Computational Statistics and Data Analysis 169 (2022) 107419

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

www.elsevier.com/locate/csda

Distributed adaptive Huber regression

Jiyu Luo [a], Qiang Sun [b], Wen-Xin Zhou [c,*]

[a] Division of Biostatistics, Herbert Wertheim School of Public Health and Human Longevity Science, University of California, San Diego, CA 92093, USA
[b] Department of Statistical Sciences, University of Toronto, Toronto, ON M5S 3G3, Canada
[c] Department of Mathematics, University of California, San Diego, La Jolla, CA 92093, USA

ARTICLE INFO

Article history:
Received 26 June 2021
Received in revised form 7 October 2021
Accepted 30 December 2021
Available online 6 January 2022

Keywords:
Adaptive Huber regression
Communication efficiency
Distributed inference
Heavy-tailed distribution
Nonasymptotic analysis

ABSTRACT

Distributed data naturally arise in scenarios involving multiple sources of observations, each stored at a different location. Directly pooling all the data together is often prohibited due to limited bandwidth and storage, or due to privacy protocols. A new robust distributed algorithm is introduced for fitting linear regressions when data are subject to heavy-tailed and/or asymmetric errors with finite second moments. The algorithm only communicates gradient information at each iteration, and therefore is communication-efficient. To achieve the bias-robustness tradeoff, the key is a novel double-robustification approach that applies on both the local and global objective functions. Statistically, the resulting estimator achieves the centralized nonasymptotic error bound as if all the data were pooled together and came from a distribution with sub-Gaussian tails. Under a finite $(2+\delta)$-th moment condition, a Berry-Esseen bound for the distributed estimator is established, based on which robust confidence intervals are constructed. In high dimensions, the proposed doubly-robustified loss function is complemented with ℓ_1-penalization for fitting sparse linear models with distributed data. Numerical studies further confirm that compared with extant distributed methods, the proposed methods achieve near-optimal accuracy with low variability and better coverage with tighter confidence width.

1. Introduction

In many applications, there are a massive number of individual agents/organizations collecting data independently. Multiple-site research has brought the possibility of studying rare outcome that require larger sample sizes, accelerating more generalizable findings, and bringing together investigators with different expertise from various backgrounds (Sidransky et al., 2009). Due to limited resources, such as bandwidth and storage, or privacy concerns, researchers across different sites are only allowed to share summary statistics without allowing collaborating parties to access raw data (Wu et al., 2012). Moreover, the collected data may often be contaminated by high level of noise, and thus of low quality. For example, in the context of gene expression data analysis, it has been observed that some gene expression levels have kurtosis values much larger than 3, despite of the normalization methods used (Wang et al., 2015). It is therefore important to develop robust and distributed learning algorithms with controlled communication cost and desirable statistical performance, measured by both efficiency and robustness.

* Corresponding author.
E-mail address: wez243@ucsd.edu (W.-X. Zhou).

https://doi.org/10.1016/j.csda.2021.107419
0167-9473/

Distributed learning algorithms have received considerable attention for multi-source studies in the past decade. Due to privacy concerns, data collected at each source, such as node, sensor or organization, must remain local. The goal is to develop efficient statistical learning methods that allow shared analyses or summary statistics without sharing individual level data. The classical divide-and-conquer principle is based on aggregating local estimators, that is, estimators computed separately on local machines, to form a final estimator; see, for example, Chen and Xie (2012), Li et al. (2013), Zhang et al. (2015), Zhao et al. (2016), Rosenblatt and Nadler (2016), Lee et al. (2017), Battey et al. (2018) and Volgushev et al. (2019), among many others. We refer to Huo and Cao (2018) for a more complete literature review. The divide-and-conquer approach, also known as one-step averaging, only takes one communication round, and therefore is convenient and has minimal communication cost. However, in order for the averaging estimator to achieve the same convergence rate as the centralized estimator, each local machine must have access to at least $\sqrt{N}$ samples, where N is the total sample size. This limits the number of machines allowed in the communication network.

To overcome this barrier of one-step averaging, multi-round procedures have been proposed for distributed data analysis with a large number of local agents (Shamir et al., 2014; Wang et al., 2017; Jordan et al., 2019; Wang et al., 2019). For linear and generalized linear models, Wang et al. (2017) and Jordan et al. (2019) proposed multi-round distributed (penalized) M-estimators that achieve optimal rates of convergence under very mild constraints on the number of machines. Chen et al. (2019) studied an iterative algorithm with proper smoothing for quantile regression under memory constraint, which may also apply under distributed computing platform. Alternatively, Dobriban and Sheng (2021) proposed an iterative weighted parameter averaging scheme for distributed linear regression when the dimension is comparable to the sample size.

For linear models under data parallelism, most of the existing distributed algorithms work with the least squares method, either by (weighted) averaging local least squares estimators or iteratively minimizing shifted (penalized) least squares loss functions. From a robustness viewpoint, distributed least squares based method inherits the sensitivity (non-robustness) of its centralized counterpart to the tails of the error distributions, hence increasing the variability of the estimator. In this paper, we propose a robust distributed algorithm for linear regression with heavy-tailed errors. Our proposal is inspired by Huber's M-estimation (Huber, 1973) but with double data-adaptive robustification parameters to achieve a balanced tradeoff between statistical optimality and communication efficiency. We refer the reader to Yohai and Maronna (1979), Portnoy (1985), Mammen (1989), He and Shao (1996) and He and Shao (2000) for the asymptotic properties of the classical Huber regression estimator in both fixed-p and increasing-p settings.

Our setup includes the heteroscedastic linear model with asymmetric errors, to which the least absolute deviation (LAD) regression does not naturally apply. Following the terminology in Catoni (2012), the type of "robustness" considered in this paper is quantified by nonasymptotic exponential deviation of the estimator versus polynomial tail of the error distribution. The ensuing procedure does sacrifice a fair amount of robustness to adversarial contamination of the data. The motivation of this work is different from and should not be confused with the classical notion of robust statistics (Huber and Ronchetti, 2009).

The distributed method is built upon the iterative, multi-round algorithm proposed by Wang et al. (2017) and Jordan et al. (2019), which only communicates gradient information at each round and therefore is communication-efficient. By a delicate choice of local and global robustification parameters, the proposed estimator satisfies exponential-type deviation bounds when the errors only have finite variance. Specifically, we show that the distributed estimator, obtained by a few rounds of communications, achieves the optimal centralized deviation bound as if the data were pooled together and subject to sub-Gaussian errors. The robustification parameters are also self-tuned, making the algorithm computationally convenient. We further derive a Berry-Esseen bound for the distributed estimator, based on which we construct robust confidence intervals. Finally, we propose a distributed penalized adaptive Huber regression estimator for high-dimensional sparse models, and establish its (near-)optimal theoretical guarantees.

Notation: For each integer $k \geq 1$, we use $\mathbb{R}^k$ to denote the k-dimensional Euclidean space. The inner product of two vectors $u = (u_1, \ldots, u_k)^{\mathrm{T}}, v = (v_1, \ldots, v_k)^{\mathrm{T}} \in \mathbb{R}^k$ is defined by $u^{\mathrm{T}}v = \langle u, v\rangle = \sum_{i=1}^k u_i v_i$. We use $\|\cdot\|_p$ $(1 \leq p \leq \infty)$ to denote the ℓ_p-norm in $\mathbb{R}^k$: $\|u\|_p = (\sum_{i=1}^k |u_i|^p)^{1/p}$ and $\|u\|_\infty = \max_{1\leq i\leq k} |u_i|$. For any $k \times k$ symmetric matrix $A \in \mathbb{R}^{k\times k}$, $\|A\|_2$ is the operator norm of A. For a positive semidefinite matrix $A \in \mathbb{R}^{k\times k}$, $\|\cdot\|_A$ denotes the norm induced by A, that is, $\|u\|_A = \|A^{1/2}u\|_2$, $u \in \mathbb{R}^k$. Moreover, we use $\mathbb{S}^{k-1} = \{u \in \mathbb{R}^k : \|u\|_2 = 1\}$ to denote the unit sphere in $\mathbb{R}^k$. For two sequences of non-negative numbers $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, $a_n \lesssim b_n$ indicates that there exists a constant $C > 0$ independent of n such that $a_n \leq Cb_n$; $a_n \gtrsim b_n$ is equivalent to $b_n \lesssim a_n$; $a_n \asymp b_n$ is equivalent to $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

2. Distributed adaptive Huber regression

2.1. Distributed Huber regression with adaptive robustification parameters

Suppose we observe independent data vectors $\{(y_i, x_i)\}_{i=1}^n$ following the linear model

$$y_i = x_i^{\mathrm{T}}\beta^* + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i|x_i) = 0, \ \ i = 1, \ldots, N, \tag{2.1}$$

where $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ with $x_{i1} \equiv 1$ is the covariate for the ith individual, $\beta^* \in \mathbb{R}^p$ is the underlying coefficient vector, and ε_i's are independent error variables. This setting allows conditional heteroscedastic models, where ε_i can depend on x_i.

For example, in a location-scale model we have $\varepsilon_i = \sigma(x_i)e_i$, where $\sigma(x_i)$ is a function of x_i, and e_i is independent of x_i. In the absence of normality assumption on the (conditional) error distribution, Huber's M-estimator (Huber, 1973) is one of the most widely used robust alternative to the least squares estimator. Given some $\tau > 0$, referred to as the robustification parameter, Huber's regression M-estimator for estimating β^* is defined as

$$\widehat{\beta} = \widehat{\beta}_\tau \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \widehat{\mathcal{L}}_\tau(\beta) := \frac{1}{N}\sum_{i=1}^{N} \ell_\tau(y_i - x_i^{\mathrm{T}}\beta),$$

where $\ell_\tau(u) = 0.5u^2 I(|u| \le \tau) + (\tau|u| - 0.5\tau^2)I(|u| > \tau)$ is the Huber loss. Traditionally, τ is often chosen to be 1.345σ with σ either determined by a robust scale estimate or simultaneously estimated by solving a system of equations, in order to achieve 95% asymptotic relative efficiency while gaining robustness when there are contaminated or heavy-tailed symmetric errors (Bickel, 1975; Western, 1995). In the presence of asymmetric heavy-tailed errors, Fan et al. (2017) and Sun et al. (2020) proposed (regularized) adaptive Huber regression estimators with τ scaling with the sample size and parametric dimension, and established exponential-type deviation bounds when ε_i's only have finite $(1+\delta)$-th moments for some $0 < \delta \le 1$.

In linear model (2.1), we allow heteroscedastic errors that are of the form $\varepsilon_i = \sigma(x_i)e_i$, where $\sigma(\cdot)$ is an unknown function on $\mathbb{R}^p$ and e_i is independent of x_i. When the error variables ε_i are heavy-tailed, asymmetric and have finite variance σ^2, Sun et al. (2020) showed that Huber's estimator $\widehat{\beta}_\tau$ with $\tau \asymp \sigma\sqrt{N/(p+\log N)}$, referred to as the adaptive Huber regression (AHR) estimator, exhibits sharp finite-sample deviation properties (Catoni, 2012), while the least squares estimator is far less concentrated around β^*. We say ε_i is heavy-tailed if it has infinite k-th absolute moment for some $k \ge 2$.

In the distributed setting, assume that the overall dataset $\{(y_i, x_i)\}_{i=1}^N$ is stored on m node machines, one central machine and $m-1$ local machines that connected to the central. For $j = 1, \ldots, m$, the jth machine stores a subsample of n_j observations, denoted by $\{(y_i, x_i)\}_{i\in\mathcal{I}_j}$, and $\mathcal{I}_j$'s are disjoint index sets that satisfy $\cup_{j=1}^m \mathcal{I}_j = \{1, \ldots, N\}$ and $N = \sum_{j=1}^m |\mathcal{I}_j| = \sum_{j=1}^m n_j$. Without loss of generality, we assume $n_1 = \cdots = n_j = n$ and $N = n \cdot m$ is divisible by m. We thus refer to n as the local sample size. When the entire dataset is available, the optimal τ scales with the total sample size N and dimension p for optimal bias and robustness tradeoff. With decentralized data, each local machine only has access to a subsample, so that the "locally optimal" τ depends on the local sample size. This, however, will lead to sub-optimal bounds for the aggregated estimator because τ is not large enough to offset the bias. To parallelize AHR in a distributed setting without compromising statistical optimality, we introduce two robustification parameters τ and κ, referred to as the global and local robustification parameters, and define the global and local Huber loss functions as $\widehat{\mathcal{L}}_\tau(\beta) = (1/N)\sum_{i=1}^N \ell_\tau(y_i - x_i^{\mathrm{T}}\beta)$ and $\widehat{\mathcal{L}}_{j,\kappa}(\beta) = (1/n)\sum_{i\in\mathcal{I}_j} \ell_\kappa(y_i - x_i^{\mathrm{T}}\beta)$ for $j = 1, \ldots, m$. Using this adaptive robustification procedure, we then extend the approximate Newton-type method (Shamir et al., 2014; Jordan et al., 2019) to robust regression with skewed heavy-tailed errors.

Starting with an initial estimator $\widetilde{\beta}^{(0)}$ of β^*, we define the shifted adaptive Huber loss

$$\begin{aligned}\widetilde{\mathcal{L}}(\beta) &= \widehat{\mathcal{L}}_{1,\kappa}(\beta) - \big\langle \nabla\widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(0)}) - \nabla\widehat{\mathcal{L}}_\tau(\widetilde{\beta}^{(0)}), \beta\big\rangle \\ &= \widehat{\mathcal{L}}_{1,\kappa}(\beta) - \Big\langle \nabla\widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(0)}) - \frac{1}{m}\sum_{j=1}^m \nabla\widehat{\mathcal{L}}_{j,\tau}(\widetilde{\beta}^{(0)}), \beta\Big\rangle, \quad \beta \in \mathbb{R}^p. \end{aligned} \tag{2.2}$$

Implicitly the shifted loss $\widetilde{\mathcal{L}}(\cdot)$ depends on both local and global robustification parameters κ and τ. It uses data available only on the first machine, used as the central machine, along with p-dimensional gradient vectors $\widehat{\mathcal{L}}_{j,\kappa}(\widetilde{\beta}^{(0)})$ $(j = 2, \ldots, m)$ that were sent from the remaining local machines. The ensuing one-step estimator is given by

$$\widetilde{\beta}^{(1)} = \widetilde{\beta}^{(1)}_{\kappa,\tau} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \widetilde{\mathcal{L}}(\beta). \tag{2.3}$$

This procedure requires one communication round of $O(pm)$ bits, and thus is communication-efficient. To investigate the statistical properties of $\widetilde{\beta}^{(1)}$, we impose the following moment condition on the data generating process.

(C1). The predictor $x \in \mathbb{R}^p$ is sub-Gaussian: there exists $\upsilon_1 \ge (2\log 2)^{-1/2}$ such that $\mathbb{P}(|z^{\mathrm{T}}u| \ge \upsilon_1 t) \le 2e^{-t^2/2}$ for every unit vector $u \in \mathbb{S}^{p-1}$ and $t \ge 0$, where $z = \Sigma^{-1/2}x$ and $\Sigma = \mathbb{E}(xx^{\mathrm{T}})$ is positive definite. Moreover, the regression error ε satisfies $\mathbb{E}(\varepsilon|x) = 0$ and $\mathbb{E}(\varepsilon^2|x) \le \sigma^2$ almost surely.

For prespecified parameters $r, r_* > 0$, define the events

$$\mathcal{E}_0(r) = \big\{\widetilde{\beta}^{(0)} \in \Theta(r)\big\} \quad \text{and} \quad \mathcal{E}_*(r_*) = \big\{\|\nabla\widehat{\mathcal{L}}_\tau(\beta^*)\|_\Omega \le r_*\big\}, \tag{2.4}$$

where $\Theta(r) := \{\beta \in \mathbb{R}^p : \|\beta - \beta^*\|_\Sigma \le r\}$ and $\Omega := \Sigma^{-1}$. Here r quantifies the statistical accuracy of the initial estimator $\widetilde{\beta}^{(0)}$, and r^* determines the estimation error of the centralized AHR estimator which essentially depends on the score $\nabla\widehat{\mathcal{L}}_\tau(\beta^*)$ with the global robustification parameter.

Theorem 2.1. *Assume Condition (C1) holds. For any $u > 0$, let the robustification parameters satisfy $\tau \geq \kappa \asymp \sigma\sqrt{n/(p+u)}$, and suppose the local sample size satisfies $n \gtrsim p + u$. Then, conditioned on the event $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*)$ with $8r_* \leq r_0 \leq \sigma$, the one-step estimator $\widetilde{\beta}^{(1)}$ defined in (2.3) satisfies*

$$\|\widetilde{\beta}^{(1)} - \beta^*\|_\Sigma \lesssim \sqrt{\frac{p+u}{n}} \cdot r_0 + r_* \quad \text{and} \tag{2.5}$$

$$\|\widetilde{\beta}^{(1)} - \beta^* + \Sigma^{-1}\nabla\widehat{\mathcal{L}}_\tau(\beta^*)\|_\Sigma \lesssim \sqrt{\frac{p+u}{n}} \cdot r_0, \tag{2.6}$$

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

In the above theorem, the bound (2.5) reflects the delicate dependence of the one-step error on the initial error r_0 as well as the centralized error rate r_*. If we take $\widetilde{\beta}^{(0)}$ to be a local estimator constructed on a single local machine that has access to only n observations, we may expect a sub-optimal convergence rate $r_0 \asymp \sigma\sqrt{p/n}$. Moreover, it can be shown that $\|\nabla\widehat{\mathcal{L}}_\tau(\beta^*)\|_\Omega \lesssim \sigma\sqrt{p/N} + \sigma^2/\tau + \tau p/N$ with high probability, up to logarithmic factors; see Lemma Appendix A.2 in the Appendix. Hence, the choice of r_* corresponds to the optimal rate of convergence when the entire dataset is available and $\tau \asymp \sigma\sqrt{N/p}$. Under the prescribed sample size scaling $n \gtrsim p$, the one-step estimator $\widetilde{\beta}^{(1)}$ refines the statistical accuracy of $\widetilde{\beta}^{(0)}$ by a factor of order $\sqrt{p/n}$, which is strictly less than 1. We thus expect the multi-step estimator, with sufficiently many communication rounds, will achieve the optimal convergence rate obtainable on the entire dataset.

The proposed multi-round procedure for adaptive Huber regression is iterative, starting at iteration 0 with an initial estimate $\widetilde{\beta}^{(0)} \in \mathbb{R}^p$. At iteration $t \geq 1$, it updates the estimate $\widetilde{\beta}^{(t)}$ by fitting a shifted adaptive Huber regression which leverages global first-order information, depending on τ, and local higher-order information, depending on κ. The procedure involves two steps.

1. Communicating gradient information. The central machine broadcasts $\widetilde{\beta}^{(t-1)}$ to every local machine. The jth machine, $2 \leq j \leq m$, computes the gradient $\nabla\widehat{\mathcal{L}}_{j,\tau}(\widetilde{\beta}^{(t-1)})$, and sends it back to the central machine. This step requires a communication of $2(m-1)p$ bits.

