - v_\delta(2/\tau)^{1+\delta}\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\delta} - (4r/\tau)^2\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}^{-2}\frac{1}{n}\sum_{i=1}^n \mathbb{E}\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle^4 \\ &\geq \|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}^2\big\{1 - v_\delta(2/\tau)^{1+\delta} - (4A_1^2 r/\tau)^2\big\}.\end{aligned}$$

Provided $\tau \geq 2\max\{(4v_\delta)^{1/(1+\delta)}, 4A_1^2 r\}$,

$$\mathbb{E}g(\boldsymbol{\beta}) \geq \frac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2}^2 \quad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^d. \tag{C.9}$$

Next we bound the supremum $\Delta(r)$. Write $g(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n g_i(\boldsymbol{\beta})$. Noting that $0 \leq \phi_R(x) \leq R^2/4$ and $0 \leq \varphi(y) \leq 1$, we have

$$0 \leq g_i(\boldsymbol{\beta}) \leq (\tau/4r)^2\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2}^2.$$

By Theorem 7.3 in Bousquet (2003), for any $x > 0$, $\Delta(r)$ satisfies the bound

$$\Delta(r) \leq \mathbb{E}\Delta(r) + \{\mathbb{E}\Delta(r)\}^{1/2}(\tau/2r)\sqrt{\frac{x}{n}} + \sigma_n\sqrt{\frac{2x}{n}} + (\tau/4r)^2\frac{x}{3n} \tag{C.10}$$

with probability at least $1 - e^{-x}$, where by (C.6),

$$\sigma_n^2 := \frac{1}{n}\sum_{i=1}^n \sup_{\boldsymbol{\beta}\in\Theta_0(r)} \frac{\mathbb{E}g_i^2(\boldsymbol{\beta})}{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2}^4} \leq A_1^4.$$

For the expected value $\mathbb{E}\Delta(r)$, using the symmetrization inequality and the connection between Gaussian complexity and Rademacher complexity, we obtain that $\mathbb{E}\Delta(r) \leq \sqrt{2\pi}\,\mathbb{E}\{\sup_{\boldsymbol{\beta}\in\Theta_0(r)}|\mathbb{G}_{\boldsymbol{\beta}}|\}$, where

$$\mathbb{G}_{\boldsymbol{\beta}} = \frac{1}{n}\sum_{i=1}^n \frac{G_i}{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2}^2}\phi_{\tau\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2}/(2r)}(\langle \boldsymbol{x}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle)\varphi_{\tau/2}(\varepsilon_i)$$

and G_i are i.i.d. standard normal random variables that are independent of $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$. Let $\mathbb{E}^*$ be the conditional expectation given $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$. Since $\{\mathbb{G}_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \Theta_0(r)\}$ is a conditional Gaussian process, for any $\boldsymbol{\beta}_0 \in \Theta_0(r)$ we have

$$\mathbb{E}^*\left\{ \sup_{\boldsymbol{\beta}\in\Theta_0(r)} |\mathbb{G}_{\boldsymbol{\beta}}| \right\} \le \mathbb{E}^*|\mathbb{G}_{\boldsymbol{\beta}_0}| + 2\mathbb{E}^*\left\{ \sup_{\boldsymbol{\beta}\in\Theta_0(r)} \mathbb{G}_{\boldsymbol{\beta}} \right\}. \tag{C.11}$$

Further, taking the expectation with respect to $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ on both sides, (C.11) remains valid with $\mathbb{E}^*$ replaced by $\mathbb{E}$. We write $\boldsymbol{\beta}^*$ as $(\beta_1^*, \widetilde{\boldsymbol{\beta}}^{*\mathrm{T}})^{\mathrm{T}}$ with β_1^* denoting the first coordinate of $\boldsymbol{\beta}^*$ and $\widetilde{\boldsymbol{\beta}}^* \in \mathbb{R}^{d-1}$. Recalling $\phi_R(u) \le \min(u^2, R^2/4)$, we take $\boldsymbol{\beta}_0 = (\beta_1^* + (\mathbb{E}x_1^2)^{-1/2} r, \widetilde{\boldsymbol{\beta}}^{*\mathrm{T}})^{\mathrm{T}}$ so that $\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2} = r$ and $\mathbb{E}|\mathbb{G}_{\boldsymbol{\beta}_0}| \le (\mathbb{E}\mathbb{G}_{\boldsymbol{\beta}_0}^2)^{1/2} \le (4r)^{-1}\tau n^{-1/2}$. To bound the conditional expectation $\mathbb{E}^*\{\sup_{\boldsymbol{\beta}\in\Theta_0(r)} \mathbb{G}_{\boldsymbol{\beta}}\}$ in (C.11), we employ the Gaussian comparison theorem as in the proof of Lemma 11 in Loh and Wainwright (2015)

Denote by var* the conditional variance given $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$. For $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \Theta_0(r)$, write $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$ and $\boldsymbol{\delta}' = \boldsymbol{\beta}' - \boldsymbol{\beta}^*$. By conditional normality, we quickly compute and bound the variance of $\mathbb{G}_{\boldsymbol{\beta}} - \mathbb{G}_{\boldsymbol{\beta}'}$:

$$\mathrm{var}^*(\mathbb{G}_{\boldsymbol{\beta}} - \mathbb{G}_{\boldsymbol{\beta}'}) \le \frac{1}{n^2}\sum_{i=1}^n \varphi_{\tau/2}^2(\varepsilon_i)\left\{ \frac{\phi_{\tau\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}/(2r)}(\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle)}{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}^2} - \frac{\phi_{\tau\|\boldsymbol{\delta}'\|_{\boldsymbol{\Sigma},2}/(2r)}(\langle \boldsymbol{x}_i, \boldsymbol{\delta}'\rangle)}{\|\boldsymbol{\delta}'\|_{\boldsymbol{\Sigma},2}^2} \right\}^2.$$

Using the property $\phi_{cR}(cx) = c^2\phi_R(x)$ for any $c > 0$, we find that

$$\phi_{\tau\|\boldsymbol{\delta}'\|_{\boldsymbol{\Sigma},2}/(2r)}(\langle \boldsymbol{x}_i, \boldsymbol{\delta}'\rangle) = \frac{\|\boldsymbol{\delta}'\|_{\boldsymbol{\Sigma},2}^2}{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}^2}\phi_{\tau\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}/(2r)}\left(\frac{\|\boldsymbol{\delta}'\|_{\boldsymbol{\Sigma},2}}{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}}\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle \right).$$

It follows from the above calculations and the Lipschitz property of ϕ_R that

$$\begin{aligned} &\mathrm{var}^*(\mathbb{G}_{\boldsymbol{\beta}} - \mathbb{G}_{\boldsymbol{\beta}'}) \\ &\le \frac{1}{n^2}\sum_{i=1}^n \frac{1}{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}^4}\left\{ \phi_{\tau\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}/(2r)}(\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle) - \phi_{\tau\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}/(2r)}\left(\frac{\|\boldsymbol{\delta}'\|_{\boldsymbol{\Sigma},2}}{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}}\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle \right) \right\}^2 \\ &\le \frac{1}{n^2}\sum_{i=1}^n \frac{\tau^2}{4r^2}\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle}{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma},2}} - \frac{\langle \boldsymbol{x}_i, \boldsymbol{\delta}'\rangle}{\|\boldsymbol{\delta}'\|_{\boldsymbol{\Sigma},2}} \right)^2. \end{aligned} \tag{C.12}$$

Let $G_1', \ldots, G_n'$ be i.i.d. standard normal random variables that are independent of all the previous variables, and define a new process

$$\mathbb{Z}_{\boldsymbol{\beta}} = \frac{\tau}{2rn}\sum_{i=1}^n G_i' \frac{\langle \boldsymbol{x}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle}{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2}}.$$

As an immediate consequence of (C.12), we have $\mathrm{var}^*(\mathbb{G}_{\boldsymbol{\beta}} - \mathbb{G}_{\boldsymbol{\beta}'}) \le \mathrm{var}^*(\mathbb{Z}_{\boldsymbol{\beta}} - \mathbb{Z}_{\boldsymbol{\beta}'})$.

Therefore, by the Gaussian comparison inequality (Ledoux and Talagrand, 1991),

$$\mathbb{E}^*\bigg\{\sup_{\boldsymbol{\beta}\in\Theta_0(r)} \mathbb{G}_{\boldsymbol{\beta}}\bigg\} \leq 2\mathbb{E}^*\bigg\{\sup_{\boldsymbol{\beta}\in\Theta_0(r)} \mathbb{Z}_{\boldsymbol{\beta}}\bigg\} \leq \frac{\tau}{r}\mathbb{E}^*\bigg\|\frac{1}{n}\sum_{i=1}^n G_i'\widetilde{\boldsymbol{x}}_i\bigg\|_2,$$

where $\widetilde{\boldsymbol{x}}_i = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i$. Taking the expectation with respect to $\{(y_i,\boldsymbol{x}_i)\}_{i=1}^n$ on both sides gives $\mathbb{E}\{\sup_{\boldsymbol{\beta}\in\Theta_0(r)}\mathbb{G}_{\boldsymbol{\beta}}\} \leq (\tau/r)\mathbb{E}\|n^{-1}\sum_{i=1}^n G_i'\widetilde{\boldsymbol{x}}_i\|_2 \leq (\tau/r)\sqrt{d/n}$. From this and the unconditional version of (C.11), we obtain

$$\mathbb{E}\Delta(r) \leq \sqrt{2\pi}\bigg(\frac{2\tau}{r}\sqrt{\frac{d}{n}} + \frac{\tau}{4r\sqrt{n}}\bigg). \tag{C.13}$$

Together, (C.10) with $x=t$ and (C.13) imply that as long as $n \gtrsim (\tau/r)^2(d+t)$, $\Delta(r) \leq 1/4$ with probability at least $1-e^{-t}$. Combining this with (C.5) and (C.9) proves the stated result. □

Recall that $\Theta_0(r) = \{\boldsymbol{\beta}\in\mathbb{R}^d : \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2} \leq r\}$. Let $\mathcal{C} = \{\boldsymbol{\beta}\in\mathbb{R}^d : \|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\mathcal{S}^c}\|_1 \leq 3\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\mathcal{S}}\|_1\}$ be an ℓ_1-cone in $\mathbb{R}^d$, where $\mathcal{S}\subseteq\{1,\ldots,d\}$ denotes the support of $\boldsymbol{\beta}^*$. As a counterpart of Lemma C.3 in high dimensions, Lemma C.4 below shows that the adaptive Huber loss satisfies the restricted strong convexity condition over $\Theta_0(r)\cap\mathcal{C}$ with high probability.

Lemma C.4. Assume $v_\delta < \infty$ for some $0<\delta\leq 1$ and $(\mathbb{E}\langle\boldsymbol{u},\widetilde{\boldsymbol{x}}\rangle^4)^{1/4} \leq A_1\|\boldsymbol{u}\|_2$ for all $\boldsymbol{u}\in\mathbb{R}^d$ and some constant $A_1>0$. Let (n,d,τ,r) satisfy

$$\tau \geq 2\max\big\{(4v_\delta)^{1/(1+\delta)}, 4A_1^2 r\big\} \quad\text{and}\quad n \gtrsim \kappa_l^{-1}(A_0\tau/r)^2\max_{1\leq j\leq d}\sigma_{jj}\, s\log d, \tag{C.14}$$

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

$$\big\langle\nabla\mathcal{L}_\tau(\boldsymbol{\beta})-\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*), \boldsymbol{\beta}-\boldsymbol{\beta}^*\big\rangle \geq \frac{1}{4}\big\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\big\|_{\boldsymbol{\Sigma},2}^2 \tag{C.15}$$

uniformly over $\boldsymbol{\beta}\in\Theta_0(r)\cap\mathcal{C}$.

