Journal of Machine Learning Research 22 (2021) 1-67 Submitted 8/20; Revised 1/21; Published 3/21

Variance Reduced Median-of-Means Estimator for Byzantine-Robust Distributed Inference

Jiyuan Tu TUJY.19@GMAIL.COM
School of Mathematical Sciences
Shanghai Jiao Tong University, Shanghai, 200240, China

Weidong Liu WEIDONGL@SJTU.EDU.CN
School of Mathematical Sciences, School of Life Sciences and Biotechnology
MoE Key Lab of Artificial Intelligence
Shanghai Jiao Tong University, Shanghai, 200240, China

Xiaojun Mao MAOXJ@FUDAN.EDU.CN
School of Data Science
Fudan University, Shanghai, 200433, China

Xi Chen XC13@STERN.NYU.EDU
Stern School of Business
New York University, New York, NY 10012, USA

Editor: Qiang Liu

Abstract

This paper develops an efficient distributed inference algorithm, which is robust against a moderate fraction of Byzantine nodes, namely arbitrary and possibly adversarial machines in a distributed learning system. In robust statistics, the median-of-means (MOM) has been a popular approach to hedge against Byzantine failures due to its ease of implementation and computational efficiency. However, the MOM estimator has the shortcoming in terms of statistical efficiency. The first main contribution of the paper is to propose a variance reduced median-of-means (VRMOM) estimator, which improves the statistical efficiency over the vanilla MOM estimator and is computationally as efficient as the MOM. Based on the proposed VRMOM estimator, we develop a general distributed inference algorithm that is robust against Byzantine failures. Theoretically, our distributed algorithm achieves a fast convergence rate with only a constant number of rounds of communications. We also provide the asymptotic normality result for the purpose of statistical inference. To the best of our knowledge, this is the first normality result in the setting of Byzantine-robust distributed learning. The simulation results are also presented to illustrate the effectiveness of our method.

Keywords: Byzantine robustness, distributed inference, median-of-means, statistical efficiency

Weidong Liu and Xiaojun Mao are the co-corresponding authors.

1. Introduction

Due to the rapid increase of the scale of data, modern data set are usually too large to fit in a single device, and thus have to be stored and processed in a distributed manner. In a common distributed computing environment, data are stored across multiple machines/nodes. A single master node is in charge of maintaining and updating target parameters, and a large number of worker machines perform local computations and communicate the computed information with the master node (see Figure 2 in Section 3 for an illustration). As compared to the traditional single machine setting, where the entire data can be loaded into the memory for the centralized computation, the distributed setting poses two major challenges.

The first challenge comes from the tradeoff between communication cost and statistical accuracy. For example, one-shot communication (e.g., taking average of local estimators), though incurs low communication cost, has a poor performance for nonlinear estimation when the number of machines is large (see, e.g., Li et al. (2013); Zhang et al. (2013, 2015); Zhao et al. (2016); Rosenblatt and Nadler (2016); Shang and Cheng (2017); Lee et al. (2017)). Therefore, iterative approaches are adopted in literature (see, e.g., Shamir et al. (2014); Jordan et al. (2019); Chen et al. (2019); Fan et al. (2019); Wang et al. (2019); Chen et al. (2020b)). For iterative algorithms, since each iteration of communication requires synchronization, a communicationally efficient algorithm should run with a small number of iterations. Our goal is to develop algorithms that achieve communication efficiency without losing statistical accuracy.

The second challenge comes from the vulnerability of worker machines and communication channels. In particular, the information sent from a worker machine can be arbitrarily erroneous due to hardware or software breakdowns, data crashes, or communication failures. Such an error is usually referred to as Byzantine failures (Lamport et al., 1982). In other words, a subset of workers called Byzantine machines, may send arbitrary and even adversarial messages to the master. Distributed learning under the Byzantine setting has attracted a lot of research attentions in recent years (see, e.g., Feng et al. (2014); Chen et al. (2017); Blanchard et al. (2017); Xie et al. (2018); Alistarh et al. (2018); Yin et al. (2018, 2019); Su and Xu (2019)). However, as we will survey later, some of these methods suffer from a larger number of iterations of communications and existing analysis only focuses on the convergence rate. The statistical inference with Byzantine failures, which plays an important role in uncertainty quantification, is still largely open.

The goal of this paper is to propose a communication-efficient statistical inference method, which is robustly against Byzantine failures. We consider a general risk minimization problem,

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\operatorname{argmin}} \mathbb{E}_{X\sim\mathfrak{X}} \{f(X, \boldsymbol{\theta})\}, \tag{1}$$

where f is the convex loss function, X denotes the random sample from a probability distribution $\mathfrak{X}$, and $\boldsymbol{\theta}^* \in \mathbb{R}^p$ is the target parameter vector of interest. To infer the underlying parameter $\boldsymbol{\theta}^*$, assume that N i.i.d. observations $\{X_1, ..., X_N\}$ are collected and evenly distributed over $(m+1)$-machines $\{\mathcal{H}_0, \ldots \mathcal{H}_m\}$, where each machine contains n observations. We allow diverging N, n, and p under certain rate constraints.

In this paper, we consider the Byzantine distributed framework, which allows for Byzantine failures described as follows. In particular, we assume there exists an α_n fraction of

worker machines (a.k.a. Byzantine machines), whose indices form a subset $\mathcal{B} \subseteq \{1, \ldots, m\}$ with $Card(\mathcal{B}) = \lfloor \alpha_n m \rfloor$. The Byzantine machines are subject to the following Byzantine failures when communicating information:

Definition 1 (Byzantine Failures) *In each round of communication, assume the information produced by each machine is $\boldsymbol{v}_j$, then the actual information $\bar{\boldsymbol{v}}_j$ received from each worker machine $\mathcal{H}_j$ is as follows*

$$\bar{\boldsymbol{v}}_j = \begin{cases} \boldsymbol{v}_j & j \notin \mathcal{B}, \\ * & j \in \mathcal{B}, \end{cases}$$

where $*$ denotes an arbitrary value.

Let us start with the most fundamental setting where $\boldsymbol{\theta}^*$ is the population mean and the goal is to infer the population mean with the presence of Byzantine failures. A widely used robust estimator is the median-of-means (MOM) estimator (Nemirovsky and Yudin, 1983; Jerrum et al., 1986; Alon et al., 1999), which first computes the local sample mean on each machine and then aggregate them by taking the median. Due its ease of implementation and computational efficiency, the MOM estimator has attracted a lot of attentions (Minsker, 2015; Hsu and Sabato, 2016; Lecué and Lerasle, 2020; Lugosi and Mendelson, 2019; Minsker, 2019) and served as an important building block in distributed learning with Byzantine failures (Yin et al. (2018)). However, despite its popularity, the MOM estimator suffers from low asymptotic statistical efficiency. More precisely, the asymptotic efficiency of the MOM estimator is only $2/\pi \approx 0.637$, which is far from 1 for normal mean problem.

The main contribution of the paper is to propose a computationally efficient robust mean estimator, which greatly improves the statistical efficiency of the MOM. Our estimator is called *variance reduced median-of-means (VRMOM)* estimator. Instead of using the median in MOM, we use multiple quantile levels to improve the statistical efficiency. By formulating a carefully designed stochastic optimization problem and leveraging the idea of one-step Newton iteration, our VRMOM estimator achieves the same order of computational complexity as the MOM estimator, but improves the asymptotic efficiency from $2/\pi \approx 0.637$ of MOM to $3/\pi \approx 0.955$ (see Theorem 5). The proposed VRMOM estimator naturally serves as a more efficient substitute of MOM in all robust statistical applications that benefit from the MOM estimator.

As an application of our VRMOM estimator, we describe a communication-efficient algorithm for the general risk minimization problem in (1) based on the VRMOM estimator. In a standard distributed gradient descent (GD) approach, each local machine computes the gradient information, which takes the form of the *mean of gradients of each local data point*. Then, the master receives the transmitted gradient information and aggregates the local gradients by taking the average. However, the averaged gradient is highly sensitive to Byzantine failure, whose value can be completely skewed by a single Byzantine worker. To hedge against Byzantine failures, the work by Yin et al. (2018) proposed to take the coordinate-wise *median* of the transmitted gradients, which is essentially an MOM estimator based on gradients of local data. Our method improves this result from two aspects. First, instead of using the median, our VRMOM serves as a new *gradient aggregator*, which is statistically more efficient in a large class of distributed robust inference problems. Second,

the distributed gradient method would take a large number of iterations (i.e., $O(\log(N/p))$) to converge, which is communicationally expensive. To address this issue, we leverage the surrogate loss function in the Communication-efficient Surrogate Likelihood (CSL) framework (Jordan et al., 2019) and develop the *robust CSL (RCSL)* method. In a wide range of choices of N, m, and p, our RCSL only requires a constant order of iterations to achieve a fast convergence rate, which greatly saves the total communication cost. Theoretically, we establish the convergence rates of our estimator (see Theorem 11 and Theorem 19 in Appendix C.1.2) and provide the asymptotic normality result (see Theorems 8).

1.1 Contributions and Related Works

The median-of-means (MOM), which was introduced by Nemirovsky and Yudin (1983), has been a popular estimator in robust statistics due to its ease of implementation and convergence guarantees (Minsker, 2015; Hsu and Sabato, 2016; Lecué and Lerasle, 2020; Lugosi and Mendelson, 2019; Minsker, 2019). The MOM estimator finds a wide range of applications, including robust PCA (Minsker, 2015), linear regression (Hsu and Sabato, 2016), sparse linear regression (Minsker, 2015; Lecué and Lerasle, 2020), robust empirical risk minimization (Lecué and Lerasle, 2020; Lugosi and Mendelson, 2019; Minsker, 2019). This paper improves the MOM estimator by proposing a variance reduction scheme, which significantly boosts the statistical efficiency. Our VRMOM estimator is motivated by the idea that composite quantiles can improve the efficiency (Zou and Yuan, 2008). However, directly taking multiple sample quantiles would incur a higher computational cost and our VRMOM is carefully designed to be computationally efficient and admit a simple closed-form for the ease of theoretical analysis. The proposed VRMOM estimator can be a natural substitute for the classical MOM estimator for all aforementioned applications.

In recent years, statistical learning and optimization with the presence of Byzantine failures have attracted a lot of attentions (Feng et al., 2014; Chen et al., 2017; Blanchard et al., 2017; Xie et al., 2018; Alistarh et al., 2018; Yin et al., 2018, 2019; Su and Xu, 2019). The key idea behind these work is to let each worker machine compute the gradient (or stochastic gradient) information, and the gradients from workers are aggregated using some robust mean estimators instead of the vanilla gradient mean. There are many applicable estimators like median, trimmed mean (Yin et al., 2018, 2019), geometric median (Feng et al., 2014; Chen et al., 2017), Krum (Blanchard et al., 2017), marginal median, mean-around-median (Xie et al., 2018), and iterative filtering (Su and Xu, 2019; Yin et al., 2019). However, most existing methods are only based on gradient information, without utilizing any second order properties. In this paper, we propose the robust CSL method, which combines the new gradient aggregator — VRMOM estimator, and the approximate-Newton framework (See, e.g. Shamir et al. (2014); Jordan et al. (2019), and Fan et al. (2019)). The combination of the VRMOM and approximate-Newton greatly facilitates communication efficient estimation by reducing the total number of communication rounds. From a theoretical perspective, we establish the asymptotic normality result, which has not been well explored in previous robust distributed learning literature. A more detailed comparison of the convergence rates with the existing approaches is presented after Theorem 11, after the formal description of our convergence result.

1.2 Paper Organization and Notations

The rest of the paper is organized as follows. Section 2 describes the proposed VRMOM estimator and its theoretical results. In Section 3, we introduce the Robust CSL (RCSL) method for Byzantine robust machine learning problem as an important application of the VRMOM estimator. Simulation experiments are provided in Section 4, which demonstrate the superiority of our method over some existing methods. Finally, we conclude our work in Section 5. The proofs of the theories of the VRMOM estimator and the theories of the RCSL method are relegated to Appendices.

For every vector $\boldsymbol{v} = (v_1, ..., v_p)^{\mathrm{T}}$, denote $|\boldsymbol{v}|_2 = \sqrt{\sum_{l=1}^{p} v_l^2}$. For every matrix $\boldsymbol{A}$, define $\|\boldsymbol{A}\| = \sup_{|\boldsymbol{v}|_2=1} |\boldsymbol{A}\boldsymbol{v}|_2$ as the operator norm, $\Lambda_{\max}(\boldsymbol{A})$ and $\Lambda_{\min}(\boldsymbol{A})$ as the largest and smallest eigenvalues of $\boldsymbol{A}$ respectively. Suppose there is another matrix $\boldsymbol{B}$, and we denote $\boldsymbol{B} \preceq \boldsymbol{A}$ if and only if $\boldsymbol{A} - \boldsymbol{B}$ is positive definite. Let $\mathcal{N}(0,1)$ be the standard normal distribution. We denote $\Phi(x) = \mathbb{P}(\mathcal{N}(0,1) \leq x)$ and $\psi(x) = e^{-x^2/2}/\sqrt{2\pi}$ to be its cumulative distribution function and probability density function, respectively. Denote $\mathbb{S}^{p-1}(\boldsymbol{\theta})$ and $\mathbb{B}^p(\boldsymbol{\theta})$ as the unit sphere and the unit ball centered at $\boldsymbol{\theta} \in \mathbb{R}^p$ respectively. For simplicity, we denote $\mathbb{S}^{p-1}$ and $\mathbb{B}^p$ as unit sphere and unit ball centered at $\mathbf{0}$. We will use $\mathbb{I}(\cdot)$ as the indicator function. The symbols $\lfloor x \rfloor$ ($\lceil x \rceil$) denotes the greatest integer (the smallest integer) not larger than (not less than) x. Summation symbol will be heavily used throughout this article. For the convenience of reading, in each summand, we will use the subscripts $i(1 \leq i \leq n)$ for each data point, $j(0 \leq j \leq m)$ for each machine, $k(1 \leq k \leq K)$ for each quantile level and $l(1 \leq l \leq p)$ for each entry of a vector, respectively. Lastly, the generic constants are assumed to be independent of m, n, and p.

2. Proposed Methods

In this section, we will firstly introduce the construction of our VRMOM estimator. Then we provide theoretical guarantees for it.

2.1 Variance Reduced Median-of-Means Estimator

To motivate our estimator, let us provide a brief review of the standard MOM estimator. Let $X_1, ..., X_N$ be *i.i.d.* copies of X with $\mathbb{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$. For the ease of presentation, we assume that N observations are evenly partitioned into $(m+1)$-batches $\{\mathcal{H}_0, \ldots \mathcal{H}_m\}$, where each $\mathcal{H}_j$ denotes the indices of the samples within the j-th batch. Let $n = N/(m+1)$ be the sample size of each batch and $\overline{X}_j = \sum_{i \in \mathcal{H}_j} X_i/n$ be the sample mean of the observations in the j-th batch. To estimate the population mean μ, the MOM estimator is defined as

$$\widehat{\mu} = \mathrm{med}(\overline{X}_0, ..., \overline{X}_m), \tag{2}$$

where $\mathrm{med}(\cdot)$ denotes the sample median. The MOM estimator is computationally efficient and robust against Byzantine failures. Moreover, as shown in Minsker (2019), when $m \to \infty$ and $m = o(\sqrt{N})$, under some mild moment conditions (e.g., $\mathbb{E}|X - \mu|^3 < \infty$), the MOM estimator admits the following limiting distribution as $N \to \infty$,

$$\sqrt{N}(\widehat{\mu} - \mu) \xrightarrow{d} \mathcal{N}(0, \frac{\pi}{2}\sigma^2).$$

In addition to the robustness, the statistical efficiency is another important issue. In a classical statistical estimation setting without Byzantine failures, we can see the relative efficiency of $\widehat{\mu}$ with respect to the vanilla sample mean is $(\sigma^2)/(\frac{\pi}{2}\sigma^2) = 2/\pi \approx 0.637$, which is far from the optimal efficiency 1. Therefore, a natural question is:

> *Is it possible to construct a computationally efficient robust estimator that achieves a nearly-optimal efficiency?*

The key idea behind our VRMOM estimator To address this challenge, we first note that by the central limit theorem, for each sample mean $\overline{X}_j$, $\sqrt{n}\overline{X}_j$ asymptotically obeys the normal distribution $\mathcal{N}(\mu, \sigma^2)$. Moreover, for the ease of notation, for a fixed n, we define

$$\overline{X} = \mu + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2/n), \tag{3}$$

where μ and σ are unknown. Note that for every quantile level τ, the τ-th population quantile of the normal distribution $\mathcal{N}(\mu, \sigma^2/n)$ is exactly $\mu^\tau := \mu + \sigma\Phi^{-1}(\tau)/\sqrt{n}$. To see this,

$$\mathbb{P}(\overline{X} \le \mu^\tau) = \mathbb{P}((\overline{X} - \mu)/(\sigma/\sqrt{n}) \le \Phi^{-1}(\tau)) = \mathbb{P}(N(0,1) \le \Phi^{-1}(\tau)) = \tau.$$

Additionally, by symmetry of normal distribution, we have

$$\frac{1}{2}(\mu^\tau + \mu^{1-\tau}) = \frac{1}{2}\Big[2\mu + \frac{\sigma}{\sqrt{n}}\left\{\Phi^{-1}(\tau) + \Phi^{-1}(1-\tau)\right\}\Big] = \mu.$$

Therefore, to improve the statistical efficiency, a natural idea is to approximate μ by averaging many pairs of estimators for the $(\tau, 1-\tau)$-th quantiles of $\mathcal{N}(\mu, \sigma^2/n)$, instead of using a single quantity (i.e., median). More precisely, let K be a pre-fixed integer. For any $1 \le k \le K$, let $\tau_k := k/(K+1)$ and $\bar{\mu}^{\tau_k}$ be the τ_k-th sample quantile of $\{\overline{X}_0, \ldots, \overline{X}_m\}$. Since $\tau_{K+1-k} = 1 - \tau_k$, $\bar{\mu}^{\tau_k}$ and $\bar{\mu}^{\tau_{K+1-k}}$ are symmetrical about μ and their average is a natural estimator of μ. Based on this idea, we can take weighted average of $\{\bar{\mu}^{\tau_k}\}_{k=1}^K$ as an estimator of μ, which improves the statistical efficiency. We also illustrate the main idea of the weighted averaged estimator in Figure 1. Next, we introduce a computationally more efficient estimator for implementing this idea. Moreover, since it is a closed-form estimator, which also facilitates the theoretical analysis.