2. Fitting local shifted AHR. The central machine computes the update $\widetilde{\beta}^{(t)}$, defined as a solution to the optimization problem

$$\min_{\beta\in\mathbb{R}^p} \widetilde{\mathcal{L}}^{(t)}(\beta) := \widehat{\mathcal{L}}_{1,\kappa}(\beta) - \left\langle \nabla\widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(t-1)}) - \frac{1}{m}\sum_{j=1}^m \nabla\widehat{\mathcal{L}}_{j,\tau}(\widetilde{\beta}^{(t-1)}), \beta \right\rangle, \tag{2.7}$$

which can be solved by the method of iteratively reweighted least squares or quasi-Newton methods. Details are given in section 4.1. We summarize the procedure, with an early stopping criterion, in Algorithm 1.

Algorithm 1: Communication-Efficient Adaptive Huber Regression.

Input: data batches $\{(y_i, x_i)\}_{i\in\mathcal{I}_j}$, $j = 1, \ldots, m$, stored on m machines, robustification parameters $\tau \geq \kappa > 0$, initialization $\widetilde{\beta}^{(0)}$, number of iterations T, gradient tolerance $g_0 = 1$.

1: **for** $t = 1, 2 \ldots, T$ **do**

2: Broadcast $\widetilde{\beta}^{(t-1)}$ to all local machines;

3: The jth $(1 \leq j \leq m)$ machine computes $\nabla\widehat{\mathcal{L}}_{j,\tau}(\widetilde{\beta}^{(t-1)})$, and transmit it to the central machine;

4: Compute $\nabla\widehat{\mathcal{L}}_\tau(\widetilde{\beta}^{(t-1)}) = (1/m)\sum_{j=1}^m \nabla\widehat{\mathcal{L}}_{j,\tau}(\widetilde{\beta}^{(t-1)})$, $\nabla\widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(t-1)})$ and $g_t = \|\nabla\widehat{\mathcal{L}}_\tau(\widetilde{\beta}^{(t-1)})\|_\infty$ on the central machine;

5: If $g_t \geq g_{t-1}$ or $g_t \leq 10^{-5}$ break ; otherwise, on the central machine, solve the shifted adaptive Huber regression problem in (2.7) to update the estimate $\widetilde{\beta}^{(t)}$;

6: **end for**

Output: $\widetilde{\beta}^{(T)}$.

Theorem 2.2. *Assume the same conditions in Theorem 2.1, and let $8r_* \leq r_0 \leq \sigma$. Conditioned on event $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*)$, the distributed AHR estimator $\widetilde{\beta}^{(T)}$ with $T \gtrsim \lceil \log(r_0/r_*)/\log(n/(p+u)) \rceil$ satisfies the bounds*

$$\|\widetilde{\beta}^{(T)} - \beta^*\|_\Sigma \lesssim r_* \quad \text{and} \quad \|\widetilde{\beta}^{(T)} - \beta^* + \Sigma^{-1}\nabla\widehat{\mathcal{L}}_\tau(\beta^*)\|_\Sigma \lesssim \sqrt{\frac{p+u}{n}} \cdot r_*, \tag{2.8}$$

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

The above result shows that, with proper choices of τ and κ as well as the number of iterations, the statistical error of the multi-step distributed AHR estimator matches that of the centralized AHR estimator on the entire dataset. For the initialization, we may take $\widetilde{\beta}^{(0)}$ to be a local AHR estimator computed on the central machine. With the above preparations, we are ready to explicitly describe the estimation error and Bahadur linearization error of the proposed distributed AHR estimator. The result is nonasymptotic, and carefully tracks the impact of the parametric dimension p, local sample size n and the number of machines m.

Corollary 2.1. *Assume Condition (C1) holds, and suppose the local sample size satisfies* $n \gtrsim p + \log n + \log_2 m$, *where* $\log_2 m := \log(\log m)$ *and* $m = N/n$. *Choose the robustification parameters* $\tau \geq \kappa > 0$ *as* $\tau \asymp \sigma\sqrt{N/(p+\log n+\log_2 m)}$ *and* $\kappa \asymp \sigma\sqrt{n/(p+\log n+\log_2 m)}$. *Then, starting at iteration 0 with a local AHR estimate* $\widetilde{\beta}^{(0)}$, *the distributed estimator* $\widetilde{\beta} = \widetilde{\beta}^{(T)}$ *with* $T \asymp \lceil \frac{\log(m)}{\log(n/(p+\log n+\log_2 m))} \rceil$ *satisfies*

$$\|\widetilde{\beta} - \beta^*\|_\Sigma \lesssim \sigma\sqrt{\frac{p+\log n+\log_2 m}{N}} \quad \text{and} \tag{2.9}$$

$$\left\|\widetilde{\beta} - \beta^* - \Sigma^{-1}\frac{1}{N}\sum_{i=1}^{N}\psi_\tau(\varepsilon_i)x_i\right\|_\Sigma \lesssim \sigma\frac{p+\log n+\log_2 m}{(nN)^{1/2}}, \tag{2.10}$$

with probability at least $1 - Cn^{-1}$, *where* $\psi_\tau(u) := \ell'_\tau(u) = \mathrm{sign}(u)\min(|u|, \tau)$.

The above corollary indicates that the multi-step distributed AHR estimator $\widetilde{\beta}$ achieves the optimal statistical rate of convergence by a delicate combination of the local robustification parameter, the global robustification parameter, and number of communication rounds. The second bound, (2.10), explicitly describes the error term of the Bahadur linearization. This allows to establish the asymptotic distribution of $\widetilde{\beta}$ when both p, n tend to infinity. Moreover, to achieve statistical optimality and communication efficiently simultaneously, the above results impose minimal conditions on the number of machines m. In summary, when data are heavy-tailed and collected on each machine remain local, the proposed procedure delivers a statistically optimal estimate by communicating as many as $O(pm\log(m))$ bits.

2.2. Distributed confidence estimation

In this section, we consider uncertainty quantification of the multi-step estimator in a distributed setting, with a particular focus on statistical confidence estimation. We first establish a Berry-Esseen bound for linear functions of the distributed AHR estimator $\widetilde{\beta}$, which explicitly quantifies the normal approximation error.

Theorem 2.3. *In addition to the conditions in Theorem 2.1, assume* $\mathbb{E}(\varepsilon^2|x) = \sigma^2$ *and* $\mathbb{E}(|\varepsilon|^{2+\delta}|x) \leq v_{2+\delta}$ *almost surely for some* $0 < \delta \leq 1$. *Then, the distributed estimator* $\widetilde{\beta} = \widetilde{\beta}^{(T)}$ *satisfies*

$$\sup_{t\in\mathbb{R}, a\in\mathbb{R}^p}\left|\mathbb{P}\left[\frac{N^{1/2}a^{\mathrm{T}}(\widetilde{\beta}-\beta^*)}{\sqrt{\mathbb{E}\{\psi_\tau(\varepsilon)a^{\mathrm{T}}\Sigma^{-1}x\}^2}} \leq t\right] - \Phi(t)\right| \lesssim \frac{p+\log n+\log_2 m}{n^{1/2}} + \frac{v_{2+\delta}(p+\log n+\log_2 m)^{(1+\delta)/2}}{\sigma^{2+\delta}N^{\delta/2}}, \tag{2.11}$$

where $\Phi(\cdot)$ *is the standard normal distribution function. In particular, assume* $\mathbb{E}(|\varepsilon|^3|x) \leq v_3 < \infty$ *almost surely. Then, under the dimension constraint* $p + \log_2 m = o(n^{1/2})$,

$$\frac{N^{1/2}a^{\mathrm{T}}(\widetilde{\beta}-\beta^*)}{\sqrt{\mathbb{E}\{\psi_\tau(\varepsilon)a^{\mathrm{T}}\Sigma^{-1}x\}^2}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1) \quad \text{and} \quad \frac{N^{1/2}a^{\mathrm{T}}(\widetilde{\beta}-\beta^*)}{\sigma(a^{\mathrm{T}}\Sigma^{-1}a)^{1/2}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1), \tag{2.12}$$

uniformly over $a \in \mathbb{R}^p$ *as* $n \to \infty$.

Let $\widetilde{\beta} = (\widetilde{\beta}_1, \ldots, \widetilde{\beta}_p)^{\mathrm{T}}$ be the distributed estimator described in the previous subsection. Theorem 2.3 implies that, for each $1 \leq j \leq p$, $N^{1/2}(\widetilde{\beta}_j - \beta^*_j)$ is asymptotically normal with zero mean and variance $(\Sigma^{-1}\mathbb{E}\{\psi_\tau(\varepsilon)xx^{\mathrm{T}}\}^2\Sigma^{-1})_{jj}$. Let $\widehat{\Sigma} = (1/N)\sum_{i=1}^N x_i x_i^{\mathrm{T}}$ be the sample version of Σ, and $\widehat{\varepsilon}_i = y_i - x_i^{\mathrm{T}}\widetilde{\beta}$ be the fitted residuals. It can be shown that $(\widehat{\Sigma}^{-1}N^{-1}\sum_{i=1}^N \psi_\tau^2(\widehat{\varepsilon}_i)x_i x_i^{\mathrm{T}}\widehat{\Sigma}^{-1})_{jj}$ provides a consistent estimator of $(\Sigma^{-1}\mathbb{E}\{\psi_\tau(\varepsilon)xx^{\mathrm{T}}\}^2\Sigma^{-1})_{jj}$. In a distributed setting, the computation of this variance estimator requires communicating $O(p^2m)$ bits, thus incurring exorbitant communication costs when p is large.

To achieve a tradeoff between communication and statistical efficiency, we propose averaging pointwise variance estimators, defined by $\widehat{\sigma}_j^2 := (1/m)\sum_{k=1}^m \widehat{\sigma}_{jm}^2$ for $j = 1, \ldots, p$, where

$$\widehat{\sigma}_{jk}^2 = (\widehat{\Sigma}_k^{-1}\widehat{\Lambda}_k\widehat{\Sigma}_k^{-1})_{jj}, \quad \widehat{\Lambda}_k = \frac{1}{n}\sum_{i\in\mathcal{I}_k}\psi_\tau^2(\widehat{\varepsilon}_i)x_i x_i^{\mathrm{T}} \text{ and } \widehat{\Sigma}_k = \frac{1}{n}\sum_{i\in\mathcal{I}_k}x_i x_i^{\mathrm{T}}.$$

This approach takes one additional round of communication, and is robust against heteroscedastic errors that are of the form $\varepsilon_i = \sigma(x_i)e_i$. When ε_i is independent of x_i, the asymptotic variances reduce to $\mathbb{E}\{\psi_\tau^2(\varepsilon)\}(\Sigma^{-1})_{jj}$, and thus can be consistently estimated by $\widetilde{\sigma}_j^2 := (\widehat{\sigma}_\varepsilon^2/m)\sum_{k=1}^m(\widehat{\Sigma}_k^{-1})_{jj}$, where $\widehat{\sigma}_\varepsilon^2 = (N-p)^{-1}\sum_{i=1}^N\psi_\tau^2(\widehat{\varepsilon}_i)$. For $\alpha \in (0,1)$, the distributed

$100(1-\alpha)\%$ normal-based confidence intervals for β_j^*, $j=1,\ldots,p$, are given by $[\widetilde{\beta}_j - z_{\alpha/2}\widehat{\sigma}_j N^{-1/2}, \widetilde{\beta}_j + z_{\alpha/2}\widehat{\sigma}_j N^{-1/2}]$ or $[\widetilde{\beta}_j - z_{\alpha/2}\widetilde{\sigma}_j N^{-1/2}, \widetilde{\beta}_j + z_{\alpha/2}\widetilde{\sigma}_j N^{-1/2}]$, where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$.

The consistency of $\widehat{\sigma}_\varepsilon^2$ is established in the following proposition.

Proposition 2.1. *In addition to Condition (C1), assume $\mathbb{E}(\varepsilon^2|x) = \sigma^2$ and $\mathbb{E}(|\varepsilon|^{2+\delta}|x) \le v_{2+\delta}$ almost surely for some $0 < \delta \le 1$. Then, conditioned on the event $\{\widetilde{\beta} \in \Theta(r)\}$ with $0 < r \lesssim \sigma$, the variance estimator $\widehat{\sigma}_\varepsilon^2$ satisfies the bound*

$$|\widehat{\sigma}_\varepsilon^2 - \sigma^2| \lesssim v_{2+\delta}^{1/2}\tau^{1-\delta/2}\sqrt{\frac{p\log(N)+t}{N}} + \tau^2\frac{p\log(N)+t}{N} + \sigma r$$

with probability at least $1-2e^{-t}$ as long as $N \gtrsim p+t$. In particular, assume $\mathbb{E}(|\varepsilon|^3|x) \le v_3$ almost surely, and choose the robustification parameter $\tau \asymp \sigma\{N/(p+\log n)\}^{1/3}$. Then, conditioned on $\{\widetilde{\beta} \in \Theta(r)\}$ we have $|\widehat{\sigma}_\varepsilon^2 - \sigma^2| \lesssim v_3^{1/2}\sigma^{1/2}(p/N)^{1/3}\log(N) + \sigma r$ with probability at least $1-2N^{-1}$.

3. Distributed regularized adaptive Huber regression

In this section, we consider high-dimensional linear models under sparsity. Specifically, we allow the parametric dimension p to be much larger than the local sample size n, and assume β^* is s-sparse, where $s = |\mathcal{S}|$ and $\mathcal{S} = \mathrm{supp}(\beta^*) = \{1 \le j \le p : \beta_j^* \ne 0\}$ denotes the true active set.

Given independent observations $\{(y_i, x_i)\}_{i=1}^N$ from the linear model (2.1), the centralized/global ℓ_1-penalized Huber regression estimator (ℓ_1-Huber) is defined as

$$\widehat{\beta} = \widehat{\beta}_\tau(\lambda) \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p}\big\{\widehat{\mathcal{L}}_\tau(\beta) + \lambda\|\beta\|_1\big\}, \tag{3.1}$$

where $\lambda > 0$ is a regularization parameter. Statistical properties of ℓ_1-penalized Huber regression have been thoroughly studied by Lambert-Lacroix and Zwald (2011), Fan et al. (2017), Loh (2017) and Chinot et al. (2020) from different perspectives. To deal with asymmetric heavy-tailed errors, Fan et al. (2017) established high probability bounds for the ℓ_1-Huber estimator with $\tau \asymp \sigma\sqrt{N/\log(p)}$ in the high-dimensional regime $p \gg n \gtrsim s\log(p)$.

Remark 3.1. In practice, it is natural to leave the intercept or a given subset of the parameters unpenalized in the penalized M-estimation framework (3.1). Denote by $\mathcal{R} \subseteq \{1,\ldots,p\}$ the index set of unpenalized parameters, which is typically user-specified and contains at least index 1. A more flexible ℓ_1-Huber estimator can be obtained by solving $\min_{\beta\in\mathbb{R}^p}\{\widehat{\mathcal{L}}_\tau(\beta) + \lambda\|\beta_{\mathcal{R}^c}\|_1\} = \min_{\beta\in\mathbb{R}^p}\{\widehat{\mathcal{L}}_\tau(\beta) + \lambda\sum_{j\in\mathcal{R}^c}|\beta_j|\}$. Similar theoretical analysis can be carried out with slight modifications, and thus will be omitted.

In a distributed setting, we integrate the ideas of Wang et al. (2017) and Jordan et al. (2019) with adaptive robustification to parallelize regularized Huber regression with controlled communication cost and optimal statistical guarantees. As before, let τ and κ be the global and local robustification parameters. Recall that $\widehat{\mathcal{L}}_{j,\kappa}(\cdot)$, $j=1,\ldots,m$, denote local Huber loss functions. Commenced with a regularized estimator $\widetilde{\beta}^{(0)}$, let $\widetilde{\mathcal{L}}(\beta) = \widehat{\mathcal{L}}_{1,\kappa}(\beta) - \langle\nabla\widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(0)}) - \nabla\widehat{\mathcal{L}}_\tau(\widetilde{\beta}^{(0)}), \beta\rangle$ be the shifted adaptive Huber loss as in (2.2). With slight abuse of notation, we define the one-step ℓ_1-penalized Huber regression estimator as

$$\widetilde{\beta}^{(1)} = \widetilde{\beta}^{(1)}_{\kappa,\tau}(\lambda) \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p}\big\{\widetilde{\mathcal{L}}(\beta) + \lambda\|\beta\|_1\big\}. \tag{3.2}$$

To assess the statistical properties of the one-step estimator $\widetilde{\beta}^{(1)}$, we work under the following moment condition on the random covariate vector in high dimensions.

(C2). The covariate vector $x = (x_1,\ldots,x_p)^{\mathrm{T}} \in \mathbb{R}^p$ with $x_1 \equiv 1$ has bounded components and uniformly bounded kurtosis. That is, $\max_{1\le j\le p}|x_j| \le B$ for some $B \ge 1$ and $\mu_4 = \sup_{u\in\mathbb{S}^{p-1}}\mathbb{E}(z^{\mathrm{T}}u)^4 < \infty$, where $z = \Sigma^{-1/2}x$ and $\Sigma = (\sigma_{jk})_{1\le j,k\le p} = \mathbb{E}(xx^{\mathrm{T}})$. Write $\sigma_u = \max_{1\le j\le p}\sigma_{jj}^{1/2}$ and $\lambda_l = \lambda_{\min}(\Sigma) > 0$. For simplicity, we assume $\lambda_l = 1$. Moreover, the error variables ε_i satisfy $\mathbb{E}(\varepsilon_i|x_i) = 0$ and $\mathbb{E}(\varepsilon_i^2|x_i) \le \sigma^2$ almost surely.

As before, we first examine the performance of $\widetilde{\beta}^{(1)}$ conditioned on certain "good" events in regard of the initialization and the centralized ℓ_1-Huber estimator. For $r_0, \lambda_* > 0$, define

$$\mathcal{E}_0(r_0) = \big\{\widetilde{\beta}^{(0)} \in \Theta(r_0)\cap\Lambda\big\} \quad\text{and}\quad \mathcal{E}_*(\lambda_*) = \big\{\|\nabla\widehat{\mathcal{L}}_\tau(\beta^*) - \nabla\mathcal{L}_\tau(\beta^*)\|_\infty \le \lambda_*\big\}, \tag{3.3}$$

where $\Lambda := \{\beta\in\mathbb{R}^p : \|\beta-\beta^*\|_1 \le 4s^{1/2}\|\beta-\beta^*\|_\Sigma\}$ is an ℓ_1-cone.