Proof of Lemma C.4. The proof is based on an argument similar to that in the proof of Lemma C.3. With slight abuse of notation, we keep using $\Delta(r)$ as the supremum of a random process:

$$\Delta(r) = \sup_{\boldsymbol{\beta}\in\Theta_0(r)\cap\mathcal{C}} \frac{|g(\boldsymbol{\beta})-\mathbb{E}g(\boldsymbol{\beta})|}{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2}^2}.$$

Provided $\tau \geq 2\max\{(4v_\delta)^{1/(1+\delta)}, 4A_1^2 r\}$, it can be shown that

$$\frac{\mathcal{T}(\boldsymbol{\beta})}{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2}^2} \geq \frac{1}{2} - \Delta(r) \quad\text{for all}\quad \boldsymbol{\beta}\in\Theta_0(r)\cap\mathcal{C}. \tag{C.16}$$

According to (C.10), it remains to bound $\mathbb{E}\Delta(r)$. Following the proof of Lemma C.3, it suffices to focus on the (conditional) Gaussian process

$$\mathbb{Z}_{\boldsymbol{\beta}} = \frac{\tau}{2rn}\sum_{i=1}^{n} G_i' \frac{\langle \boldsymbol{x}_i, \boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle}{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2}}, \quad \boldsymbol{\beta} \in \Theta_0(r)\cap\mathcal{C},$$

where G_i' are i.i.d. standard normal random variables that are independent of all other random variables. For every $\boldsymbol{\beta} \in \Theta_0(r)\cap\mathcal{C}$, it is easy to see that

$$\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1 \le 4\sqrt{s}\,\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \le 4\kappa_l^{-1/2}\sqrt{s}\,\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2},$$

implying

$$\sup_{\boldsymbol{\beta}\in\Theta_0(r)\cap\mathcal{C}} \mathbb{Z}_{\boldsymbol{\beta}} \le 2\kappa_l^{-1/2}\sqrt{s}\,\frac{\tau}{r}\left\|\frac{1}{n}\sum_{i=1}^{n} G_i'\boldsymbol{x}_i\right\|_\infty .$$

Keep all other statements the same, we obtain

$$\mathbb{E}\Delta(r) \le \sqrt{2\pi}\left(8\kappa_l^{-1/2}\sqrt{s}\,\frac{\tau}{r}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} G_i'\boldsymbol{x}_i\right\|_\infty + \frac{\tau}{4r\sqrt{n}}\right).$$

With $\boldsymbol{x}_i = (x_{i1},\ldots,x_{id})^{\mathrm{T}} \in \mathbb{R}^d$, note that

$$\left\|\frac{1}{n}\sum_{i=1}^{n} G_i'\boldsymbol{x}_i\right\|_\infty = \max_{1\le j\le d}\left|\frac{1}{n}\sum_{i=1}^{n} G_i' x_{ij}\right|.$$

Since $G_i'x_{ij}$ are sub-exponential/sub-gamma random variables, from Corollary 2.6 in Boucheron, Lugosi and Massart (2013) we find that

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} G_i'\boldsymbol{x}_i\right\|_\infty \lesssim A_0 \max_{1\le j\le d}\sigma_{jj}^{1/2}\left(\sqrt{\frac{\log d}{n}} + \frac{\log d}{n}\right).$$

Substituting this into (C.10) and taking $x = \log d$, we obtain that with probability at least $1-d^{-1}$,

$$\frac{\mathcal{T}(\boldsymbol{\beta})}{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2}^2} \ge \frac{1}{4} \quad \text{uniformly over } \boldsymbol{\beta}\in\Theta_0(r)\cap\mathcal{C}$$

for all sufficiently large n that scales as $\kappa_l^{-1}(A_0\tau/r)^2 \max_{1\le j\le d}\sigma_{jj}\, s\log d$ up to an absolute constant. This proves (C.15). □

Lemmas C.5 and C.6 provide concentration inequalities for $\|\boldsymbol{\Sigma}^{-1/2}\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_2$ and $\|\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_\infty$, respetively.

Lemma C.5. Assume Condition 1 holds with $0<\delta\le 1$. Then with probability at least $1 - 2e^{-t}$,

$$\big\|\boldsymbol{\Sigma}^{-1/2}\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)\big\|_2 \le 4\sqrt{2}A_0\sqrt{\frac{v_\delta\tau^{1-\delta}(d+t)}{n}} + 2A_0\tau\frac{d+t}{n} + v_\delta\tau^{-\delta}. \tag{C.17}$$

Proof of C.5. Assume without loss of generality that $t \ge \log 2$, or equivalently, $2e^{-t} \le 1$; otherwise $2e^{-t} > 1$ so that the bound is trivial. To bound $\|\boldsymbol{\Sigma}^{-1/2}\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_2$, first define the centered random vector

$$\boldsymbol{\xi}^* = \boldsymbol{\Sigma}^{-1/2}\big\{\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*) - \nabla\mathbb{E}\mathcal{L}_\tau(\boldsymbol{\beta}^*)\big\} = -\frac{1}{n}\sum_{i=1}^n\big\{\xi_i\widetilde{\boldsymbol{x}}_i - \mathbb{E}(\xi_i\widetilde{\boldsymbol{x}}_i)\big\},$$

where $\xi_i = \psi_\tau(\varepsilon_i)$. To evaluate the ℓ_2-norm, there exits a $1/2$-net $\mathcal{N}_{1/2}$ of the unit sphere $\mathbb{S}^{d-1}$ in $\mathbb{R}^d$ with $|\mathcal{N}_{1/2}| \le 5^d$ such that $\|\boldsymbol{\xi}^*\|_2 \le 2\max_{\boldsymbol{u}\in\mathcal{N}_{1/2}}|\langle\boldsymbol{u},\boldsymbol{\xi}^*\rangle|$. Under Condition 1, it holds for every $\boldsymbol{u}\in\mathbb{S}^{d-1}$ that $\mathbb{E}|\langle\boldsymbol{u},\widetilde{\boldsymbol{x}}\rangle|^k \le A_0^k k\Gamma(k/2)$ for all $k \ge 1$. By direct calculations,

$$\sum_{i=1}^n\mathbb{E}(\xi_i\langle\boldsymbol{u},\widetilde{\boldsymbol{x}}_i\rangle)^2 \le 2A_0^2\tau^{1-\delta}\sum_{i=1}^n v_{i,1} = 2A_0^2 n v_\delta\tau^{1-\delta},$$

$$\sum_{i=1}^n\mathbb{E}|\xi_i\langle\boldsymbol{u},\widetilde{\boldsymbol{x}}_i\rangle|^k \le \frac{k!}{2}(A_0\tau/2)^{k-2}2A_0^2 n v_\delta\tau^{1-\delta} \quad \text{for all } k \ge 3.$$

It then follows from Bernstein's inequality that

$$\mathbb{P}\left\{|\langle\boldsymbol{u},\boldsymbol{\xi}^*\rangle| \ge 2A_0\sqrt{\frac{v_\delta\tau^{1-\delta}x}{n}} + (A_0/2)\frac{\tau x}{n}\right\} \le 2e^{-x} \quad \text{for any } x > 0.$$

Taking the union bound over $\boldsymbol{u}\in\mathcal{N}_{1/2}$, we obtain that with probability at least $1 - 5^d\cdot 2e^{-x}$,

$$\|\boldsymbol{\xi}^*\|_2 \le 4A_0\sqrt{\frac{v_\delta\tau^{1-\delta}x}{n}} + A_0\frac{\tau x}{n}. \tag{C.18}$$

Next, for the deterministic part $\|\boldsymbol{\Sigma}^{-1/2}\nabla\mathbb{E}\mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_2$, it is easy to see that

$$\big\|\boldsymbol{\Sigma}^{-1/2}\nabla\mathbb{E}\mathcal{L}_\tau(\boldsymbol{\beta}^*)\big\|_2 = \sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}}\frac{1}{n}\sum_{i=1}^n\mathbb{E}|\xi_i\langle\boldsymbol{u},\widetilde{\boldsymbol{x}}_i\rangle| \le v_\delta\tau^{-\delta}.$$

Combining this and (C.18) with $x = 2(d+t)$, we reach the bound (C.17) which holds with probability at least $1 - 2e^{-2t} \ge 1 - e^{-t}$. □

Lemma C.6. Assume Condition 1 holds with $0 < \delta \le 1$. Then with probability at least $1-2d^{-1}$,

$$\|\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_\infty \le \max_{1\le j\le d}\sigma_{jj}^{1/2}\Bigg(2\sqrt{2}A_0\sqrt{\frac{v_\delta\tau^{1-\delta}\log d}{n}}+A_0\frac{\tau\log d}{n}+v_\delta\tau^{-\delta}\Bigg).$$

Proof of Lemma C.6. The proof is based on Bernstein's inequality and the union bound. Define $\xi_i=\psi_\tau(\varepsilon_i)$ for $i=1,\ldots,n$ such that $\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)=-n^{-1}\sum_{i=1}^n\xi_i\boldsymbol{x}_i$. For every $1\le j\le d$, note that $|\mathbb{E}(\xi_i x_{ij})|=|\mathbb{E}\{\mathbb{E}(\xi_i|x_{ij})x_{ij}\}|\le\sigma_{jj}^{1/2}v_\delta\tau^{-\delta}$. Moreover, from the proof of Lemma C.5 we see that

$$\sum_{i=1}^n\mathbb{E}(\xi_i x_{ij})^2\le\sigma_{jj}nv_\delta\tau^{1-\delta},$$

$$\sum_{i=1}^n\mathbb{E}|\xi_i x_{ij}|^k\le\frac{k!}{2}2A_0^2\sigma_{jj}nv_\delta\tau^{1-\delta}(A_0\sigma_{jj}^{1/2}\tau/2)^{k-2}\quad\text{for } k\ge3.$$

By Bernstein's inequality, for any $x>0$ it holds

$$\Bigg|\frac{1}{n}\sum_{i=1}^n(\xi_i x_{ij}-\mathbb{E}\xi_i x_{ij})\Bigg|\le2A_0\sigma_{jj}^{1/2}\sqrt{\frac{v_\delta\tau^{1-\delta}x}{n}}+A_0\sigma_{jj}^{1/2}\frac{\tau x}{2n}$$

with probability at least $1-2e^{-x}$. By the union bound and taking $x=2\log d$ in the last display, we arrive at the stated result. □

C.2 Proof of Proposition 1

Define the error vector $\boldsymbol{\Delta}=\boldsymbol{\beta}^*-\boldsymbol{\beta}^*_\tau$ and function $h(\boldsymbol{\beta})=n^{-1}\sum_{i=1}^n\mathbb{E}\{\ell_\tau(y_i-\langle\boldsymbol{x}_i,\boldsymbol{\beta}\rangle)\}$, $\boldsymbol{\beta}\in\mathbb{R}^d$. By the optimality of $\boldsymbol{\beta}^*_\tau$ and the mean value theorem, we have $\nabla h(\boldsymbol{\beta}^*_\tau)=\mathbf{0}$ and thus

$$\langle\boldsymbol{\Delta},\nabla^2h(\widetilde{\boldsymbol{\beta}}_1)\boldsymbol{\Delta}\rangle=\langle\nabla h(\boldsymbol{\beta}^*)-\nabla h(\boldsymbol{\beta}^*_\tau),\boldsymbol{\Delta}\rangle=\langle\nabla h(\boldsymbol{\beta}^*),\boldsymbol{\Delta}\rangle=-\frac{1}{n}\sum_{i=1}^n\mathbb{E}\{\psi_\tau(\varepsilon_i)\}\langle\boldsymbol{x}_i,\boldsymbol{\Delta}\rangle, \tag{C.19}$$

where $\widetilde{\boldsymbol{\beta}}_1=\lambda\boldsymbol{\beta}^*+(1-\lambda)\boldsymbol{\beta}^*_\tau$ for some $0\le\lambda\le1$.