Computationally efficient VRMOM estimator Denote the quantile loss function as $\rho_\tau(z) = z(\tau - \mathbb{I}(z \le 0))$, we consider the following stochastic optimization problem:

$$\operatorname*{argmin}_{x\in\mathbb{R}} \left[\mathbb{E}\left\{G(\overline{X}, x)\right\}\right] := \operatorname*{argmin}_{x\in\mathbb{R}} \left[\mathbb{E}\left\{\sum_{k=1}^K \rho_{\tau_k}\left(\overline{X} - \frac{\sigma\Delta_k}{\sqrt{n}} - x\right)\right\}\right], \tag{4}$$

where $\Delta_k := \Phi^{-1}(\tau_k)$, and the expectation is taken over $\overline{X}$ in (3). We can easily see that μ is the solution of (4). To approximate μ from (4), we adopt the idea of one-step estimator

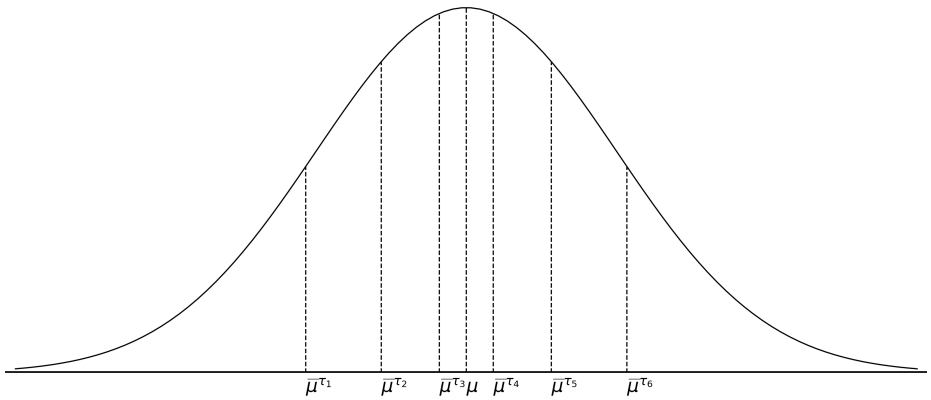

Figure 1: Let $K = 6$. For $1 \le k \le K$, $\bar{\mu}^{\tau_k}$ is defined as the $k/7$-th sample quantile of $\{\overline{X}_0, \ldots, \overline{X}_m\}$. We can see that the pairs $(\bar{\mu}^{\tau_1}, \bar{\mu}^{\tau_6}), (\bar{\mu}^{\tau_2}, \bar{\mu}^{\tau_5}), (\bar{\mu}^{\tau_3}, \bar{\mu}^{\tau_4})$ are nearly symmetrical about the targeting parameter μ. Thus the weighted average of $\{\bar{\mu}^{\tau_k}\}_{k=1}^K$ serves as an estimator of μ.

as follows. Define

$$g(x) := \frac{\mathrm{d}}{\mathrm{d}x}\mathbb{E}\left\{G(\overline{X}, x)\right\} = \mathbb{E}\left\{\sum_{k=1}^{K}\mathbb{I}\left(\overline{X} \le x + \frac{\sigma\Delta_k}{\sqrt{n}}\right) - \tau_k\right\},$$
$$H(x) := \frac{\mathrm{d}}{\mathrm{d}x}g(x) = \sum_{k=1}^{K}\mathfrak{p}\left(x - \mu + \frac{\sigma\Delta_k}{\sqrt{n}}\right),$$

as the gradient and Hessian of the loss function in (4) respectively. Here, $\mathfrak{p}(x) = \sqrt{n}\psi(\sqrt{n}x/\sigma)/\sigma$ denotes the probability density function of the noise $\epsilon \sim N(0, \sigma^2/n)$ in (3), and $\psi(\cdot)$ denotes the density function of $\mathcal{N}(0,1)$. Given an initial crude estimator μ_0 of μ, the one-step estimator essentially takes the following Newton-Raphson step:

$$\widetilde{\mu}_1 := \mu_0 - g(\mu_0)/H(\mu_0) = \mu_0 - \frac{\mathbb{E}\{\sum_{k=1}^{K}\mathbb{I}\left(\overline{X} \le \mu_0 + \sigma\Delta_k/\sqrt{n}\right) - \tau_k\}}{\sum_{k=1}^{K}\mathfrak{p}(\mu_0 - \mu + \sigma\Delta_k/\sqrt{n})}. \tag{5}$$

Next, we use the MOM estimator $\widehat{\mu}$ in (2) as the initial estimator of (4). The unknown parameter σ^2 can be estimated by $\widehat{\sigma}^2 := \sum_{i\in\mathcal{H}_0}(X_i - \overline{X}_0)^2/n$, the sample variance of the first batch of observations $\mathcal{H}_0$. With the initial estimator μ_0 in place, replacing $g(\mu_0)$ in (5) with its empirical counterpart and approximating $H(\mu_0)$ by $\sum_{k=1}^{K}\mathfrak{p}(\widehat{\sigma}\Delta_k/\sqrt{n}) \approx \sqrt{n}\sum_{k=1}^{K}\psi(\Delta_k)/\widehat{\sigma}$, we derive the following one-step estimator of (4) from (5)

$$\bar{\mu} = \widehat{\mu} - \frac{\widehat{\sigma}}{(m+1)\sqrt{n}\sum_{k=1}^{K}\psi(\Delta_k)}\sum_{j=0}^{m}\sum_{k=1}^{K}\left\{\mathbb{I}\left(\overline{X}_j \le \widehat{\mu} + \frac{\widehat{\sigma}\Delta_k}{\sqrt{n}}\right) - \frac{k}{K+1}\right\}. \tag{6}$$

To further alleviate the burden of computation, we choose one summand in (6) and simplify it as follows

$$\begin{aligned}
& \sum_{k=1}^{K}\left\{\mathbb{I}\left(\overline{X}_j \le \widehat{\mu}+\frac{\widehat{\sigma}\Delta_k}{\sqrt{n}}\right)-\frac{k}{K+1}\right\} \\
= & \sum_{k=1}^{K}\mathbb{I}\left(\overline{X}_j \le \widehat{\mu}+\frac{\widehat{\sigma}\Delta_k}{\sqrt{n}}\right)-\frac{K}{2} \\
= & \mathrm{Card}\left\{k\ :\ \frac{\sqrt{n}(\overline{X}_j-\widehat{\mu})}{\widehat{\sigma}} \le \Delta_k,\ 1\le k\le K\right\}-\frac{K}{2} \\
= & \mathrm{Card}\left\{k\ :\ \Phi\left(\frac{\sqrt{n}(\overline{X}_j-\widehat{\mu})}{\widehat{\sigma}}\right) \le \frac{k}{K+1},\ 1\le k\le K\right\}-\frac{K}{2} \\
= & \frac{K}{2}+1-\left\lceil (K+1)\Phi\left(\frac{\sqrt{n}(\overline{X}_j-\widehat{\mu})}{\widehat{\sigma}}\right)\right\rceil .
\end{aligned}$$

In fact, our theoretical result will show that a larger K leads to a better statistical efficiency. This derivation shows that it is possible to enhance the efficiency by taking a larger K without incurring additional computational cost. Our VRMOM estimator in (6) can be rewritten as

$$\bar{\mu}=\widehat{\mu}-\frac{\widehat{\sigma}}{(m+1)\sqrt{n}\sum_{k=1}^{K}\psi(\Delta_k)}\sum_{j=0}^{m}\left\{\frac{K}{2}+1-\left\lceil (K+1)\Phi\left(\frac{\sqrt{n}(\overline{X}_j-\widehat{\mu})}{\widehat{\sigma}}\right)\right\rceil\right\}. \tag{7}$$

Although the expression of VRMOM $\bar{\mu}$ in (7) seems more complicated than the MOM estimator $\widehat{\mu}$, the time complexity of $\bar{\mu}$ is at the same order as the MOM estimator. In particular, at each iteration, each worker machine computes a local sample mean in parallel, which takes $O(n)$ time complexity. In MOM estimator, it takes another $O(m)$ operations to find the median $\widehat{\mu}$ (see Paterson (1996)). While in VRMOM estimator, we only need an extra $O(m+n+K)$ time complexity ($O(n)$ for sample variance computed in $\mathcal{H}_0$, $O(m+K)$ for the variance reduction term in (7)). Therefore the time complexity of both methods is $O(m+n)$ (here K is fixed). While keeping the same order of computational complexity, the VRMOM greatly improves the statistical efficiency. As we can see from Theorem 5 in the following section, the asymptotic efficiency of $\bar{\mu}$ approaches $3/\pi \approx 0.955$ as K grows to infinity, which is nearly optimal. In fact, by taking $K=5$, the efficiency has already been more than 0.9 as compared to 0.637 of $\widehat{\mu}$.

Remark 2 *As illustrated in Figure 1, we can also find the τ_k-th sample quantile $\bar{\mu}^{\tau_k}$ and take a weighted average of $\{\bar{\mu}^{\tau_k}\}_{k=1}^{K}$ as an estimator of μ. However, to give the sample quantiles at K different levels, we need to perform a sorting algorithm among the set $\{\overline{X}_j\}_{j=0}^{m}$, which takes $O(m\log m)$ operations. Therefore the total complexity would be $O(n+m\log m)$, which it is more costly compared with $O(m+n)$ complexity of our VRMOM estimator in (7). The inferior in complexity is exacerbated in multivariate case. When the dimension p is very large, our coordinate-wise VRMOM estimator has complexity $O(p(m+n))$, while the average of sample quantiles would take complexity $O(p(m\log m+n))$.*

Remark 3 *Although we utilize averaging sample quantiles to motivate our method, the direct average among sample quantiles cannot tolerate even a small fraction of Byzantine*

machines. For example, we assume $K = 6$, and there are $1/6$ of machines are Byzantine. In this case, the $1/7$-th and $6/7$-th sample quantiles can be completely ruined by the Byzantine machines, and further foil the weighted average. In contrast, our VRMOM estimator tolerates $1/2 - \delta$ (where $\delta \in (0, 1/2)$ can be arbitrarily small) fraction of Byzantine machines, which is better than the direct weighted average of sample quantiles (see Theorem 6 for more details). To see this in a more intuitive way, we can take a closer look at (6). Each summand of the correction term is bounded in the interval $[-1, 1]$. Noticing that there is a factor of order $O(1/(m\sqrt{n}))$ multiplying the summation, the overall magnitude of the correction term is only of the order $O(1/\sqrt{n})$. In consequence, as long as the initial estimator (e.g., the MOM estimator) is robust, our proposed VRMOM estimator is Byzantine robust.

Remark 4 *To approximate μ from (4), we can also directly solve the following optimization problem*

$$\underset{x\in\mathbb{R}}{\operatorname{argmin}} \left[\sum_{k=1}^{K} \sum_{j=0}^{m} \rho_{\tau_k} \left(\overline{X}_j - \frac{\widehat{\sigma}\Delta_k}{\sqrt{n}} - x \right) \right],$$

which is the empirical version of (4). This formula is similar as the univariate composite quantile regression in Zou and Yuan (2008). However, it is much costly to solve such non-smooth optimization problem in a direct way. Instead, we leverage the idea of Newton-Raphson step, which greatly improves computation efficiency.

2.2 Theories for VRMOM Estimator

Now, we present several theoretical results for the proposed VRMOM estimator. We firstly provide the asymptotic normality and convergence result of the VRMOM estimator in one-dimensional case. Then we extend these results to its multi-dimensional variant. The proofs of results in this section are all relegated in Appendix A.

Theorem 5 (Asymptotic normality of VRMOM) *Let $N = (m+1)n$ i.i.d. random variables $X_1, ..., X_N$ be evenly distributed in $m+1$ subsets $\mathcal{H}_0, ..., \mathcal{H}_m$. There is a subset of Byzantine machine indices $\mathcal{B} \subseteq \{1, \ldots, m\}$ with $\mathrm{Card}(\mathcal{B}) = \lfloor \alpha_n m \rfloor$, where $\alpha_n = o(m^{-1/2})$. Let*

$$\overline{X}_j = \begin{cases} \frac{1}{n}\sum_{i\in\mathcal{H}_j} X_i & j \notin \mathcal{B}, \\ * & j \in \mathcal{B}, \end{cases} \tag{8}$$

and $\bar{\mu}$ be defined as in (6). Suppose X satisfies $\mathbb{E}(X) = \mu, \mathrm{Var}(X) = \sigma^2$. Assume there exists some $\kappa > 0$ such that $\mathbb{E}[|X - \mu|^{2+\kappa}] < \infty$, and $m = o(\min\{n, n^{2\kappa/(2+\kappa)}\}), \log^3 n = o(m)$. Then we have

$$\sqrt{N}(\bar{\mu} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma_K^2),$$

where

$$\sigma_K^2 = \frac{\sum_{k_1,k_2=1}^{K} \min(\tau_{k_1}, \tau_{k_2})\{1 - \max(\tau_{k_1}, \tau_{k_2})\}}{\{\sum_{k=1}^{K} \psi(\Delta_k)\}^2} \sigma^2, \tag{9}$$

with $\tau_k = k/(K+1)$. Moreover, $\lim_{K\to\infty} \sigma_K^2 = \pi\sigma^2/3$.

This theorem provides the asymptotic normality result of our VRMOM estimator $\bar{\mu}$ and characterizes the asymptotic variance. In particular, it shows that $\bar{\mu}$ is a consistent estimator of μ. Comparing with the MOM estimator, we *improve the efficiency* of the estimator by reducing the variance from $\pi\sigma^2/2$ (Minsker, 2019) to $\pi\sigma^2/3$ when K goes to infinity. It should be noted that, we impose the rate constraints $\alpha_n = o(m^{-1/2}), m = o(\min\{n, n^{2\kappa/(2+\kappa)}\})$, and $\log^3 n = o(m)$ in order to obtain asymptotic normality. In the following theorem, we drop out these conditions and investigate the convergence rate of the VRMOM estimator.

Theorem 6 (Convergence rate of VRMOM) *Let $N = (m+1)n$ i.i.d. random variables $X_1, ..., X_N$ be evenly distributed in $m+1$ subsets $\mathcal{H}_0, ..., \mathcal{H}_m$. There is a subset of Byzantine machine indices $\mathcal{B} \subseteq \{1, \ldots, m\}$ with $\mathrm{Card}(\mathcal{B}) = \lfloor \alpha_n m \rfloor$, where $\alpha_n \leq 1/2 - \delta$ for some fixed $\delta \in (0, 1/2)$. Let*

$$\overline{X}_j = \begin{cases} \frac{1}{n}\sum_{i \in \mathcal{H}_j} X_i & j \notin \mathcal{B}, \\ * & j \in \mathcal{B}, \end{cases}$$

and $\bar{\mu}$ be defined as in (6). Suppose X satisfies $\mathbb{E}(X) = \mu$, $Var(X) = \sigma^2$. Assume there exists some $\kappa > 0$ such that $\mathbb{E}[|X - \mu|^{2+\kappa}] < \infty$. Then we have

$$|\,\bar{\mu} - \mu| = O_{\mathbb{P}}\left(\frac{\alpha_n}{\sqrt{n}} + \frac{1}{\sqrt{mn}} + \frac{1}{n^{(3\kappa_2+2)/(2\kappa_2+4)}} + \frac{\log^{3/4} n}{n^{1/2} m^{3/4}}\right), \tag{10}$$

where $\kappa_2 = \min(\kappa, 2)$.

This convergence result shows that our VRMOM estimator is consistent as long as α_n is strictly smaller than $1/2$. The condition on α_n is also necessary because clearly the sample median can be ruined when there are more than $\lceil m/2 \rceil$ corruptions. From (10), when $m = O(\min\{n, n^{2\kappa/(2+\kappa)}\}), \log^3 n = O(m)$, the rate matches the optimal rate $O(\alpha_n/\sqrt{n} + 1/\sqrt{mn})$ (See Observation 1 in Yin et al. (2018)). Further assume that $\alpha_n = O(1/\sqrt{m})$, the VRMOM achieves square root-N consistency.

Next we extend our VRMOM estimator to the multi-dimensional extension setting. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ be i.i.d. copies of the p-dimensional random vectors $\boldsymbol{X} = (X^{(1)}, \ldots, X^{(p)})^{\mathrm{T}}$ with $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu} = (\mu^{(1)}, \ldots, \mu^{(p)})^{\mathrm{T}}$ and $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma} = (\sigma_{l_1,l_2})_{l_1,l_2=1}^{p}$. Then the multi-dimensional VRMOM estimator $\bar{\boldsymbol{\mu}}$ is defined by applying (7) on each coordinate l, where $1 \leq l \leq p$. We first obtain the convergence rate of the multi-dimensional VRMOM estimator in terms of ℓ_2-norm.