Theorem 3.1. *Assume Condition (C2) holds. Given $\delta \in (0,1)$ and $0 < r_0, \lambda_* \lesssim \sigma$, let (τ, κ, λ) satisfy $\tau \geq \kappa \asymp \sigma\sqrt{n/\log(p/\delta)}$ and $\lambda = 2.5(\lambda_* + \rho)$ with*

$$\rho \asymp \max\left\{ r_0\sqrt{\frac{s\log(p/\delta)}{n}}, s^{-1/2}\sigma^2\tau^{-1} \right\}.$$

Moreover, suppose the local sample size satisfies $n \gtrsim s\log(p/\delta)$. Then, conditioned on the event $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(\lambda_*)$, the one-step regularized estimator $\widetilde{\beta}^{(1)}$ defined in (3.2) satisfies $\widetilde{\beta}^{(1)} \in \Lambda$ and

$$\|\widetilde{\beta}^{(1)} - \beta^*\|_\Sigma \lesssim s\sqrt{\frac{\log(p/\delta)}{n}} \cdot r_0 + \sigma^2\tau^{-1} + s^{1/2}\lambda_*, \tag{3.4}$$

with probability at least $1 - \delta$.

Theorem 3.1 indicates that the one-step procedure is able to reduce the statistical error of the initial estimator by a factor of $s\sqrt{\log(p)/n}$ when the local sample size satisfies $n \gtrsim s^2\log(p)$; see the first term on the right-hand of (3.4). The second term, $\sigma^2\tau^{-1} + s^{1/2}\lambda_*$, corresponds to the global error rate achievable on the entire dataset. In view of Theorem B.2 (with $\delta = 1$) in Sun et al. (2020), if we take $\lambda_* \asymp \sigma\sqrt{\log(p)/N}$ and $\tau \asymp \sigma\sqrt{N/\log(p)}$, the centralized ℓ_1-Huber estimator given in (3.1) satisfies $\|\widehat{\beta} - \beta^*\|_\Sigma \lesssim \sigma^2\tau^{-1} + s^{1/2}\lambda_* \asymp \sigma\sqrt{s\log(p)/N}$ with probability at least $1 - Cp^{-1}$.

Now we extend the iterative procedure in Section 2 to high-dimensional settings, starting at iteration 0 with an initial estimate $\widetilde{\beta}^{(0)} \in \mathbb{R}^p$. At iteration $t = 1, 2, \ldots$, it proceeds as follows:

Communicating gradient information. The jth $(2 \leq j \leq m)$ machine receives $\widetilde{\beta}^{(t-1)}$ from the central machine, computes the local gradient $\nabla\widehat{\mathcal{L}}_{j,\tau}(\widetilde{\beta}^{(t-1)})$, and sends it back to the central.

Fitting local regularized AHR: On the central machine, solve $\min_{\beta\in\mathbb{R}^p}\{\widetilde{\mathcal{L}}^{(t)}(\beta) + \lambda_t\|\beta\|_1\}$ to obtain $\widetilde{\beta}^{(t)}$, where $\widetilde{\mathcal{L}}^{(t)}(\beta) = \widehat{\mathcal{L}}_{1,\kappa}(\beta) - \langle\nabla\widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(t-1)}) - (1/m)\sum_{j=1}^m \nabla\widehat{\mathcal{L}}_{j,\tau}(\widetilde{\beta}^{(t-1)}), \beta\rangle$ and $\lambda_t > 0$ is a regularization parameter.

Computationally, we use a variant of the majorize-minimize algorithm (Lange et al., 2000), a proximal gradient descent type method, to solve the regularized optimization problem at each iteration. Details are provided in section 4.2. Theorem 3.2 below describes the statistical properties of the solution path $\{\widetilde{\beta}^{(t)}\}_{t\geq 1}$ conditioned on a prespecified level of accuracy of the initial estimator.

Theorem 3.2. *Assume Condition (C2) holds. Given $\delta \in (0,1)$ and $0 < r_0, \lambda_* \lesssim \sigma$, let (τ, κ) satisfy $\tau \geq \kappa \asymp \sigma\sqrt{n/\log(p/\delta)}$. For $t = 1, 2, \ldots$, set $\lambda_t = 2.5(\lambda_* + \rho_t) > 0$ with $\rho_t \asymp s^{-1/2}\max\{\alpha^t r_0, \sigma^2\tau^{-1}\}$ and $\alpha \asymp s\sqrt{\log(p/\delta)/n}$. Suppose the local sample size satisfies $n \gtrsim s^2\log(p/\delta)$, and let $r_* \asymp \sigma^2\tau^{-1} + s^{1/2}\lambda_*$. Then, conditioned on event $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(\lambda_*)$, the distributed regularized estimator $\widetilde{\beta}^{(T)}$ with $T \asymp \frac{\log(r_0/r_*)}{\log(1/\alpha)}$ satisfies $\widetilde{\beta}^{(T)} \in \Lambda$ and $\|\widetilde{\beta}^{(T)} - \beta^*\|_\Sigma \lesssim r_*$ with probability at least $1 - T\delta$.*

With sufficiently many samples on the central machine—$n \gtrsim s^2\log(p)$, Theorems 3.1 and 3.2 ensure that the initial estimation error, albeit being sub-optimal, can be repeatedly refined by a factor of order $s\sqrt{\log(p)/n}$ until it reaches the optimal rate. For simplicity, we take $\widetilde{\beta}^{(0)}$ to be a local ℓ_1-penalized AHR estimator, that is, $\widetilde{\beta}^{(0)} \in \operatorname{argmin}_{\beta\in\mathbb{R}^p}\{\widehat{\mathcal{L}}_{1,\kappa}(\beta) + \lambda_0\|\beta\|_1\}$.

Corollary 3.1. *Assume Condition (C2) holds, and the sample size per machine satisfies $n \gtrsim s^2\log p$. Choose the robustification and regularization parameters as $\tau \asymp \sigma\sqrt{N/\log(p)}$, $\kappa \asymp \sigma\sqrt{n/\log(p)}$ and*

$$\lambda_t \asymp \sigma\sqrt{\frac{\log p}{N}} + \sigma\left(\frac{s^2\log p}{n}\right)^{t/2}\sqrt{\frac{\log p}{n}}, \quad t = 0, 1, 2, \ldots.$$

Starting at iteration 0 with a local ℓ_1-penalized AHR estimator, the multi-step estimator $\widetilde{\beta}^{(T)}$ after $T \asymp \lceil\log(m)\rceil$ rounds of communication satisfies the bounds

$$\|\widetilde{\beta}^{(T)} - \beta^*\|_\Sigma \lesssim \sigma\sqrt{\frac{s\log p}{N}} \quad \text{and} \quad \|\widetilde{\beta}^{(T)} - \beta^*\|_1 \lesssim \sigma s\sqrt{\frac{\log p}{N}},$$

with probability at least $1 - C\log(m)/p$.

Corollary 3.1, along with the global error analysis in Fan et al. (2017) and Loh (2017), implies the optimality of distributed adaptive Huber regression in terms of the tradeoff between communication cost and statistical accuracy.

Remark 3.2. Under light-tailed error distributions (e.g., sub-Gaussian errors), Lee et al. (2017) and Battey et al. (2018) studied a one-shot approach based on averaging debiased Lasso estimators (Zhang and Zhang, 2014; van de Geer et al., 2014). Theoretically, averaged debiased Lasso achieves the optimal error rate when the local size satisfies $n \gtrsim ms^2\log(p)$;

and computationally, each local machine needs to estimate a $p \times p$ matrix for debiasing the Lasso. We may expect the same issues for the robust one-shot method that averages debiased ℓ_1-Huber estimators. The proposed distributed AHR method not only requires the minimum sample complexity but also is computationally efficient.

4. Optimization methods

4.1. Barzilai-Borwein gradient descent for distributed AHR

Let us first recall the multi-round distributed procedure for adaptive Huber regression. Starting with an initial estimator $\widetilde{\beta}^{(0)} \in \mathbb{R}^p$, and given robustification parameters τ and κ, for $t = 1, \ldots, T$, we update

$$\widetilde{\beta}^{(t)} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \widetilde{\mathcal{L}}^{(t)}(\beta) = \widehat{\mathcal{L}}_{1,\kappa}(\beta) - \left\langle \nabla \widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(t-1)}) - \nabla \widehat{\mathcal{L}}_{\tau}(\widetilde{\beta}^{(t-1)}), \beta \right\rangle. \tag{4.1}$$

Note that the empirical loss $\widetilde{\mathcal{L}}^{(t)}(\cdot)$ is convex and continuously differentiable. Moreover, since the Huber loss is locally strongly convex around zero, we will show that $\widetilde{\mathcal{L}}^{(t)}(\cdot)$ is locally strongly convex in a neighborhood of $\widetilde{\beta}^{(t)}$ with high probability. To take advantage of such a local strong convexity, we employ the gradient descent method with a Barzilai-Borwein update step (GD-BB) (Barzilai and Borwein, 1988) to solve the optimization problem in (4.1). The Barzilai-Borwein method is motivated by quasi-Newton methods, which avoid calculating the inverse Hessian at each iteration. The latter is computationally expensive when p is large. To be specific, let us consider the optimization $\min_{\beta \in \mathbb{R}^p} \widetilde{\mathcal{L}}^{(t)}(\beta)$ for a fixed $t \geq 1$. Starting with the initialization $\widetilde{\boldsymbol{\beta}}^{(t,0)} = \widetilde{\boldsymbol{\beta}}^{(t-1)}$, at (inner) iteration $k = 1, 2, \ldots$, compute the update $\widetilde{\boldsymbol{\beta}}^{(t,k+1)} = \widetilde{\boldsymbol{\beta}}^{(t,k)} - \min\{\eta_k, 10\} \nabla \widetilde{\mathcal{L}}^{(t)}(\widetilde{\boldsymbol{\beta}}^{(t,k)})$, where $\eta_1 = 1$ and for $k \geq 2$,

$$\eta_k = \frac{\langle \widetilde{\beta}^{(t,k)} - \widetilde{\beta}^{(t,k-1)}, \widetilde{\beta}^{(t,k)} - \widetilde{\beta}^{(t,k-1)} \rangle}{\langle \widetilde{\beta}^{(t,k)} - \widetilde{\beta}^{(t,k-1)}, \nabla \widetilde{\mathcal{L}}^{(t)}(\widetilde{\beta}^{(t,k)}) - \nabla \widetilde{\mathcal{L}}^{(t)}(\widetilde{\beta}^{(t,k-1)}) \rangle} \tag{4.2}$$

or

$$\eta_k = \frac{\langle \widetilde{\beta}^{(t,k)} - \widetilde{\beta}^{(t,k-1)}, \nabla \widetilde{\mathcal{L}}^{(t)}(\widetilde{\beta}^{(t,k)}) - \nabla \widetilde{\mathcal{L}}^{(t)}(\widetilde{\beta}^{(t,k-1)}) \rangle}{\| \nabla \widetilde{\mathcal{L}}^{(t)}(\widetilde{\beta}^{(t,k)}) - \nabla \widetilde{\mathcal{L}}^{(t)}(\widetilde{\beta}^{(t,k-1)}) \|_2^2}.$$

In practice, the step size computed in GD-BB may sometimes vibrate to some extent, and this may cause instability of the algorithm. Therefore, we set a upper bound for the step sizes by taking $\min\{\eta_k, 10\}$. This procedure is summarized in Algorithm 2.

4.2. Majorize-minimize algorithm for distributed penalized AHR

In the high-dimensional setting, we need to solve ℓ_1-penalized shifted Huber loss minimization problems.

Algorithm 2: Gradient Descent with Barzilai-Borwein stepsize for solving (4.1).

Input: Local data vectors $\{(y_i, x_i)\}_{i \in \mathcal{I}_1}$, initial estimator $\widehat{\beta}^0 = \widetilde{\beta}^{(t-1)}$, gradient $\nabla \widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(t-1)})$ and $\nabla \widehat{\mathcal{L}}_{j,\tau}(\widetilde{\beta}^{(t-1)})$ for $j = 1, \ldots, m$, and gradient tolerance level $\delta = 10^{-4}$.

1: Compute $\widehat{\beta}^1 \leftarrow \widehat{\beta}^0 - \nabla \widetilde{\mathcal{L}}^{(t)}(\widehat{\beta}^0)$
2: **for** $k = 1, 2 \ldots$ **do**
3: Compute η_k as defined in (4.2).
4: Update $\widehat{\beta}^{k+1} \leftarrow \widehat{\beta}^k - \min\{\eta_k, 10\} \nabla \widetilde{\mathcal{L}}^{(t)}(\widehat{\beta}^k)$;
5: **end for** when $\| \nabla \widetilde{\mathcal{L}}^{(t)}(\widehat{\beta}^k) \|_\infty \leq \delta$

With slight abuse of notation, given an initial regularized estimator $\widetilde{\beta}^{(0)}$, at each iteration $t = 1, 2, \ldots, T$, define the update as

$$\widetilde{\beta}^{(t)} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \left\{ \widetilde{\mathcal{L}}^{(t)}(\beta) + \lambda \|\beta_-\|_1 = \widehat{\mathcal{L}}_{1,\kappa}(\beta) - \left\langle \nabla \widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(t-1)}) - \nabla \widehat{\mathcal{L}}_{\tau}(\widetilde{\beta}^{(t-1)}), \beta \right\rangle + \lambda \|\beta_-\|_1 \right\}. \tag{4.3}$$

Here we use $\beta_- \in \mathbb{R}^{p-1}$ to denote the subvector of β with its first component removed. To solve the optimization problem in (4.3), we employ the locally adaptive majorize-minimize (LAMM) principle Fan et al. (2018), which extends the classical MM algorithm (Hunter and Lange, 2000) to accommodate ℓ_1 penalty. This procedure minimizes a surrogate ℓ_1-penalized isotropic quadratic function at each iteration, thus permitting an analytical solution.

Let $\widetilde{\mathcal{L}}(\cdot)$ be the loss function of interest. For $k = 1, 2, \ldots$, define

$$g_k(\beta; \beta^{k-1}, \phi_k) = \widetilde{\mathcal{L}}(\beta^{k-1}) + \left\langle \nabla \widetilde{\mathcal{L}}(\beta^{k-1}), \beta - \beta^{k-1} \right\rangle + \frac{\phi_k}{2} \|\beta - \beta^{k-1}\|_2^2.$$

We say $g_k(\beta;\beta^{k-1},\phi_k)$ majorizes $\widetilde{\mathcal{L}}(\beta)$ at β^{k-1} if

$$g_k(\beta;\beta^{k-1},\phi_k)\geq\widetilde{\mathcal{L}}(\beta)\ \ \forall\beta\in\mathbb{R}^p\quad\text{and}\quad g_k(\beta^{k-1};\beta^{k-1},\phi_k)=\widetilde{\mathcal{L}}(\beta^{k-1}).\tag{4.4}$$

By choosing ϕ_k large enough, $g_k(\cdot;\beta^{k-1},\phi_k)$ is guaranteed to satisfy (4.4). To find the smallest such ϕ_k, we start with $\phi_0=0.0001$, and repeatedly inflate it by a constant factor, say 1.1, until (4.4) is satisfied. Finally, we update β^k by minimizing

$$g_k(\beta;\beta^{k-1},\phi_k)+\lambda\|\beta_-\|_1.\tag{4.5}$$

Due to the isotropic quadratic term in $g_k(\beta;\beta^{k-1},\phi_k)$, β^k can be obtained by a simple analytic formula:

$$\begin{cases}\beta_1^k=\beta_1^{k-1}-\phi_k^{-1}(\nabla\widetilde{\mathcal{L}}(\beta^{k-1}))_1\\ \beta_j^k=S(\beta_j^{k-1}-\phi_k^{-1}(\nabla\widetilde{\mathcal{L}}(\beta^{k-1}))_j,\phi_k^{-1}\lambda),\quad j=2,\ldots,p,\end{cases}$$

where $S(u,\lambda)=\mathrm{sign}(u)\max(|u|-\lambda,0)$ denotes the soft-thresholding operator. This algorithm also guarantees a descent in the overall loss function at every iteration, which is a direct consequence of (4.4) and (4.5):

$$\begin{aligned}\widetilde{\mathcal{L}}(\beta^k)+\lambda\|\beta_-^k\|_1&\leq g_k(\beta^k;\beta^{k-1},\phi_k)+\lambda\|\beta_-^k\|_1\\&\leq g_k(\beta^{k-1};\beta^{k-1},\phi_k)+\lambda\|\beta_-^{k-1}\|_1=\widetilde{\mathcal{L}}(\beta^{k-1})+\lambda\|\beta_-^{k-1}\|_1.\end{aligned}$$

Algorithm 3 summarizes the LAMM algorithm described above.

Algorithm 3: Local adaptive majorize-minimize (LAMM) algorithm for solving (3.2).

Input: Local data vectors $\{(y_i,x_i)\}_{i\in I_1}$, initial estimator $\widehat{\beta}^0=\widetilde{\beta}^{(t-1)}$ gradient vectors $\nabla\widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(t-1)})$ and $\nabla\widehat{\mathcal{L}}_{\tau}(\widetilde{\beta}^{(t-1)})$, regularization parameter λ, initial isotropic parameter ϕ_0 and convergence tolerance δ

1: **for** $k=1,2\ldots$ **do**
2: Set $\phi_k\leftarrow\max\{\phi_0,\phi_{k-1}/1.1\}$
3: **repeat**
4: Update $\widehat{\beta}_1^k\leftarrow\widehat{\beta}_1^{k-1}-\phi_k^{-1}\nabla_{\beta_1}\widetilde{\mathcal{L}}(\widehat{\beta}^{k-1})$
5: Update $\widehat{\beta}_j^k\leftarrow S(\widehat{\beta}_j^{k-1}-\phi_k^{-1}\nabla_{\beta_j}\widetilde{\mathcal{L}}(\widehat{\beta}^{k-1}),\phi_k^{-1}\lambda)$ for $j=2,\ldots,p$
6: If $g_k(\widehat{\beta}^k;\widehat{\beta}^{k-1},\phi_k)<\widetilde{\mathcal{L}}(\widehat{\beta}^k)$, set $\phi_k\leftarrow1.1\phi_k$
7: **until** $g_k(\widehat{\beta}^k;\widehat{\beta}^{k-1},\phi_k)\geq\widetilde{\mathcal{L}}(\widehat{\beta}^k)$
8: **end for** when $\|\widehat{\beta}^k-\widehat{\beta}^{k-1}\|_2\leq\delta$

5. Numerical studies

In this section, we compare the numerical performance of the proposed method with several state-of-the-art distributed regression methods in both low and high dimensions.

5.1. Distributed robust regression and inference

In the low-dimensional setting where $n\gg p$, we consider five distributed regression methods: (i) the global adaptive Huber regression (AHR) estimator (Sun et al., 2020) that uses all the available $N=mn$ observations; (ii) divide-and-conquer AHR (DC-AHR) estimator based on averaging m local AHR estimators; (iii) DC-OLS estimator that averages m local OLS estimators; (iv) distributed OLS estimator (Shamir et al., 2014); and (v) the proposed distributed AHR estimator with early stopping.