Case 1. First we consider the case of $0<\delta<1$. Since $\mathbb{E}(\varepsilon_i)=0$, we have $-\mathbb{E}\{\psi_\tau(\varepsilon_i)\}=\mathbb{E}\{\varepsilon_i 1(|\varepsilon_i|>\tau)-\tau1(\varepsilon_i>\tau)+\tau1(\varepsilon_i<-\tau)\}$ and therefore

$$|\mathbb{E}\{\psi_\tau(\varepsilon_i)\}|\le\mathbb{E}\big\{\big(|\varepsilon_i|-\tau\big)1\big(|\varepsilon_i|>\tau\big)\big\}\le v_{i,\delta}\tau^{-\delta}. \tag{C.20}$$

Taking $\widetilde{\varepsilon}_i = y_i - \langle \boldsymbol{x}_i, \widetilde{\boldsymbol{\beta}}_1 \rangle$, we see that

$$\nabla^2 h(\widetilde{\boldsymbol{\beta}}_1) = \mathbf{S}_n - \frac{1}{n}\sum_{i=1}^n \mathbb{P}\big(|\widetilde{\varepsilon}_i| > \tau\big)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}. \tag{C.21}$$

Note that

$$\begin{aligned}&\mathbb{E}\{\ell_\tau(\varepsilon_i)\}\\ &\le \mathbb{E}\bigg\{\frac{\tau^{1-\delta}}{2}|\varepsilon_i|^{1+\delta}1\big(|\varepsilon_i| \le \tau\big) + \bigg(\tau^{1-\delta}|\varepsilon_i|^{1+\delta} - \frac{\tau^{2-\tau}}{2}|\varepsilon_i|^{\delta}\bigg)1\big(|\varepsilon_i| > \tau\big)\bigg\} \le v_{i,\delta}\tau^{1-\delta}.\end{aligned}$$

This, together with the convexity of h implies that $h(\widetilde{\boldsymbol{\beta}}_1) \le \lambda h(\boldsymbol{\beta}^*) + (1-\lambda)h(\boldsymbol{\beta}^*_\tau) \le h(\boldsymbol{\beta}^*) \le v_\delta \tau^{1-\delta}$, where $v_\delta = n^{-1}\sum_{i=1}^n v_{i,\delta}$. For the lower bound, note that $h(\boldsymbol{\beta}) \ge n^{-1}\mathbb{E}\{(\tau|y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle| - \tau^2/2)\}1(|y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle| > \tau)\}$ for all $\boldsymbol{\beta} \in \mathbb{R}^d$. Putting these upper and lower bounds on $h(\widetilde{\boldsymbol{\beta}}_1)$ together yields

$$\frac{\tau}{n}\sum_{i=1}^n \mathbb{E}|\widetilde{\varepsilon}_i|1\big(|\widetilde{\varepsilon}_i| > \tau\big) \le \frac{\tau^2}{2n}\sum_{i=1}^n \mathbb{P}\big(|\widetilde{\varepsilon}_i| > \tau\big) + v_\delta\tau^{1-\delta},$$

as a consequence of which $n^{-1}\sum_{i=1}^n \mathbb{P}(|\widetilde{\varepsilon}_i| > \tau) \le 2v_\delta\tau^{-1-\delta}$. Combining this with (C.21), we deduce that as long as $\tau > (2v_\delta\widetilde{M}^2)^{1/(1+\delta)}$,

$$\begin{aligned}\boldsymbol{\Delta}^{\mathrm{T}}\,\nabla^2 h(\widetilde{\boldsymbol{\beta}}_1)\boldsymbol{\Delta} &\ge \|\mathbf{S}_n^{1/2}\boldsymbol{\Delta}\|_2^2 - \frac{1}{n}\sum_{i=1}^n \mathbb{P}\big(|\widetilde{\varepsilon}_i| > \tau\big)\langle \boldsymbol{x}_i, \boldsymbol{\Delta}\rangle^2\\ &\ge \|\mathbf{S}_n^{1/2}\boldsymbol{\Delta}\|_2^2 - 2\|\mathbf{S}_n^{1/2}\boldsymbol{\Delta}\|_2^2 \max_{1\le i\le n}\|\mathbf{S}_n^{-1/2}\boldsymbol{x}_i\|_2^2\, v_\delta\tau^{-1-\delta}\\ &\ge \big(1 - 2v_\delta\widetilde{M}^2\tau^{-1-\delta}\big)\|\mathbf{S}_n^{1/2}\boldsymbol{\Delta}\|_2^2.\end{aligned}$$

This provides a lower bound for the left-hand side of (C.19). On the other hand, using (C.20) and Hölder's inequality to bound the right-hand side of (C.19), the claim (7) for $0 < \delta < 1$ follows immediately.

Case 2. Next we assume $\delta \ge 1$ and note that $v_{i,1} = \mathbb{E}(\varepsilon_i^2)$. In this case, we have $\mathbb{E}\{\ell_\tau(\varepsilon_i)\} \le \frac{1}{2}v_{i,1}$ and $|\mathbb{E}\{\psi_\tau(\varepsilon_i)\}| \le v_{i,\delta}\tau^{-\delta}$. Then, following the same arguments as above, it can be shown that as long as $\tau > v_1^{1/2}m_n$,

$$\big(1 - v_1 m_n^2\tau^{-2}\big)\|\mathbf{S}_n^{1/2}\boldsymbol{\Delta}\|_2^2 \le \langle \boldsymbol{\Delta}, \nabla^2 h(\widetilde{\boldsymbol{\beta}}_1)\boldsymbol{\Delta}\rangle \le \|\mathbf{S}_n^{1/2}\boldsymbol{\Delta}\|_2\, v_\delta\tau^{-\delta}. \tag{C.22}$$

This proves (7) for $\delta \ge 1$ and hence completes the proof. □

C.3 Proof of Theorem 1

Without loss of generality, we assume $t \geq 1$ throughout the proof; otherwise, $3e^{-t} \geq 1$ and the stated result holds trivially. For simplicity, we write $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_\tau$. Note that for any prespecified $r > 0$, we can construct an intermediate estimator, denoted by $\widehat{\boldsymbol{\beta}}_{\tau,\eta} = \boldsymbol{\beta}^* + \eta(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$, such that $\|\widehat{\boldsymbol{\beta}}_{\tau,\eta} - \boldsymbol{\beta}^*\|_2 \leq r$. To see that, we take $\eta = 1$ if $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq r$; otherwise, we can always choose some $\eta \in (0,1)$ so that $\|\widehat{\boldsymbol{\beta}}_{\tau,\eta} - \boldsymbol{\beta}^*\|_2 = r$. Applying Lemma C.1 gives

$$\big\langle \nabla \mathcal{L}_\tau(\widehat{\boldsymbol{\beta}}_{\tau,\eta}) - \nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*), \widehat{\boldsymbol{\beta}}_{\tau,\eta} - \boldsymbol{\beta}^* \big\rangle \leq \eta \big\langle \nabla \mathcal{L}_\tau(\widehat{\boldsymbol{\beta}}) - \nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*), \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \big\rangle, \tag{C.23}$$

where $\nabla \mathcal{L}_\tau(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$ according to the Karush-Kuhn-Tucker condition. By the mean value theorem for vector-valued functions, we have

$$\nabla \mathcal{L}_\tau(\widehat{\boldsymbol{\beta}}_{\tau,\eta}) - \nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*) = \int_0^1 \nabla^2 \mathcal{L}_\tau((1-t)\boldsymbol{\beta}^* + t\widehat{\boldsymbol{\beta}}_{\tau,\eta})\, dt\, (\widehat{\boldsymbol{\beta}}_{\tau,\eta} - \boldsymbol{\beta}^*).$$

If, there exists some $a_0 > 0$ such that

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^d : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq r} \lambda_{\min}\big(\nabla^2 \mathcal{L}_\tau(\boldsymbol{\beta})\big) \geq a_0, \tag{C.24}$$

then we have $a_0 \|\widehat{\boldsymbol{\beta}}_{\tau,\eta} - \boldsymbol{\beta}^*\|_2^2 \leq \|\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_2 \|\widehat{\boldsymbol{\beta}}_{\tau,\eta} - \boldsymbol{\beta}^*\|_2$. Canceling the common factor on both sides yields

$$\big\|\widehat{\boldsymbol{\beta}}_{\tau,\eta} - \boldsymbol{\beta}\big\|_2 \leq a_0^{-1} \big\|\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)\big\|_2. \tag{C.25}$$

Define the random vector $\boldsymbol{\xi}^* = \nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)$, which can be written as

$$\boldsymbol{\xi}^* = -\frac{1}{n} \sum_{i=1}^n \psi_\tau(\varepsilon_i) \boldsymbol{x}_i.$$

By definition (B.1), $\psi_1(x) = \tau^{-1}\psi_\tau(\tau x)$. We write $\Psi_j = n^{-1} \sum_{i=1}^n (x_{ij}/L)\psi_1(\varepsilon_i/\tau)$ for $j = 1, \ldots, d$, such that $\|\boldsymbol{\xi}^*\|_2 \leq d^{1/2} \|\boldsymbol{\xi}^*\|_\infty = L d^{1/2} \tau \max_{1 \leq j \leq d} |\Psi_j|$. With $0 < \delta \leq 1$, it is easy to see that the function $\psi_1(\cdot)$ satisfies

$$-\log(1 - u + |u|^{1+\delta}) \leq \psi_1(u) \leq \log(1 + u + |u|^{1+\delta}) \tag{C.26}$$

for all $u \in \mathbb{R}$. It follows that

$$\begin{aligned}(x_{ij}/L)\psi_1(\varepsilon_i/\tau) \leq{}& (x_{ij}/L)1(x_{ij} \geq 0) \log(1 + \varepsilon_i/\tau + |\varepsilon_i/\tau|^{1+\delta}) \\ & - (x_{ij}/L)1(x_{ij} < 0) \log(1 - \varepsilon_i/\tau + |\varepsilon_i/\tau|^{1+\delta}).\end{aligned}$$

This, together with the inequality $(1+u)^v \leq 1 + uv$ for $u \geq -1$ and $0 < v \leq 1$, implies

$$
\begin{aligned}
&\exp\{(x_{ij}/L)\psi_1(\varepsilon_i/\tau)\} \\
&\le (1+\varepsilon_i/\tau+|\varepsilon_i/\tau|^{1+\delta})^{(x_{ij}/L)1(x_{ij}\ge 0)} + (1-\varepsilon_i/\tau+|\varepsilon_i/\tau|^{1+\delta})^{-(x_{ij}/L)1(x_{ij}<0)} \\
&\le 1+(\varepsilon_i/\tau)(x_{ij}/L)+|\varepsilon_i/\tau|^{1+\delta}.
\end{aligned}
$$

Consequently, we have

$$
\mathbb{E}\{\exp(n\Psi_j)\} = \prod_{i=1}^{n} \mathbb{E}\exp\{(x_{ij}/L)\psi_1(\varepsilon_i/\tau)\} \le \prod_{i=1}^{n}(1+v_{i,\delta}\tau^{-1-\delta}) \le \exp(v_\delta n\tau^{-1-\delta}),
$$

where we used the inequality $1+u \le e^u$ in the last step. For any $z \ge 0$, using Markov's inequality gives

$$
\mathbb{P}(\Psi_j \ge v_\delta z) \le \exp(-v_\delta n z)\mathbb{E}\{\exp(n\Psi_j)\} \le \exp\{v_\delta n(\tau^{-1-\delta}-z)\}.
$$