Theorem 7 (Convergence rate of multi-dimensional VRMOM) *Let $N = (m+1)n$ i.i.d. random vectors $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$ be evenly distributed in $m+1$ subsets $\mathcal{H}_0, ..., \mathcal{H}_m$. There is a subset of Byzantine machine indices $\mathcal{B} \subseteq \{1, \ldots, m\}$ with $\mathrm{Card}(\mathcal{B}) = \lfloor \alpha_n m \rfloor$, where $\alpha_n \leq 1/2 - \delta$ for some fixed $\delta \in (0, 1/2)$. Let $\bar{\boldsymbol{\mu}}$ be the multi-dimensional VRMOM estimator defined in (7). Suppose $\boldsymbol{X}$ satisfies $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu}, \mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$. Moreover, for each coordinate $l \in \{1, \ldots, p\}$, we assume there exists some $\kappa > 0$ such that $\mathbb{E}[|X^{(l)} - \mu^{(l)}|^{2+\kappa}] < \infty$. Then we have*

$$|\,\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2 = O_{\mathbb{P}}\left(\frac{\alpha_n\sqrt{p}}{\sqrt{n}} + \sqrt{\frac{p}{mn}} + \frac{\sqrt{p}}{n^{(3\kappa_2+2)/(2\kappa_2+4)}} + \frac{p^{1/2}\log^{3/4} n}{n^{1/2} m^{3/4}}\right), \tag{11}$$

where $\kappa_2 = \min(\kappa, 2)$.

As we can see from the theorem, the convergence rate of multi-dimensional VRMOM estimator is simply the rate in Theorem 6 multiplied with $\sqrt{p}$, which is not surprising because the VRMOM estimator is applied coordinate-wisely. Moreover, to guarantee consistency of the proposed estimator, we require the rate on the right hand side of (11) to be the order of $o_{\mathbb{P}}(1)$, which implies that

$$p = o\Big(\min\Big\{ \frac{n}{\alpha_n^2}, mn, n^{\frac{3\kappa_2+2}{\kappa_2+2}}, \frac{nm^{3/2}}{\log^{3/2} n} \Big\} \Big). \tag{12}$$

In particular, we are more interested in the asymptotic normality of the multi-dimensional VRMOM estimator, which will be presented in the next theorem.

By definition we know $\sigma_{l_1,l_2} = Cov\{X^{(l_1)}, X^{(l_2)}\}$ is the (l_1, l_2)-entry of covariance matrix of $\boldsymbol{X}$. Let (Z_{l_1}, Z_{l_2}) admit the following bivariate normal distribution

$$\mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{l_1,l_2}\right), \quad \text{where} \quad \boldsymbol{\Sigma}_{l_1,l_2} = \begin{pmatrix} 1 & \frac{\sigma_{l_1,l_2}}{\sqrt{\sigma_{l_1,l_1}\sigma_{l_2,l_2}}} \\ \frac{\sigma_{l_1,l_2}}{\sqrt{\sigma_{l_1,l_1}\sigma_{l_2,l_2}}} & 1 \end{pmatrix}. \tag{13}$$

Next we define the $p \times p$ matrix $\boldsymbol{\mathcal{C}}$ with its (l_1, l_2)-entry given by the following formula:

$$\mathcal{C}_{l_1,l_2} = \frac{\sum_{k_1,k_2=1}^{K} (\tau_{k_1,k_2}^{l_1,l_2} - \tau_{k_1}\tau_{k_2})}{\{\sum_{k=1}^{K} \psi(\Delta_k)\}^2} \sqrt{\sigma_{l_1,l_1}\sigma_{l_2,l_2}}, \tag{14}$$

where $\tau_k = k/(K+1)$, and $\tau_{k_1,k_2}^{l_1,l_2} = \mathbb{P}(Z_{l_1} \le \Delta_{k_1}, Z_{l_2} \le \Delta_{k_2})$. Then we can prove the following asymptotic normality result:

Theorem 8 (Asymptotic normality of multi-dimensional VRMOM estimator) *Under the same assumption as in Theorem 7, and additionally, we assume the rate constraints* $p = o(\min\{\frac{m^{1/2}}{\log^{3/2} n}, \frac{n^{2\kappa_2/(\kappa_2+2)}}{m}\})$, *and* $\alpha_n = o(1/\sqrt{mp})$. *Then for any vector* $\boldsymbol{v} \in \mathbb{R}^p$ *with* $|\boldsymbol{v}|_2 = 1$, *we have that*

$$\frac{\sqrt{N}}{\sigma_{\boldsymbol{v}}} \langle \boldsymbol{v},\, \bar{\boldsymbol{\mu}} - \boldsymbol{\mu} \rangle \xrightarrow{d} \mathcal{N}(0,1), \tag{15}$$

as $n \to \infty$, *where* $\sigma_{\boldsymbol{v}}^2 = \boldsymbol{v}^{\mathrm{T}} \boldsymbol{\mathcal{C}} \boldsymbol{v}$.

To prove asymptotic normality result for the multi-dimensional VRMOM estimator, we need a more restrictive constraint on the dimension p than the one in (12). Moreover, we require the number of Byzantine machines is $o(\sqrt{m/p})$, i.e., the fraction $\alpha_n = o(1/\sqrt{mp})$. As compared to the condition $\alpha_n = o(1/\sqrt{m})$ in Theorem 5, there is an extra $1/\sqrt{p}$ in the condition since we are dealing with a p-dimensional multivariate inference problem.

In order to illustrate the efficiency of our VRMOM estimator in multi-dimensional case, let us consider the multi-dimensional median-of-means (MOM) estimator $\widehat{\boldsymbol{\mu}}_{\mathrm{MOM}}$. More specifically, we also apply the MOM estimator at each coordinate and establish the following parallel asymptotic normality result for the multi-dimensional MOM estimator.

Proposition 9 (Asymptotic normality of multi-dimensional MOM estimator) *Under the same assumption as in Theorem 7, and additionally, we assume the rate constraints* $p = o(\min\{\frac{m^{1/2}}{\log^{3/2} n}, \frac{n^{2\kappa_2/(\kappa_2+2)}}{m}\})$, *and* $\alpha_n = o(1/\sqrt{mp})$. *Then for any vector* $\boldsymbol{v} \in \mathbb{R}^p$ *with* $|\boldsymbol{v}|_2 = 1$, *we have that*

$$\frac{\sqrt{N}}{\sigma_{\mathrm{MOM},\boldsymbol{v}}} \langle \boldsymbol{v}, \widehat{\boldsymbol{\mu}}_{\mathrm{MOM}} - \boldsymbol{\mu} \rangle \xrightarrow{d} \mathcal{N}(0,1), \tag{16}$$

as $n \to \infty$, *where* $\sigma^2_{\mathrm{MOM},\boldsymbol{v}} = \boldsymbol{v}^{\mathrm{T}} \boldsymbol{\mathcal{C}}_{\mathrm{MOM}} \boldsymbol{v}$, *and* $\boldsymbol{\mathcal{C}}_{\mathrm{MOM}}$ *is a* $p \times p$ *matrix with each* (l_1, l_2)*-entry taking the following form,*

$$\mathcal{C}_{\mathrm{MOM},l_1,l_2} = \left(2\pi \tau^{l_1,l_2}_{(K+1)/2,(K+1)/2} - \frac{\pi}{2}\right) \sqrt{\sigma_{l_1,l_1}\sigma_{l_2,l_2}}. \tag{17}$$

When each coordinate of random vector $\boldsymbol{X}$ is independent, the off-diagonal entries of the matrices $\boldsymbol{\mathcal{C}}$ and $\boldsymbol{\mathcal{C}}_{\mathrm{MOM}}$ are all zero (in this case $\tau^{l_1,l_2}_{k_1,k_2} = \tau_{k_1}\tau_{k_2}$ for $l_1 \neq l_2$). For diagonal entries, we can readily compute that

$$\mathcal{C}_{l,l} = \frac{\sum_{k_1,k_2=1}^{K} \min(\tau_{k_1}, \tau_{k_2})\{1 - \max(\tau_{k_1}, \tau_{k_2})\}}{\{\sum_{k=1}^{K} \psi(\Delta_k)\}^2} \sigma_{l,l}, \quad \mathcal{C}_{\mathrm{MOM},l,l} = \frac{\pi}{2}\sigma_{l,l}.$$

According to Theorem 5, when $K \to \infty$, we have $\mathcal{C}_{l,l} \to \frac{\pi}{3}\sigma_{l,l}$, which suggests that our multi-dimensional VRMOM estimator has a higher statistical efficiency than the corresponding MOM estimator.

Remark 10 *In two-dimensional case, the covariance matrix of* $\boldsymbol{X}$ *can be written as the following form*

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{1,1} & \sin\phi\sqrt{\sigma_{1,1}\sigma_{2,2}} \\ \sin\phi\sqrt{\sigma_{1,1}\sigma_{2,2}} & \sigma_{2,2} \end{pmatrix},$$

where $\phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. *Then we have that our 2-dimensional* VRMOM *estimator* $\overline{\boldsymbol{\mu}}$ *has higher statistical efficiency than the 2-dimensional* MOM *estimator* $\widehat{\boldsymbol{\mu}}$ *as* K *tends to infinity. The detailed argument is relegated to Appendix B. In the higher dimension case when* $p > 2$, *we believe that the superiority in efficiency of our* VRMOM *estimator still holds. We leave the theoretical investigation as a future work.*

3. Application for Byzantine Distributed Statistical Optimization

As an important application of the proposed VRMOM estimator, in this section, we consider the general distributed statistical optimization problem in (1) under Byzantine setup. In particular, we propose a Byzantine robust distributed approximate newton method, called Robust Communication-efficient Surrogate Likelihood (RCSL) Method.

For the ease of presentation, we adopt the master/worker setting in Jordan et al. (2019), where $\mathcal{H}_0$ denotes the master machine and the rest are worker machines. We assume that the master machine $\mathcal{H}_0$ stores n observations as each local worker and the data on the master machine will not be corrupted. In practice, it is easier to use one powerful machine as the master machine that is robust. Let $\widehat{\boldsymbol{\theta}}^{(0)}$ be an initial estimator of $\boldsymbol{\theta}^*$. At the beginning, the master machine $\mathcal{H}_0$ broadcasts the parameter $\widehat{\boldsymbol{\theta}}^{(0)}$ to all worker machines $\mathcal{H}_1, \ldots, \mathcal{H}_m$. The j-th worker computes a local gradient $\boldsymbol{g}_j^{(0)} = n^{-1}\sum_{i \in \mathcal{H}_j} \nabla|_{\boldsymbol{\theta}} f(X_i, \widehat{\boldsymbol{\theta}}^{(0)})$

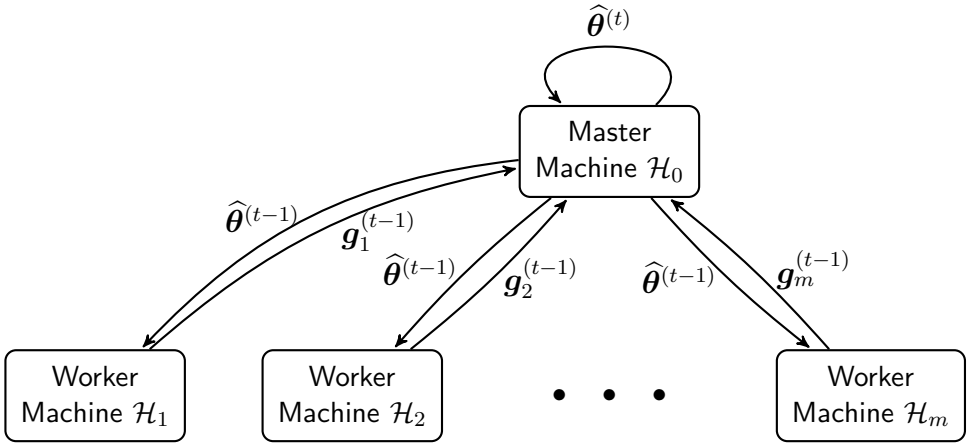

Figure 2: Communication protocol of the robust CSL (RCSL) method. In the t-th iteration, the master machine $\mathcal{H}_0$ distributes the parameter $\widehat{\boldsymbol{\theta}}^{(t-1)}$ to each worker machine. The j-th worker machine computes the local gradient $\boldsymbol{g}_j^{(t-1)}$ and sends it back to the master machine. Then $\mathcal{H}_0$ updates the new parameter $\widehat{\boldsymbol{\theta}}^{(t)}$ and repeats the procedure.

and sends it back to master. Then the master machine applies the VRMOM estimator to every coordinate. More precisely, for each of the l-th coordinate (we will use the subscript l to represent the entry of a vector), master machine computes the VRMOM estimator

$$\begin{aligned}\bar{g}_l^{(0)} =& \widehat{g}_l^{(0)} - \frac{\widehat{\sigma}_l^{(0)}}{(m+1)\sqrt{n}\sum_{k=1}^{K}\psi(\Delta_k)}\times \\ & \sum_{j=0}^{m}\left\{\frac{K}{2}+1-\left\lceil (K+1)\Phi\left(\frac{\sqrt{n}(g_{j,l}^{(0)}-\widehat{g}_l^{(0)})}{\widehat{\sigma}_l^{(0)}}\right)\right\rceil\right\},\end{aligned} \tag{18}$$

where $\widehat{g}_l^{(0)} = \text{med}\{g_{0,l}^{(0)}, ..., g_{m,l}^{(0)}\}$ is the median and

$$(\widehat{\sigma}_l^{(0)})^2 = \frac{1}{n}\sum_{i\in\mathcal{H}_0}\left\{\nabla|_{\boldsymbol{\theta}} f_l(X_i, \widehat{\boldsymbol{\theta}}^{(0)}) - g_{0,l}^{(0)}\right\}^2,$$

is the local sample variance. Here, $\nabla|_{\boldsymbol{\theta}} f_l(X_i, \widehat{\boldsymbol{\theta}}^{(0)})$ denotes the l-th coordinate of $\nabla|_{\boldsymbol{\theta}} f(X_i, \widehat{\boldsymbol{\theta}}^{(0)})$. Let $\bar{\boldsymbol{g}}^{(0)} = (\bar{g}_1^{(0)}, \dots \bar{g}_p^{(0)})^{\mathrm{T}}$ denote the VRMOM aggregated gradient. The master machine solves the following surrogate loss introduced by Jordan et al. (2019) to update the parameter,

$$\widehat{\boldsymbol{\theta}}^{(1)} = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\text{argmin}}\left\{\frac{1}{n}\sum_{i\in\mathcal{H}_0} f(X_i, \boldsymbol{\theta}) - \left\langle \boldsymbol{g}_0^{(0)} - \bar{\boldsymbol{g}}^{(0)}, \boldsymbol{\theta}\right\rangle\right\}, \tag{19}$$

Algorithm 1 Robust CSL (RCSL) Method

Input: The data $\{X_1, ..., X_N\}$ is evenly distributed on $m+1$ machines $\{\mathcal{H}_0, \ldots, \mathcal{H}_m\}$. Let $\mathcal{H}_0$ be the master and the rest be workers.

1: Compute an initial estimator $\widehat{\boldsymbol{\theta}}^{(0)}$ on the master machine $\mathcal{H}_0$.

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

3: Distribute $\widehat{\boldsymbol{\theta}}^{(t-1)}$ to each local machine $j = 1, 2, \ldots, m$.

4: **for** $j = 0, \ldots, m$ **do**

5: The j-th worker machine computes the local gradient

$$\boldsymbol{g}_j^{(t-1)} = \begin{cases} n^{-1} \sum_{i \in \mathcal{H}_j} \nabla|_{\boldsymbol{\theta}} f(X_i, \widehat{\boldsymbol{\theta}}^{(t-1)}) & \text{if } \mathcal{H}_j \text{ is normal,} \\ * & \text{if } \mathcal{H}_j \text{ is Byzantine,} \end{cases}$$

where $*$ denotes arbitrary values. Then the j-th worker sends $\boldsymbol{g}_j^{(t-1)}$ back to master machine.

6: **end for**

7: Master machine constructs the VRMOM aggregated gradient $\bar{\boldsymbol{g}}^{(t-1)} = (\bar{g}_1^{(t-1)}, \ldots \bar{g}_p^{(t-1)})^{\mathrm{T}}$, where each l-th coordinate takes the following form,

$$\begin{aligned} \bar{g}_l^{(t-1)} =& \widehat{g}_l^{(t-1)} - \frac{\widehat{\sigma}_l^{(t-1)}}{(m+1)\sqrt{n} \sum_{k=1}^{K} \psi(\Delta_k)} \times \\ & \sum_{j=0}^{m} \left\{ \frac{K}{2} + 1 - \left\lfloor (K+1) \Phi\left(\frac{\sqrt{n}(g_{j,l}^{(t-1)} - \widehat{g}_l^{(t-1)})}{\widehat{\sigma}_l^{(t-1)}} \right) \right\rfloor \right\}, \end{aligned} \tag{20}$$

where $\widehat{g}_l^{(t-1)} = \mathrm{med}\{g_{0,l}^{(t-1)}, ..., g_{m,l}^{(t-1)}\}$, and

$$\left(\widehat{\sigma}_l^{(t-1)}\right)^2 = \frac{1}{n} \sum_{i \in \mathcal{H}_0} \left\{ \nabla|_{\boldsymbol{\theta}} f_l(X_i, \widehat{\boldsymbol{\theta}}^{(t-1)}) - g_{0,l}^{(t-1)} \right\}^2.$$

8: Master machine solves the following surrogate loss minimization problem

$$\widehat{\boldsymbol{\theta}}^{(t)} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\mathrm{argmin}} \left\{ \frac{1}{n} \sum_{i \in \mathcal{H}_0} f(X_i, \boldsymbol{\theta}) - \left\langle \boldsymbol{g}_0^{(t-1)} - \bar{\boldsymbol{g}}^{(t-1)}, \boldsymbol{\theta} \right\rangle \right\}. \tag{21}$$

9: **end for**

Output: The final estimator $\widehat{\boldsymbol{\theta}}^{(T)}$.

where $\boldsymbol{g}_0^{(0)} = n^{-1} \sum_{i \in \mathcal{H}_0} \nabla|_{\boldsymbol{\theta}} f(X_i, \widehat{\boldsymbol{\theta}}^{(0)})$ is the local gradient computed on the master machine. As shown in Jordan et al. (2019) and our experiments, for a wide range of statistical learning problems, the surrogate loss in (19) can be easily minimized by existing optimization solvers. Moreover, this surrogate loss minimization is done on the master machine, and thus does not involve any communication.