To implement methods (i) and (ii), we use the self-tuning principle proposed by Wang et al. (2021) which automatically selects the robustification parameter τ. The distributed procedures (iv) and (v) are iterative, and require a reasonably well initial estimator, say $\widetilde{\beta}^{(0)}$. In our simulations, we take $\widetilde{\beta}^{(0)}$ to be either the DC-AHR or the DC-OLS estimator, which only requires one communication round. When the error distribution is heavy-tailed and symmetric, DC-AHR often has better finite-sample performance than DC-OLS. However, it produces biased estimate when the error is asymmetric. In contrast, although the DC-OLS exhibits larger variability due to heavy-tailedness, it has smaller bias on average. Therefore, we use DC-OLS estimator as the initialization for both methods (iv) and (v). Recall that the distributed AHR estimator involves two robustification parameters κ and τ. The local parameter κ can be automatically obtained by the self-tuning procedure (Wang et al., 2021). Guided by theoretical orders of (κ,τ) stated in Theorem 2.1, we choose the global parameter τ to be $cm^{1/2}\kappa$, where $c\geq1$ is a numerical constant that can be tuned by the validation set approach. We suggest to choose c from $\{1,2,3,4,5\}$, which suffices to achieve promising performance in a wide range of simulation settings.

We generate data vectors $\{(y_i,x_i)\}_{i=1}^N$ from a heteroscedastic model $y_i=x_i^{\mathrm{T}}\beta^*+c^{-1}(x_i^{\mathrm{T}}\beta^*)^2\varepsilon_i$, where $\beta^*=(1.5,\ldots,1.5)^{\mathrm{T}}\in\mathbb{R}^p$, $x_i=(1,x_{i2},\ldots,x_{ip})^{\mathrm{T}}$ with $x_{ij}\sim\mathcal{N}(0,1)$ for $j=2,\ldots,p$ and $c=\sqrt{3}\|\beta^*\|_2^2$ that makes $\mathbb{E}\{c^{-1}(x_i^{\mathrm{T}}\beta^*)^2\}^2=1$. The regression errors ε_i are generated from one of the following four distributions (centered if the mean is nonzero): (a) $\mathcal{N}(0,1)$

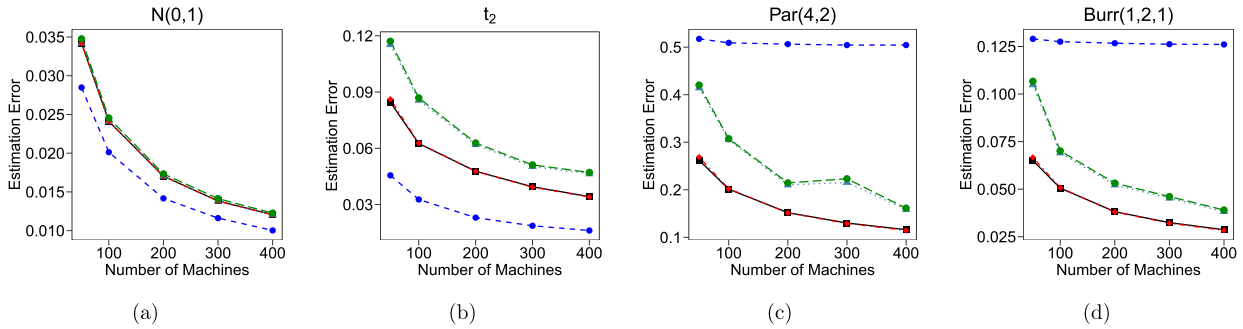

Fig. 1. Plots of estimation error (under ℓ_2-norm) versus number of machines when $(n, p) = (400, 20)$, averaged over 500 replications. Five estimators are presented: global AHR estimator (—■—); DC-AHR estimator (-●- -●-); DC-OLS estimator (—●—●—); distributed OLS estimator (··▲··▲··); and distributed AHR estimator (-·◆·◆·-).

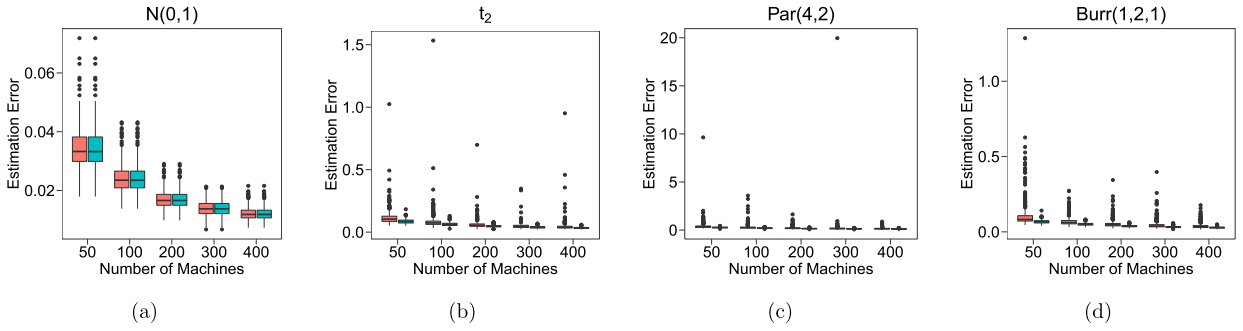

Fig. 2. Boxplots of estimation error (under ℓ_2-norm) versus the number of machines when $(n, p) = (400, 20)$ for distributed OLS estimator (■) and distributed AHR estimator (■), averaged over 500 replications.

(standard normal), (b) t_2 (t-distribution with 2 degrees of freedom), (c) Par$(4, 2)$–Pareto distribution with scale parameter 4 and shape parameter 2, and (d) Burr$(1, 2, 1)$–Burr distribution or the Singh-Maddala distribution (Singh and Maddala, 1976), which is commonly used to model household income. First, we fix $(n, p) = (400, 20)$ and let the number of machines m increase from 10 to 500. Fig. 1 plots the ℓ_2-error $\|\widehat{\beta} - \beta^*\|_2$ versus the number of machines, averaged over 500 replications, for all five methods. The global and distributed AHR estimators have almost identical performance, thus corroborating our theoretical results. The DC-AHR estimator only performs well under symmetric errors and suffers from non-negligible bias if the errors come from asymmetric distributions. This is largely expected because the robustification parameter for a local AHR estimator is tuned by a small subset of the data and results in a bias scaling with the local sample size. After averaging, this bias will not be offset when the number of machines increases. This points out a key drawback of the one-shot averaging approach when dealing with skewed data distributed across local machines. It is worth noticing that the distributed OLS and DC-OLS estimators perform almost identically in all the settings, which is as expected according to Jordan et al. (2019). They have decaying estimation errors as m grows, but at a slower rate compared to the global and the distributed AHR estimators for heavy-tailed data. The boxplots in Fig. 2 further reveal that the distributed OLS method often produces very poor estimates with high variability, while the distributed AHR method exhibits high degree of robustness.

Interestingly, under symmetric errors such as $\mathcal{N}(0, 1)$ and t_2, the DC-AHR estimator even outperforms the global AHR estimator, which may be due to the following reasons. Recall that the data is generated from a heteroscedastic model. The global AHR estimator chooses only one τ value using all the data, while for the DC approach, each local AHR estimator is based on a self-tuned κ using the local data. Due to symmetry, local AHR estimators gain robustness without sacrificing bias; moreover, averaging independent estimators reduces the variance. On the other hand, in the presence of asymmetric errors, each local AHR suffers from a bias depending only on the local sample size. Although averaging reduces variance, the bias remains and therefore the performance of DC-AHR barely improves as the number of machines increases.

Turning to uncertainty quantification, we construct approximate 95% confidence intervals for the slope coefficients based on distributed OLS and AHR methods. As before, we set $(n, p) = (400, 20)$ and let m increase from 10 to 500. Table 1 shows the average coverage probabilities and widths, with standard errors in parentheses, across all slope coefficients based on 500 Monte Carlo simulations. Across all the settings, the AHR-based confidence intervals are consistently accurate with tight width and reliable with high coverage. In the presence of heavy-tailed errors, the OLS-based confidence intervals

Table 1
Coverage probabilities and widths (with standard errors in parentheses) of the normal-based CIs (averaged over all slope coefficients) for the distributed OLS and distributed AHR methods, based on 500 Monte Carlo simulations.

		$N(0,1)$		t_2		Par(4,2)		Burr(1,2,1)	
		Coverage mean (sd)	Width mean (sd)	Coverage mean (sd)	Width mean (sd)	Coverage mean (sd)	Width mean (sd)	Coverage mean (sd)	Width mean (sd)
$t_{m=50}$	Dist-OLS	0.93(0.011)	0.029(0.001)	0.93(0.011)	0.097(0.056)	0.93(0.012)	0.35(0.420)	0.94(0.011)	0.088(0.068)
	Dist-AHR	0.95(0.007)	0.031(0.001)	0.95(0.009)	0.077(0.007)	0.95(0.008)	0.23(0.025)	0.95(0.009)	0.058(0.006)
$m = 100$	Dist-OLS	0.93(0.012)	0.020(0.000)	0.94(0.010)	0.072(0.056)	0.93(0.012)	0.25(0.220)	0.93(0.008)	0.058(0.021)
	Dist-AHR	0.95(0.010)	0.022(0.001)	0.96(0.008)	0.058(0.005)	0.95(0.009)	0.18(0.017)	0.95(0.009)	0.044(0.004)
$m = 200$	Dist-OLS	0.93(0.011)	0.014(0.000)	0.93(0.013)	0.052(0.031)	0.93(0.010)	0.18(0.095)	0.94(0.015)	0.044(0.021)
	Dist-AHR	0.96(0.007)	0.015(0.000)	0.95(0.011)	0.043(0.003)	0.95(0.009)	0.13(0.012)	0.96(0.012)	0.034(0.003)
$m = 300$	Dist-OLS	0.93(0.013)	0.012(0.000)	0.94(0.011)	0.043(0.022)	0.94(0.011)	0.18(0.820)	0.93(0.008)	0.038(0.020)
	Dist-AHR	0.95(0.010)	0.013(0.000)	0.96(0.010)	0.036(0.003)	0.95(0.012)	0.11(0.009)	0.96(0.009)	0.028(0.002)
$m = 400$	Dist-OLS	0.93(0.010)	0.010(0.000)	0.94(0.011)	0.040(0.046)	0.93(0.008)	0.13(0.071)	0.94(0.010)	0.032(0.014)
	Dist-AHR	0.95(0.009)	0.011(0.000)	0.96(0.009)	0.031(0.002)	0.95(0.012)	0.10(0.008)	0.96(0.009)	0.025(0.002)

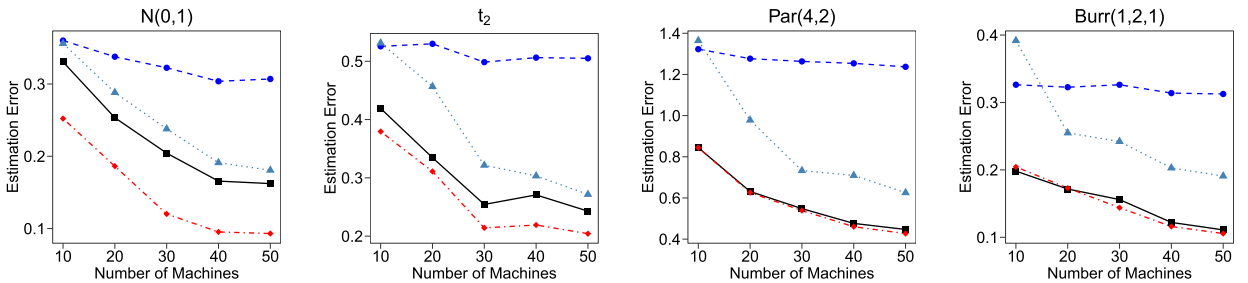

Fig. 3. Plots of estimation error (under ℓ_2-norm) versus the number of machines, over 100 replications, under a high-dimensional heteroscedastic model when $(n, p, s) = (250, 1000, 5)$. Four estimators are presented: centralized ℓ_1-penalized AHR estimator (—■—); DC ℓ_1-penalized AHR estimator (-●- ●-); centralized Lasso estimator (··▲··▲··); and proposed distributed regularized AHR estimator (-◆-◆-).

tend to be wider, and standard errors of the interval width are also larger than those of the AHR method by one order of magnitude.

5.2. Distributed regularized Huber regression

In the high-dimensional setting where the dimension p exceeds the sample size n, we compare four methods across a range of settings: (1) centralized ℓ_1-penalized AHR estimator; (2) DC ℓ_1-penalized AHR estimator; (3) centralized Lasso; and (4) distributed regularized AHR estimator with $T = \lfloor \log(m) \rfloor$ rounds of communication and with a local Lasso estimator as the initialization. All four methods involve a regularization parameter λ, which will be tuned by a held-out validation set of size $\lfloor 0.25N \rfloor$. Similarly to the low-dimensional case, the robustification parameters τ in methods (1), (2) and κ in method (4) are also determined by a self-tuning principle; see equation (3.10) in Wang et al. (2021). The τ value for method (4) is chosen by the validation set approach and the theoretical scaling stated in Corollary 3.1.

As before, we generate $\{(y_i, x_i)\}_{i=1}^N$ from the heteroscedastic model $y_i = x_i^{\mathrm{T}}\beta^* + c^{-1}(x_i^{\mathrm{T}}\beta^*)^2\varepsilon_i$, where $\beta^* = (1.5, 1.5, 1.5, 1.5, 1.5, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^p$, $x_i = (1, x_{i2}, \ldots, x_{ip})^{\mathrm{T}}$ with $x_{ij} \sim \mathcal{N}(0,1)$ for $j = 2, \ldots, p$, and $c = \sqrt{3}\|\beta^*\|_2^2$. The regression errors ε_i are generated from one of the four distributions considered in Section 5.1, which are $\mathcal{N}(0,1)$, t_2 (heavy-tailed and symmetric), Par(4, 2) and Burr(1, 2, 1) (heavy-tailed and skewed). We fix $(n, p) = (250, 1000)$ and let m increase from 10 to 50. Fig. 3 plots the ℓ_2 error $\|\widehat{\beta} - \beta^*\|_2$ versus the number of machines m, averaged over 100 replications, for all four methods. The averaging ℓ_1-penalized AHR estimator has a nondecaying estimation error as m increases, which is expected because of its sub-optimal convergence rate that scales with the local sample size n. The distributed AHR estimator with $T = \lfloor \log(m) \rfloor$ rounds of communication performs as good as the centralized AHR on the entire data set, and has much smaller estimation errors than the centralized Lasso in heavy-tailed cases. Furthermore, from the boxplots displayed in Fig. 4 we see that the distributed AHR improves upon centralized Lasso in terms of both average performance and variability.

6. Conclusion

Distributed inference aims at efficiently combining local information (statistics computed on each local machine) to obtain a global solution that is satisfactory, both in terms of communication costs between the machines and in terms of

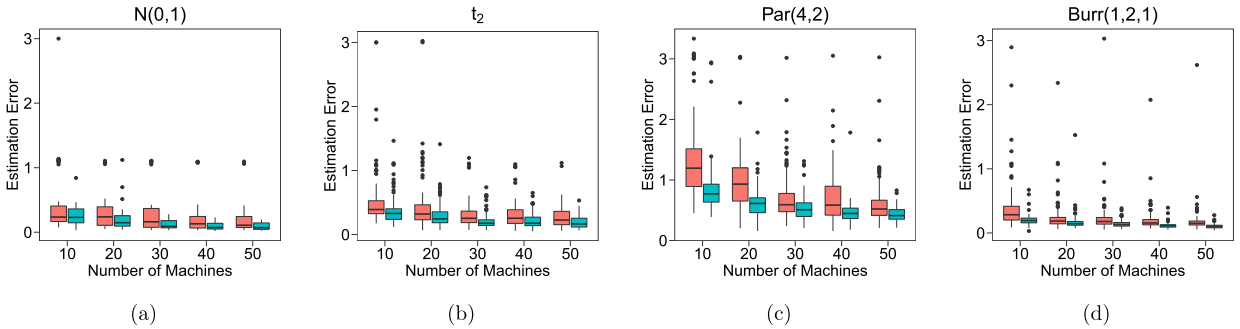

Fig. 4. Boxplots of estimation errors (under ℓ_2-norm) versus the number of machines, over 100 replications, for centralized Lasso ( ) and distributed AHR ( ) under a high-dimensional heteroscedastic model when $(n, p, s) = (250, 1000, 5)$.

statistical accuracy of the final estimator. This paper proposes a new robust algorithm for distributed linear and sparse regressions when data are subject to asymmetric heavy-tailed errors. Founded on the communication-efficient framework proposed by Wang et al. (2017) and Jordan et al. (2019), the new proposal relies on a novel double-robustification approach that applies on both the local and global objective functions. The proposed procedure iteratively minimizes a one-step combination of local and global objectives to improve statistical accuracy. With properly chosen local and global robustification parameters, convergence rates and Bahadur representations are derived for the multi-step estimator. These results show that the optimal rate can be achieved after as many as $\log(m)$ rounds of communication, where m is the number of machines. Under slightly stronger moment conditions, an explicit Berry-Esseen bound is established for the final estimator, based on which asymptotic confidence sets are constructed. In high dimensions, a sparse framework is adopted, where the proposed low-dimensional doubly-robustified objective function is complemented with an ℓ_1-penalty. Near-optimal convergence rates under ℓ_1- and ℓ_2-norms are obtained. Computationally, the proposed procedure employs gradient descent with Barzilai-Borwein step size and the locally adaptive majorize-minimization algorithm to solve the optimization problems, respectively, in low- and high-dimensional settings. To highlight the importance of robustness in distributed inference, this paper closes with extensive numerical studies under models with light- and heavy-tailed, symmetric and asymmetric errors.

Acknowledgements

We sincerely thank the associate editor and the two anonymous reviewers for their many constructive comments and valuable suggestions. Sun was supported in part by the NSERC Grant RGPIN-2018-06484. Zhou acknowledges the support from the National Science Foundation Grant DMS-2113409.

Appendix A. Preliminaries

For any convex function $\psi : \mathbb{R}^k \to \mathbb{R}$, define the corresponding Bregman divergence $D_\psi(w', w) = \psi(w') - \psi(w) - \langle \nabla\psi(w), w' - w\rangle$ and its symmetrized version

$$\overline{D}_\psi(w, w') = D_\psi(w, w') + D_\psi(w', w) = \big\langle \nabla\psi(w) - \nabla\psi(w'), w - w' \big\rangle, \quad w, w' \in \mathbb{R}^k. \tag{A.1}$$

Let $z = \Sigma^{-1/2}x \in \mathbb{R}^p$ be the standardized vector of covariates such that $\mathbb{E}(zz^{\mathrm{T}}) = I_p$, and define $\mu_k = \sup_{u\in\mathbb{S}^{p-1}} \mathbb{E}|z^{\mathrm{T}}u|^k$ for $k \geq 1$. In particular, $\mu_2 = 1$. For every $\delta \in (0, 1]$, define

$$\eta_\delta = \inf\left\{ \eta > 0 : \sup_{u\in\mathbb{S}^{p-1}} \mathbb{E}\big\{(z^{\mathrm{T}}u)^2 \mathbb{1}(|z^{\mathrm{T}}u| > \eta)\big\} \leq \delta \right\}. \tag{A.2}$$

Under Condition (C1), η_δ depends only on δ and υ_1, and the map $\delta \mapsto \eta_\delta$ is non-increasing with $\eta_\delta \downarrow 0$ as $\delta \to 1$. By Markov's inequality,

$$\mathbb{E}\big\{(z^{\mathrm{T}}u)^2 \mathbb{1}(|z^{\mathrm{T}}u| > \eta)\big\} \leq \eta^{-2}\mathbb{E}(z^{\mathrm{T}}u)^4 \leq \eta^{-2}\mu_4 \ \text{ for all } u \in \mathbb{S}^{p-1}.$$

Therefore, a crude bound for η_δ, as a function of δ, is $\eta_\delta \leq (\mu_4/\delta)^{1/2}$.