As long as $\tau \ge (2/z)^{1/(1+\delta)}$, we have $\mathbb{P}(\Psi_j \ge v_\delta z) \le e^{-v_\delta n z/2}$. On the other hand, it can be similarly shown that $\mathbb{P}(-\Psi_j \ge v_\delta z) \le e^{-v_\delta n z/2}$. For any $t>0$, taking $z = 2t/(v_\delta n)$ in these two inequalities yields that as long as $\tau \ge (v_\delta n/t)^{1/(1+\delta)}$,

$$
\begin{aligned}
&\mathbb{P}\big(\|\boldsymbol{\xi}^*\|_2 \ge 2Ld^{1/2}\tau n^{-1}t\big) \\
&\le \mathbb{P}\big(\|\boldsymbol{\xi}^*\|_\infty \ge 2L\tau n^{-1}t\big) \le \sum_{j=1}^{d}\mathbb{P}\big(|\Psi_j| \ge 2n^{-1}t\big) \le 2d\exp(-t).
\end{aligned}
\tag{C.27}
$$

Taking $r = \tau/(4\sqrt{2}M)$, it follows from Lemma C.2 and the definition of τ that with probability at least $1-e^{-t}$, (C.24) holds with $a_0 = c_l/2$ provided

$$
n \ge \max\big(8M^4c_l^{-2}, 2^{4+\delta}M^2c_l^{-1}\big)t.
$$

Combining (C.25) and (C.27) implies that, with probability at least $1-(2d+1)e^{-t}$,

$$
\big\|\widehat{\boldsymbol{\beta}}_{\tau,\eta} - \boldsymbol{\beta}^*\big\|_2 < 4Lc_l^{-1}d^{1/2}\tau n^{-1}t.
$$

Provided $n \ge 16\sqrt{2}c_l^{-1}LMd^{1/2}t$, the intermediate estimator $\widehat{\boldsymbol{\beta}}_{\tau,\eta}$ will lie in the interior of the ball with radius r. By our construction in the beginning of the proof, this enforces $\eta = 1$ and thus $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{\tau,\eta}$. □

C.4 Proof of Theorem 2

We start by defining a simple class of distributions for the response variable y as $\mathcal{P}_{c,\gamma} = \{\mathbb{P}_{c+}, \mathbb{P}_{c-}\}$, where

$$\mathbb{P}_c^+(\{0\}) = 1-\gamma,\ \mathbb{P}_c^+(\{c\}) = \gamma,\ \text{and } \mathbb{P}_c^-(\{0\}) = 1-\gamma,\ \mathbb{P}_c^-(\{-c\}) = \gamma.$$

Here, we suppress the dependence of $\mathbb{P}_c^+$ and $\mathbb{P}_c^-$ on γ for convenience. It follows that, for any $0 < \delta \le 1$, the $(1+\delta)$-th absolute central moment v_δ of y with law either $\mathbb{P}_c^+$ or $\mathbb{P}_c^-$ is

$$v_\delta = |c|^{1+\delta}\gamma(1-\gamma)\{\gamma^\delta + (1-\gamma)^\delta\}. \tag{C.28}$$

For $i = 1,\dots,n$, let (y_{1i}, y_{2i}) be independent pairs of real-valued random variables satisfying

$$\mathbb{P}(y_{1i} = y_{2i} = 0) = 1-\gamma,\ \mathbb{P}(y_{1i} = c_i, y_{2i} = -c_i) = \gamma,\ \text{and } y_{1i} \sim \mathbb{P}_{c_i}^+,\ y_{2i} \sim \mathbb{P}_{c_i}^-.$$

Let $\boldsymbol{y}_k = (y_{k1},\dots,y_{kn})^{\mathrm{T}}$ for $k = 1,2$, and $\xi \in (0, 1/2]$. Taking $\gamma = \log\{1/(2\xi)\}/(2n)$ with $\xi \ge e^{-n}/2$, we obtain $1-\gamma \ge 1/2$ and

$$\mathbb{P}\big(\boldsymbol{y}_1 = \boldsymbol{y}_2 = \mathbf{0}\big) = (1-\gamma)^n \ge \left\{\exp\left(\frac{-\gamma}{1-\gamma}\right)\right\}^n \ge 2\xi.$$

By assumption, we know that there is an n-dimensional vector $\mathbf{u} \in \{-1,+1\}^n$ with each coordinate taking -1 or 1 such that $\frac{1}{n}\|\mathbf{X}^{\mathrm{T}}\mathbf{u}\|_{\min} \ge \alpha$. Note that this assumption naturally holds for the mean model, where $\mathbf{X} = (1,\dots,1)^{\mathrm{T}}$ and α can be taken as 1. Now we take $\boldsymbol{c}$, $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ such that $\boldsymbol{c} = c\mathbf{u}$ for a $c > 0$, $\mathbf{X}\boldsymbol{\beta}_1^* = \boldsymbol{c}\gamma$ and $\boldsymbol{\beta}_2^* = -\boldsymbol{\beta}_1^*$, which indicates that

$$\boldsymbol{\beta}_1^* = \left(\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{u}, \quad \text{and}$$
$$\|\boldsymbol{\beta}_1^*\|_2 \ge c\gamma\left\|\left(\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{u}\right\|_2 \ge c\gamma\frac{d^{1/2}}{c_u}\|\mathbf{X}^{\mathrm{T}}\mathbf{u}/n\|_{\min} \ge c\gamma\frac{d^{1/2}\alpha}{c_u}.$$

Let $\widehat{\boldsymbol{\beta}}_k(\boldsymbol{y}_k)$ be any estimator possibly depending on ξ, then the above calculation yields

$$\begin{aligned}
&\max\left\{\mathbb{P}\big(\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \ge c\gamma c_u^{-1}d^{1/2}\alpha\big), \mathbb{P}\big(\|\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2^*\|_2 \ge c\gamma c_u^{-1}d^{1/2}\alpha\big)\right\}\\
&\ge \frac{1}{2}\mathbb{P}\Big(\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \ge c\gamma c_u^{-1}d^{1/2}\alpha \text{ or } \|\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2^*\|_2 \ge c\gamma c_u^{-1}d^{1/2}\alpha\Big)\\
&\ge \frac{1}{2}\mathbb{P}\big(\widehat{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{\beta}}_2\big) \ge \frac{1}{2}\mathbb{P}\big(\boldsymbol{y}_1 = \boldsymbol{y}_2\big) \ge \frac{1}{2}(1-\gamma)^n \ge \xi,
\end{aligned} \tag{C.29}$$

where we suppress the dependence of $\widehat{\boldsymbol{\beta}}_k$ on $\boldsymbol{y}_k$ for simplicity. Using the fact that $c\gamma \geq v_\delta^{1/(1+\delta)}(\gamma/2)^{\delta/(1+\delta)}$ further implies

$$\mathbb{P}\left[\big\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\big\|_2 \geq v_\delta^{1/(1+\delta)} \frac{d^{1/2}\alpha}{c_u}\left\{\frac{\log(1/(2\xi))}{2n}\right\}^{\delta/(1+\delta)}\right] \\ \bigvee \mathbb{P}\left[\big\|\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2^*\big\|_2 \geq v_\delta^{1/(1+\delta)} \frac{d^{1/2}\alpha}{c_u}\left\{\frac{\log(1/(2\xi))}{2n}\right\}^{\delta/(1+\delta)}\right] \geq \xi.$$

Now since $\mathcal{P}_{c,\gamma} \subseteq \mathcal{P}_\delta^{v_\delta}$, taking $\log\{1/(2\xi)\} = 2t$ implies the result for the case where $\delta \in (0,1]$. When $\delta > 1$, the second moment exists, and therefore using the fact that $v_1 < \infty$ completes the proof. □

C.5 Proof of Theorem 3

We start with the proof of Lemma 1.

Proof of Lemma 1. Let $\mathbf{H}_\tau = \nabla^2\mathcal{L}_\tau(\boldsymbol{\beta})$, where we suppress the dependence on $\boldsymbol{\beta}$. Then for any $(\boldsymbol{u},\boldsymbol{\beta}) \in \mathcal{C}(k,\gamma,r)$, we have

$$\begin{aligned} \langle \boldsymbol{u}, \mathbf{H}_\tau \boldsymbol{u}\rangle &= \boldsymbol{u}^{\mathrm{T}}\left\{\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} 1\big(|y_i - \langle \boldsymbol{x}, \boldsymbol{\beta}_i\rangle| \leq \tau\big)\right\}\boldsymbol{u} \\ &\geq \big\|\mathbf{S}_n^{1/2}\boldsymbol{u}\big\|_2^2 - \frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{u}, \boldsymbol{x}_i\rangle^2 1\big(|\langle \boldsymbol{x}_i, \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle| \geq \tau/2\big) - \frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{u}, \boldsymbol{x}_i\rangle^2 1\big(|\varepsilon_i| > \tau/2\big) \\ &\geq \big\|\mathbf{S}_n^{1/2}\boldsymbol{u}\big\|_2^2 - \frac{2r}{\tau}\max_{1\leq i\leq n}\|\boldsymbol{x}_i\|_\infty \|\mathbf{S}_n^{1/2}\boldsymbol{u}\|_2^2 - \max_{1\leq i\leq n}\langle \boldsymbol{u}, \boldsymbol{x}_i\rangle^2\,\frac{1}{n}\sum_{i=1}^n 1\big(|\varepsilon_i| > \tau/2\big). \end{aligned} \tag{C.30}$$

As $\|\boldsymbol{x}_i\|_\infty \leq L$ for any $1 \leq i \leq n$, we have

$$|\langle \boldsymbol{u}, \boldsymbol{x}_i\rangle| \leq \|\boldsymbol{x}_i\|_\infty \|\boldsymbol{u}\|_1 \leq (1+\gamma)\|\boldsymbol{x}_i\|_\infty\|\boldsymbol{u}_J\|_1 \leq Lk^{1/2}(1+\gamma).$$

Moreover, for any $t \geq 0$, applying Hoeffding's inequality yields that, with probability at least $1 - e^{-t}$,

$$\frac{1}{n}\sum_{i=1}^n 1\big(|\varepsilon_i| > \tau/2\big) \leq \left(\frac{2}{\tau}\right)^{1+\delta}\frac{1}{n}\sum_{i=1}^n v_{i,\delta} + \sqrt{\frac{t}{2n}} = \left(\frac{2}{\tau}\right)^{1+\delta} v_\delta + \sqrt{\frac{t}{2n}}.$$

Putting the above calculations together, we obtain

$$\langle \boldsymbol{u}, \mathbf{H}_\tau \boldsymbol{u}\rangle \geq \big\|\mathbf{S}_n^{1/2}\boldsymbol{u}\big\|_2^2 - 2\tau^{-1} rL\big\|\mathbf{S}_n^{1/2}\boldsymbol{u}\big\|_2^2 - k(1+\gamma)^2L^2\big(2^{1+\delta}v_\delta\tau^{-1-\delta} + \sqrt{t/2}\,n^{-1/2}\big).$$

Consequently, as long as $\tau \geq 8Lr$, the following inequality

$$\langle \boldsymbol{u}, \mathbf{H}_\tau \boldsymbol{u} \rangle \geq \frac{3}{4}\kappa_l - k(1+\gamma)^2 L^2 \big(2^{1+\delta} v_\delta \tau^{-1-\delta} + \sqrt{t/2}\, n^{-1/2}\big) \geq \frac{1}{2}\kappa_l, \tag{C.31}$$

holds uniformly over $(\boldsymbol{u}, \boldsymbol{\beta}) \in \mathcal{C}(k, \gamma, r)$ with probability at least $1 - e^{-t}$, where the last inequality in (C.31) holds whenever $\tau \gtrsim (1+\gamma)^{2/(1+\delta)} \kappa_l^{-1/(1+\delta)} (L^2 k v_\delta)^{1/(1+\delta)}$ and $n \gtrsim (1+\gamma)^4 \kappa_l^{-2} L^4 k^2 t$. On the other side, it can be easily shown that $\langle \boldsymbol{u}, \mathbf{H}_\tau \boldsymbol{u} \rangle \leq \kappa_u$. This completes the proof of the lemma. □

The following lemma is taken from Fan et al. (2018) with slight modification, which shows that the solution $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{\tau,\lambda}$ falls in a ℓ_1-cone.