Repeating the above procedure, we develop a multi-round algorithm named Robust CSL (RCSL), which is presented in Algorithm 1. We note that in the Byzantine setting, there is a subset of workers $\mathcal{B} \subseteq \{1, \ldots, m\}$, which return arbitrary values in each iteration. For $j \in \mathcal{B}$, we will use $\boldsymbol{g}_j^{(t-1)} = *$ to represent these nuisance values. To guarantee the consistency of the initial estimator, in Step 1 of Algorithm 1, we can compute $\widehat{\boldsymbol{\theta}}^{(0)}$ by the local empirical risk minimization on the master machine $\mathcal{H}_0$, i.e.,

$$\widehat{\boldsymbol{\theta}}^{(0)} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i \in \mathcal{H}_0} f(X_i, \boldsymbol{\theta}) \right\}. \tag{22}$$

We note that our theoretical result only requires the consistency of the initial estimator, and thus other consistent estimators could also be used as the initial estimator.

Now we briefly comment on the communication cost of Algorithm 1. In each round, the communication cost is $O(mp)$, which is at the same order as other gradient descent algorithms in the literature (e.g., Yin et al. (2018, 2019)). Moreover, from Theorem 11 below, our RCSL only takes a constant number of rounds of communication, as opposed to the order of $O(\log(N/p))$ rounds in other gradient-based methods. Therefore, the total communication cost of RCSL is only $O(mp)$. For the sake of clarity, we only present the result of multi-round convergence rate of the RCSL method in the following. More detailed technical conditions and theoretical results are relegated to Appendix C.

Theorem 11 (Multi-round convergence rate of RCSL method) *Suppose Assumption A-G (see Appendix C.1.1) hold and the initial estimator $\widehat{\boldsymbol{\theta}}^{(0)}$ satisfies $|\widehat{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\theta}^*|_2 = O_{\mathbb{P}}(r_n)$. Further assume the fraction α_n of Byzantine machines satisfies $\alpha_n \leq 1/2 - \delta$ for some fixed $\delta \in (0, 1/2)$. The RCSL estimator in the t-th iteration $\widehat{\boldsymbol{\theta}}^{(t)}$ defined in (21) satisfies*

$$|\widehat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*|_2 = O_{\mathbb{P}}\left(\frac{\alpha_n \sqrt{p}}{\sqrt{n}} + \sqrt{\frac{p \log n}{mn}} + \frac{p^{1/2} \log^{3/4} n}{n^{1/2} m^{3/4}} + r_n p^t \left(\frac{\log n}{n} \right)^{t/2} \right). \tag{23}$$

The proof of this theorem can be found in Appendix C.1.2. We note that the second term $\sqrt{p \log n/(mn)}$ in the right hand side of (23) matches the optimal rate $\sqrt{p/(mn)}$ up to a logarithmic factor. The third term is inherited from the last term in (10) of Theorem 6. It becomes $O(\sqrt{p \log n/(mn)})$ when $\log n = O(m)$. The first term is the price paid for the Byzantine failures. As we can see, after each iteration, the fourth term in (23) is improved by a factor $p\sqrt{\log n/n}$, which is of the order $o(1)$ by the rate constraints in Assumption G (see Appendix C.1.1). In stark contrast, for the Byzantine robust gradient descent, the convergence rate is only improved by a constant factor $c < 1$ after each round of communication (See, e.g. Theorem 3 in Yin et al. (2019) and Theorem 3.4 in Alistarh et al. (2018)). Therefore, Theorem 11 suggests that the proposed RCSL method enjoys faster convergence rate than vanilla Byzantine robust gradient descent, therefore is more communication-efficient. We can also demonstrate the communication-efficiency of our RCSL method in another way. From the expression of (23), we can see that when the number of t is sufficiently large, i.e.,

$$t \geq \frac{\log m + \log p + 2 \log r_n}{\log n - 2 \log p - \log \log n} + 1, \tag{24}$$

the last term will become the order of $O(\sqrt{p\log n/(mn)})$. Moreover, with $p = O(n^{1/3}/\sqrt{\log n})$ and $r_n = o(1)$ in Assumption G, we have

$$\frac{\log m + \log p + 2\log r_n}{\log n - 2\log p - \log\log n} + 1 \le \frac{3\log m}{\log n} + c_0, \quad \text{for some constant } c_0 > 0.$$

Therefore, when $t \ge c_0 + 3\log m/\log n$, the rate of t-th iteration is dominated by the first three terms. It would save a lot of communication cost as compared to gradient-based algorithms, where at least $O(\log(N/p))$ iterations of communications are necessary.

Assume $\log n = O(m)$ and the number of iterations t is sufficiently large, our rate in (23) will become $O_{\mathbb{P}}(\alpha_n\sqrt{p}/\sqrt{n} + \sqrt{p\log n/(mn)})$. It is interesting to compare this rate with contemporary results of the Byzantine perturbed gradient method with different aggregators (see Theorem 2 in Yin et al. (2019)). For example, median aggregator leads to the rate $\alpha_n\sqrt{p}/\sqrt{n} + p/\sqrt{mn}$ (we save a $\sqrt{p}$ in the second term), the trimmed-mean aggregator to the rate $\alpha_n p/\sqrt{n} + p/\sqrt{mn}$ (saving a $\sqrt{p}$ in both terms), and the iterative filtering to the rate $\sqrt{\alpha_n/n} + \sqrt{p/mn}$ (saving a $\sqrt{\alpha_n}$ but losing a $\sqrt{p}$ in the first term). However, their filtering estimator involves solving convex programs iteratively. Moreover, in the theory of the filtering estimator, the Byzantine fraction α_n is required to be not larger than $1/4$ (see, *e.g.* Theorem 1 in Su and Xu (2019) and Theorem 5 in Yin et al. (2019)), which is more restrictive than ours ($\alpha_n \le 1/2 - \delta$). We would also like to note that the lower bound on the convergence rate is known to be $\Omega(\alpha_n/\sqrt{n} + \sqrt{p/mn})$. As compared to the lower bound, our upper bound is missing a $\sqrt{p}$ factor in the first term. When there is no Byzantine worker (i.e., $\alpha_n = 0$) or the dimensionality p is a constant, our rate is optimal upto logarithmic factors. We believe this extra $\sqrt{p}$ in the first term comes out because the gradients are aggregated coordinate-wisely. It would be interesting as a future direction to develop a new multi-variate aggregator based on VRMOM, which is both statistically and computationally efficient and achieves the optimal convergence rate.

It is also worthwhile noting that we can extend our algorithm to a general scheme by replacing (19) with the following surrogate loss minimization,

$$\operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{n}\sum_{i\in\mathcal{H}_0} f(X_i,\boldsymbol{\theta}) - \left\langle \boldsymbol{g}_0^{(0)} - \mathrm{Aggr}(\boldsymbol{g}_0^{(0)},\ldots,\boldsymbol{g}_m^{(0)}), \boldsymbol{\theta} \right\rangle \right\}, \tag{25}$$

where $\mathrm{Aggr}(\boldsymbol{g}_0^{(0)},\ldots,\boldsymbol{g}_m^{(0)})$ can be any consistent estimator of the population gradient $\mathbb{E}\{\nabla|_{\boldsymbol{\theta}} f(X,\widehat{\boldsymbol{\theta}}^{(0)})\}$ given $\widehat{\boldsymbol{\theta}}^{(0)}$. Any robust aggregator in the literature can be adopted in (25), e.g., median-of-means, trimmed mean (Yin et al., 2018), geometric median (Feng et al., 2014; Chen et al., 2017), Krum (Blanchard et al., 2017), marginal median (Xie et al., 2018). In the original CSL framework (Jordan et al., 2019), the aggregator is chosen as the coordinate-wise average, which is sensitive to corruptions.

Remark 12 *It is worthwhile noting that when the target parameter admits some specific structures, e.g., sparsity structure, it is straightforward to extend the proposed framework* (25) *to the following regularized problem*

$$\operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{n}\sum_{i\in\mathcal{H}_0} f(X_i,\boldsymbol{\theta}) - \left\langle \boldsymbol{g}_0^{(0)} - \mathrm{Aggr}(\boldsymbol{g}_0^{(0)},\ldots,\boldsymbol{g}_m^{(0)}), \boldsymbol{\theta} \right\rangle + \lambda_n\mathcal{R}(\boldsymbol{\theta}) \right\}, \tag{26}$$

where $\mathcal{R}(\boldsymbol{\theta})$ is some regularizer, for example, the ℓ_1-penalty (Tibshirani, 1996), smooth clipped absolute deviation (SCAD) (Fan and Li, 2001), and minimax concave penalty (MCP) (Zhang, 2010). With the above formulation, we are able to address the sparse learning problem in Byzantine robust setup. We leave more rigorous theoretical investigation of the penalized Byzantine robust estimation to future research.

4. Simulation Studies

In our simulation studies, we conduct several experiments to show the effectiveness of the VRMOM and the robust CSL (RCSL) Method.

4.1 Results for VRMOM

In this section, we show the performance of the proposed VRMOM estimator for robust mean estimation problem. We first demonstrate how the number of quantile levels K in (7) affects the estimation accuracy of the VRMOM estimator, and then compare the statistical efficiency of our VRMOM estimator and the MOM estimator defined in (2). We generate the random vectors X_is from the normal distribution $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}_X)$ where $\boldsymbol{\mu}^* = p^{-1/2}(1, (p-2)/(p-1), (p-3)/(p-1), \ldots, 0)$ and $\boldsymbol{\Sigma}_{\boldsymbol{X}} = \mathrm{diag}(1, \ldots, 1)$. We choose $p = 1$ and $p = 30$ to consider both univariate and multivariate cases. The entire sample size is $N = 1000 \times (100 + 1)$. By dividing the data into one master machine $\mathcal{H}_0$ and 100 worker machines $\{\mathcal{H}_1, \ldots, \mathcal{H}_{100}\}$. Note that the master machine $\mathcal{H}_0$ can never be corrupted in our setup. Therefore, each local sample size is $n = 1000$. We consider the following settings: (1) $\alpha_n = 0$ which denotes no Byzantine machine, (2) $\alpha_n > 0$, which denotes the existence of Byzantine machine case. We vary the fraction of Byzantine machines $\alpha_n = 0.05, 0.1, 0.15$. When $\alpha_n > 0$, we replace the sample means in each Byzantine machine by a random vector whose entries are generated from $\mathcal{N}(0, 200\mathbb{I})$ independently. For each experiment, we repeat 500 independent simulations and report averaged root mean square estimation errors and standard deviations.

4.1.1 Effect of K

In the first experiment, we vary the number of quantile levels K from $\{10, 20, 50, 100\}$ and investigate the estimation accuracy of the VRMOM estimator for different dimensions $p \in \{1, 30\}$ and different fractions $\alpha_n \in \{0, 0.05, 0.1, 0.15\}$. The results of the root mean square errors and the standard errors are presented in Table 1. As we can see from the table, for each fraction of Byzantine machine α_n, the root mean square errors of the VRMOM estimator for different Ks are almost the same. Based on this observation, in the following experiment, we fix K to be 10 for the ease of computation.

4.1.2 Comparison between VRMOM and MOM

In the second experiment, we compare the performance of the VRMOM estimator and the MOM estimator in terms of the root mean square errors and their standard errors. We fixed the total sample size as $N = 1000 \times (100 + 1)$, local sample size $n = 1000$. We let the dimension $p \in \{1, 30\}$ and the fraction of Byzantine machines varies from $\alpha_n \in$

Table 1: The root mean square errors (RMSEs) and their standard errors (in parentheses) of the VRMOM under sample size $N = 1000 \times (100+1)$, local sample size $n = 1000$ and number of quantile levels $K = 10, 20, 50, 100$. The sample means sent from Byzantine machines are generated from Gaussian $\mathcal{N}(0, 200\mathbb{I})$.

p	K	$\alpha_n = 0$	$\alpha_n = 0.05$	$\alpha_n = 0.1$	$\alpha_n = 0.15$
1	10	0.0025 (0.0018)	0.0027 (0.0020)	0.0030 (0.0022)	0.0032 (0.0024)
	20	0.0025 (0.0018)	0.0027 (0.0021)	0.0030 (0.0022)	0.0034 (0.0026)
	50	0.0025 (0.0018)	0.0028 (0.0021)	0.0030 (0.0022)	0.0033 (0.0024)
	100	0.0025 (0.0018)	0.0028 (0.0020)	0.0030 (0.0022)	0.0032 (0.0024)
30	10	0.0175 (0.0022)	0.0192 (0.0024)	0.0209 (0.0026)	0.0227 (0.0028)
	20	0.0174 (0.0022)	0.0192 (0.0024)	0.0209 (0.0026)	0.0228 (0.0030)
	50	0.0174 (0.0022)	0.0192 (0.0025)	0.0208 (0.0026)	0.0230 (0.0029)
	100	0.0174 (0.0022)	0.0192 (0.0024)	0.0208 (0.0027)	0.0230 (0.0030)

Table 2: The root mean square errors (RMSEs) and their standard errors (in parentheses) of the VRMOM and MOM, and the ratios of RMSEs between VRMOM and MOM under sample size $N = 1000 \times (100 + 1)$, local sample size $n = 1000$ and integer $K = 10$. The sample means sent from Byzantine machines are generated from Gaussian $\mathcal{N}(0, 200\mathbb{I})$.

p		$\alpha_n = 0$	$\alpha_n = 0.05$	$\alpha_n = 0.1$	$\alpha_n = 0.15$
1	VRMOM	0.0025 (0.0018)	0.0027 (0.0020)	0.0030 (0.0022)	0.0032 (0.0024)
	MOM	0.0030 (0.0022)	0.0031 (0.0023)	0.0033 (0.0025)	0.0035 (0.0026)
	Ratio	0.8613	0.8901	0.9044	0.9129
30	VRMOM	0.0175 (0.0022)	0.0192 (0.0024)	0.0209 (0.0026)	0.0227 (0.0028)
	MOM	0.0211 (0.0028)	0.0223 (0.0028)	0.0234 (0.0030)	0.0249 (0.0034)
	Ratio	0.8285	0.8601	0.8921	0.9108

$\{0, 0.05, 0.1, 0.15\}$. Throughout our experiment, we fix the number of quantiles K in (7) to be $K = 10$.

From Table 2, we observe that VRMOM has smaller root mean square errors than MOM as all the ratios are greater than 1. With the increase of the fraction of Byzantine machines, both methods has larger root mean square errors. The difference between VRMOM and MOM tends to be smaller with more Byzantine machines. It is interesting to note that, when the dimension p is 30, the ratio of the root mean square errors between VRMOM and MOM is smaller than that when $p = 1$. It suggests that the variance reduction effect of our VRMOM estimator becomes better for higher dimensions, although we have only proved the superiority when $p = 1$ and 2 in this paper (see Remark 10).

4.2 Results for Robust CSL Method

In this section, we consider the linear model and logistic regression model to demonstrate our robust CSL method.

Settings for the linear regression model For the linear model experiment, the data are generated as follows:

$$Y_i = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\theta}^* + \epsilon_i, \qquad i = 1, 2, \ldots, n,$$

where each $\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,p})^{\mathrm{T}}$ is a p-dimensional covariate vector and $(X_{i,1}, \ldots, X_{i,p})$s are drawn *i.i.d.* from a multivariate normal distribution $\mathcal{N}(0, \boldsymbol{\Sigma}_X)$. The covariance $\boldsymbol{\Sigma}_X$ is a symmetric Toeplitz matrix with $\Sigma_{ij} = 0.5^{|i-j|}$ for $1 \le i, j \le p$, which enforces correlation structure among covariates. We fix the dimension $p = 30$ and generate the entries of the true coefficient vector $\boldsymbol{\theta}^*$ to be $p^{-1/2}(1, (p-2)/(p-1), (p-3)/(p-1), \ldots, 0)$. Similar to the previous experiment, the entire sample size is $N = 1000 \times (100 + 1)$ and we divide the data into one master machine $\mathcal{H}_0$ and 100 worker machines $\{\mathcal{H}_1, \ldots, \mathcal{H}_{100}\}$. so that each local sample size $n = 1000$. We consider the standard normal noise distribution where the noise $\epsilon_i \sim \mathcal{N}(0, 1)$. For the initial estimator $\widehat{\boldsymbol{\theta}}^{(0)}$, according to (22), we use the least square estimator with the data only on the master machine $\mathcal{H}_0$, i.e., $\widehat{\boldsymbol{\theta}}^{(0)} = (\sum_{i\in\mathcal{H}_0} \boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}})^{-1}(\sum_{i\in\mathcal{H}_0} \boldsymbol{X}_i Y_i)$. The computation of the initial estimator is very efficient since it has closed form and does not require any communication. We also note that for the surrogate loss minimization problem (21) at $t-$th iteration, we directly obtain the closed-form solution $\widehat{\boldsymbol{\theta}}^{(t)} = (2n^{-1}\sum_{i\in\mathcal{H}_0} \boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}})^{-1}(2n^{-1}\sum_{i\in\mathcal{H}_0} \boldsymbol{X}_i Y_i + \boldsymbol{g}_0^{(t-1)} - \bar{\boldsymbol{g}}^{(t-1)})$, which is also computationally efficient. Moreover, we generate corrupted gradients sent from Byzantine machines from the following attack models,

(a) Gaussian attack: We replace the gradient vectors in the Byzantine machines by random vectors in which all the entries are generated from $\mathcal{N}(0, 200\mathbb{I})$ independently.