In Lemmas Appendix A.1 and Appendix A.2 below, we provide a lower bound on the symmetrized Bregman divergence and an upper bound on the score, respectively. The former is a direct consequence of Lemmas C.3 and C.4 in Sun et al. (2020) with slight modifications, and the latter combines Lemmas C.5 and C.6 in Sun et al. (2020) with $\delta = 1$. For the shifted Huber loss $\widetilde{\mathcal{L}}(\cdot)$, note that

$$\overline{D}_{\widetilde{\mathcal{L}}}(\beta,\beta^*)=\big\langle\nabla\widehat{\mathcal{L}}_{1,\kappa}(\beta)-\nabla\widehat{\mathcal{L}}_{1,\kappa}(\beta^*),\beta-\beta^*\big\rangle.$$

Moreover, define the ℓ_1-cone

$$\Lambda=\big\{\beta\in\mathbb{R}^p:\|\beta-\beta^*\|_1\le 4s^{1/2}\|\beta-\beta^*\|_\Sigma\big\}.$$

Lemma Appendix A.1. *Let $\kappa, r>0$ satisfy $\kappa\ge 4\max(\eta_{0.25}r,\sigma)$.*

(i) *Condition (C1) ensures that, with probability at least $1-e^{-u}$,*

$$\overline{D}_{\widetilde{\mathcal{L}}}(\beta,\beta^*)\ge\frac{1}{4}\|\beta-\beta^*\|_\Sigma^2 \ \text{ holds uniformly over } \ \beta\in\Theta(r), \tag{A.3}$$

as long as $n\gtrsim(\kappa/r)^2(p+u)$.

(ii) *Condition (C2) ensures that, with probability at least $1-e^{-u}$,*

$$\overline{D}_{\widetilde{\mathcal{L}}}(\beta,\beta^*)\ge\frac{1}{4}\|\beta-\beta^*\|_\Sigma^2 \ \text{ holds uniformly over } \ \beta\in\Theta(r)\cap\Lambda, \tag{A.4}$$

as long as $n\gtrsim(\kappa/r)^2(s\log p+u)$.

Proof. Without loss of generality, assume $\mathcal{I}_1=\{1,\ldots,n\}$. It suffices to prove (A.4) under Condition (C2). Following the proof of Lemma C.4 in Sun et al. (2020), the key is to upper bound the expected value of the maximum $\|(1/n)\sum_{i=1}^n e_i x_i\|_\infty$, where $e_1,\ldots,e_n$ are independent Rademacher random variables. Let $\mathbb{E}_e$ be the expectation with respect to $e_1,\ldots,e_n$ conditional on the remaining variables. By Hoeffding's moment inequality (see, e.g. Lemma 14.14 in Bühlmann and van de Geer (2011)),

$$\mathbb{E}_e\left\|\frac{1}{n}\sum_{i=1}^n e_i x_i\right\|_\infty\le\max_{1\le j\le p}\left(\frac{1}{n}\sum_{i=1}^n x_{ij}^2\right)^{1/2}\sqrt{\frac{2\log(2p)}{n}}\le B\sqrt{\frac{2\log(2p)}{n}},$$

which in turns implies $\mathbb{E}\|(1/n)\sum_{i=1}^n e_i x_i\|_\infty\le B\sqrt{2\log(2p)/n}$. Keep the rest of the proof the same proves the claimed bound. □

Consider the gradient $\nabla\widehat{\mathcal{L}}_\tau(\cdot)$ evaluated at β^*, namely,

$$\nabla\widehat{\mathcal{L}}_\tau(\beta^*)=-\frac{1}{N}\sum_{i=1}^N\psi_\tau(\varepsilon_i)x_i,$$

where $\psi_\tau(u)=\ell_\tau'(u)$. The following lemma provides high probability bounds on both ℓ_2- and ℓ_∞-norms of $\nabla\widehat{\mathcal{L}}_\tau(\beta^*)$. Recall that $\Omega=\Sigma^{-1}$.

Lemma Appendix A.2. *Let $u>0$ and write $\mathcal{L}_\tau(\cdot)=\mathbb{E}\widehat{\mathcal{L}}_\tau(\cdot)$.*

(i) *Condition (C1) ensures that, with probability at least $1-e^{-u}$,*

$$\|\nabla\widehat{\mathcal{L}}_\tau(\beta^*)-\nabla\mathcal{L}_\tau(\beta^*)\|_{\Sigma^{-1}}\le C_0\Big\{\sigma\sqrt{(p+u)/N}+\tau(p+u)/N\Big\}, \tag{A.5}$$

where $C_0>0$ is a constant depending only on υ_1. Moreover, $\|\nabla\mathcal{L}_\tau(\beta^*)\|_\Omega\le\sigma^2/\tau$.

(ii) *Condition (C2) ensures that, with probability at least $1-e^{-u}$,*

$$\|\nabla\widehat{\mathcal{L}}_\tau(\beta^*)-\nabla\mathcal{L}_\tau(\beta^*)\|_\infty\le\sigma\sigma_u\sqrt{\frac{2\{\log(2p)+u\}}{N}}+\frac{B\tau}{3}\frac{\log(2p)+u}{N}. \tag{A.6}$$

Proof. The bound (A.5) is an immediate consequence of Lemma C.5 in Sun et al. (2020). It suffices to prove (A.6) under Condition (C2). Note that

$$\|\nabla\widehat{\mathcal{L}}_\tau(\beta^*)-\nabla\mathcal{L}_\tau(\beta^*)\|_\infty=\max_{1\le j\le p}\left|\frac{1}{N}\sum_{i=1}^N(1-\mathbb{E})\xi_i x_{ij}\right|,$$

where $\xi_i:=\psi_\tau(\varepsilon_i)$ satisfy $|\xi_i|\le\tau$ and $\mathbb{E}(\xi_i^2|x_i)\le\mathbb{E}(\varepsilon_i^2|x_i)\le\sigma^2$. For any $1\le j\le p$ and $z\ge 0$, applying Bernstein's inequality yields that with probability at least $1-2e^{-z}$,

$$\left|\frac{1}{N}\sum_{i=1}^{N}(1-\mathbb{E})\xi_i x_{ij}\right| \le \sigma_{jj}^{1/2}\sigma\sqrt{\frac{2z}{N}}+\frac{B\tau}{3}\frac{z}{N}.$$

Taking $z=\log(2p)+u$, the claimed bound (A.6) then follows from the union bound. □

Appendix B. Proof of main results

B.1. Proof of Theorem 2.1

Proof of (2.5). For simplicity, we write $\widetilde{\beta}=\widetilde{\beta}^{(1)}$, which minimizes the shifted Huber loss $\widetilde{\mathcal{L}}(\cdot)$ and thus satisfies the first-order condition $\nabla\widetilde{\mathcal{L}}(\widetilde{\beta})=0$. Throughout the proof we assume the event $\mathcal{E}_0(r_0)\cap\mathcal{E}_*(r_*)$ occurs. In view of Lemma Appendix A.1, we consider a local region $\Theta(r_{\mathrm{loc}})$ with $r_{\mathrm{loc}}=\kappa/(4\eta_{0.25})$, and define an intermediate estimator $\widetilde{\beta}_c=(1-c)\beta^*+c\widetilde{\beta}$, where

$$c:=\sup\left\{u\in[0,1]:(1-u)\beta^*+u\widetilde{\beta}\in\Theta(r_{\mathrm{loc}})\right\}\begin{cases}=1 & \text{if } \widetilde{\beta}\in\Theta(r_{\mathrm{loc}}),\\ \in(0,1) & \text{otherwise.}\end{cases}$$

By construction, $\widetilde{\beta}_c\in\Theta(r_{\mathrm{loc}})$. In particular, if $\widetilde{\beta}\notin\Theta(r_{\mathrm{loc}})$, we must have $\widetilde{\beta}_c$ lying on the boundary of $\Theta(r_{\mathrm{loc}})$, i.e. $\|\widetilde{\beta}_c-\beta^*\|_\Sigma=r_{\mathrm{loc}}$.

Applying Lemma C.1 in Sun et al. (2020), we see that the three points $\widetilde{\beta}$, $\widetilde{\beta}_c$ and β^* satisfy $\overline{D}_{\widetilde{\mathcal{L}}}(\widetilde{\beta}_c,\beta^*)\le c\overline{D}_{\widetilde{\mathcal{L}}}(\widetilde{\beta},\beta^*)$, where $\overline{D}_{\widetilde{\mathcal{L}}}(\beta,\beta^*)=\langle\nabla\widetilde{\mathcal{L}}(\beta)-\nabla\widetilde{\mathcal{L}}(\beta^*),\beta-\beta^*\rangle=\langle\nabla\widehat{\mathcal{L}}_{1,\kappa}(\beta)-\nabla\widehat{\mathcal{L}}_{1,\kappa}(\beta^*),\beta-\beta^*\rangle$. Together with the first-order condition $\nabla\widetilde{\mathcal{L}}(\widetilde{\beta})=0$, this implies

$$\overline{D}_{\widetilde{\mathcal{L}}}(\widetilde{\beta}_c,\beta^*)\le -c\left\langle\nabla\widetilde{\mathcal{L}}(\beta^*),\widetilde{\beta}-\beta^*\right\rangle\le\|\nabla\widetilde{\mathcal{L}}(\beta^*)\|_\Omega\cdot\|\widetilde{\beta}_c-\beta^*\|_\Sigma. \tag{B.1}$$

For the left-hand side of (B.1), applying Lemma Appendix A.1 with $r=r_{\mathrm{loc}}$ and the fact $\widetilde{\beta}_c\in\Theta(r_{\mathrm{loc}})$ yields that with probability at least $1-e^{-u}$,

$$\overline{D}_{\widetilde{\mathcal{L}}}(\widetilde{\beta}_c,\beta^*)\ge\frac{1}{4}\|\widetilde{\beta}_c-\beta^*\|_\Sigma^2, \tag{B.2}$$

as long as $n\gtrsim p+u$.

To bound the right-hand side of (B.1), we define vector-valued random processes

$$\begin{cases}\Delta_1(\beta)=\Sigma^{-1/2}\left\{\nabla\widehat{\mathcal{L}}_{1,\kappa}(\beta)-\nabla\widehat{\mathcal{L}}_{1,\kappa}(\beta^*)\right\}-\Sigma^{1/2}(\beta-\beta^*),\\ \Delta(\beta)=\Sigma^{-1/2}\left\{\nabla\widehat{\mathcal{L}}_{\tau}(\beta)-\nabla\widehat{\mathcal{L}}_{\tau}(\beta^*)\right\}-\Sigma^{1/2}(\beta-\beta^*).\end{cases} \tag{B.3}$$

Let $0<r_0\le\sigma$. Following the proof of Theorem B.1 in the supplement of Sun et al. (2020) with $\boldsymbol{B}(\beta)$ therein replaced by $\Delta_1(\beta)$ or $\Delta(\beta)$, it can be similarly shown that, with probability at least $1-2e^{-u}$,

$$\sup_{\beta\in\Theta(r_0)}\|\Delta_1(\beta)\|_2\le C_1\left(\sqrt{\frac{p+u}{n}}+\frac{\sigma^2}{\kappa^2}\right)r_0 \text{ and } \sup_{\beta\in\Theta(r_0)}\|\Delta(\beta)\|_2\le C_1\left(\sqrt{\frac{p+u}{N}}+\frac{\sigma^2}{\tau^2}\right)r_0 \tag{B.4}$$

as long as $n\gtrsim p+u$, where $C_1>0$ is a constant depending only on υ_1. Recall that $\tau\ge\kappa\asymp\sigma\sqrt{n/(p+u)}$. Conditioned on event $\mathcal{E}_0(r_0)\cap\mathcal{E}_*(r_*)$, it follows that

$$\begin{aligned}\|\nabla\widetilde{\mathcal{L}}(\beta^*)\|_\Omega&=\|\Delta(\widetilde{\beta}^{(0)})-\Delta_1(\widetilde{\beta}^{(0)})+\Sigma^{-1/2}\nabla\widehat{\mathcal{L}}_\tau(\beta^*)\|_2\\ &\le\|\Delta(\widetilde{\beta}^{(0)})-\Delta_1(\widetilde{\beta}^{(0)})\|_2+\|\nabla\widehat{\mathcal{L}}_\tau(\beta^*)\|_\Omega\\ &\le C_2 r_0\sqrt{\frac{p+u}{n}}+r_*.\end{aligned} \tag{B.5}$$

Together, the bounds (B.1), (B.2) and (B.5) imply that, conditioning on $\mathcal{E}_0(r_0)\cap\mathcal{E}_*(r_*)$,

$$\|\widetilde{\beta}_c-\beta^*\|_\Sigma\le 4\|\nabla\widetilde{\mathcal{L}}(\beta^*)\|_\Omega\le 4\left(C_2 r_0\sqrt{\frac{p+u}{n}}+r^*\right), \tag{B.6}$$

with probability at least $1-3e^{-u}$. Provided that the sample size is sufficiently large—$n\gtrsim p+u$, the right-hand side of the above inequality is strictly less than $r_{\mathrm{loc}}=\kappa/(4\eta_{0.25})$ with $\kappa\asymp\sigma\sqrt{n/(p+u)}$. As a result, the intermediate estimator $\widetilde{\beta}_c$ falls into the interior of $\Theta(r_{\mathrm{loc}})$ with high probability conditioned on $\mathcal{E}_0(r_0)\cap\mathcal{E}_*(r_*)$. Via proof by contradiction, we must have $\widetilde{\beta}\in\Theta(r_{\mathrm{loc}})$ and hence $\widetilde{\beta}=\widetilde{\beta}_c$; otherwise if $\widetilde{\beta}\notin\Theta(r_{\mathrm{loc}})$, we have demonstrated that $\widetilde{\beta}_c$ must lie on the boundary of $\Theta(r_{\mathrm{loc}})$, which is a contradiction. Consequently, the bound (B.6) also applies to $\widetilde{\beta}$, as claimed.

Proof of (2.6). To establish the Bahadur representation, note that the random process $\Delta_1(\cdot)$ defined in (B.3) can be written as $\Delta_1(\beta) = \Sigma^{-1/2}\{\nabla\widetilde{\mathcal{L}}(\beta) - \nabla\widetilde{\mathcal{L}}(\beta^*)\} - \Sigma^{1/2}(\beta - \beta^*)$. Moreover, note that

$$\nabla\widetilde{\mathcal{L}}(\beta^*) = \nabla\widehat{\mathcal{L}}_{1,\kappa}(\beta^*) - \nabla\widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(0)}) + \nabla\widehat{\mathcal{L}}_{\tau}(\widetilde{\beta}^{(0)}) - \nabla\widehat{\mathcal{L}}_{\tau}(\beta^*) + \nabla\widehat{\mathcal{L}}_{\tau}(\beta^*),$$

which in turn implies

$$\|\nabla\widetilde{\mathcal{L}}(\beta^*) - \nabla\widehat{\mathcal{L}}_{\tau}(\beta^*)\|_{\Omega} \le \|\Delta_1(\widetilde{\beta}^{(0)})\|_2 + \|\Delta(\widetilde{\beta}^{(0)})\|_2.$$

Recall that $\nabla\widetilde{\mathcal{L}}(\widetilde{\beta}) = 0$, and by (B.6), $\|\widetilde{\beta} - \beta^*\|_{\Sigma} \le r_1 := 4C_2 r_0\sqrt{(p+u)/n} + 4r_*$ with high probability conditioned on $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*)$. For $r_0 \ge 8r_*$, we have $r_1 \le r_0/2 + r_0/2 = r_0$ as long as $n \gtrsim p + u$, and hence $\widetilde{\beta} \in \Theta(r_0)$. Applying the bounds in (B.4) again, we obtain that conditioned on $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*)$,

$$\begin{aligned}
&\|\Sigma^{1/2}(\widetilde{\beta} - \beta^*) + \Sigma^{-1/2}\nabla\widehat{\mathcal{L}}_{\tau}(\beta^*)\|_2 \\
&= \|\Delta_1(\widetilde{\beta}) + \Sigma^{-1/2}\nabla\widetilde{\mathcal{L}}(\beta^*) - \Sigma^{-1/2}\nabla\widehat{\mathcal{L}}_{\tau}(\beta^*)\|_2 \\
&\le \|\Delta_1(\widetilde{\beta})\|_2 + \|\Delta_1(\widetilde{\beta}^{(0)})\|_2 + \|\Delta(\widetilde{\beta}^{(0)})\|_2 \\
&\le 2\sup_{\beta\in\Theta(r_0)} \|\Delta_1(\beta)\|_2 + \sup_{\beta\in\Theta(r_0)} \|\Delta(\beta)\|_2 \\
&\lesssim \sqrt{\frac{p+u}{n}} \cdot r_0,
\end{aligned}$$

with probability at least $1 - 3e^{-u}$. This completes the proof. □

B.2. Proof of Theorem 2.2

Given a sequence of iterates $\{\widetilde{\beta}^{(t)}\}_{t=0,1,\ldots,T}$, we define "good" events

$$\mathcal{E}_t(r_t) = \big\{\widetilde{\beta}^{(t)} \in \Theta(r_t)\big\}, \quad t = 0, \ldots, T, \tag{B.7}$$

for some sequence of radii $r_0 \ge r_1 \ge \cdots \ge r_T > 0$ to be determined. Examine the proof of Theorem 2.1, we see that the statistical properties of $\widetilde{\beta}^{(t)}$ depend on both first-order and second-order information of the loss function $\widetilde{\mathcal{L}}^{(t)}(\cdot)$, namely, the ℓ_2-norm of the gradient $\nabla\widetilde{\mathcal{L}}^{(t)}(\beta^*)$ and the (symmetrized) Bregman divergence of $\widetilde{\mathcal{L}}^{(t)}(\cdot)$. For the former, we have

$$\nabla\widetilde{\mathcal{L}}^{(t)}(\beta^*) = \nabla\widehat{\mathcal{L}}_{1,\kappa}(\beta^*) - \nabla\widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(t-1)}) + \nabla\widehat{\mathcal{L}}_{\tau}(\widetilde{\beta}^{(t-1)}). \tag{B.8}$$

Let $\Delta_1(\cdot)$ and $\Delta(\cdot)$ be the random processes defined in (B.3), and observe that $\Sigma^{-1/2}\nabla\widetilde{\mathcal{L}}^{(t)}(\beta^*) = \Delta(\widetilde{\beta}^{(t-1)}) - \Delta_1(\widetilde{\beta}^{(t-1)}) + \Sigma^{-1/2}\nabla\widehat{\mathcal{L}}_{\tau}(\beta^*)$. By the triangle inequality,

$$\|\nabla\widetilde{\mathcal{L}}^{(t)}(\beta^*)\|_{\Omega} \le \|\Delta(\widetilde{\beta}^{(t-1)})\|_2 + \|\Delta_1(\widetilde{\beta}^{(t-1)})\|_2 + \|\nabla\widehat{\mathcal{L}}_{\tau}(\beta^*)\|_{\Omega}. \tag{B.9}$$

On the other hand, note that the shifted Huber losses $\widetilde{\mathcal{L}}^{(t)}(\cdot)$ have the same Bregman divergence, denoted by

$$\overline{D}(\beta_1, \beta_2) = \langle\nabla\widetilde{\mathcal{L}}^{(t)}(\beta_1) - \nabla\widetilde{\mathcal{L}}^{(t)}(\beta_2), \beta_1 - \beta_2\rangle = \langle\nabla\widehat{\mathcal{L}}_{1,\kappa}(\beta_1) - \nabla\widehat{\mathcal{L}}_{1,\kappa}(\beta_2), \beta_1 - \beta_2\rangle.$$

Define the local radius $r_{\mathrm{loc}} = \kappa/(4\eta_{0.25})$. Then, applying Lemma Appendix A.1 with $r = r_{\mathrm{loc}}$ yields that, with probability at least $1 - e^{-u}$,

$$\overline{D}(\beta, \beta^*) \ge \frac{1}{4}\|\beta - \beta^*\|_{\Sigma}^2 \tag{B.10}$$

holds uniformly over $\beta \in \Theta(r_{\mathrm{loc}})$. Let $\mathcal{E}_{\mathrm{lsc}}$ be the event that the local strong convexity (B.10) holds.