Lemma C.7 (ℓ_1-cone Property). For any $\mathcal{E}$ such that $\mathcal{S} \subseteq \mathcal{E}$, if $\|\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_\infty \leq \lambda/2$, then $\|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{E}^c}\|_1 \leq 3\|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{E}}\|_1$.

Now we are ready to prove the theorem.

Proof of Theorem 3. It suffices to prove the statement for $\delta \in (0, 1]$. We start by constructing an intermediate estimator $\widehat{\boldsymbol{\beta}}_\eta = \boldsymbol{\beta}^* + \eta(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ such that $\|\widehat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_1 \leq r$ for some $r > 0$ to be specified. We take $\eta = 1$ if $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq r$, and choose $\eta \in (0, 1)$ so that $\|\widehat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_1 = r$ otherwise. Lemma C.7, $\widehat{\boldsymbol{\beta}}_\eta$ also falls in a ℓ_1-cone:

$$\|(\widehat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*)_{\mathcal{S}^c}\|_1 \leq 3\|(\widehat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*)_{\mathcal{S}}\|_1. \tag{C.32}$$

Under Condition 3, it follows from Lemma 1 that with probability at least $1 - e^{-t}$,

$$\frac{\kappa_l}{2}\|\widehat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_2^2 \leq \big\langle \nabla \mathcal{L}_\tau(\widehat{\boldsymbol{\beta}}_\eta) - \nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*), \widehat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^* \big\rangle$$

as long as $\tau \gtrsim \max\{(L^2 k v_\delta)^{1/(1+\delta)}, Lr\}$ and $n \gtrsim L^4 k^2 t$. Applying Lemma C.1 and following the same calculations as in Lemma B.7 of Fan et al. (2018), we obtain

$$\frac{\kappa_l}{2}\|\widehat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_2^2 \leq \big\{s^{1/2}\lambda + \|\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)_{\mathcal{S}}\|_2\big\}\|(\widehat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*)_{\mathcal{S}}\|_2,$$

which, combined with $\|\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)_{\mathcal{S}}\|_\infty \leq \lambda/2$, implies that

$$\|\widehat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_2 \leq 3\kappa_l^{-1} s^{1/2} \lambda. \tag{C.33}$$

Inequalities in (C.32) imply that $\|\widehat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_1 \leq 4\|(\widehat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*)_{\mathcal{S}}\|_1 \leq 4s^{1/2}\|\widehat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_2 \leq 12\kappa_l^{-1} s\lambda < r$. By the construction of $\widehat{\boldsymbol{\beta}}_\eta$, we conclude that $\widehat{\boldsymbol{\beta}}_\eta = \widehat{\boldsymbol{\beta}}$, and thus the stated result holds. It remains to bound the probability that event $\{\|\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)_{\mathcal{S}}\|_\infty \leq \lambda/2\}$ occurs. Recall the gradient of $\mathcal{L}_\tau$ evaluated at $\boldsymbol{\beta}^*$, i.e. $\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*) = -n^{-1}\sum_{i=1}^n \psi_\tau(\varepsilon_i)\boldsymbol{x}_i$. Following the same argument used in the proof of Theorem 1, we take $\tau = \tau_0(n/t)^{1/(1+\delta)}$ for some $\tau_0 \geq \nu_\delta$ and reach $\mathbb{P}\{\|\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)_{\mathcal{S}}\|_\infty \geq 2L\tau n^{-1} t\} \leq 2se^{-t}$. This, together with (C.33), proves (9).

Finally, taking $t = (1+c)\log d$ for some $c > 0$ yields that with probability at least $1-(2s+1)d^{-1-c}$, $\|\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)_{\mathcal{S}}\|_\infty \le 2L\tau_0\{(1+c)(\log d)/n\}^{\delta/(1+\delta)}$. As implied by Condition 3 with $k=2s$, we have $2s+1 \le d$ and thus (10) follows immediately. □

C.6 Proof of Theorem 4

The proof of this theorem follows the similar argument to that of Theorem 2. It suffices to prove the result for $\delta \in (0,1]$. Similar to the proof of Theorem 2, We start by defining a simple class of distributions for the response variable y as $\mathcal{P}_{c,\gamma} = \{\mathbb{P}_{c+}, \mathbb{P}_{c-}\}$, where

$$\mathbb{P}_c^+(\{0\}) = 1-\gamma,\ \mathbb{P}_c^+(\{c\}) = \gamma,\ \text{and}\ \mathbb{P}_c^-(\{0\}) = 1-\gamma,\ \mathbb{P}_c^-(\{-c\}) = \gamma.$$

Here, we suppress the dependence of $\mathbb{P}_c^+$ and $\mathbb{P}_c^-$ on γ for convenience. It follows that, for any $0<\delta\le 1$, the $(1+\delta)$-th absolute central moment v_δ of y with law either $\mathbb{P}_c^+$ or $\mathbb{P}_c^-$ is

$$v_\delta = |c|^{1+\delta}\gamma(1-\gamma)\{\gamma^\delta + (1-\gamma)^\delta\}. \tag{C.34}$$

We define the following s-sparse sign-ball $\mathcal{U}_n$ as

$$\mathcal{U}_n = \big\{\mathbf{u} : \mathbf{u} \in \{-1,1\}^n\big\}.$$

By assumption, there exist $\mathbf{u} \in \mathcal{U}_n$ and $\mathcal{A}$ with $|\mathcal{A}| = s$ such that $\|\mathbf{X}_{\mathcal{A}}^{\mathrm{T}}\mathbf{u}\|_{\min}/n \ge \alpha$. Take $\boldsymbol{\beta}_1^*$, $\boldsymbol{\beta}_2^*$ supported on $\mathcal{A}$ and $\boldsymbol{c} \in \mathbb{R}^n$ such that $\boldsymbol{c} = c\mathbf{u}$ for a $c>0$, $\mathbf{X}\boldsymbol{\beta}_1^* = \boldsymbol{c}\gamma$ and $\boldsymbol{\beta}_2^* = -\boldsymbol{\beta}_1^*$. Let $\mathbb{P}^+$ be the distribution of $\boldsymbol{y}_1 = \mathbf{X}\boldsymbol{\beta}_1^* + \boldsymbol{\varepsilon}$ and $\mathbb{P}_-$ that of $\boldsymbol{y}_2 = \mathbf{X}\boldsymbol{\beta}_2^* + \boldsymbol{\varepsilon}$. Clearly, we have

$$\mathbb{E}(\varepsilon_i) = 0 \ \text{ and } \ \mathbb{E}(|\varepsilon_i|^{1+\delta}) = c^{1+\delta}\gamma(1-\gamma)\{\gamma^\delta + (1-\gamma)^\delta\}.$$

Let $\mathcal{A}$ be the support of $\boldsymbol{\beta}_1^*$. Then, we have

$$(\boldsymbol{\beta}_1^*)_{\mathcal{A}} = c\gamma\Big(\frac{1}{n}\mathbf{X}_{\mathcal{A}}^{\mathrm{T}}\mathbf{X}_{\mathcal{A}}\Big)^{-1}\frac{1}{n}\mathbf{X}_{\mathcal{A}}^{\mathrm{T}}\mathbf{u},\ \text{and}$$
$$\|\boldsymbol{\beta}_1^*\|_2 \ge c\gamma\,\kappa_u^{-1}s^{1/2}\|\mathbf{X}_{\mathcal{A}}^{\mathrm{T}}\mathbf{u}/n\|_{\min} \ge c\gamma\,\kappa_u^{-1}s^{1/2}\alpha.$$

Let $\widehat{\boldsymbol{\beta}}_k(\boldsymbol{y}_k)$ be any s-sparse estimator. With the above setup, we have

$$\begin{aligned}&\max\Big\{\mathbb{P}\big(\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \ge c\gamma\,\kappa_u^{-1}s^{1/2}\alpha\big), \mathbb{P}\big(\|\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2^*\|_2 \ge c\gamma\,\kappa_u^{-1}s^{1/2}\alpha\big)\Big\}\\ &\ge \frac{1}{2}\mathbb{P}\Big(\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2^2 \ge c\gamma\,\kappa_u^{-1}s^{1/2}\alpha \text{ or } \|\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2^*\|_2 \ge c\gamma\,\kappa_u^{-1}s^{1/2}\alpha\Big)\\ &\ge \frac{1}{2}\mathbb{P}\big(\widehat{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{\beta}}_2\big) \ge \frac{1}{2}\mathbb{P}\big(\boldsymbol{y}_1 = \boldsymbol{y}_2 = \mathbf{0}\big)\end{aligned} \tag{C.35}$$

where we suppress the dependence of $\widehat{\boldsymbol{\beta}}_k$ on $\boldsymbol{y}_k$ for simplicity. For the last quantity in the displayed inequality above, taking $\gamma = \log\{1/(2t)\}/(2n)$ with $t \geq e^{-n}/2$, we obtain $1-\gamma \geq 1/2$ and

$$\mathbb{P}\big(\boldsymbol{y}_1 = \boldsymbol{y}_2 = \mathbf{0}\big) = (1-\gamma)^n \geq \left\{\exp\left(\frac{-\gamma}{1-\gamma}\right)\right\}^n \geq 2t.$$

Using the fact that $c\gamma \geq v_\delta^{1/(1+\delta)}(\gamma/2)^{\delta/(1+\delta)}$, this further implies

$$\begin{aligned}&\mathbb{P}\left[\big\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\big\|_2 \geq v_\delta^{1/(1+\delta)}\kappa_u^{-1}\alpha s^{1/2}\left\{\frac{\log\{1/(2t)\}}{2n}\right\}^{\delta/(1+\delta)}\right] \\ &\qquad \bigvee \mathbb{P}\left[\big\|\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2^*\big\|_2 \geq v_\delta^{1/(1+\delta)}\kappa_u^{-1}\alpha s^{1/2}\left\{\frac{\log\{1/(2t)\}}{2n}\right\}^{\delta/(1+\delta)}\right] \geq t.\end{aligned}$$

Now since $\mathcal{P}_{c,\gamma} \subseteq \mathcal{P}_\delta^{v_\delta}$, taking $t = d^{-A}/2$ implies the result for the case where $\delta \in (0,1]$. When $\delta > 1$, the second moment exists. Thus using $v_1 < \infty$ completes the proof. □

C.7 Proof of Theorem 5

The proof is almost identical to that of Theorem 3. We only need to derive a probability bound for the event $\{\|\boldsymbol{\xi}_{\mathcal{S}}^*\|_\infty \leq \lambda/2\}$ under the assumed scaling and moment conditions, where $\boldsymbol{\xi}^* := \nabla\mathcal{L}_\tau^\varpi(\boldsymbol{\beta}^*)$.