(b) Omniscient attack: For the Byzantine machines, we replace the true gradient vectors by the scaled negative gradients where the scale constant is extremely large ($1e10$ in our experiment).

(c) Bit-flip attack: For the Byzantine machines, we replace the true gradient vectors by flipping the sign fo the first five dimensions.

Settings for the logistic regression model For the logistic regression model experiment, the data are generated from the following:

$$Y_i = \begin{cases} 1 & \text{with probability } \mathcal{L}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\theta}^*), \\ 0 & \text{with probability } 1 - \mathcal{L}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\theta}^*), \end{cases} \qquad i = 1, 2, \ldots, n$$

where the link function $\mathcal{L}(x) = e^x/(1+e^x)$ and each $\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,p})^{\mathrm{T}}$ is a p-dimensional covariate vector which is drawn *i.i.d.* from a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_X)$, which is coincident with the setting in the linear regression model. We choose $\boldsymbol{\mu}_x = (\mu_x, \ldots, \mu_x)^{\mathrm{T}}$ and adopt two settings that $\mu_x = 0$ and $\mu_x = 0.5$. Here $\mu_x = 0$ corresponds to the balanced response case that 50% Y_is are 1 and 50% Y_is are 0. And $\mu_x = 0.5$ corresponds to the imbalanced response case where 76% Y_is are 1 and 24% Y_is are 0. We also fix the dimension $p = 30$ and adopt the same true coefficient vector $\boldsymbol{\theta}^*$ as before. For each setting, we repeat 500 independent simulations and report averaged root mean square estimation errors and standard deviations. For the initial estimator $\widehat{\boldsymbol{\theta}}^{(0)}$, we use the logistic regression estimator with the data only on master machine $\mathcal{H}_0$, i.e.,

Table 3: The root mean square errors (RMSEs) and their standard errors (in parentheses) of the RCSL and MOM-RCSL, and the ratios of RMSEs between RCSL and MOM-RCSL under sample size $N = 1000 \times (100+1)$, local sample size $n = 1000$ and integer $K = 10$. The corrupted gradients sent from Byzantine machines are generated from Gaussian, Omniscient and Bit-flip attacks. The tolerance parameter for the stopping rule is set to $e_r = 10^{-4}$.

Attack	None		
α_n	0		
RCSL	0.0231 (0.0036)		
MOM-RCSL	0.0319 (0.0050)		
Ratio	0.7243		
Attack	Gaussian		
α_n	0.05	0.1	0.15
RCSL	0.0270 (0.0044)	0.0312 (0.0049)	0.0351 (0.0060)
MOM-RCSL	0.0343 (0.0054)	0.0369 (0.0058)	0.0398 (0.0063)
Ratio	0.7863	0.8434	0.8817
Attack	Omniscient		
α_n	0.05	0.1	0.15
RCSL	0.0276 (0.0042)	0.0328 (0.0051)	0.0396 (0.0060)
MOM-RCSL	0.0355 (0.0057)	0.0395 (0.0061)	0.0449 (0.0069)
Ratio	0.7774	0.8296	0.8815
Attack	Bit-flip		
α_n	0.05	0.1	0.15
RCSL	0.0236 (0.0037)	0.0242 (0.0039)	0.0250 (0.0041)
MOM-RCSL	0.0325 (0.0051)	0.0334 (0.0053)	0.0343 (0.0058)
Ratio	0.7276	0.7248	0.7291

$\widehat{\boldsymbol{\theta}}^{(0)} = \operatorname{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \left[\frac{1}{n}\sum_{i\in\mathcal{H}_0}\left\{\log\left(1+e^{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\theta}}\right) - Y_i\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\theta}\right\}\right]$. The surrogate loss minimization problem (21) at $t-$th iteration, i.e.,

$$\widehat{\boldsymbol{\theta}}^{(t)} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \Big[\frac{1}{n}\sum_{i\in\mathcal{H}_0}\Big\{\log\Big(1+e^{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\theta}}\Big)\Big\} - \Big\{\Big(\frac{1}{n}\sum_{i\in\mathcal{H}_0} Y_i\boldsymbol{X}_i^{\mathrm{T}}\Big) + \boldsymbol{g}_0^{(t-1)} - \bar{\boldsymbol{g}}^{(t-1)}\Big\}\boldsymbol{\theta}\Big],$$

can be efficiently solved by standard gradient descent or quasi-Newton approaches (Nocedal and Wright, 2006) on the center machine without any communication.

As for the attack mode of Byzantine machines in the logistic regression model, we simulate the transmitted message in the following way. We replace every response Y by $1-Y$ and compute the gradients based on these transformed local data on each Byzantine machine.

More settings in the simulation For the fraction of Byzantine machines, we consider the following settings: (1) $\alpha_n = 0$ which denotes no Byzantine machine, (2) $\alpha_n > 0$, which denotes the existence of Byzantine machine case. We vary the fraction of Byzantine machines $\alpha_n = 0.05, 0.1, 0.15$. We also discuss the stopping criteria for Algorithm 1. Throughout

the experiments, we use the tolerance parameter $e_r = 10^{-4}$ as the stopping criterion. In particular, at the t-th iteration of Algorithm 1, we compute $e = |\widehat{\boldsymbol{\theta}}^{(t)} - \widehat{\boldsymbol{\theta}}^{(t-1)}|_2^2 / |\widehat{\boldsymbol{\theta}}^{(t-1)}|_2^2$ and stop the algorithm once $e \leq e_r$. In our experiments, it only requires 4 to 8 iterations to stop. We also provide the results with simple fixed number of iterations with $T = 5$ and $T = 10$.

Since the main focus of this paper is the variance reduction effect of the proposed VRMOM estimator, in the simulation study, we mainly compare the performance of our Robust CSL algorithm (RCSL) with the MOM-based Robust CSL algorithm (MOM-RCSL). More specifically, let $\widehat{\boldsymbol{\theta}}^{(0)}$ be the initial estimator obtained in the master machine $\mathcal{H}_0$, the refined estimator $\widehat{\boldsymbol{\theta}}^{(1)}_{\text{MOM}}$ is defined as the solution of (25) with $\text{Aggr}(\boldsymbol{g}_0^{(0)}, \ldots, \boldsymbol{g}_m^{(0)})$ being the MOM aggregator of the gradients. Repeat the procedure, and we denote the t-th round MOM-RCSL estimator by $\widehat{\boldsymbol{\theta}}^{(t)}_{\text{MOM}}$.

4.2.1 Results for linear regression model

The results for linear regression model are presented in Table 3 (for adaptive stopping criterion) and 4 (for fixed number of iterations). From Table 3, RCSL has smaller root mean square errors than MOM-RCSL and all the ratios are greater than 1 for different kinds of attacks. With the increase of Byzantine fractions α_n, both methods have larger root mean square errors. The difference between RCSL and MOM-RCSL tends to be smaller with more Byzantine machines, which is coincident with the phenomenon of the mean estimation problem (See Table 2). Among these three attack models, it seems that the omniscient attacker has the largest root mean square error in general. It is quite natural since this attacker makes the parameter vector go into the opposite direction by negative gradients with an extremely large scale factor (i.e., $1e10$), which exerts the most negative impact on the gradient aggregation step. On the other hand, even for such a strong attack model, our RCSL method still performs quite well. It is also interesting to compare Table 2 with the Gaussian attacker part in Table 3 and 4. All simulations share the same attack mode and the same dimension. It seems that the variance reduction effect of VRMOM is more significant when it is performed as an iterative gradient aggregator than as a mean estimator.

In Table 4, we fix the iteration number to be 5 and 10. Table 4 shows similar patterns as Table 3. There are almost no difference between $T = 5$ and $T = 10$ iterations. In other words, it shows that only using a very small number of iterations (i.e., $T = 5$), our RCSL estimator has already converged. This experiment suggests that the RCSL is communicationally efficient.

4.2.2 Results for logistic regression model

The results for logistic regression model are presented in Table 5 (for adaptive stopping criterion) and 6 (for fixed number of iterations). From Table 5, RCSL has smaller root mean square errors than MOM-RCSL since all the ratios are greater than 1. With the increase of Byzantine fractions α_n, both methods has larger root mean square errors. Moreover, the RCSL leads to a more significant improvement over the MOM-RCSL for imbalanced class case. Similar observations can be made from Table 6, where the number of iterations have

Table 4: The root mean square errors (RMSEs) and their standard errors (in parentheses) of the RCSL and MOM-RCSL, and the ratios of RMSEs between RCSL and MOM-RCSL under sample size $N = 1000 \times (100 + 1)$, local sample size $n = 1000$ and integer $K = 10$. The iteration numbers T are fixed to be 5 and 10. The corrupted gradients sent from Byzantine machines are generated from Gaussian, omniscient and bit-flip attacks.

T	Attack	None		
	α_n	0		
5	RCSL	0.0231 (0.0036)		
	MOM-RCSL	0.0319 (0.0051)		
	Ratio	0.7236		
10	RCSL	0.0231 (0.0036)		
	MOM-RCSL	0.0319 (0.0051)		
	Ratio	0.7233		
T	Attack	Gaussian		
	α_n	0.05	0.1	0.15
5	RCSL	0.0271 (0.0043)	0.0310 (0.005)	0.0354 (0.0057)
	MOM-RCSL	0.0343 (0.0056)	0.0373 (0.006)	0.0400 (0.0063)
	Ratio	0.7897	0.8305	0.8859
10	RCSL	0.0272 (0.0042)	0.0313 (0.0051)	0.0348 (0.0058)
	MOM-RCSL	0.0344 (0.0053)	0.0368 (0.0058)	0.0398 (0.0065)
	Ratio	0.7905	0.8483	0.8750
T	Attack	Omniscient		
	α_n	0.05	0.1	0.15
5	RCSL	0.0276 (0.0042)	0.0328 (0.0052)	0.0396 (0.0061)
	MOM-RCSL	0.0355 (0.0057)	0.0395 (0.0061)	0.0450 (0.0069)
	Ratio	0.7768	0.8304	0.8811
10	RCSL	0.0276 (0.0042)	0.0329 (0.0052)	0.0398 (0.0061)
	MOM-RCSL	0.0355 (0.0057)	0.0396 (0.0061)	0.0451 (0.0069)
	Ratio	0.7769	0.8311	0.8820
T	Attack	Bit-flip		
	α_n	0.05	0.1	0.15
5	RCSL	0.0236 (0.0037)	0.0242 (0.0039)	0.0250 (0.0041)
	MOM-RCSL	0.0325 (0.0051)	0.0334 (0.0053)	0.0343 (0.0058)
	Ratio	0.7268	0.7240	0.7281
10	RCSL	0.0236 (0.0037)	0.0242 (0.0039)	0.0250 (0.0041)
	MOM-RCSL	0.0325 (0.0051)	0.0335 (0.0053)	0.0344 (0.0058)
	Ratio	0.7259	0.7235	0.7260

been pre-determined. Table 6 also shows that our RCSL is communicationally efficient since $T = 5$ iterations have been sufficient for the convergence.

5. Conclusions and Future Work

In this paper, we design a Byzantine tolerant algorithm to address a general class of estimation and inference problems in a distributed setting. The first contribution is to improve the statistical efficiency of the widely used Median-of-Means (MOM) estimator by proposing a

Table 5: The root mean square errors (RMSEs) and their standard errors (in parentheses) of the RCSL and MOM-RCSL, and the ratios of RMSEs between RCSL and MOM-RCSL under sample size $N = 1000 \times (100 + 1)$, local sample size $n = 1000$ and integer $K = 10$. The tolerance parameter for the stopping rule is set to $e_r = 10^{-4}$. The parameter μ_x controls the class balance.

μ_x	0			
α_n	0	0.05	0.1	0.15
RCSL	0.0504 (0.0075)	0.0531 (0.0075)	0.0600 (0.0072)	0.0701 (0.0075)
MOM-RCSL	0.0699 (0.0109)	0.0716 (0.0107)	0.0765 (0.0108)	0.0830 (0.0112)
Ratio	0.7215	0.7418	0.7844	0.8452
μ_x	0.5			
α_n	0	0.05	0.1	0.15
RCSL	0.0583 (0.0087)	0.0601 (0.0087)	0.0632 (0.0091)	0.0669 (0.0096)
MOM-RCSL	0.0830 (0.0135)	0.0841 (0.0132)	0.0868 (0.0134)	0.0905 (0.0141)
Ratio	0.7024	0.7142	0.7281	0.7395

Table 6: The root mean square errors (RMSEs) and their standard errors (in parentheses) of the RCSL and MOM-RCSL, and the ratios of RMSEs between RCSL and MOM-RCSL under sample size $N = 1000 \times (100 + 1)$, local sample size $n = 1000$ and integer $K = 10$. The iteration numbers T are fixed to be 5 and 10. The parameter μ_x controls the class balance.

T	μ_x	0			
	α_n	0	0.05	0.1	0.15
5	RCSL	0.0505 (0.0075)	0.0531 (0.0076)	0.0600 (0.0072)	0.0702 (0.0075)
	MOM-RCSL	0.0699 (0.0109)	0.0716 (0.0107)	0.0765 (0.0108)	0.0830 (0.0113)
	Ratio	0.7112	0.7371	0.7901	0.8646
10	RCSL	0.0504 (0.0075)	0.0531 (0.0075)	0.0600 (0.0072)	0.0701 (0.0075)
	MOM-RCSL	0.0699 (0.0109)	0.0716 (0.0107)	0.0765 (0.0108)	0.0830 (0.0112)
	Ratio	0.7218	0.7418	0.7845	0.8452
T	μ_x	0.5			
	α_n	0	0.05	0.1	0.15
5	RCSL	0.0583 (0.0087)	0.0601 (0.0087)	0.0632 (0.0091)	0.0670 (0.0096)
	MOM-RCSL	0.0829 (0.0135)	0.0841 (0.0132)	0.0868 (0.0134)	0.0906 (0.0140)
	Ratio	0.7028	0.7145	0.7281	0.7395
10	RCSL	0.0583 (0.0087)	0.0601 (0.0087)	0.0632 (0.0091)	0.0670 (0.0096)
	MOM-RCSL	0.0830 (0.0135)	0.0841 (0.0132)	0.0868 (0.0134)	0.0905 (0.0141)
	Ratio	0.7025	0.7145	0.7281	0.7395

new Variance Reduced MOM (VRMOM) estimator. It achieves a nearly optimal convergence rate upto a logarithmic factor and has the same order of computational complexity as the MOM estimator. Inspired by the VRMOM estimator, we further develop the Robust CSL (RCSL) method for general statistical inference problem. The convergence rate improves the previous results in Yin et al. (2018) using either median or trimmed-mean as the

gradient aggregator. Moreover, we establish the asymptotic normality result for our RCSL method.

To highlight our main idea behind VRMOM, we choose to focus on the one-dimensional case and extend to multi-variate case in a coordinate-wise way. Therefore, the convergence rate in Theorem 11 has an extra $\sqrt{p}$ in the term related to Byzantine failures (i.e., $\alpha_n\sqrt{p}/\sqrt{n}$). How to remove this extra $\sqrt{p}$ in a computationally efficient way while still achieving the same statistical efficiency is still an open problem. As an important future direction, we will address this challenge by developing new multivariate efficient aggregators based on the VRMOM estimator. On the other hand, by adding additional regularizers as in (26), we can address a wider class of problems in Byzantine setup. We leave these problems in the future study.

Acknowledgments

Weidong Liu's research is supported by National Program on Key Basic Research Project (973 Program, 2018AAA0100704), NSFC Grant No. 11825104 and 11690013, Youth Talent Support Program, and a grant from Australian Research Council. Xiaojun Mao's research is supported by NSFC Grant No. 12001109, Shanghai Sailing Program 19YF1402800, Major Research Plan of NSFC Grant No. 92046021, and the Science and Technology Commission of Shanghai Municipality grant 20dz1200600. The authors would like to thank the action editor and two anonymous referees for their constructive comments, which greatly improves the quality of the paper.

Appendix

The appendix consists of four parts. In Appendix A, we provide detailed proof for the theoretical results of VRMOM estimator presented in Section 2.2. In Appendix B, we prove the positive definiteness of $\boldsymbol{\mathcal{C}}_{\text{MOM}} - \boldsymbol{\mathcal{C}}$ in two-dimensional case, which have been mentioned in Remark 10 of the main paper. In Appendix C, we present the theoretical results and some technical assumptions for the RCSL method. Lastly, in Appendix D, we will show that a large class of generalized linear models and M-estimators suffice the proposed assumptions in Appendix C.1.1.

Appendix A. Proof of Theories for VRMOM Estimator

In this appendix, we mainly prove the theoretical results of the proposed VRMOM estimator in Section 2.2. In Appendix A.1, we introduce several lemmas which are useful for proofing the main theorems. Next, we present the main proofs in Appendix A.2.

A.1 Technical Lemmas

Lemma 13 *(Berry-Esseen Theorem, Theorem 9.1.3 in Chow and Teicher (2012)) If* $\{X_i,\ i \geq 1\}$ *are i.i.d. mean-zero random variables with* $\mathbb{E}|X_1|^2 = \sigma^2, \mathbb{E}|X_1|^{2+\kappa} < \infty$*, where* $\kappa \in (0, 1]$ *. Then there exists a constant* $C_\kappa > 0$ *such that*

$$\sup_{-\infty<x<\infty} \left| \mathbb{P}\left\{ \sum_{i=1}^{n} X_i < x\sigma n^{1/2} \right\} - \Phi(x) \right| \leq C_\kappa \frac{\mathbb{E}|X_1|^{2+\kappa}}{\sigma^{2+\kappa} n^{\kappa/2}}.$$

Lemma 14 *(Exponential Inequality, Lemma 1 in Cai and Liu (2011)) Let* $X_1, ..., X_n$ *be i.i.d. random variables with zero mean. Suppose that there exist some* $\eta > 0$ *and* $C > 0$ *such that* $\mathbb{E}(X_1^2 e^{\eta|X_1|}) \leq C$*. Then uniformly for* $0 < x \leq C$ *and* $n \geq 1$*, there is*

$$\mathbb{P}\left\{ \frac{1}{n} \sum_{i=1}^{n} X_i \geq (\eta + \eta^{-1})x \right\} \leq \exp\left(-\frac{nx^2}{C} \right).$$

This lemma will be the workhorse throughout all proofs in this article. For ease of notations, we will use

$$\mathfrak{E}(\eta, X) := \mathbb{E}\left(X^2 e^{\eta|X|} \right), \tag{27}$$

when the expression of X is too complicated.