With the above preparations, we are ready to extend the argument in the proof of Theorem 2.1 to deal with $\widetilde{\beta}^{(t)}$ sequentially. At each iteration, we construct an intermediate estimator $\widetilde{\beta}^{(t)}_{\mathrm{imd}}$—a convex combination of $\widetilde{\beta}^{(t)}$ and β^*—which falls in $\Theta(r_{\mathrm{loc}})$ and satisfies

$$\overline{D}(\widetilde{\beta}^{(t)}_{\mathrm{imd}}, \beta^*) \le \|\nabla\widetilde{\mathcal{L}}^{(t)}(\beta^*)\|_{\Omega} \cdot \|\widetilde{\beta}^{(t)}_{\mathrm{imd}} - \beta^*\|_{\Sigma}.$$

If event $\mathcal{E}_*(r_*) \cap \mathcal{E}_{\mathrm{lsc}}$ occurs, the bounds (B.9) and (B.10) imply

$$\|\widetilde{\beta}^{(t)}_{\mathrm{imd}} - \beta^*\|_{\Sigma} \le 4\big\{\|\Delta_1(\widetilde{\beta}^{(t-1)})\|_2 + \|\Delta(\widetilde{\beta}^{(t-1)})\|_2\big\} + 4r_*. \tag{B.11}$$

Moreover, it follows from (B.8) and the first-order condition $\nabla\widetilde{\mathcal{L}}^{(t)}(\widetilde{\beta}^{(t)}) = 0$ that

$$\begin{aligned}
&\|\Sigma^{1/2}(\widetilde{\beta}^{(t)}-\beta^*)+\Sigma^{-1/2}\nabla\widehat{\mathcal{L}}_\tau(\beta^*)\|_2\\
&=\|\Sigma^{-1/2}\{\nabla\widetilde{\mathcal{L}}^{(t)}(\widetilde{\beta}^{(t)})-\nabla\widetilde{\mathcal{L}}^{(t)}(\beta^*)\}-\Sigma^{1/2}(\widetilde{\beta}^{(t)}-\beta^*)+\Sigma^{-1/2}\{\nabla\widetilde{\mathcal{L}}^{(t)}(\beta^*)-\nabla\widehat{\mathcal{L}}_\tau(\beta^*)\}\|_2\\
&\le\|\Delta_1(\widetilde{\beta}^{(t)})\|_2+\|\Delta_1(\widetilde{\beta}^{(t-1)})\|_2+\|\Delta(\widetilde{\beta}^{(t-1)})\|_2. \qquad \text{(B.12)}
\end{aligned}$$

In view of the bounds in (B.4), for every $0<r\le\sigma$ we define the event

$$\mathcal{F}(r)=\Big\{\sup_{\beta\in\Theta(r)}\big\{\|\Delta_1(\beta)\|_2+\|\Delta(\beta)\|_2\big\}\le\gamma(u)\cdot r\Big\}, \qquad \text{(B.13)}$$

with $\gamma(u)=C\sqrt{(p+u)/n}$ for some $C>0$, which satisfies $\mathbb{P}\{\mathcal{F}(r)\}\ge 1-2e^{-u}$.

Let $8r^*\le r_0\le\sigma$. In the following, we deal with $\{(\widetilde{\beta}_{\rm imd}^{(t)},\widetilde{\beta}^{(t)}),t=1,2,\ldots,T\}$ sequentially conditioning on the event $\mathcal{E}_0(r_0)\cap\mathcal{E}_*(r_*)\cap\mathcal{E}_{\rm lsc}$. At iteration 1, it follows from (B.11) that, conditioned on $\mathcal{F}(r_0)$,

$$\|\widetilde{\beta}_{\rm imd}^{(1)}-\beta^*\|_\Sigma\le r_1:=4\gamma(u)\cdot r_0+4r_*.$$

Provided that $n\gtrsim p+u$, we have $4\gamma(u)\le 1/2<1$ and $r_1\le r_0<r_{\rm loc}=\kappa/(4\eta_{0.25})$, so that $\widetilde{\beta}_{\rm imd}^{(1)}\in\Theta(r_1)\subseteq{\rm int}(\Theta(r_{\rm loc}))$. Via proof by contradiction, we must have $\widetilde{\beta}^{(1)}=\widetilde{\beta}_{\rm imd}^{(1)}\in\Theta(r_{\rm loc})$, which in turns certifies event $\mathcal{E}(r_1)$. Combining this with (B.12), we see that conditioned on $\mathcal{F}(r_0)$, the event $\mathcal{E}_1(r_1)$ must happen and hence

$$\begin{cases}\|\widetilde{\beta}^{(1)}-\beta^*\|_\Sigma\le r_1=4\gamma(u)\cdot r_0+4r_*\le r_0,\\ \|\widetilde{\beta}^{(1)}-\beta^*+\Sigma^{-1}\nabla\widehat{\mathcal{L}}_\tau(\beta^*)\|_\Sigma\le 2\gamma(u)\cdot r_0.\end{cases}.$$

Now assume that for some $t\ge 1$, $\widetilde{\beta}^{(t)}\in\Theta(r_t)$ with $r_t=4\gamma(u)\cdot r_{t-1}+4r_*\le r_{t-1}<r_{\rm loc}$. At $(t+1)$-th iteration, applying (B.11) again yields that, conditioned on event $\mathcal{E}_t(r_t)\cap\mathcal{F}(r_t)$,

$$\|\widetilde{\beta}_{\rm imd}^{(t+1)}-\beta^*\|_\Sigma\le r_{t+1}:=4\gamma(u)\cdot r_t+4r_*.$$

By induction, $r_t\le r_{t-1}<r_{\rm loc}$ so that $r_{t+1}\le 4\gamma(u)\cdot r_{t-1}+4r_*=r_t<r_{\rm loc}$. This implies that $\widetilde{\beta}_{\rm imd}^{(t+1)}$ falls into the interior of $\Theta(r_{\rm loc})$, which enforces $\widetilde{\beta}^{(t+1)}=\widetilde{\beta}_{\rm imd}^{(t+1)}\in\Theta(r_{t+1})$ and thus certifies event $\mathcal{E}_{t+1}(r_{t+1})$. Combining this with the bound (B.12), we find that

$$\begin{cases}\|\widetilde{\beta}^{(t+1)}-\beta^*\|_\Sigma\le r_{t+1}=4\gamma(u)\cdot r_t+4r_*\le r_t,\\ \|\widetilde{\beta}^{(t+1)}-\beta^*+\Sigma^{-1}\nabla\widehat{\mathcal{L}}_\tau(\beta^*)\|_\Sigma\le 2\gamma(u)\cdot r_t.\end{cases}.$$

Repeat the above argument until we obtain $\widetilde{\beta}^{(T)}$. We have shown that conditioned on $\mathcal{E}_*(r_*)\cap\mathcal{E}_{\rm lsc}\cap\mathcal{E}_{t-1}(r_{t-1})\cap\mathcal{F}(r_{t-1})$ for every $0\le t\le T-1$, the event $\mathcal{E}_t(r_t)$ must occur. Therefore, conditioned on $\mathcal{E}_*(r_*)\cap\mathcal{E}_{\rm lsc}\cap\mathcal{E}_0(r_0)\cap\{\cap_{t=0}^{T-1}\mathcal{F}(r_t)\}$, $\widetilde{\beta}^{(T)}$ satisfies the bounds

$$\begin{cases}\|\widetilde{\beta}^{(T)}-\beta^*\|_\Sigma\le r_T=4\gamma(u)\cdot r_{T-1}+4r_*,\\ \|\widetilde{\beta}^{(T)}-\beta^*+\Sigma^{-1}\nabla\widehat{\mathcal{L}}_\tau(\beta^*)\|_\Sigma\le 2\gamma(u)\cdot r_{T-1}.\end{cases} \qquad \text{(B.14)}$$

Observe that $r_t=\{4\gamma(u)\}^t r_0+\frac{1-\{4\gamma(u)\}^t}{1-4\gamma(u)}4r_*$ for $t=1,\ldots,T$. We choose T to be the smallest integer such that $\{4\gamma(u)\}^{T-1}r_0\le r_*$, that is, $T=\lceil\log(r_0/r_*)/\log(1/\{4\gamma(u)\})\rceil+1$. Consequently, the bounds in (B.14) become

$$\begin{cases}\|\widetilde{\beta}^{(T)}-\beta^*\|_\Sigma\le\big\{\gamma(u)+\frac{1}{1-4\gamma(u)}\big\}4r_*\le\{4\gamma(u)+8\}r_*,\\ \|\widetilde{\beta}^{(T)}-\beta^*+\Sigma^{-1}\nabla\widehat{\mathcal{L}}_\tau(\beta^*)\|_\Sigma\le 18\gamma(u)\cdot r_*.\end{cases} \qquad \text{(B.15)}$$

Finally, it suffices to show that the event $\mathcal{E}_{\rm lsc}\cap\{\cap_{t=0}^{T-1}\mathcal{F}(r_t)\}$ occurs with high probability. Recall from (B.10) and (B.13) that $\mathbb{P}(\mathcal{E}_{\rm lsc})\ge 1-e^{-u}$ and $\mathbb{P}\{\mathcal{F}(r_t)\}\ge 1-2e^{-u}$ for every $t=0,1,\ldots,T-1$. The claimed result then follows from (B.15) and the union bound. □

B.3. Proof of Theorem 2.3

For simplicity, we write $q=p+\log n+\log_2 m$ throughout the proof. For every vector $a\in\mathbb{R}^p$, define $S_a=N^{-1/2}\sum_{i=1}^N\xi_i w_i$ and $S_a^0=S_a-\mathbb{E}S_a$, where $\xi_i=\psi_\tau(\varepsilon_i)$ and $w_i=a^{\rm T}\Sigma^{-1}x_i$. Under the moment condition $\mathbb{E}(|\varepsilon|^{2+\delta}|x)\le\nu_{2+\delta}$, using Markov's inequality yields $|\mathbb{E}(\xi_i|x_i)|\le\tau^{-1-\delta}\mathbb{E}(|\varepsilon_i|^{2+\delta}|x_i)\le\nu_{2+\delta}\tau^{-1-\delta}$. Hence, $|\mathbb{E}(\xi_i w_i)|\le\nu_{2+\delta}\|a\|_\Omega\cdot\tau^{-1-\delta}$ and $|\mathbb{E}S_a|\le\nu_{2+\delta}\|a\|_\Omega\cdot N^{1/2}\tau^{-1-\delta}$.

With the above preparations, we are ready to prove the normal approximation for $\widetilde{\beta}$. Note that

$$\begin{aligned}
&|N^{1/2}a^{\mathrm{T}}(\widetilde{\beta}-\beta^*)-S_a^0| \\
&\le N^{1/2}\left|\left\langle \Sigma^{-1/2}a, \Sigma^{1/2}(\widetilde{\beta}-\beta^*)-\Sigma^{-1/2}\frac{1}{N}\sum_{i=1}^{N}\psi_\tau(\varepsilon_i)x_i\right\rangle\right| + |\mathbb{E}S_a| \\
&\le N^{1/2}\|a\|_\Omega \cdot \left\|\widetilde{\beta}-\beta^*-\Sigma^{-1}\frac{1}{N}\sum_{i=1}^{N}\psi_\tau(\varepsilon_i)x_i\right\|_\Sigma + \nu_{2+\delta}\|a\|_\Omega \cdot N^{1/2}\tau^{-1-\delta}.
\end{aligned}$$

Applying (2.10) in Theorem 2.1, we find that with probability at least $1-Cn^{-1}$,

$$|N^{1/2}a^{\mathrm{T}}(\widetilde{\beta}-\beta^*)-S_a^0| \le C_1\|a\|_\Omega \cdot \big(\sigma qn^{-1/2}+N^{1/2}\nu_{2+\delta}\tau^{-1-\delta}\big), \tag{B.16}$$

where $C_1>0$ is a constant independent of (N,n,p).

For the centered partial sum S_a^0, it follows from the Berry-Esseen inequality (see, e.g. Theorem 2.1 in Chen and Shao (2001)) that

$$\sup_{t\in\mathbb{R}}\big|\mathbb{P}\big\{S_a^0 \le \mathrm{var}(S_a^0)^{1/2}t\big\}-\Phi(t)\big| \le 4.1\frac{\mathbb{E}|\xi w-\mathbb{E}(\xi w)|^{2+\delta}}{\mathrm{var}(\xi w)^{1+\delta/2}N^{\delta/2}}, \tag{B.17}$$

where $\xi=\psi_\tau(\varepsilon)$ and $w=a^{\mathrm{T}}\Sigma^{-1}x$. Recall that $\tau \asymp \sigma\sqrt{N/q}$, and write $\sigma_{\tau,a}^2=\mathbb{E}(\xi w)^2$. By Proposition A.2 in Zhou et al. (2018), $|\mathbb{E}(\xi^2|x)-\sigma^2| \le 2\delta^{-1}\nu_{2+\delta}\tau^{-\delta} \asymp \delta^{-1}\nu_{2+\delta}\sigma^{-\delta}(q/N)^{\delta/2}$, and hence

$$\big|\sigma_{\tau,a}^2/(\sigma\|a\|_\Omega)^2-1\big| \lesssim \frac{\nu_{2+\delta}}{\delta\sigma^{2+\delta}}\left(\frac{q}{N}\right)^{\delta/2}. \tag{B.18}$$

Moreover, $\mathbb{E}|\xi w|^{2+\delta} \le \mathbb{E}|\varepsilon w|^{2+\delta} \le \mu_{2+\delta}\|a\|_\Omega^{2+\delta}\nu_{2+\delta}$, where $\mu_{2+\delta}:=\sup_{u\in\mathbb{S}^{p-1}}\mathbb{E}|z^{\mathrm{T}}u|^{2+\delta}$ depends only on υ_1 under Condition (C1). Substituting these bounds into (B.17) yields

$$\sup_{t\in\mathbb{R}}\big|\mathbb{P}\big\{S_a^0 \le \mathrm{var}(S_a^0)^{1/2}t\big\}-\Phi(t)\big| \le C_2\frac{\nu_{2+\delta}}{\sigma^{2+\delta}N^{\delta/2}}, \tag{B.19}$$

provided that $N \gtrsim q$. For the variance term, the bound $|\mathbb{E}(\xi|x)| \le \sigma^2\tau^{-1}$ guarantees that

$$\mathbb{E}(\xi w)^2 \ge \mathrm{var}(S_a^0)=\mathbb{E}(\xi w)^2-(\mathbb{E}\xi w)^2 \ge \mathbb{E}(\xi w)^2-(\sigma\|a\|_\Omega)^2\cdot\sigma^2\tau^{-2}.$$

Combined with (B.18), this implies $|\mathrm{var}(S_a^0)/\sigma_{\tau,a}^2-1| \lesssim \sigma^2\tau^{-2}$, from which it follows that

$$\sup_{t\in\mathbb{R}}\big|\Phi(t/\mathrm{var}(S_a^0)^{1/2})-\Phi(t/\sigma_{\tau,a})\big| \le C_3\frac{\sigma^2}{\tau^2}. \tag{B.20}$$

Let $G\sim\mathcal{N}(0,1)$ and $t\in\mathbb{R}$. Combining the bounds (B.16), (B.19) and (B.20), we obtain

$$\begin{aligned}
&\mathbb{P}\big\{N^{1/2}a^{\mathrm{T}}(\widetilde{\beta}-\beta^*)\le t\big\} \\
&\le \mathbb{P}\big\{S_a^0 \le x+C_1\|a\|_\Omega\cdot\big(\sigma qn^{-1/2}+N^{1/2}\nu_{2+\delta}\tau^{-1-\delta}\big)\big\}+Cn^{-1} \\
&\le \mathbb{P}\big\{\mathrm{var}(S_a^0)^{1/2}G \le t+C_1\|a\|_\Omega\cdot\big(\sigma qn^{-1/2}+N^{1/2}\nu_{2+\delta}\tau^{-1-\delta}\big)\big\}+Cn^{-1}+C_2\frac{\nu_{2+\delta}}{\sigma^{2+\delta}N^{\delta/2}} \\
&\le \mathbb{P}\big\{\sigma_{\tau,a}G \le t+C_1\|a\|_\Omega\cdot\big(\sigma qn^{-1/2}+N^{1/2}\nu_{2+\delta}\tau^{-1-\delta}\big)\big\}+C_2\frac{\nu_{2+\delta}}{\sigma^{2+\delta}N^{\delta/2}}+C_3\frac{\sigma^2}{\tau^2} \\
&\le \mathbb{P}\big(\sigma_{\tau,a}G\le t\big)+Cn^{-1}+C_1(2\pi)^{-1/2}\big(qn^{-1/2}+N^{1/2}\nu_{2+\delta}\sigma^{-1}\tau^{-1-\delta}\big)+C_2\frac{\nu_{2+\delta}}{\sigma^{2+\delta}N^{\delta/2}}+C_3\frac{\sigma^2}{\tau^2}.
\end{aligned}$$

A similar argument leads to a series of reverse inequalities, and thus completes the proof. □

B.4. Proof of Proposition 2.1

Consider the change of variable $\delta=\Sigma^{1/2}(\beta-\beta^*)$, so that $\beta\in\Theta(r)$ is equivalent to $\delta\in\mathbb{B}^p(r)$—the ℓ_2-ball in $\mathbb{R}^p$ with center 0 and radius r. For $\delta\in\mathbb{R}^p$, define

$$\widehat{\sigma}^2(\delta)=\frac{1}{N}\sum_{i=1}^{N}\psi_\tau^2(\varepsilon_i-z_i^{\mathrm{T}}\delta) \text{ and } \sigma^2(\delta)=\mathbb{E}\widehat{\sigma}^2(\delta), \tag{B.21}$$

where $z_i = \Sigma^{-1/2} x_i$. Then $\widehat{\sigma}_\varepsilon^2 = \widehat{\sigma}^2(\widehat{\delta})$ with $\widehat{\delta} = \widetilde{\beta} - \beta^*$. Conditioned on the event $\{\widetilde{\beta} \in \Theta(r)\}$ for some predetermined $r > 0$, it suffices to bound $\sup_{\delta \in \mathbb{B}^p(r)} |\widehat{\sigma}^2(\delta) - \sigma^2|$.