Recall that $\boldsymbol{x}_i^\varpi = (x_{i1}^\varpi, \ldots, x_{id}^\varpi)^{\mathrm{T}}$ with $x_{ij}^\varpi = \psi_\varpi(x_{ij})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, d$. Define $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{id})^{\mathrm{T}} = \boldsymbol{x}_i - \boldsymbol{x}_i^\varpi$, where $z_{ij} = \{x_{ij} - \varpi\,\mathrm{sign}(x_{ij})\}1(|x_{ij}| > \varpi)$. Moreover, write $\boldsymbol{z}_{i\mathcal{S}} = (z_{ij}1(j \in \mathcal{S})) \in \mathbb{R}^d$ and $\epsilon_i = \varepsilon_i + \langle \boldsymbol{z}_i, \boldsymbol{\beta}^*\rangle$. In this notation, we have $\boldsymbol{\xi}^* = -n^{-1}\sum_{i=1}^n \psi_\tau(\epsilon_i)\boldsymbol{x}_i^\varpi$. From the identity $\mathbb{E}\{\psi_\tau(\epsilon_i)x_{ij}^\varpi\} = \mathbb{E}\{\langle \boldsymbol{z}_i, \boldsymbol{\beta}^*\rangle x_{ij}^\varpi\} - \mathbb{E}\{\epsilon_i - \tau\,\mathrm{sign}(\epsilon_i)\}x_{ij}^\varpi 1(|\epsilon_i| > \tau)$, we see that

$$\begin{aligned}&|\mathbb{E}\{\psi_\tau(\epsilon_i)x_{ij}^\varpi\}| \\ &\leq M_4\|\boldsymbol{\beta}^*\|_1\varpi^{-2} + \tau^{-2}\mathbb{E}(|\epsilon_i|^3|x_{ij}^\varpi|) \\ &\leq M_4\|\boldsymbol{\beta}^*\|_1\varpi^{-2} + 4\tau^{-2}\big\{\mathbb{E}(|\varepsilon_i|^3|x_{ij}^\varpi|) + \|\boldsymbol{\beta}^*\|_2^3\,\mathbb{E}(\|\boldsymbol{z}_i\|_2^3|x_{ij}^\varpi|)\big\} \\ &\leq M_4\|\boldsymbol{\beta}^*\|_2\, s^{1/2}\varpi^{-2} + 4\tau^{-2}\big\{v_2M_2^{1/2} + M_4\|\boldsymbol{\beta}^*\|_2^3\, s^{3/2}\big\}.\end{aligned}$$

Then it holds

$$\|\mathbb{E}(\boldsymbol{\xi}^*)\|_\infty \leq M_4\|\boldsymbol{\beta}^*\|_2\, s^{1/2}\varpi^{-2} + 4\tau^{-2}\big\{v_2M_2^{1/2} + M_4\|\boldsymbol{\beta}^*\|_2^3\, s^{3/2}\big\}. \tag{C.36}$$

For each j fixed, note that

$$\sum_{i=1}^n \mathbb{E}\{x_{ij}^{\varpi}\psi_\tau(\epsilon_i)\}^2 \le \sum_{i=1}^n \mathbb{E}\{x_{ij}^2(\varepsilon_i^2 + \langle \boldsymbol{z}_i, \boldsymbol{\beta}^*\rangle^2)\} \le n\big(\sigma^2 M_2 + M_4 \|\boldsymbol{\beta}^*\|_2^2\, s\big),$$

$$\text{and} \quad \sum_{i=1}^n \mathbb{E}|x_{ij}^{\varpi}\psi_\tau(\epsilon_i)|^k \le \frac{k!}{2}(\varpi\tau/2)^{k-2} n\big(\sigma^2 M_2 + M_4\|\boldsymbol{\beta}^*\|_2^2\, s\big) \quad \text{for all } k \ge 3.$$

Applying Bernstein's inequality gives

$$\left|\frac{1}{n}\sum_{i=1}^n \big[x_{ij}^{\varpi}\psi_\tau(\epsilon_i) - \mathbb{E}\{x_{ij}^{\varpi}\psi_\tau(\epsilon_i)\}\big]\right| \le \big(\sigma^2 M_2 + M_4\|\boldsymbol{\beta}^*\|_2^2\, s\big)^{1/2}\sqrt{\frac{2t}{n}} + \varpi\tau\frac{t}{2n}$$

with probability at least $1 - 2e^{-t}$. Taking the union bound over $j \in \mathcal{S}$, we obtain that, with probability at least $1 - 2se^{-t}$,

$$\|\boldsymbol{\xi}_{\mathcal{S}}^* - \mathbb{E}(\boldsymbol{\xi}_{\mathcal{S}}^*)\|_\infty \le \big(2\sigma^2 M_2 + 2M_4\|\boldsymbol{\beta}^*\|_2^2\, s\big)^{1/2}\sqrt{\frac{t}{n}} + \varpi\tau\frac{t}{2n}.$$

This, together with (C.36), implies that $\mathbb{P}\{\mathcal{E}(\tau,\varpi,\lambda)\} \ge 1 - 2se^{-t}$ provided

$$\lambda \ge 2M_4\|\boldsymbol{\beta}^*\|_2\, s^{1/2}\varpi^{-2} + 8\big\{v_2 M_2^{1/2} + M_4\|\boldsymbol{\beta}^*\|_2^3\, s^{3/2}\big\}\tau^{-2}$$
$$+ 2\big(2\sigma^2 M_2 + 2M_4\|\boldsymbol{\beta}^*\|_2^2\, s\big)^{1/2}\sqrt{\frac{t}{n}} + \varpi\tau\frac{t}{n}.$$

This is the stated result. □

C.8 Proof of Theorem B.1

To begin with, define the parameter set $\Theta_0(r) = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2} \le r\}$ for some $r > 0$ to be specified, and let $\widehat{\boldsymbol{\beta}}_{\tau,\eta} \in \Theta_0(r)$ be the intermediate estimator introduced in the proof of Theorem 1.

Proof of (B.2). In view of (C.23) and (C.25), lying in the heart of the arguments is to derive deviation inequalities for $\|\boldsymbol{\Sigma}^{-1/2}\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_2$ under the moment condition that $v_\delta < \infty$ for some $0 < \delta \le 1$, and to establish the restricted strong convexity for the Huber loss $\mathcal{L}_\tau$, i.e. there exists some $\kappa > 0$ such that

$$\langle \nabla\mathcal{L}_\tau(\boldsymbol{\beta}) - \nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle \ge \kappa\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2$$

holds uniformly over $\boldsymbol{\beta}$ in a neighborhood of $\boldsymbol{\beta}^*$.

First, from (C.17) in Lemma C.5 we see that

$$\big\|\boldsymbol{\Sigma}^{-1/2}\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)\big\|_2 < r_0 := 4\sqrt{2}A_0 v_\delta^{1/2}\tau^{(1-\delta)/2}\sqrt{\frac{d+t}{n}} + 2A_0\tau\frac{d+t}{n} + \frac{v_\delta}{\tau^\delta}$$

with probability at least $1-e^{-t}$. Next, since $\widehat{\boldsymbol{\beta}}_{\tau,\eta} \in \Theta_0(r)$ and according to Lemma C.3, we take $r = \tau/(4A_1^2)$ such that under the scaling (C.2),

$$\big\langle \nabla \mathcal{L}_\tau(\widehat{\boldsymbol{\beta}}_{\tau,\eta}) - \nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*), \widehat{\boldsymbol{\beta}}_{\tau,\eta} - \boldsymbol{\beta}^* \big\rangle \geq \frac{1}{4} \big\| \widehat{\boldsymbol{\beta}}_{\tau,\eta} - \boldsymbol{\beta} \big\|_{\boldsymbol{\Sigma},2}^2$$

with probability at least $1-e^{-t}$. Together, the last two displays and (C.23) imply that with probability at least $1-2e^{-t}$, $\|\widehat{\boldsymbol{\beta}}_{\tau,\eta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2} \leq 4r_0 < r$ provided $n \geq C_1(d+t)$, where $C_1 > 0$ is a constant depending only on A_0. Following the same arguments as we used in the proof of Theorem 1, this proves (B.2).

Proof of (B.3). From the preceding proof, we see that

$$\mathbb{P}\big\{ \widehat{\boldsymbol{\beta}} \in \Theta_0(r_1) \big\} \geq 1 - 2e^{-t} \tag{C.37}$$

as long as $n \geq C_1(d+t)$, where $r_1 = 4r_0$. Moreover, define random processes $\boldsymbol{\zeta}(\boldsymbol{\beta}) = \mathcal{L}_\tau(\boldsymbol{\beta}) - \mathbb{E}\mathcal{L}_\tau(\boldsymbol{\beta})$ and

$$\boldsymbol{B}(\boldsymbol{\beta}) = \boldsymbol{\Sigma}^{-1/2}\big\{ \nabla \mathcal{L}_\tau(\boldsymbol{\beta}) - \nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*) \big\} - \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*). \tag{C.38}$$

To bound $\|\boldsymbol{B}(\widehat{\boldsymbol{\beta}}_\tau)\|_2 = \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*) + \boldsymbol{\Sigma}^{-1/2}\nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_2$, the key is to bound the supremum of the empirical process $\{\boldsymbol{B}(\boldsymbol{\beta}) : \boldsymbol{\beta} \in \Theta_0(r)\}$. To that end, we deal with $\boldsymbol{B}(\boldsymbol{\beta}) - \mathbb{E}\{\boldsymbol{B}(\boldsymbol{\beta})\}$ and $\mathbb{E}\{\boldsymbol{B}(\boldsymbol{\beta})\}$ separately, starting with the latter. By the mean value theorem,

$$\begin{aligned} \mathbb{E}\{\boldsymbol{B}(\boldsymbol{\beta})\} &= \boldsymbol{\Sigma}^{-1/2}\big\{ \nabla \mathbb{E}\mathcal{L}_\tau(\boldsymbol{\beta}) - \nabla \mathbb{E}\mathcal{L}_\tau(\boldsymbol{\beta}^*) \big\} - \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \\ &= \big\{ \boldsymbol{\Sigma}^{-1/2} \nabla^2 \mathbb{E}\mathcal{L}_\tau(\widetilde{\boldsymbol{\beta}}) \boldsymbol{\Sigma}^{-1/2} - \mathbf{I}_d \big\} \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*), \end{aligned}$$

where $\widetilde{\boldsymbol{\beta}}$ is a convex combination of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$. Therefore,

$$\sup_{\boldsymbol{\beta} \in \Theta_0(r)} \big\| \mathbb{E}\{\boldsymbol{B}(\boldsymbol{\beta})\} \big\|_2 \leq r \times \sup_{\boldsymbol{\beta} \in \Theta_0(r)} \big\| \boldsymbol{\Sigma}^{-1/2} \nabla^2 \mathbb{E}\mathcal{L}_\tau(\boldsymbol{\beta}) \boldsymbol{\Sigma}^{-1/2} - \mathbf{I}_d \big\|.$$