Lemma 15 *Let* $X_1, ..., X_m$ *be i.i.d. random variables with cumulative distribution function* $G(x) = \mathbb{P}(X_1 \leq x)$*. And there exists some constants* $C > 0$ *and* $\kappa \in (0, 1]$ *such that there is*

$$|G(x_1) - G(x_2)| \leq C|x_1 - x_2| + \frac{C}{n^{\kappa/2}}, \tag{28}$$

holds for any $x_1, x_2 \in \mathbb{R}$*. Further define the function*

$$Z_i(x) = \mathbb{I}\{X_i \leq x\} - G(x).$$

Let $\delta_n = O(1)$ be some rate, then there exists $\widetilde{C}$ large enough such that

$$\mathbb{P}\left[\frac{1}{m}\sup_{|x|\le\delta_n}\left|\sum_{i=1}^{m}\{Z_i(x)-Z_i(0)\}\right| \ge \widetilde{C}\left(\sqrt{\frac{\delta_n\log n}{m}}+\frac{1}{n^{\kappa/2}}\right)\right] = O(n^{-\gamma}),$$

provided the rates constraints $\log n = O(m)$ and $\max\{m^{-1}\log n, n^{-\kappa/2}\} = O(\delta_n)$.

Proof Evenly divide the interval $[-\delta_n, \delta_n]$ into $2n$ pieces and denote the set $\mathfrak{N} = \{-\delta_n, -\frac{n-1}{n}\delta_n, ..., \delta_n\}$. Then

$$\begin{aligned}\frac{1}{m}\sup_{|x|\le\delta_n}\left|\sum_{i=1}^{m}\{Z_i(x)-Z_i(0)\}\right| \le& \max_{\widetilde{x}\in\mathfrak{N}}\left|\frac{1}{m}\sum_{i=1}^{m}\{Z_i(\widetilde{x})-Z_i(0)\}\right| \\ &+\max_{\widetilde{x}\in\mathfrak{N}}\sup_{\{x:|\widetilde{x}-x|\le\delta_n/n\}}\left|\frac{1}{m}\sum_{i=1}^{m}\{Z_i(\widetilde{x})-Z_i(x)\}\right|.\end{aligned} \tag{29}$$

For the second term, notice that

$$\sup_{\{x:|\widetilde{x}-x|\le\delta_n/n\}}\left|\frac{1}{m}\sum_{i=1}^{m}\{Z_i(\widetilde{x})-Z_i(x)\}\right| \le \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left\{|X_i-\widetilde{x}|\le\frac{\delta_n}{n}\right\}+C\left(\frac{2\delta_n}{n}+\frac{1}{n^{\kappa/2}}\right).$$

For every fixed $\widetilde{x}\in\mathfrak{N}$, we know

$$\begin{aligned}&\mathbb{E}\left\{\mathbb{I}\left(|X_i-\widetilde{x}|\le\frac{\delta_n}{n}\right)\right\} = \mathbb{P}\left(|X_i-\widetilde{x}|\le\frac{\delta_n}{n}\right) \le C\left(\frac{2\delta_n}{n}+\frac{1}{n^{\kappa/2}}\right),\\ &\mathfrak{E}\left\{1,\mathbb{I}\left(|X_i-\widetilde{x}|\le\frac{\delta_n}{n}\right)-\mathbb{P}\left(|X_i-\widetilde{x}|\le\frac{\delta_n}{n}\right)\right\}\\ \le& e\mathbb{P}\left(|X_i-\widetilde{x}|\le\frac{\delta_n}{n}\right)\left\{\mathbb{P}\left(|X_i-\widetilde{x}|\le\frac{\delta_n}{n}\right)+1\right\} = O\left(\frac{\log n}{m}+\frac{1}{n^{\kappa/2}}\right),\end{aligned}$$

since $\delta_n/n = O(n^{-\kappa/2})$ from the rates constraints. Then applying Lemma 14 we have

$$\begin{aligned}&\mathbb{P}\left[\max_{\widetilde{x}\in\mathfrak{N}}\sup_{\{x:|\widetilde{x}-x|\le\delta_n/n\}}\frac{1}{m}\left|\sum_{i=1}^{m}\{Z_i(\widetilde{x})-Z_i(x)\}\right| \ge C_1\left(\frac{1}{n^{\kappa/2}}+\frac{\log n}{m}\right)+C\left(\frac{2\delta_n}{n}+\frac{1}{n^{\kappa/2}}\right)\right]\\ \le& 2n\max_{\widetilde{x}\in\mathfrak{N}}\mathbb{P}\left\{\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(|X_i-\widetilde{x}|\le\frac{\delta_n}{n}\right)-\mathbb{P}\left(|X_1-\widetilde{x}|\le\frac{\delta_n}{n}\right)\ge C_1\left(\frac{\log n}{m}+\frac{1}{n^{\kappa/2}}\right)\right\}\\ =& O(n^{-\gamma}),\end{aligned} \tag{30}$$

by letting C_1 large enough. Now for the first term in (29), again we apply Lemma 14 with

$$\begin{aligned}\mathbb{E}[Z_i(x)-Z_i(0)] &= 0,\\ \sup_{-\delta_n\le x\le\delta_n}\mathbb{E}\left[\{Z_i(x)-Z_i(0)\}^2 e^{|Z_i(x)-Z_i(0)|}\right] &= O(\delta_n),\end{aligned}$$

there is

$$\begin{aligned}&\mathbb{P}\left[\max_{\widetilde{x}\in\mathfrak{N}}\left|\frac{1}{m}\sum_{i=1}^{m}\{Z_i(\widetilde{x})-Z_i(0)\}\right|\geq C_2\sqrt{\frac{\delta_n\log n}{m}}\right]\\ \leq&2n\max_{\widetilde{x}\in\mathfrak{N}}\mathbb{P}\left[\frac{1}{m}\left|\sum_{i=1}^{m}\{Z_i(\widetilde{x})-Z_i(0)\}\right|\geq C_2\sqrt{\frac{\delta_n\log n}{m}}\right]=O(n^{-\gamma}),\end{aligned}\tag{31}$$

for some C_2 large enough. Then the lemma is proved by combining (30) and (31). ■

Lemma 16 *(Concentration of median-of-means with Byzantine machines) Let $N(=(m+1)n)$ i.i.d. random variables $X_1,...,X_N$ evenly distributed in $m+1$ subsets $\mathcal{H}_0,...,\mathcal{H}_m$. There is a subset of index $\mathcal{B}\subseteq\{1,\ldots,m\}$ with $\mathrm{Card}(\mathcal{B})=\lfloor\alpha_n m\rfloor$. Define*

$$\overline{X}_j=\begin{cases}\frac{1}{n}\sum_{i\in\mathcal{H}_j}X_i & j\notin\mathcal{B},\\ * & j\in\mathcal{B},\end{cases}\qquad \widehat{X}=med\left(\overline{X}_j\mid 0\leq j\leq m\right).$$

Suppose $\mathbb{E}(X_1)=0,\mathrm{Var}(X_1)=\sigma^2,\mathbb{E}|X_1|^{2+\kappa}<\infty$, where $\kappa\in(0,1]$. The fraction α_n satisfies $\alpha_n\leq 1/2-\delta$ for some $\delta>0$. Then for every $\gamma>1$, there exists $\widetilde{C}>0$ large enough, such that

$$\mathbb{P}\left\{\left|\widehat{X}\right|\geq\frac{\widetilde{C}}{\sqrt{n}}\left(\alpha_n+\frac{1}{n^{\kappa/2}}+\sqrt{\frac{\log n}{m}}\right)\right\}=O(n^{-\gamma}).$$

Proof By definition of medians, for any $x>0$, there is

$$\begin{aligned}&\mathbb{P}\left(\widehat{X}\geq x\right)=\mathbb{P}\left\{\sum_{j=0}^{m}\mathbb{I}\left(\overline{X}_j<x\right)\leq\frac{m+1}{2}\right\}\leq\mathbb{P}\left\{\sum_{j\notin\mathcal{B}}\mathbb{I}\left(\overline{X}_j<x\right)\leq\frac{m+1}{2}\right\}\\ =&\mathbb{P}\left\{\frac{1}{(1-\alpha_n)(m+1)}\sum_{j\notin\mathcal{B}}\mathbb{I}\left(\overline{X}_j<x\right)-\mathbb{P}\left(\overline{X}_1<x\right)\leq\frac{1}{2}+\frac{\alpha_n}{2(1-\alpha_n)}-\mathbb{P}\left(\overline{X}_1<x\right)\right\}\\ \leq&\mathbb{P}\left\{\frac{1}{(1-\alpha_n)(m+1)}\sum_{j\notin\mathcal{B}}\mathbb{I}\left(\overline{X}_j<x\right)-\mathbb{P}\left(\overline{X}_1<x\right)\leq\frac{1}{2}+\frac{\alpha_n}{2(1-\alpha_n)}+C_\kappa\frac{\mathbb{E}|X_1|^{2+\kappa}}{\sigma^{2+\kappa}n^{\kappa/2}}-\Phi\left(\frac{\sqrt{n}x}{\sigma}\right)\right\}\\ =&\mathbb{P}\left\{\frac{1}{(1-\alpha_n)(m+1)}\sum_{j\notin\mathcal{B}}\mathbb{I}\left(\overline{X}_j<x\right)-\mathbb{P}\left(\overline{X}_1<x\right)\leq-\sqrt{\frac{c\log n}{(1-\alpha_n)(m+1)}}\right\},\end{aligned}$$

where the last line uses Berry-Esseen Theorem (Lemma 13), and x is given by

$$x=\frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{1}{2}+\frac{\alpha_n}{2(1-\alpha_n)}+C_\kappa\frac{\mathbb{E}|X_1|^{2+\kappa}}{\sigma^{2+\kappa}n^{\kappa/2}}+\sqrt{\frac{c\log n}{(1-\alpha_n)(m+1)}}\right).$$

Now apply Lemma 14 for the i.i.d. sequence $\mathbb{I}\left(\overline{X}_j<x\right)-\mathbb{P}\left(\overline{X}_1<x\right)$ with

$$\mathfrak{E}\left\{1,\mathbb{I}\left(\overline{X}_j<x\right)-\mathbb{P}\left(\overline{X}_1<x\right)\right\}\leq e,$$

we have

$$\mathbb{P}\left\{\frac{1}{(1-\alpha_n)(m+1)}\sum_{j\notin\mathcal{B}}\mathbb{I}\left(\overline{X}_j<x\right)-\mathbb{P}\left(\overline{X}_1<x\right)\leq-\sqrt{\frac{c\log n}{(1-\alpha_n)(m+1)}}\right\}=O(n^{-\gamma}),$$

with $c=4e\gamma$. Moreover, we have the following elementary facts

$$\Phi^{-1}(x_0)=\Phi^{-1}(x_0)-\Phi^{-1}(1/2)\leq(x_0-1/2)\left(\Phi^{-1}\right)'(x_0)\leq\frac{x_0-1/2}{\psi\left\{\Phi^{-1}(1-\delta/2)\right\}},$$

holds for any $1/2\leq x_0<1-\delta/2$. Also we know $1/2\leq\Phi(\sqrt{n}x/\sigma)<1-\delta/2$ holds for m,n sufficiently large. Denote $C_\delta=1/\psi\{\Phi^{-1}(1-\delta/2)\}$, then there is

$$\begin{aligned}&\mathbb{P}\left\{\widehat{X}\geq\frac{\sigma C_\delta}{\sqrt{n}}\left(\frac{\alpha_n}{2(1-\alpha_n)}+C_\kappa\frac{\mathbb{E}|X_1|^{2+\kappa}}{\sigma^{2+\kappa}n^{\kappa/2}}+\sqrt{\frac{c\log n}{(1-\alpha_n)(m+1)}}\right)\right\}\leq\mathbb{P}\left(\widehat{X}\geq x\right)\\ \leq&\mathbb{P}\left\{\frac{1}{(1-\alpha_n)(m+1)}\sum_{j\notin\mathcal{B}}\mathbb{I}\left(\overline{X}_j<x\right)-\mathbb{P}\left(\overline{X}_1<x\right)\leq-\sqrt{\frac{c\log n}{(1-\alpha_n)(m+1)}}\right\}\leq n^{-\gamma}.\end{aligned}$$

Now do the same thing for $\mathbb{P}(\widehat{X}\leq-x)$, we finally get the desired result. ■

A.2 Proofs of the results in Section 2.2

We firstly provide the proofs related to the univariate VRMOM estimator.

Proof [Proof of Theorem 5, Theorem 6, and Theorem 7] Denote $\kappa_1=\min\{1,\kappa\}$ and $\kappa_2=\min\{2,\kappa\}$. To prove this result, let's firstly give a convergence rate for sample variance $\widehat{\sigma}-\sigma$. From the moment bound $\mathbb{E}|X_1-\mu|^{2+\kappa}<\infty$, by Marcinkiewicz-Zygmund theorem (Theorem 5.2.2 in Chow and Teicher (2012)), we have

$$\frac{1}{n}\sum_{i\in\mathcal{H}_0}(X_i-\mu)^2-\sigma^2=o_{\mathbb{P}}\left(\frac{1}{n^{\kappa_2/(2+\kappa_2)}}\right),\quad\frac{1}{n}\sum_{i\in\mathcal{H}_0}(X_i-\mu)=o_{\mathbb{P}}\left(\frac{1}{n^{\kappa_2/(2+\kappa_2)}}\right).$$

Therefore we have

$$\begin{aligned}|\widehat{\sigma}-\sigma|=&\frac{1}{\widehat{\sigma}+\sigma}\left|\frac{1}{n}\sum_{i\in\mathcal{H}_0}\left(X_i-\overline{X}_0\right)^2-\sigma^2\right|\\ \leq&\frac{1}{\sigma}\left|\frac{1}{n}\sum_{i\in\mathcal{H}_0}(X_i-\mu)^2-\sigma^2-\left(\overline{X}_0-\mu\right)^2\right|=o_{\mathbb{P}}\left(\frac{1}{n^{\kappa_2/(2+\kappa_2)}}\right).\end{aligned}\tag{32}$$

Next pick out one summand in (6). Denote

$$G_j(x):=\mathbb{P}\left\{\frac{\sqrt{n}(\overline{X}_j-\mu)}{\sigma}\leq x\right\},\quad I_j(x):=\mathbb{I}\left\{\frac{\sqrt{n}(\overline{X}_j-\mu)}{\sigma}\leq x\right\},$$

then

$$\begin{aligned}
&\mathbb{I}\left\{\overline{X}_j \le \widehat{\mu} + \frac{\widehat{\sigma}\Delta_k}{\sqrt{n}}\right\} - \frac{k}{K+1} \\
=&I_j\left\{\frac{\sqrt{n}(\widehat{\mu}-\mu)}{\sigma} + \frac{\widehat{\sigma}}{\sigma}\Delta_k\right\} - G_j\left\{\frac{\sqrt{n}(\widehat{\mu}-\mu)}{\sigma} + \frac{\widehat{\sigma}}{\sigma}\Delta_k\right\} + \underbrace{G_j\left\{\frac{\sqrt{n}(\widehat{\mu}-\mu)}{\sigma} + \frac{\widehat{\sigma}}{\sigma}\Delta_k\right\} - \Phi(\Delta_k)}_{T}.
\end{aligned}$$

For the term T,

$$\begin{aligned}
T =&G_j\left\{\frac{\sqrt{n}(\widehat{\mu}-\mu)}{\sigma} + \frac{\widehat{\sigma}}{\sigma}\Delta_k\right\} - \Phi\left\{\frac{\sqrt{n}(\widehat{\mu}-\mu)}{\sigma} + \frac{\widehat{\sigma}}{\sigma}\Delta_k\right\} + \Phi\left\{\frac{\sqrt{n}(\widehat{\mu}-\mu)}{\sigma} + \frac{\widehat{\sigma}}{\sigma}\Delta_k\right\} - \Phi(\Delta_k) \\
=&O(n^{-\kappa_1/2}) + \psi(\Delta_k)\frac{\sqrt{n}(\widehat{\mu}-\mu)}{\sigma} + \psi(\Delta_k)\frac{\widehat{\sigma}-\sigma}{\sigma}\Delta_k + O\left(\frac{\sqrt{n}(\widehat{\mu}-\mu)}{\sigma} + \frac{\widehat{\sigma}-\sigma}{\sigma}\Delta_k\right)^2 \\
=&\psi(\Delta_k)\frac{\sqrt{n}(\widehat{\mu}-\mu)}{\sigma} + O_{\mathbb{P}}\left(\alpha_n^2 + \frac{\log n}{m} + \frac{1}{n^{\kappa_2/(2+\kappa_2)}}\right),
\end{aligned} \tag{33}$$

where line 1 to line 2 uses Berry-Esseen Inequality (Lemma 13) and Taylor expansion, line 2 to line 3 uses (32) and concentration inequalities for $\widehat{\mu}$ (Lemma 16). By using Lemma 16 and (32) again we have that

$$\frac{\sqrt{n}(\widehat{\mu}-\mu)}{\sigma} + \frac{\widehat{\sigma}-\sigma}{\sigma}\Delta_k = O_{\mathbb{P}}\left(\alpha_n + \sqrt{\frac{\log n}{m}} + \frac{1}{n^{\kappa_2/(2+\kappa_2)}}\right).$$