For any $\epsilon \in (0, r)$, there exists an ϵ-net $\{\delta_1, \ldots, \delta_{K_\epsilon}\}$ with $K_\epsilon \le (1 + 2r/\epsilon)^p$ satisfying that, for every $\delta \in \mathbb{B}^p(r)$, there exists some $1 \le k \le K_\epsilon$ such that $\|\delta - \delta_k\|_2 \le \epsilon$. Consequently,

$$|\widehat{\sigma}^2(\delta) - \sigma^2| \le |\widehat{\sigma}^2(\delta) - \widehat{\sigma}^2(\delta_k)| + |\widehat{\sigma}^2(\delta_k) - \sigma^2(\delta_k)| + |\sigma^2(\delta_k) - \sigma^2|. \tag{B.22}$$

Recall that the function $\psi_\tau(\cdot)$ satisfies $\sup_t |\psi_\tau(t)| \le \tau$ and $|\psi_\tau(t_1) - \psi_\tau(t_2)| \le |t_1 - t_2|$ for any $t_1, t_2 \in \mathbb{R}$. Hence,

$$\begin{aligned} |\widehat{\sigma}^2(\delta) - \sigma^2| &\le \frac{1}{N} \sum_{i=1}^N \big| \psi_\tau^2(\varepsilon_i - z_i^{\mathrm{T}} \delta) - \psi_\tau^2(\varepsilon_i - z_i^{\mathrm{T}} \delta_k) \big| \\ &\le \frac{2\tau}{N} \sum_{i=1}^N |z_i^{\mathrm{T}}(\delta - \delta_k)| \le 2\tau\epsilon \cdot \bigg\| \frac{1}{N} \sum_{i=1}^N z_i z_i^{\mathrm{T}} \bigg\|_2 \end{aligned}$$

holds uniformly over all (δ, δ_k) pairs. For the last term on the right-hand side of (B.22), since $\delta_k \in \mathbb{B}^p(r)$ and $|\psi_\tau(t) \le |t|$, we have

$$|\sigma^2(\delta_k) - \sigma^2| \le \mathbb{E}(2|\varepsilon| + |z^{\mathrm{T}} \delta_k|) \cdot |z^{\mathrm{T}} \delta_k| \le 2\sigma r + r^2.$$

Back to (B.22), first taking the maximum over $k \in \{1, \ldots, K_\epsilon\}$, and then taking the supremum over $\delta \in \mathbb{B}^p(r)$, we conclude that

$$\sup_{\delta \in \mathbb{B}^p(r)} |\widehat{\sigma}^2(\delta) - \sigma^2| \le 2\tau\epsilon \cdot \bigg\| \frac{1}{N} \sum_{i=1}^N z_i z_i^{\mathrm{T}} \bigg\|_2 + \max_{1 \le k \le K_\epsilon} |\widehat{\sigma}^2(\delta_k) - \sigma^2(\delta_k)| + r(2\sigma + r). \tag{B.23}$$

For $\|(1/N) \sum_{i=1}^N z_i z_i^{\mathrm{T}}\|_2$, using the same covering argument along with Bernstein's inequality (see, e.g. Theorem 5.39 and Remark 5.40 in Vershynin (2012)), it can be shown that with probability at least $1 - e^{-t}$,

$$\bigg\| \frac{1}{N} \sum_{i=1}^N z_i z_i^{\mathrm{T}} - \mathrm{I}_p \bigg\|_2 \lesssim \sqrt{\frac{p+t}{N}} \bigvee \frac{p+t}{N}. \tag{B.24}$$

It remains to bound $|\widehat{\sigma}^2(\delta_k) - \sigma^2(\delta_k)|$ for each k. Note that $\psi_\tau^2(\varepsilon_i - z_i^{\mathrm{T}} \delta_k) \le \tau^2$ and

$$\begin{aligned} \mathbb{E}\big\{ \psi_\tau^4(\varepsilon_i - z_i^{\mathrm{T}} \delta_k) \big\} &\le \tau^{2-\delta} \mathbb{E}\big\{ |\psi_\tau(\varepsilon_i - z_i^{\mathrm{T}} \delta_k)|^{2+\delta} \big\} \\ &\le \tau^{2-\delta} 2^{1+\delta} \mathbb{E}\big(|\varepsilon_i|^{2+\delta} + |z_i^{\mathrm{T}} \delta_k|^{2+\delta} \big) \\ &\le \tau^{2-\delta} 2^{1+\delta} \big(\nu_{2+\delta} + \mu_{2+\delta} r^{2+\delta} \big). \end{aligned}$$

By Bernstein's inequality, we have that with probability at least $1 - 2e^{-t}$,

$$|\widehat{\sigma}^2(\delta_k) - \sigma^2(\delta_k)| \le 2^{1+\delta/2} \big(\nu_{2+\delta} + \mu_{2+\delta} r^{2+\delta} \big)^{1/2} \tau^{1-\delta/2} \sqrt{\frac{t}{N}} + \tau^2 \frac{t}{3N}.$$

Taking the union bound over $k = 1, \ldots, K_\epsilon$ yields

$$\begin{aligned} &\max_{1 \le k \le K_\epsilon} |\widehat{\sigma}^2(\delta_k) - \sigma^2(\delta_k)| \\ &\le 2^{1+\delta/2} \big(\nu_{2+\delta} + \mu_{2+\delta} r^{2+\delta} \big)^{1/2} \tau^{1-\delta/2} \sqrt{\frac{\log(2K_\epsilon) + t}{N}} + \tau^2 \frac{\log(2K_\epsilon) + t}{3N} \end{aligned} \tag{B.25}$$

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

Finally, we set $\epsilon = r/N$ so that $K_\epsilon \le (1 + 2N)^p$. Together, (B.23), (B.24) and (B.25) with $r \lesssim \sigma$ prove the claimed bound. □

B.5. Proof of Theorem 3.1

As before, we assume without loss of generality that $\mathcal{I}_1 = \{1, \ldots, n\}$. Write $\widetilde{\beta} = \widetilde{\beta}^{(1)}$ for simplicity, and let $h = \widetilde{\beta} - \beta^*$ be the error vector. By the first-order optimality condition, there exists a subgradient $g \in \partial \|\widetilde{\beta}\|_1$ such that $g^{\mathrm{T}} \widetilde{\beta} = \|\widetilde{\beta}\|_1$ and $\nabla \widetilde{\mathcal{L}}(\widetilde{\beta}) + \lambda \cdot g = 0$. Moreover, the convexity of $\widetilde{\mathcal{L}}(\cdot)$ implies

$$0 \le \overline{D}_{\widetilde{\mathcal{L}}}(\widetilde{\beta}, \beta^*) = h^{\mathrm{T}} \big\{ \nabla \widetilde{\mathcal{L}}(\widetilde{\beta}) - \nabla \widetilde{\mathcal{L}}(\beta^*) \big\} = -\lambda \cdot h^{\mathrm{T}} g - h^{\mathrm{T}} \nabla \widetilde{\mathcal{L}}(\beta^*).$$

Recall the true active set $\mathcal{S} = \mathrm{supp}(\beta^*) \subseteq \{1, \ldots, p\}$, we have

$$-h^{\mathrm{T}} g \le \|\beta^*\|_1 - \|\widetilde{\beta}\|_1 = \|\beta^*_{\mathcal{S}}\|_1 - \|h_{\mathcal{S}^{\mathrm{c}}}\|_1 - \|h_{\mathcal{S}} + \beta^*_{\mathcal{S}}\|_1 \le \|h_{\mathcal{S}}\|_1 - \|h_{\mathcal{S}^{\mathrm{c}}}\|_1.$$

Together, the above two displays yield

$$0 \le \overline{D}_{\widetilde{\mathcal{L}}}(\widetilde{\beta}, \beta^*) \le \lambda\big(\|h_{\mathcal{S}}\|_1 - \|h_{\mathcal{S}^{\mathrm{c}}}\|_1\big) - h^{\mathrm{T}} \nabla \widetilde{\mathcal{L}}(\beta^*). \tag{B.26}$$

To deal with $\nabla \widetilde{\mathcal{L}}(\beta^*) = \nabla \widehat{\mathcal{L}}_{1,\kappa}(\beta^*) - \nabla \widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(0)}) + \nabla \widehat{\mathcal{L}}_{\tau}(\widetilde{\beta}^{(0)})$, we define random processes

$$\widehat{D}_1(\beta) = \nabla \widehat{\mathcal{L}}_{1,\kappa}(\beta) - \nabla \widehat{\mathcal{L}}_{1,\kappa}(\beta^*), \quad \widehat{D}(\beta) = \nabla \widehat{\mathcal{L}}_{\tau}(\beta) - \nabla \widehat{\mathcal{L}}_{\tau}(\beta^*),$$

and write $D_1(\beta) = \mathbb{E}\widehat{D}_1(\beta)$ and $D(\beta) = \mathbb{E}\widehat{D}(\beta)$. The gradient $\nabla \widetilde{\mathcal{L}}(\beta^*)$ can thus be written as

$$\big\{\widehat{D}(\beta) - D(\beta)\big\}\Big|_{\beta=\widetilde{\beta}^{(0)}} + \big\{D_1(\beta) - \widehat{D}_1(\beta)\big\}\Big|_{\beta=\widetilde{\beta}^{(0)}} + \nabla \widehat{\mathcal{L}}_{\tau}(\beta^*) - \nabla \mathcal{L}_{\tau}(\beta^*)$$
$$+ \big\{D(\beta) - D_1(\beta)\big\}\Big|_{\beta=\widetilde{\beta}^{(0)}} + \nabla \mathcal{L}_{\tau}(\beta^*).$$

For any $r > 0$, define

$$\Delta_1(r) = \sup_{\beta \in \Theta(r) \cap \Lambda} \|\widehat{D}_1(\beta) - D_1(\beta)\|_\infty, \quad \Delta(r) = \sup_{\beta \in \Theta(r) \cap \Lambda} \|\widehat{D}(\beta) - D(\beta)\|_\infty, \tag{B.27}$$

$$\delta(r) = \sup_{\beta \in \Theta(r)} \|D_1(\beta) - D(\beta)\|_\Omega \text{ and } b^* = \|\nabla \mathcal{L}_{\tau}(\beta^*)\|_\Omega. \tag{B.28}$$

The quantity b^* can be viewed as the robustification bias and by Lemma Appendix A.2, $b^* \le \sigma^2 \tau^{-1}$.

Back to the right-hand of (B.26), conditioning on the event $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(\lambda_*)$, it follows from Hölder's inequality that

$$|h^{\mathrm{T}} \nabla \widetilde{\mathcal{L}}(\beta^*)| \le \big\{\Delta(r_0) + \Delta_1(r_0) + \lambda_*\big\}\|h\|_1 + \big\{\delta(r_0) + b^*\big\}\|h\|_\Sigma. \tag{B.29}$$

Let $\lambda = 2.5(\lambda_* + \rho)$ for some $\rho > 0$. Provided that

$$\rho \ge \max\big[\Delta(r_0) + \Delta_1(r_0), s^{-1/2}\big\{\delta(r_0) + b^*\big\}\big], \tag{B.30}$$

we have $|h^{\mathrm{T}} \nabla \widetilde{\mathcal{L}}(\beta^*)| \le 0.4\lambda\|h\|_1 + 0.4s^{1/2}\lambda\|h\|_\Sigma$. Combined with (B.26), this yields $0 \le 1.4\|h_{\mathcal{S}}\|_1 - 0.6\|h_{\mathcal{S}^{\mathrm{c}}}\|_1 + 0.4s^{1/2}\|h\|_\Sigma$. Consequently, with $\lambda_l = \lambda_{\min}(\Sigma) = 1$, we have $\|h\|_1 \le (10/3)\|h_{\mathcal{S}}\|_1 + (2/3)s^{1/2}\|h\|_\Sigma \le 4s^{1/2}\|h\|_\Sigma$, and hence $\widetilde{\beta} \in \Lambda$. Throughout the rest of the proof, we assume that the constraint (B.30) holds.

Next, we apply Lemma Appendix A.1 to bound the left-hand side of (B.26) from below. As in the proof of Theorem 2.1, we set $r_{\mathrm{loc}} = \kappa/(4\eta_{0.25})$ and define $\widetilde{\beta}_c = (1-c)\beta^* + c\widetilde{\beta}$, where $c = \sup\{u \in [0,1] : (1-u)\beta^* + u\widetilde{\beta} \in \Theta(r_{\mathrm{loc}})\}$. The same argument therein implies $\overline{D}_{\widetilde{\mathcal{L}}}(\widetilde{\beta}_c, \beta^*) \le c\overline{D}_{\widetilde{\mathcal{L}}}(\widetilde{\beta}, \beta^*)$. Recall that conditioned on $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(\lambda_*)$, $\widetilde{\beta}$ falls in the ℓ_1-cone Λ and thus so does $\widetilde{\beta}_c$. Moreover, $\widetilde{\beta}_c \in \Theta(r_{\mathrm{loc}})$ by construction. Then it follows from Lemma Appendix A.1 that, with probability at least $1 - e^{-u}$,

$$\overline{D}_{\widetilde{\mathcal{L}}}(\widetilde{\beta}_c, \beta^*) \ge \frac{1}{4}\|\widetilde{\beta}_c - \beta^*\|_\Sigma^2,$$

as long as $n \gtrsim s\log p + u$. Combining this with (B.26), (B.29) and (B.30), we obtain that

$$\frac{1}{4}\|\widetilde{\beta}_c - \beta^*\|_\Sigma^2 \le c\lambda\big(1.4\|h_{\mathcal{S}}\|_1 + 0.4s^{1/2}\|h\|_\Sigma\big) \le 1.8s^{1/2}\lambda\|\widetilde{\beta}_c - \beta^*\|_\Sigma.$$

Canceling $\|\widetilde{\beta}_c - \beta^*\|_\Sigma$ on both sides yields

$$\|\widetilde{\beta}_c - \beta^*\|_\Sigma \le 7.2s^{1/2}\lambda. \tag{B.31}$$

Provided that $\kappa > 28.8\eta_{0.25}s^{1/2}\lambda$, the right-hand side is strictly less than r_{loc}. Via proof by contradiction, we must have $\widetilde{\beta} = \widetilde{\beta}_c \in \Theta(r_{\mathrm{loc}})$, and hence the bound (B.31) also applies to $\widetilde{\beta}$.

It remains to choose ρ properly so that the constraint (B.30) holds with high probability. Recall from Lemma Appendix A.2 that $b^* \le \sigma^2\tau^{-1}$. The following two lemmas provide upper bounds on the suprema $\Delta(r_0)$, $\Delta_1(r_0)$ and $\delta(r_0)$ defined in (B.27) and (B.28).

Lemma Appendix B.1. *Assume Condition (C2) holds. Then, for any $r, u > 0$,*

$$\Delta(r) \le C_1 B^2 r\left\{\sqrt{\frac{s\log(2p)}{N}} + s^{1/2}\frac{\log(2p)+u}{N}\right\} + C_2(\sigma_u\mu_4)^{1/2} r\sqrt{\frac{\log(2p)+u}{N}}, \tag{B.32}$$

with probability at least $1 - e^{-u}$, where $C_1, C_2 > 0$ are absolute constants. The same bound, with N replaced by n, holds for $\Delta_1(r)$.

Lemma Appendix B.2. *Condition (C2) guarantees* $\delta(r) \le \kappa^{-2} r(\sigma^2 + \mu_4 r^2/3)$ *for any* $r > 0$.

Let $0 < r_0 \lesssim \sigma$ and set $\delta = 2e^{-u}$, so that $\log p + u \asymp \log(p/\delta)$. Suppose the sample size per machine satisfies $n \gtrsim s\log(p/\delta)$. Then, in view of Lemmas Appendix B.1 and Appendix B.2, a sufficiently large ρ, which is of order

$$\rho \asymp \max\left\{ r_0\sqrt{\frac{s\log(p/\delta)}{n}}, s^{-1/2}\sigma^2(\kappa^{-2}r_0 + \tau^{-1}) \right\},$$

guarantees that (B.30) holds with probability at least $1 - \delta/2$. With this choice of ρ, we see that the right-hand of (B.31) is strictly less than r_{loc} as long as $\kappa \gtrsim s^{1/2}\{\lambda^* + r_0\sqrt{s\log(p/\delta)/n}\} + \sigma^2(\kappa^{-2}r_0 + \tau^{-1})$. Since $\kappa \asymp \sigma\sqrt{n/\log(p/\delta)}$ and $\sigma^2\kappa^{-2}r_0$ is negligible compared to $r_0\sqrt{s\log(p/\delta)/n}$, this holds trivially under the assumed sample size scaling, and thus completes the proof. □

We end this subsection with the proofs of Lemmas Appendix B.1 and Appendix B.2.