For $\boldsymbol{\beta} \in \Theta_0(r)$ and $\boldsymbol{u} \in \mathbb{S}^{d-1}$, write $\boldsymbol{\delta} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$ such that $\|\boldsymbol{\delta}\|_2 \leq r$. Let $A_1 > 0$ be the constant in Lemma C.3 that scales as A_0. It follows that

$$\begin{aligned} &\big| \boldsymbol{u}^{\mathrm{T}} \big\{ \boldsymbol{\Sigma}^{-1/2} \nabla \mathbb{E}\mathcal{L}_\tau(\boldsymbol{\beta}) \boldsymbol{\Sigma}^{-1/2} - \mathbf{I}_d \big\} \boldsymbol{u} \big| = \frac{1}{n} \sum_{i=1}^n \mathbb{E}\big\{ 1\big(|y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle| > \tau \big) \langle \boldsymbol{u}, \widetilde{\boldsymbol{x}}_i \rangle^2 \big\} \\ &\leq \frac{1}{n\tau^2} \sum_{i=1}^n \big\{ v_{i,1} + \mathbb{E} \langle \boldsymbol{\delta}, \widetilde{\boldsymbol{x}}_i \rangle^2 \langle \boldsymbol{u}, \widetilde{\boldsymbol{x}}_i \rangle^2 \big\} \leq v_1 \tau^{-2} + A_1^4 \tau^{-2} \|\boldsymbol{\delta}\|_2^2 \leq v_1 \tau^{-2} + A_1^4 r^2 \tau^{-2}, \end{aligned}$$

which, further implies

$$\sup_{\boldsymbol{\beta}\in\Theta_0(r)} \big\|\mathbb{E}\{\boldsymbol{B}(\boldsymbol{\beta})\}\big\|_2 \le v_1\tau^{-2} + A_1^4 r^2\tau^{-2}. \tag{C.39}$$

Next, we consider $\boldsymbol{B}(\boldsymbol{\beta}) - \mathbb{E}\{\boldsymbol{B}(\boldsymbol{\beta})\} = \boldsymbol{\Sigma}^{-1/2}\{\nabla\boldsymbol{\zeta}(\boldsymbol{\beta}) - \nabla\boldsymbol{\zeta}(\boldsymbol{\beta}^*)\}$. With $\boldsymbol{\delta} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)$, define a new process $\overline{\boldsymbol{B}}(\boldsymbol{\delta}) = \boldsymbol{B}(\boldsymbol{\beta}) - \mathbb{E}\{\boldsymbol{B}(\boldsymbol{\beta})\}$, satisfying $\overline{\boldsymbol{B}}(\mathbf{0}) = \mathbf{0}$ and $\mathbb{E}\{\overline{\boldsymbol{B}}(\boldsymbol{\delta})\} = \mathbf{0}$. Note that, for every $\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{d-1}$ and $\lambda\in\mathbb{R}$,

$$\begin{aligned}
&\mathbb{E}\exp\big\{\lambda\sqrt{n}\,\boldsymbol{u}^{\mathrm{T}}\nabla_{\boldsymbol{\delta}}\overline{\boldsymbol{B}}(\boldsymbol{\delta})\boldsymbol{v}\big\}\\
&\le \prod_{i=1}^n\bigg(1+\frac{\lambda^2}{n}\mathbb{E}\Big[\big\{\langle\boldsymbol{u},\widetilde{\boldsymbol{x}}_i\rangle^2\langle\boldsymbol{v},\widetilde{\boldsymbol{x}}_i\rangle^2 + \big(\mathbb{E}|\langle\boldsymbol{u},\widetilde{\boldsymbol{x}}\rangle\langle\boldsymbol{v},\widetilde{\boldsymbol{x}}\rangle|\big)^2\big\}e^{\frac{|\lambda|}{\sqrt{n}}(|\langle\boldsymbol{u},\widetilde{\boldsymbol{x}}_i\rangle\langle\boldsymbol{v},\widetilde{\boldsymbol{x}}_i\rangle|+\mathbb{E}|\langle\boldsymbol{u},\widetilde{\boldsymbol{x}}\rangle\langle\boldsymbol{v},\widetilde{\boldsymbol{x}}\rangle|)}\Big]\bigg)\\
&\le \prod_{i=1}^n\bigg\{1+e^{\frac{|\lambda|}{\sqrt{n}}}\frac{\lambda^2}{n}\mathbb{E}\big(e^{\frac{|\lambda|}{\sqrt{n}}|\langle\boldsymbol{u},\widetilde{\boldsymbol{x}}_i\rangle\langle\boldsymbol{v},\widetilde{\boldsymbol{x}}_i\rangle|}\big) + e^{\frac{|\lambda|}{\sqrt{n}}}\frac{\lambda^2}{n}\mathbb{E}\big(\langle\boldsymbol{u},\widetilde{\boldsymbol{x}}_i\rangle^2\langle\boldsymbol{v},\widetilde{\boldsymbol{x}}_i\rangle^2 e^{\frac{|\lambda|}{\sqrt{n}}|\langle\boldsymbol{u},\widetilde{\boldsymbol{x}}_i\rangle\langle\boldsymbol{v},\widetilde{\boldsymbol{x}}_i\rangle|}\big)\bigg\}\\
&\le \prod_{i=1}^n\bigg\{1+e^{\frac{|\lambda|}{\sqrt{n}}}\frac{\lambda^2}{n}\max_{\boldsymbol{w}\in\mathbb{S}^{d-1}}\mathbb{E}\big(e^{\frac{|\lambda|}{\sqrt{n}}\langle\boldsymbol{w},\widetilde{\boldsymbol{x}}\rangle^2}\big) + e^{\frac{|\lambda|}{\sqrt{n}}}\frac{\lambda^2}{n}\max_{\boldsymbol{w}\in\mathbb{S}^{d-1}}\mathbb{E}\big(\langle\boldsymbol{w},\widetilde{\boldsymbol{x}}\rangle^4 e^{\frac{|\lambda|}{\sqrt{n}}\langle\boldsymbol{w},\widetilde{\boldsymbol{x}}\rangle^2}\big)\bigg\}\\
&\le \exp\bigg\{e^{\frac{|\lambda|}{\sqrt{n}}}\lambda^2\max_{\boldsymbol{w}\in\mathbb{S}^{d-1}}\mathbb{E}\big(e^{\frac{|\lambda|}{\sqrt{n}}\langle\boldsymbol{w},\widetilde{\boldsymbol{x}}\rangle^2}\big) + e^{\frac{|\lambda|}{\sqrt{n}}}\lambda^2\max_{\boldsymbol{w}\in\mathbb{S}^{d-1}}\mathbb{E}\big(\langle\boldsymbol{w},\widetilde{\boldsymbol{x}}\rangle^4 e^{\frac{|\lambda|}{\sqrt{n}}\langle\boldsymbol{w},\widetilde{\boldsymbol{x}}\rangle^2}\big)\bigg\}.
\end{aligned}$$

Under Condition 1, there exist constants $C_2, C_3 > 0$ depending only on A_0 such that, for any $|\lambda| \le \sqrt{n/C_2}$,

$$\sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{d-1}}\mathbb{E}\exp\big\{\lambda\sqrt{n}\,\boldsymbol{u}^{\mathrm{T}}\nabla_{\boldsymbol{\delta}}\overline{\boldsymbol{B}}(\boldsymbol{\delta})\boldsymbol{v}\big\} \le \exp(C_3^2\lambda^2/2).$$

With the above preparations and applying Theorem A.3 in Spokoiny (2013), we reach

$$\mathbb{P}\bigg\{\sup_{\boldsymbol{\beta}\in\Theta_0(r)}\|\boldsymbol{B}(\boldsymbol{\beta}) - \mathbb{E}\{\boldsymbol{B}(\boldsymbol{\beta})\}\|_2 \ge 6C_3(8d+2t)^{1/2}r\bigg\} \le e^{-t}$$

as long as $n \ge C_2(8d+2t)$. Together with (C.39), this yields

$$\begin{aligned}
\sup_{\boldsymbol{\beta}\in\Theta_0(r_1)}&\big\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*) - \boldsymbol{\Sigma}^{-1/2}\big\{\nabla\mathcal{L}_\tau(\boldsymbol{\beta}) - \nabla\mathcal{L}_\tau(\boldsymbol{\beta}^*)\big\}\big\|_2\\
&\le v_1\tau^{-2}r_1 + A_1^4\tau^{-2}r_1^3 + 6C_3(8d+2t)^{1/2}n^{-1/2}r_1
\end{aligned}$$

with probability at least $1-e^{-t}$. Combine this bound with (C.37) to obtain the stated result (B.3). □

C.9 Proof of Proposition B.1

Since $\mathbb{E}(\varepsilon) = 0$, we have $\mathbb{E}\{\psi_\tau(\varepsilon)\} = -\mathbb{E}\{(\varepsilon-\tau)1(\varepsilon>\tau)\} + \mathbb{E}\{(-\varepsilon-\tau)1(\varepsilon<-\tau)\}$. Thus, for any $2\le q\le 2+\kappa$, $|\mathbb{E}\psi_\tau(\varepsilon)| \le \mathbb{E}\{|\varepsilon|-\tau)1(|\varepsilon|>\tau)\} \le \tau^{1-q}\,\mathbb{E}(|\varepsilon|^q)$. In particular, taking q to be 2 and $2+\kappa$ proves the first conclusion. Next, note that $\mathbb{E}\{\psi_\tau^2(\varepsilon)\} = \mathbb{E}(\varepsilon^2) - \{\mathbb{E}\varepsilon^2 1(|\varepsilon| > \tau) - \tau^2 \mathbb{P}(|\varepsilon| > \tau)\}$. Letting $\eta = |\varepsilon|$, we deduce that

$$\mathbb{E}\{\eta^2 1(\eta > \tau)\} = 2\mathbb{E}\int_0^\infty 1(\eta > y) 1(\eta > \tau) y \, dy$$
$$= 2\mathbb{P}(\eta > \tau)\int_0^\tau y\,dy + 2\int_\tau^\infty y\mathbb{P}(\eta > y)\,dy = \tau^2\mathbb{P}(\eta > \tau) + 2\int_\tau^\infty y\mathbb{P}(\eta > y)\,dy.$$

By Markov's inequality, $\int_\tau^\infty y\mathbb{P}(\eta > y)\,dy \le \mathbb{E}(\eta^{2+\kappa})\int_\tau^\infty y^{-1-\kappa}\,dy = \kappa^{-1}\tau^{-\kappa}\,\mathbb{E}(\eta^{2+\kappa})$. Putting the above calculations together proves the second inequality. □

C.10 Proof of Theorem B.2

For simplicity, we write $\widehat{\boldsymbol\beta} = \widehat{\boldsymbol\beta}_{\tau,\lambda}$ and assume without loss of generality that $0 < \delta \le 1$. As in the proof of Theorem B.1, we construct an intermediate estimator $\widetilde{\boldsymbol\beta}_\eta = \boldsymbol\beta^* + \eta(\widehat{\boldsymbol\beta} - \boldsymbol\beta^*)$ satisfying $\|\widetilde{\boldsymbol\beta}_\eta - \boldsymbol\beta^*\|_{\boldsymbol\Sigma,2} \le r$ for some $r > 0$ to be specified. We take $\eta = 1$ if $\|\widehat{\boldsymbol\beta} - \boldsymbol\beta^*\|_{\boldsymbol\Sigma,2} \le r$; otherwise if $\|\widehat{\boldsymbol\beta} - \boldsymbol\beta^*\|_{\boldsymbol\Sigma,2} > r$, there exists $\eta \in (0,1)$ such that $\|\widetilde{\boldsymbol\beta}_\eta - \boldsymbol\beta^*\|_{\boldsymbol\Sigma,2} = r$. Lemma C.1 demonstrates that

$$\langle \nabla\mathcal{L}_\tau(\widetilde{\boldsymbol\beta}_\eta) - \nabla\mathcal{L}_\tau(\boldsymbol\beta^*), \widetilde{\boldsymbol\beta}_\eta - \boldsymbol\beta^*\rangle \le \eta\langle \nabla\mathcal{L}_\tau(\widehat{\boldsymbol\beta}) - \nabla\mathcal{L}_\tau(\boldsymbol\beta^*), \widehat{\boldsymbol\beta} - \boldsymbol\beta^*\rangle. \tag{C.40}$$