Also note that Berry-Esseen Inequality guarantees that $\sqrt{n}(Y_j-\mu)/\sigma$ satisfies

$$|G_j(x_1) - G_j(x_2)| \le \psi(0)|x_1 - x_2| + 2C_{\kappa_1}\frac{\mathbb{E}|X_1-\mu|^{2+\kappa_1}}{\sigma^{2+\kappa_1}n^{\kappa_1/2}}.$$

So we can apply Lemma 15 with $\delta_n = O(\alpha_n + \sqrt{\frac{\log n}{m}} + \frac{1}{n^{\kappa_2/(2+\kappa_2)}})$, which yields

$$\begin{aligned}
&\frac{1}{m+1}\left|\sum_{j\notin\mathcal{B}}\left[I_j\left\{\frac{\sqrt{n}(\widehat{\mu}-\mu)}{\sigma} + \frac{\widehat{\sigma}}{\sigma}\Delta_k\right\} - G_j\left\{\frac{\sqrt{n}(\widehat{\mu}-\mu)}{\sigma} + \frac{\widehat{\sigma}}{\sigma}\Delta_k\right\} - I_j(\Delta_k) + G_j(\Delta_k)\right]\right| \\
\le&\frac{1}{m+1}\sup_{|x-\Delta_k|\le\delta_n}\left|\sum_{j\notin\mathcal{B}}\{I_j(x+\Delta_k) - G_j(x+\Delta_k) - I_j(\Delta_k) + G_j(\Delta_k)\}\right| \\
=&O_{\mathbb{P}}\left(\sqrt{\frac{\delta_n\log n}{m}} + \frac{1}{n^{\kappa_1/2}}\right) \\
=&O_{\mathbb{P}}\left(\sqrt{\frac{\alpha_n\log n}{m}} + \left(\frac{\log n}{m}\right)^{3/4} + \frac{\log^{1/2} n}{m^{1/2}n^{\kappa_2/(4+2\kappa_2)}} + \frac{1}{n^{\kappa_1/2}}\right).
\end{aligned}$$

Thus it implies that

$$\begin{aligned}
&\frac{1}{m+1}\sum_{j=0}^{m}\left[I_j\left\{\frac{\sqrt{n}(\widehat{\mu}-\mu)}{\sigma}+\frac{\widehat{\sigma}}{\sigma}\Delta_k\right\}-G_j\left\{\frac{\sqrt{n}(\widehat{\mu}-\mu)}{\sigma}+\frac{\widehat{\sigma}}{\sigma}\Delta_k\right\}\right]\\
=&\frac{1}{m+1}\sum_{j=0}^{m}\{I_j(\Delta_k)-G_j(\Delta_k)\}\\
&+O_{\mathbb{P}}\left(\sqrt{\frac{\alpha_n\log n}{m}}+\Big(\frac{\log n}{m}\Big)^{3/4}+\frac{\log^{1/2}n}{m^{1/2}n^{\kappa_2/(4+2\kappa_2)}}+\frac{1}{n^{\kappa_1/2}}\right).
\end{aligned}\tag{34}$$

Again from (32) we have

$$\begin{aligned}
&\left|\frac{\widehat{\sigma}-\sigma}{(m+1)\sqrt{n}\sum_{k=1}^{K}\psi(\Delta_k)}\sum_{k=1}^{K}\sum_{j=0}^{m}\left[\mathbb{I}\left(\overline{X}_j\le\widehat{\mu}+\frac{\widehat{\sigma}\Delta_k}{\sqrt{n}}\right)-\frac{k}{K+1}\right]\right|\\
\le&\frac{K|\widehat{\sigma}-\sigma|}{\sqrt{n}\sum_{k=1}^{K}\psi(\Delta_k)}=O_{\mathbb{P}}\left(\frac{1}{n^{(3\kappa_2+2)/(2\kappa_2+4)}}\right).
\end{aligned}\tag{35}$$

Combining equations (33) (34) (35), we have

$$\begin{aligned}
\bar{\mu}-\mu=&\widehat{\mu}-\mu-\frac{\widehat{\sigma}}{(m+1)\sqrt{n}\sum_{k=1}^{K}\psi(\Delta_k)}\sum_{k=1}^{K}\sum_{j=0}^{m}\left\{\mathbb{I}\left(\overline{X}_j\le\widehat{\mu}+\frac{\widehat{\sigma}\Delta_k}{\sqrt{n}}\right)-\frac{k}{K+1}\right\}\\
=&\widehat{\mu}-\mu-\frac{\sigma}{(m+1)\sqrt{n}\sum_{k=1}^{K}\psi(\Delta_k)}\sum_{k=1}^{K}\sum_{j\notin\mathcal{B}}\left\{\mathbb{I}\left(\overline{X}_j\le\widehat{\mu}+\frac{\widehat{\sigma}\Delta_k}{\sqrt{n}}\right)-\frac{k}{K+1}\right\}\\
&+O_{\mathbb{P}}\left(\frac{\alpha_n}{\sqrt{n}}+\frac{1}{n^{(3\kappa_2+2)/(2\kappa_2+4)}}\right)\\
=&\frac{\sigma}{\sqrt{n}\sum_{k=1}^{K}\psi(\Delta_k)}\sum_{k=1}^{K}\frac{1}{m+1}\sum_{j\notin\mathcal{B}}\left[G_j\left\{\frac{\sqrt{n}(\widehat{\mu}-\mu)}{\sigma}+\frac{\widehat{\sigma}}{\sigma}\Delta_k\right\}-I_j\left\{\frac{\sqrt{n}(\widehat{\mu}-\mu)}{\sigma}+\frac{\widehat{\sigma}}{\sigma}\Delta_k\right\}\right]\\
&+O_{\mathbb{P}}\left(\frac{\alpha_n}{\sqrt{n}}+\frac{1}{n^{(3\kappa_2+2)/(2\kappa_2+4)}}+\frac{\log n}{m\sqrt{n}}\right)\\
=&\frac{1}{m+1}\sum_{j\notin\mathcal{B}}\frac{\sigma}{\sqrt{n}\sum_{k=1}^{K}\psi(\Delta_k)}\sum_{k=1}^{K}\{G_j(\Delta_k)-I_j(\Delta_k)\}\\
&+O_{\mathbb{P}}\left(\frac{\alpha_n}{\sqrt{n}}+\frac{\log n}{m\sqrt{n}}+\frac{1}{n^{(3\kappa_2+2)/(2\kappa_2+4)}}+\frac{\log^{3/4}n}{n^{1/2}m^{3/4}}\right).
\end{aligned}\tag{36}$$

From central limit theorem, we have

$$\bar{\mu}-\mu=O_{\mathbb{P}}\left(\frac{\alpha_n}{\sqrt{n}}+\frac{1}{\sqrt{mn}}+\frac{1}{n^{(3\kappa_2+2)/(2\kappa_2+4)}}+\frac{\log^{3/4}n}{n^{1/2}m^{3/4}}\right),\tag{37}$$

which proves Theorem 6. Moreover, when $m=o(n^{2\kappa_2/(\kappa_2+2)})=o(\min\{n^{2\kappa/(\kappa+2)},n\}),\log^3 n=o(m)$ and $\alpha_n=o(1/\sqrt{m})$, the remainder in (36) will be of the order $o_{\mathbb{P}}(\frac{1}{\sqrt{mn}})$, while the

major term becomes a summation of i.i.d. sequence with

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K}\{G_j(\Delta_k)-I_j(\Delta_k)\}\right]=0, \text{ and}\\
&\mathrm{Var}\left[\sum_{k=1}^{K}\{G_j(\Delta_k)-I_j(\Delta_k)\}\right]\\
=&\sum_{k_1,k_2=1}^{K} G_j\left\{\min(\Delta_{k_1},\Delta_{k_2})\right\}-G_j(\Delta_{k_1})G_j(\Delta_{k_2})\\
=&\sum_{k_1,k_2=1}^{K}\min(\tau_{k_1},\tau_{k_2})\{1-\max(\tau_{k_1},\tau_{k_2})\}+O(n^{-1/2}),
\end{aligned}$$

where the last line again follows from Berry-Esseen theorem (Theorem 13). Applying standard central limit theorem we have

$$\sqrt{N}(\,\bar{\mu}-\mu)\xrightarrow{d}\mathcal{N}(0,\sigma_K^2),$$

where σ_K^2, as defined in (9), tends to $\pi\sigma^2/3$ according to Lemma 18 in Appendix B. Therefore Theorem 5 is proved.

To prove Theorem 7, we just apply (37) to each coordinate and obtain that

$$|\bar{\boldsymbol{\mu}}-\boldsymbol{\mu}|_2=\sqrt{\sum_{l=1}^{p}|\bar{\mu}_l-\mu_l|^2}=O_{\mathbb{P}}\left(\frac{\alpha_n\sqrt{p}}{\sqrt{n}}+\sqrt{\frac{p}{mn}}+\frac{\sqrt{p}}{n^{(3\kappa_2+2)/(2\kappa_2+4)}}+\frac{p^{1/2}\log^{3/4}n}{n^{1/2}m^{3/4}}\right).$$

■

Proof [Proof of Theorem 8] To prove asymptotic normality of multi-dimensional VRMOM estimator, using (36) and the rate constraints, for each coordinate l (where $1\le l\le p$), there is

$$\bar{\mu}_l-\mu_l=\frac{1}{m+1}\sum_{j\notin\mathcal{B}}\frac{\sqrt{\sigma_{l,l}}}{\sqrt{n}\sum_{k=1}^{K}\psi(\Delta_k)}\sum_{k=1}^{K}\{G_{j,l}(\Delta_k)-I_{j,l}(\Delta_k)\}+o_{\mathbb{P}}\left(\frac{1}{\sqrt{pmn}}\right),$$

where

$$G_{j,l}(x):=\mathbb{P}\left\{\frac{\sqrt{n}(\overline{X}_{j,l}-\mu_l)}{\sqrt{\sigma_{l,l}}}\le x\right\},\quad I_{j,l}(x):=\mathbb{I}\left\{\frac{\sqrt{n}(\overline{X}_{j,l}-\mu_l)}{\sqrt{\sigma_{l,l}}}\le x\right\}.$$

For any vector $|\boldsymbol{v}|_2=1$, there is

$$\langle\bar{\boldsymbol{\mu}}-\boldsymbol{\mu},\boldsymbol{v}\rangle=\frac{1}{m+1}\sum_{j\notin\mathcal{B}}\frac{1}{\sqrt{n}\sum_{k=1}^{K}\psi(\Delta_k)}\sum_{k=1}^{K}\sum_{l=1}^{p}\sqrt{\sigma_{l,l}}\{I_{j,l}(\Delta_k)-G_{j,l}(\Delta_k)\}v_l+o_{\mathbb{P}}\left(\frac{1}{\sqrt{mn}}\right).$$

Now we apply central limit theorem and yield

$$\frac{\sqrt{(m+1)n}}{\widetilde{\sigma}_{\boldsymbol{v}}}\langle \bar{\boldsymbol{\mu}} - \boldsymbol{\mu}, \boldsymbol{v}\rangle \xrightarrow{d} \mathcal{N}(0,1),$$
$$\text{where} \quad \widetilde{\sigma}_{\boldsymbol{v}}^2 = \boldsymbol{v}^{\mathrm{T}}\widetilde{\boldsymbol{\mathcal{C}}}\boldsymbol{v}.$$

Here $\widetilde{\boldsymbol{\mathcal{C}}} \in \mathbb{R}^{p\times p}$ has its (l_1, l_2)-entry defined as

$$\begin{aligned}\widetilde{\mathcal{C}}_{l_1,l_2} =& \frac{\sqrt{\sigma_{l_1,l_1}\sigma_{l_2,l_2}}}{\{\sum_{k=1}^K \psi(\Delta_k)\}^2}\mathbb{E}\left[\sum_{k=1}^K\{I_{0,l_1}(\Delta_k) - G_{0,l_1}(\Delta_k)\}\sum_{k=1}^K\{I_{0,l_2}(\Delta_k) - G_{0,l_2}(\Delta_k)\}\right]\\ =& \frac{\sqrt{\sigma_{l_1,l_1}\sigma_{l_2,l_2}}}{\{\sum_{k=1}^K \psi(\Delta_k)\}^2}\sum_{k_1,k_2}\left\{\mathbb{P}\left(\frac{\sqrt{n}(\overline{X}_{j,l_1} - \mu_{l_1})}{\sqrt{\sigma_{l_1,l_1}}} \le \Delta_{k_1}, \frac{\sqrt{n}(\overline{X}_{j,l_2} - \mu_{l_2})}{\sqrt{\sigma_{l_2,l_2}}} \le \Delta_{k_2}\right) - G_{0,l_1}(\Delta_{k_1})G_{0,l_2}(\Delta_{k_2})\right\},\end{aligned}$$

Moreover, we can apply multivariate Berry-Esseen theorem (See Theorem 1.3 in Götze (1991)) and give

$$\widetilde{\mathcal{C}}_{l_1,l_2} = \mathcal{C}_{l_1,l_2} + O(n^{-1/2}),$$

where $\mathcal{C}_{l_1,l_2}$ is defined in (14). Thus the theorem is proved. ■

Proof [Proof of Proposition 9] For the asymptotic normality of multi-dimensional MOM estimator, together with Lemma 17 below and the rate constraint, we can show that

$$\widehat{\mu}_{\mathrm{MOM},l} - \mu_l = \frac{\sqrt{2\pi\sigma_{l,l}}}{(m+1)\sqrt{n}}\sum_{j\notin\mathcal{B}}\{G_{j,l}(0) - I_{j,l}(0)\} + o_{\mathbb{P}}\left(\frac{1}{\sqrt{pmn}}\right).$$

For any vector $|\boldsymbol{v}|_2 = 1$, there is

$$\langle \widehat{\boldsymbol{\mu}}_{\mathrm{MOM}} - \boldsymbol{\mu}, \boldsymbol{v}\rangle = \frac{\sqrt{2\pi}}{(m+1)\sqrt{n}}\sum_{j\notin\mathcal{B}}\sum_{l=1}^p \sqrt{\sigma_{l,l}}\big\{I_{j,l}(\Delta_k) - G_{j,l}(\Delta_k)\big\}v_l + o_{\mathbb{P}}\left(\frac{1}{\sqrt{mn}}\right).$$

Now we apply central limit theorem and yield

$$\frac{\sqrt{(m+1)n}}{\widetilde{\sigma}_{\boldsymbol{v}}}\langle \widehat{\boldsymbol{\mu}}_{\mathrm{MOM}} - \boldsymbol{\mu}, \boldsymbol{v}\rangle \xrightarrow{d} \mathcal{N}(0,1),$$
$$\text{where} \quad \widetilde{\sigma}_{\boldsymbol{v}}^2 = \boldsymbol{v}^{\mathrm{T}}\widetilde{\boldsymbol{\mathcal{C}}}_{\mathrm{MOM}}\boldsymbol{v}.$$

Here $\widetilde{\boldsymbol{\mathcal{C}}}_{\mathrm{MOM}} \in \mathbb{R}^{p\times p}$ has its (l_1, l_2)-entry defined as

$$\begin{aligned}\widetilde{\mathcal{C}}_{\mathrm{MOM},l_1,l_2} =& 2\pi\sqrt{\sigma_{l_1,l_1}\sigma_{l_2,l_2}}\,\mathbb{E}\left[\{I_{0,l_1}(0) - G_{0,l_1}(0)\}\{I_{0,l_2}(0) - G_{0,l_2}(0)\}\right]\\ =& 2\pi\sqrt{\sigma_{l_1,l_1}\sigma_{l_2,l_2}}\left\{\mathbb{P}\left(\frac{\sqrt{n}(\overline{X}_{j,l_1} - \mu_{l_1})}{\sqrt{\sigma_{l_1,l_1}}} \le 0, \frac{\sqrt{n}(\overline{X}_{j,l_2} - \mu_{l_2})}{\sqrt{\sigma_{l_2,l_2}}} \le 0\right) - G_{0,l_1}(0)G_{0,l_2}(0)\right\},\end{aligned}$$

Moreover, we can apply multivariate Berry-Esseen theorem (See Theorem 1.3 in Götze (1991)) and give

$$\widetilde{\mathcal{C}}_{\mathrm{MOM},l_1,l_2} = \mathcal{C}_{\mathrm{MOM},l_1,l_2} + O(n^{-1/2}),$$

where $\mathcal{C}_{\mathrm{MOM},l_1,l_2}$ is defined in (17). Thus the proposition is proved. ■

Lemma 17 *Let $N = (m+1)n$ i.i.d. random variables $X_1, ..., X_N$ evenly distributed in $(m+1)$ subsets $\mathcal{H}_0, \ldots, \mathcal{H}_m$. There is a subset of index $\mathcal{B} \subset \{1, \ldots, m\}$ with $\mathrm{Card}(\mathcal{B}) = \lfloor \alpha_n m \rfloor$. Define*

$$\overline{X}_j = \begin{cases} \frac{1}{n}\sum_{i\in\mathcal{H}_j} X_i & j \notin \mathcal{B}, \\ * & j \in \mathcal{B}, \end{cases} \quad \widehat{\mu} = \mathrm{med}(\overline{X}_j \mid 0 \le j \le m).$$