B.5.1. Proof of Lemma Appendix B.1

For any $r_1, r_2 > 0$, define the ℓ_1/ℓ_2-ball $\mathbb{B}(r_1, r_2) = \{\beta \in \mathbb{R}^p : \|\beta\|_1 \le r_1, \|\beta\|_2 \le r_2\}$. Consider the change of variable $v = \beta - \beta^*$, so that $v \in \mathbb{B}(4s^{1/2}r, r)$ for $\beta \in \Theta(r) \cap \Lambda$. It follows that

$$\begin{aligned}
&\sup_{\beta\in\Theta(r)\cap\Lambda} \|\widehat{D}(\beta) - D(\beta)\|_\infty \\
&\le \max_{1\le j\le p} \sup_{v\in\mathbb{B}(4s^{1/2}r,r)} \left| \frac{1}{N}\sum_{i=1}^N (1-\mathbb{E}) \underbrace{\{\psi_\tau(\varepsilon_i - x_i^{\mathrm{T}}v) - \psi_\tau(\varepsilon_i)\}x_{ij}}_{=:\phi_{ij}(v)} \right| = \max_{1\le j\le p} \Phi_j,
\end{aligned} \tag{B.33}$$

where $\Phi_j := \sup_{v\in\mathbb{B}(4s^{1/2}r,r)} |(1/N)\sum_{i=1}^N (1-\mathbb{E})\phi_{ij}(v)|$ and $\psi_\tau(u) = \mathrm{sign}(u)\min(|u|, \tau)$. By the Lipschitz continuity of $\psi_\tau(\cdot)$, $\sup_{v\in\mathbb{B}(4s^{1/2}r,r)} |\phi_{ij}(v)| \le \sup_{v\in\mathbb{B}(4s^{1/2}r,r)} |x_i^{\mathrm{T}}v| \cdot |x_{ij}| \le 4B^2 s^{1/2} r$ and, for each $v \in \mathbb{B}(4s^{1/2}r, r)$,

$$\mathbb{E}\phi_{ij}^2(v) \le \mathbb{E}\{x_{ij}^2 (x_i^{\mathrm{T}}v)^2\} \le \left(\mathbb{E}x_{ij}^4\right)^{1/2} \left\{\mathbb{E}(x_i^{\mathrm{T}}v)^4\right\}^{1/2} \le \sigma_{jj}\mu_4 \cdot r^2.$$

We then apply Bousquet's version of Talagrand's inequality (Bousquet, 2003) and obtain that, for any $z > 0$,

$$\begin{aligned}
\Phi_j &\le \mathbb{E}\Phi_j + \sup_{v\in\mathbb{B}(4s^{1/2}r,r)} \left\{\mathbb{E}\phi_{ij}^2(v)\right\}^{1/2} \sqrt{\frac{2z}{N}} + 4\sqrt{\mathbb{E}\Phi_j \cdot B^2 s^{1/2} r \frac{z}{N}} + (4/3)B^2 s^{1/2} r \frac{z}{N} \\
&\le \mathbb{E}\Phi_j + (2\sigma_{jj}\mu_4)^{1/2} r\sqrt{\frac{z}{N}} + 4\sqrt{\mathbb{E}\Phi_j \cdot B^2 s^{1/2} r \frac{z}{N}} + (4/3)B^2 s^{1/2} r \frac{z}{N},
\end{aligned} \tag{B.34}$$

with probability at least $1 - 2e^{-z}$. For the expected value $\mathbb{E}\Phi_j$, by Rademacher symmetrization we have

$$\mathbb{E}\Phi_j \le 2\mathbb{E} \sup_{v\in\mathbb{B}(4s^{1/2}r,r)} \left| \frac{1}{N}\sum_{i=1}^N e_i\phi_{ij}(v) \right| = 2\mathbb{E}\left\{ \mathbb{E}_e \sup_{v\in\mathbb{B}(4s^{1/2}r,r)} \left| \frac{1}{N}\sum_{i=1}^N e_i\phi_{ij}(v) \right| \right\},$$

where $e_1, \ldots, e_N$ are independent Rademacher random variables. For each i, write $\phi_{ij}(v) = \phi_j(x_i^{\mathrm{T}}v)$, where $\phi_j(\cdot)$ is such that $\phi_j(0) = 0$ and $|\phi_j(t_1) - \phi_j(t_2)| \le |x_{ij}| \cdot |t_1 - t_2| \le B|t_1 - t_2|$. It thus follows from Talagrand's contraction principle that

$$\mathbb{E}_e \sup_{v\in\mathbb{B}(4s^{1/2}r,r)} \left| \frac{1}{N}\sum_{i=1}^N e_i\phi_{ij}(v) \right| \le 2B \cdot \mathbb{E}_e \sup_{v\in\mathbb{B}(4s^{1/2}r,r)} \left| \frac{1}{N}\sum_{i=1}^N e_i x_i^{\mathrm{T}}v \right| \le 8Bs^{1/2}r \cdot \mathbb{E}_e \left\| \frac{1}{N}\sum_{i=1}^N e_i x_i \right\|_\infty.$$

Again, applying Lemma 14.14 in Bühlmann and van de Geer (2011) yields $\mathbb{E}_e\|(1/N)\sum_{i=1}^N e_i x_i\|_\infty \le B\sqrt{2\log(2p)/N}$. Putting together the pieces, we conclude that, for $j = 1, \ldots, p$,

$$\mathbb{E}\Phi_j \le 16B^2 r\sqrt{\frac{2s\log(2p)}{N}}.$$

Finally, taking $z = \log(2p) + u$ in (B.34), the claimed bound follows from the union bound. □

B.5.2. Proof of Lemma Appendix B.2

Let $\mathcal{L}_\tau(\beta) = \mathbb{E}\widehat{\mathcal{L}}_\tau(\beta)$ be the population loss, so that

$$D_1(\beta) = \nabla\mathcal{L}_\kappa(\beta) - \nabla\mathcal{L}_\kappa(\beta^*) \quad \text{and} \quad D(\beta) = \nabla\mathcal{L}_\tau(\beta) - \nabla\mathcal{L}_\tau(\beta^*).$$

Starting with $D_1(\beta)$, consider the change of variable $v = \Sigma^{1/2}(\beta - \beta^*)$. Then, by the mean value theorem for vector-valued functions,

$$\begin{aligned}
&\Sigma^{-1/2}D_1(\beta) - \Sigma^{1/2}(\beta - \beta^*) \\
&= \Sigma^{-1/2}\int_0^1 \nabla^2\mathcal{L}_\kappa\big((1-t)\beta^* + t\beta\big)\mathrm{d}t\,\Sigma^{-1/2}\cdot v - v \\
&= -\int_0^1 \mathbb{E}\big\{\mathbb{P}\big(|\varepsilon - tz^{\mathrm{T}}v| > \kappa\,|x\big)zz^{\mathrm{T}}\big\}\mathrm{d}t\cdot v.
\end{aligned}$$

Similarly, it can be obtained that

$$\Sigma^{-1/2}D(\beta) - \Sigma^{1/2}(\beta - \beta^*) = -\int_0^1 \mathbb{E}\big\{\mathbb{P}\big(|\varepsilon - tz^{\mathrm{T}}v| > \tau\,|x\big)zz^{\mathrm{T}}\big\}\mathrm{d}t\cdot v.$$

Recall that $\tau \geq \kappa > 0$. We have

$$\Sigma^{-1/2}\{D_1(\beta) - D(\beta)\} = -\int_0^1 \mathbb{E}\big\{\mathbb{P}\big(\kappa < |\varepsilon - tz^{\mathrm{T}}v| \leq \tau\,|x\big)zz^{\mathrm{T}}\big\}\mathrm{d}t\cdot v.$$

By Markov's inequality and the fact that $\mathbb{E}(\varepsilon|x) = 0$, $\mathbb{P}(|\varepsilon - tz^{\mathrm{T}}v| > \kappa\,|x) \leq \kappa^{-2}\{\mathbb{E}(\varepsilon^2|x) + t^2(z^{\mathrm{T}}v)^2\} \leq \kappa^{-2}\{\sigma^2 + t^2(z^{\mathrm{T}}v)^2\}$. Substituting this into the above bound yields

$$\begin{aligned}
\sup_{\beta\in\Theta(r)} \|D_1(\beta) - D(\beta)\|_\Omega &\leq r\bigg\|\int_0^1 \kappa^{-2}\big[\sigma^2 + t^2\mathbb{E}\{(z^{\mathrm{T}}v)^2zz^{\mathrm{T}}\}\big]dt\bigg\|_2 \\
&\leq \kappa^{-2}r\big[\sigma^2 + \tfrac{1}{3}\|\mathbb{E}\{(z^{\mathrm{T}}v)^2zz^{\mathrm{T}}\}\|_2\big] \\
&\leq \kappa^{-2}r\big(\sigma^2 + \mu_4 r^2/3\big),
\end{aligned}$$

as desired. □

B.6. Proof of Theorem 3.2

The proof will be carried out conditioning on the "good event" $\mathcal{E}_0(r_0)\cap\mathcal{E}_*(\lambda_*)$ for some predetermined $0 < r_0, \lambda_* \lesssim \sigma$. Given $\delta \in (0,1)$, let the robustification parameters satisfy $\tau \geq \kappa \asymp \sigma\sqrt{n/\log(p/\delta)}$. Theorem 3.1 implies that the first iterate $\widetilde{\beta}^{(1)} \in \operatorname{argmin}_{\beta\in\mathbb{R}^p}\{\widetilde{\mathcal{L}}^{(1)}(\beta) + \lambda_1\|\beta\|_1\}$ with

$$\lambda_1 = 2.5(\lambda_* + \rho_1) \quad \text{and} \quad \rho_1 \asymp \max\left\{r_0\sqrt{\frac{s\log(p/\delta)}{n}}, s^{-1/2}\sigma^2\tau^{-1}\right\},$$

satisfies the cone property $\widetilde{\beta}^{(1)} \in \Lambda$ and the error bound

$$\|\widetilde{\beta}^{(1)} - \beta^*\|_\Sigma \leq C_1 s\sqrt{\log(p/\delta)/n}\cdot r_0 + C_2(\sigma^2\tau^{-1} + s^{1/2}\lambda_*) =: r_1, \tag{B.35}$$

with probability at least $1-\delta$. In (B.35), we set $\alpha = \alpha(s,p,n,\delta) = C_1 s\sqrt{\log(p/\delta)/n}$ and $r_* = C_2(\sigma^2\tau^{-1} + s^{1/2}\lambda_*)$, so that $r_1 = \alpha r_0 + r_*$. Provided the sample size per machine is sufficiently large, namely, $n \gtrsim s^2\log(p/\delta)$, the contraction factor α is strictly less than 1, and hence the initial estimation error r_0 is reduced by a factor of α after one round of communication.

For $t = 2, 3, \ldots, T$, define the events $\mathcal{E}_t(r_t) = \{\widetilde{\beta}^{(t)} \in \Theta(r_t)\cap\Lambda\}$ and radius parameters

$$r_t = \alpha r_{t-1} + r_* = \alpha^2 r_{t-2} + (1+\alpha)r_* = \cdots = \alpha^t r_0 + \frac{1-\alpha^t}{1-\alpha}r_*.$$

In the t-th iteration, we choose the regularization parameter $\lambda_t = 2.5(\lambda_* + \rho_t)$ with

$$\rho_t \asymp \max\left\{ r_{t-1}\sqrt{\frac{s\log(p/\delta)}{n}}, s^{-1/2}\sigma^2\tau^{-1} \right\} \asymp s^{-1/2}\max\{\alpha^t r_0, \sigma^2\tau^{-1}\}.$$

Commenced with $\widetilde{\beta}^{(t-1)}$ at iteration $t \geq 2$, we apply Theorem 3.1 to obtain that conditioned on event $\mathcal{E}_{t-1}(r_{t-1}) \cap \mathcal{E}_*(\lambda_*)$,

$$\widetilde{\beta}^{(t)} \in \Lambda \quad \text{and} \quad \|\widetilde{\beta}^{(t)} - \beta^*\|_\Sigma \leq \alpha r_{t-1} + r_* = r_t, \tag{B.36}$$

with probability at least $1-\delta$. In other words, event $\mathcal{E}_t(r_t)$ occurs with probability at least $1-\delta$ conditioned on $\mathcal{E}_{t-1}(r_{t-1}) \cap \mathcal{E}_*(\lambda_*)$.

Finally, we choose $T = \lceil \log(r_0/r_*)/\log(1/\alpha) \rceil$ so that $\alpha^T r_0 \leq r_*$. Then, applying (B.35), (B.36) and the union bound over $t = 1, \ldots, T$ yields that, conditioned on $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*)$, the T-th iterate $\widetilde{\beta}^{(T)}$ falls into the cone Λ and satisfies the error bound

$$\|\widetilde{\beta}^{(T)} - \beta^*\|_\Sigma \leq r_T \asymp r_*,$$

with probability at least $1 - T\delta$. This completes the proof of the theorem. □

References

Barzilai, J., Borwein, J.M., 1988. Two-point step size gradient methods. IMA J. Numer. Anal. 8, 141–148.

Battey, H., Fan, J., Liu, H., Lu, J., Zhu, Z., 2018. Distributed testing and estimation under sparse high dimensional models. Ann. Stat. 46, 1352–1382.

Bickel, P.J., 1975. One-step Huber estimates in the linear model. J. Am. Stat. Assoc. 70, 428–434.

Bousquet, O., 2003. Concentration inequalities for sub-additive functions using the entropy method. In: Stochastic Inequalities and Applications. In: Progress in Probability, vol. 56. Birkhäuser, Basel, pp. 213–247.

Bühlmann, P., 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. Ann. Inst. Henri Poincaré Probab. Stat. 48, 1148–1185.

Chen, L.H.Y., Shao, Q.-M., 2001. A non-uniform Berry-Esseen bound via Stein's method. Probab. Theory Relat. Fields 120, 236–254.

Chen, X., Liu, W., Zhang, Y., 2019. Quantile regression under memory constraint. Ann. Stat. 47, 3244–3273.

Chen, X., Xie, M.G., 2012. A split-and-conquer approach for analysis of extraordinarily large data. Stat. Sin. 24, 1655–1684.

Chinot, G., Lecué, G., Lerasle, M., 2020. Robust statistical learning with Lipschitz and convex loss functions. Probab. Theory Relat. Fields 45, 866–896.

Dobriban, E., Sheng, Y., 2021. Distributed linear regression by averaging. Ann. Stat. 49, 918–943.

Fan, J., Li, Q., Wang, Y., 2017. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. J. R. Stat. Soc., Ser. B, Stat. Methodol. 79, 247–265.

Fan, J., Liu, H., Sun, Q., Zhang, T., 2018. I-LAMM for sparse learning: simultaneous control of algorithmic complexity and statistical error. Ann. Stat. 46, 814–841.

He, X., Shao, Q.-M., 1996. A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. Ann. Stat. 24, 2608–2630.

He, X., Shao, Q.-M., 2000. On parameters of increasing dimensions. J. Multivar. Anal. 73, 120–135.

Huber, P., 1973. Robust regression: asymptotics, conjectures and Monte Carlo. Ann. Stat. 1, 799–821.

Huber, P.J., Ronchetti, E.M., 2009. Robust Statistics, 2nd ed. Wiley, New York.

Hunter, D.R., Lange, K., 2000. Quantile regression via an MM algorithm. J. Comput. Graph. Stat. 9, 60–77.

Huo, X., Cao, S., 2018. Aggregated inference. Wiley Interdiscip. Rev.: Comput. Stat. 11 (1), e1451.

Jordan, M.I., Lee, J.D., Yang, Y., 2019. Communication-efficient distributed statistical inference. J. Am. Stat. Assoc. 114, 668–681.

Lambert-Lacroix, S., Zwald, L., 2011. Robust regression through the Huber's criterion and adaptive lasso penalty. Electron. J. Stat. 5, 1015–1053.

Lange, K., Hunter, D.R., Yang, I., 2000. Optimization transfer using surrogate objective functions. J. Comput. Graph. Stat. 9, 1–20.

Lee, J.D., Liu, Q., Sun, Y., Taylor, J.E., 2017. Communication-efficient sparse regression. J. Mach. Learn. Res. 18 (5), 1–30.

Li, R., Lin, D.K., Li, B., 2013. Statistical inference in massive data sets. Appl. Stoch. Model Bus. 29, 399–409.

Loh, P., 2017. Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. Ann. Stat. 45, 866–896.

Mammen, E., 1989. Asymptotics with increasing dimension for robust regression with applications to the bootstrap. Ann. Stat. 17, 382–400.

Portnoy, S., 1985. Asymptotic behavior of M estimators of p regression parameters when p^2/n is large; II. Norm. Approx. Ann. Statist. 13, 1403–1417.

Rosenblatt, J.D., Nadler, B., 2016. On the optimality of averaging in distributed statistical learning. Inf. Inference 5 (4), 379–404.

Shamir, O., Srebro, N., Zhang, T., 2014. Communication efficient distributed optimization using an approximate Newton-type method. In: Proceedings of the 31st International Conference on Machine Learning, vol. 32, pp. 1000–1008.

Sidransky, E., Nalls, M.A., Aasly, J.O., Aharon-Peretz, J., Annesi, G., Barbosa, E.R., Bar-Shira, A., Berg, D., Bras, J., Brice, A., 2009. Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. N. Engl. J. Med. 361, 1651–1661.

Singh, S.K., Maddala, G.S., 1976. A function for size distribution of incomes. Econometrica 44, 963–970.

Sun, Q., Zhou, W.-X., Fan, J., 2020. Adaptive Huber regression. J. Am. Stat. Assoc. 115, 254–265.

van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Stat. 42, 1166–1202.

Vershynin, R., 2012. Introduction to the non-asymptotic analysis of random matrices. In: Eldar, Y., Kutyniok, G. (Eds.), Compressed Sensing. Cambridge Univ. Press, Cambridge, pp. 210–268.

Volgushev, S., Chao, S.-K., Cheng, G., 2019. Distributed inference for quantile regression processes. Ann. Stat. 47, 1634–1662.

Wang, J., Kolar, M., Srebro, N., Zhang, T., 2017. Efficient distributed learning with sparsity. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 3636–3645.

Wang, L., Peng, B., Li, R., 2015. A high-dimensional nonparametric multivariate test for mean vector. J. Am. Stat. Assoc. 110, 1658–1669.

Wang, L., Zheng, C., Zhou, W., Zhou, W.-X., 2021. A new principle for tuning-free Huber regression. Stat. Sin. 31, 2153–2177.

Wang, X., Yang, Z., Chen, X., Liu, W., 2019. Distributed inference for linear support vector machine. J. Mach. Learn. Res. 20 (113), 1–41.

Western, B., 1995. Concepts and suggestions for robust regression analysis. Am. J. Polit. Sci. 39, 786–817.

Wu, Y., Jiang, X., Kim, J., Ohno-Machado, L., 2012. Grid binary LOgistic REgression (GLORE): building shared models without sharing data. J. Am. Med. Inform. Assoc. 19, 758–764.

Yohai, V.J., Maronna, R.A., 1979. Asymptotic behavior of M-estimators for the linear model. Ann. Stat. 7, 258–268.

Zhang, C.-H., Zhang, S.-S., 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. J. R. Stat. Soc., Ser. B, Stat. Methodol. 76, 217–242.
Zhang, Y., Duchi, J., Wainwright, M., 2015. Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. J. Mach. Learn. Res. 16 (102), 3299–3340.
Zhao, T., Cheng, G., Liu, H., 2016. A partially linear framework for massive heterogeneous data. Ann. Stat. 44, 1400–1437.
Zhou, W.-X., Bose, K., Fan, J., Liu, H., 2018. A new perspective on robust *M*-estimation: finite sample theory and applications to dependence-adjusted multiple testing. Ann. Stat. 46, 1904–1931.