Next, let $\mathcal{S} \subseteq \{1,\ldots,d\}$ be the support of $\boldsymbol\beta^*$ and define the ℓ_1-cone $\mathcal{C} \subseteq \mathbb{R}^d$:

$$\mathcal{C} = \big\{\boldsymbol\beta \in \mathbb{R}^d : \|(\boldsymbol\beta - \boldsymbol\beta^*)_{\mathcal{S}^c}\|_1 \le 3\|(\boldsymbol\beta - \boldsymbol\beta^*)_{\mathcal{S}}\|_1\big\}.$$

We claim that

$$\widehat{\boldsymbol\beta} \in \mathcal{C} \ \text{ on the event } \ \{\lambda \ge 2\|\nabla\mathcal{L}_\tau(\boldsymbol\beta^*)\|_\infty\}, \tag{C.41}$$

from which it follows

$$\|\widehat{\boldsymbol\delta}\|_1 = \|\widehat{\boldsymbol\delta}_{\mathcal{S}}\|_1 + \|\widehat{\boldsymbol\delta}_{\mathcal{S}^c}\|_1 \le 4\|\widehat{\boldsymbol\delta}_{\mathcal{S}}\|_1 \le 4\sqrt{s}\,\|\widehat{\boldsymbol\delta}\|_2, \tag{C.42}$$

where $\widehat{\boldsymbol\delta} := \widehat{\boldsymbol\beta} - \boldsymbol\beta^*$. To prove (C.41), first, from the optimality of $\widehat{\boldsymbol\beta}$ we see that

$$\mathcal{L}_\tau(\widehat{\boldsymbol\beta}) - \mathcal{L}_\tau(\boldsymbol\beta^*) \le \lambda\big(\|\boldsymbol\beta^*\|_1 - \|\widehat{\boldsymbol\beta}\|_1\big). \tag{C.43}$$

By direct calculation, we have

$$\|\widehat{\boldsymbol\beta}\|_1 - \|\boldsymbol\beta^*\|_1 \ge \|\boldsymbol\beta^*_{\mathcal{S}} + \widehat{\boldsymbol\delta}_{\mathcal{S}^c}\|_1 - \|\boldsymbol\beta^*_{\mathcal{S}^c}\|_1 - \|\widehat{\boldsymbol\delta}_{\mathcal{S}}\|_1 - \big(\|\boldsymbol\beta^*_{\mathcal{S}}\|_1 + \|\boldsymbol\beta^*_{\mathcal{S}^c}\|_1\big)$$
$$\ge \|\widehat{\boldsymbol\delta}_{\mathcal{S}^c}\|_1 - \|\widehat{\boldsymbol\delta}_{\mathcal{S}}\|_1.$$

Under the scaling $\lambda \ge 2\|\nabla\mathcal{L}_\tau(\boldsymbol\beta^*)\|_\infty$, it follows from the convexity of $\mathcal{L}_\tau$ and Cauchy-

Schwarz inequality that

$$\mathcal{L}_\tau(\widehat{\boldsymbol{\beta}}) - \mathcal{L}_\tau(\boldsymbol{\beta}^*) \ge \langle \nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*), \widehat{\boldsymbol{\delta}} \rangle \ge -\|\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_\infty \|\widehat{\boldsymbol{\delta}}\|_1 \\ \ge -\frac{\lambda}{2}\big(\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}^c}\|_1 + \|\widehat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1\big). \tag{C.44}$$

Together, (C.43) and (C.44) imply $0 \le \frac{\lambda}{2}(3\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 - \|\widehat{\boldsymbol{\delta}}_{\mathcal{S}^c}\|_1)$ and thus $\widehat{\boldsymbol{\beta}} \in \mathcal{C}$.

By necessary conditions of extrema in the convex optimization problem (B.4),

$$\langle \nabla \mathcal{L}_\tau(\widehat{\boldsymbol{\beta}}) + \lambda \widehat{\boldsymbol{z}}, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \le 0,$$

where $\widehat{\boldsymbol{z}} \in \partial \|\widehat{\boldsymbol{\beta}}\|_1$ satisfies $\langle \widehat{\boldsymbol{z}}, \boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}} \rangle \le \|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1$. Under the scaling $\lambda \ge 2\|\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_\infty$, it holds

$$\langle \nabla \mathcal{L}_\tau(\widehat{\boldsymbol{\beta}}) - \nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*), \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \le \lambda\big(\|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1\big) + \frac{\lambda}{2}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \\ \le \lambda\big(\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 - \|\widehat{\boldsymbol{\delta}}_{\mathcal{S}^c}\|_1\big) + \frac{\lambda}{2}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \le \frac{\lambda}{2}\big(3\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 - \|\widehat{\boldsymbol{\delta}}_{\mathcal{S}^c}\|_1\big).$$

Together with (C.40), this implies

$$\langle \nabla \mathcal{L}_\tau(\widetilde{\boldsymbol{\beta}}_\eta) - \nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*), \widetilde{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^* \rangle \le \frac{1}{2}\lambda\eta\big(3\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 - \|\widehat{\boldsymbol{\delta}}_{\mathcal{S}^c}\|_1\big). \tag{C.45}$$

Moreover, we introduce $\widetilde{\boldsymbol{\delta}}_\eta = \widetilde{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*$ and note that $\widetilde{\boldsymbol{\delta}}_\eta = \eta\widehat{\boldsymbol{\delta}}$. By (C.41), we also have $\widetilde{\boldsymbol{\beta}}_\eta \in \mathcal{C}$ under the assumed scaling.

Let Ω_r be the event on which (C.15) holds. Then $\mathbb{P}(\Omega_r^c) \le d^{-1}$ under the scaling (C.14) and it holds on $\Omega_r \cap \{\lambda \ge 2\|\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_\infty\}$ that

$$\langle \nabla \mathcal{L}_\tau(\widetilde{\boldsymbol{\beta}}_\eta) - \nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*), \widetilde{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^* \rangle \ge \frac{1}{4}\|\widetilde{\boldsymbol{\delta}}_\eta\|_{\boldsymbol{\Sigma},2}^2 \ge \frac{1}{4}\kappa_l^{1/2}\|\widetilde{\boldsymbol{\delta}}_\eta\|_2 \|\widetilde{\boldsymbol{\delta}}_\eta\|_{\boldsymbol{\Sigma},2}.$$

Substituting this lower bound into (C.45) yields

$$\frac{1}{4}\kappa_l^{1/2}\|\widetilde{\boldsymbol{\delta}}_\eta\|_2 \|\widetilde{\boldsymbol{\delta}}_\eta\|_{\boldsymbol{\Sigma},2} \le \frac{3}{2}\lambda\eta\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 \le \frac{3}{2}\lambda s^{1/2}\|\eta\widehat{\boldsymbol{\delta}}\|_2 = \frac{3}{2}\lambda s^{1/2}\|\widetilde{\boldsymbol{\delta}}_\eta\|_2.$$

Canceling $\|\widetilde{\boldsymbol{\delta}}_\eta\|_2$ on both sides delivers

$$\|\widetilde{\boldsymbol{\delta}}_\eta\|_{\boldsymbol{\Sigma},2} \le 6\kappa_l^{-1/2}s^{1/2}\lambda \quad \text{and} \quad \|\widetilde{\boldsymbol{\delta}}_\eta\|_1 \le 24\kappa_l^{-1}s\lambda \tag{C.46}$$

under the scaling $\lambda \ge 2\|\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_\infty$ and (C.14) .

It remains to calibrate the parameters τ, λ and r. First, applying Lemma C.6 with

$\tau = \tau_0(n/\log d)^{1/(1+\delta)}$, we see that

$$\|\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_\infty \le c_1 \max_{1\le j\le d} \sigma_{jj}^{1/2} \tau_0 \left(\frac{\log d}{n}\right)^{\delta/(1+\delta)}$$

with probability at least $1-2d^{-1}$, where $c_1 = (2\sqrt{2}+1)A_0 + 1$. We therefore choose $\lambda = c_2 \max_{1\le j\le d} \sigma_{jj}^{1/2} \tau_0 \{(\log d)/n\}^{\delta/(1+\delta)}$ for some constant $c_2 \ge 2c_1$, such that $\lambda \ge 2\|\nabla \mathcal{L}_\tau(\boldsymbol{\beta}^*)\|_\infty$ with probability at least $1-2d^{-1}$. Next, according to (C.14), the restricted strong convexity (C.15) holds with $r \asymp \kappa_l^{-1/2} A_0 \max_{1\le j\le d} \sigma_{jj}^{1/2} \tau \sqrt{(\log d)/n}$. Putting the above calculations together, we conclude that

$$\|\widetilde{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2} \le 6c_2 \kappa_l^{-1/2} \max_{1\le j\le d} \sigma_{jj}^{1/2} \tau_0\, s^{1/2} \left(\frac{\log d}{n}\right)^{\delta/(1+\delta)} < r \tag{C.47}$$

with probability at least $1-3d^{-1}$, assuming the scaling $n \gtrsim \kappa_l^{-1} A_0^2 A_1^4 \max_{1\le j\le d} \sigma_{jj}\, s \log d$. By the construction of $\widetilde{\boldsymbol{\beta}}_\eta$, with the same probability we must have $\eta = 1$ and therefore $\widehat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}_\eta$. The stated result (B.6) then follows from (C.46). □

C.11 Proof of Corollary B.1

Recall that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are i.i.d. random vectors from a sub-Gaussian vector $\boldsymbol{x} = (x_1, \ldots, x_d)^{\mathrm{T}}$ with $\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}) = \boldsymbol{\Sigma}$. Let $\boldsymbol{\Psi} = \mathbf{X}\boldsymbol{\Sigma}^{-1/2}$ be an $n \times d$ matrix whose rows are independent isotropic sub-Gaussian random vectors. Since $\kappa_l = \lambda_{\min}(\boldsymbol{\Sigma}) > 0$, Definition 1 in Rudelson and Zhou (2013) holds with $s_0 = s$, $k_0 = 3$, $A = \boldsymbol{\Sigma}^{1/2}$ and $K(s_0, k_0, A) = \kappa_l^{-1/2}$. Taking $\delta = 1$ in Theorem 16 of Rudelson and Zhou (2013) we obtain that, with probability at least $1-2d^{-1}$,

$$\frac{1}{\sqrt{n}} \frac{\|\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\|_2}{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma},2}} = \frac{1}{\sqrt{n}} \frac{\|\boldsymbol{\Psi}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\|_2}{\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\|_2} \le 2$$

for all $\boldsymbol{\beta} \in \mathcal{C}$ as long as $n \gtrsim \kappa_l^{-1} A_0^4 \max_{0\le j\le d} \sigma_{jj}\, s \log d$. This, together with (C.41) and (C.47), proves (B.7). □

References

Belloni, A. and Chernozhukov, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, **39** 82–130.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *The Annals of Statistics*, **37** 1705–1732.

Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press, Oxford.

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.

Fan, J., Liu, H., Sun, Q. and Zhang, T. (2018). I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *The Annals of Statistics*, **96** 1348–1360.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, Berlin.

Lepski, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *IEEE Transactions on Information Theory*, **36** 682–697.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, **16** 559–616.

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

Rudelson, M. and Zhou, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, **59** 3434–3447.

Spokoiny, V. (2013). Bernstein–von Mises theorem for growing parameter dimension. Preprint. Available at arXiv:1302.3430.

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

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, **55** 2183–2202.