Suppose X_1 satisfies $\mathbb{E}(X_1) = \mu, \mathrm{Var}(X_1) = \sigma^2$, and $\mathbb{E}|X_1 - \mu|^3 < \infty$. The fraction α_n satisfies $\alpha_n \le 1/2 - \delta$ for some $\delta > 0$. Then $\widehat{\mu}$ admits the following representation:

$$\widehat{\mu} = \mu - \frac{\sqrt{2\pi}\sigma}{(m+1)\sqrt{n}} \sum_{j\notin\mathcal{B}} \left\{ \mathbb{I}(\overline{X}_j \le \mu) - \mathbb{P}(\overline{X}_j \le \mu) \right\} + O_{\mathbb{P}}\left(\frac{\alpha_n}{\sqrt{n}} + \frac{\log n}{m\sqrt{n}} + \frac{\log^{3/4} n}{m^{3/4} n^{1/2}} + \frac{1}{n} \right).$$

Proof Using Taylor expansion we have

$$\begin{aligned} \Phi\left\{ \frac{\sqrt{n}(\widehat{\mu} - \mu)}{\sigma} \right\} - \Phi(0) =& \frac{\sqrt{n}}{\sqrt{2\pi}\sigma}(\widehat{\mu} - \mu) + O(n(\widehat{\mu} - \mu)^2) \\ =& \frac{\sqrt{n}}{\sqrt{2\pi}\sigma}(\widehat{\mu} - \mu) + O_{\mathbb{P}}\left(\alpha_n^2 + \frac{\log n}{m} \right), \end{aligned} \tag{38}$$

since $\widehat{\mu} - \mu = O_{\mathbb{P}}(\frac{\alpha_n}{\sqrt{n}} + \sqrt{\frac{\log n}{mn}})$ by Lemma 16. On the other hand, denote

$$G(x) = \mathbb{P}\left\{ \frac{\sqrt{n}(\overline{X}_1 - \mu)}{\sigma} \le x \right\}, \quad I_j(x) = \mathbb{I}\left\{ \frac{\sqrt{n}(\overline{X}_j - \mu)}{\sigma} \le x \right\}.$$

From Berry-Essen theorem (Lemma 13) we know Lemma 15 is applicable and yields

$$\begin{aligned} & \frac{1}{m+1}\sum_{j=0}^{m} I_j\left\{ \frac{\sqrt{n}(\widehat{\mu} - \mu)}{\sigma} \right\} - G\left\{ \frac{\sqrt{n}(\widehat{\mu} - \mu)}{\sigma} \right\} - \frac{1}{m+1}\sum_{j=0}^{m} I_j(0) + G(0) \\ =& O_{\mathbb{P}}\left(\alpha_n + \frac{\log^{3/4} n}{m^{3/4}} + \frac{1}{\sqrt{n}} \right). \end{aligned}$$

By definition of median we know

$$\Phi(0) = \frac{1}{2} = \frac{1}{m+1}\sum_{j=0}^{m} I_j\left\{ \frac{\sqrt{n}(\widehat{\mu} - \mu)}{\sigma} \right\} + O\left(\frac{1}{m} \right).$$

Thus

$$\begin{aligned} \Phi\left\{ \frac{\sqrt{n}(\widehat{\mu} - \mu)}{\sigma} \right\} - \Phi(0) =& G\left\{ \frac{\sqrt{n}(\widehat{\mu} - \mu)}{\sigma} \right\} - \frac{1}{m+1}\sum_{j=0}^{m} I_j\left\{ \frac{\sqrt{n}(\widehat{\mu} - \mu)}{\sigma} \right\} + O\left(\frac{1}{m} + \frac{1}{\sqrt{n}} \right) \\ =& -\frac{1}{m+1}\sum_{j\notin\mathcal{B}} \{ I_j(0) - G(0) \} + O_{\mathbb{P}}\left(\alpha_n + \frac{\log^{3/4} n}{m^{3/4}} + \frac{1}{\sqrt{n}} \right). \end{aligned} \tag{39}$$

Combining (38) and (39) we have

$$\begin{aligned}\widehat{\mu} =&\mu - \frac{\sqrt{2\pi}\sigma}{(m+1)\sqrt{n}} \sum_{j\notin\mathcal{B}} \{I_j(0) - G(0)\} + O_{\mathbb{P}}\left(\frac{\alpha_n}{\sqrt{n}} + \frac{\log n}{m\sqrt{n}} + \frac{\log^{3/4} n}{m^{3/4}n^{1/2}} + \frac{1}{n}\right)\\ =&\mu - \frac{\sqrt{2\pi}\sigma}{(m+1)\sqrt{n}} \sum_{j\notin\mathcal{B}} \big\{\mathbb{I}(\overline{X}_j \le \mu) - \mathbb{P}(\overline{X}_j \le \mu)\big\} + O_{\mathbb{P}}\left(\frac{\alpha_n}{\sqrt{n}} + \frac{\log n}{m\sqrt{n}} + \frac{\log^{3/4} n}{m^{3/4}n^{1/2}} + \frac{1}{n}\right),\end{aligned}$$

which is exactly what we want to prove. ■

Appendix B. Positive Definiteness of $\mathcal{C}_{\mathrm{MOM}} - \mathcal{C}$

From Theorem 8 and Proposition 9, in order to show that our proposed multi-dimensional VRMOM estimator $\overline{\boldsymbol{\mu}}$ has higher statistical efficiency than the corresponding MOM estimator $\widehat{\boldsymbol{\mu}}$, it is equivalent to prove that $\boldsymbol{\mathcal{C}} \preceq \boldsymbol{\mathcal{C}}_{\mathrm{MOM}}$ holds true. In this appendix, we will verify that the covariance difference $\boldsymbol{\mathcal{C}}_{\mathrm{MOM}} - \boldsymbol{\mathcal{C}}$ is positive definite in dimension 2 as K tends to infinity. First of all, we shall give a general formulation for the entry $\mathcal{C}_{l_1,l_2}$ as $K \to \infty$.

Lemma 18 *Denote $\mathcal{C}^K_{l_1,l_2}$ as the (l_1,l_2)-entry of the matrix $\boldsymbol{\mathcal{C}}$ defined in* (14)*. Then we have*

$$\lim_{K\to\infty} \mathcal{C}^K_{l_1,l_2} = \Big\{4\pi \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi(y_1)\psi(y_2)F_{l_1,l_2}(y_1,y_2)\mathrm{d}y_1\mathrm{d}y_2 - \pi\Big\}\sqrt{\sigma_{l_1,l_1}\sigma_{l_2,l_2}}. \tag{40}$$

In particular, when $l_1 = l_2 = l$, we have

$$\lim_{K\to\infty} \mathcal{C}^K_{l,l} = \frac{\pi}{3}\sigma_{l,l}. \tag{41}$$

Proof For the denominator in (14), we compute that

$$\begin{aligned}\lim_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K} \psi(\Delta_k) =& \lim_{K\to\infty} \frac{K+1}{K}\sum_{k=1}^{K} \psi\Big(\Phi^{-1}\Big(\frac{k}{K+1}\Big)\Big)\frac{1}{K+1}\\ =&\int_0^1 \psi\big(\Phi^{-1}(x)\big)\mathrm{d}x\\ (\text{change variable } x=\Phi(y)) =&\int_{-\infty}^{\infty} \psi^2(y)\mathrm{d}y\\ =&\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-x^2}\mathrm{d}x = \frac{1}{2\sqrt{\pi}}.\end{aligned} \tag{42}$$

For the numerator, on the one hand, we have

$$\begin{aligned}\lim_{K\to\infty} \frac{1}{K^2}\sum_{k_1,k_2=1}^{K} \tau_{k_1}\tau_{k_2} =& \lim_{K\to\infty} \Big(\frac{1}{K}\sum_{k=1}^{K}\tau_k\Big)^2\\ =& \lim_{K\to\infty} \Big(\frac{1}{K}\sum_{k=1}^{K}\frac{k}{K+1}\Big)^2\\ =&\Big(\int_0^1 x\mathrm{d}x\Big)^2 = \frac{1}{4}.\end{aligned} \tag{43}$$

On the other hand, we have that

$$\begin{aligned}\lim_{K\to\infty}\frac{1}{K^2}\sum_{k_1,k_2=1}^{K}\tau_{k_1,k_2}^{l_1,l_2} &= \int_0^1\int_0^1 F_{l_1,l_2}(\Phi^{-1}(x_1),\Phi^{-1}(x_2))\mathrm{d}x_1\mathrm{d}x_2\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\psi(y_1)\psi(y_2)F_{l_1,l_2}(y_1,y_2)\mathrm{d}y_1\mathrm{d}y_2,\end{aligned} \tag{44}$$

where $F_{l_1,l_2}(y_1,y_2)=\mathbb{P}(Z_{l_1}\le y_1, Z_{l_2}\le y_2)$. Combining (42), (43) and (44) we have

$$\lim_{K\to\infty}\mathcal{C}_{l_1,l_2}^{K}=\left\{4\pi\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\psi(y_1)\psi(y_2)F_{l_1,l_2}(y_1,y_2)\mathrm{d}y_1\mathrm{d}y_2-\pi\right\}\sqrt{\sigma_{l_1,l_1}\sigma_{l_2,l_2}}. \tag{45}$$

In particular, when $l_1=l_2=l$, we have

$$F_{l,l}(y_1,y_2)=\Phi(\min(y_1,y_2)).$$

Substitute it in (44), we can obtain

$$\begin{aligned}\lim_{K\to\infty}\frac{1}{K^2}\sum_{k_1,k_2=1}^{K}\tau_{k_1,k_2}^{l,l} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\psi(y_1)\psi(y_2)\Phi(\min(y_1,y_2))\mathrm{d}y_1\mathrm{d}y_2\\ &=2\int_{-\infty}^{\infty}\psi(y_1)\int_{-\infty}^{y_1}\psi(y_2)\Phi(y_2)\mathrm{d}y_2\mathrm{d}y_1\\ &=\int_{-\infty}^{\infty}\psi(y_1)\Phi^2(y_1)\mathrm{d}y_1=\frac{1}{3}.\end{aligned}$$

Therefore we have

$$\lim_{K\to\infty}\mathcal{C}_{l,l}=\frac{1/3-1/4}{1/(4\pi)}\sigma_{l,l}=\frac{\pi}{3}\sigma_{l,l},$$

which completes the proof. ■

Proof [Verification of $\boldsymbol{\mathcal{C}}\preceq\boldsymbol{\mathcal{C}}_{\mathrm{MOM}}$] In the case of dimension 2, we assume the gradient $\nabla f(X,\boldsymbol{\theta}^*)=(\nabla_1 f(X,\boldsymbol{\theta}^*),\nabla_2 f(X,\boldsymbol{\theta}^*))^{\mathrm{T}}$ has covariance matrix

$$\boldsymbol{\Sigma}=\begin{pmatrix}\sigma_{1,1} & \sin\phi\sqrt{\sigma_{1,1}\sigma_{2,2}}\\ \sin\phi\sqrt{\sigma_{1,1}\sigma_{2,2}} & \sigma_{2,2}\end{pmatrix},\quad\text{therefore,}\quad\boldsymbol{\Sigma}_{1,2}=\begin{pmatrix}1 & \sin\phi\\ \sin\phi & 1\end{pmatrix}.$$

From Lemma 18, we have that $\mathcal{C}_{\mathrm{MOM},l,l}-\mathcal{C}_{l,l}=\pi\sigma_{l,l}/6$ as $K\to\infty$, and

$$\begin{aligned}\mathcal{C}_{1,2} =&4\pi\left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\psi(y_1)\psi(y_2)F_{1,2}(y_1,y_2)\mathrm{d}y_1\mathrm{d}y_2-1/4\right\}\sqrt{\sigma_{1,1}\sigma_{2,2}}\\ =&4\pi\left\{\int_{-\infty}^{\infty}\psi(y_2)\mathrm{d}y_2\int_{-\infty}^{y_2}\mathrm{d}x_2\int_{-\infty}^{\infty}\int_{-\infty}^{y_1}\psi(y_1)\psi_\phi(x_1,x_2)\mathrm{d}x_1\mathrm{d}y_1-1/4\right\}\sqrt{\sigma_{1,1}\sigma_{2,2}}\\ =&4\pi\left\{\int_{-\infty}^{\infty}(1-\Phi(x_1))\mathrm{d}x_1\int_{-\infty}^{\infty}\int_{-\infty}^{y_2}\psi(y_2)\psi_\phi(x_1,x_2)\mathrm{d}x_2\mathrm{d}y_2-1/4\right\}\sqrt{\sigma_{1,1}\sigma_{2,2}}\\ =&\left\{4\pi\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Phi(-x_1)\Phi(-x_2)\psi_\phi(x_1,x_2)\mathrm{d}x_1\mathrm{d}x_2-\pi\right\}\sqrt{\sigma_{1,1}\sigma_{2,2}},\end{aligned}$$

where $\psi_\phi(\cdot,\cdot)$ denotes the probability density function of the multivariate normal distribution with covariance matrix $\Sigma_{1,2}$, more precisely, we have

$$\psi_\phi(x_1,x_2)=\frac{1}{2\pi\cos\phi}\exp\Big\{-\frac{x_1^2-2\sin\phi\cdot x_1x_2+x_2^2}{2\cos^2\phi}\Big\}.$$

By symmetry we clearly see that

$$\psi_\phi(x_1,x_2)=\psi_\phi(-x_1,-x_2),\quad \psi_\phi(x_1,-x_2)=\psi_\phi(-x_1,x_2)=\psi_{-\phi}(x_1,x_2),\quad \Phi(x)+\Phi(-x)=1.$$

Then $\mathcal{C}_{1,2}$ can be further simplified as follows

$$\begin{aligned}
&\mathcal{C}_{1,2}/\sqrt{\sigma_{1,1}\sigma_{2,2}}\\
=&4\pi\int_{-\infty}^0\int_{-\infty}^0\big\{\Phi(-x_1)\Phi(-x_2)+\Phi(-x_1)\Phi(x_2)+\Phi(x_1)\Phi(-x_2)+\Phi(x_1)\Phi(x_2)\big\}\psi_\phi(x_1,x_2)\mathrm{d}x_1\mathrm{d}x_2\\
&+4\pi\int_{-\infty}^0\int_{-\infty}^0\big\{\Phi(-x_1)\Phi(x_2)+\Phi(x_1)\Phi(-x_2)\big\}\big\{\psi_\phi(x_1,-x_2)-\psi_\phi(x_1,x_2)\big\}\mathrm{d}x_1\mathrm{d}x_2-\pi\\
=&2\pi\int_{-\infty}^0\int_{-\infty}^0\big\{1-2\Phi(x_2)\big\}\big\{1-2\Phi(x_1)\big\}\big\{\psi_\phi(x_1,x_2)-\psi_{-\phi}(x_1,x_2)\big\}\mathrm{d}x_1\mathrm{d}x_2.
\end{aligned}$$

Similarly, $\mathcal{C}_{\mathrm{MOM},1,2}$ can be written in a similar form

$$\mathcal{C}_{\mathrm{MOM},1,2}/\sqrt{\sigma_{1,1}\sigma_{2,2}}=\pi\int_{-\infty}^0\int_{-\infty}^0\big\{\psi_\phi(x_1,x_2)-\psi_{-\phi}(x_1,x_2)\big\}\mathrm{d}x_1\mathrm{d}x_2=\phi.$$

Therefore, to prove the positive definiteness of the matrix $\boldsymbol{\mathcal{C}}_{\mathrm{MOM}}-\boldsymbol{\mathcal{C}}$, it left to prove that

$$|\mathcal{C}_{\mathrm{MOM},1,2}-\mathcal{C}_{1,2}|/\sqrt{\sigma_{1,1}\sigma_{2,2}}\le|\mathcal{C}_{\mathrm{MOM},1,1}-\mathcal{C}_{1,1}|/\sigma_{1,1}=\frac{\pi}{6}.$$

It is equivalent to the following inequality

$$\begin{aligned}
|h(\phi)|:=&\left|\frac{\phi}{\pi}-2\int_{-\infty}^0\int_{-\infty}^0\big\{1-2\Phi(x_2)\big\}\big\{1-2\Phi(x_1)\big\}\big\{\psi_\phi(x_1,x_2)-\psi_{-\phi}(x_1,x_2)\big\}\mathrm{d}x_1\mathrm{d}x_2\right|\\
\le&\frac{1}{6},
\end{aligned}\tag{46}$$

holds for all $\phi\in[-\frac{\pi}{2},\frac{\pi}{2}]$. In order to show this bound, we can draw the graph of $h(\phi)$ numerically. As shown in Figure 3, we can see that (46) holds true, which implies that $\boldsymbol{\mathcal{C}}\preceq\boldsymbol{\mathcal{C}}_{\mathrm{MOM}}$. ■

Appendix C. Theories and Proofs for RCSL Estimator

This appendix consists of the theoretical results and proofs for the robust CSL estimator. In Appendix C.1, we present the technical assumptions and the main theories for the robust CSL estimator. The proofs of the results will be given in Appendix C.2.

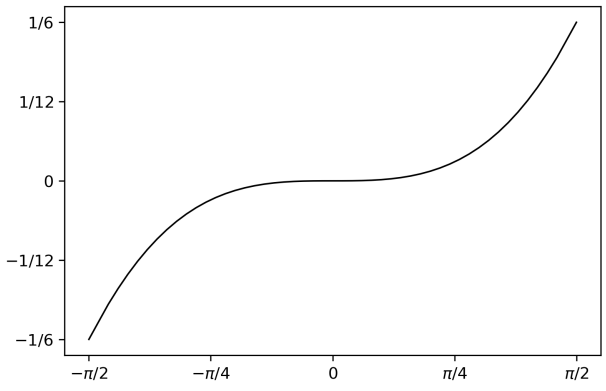

Figure 3: The graph of function $h(\phi)$ on the interval $[-\pi/2, \pi/2]$. We can see that h is increasing and has absolute value uniformly bounded by $1/6$.

C.1 Theoretical Results for Robust CSL Estimator

In this part, we present the theoretical results for the robust CSL estimator. Before that, we first introduce some notations and our technical assumptions.

C.1.1 Notations and technical assumptions

As we will demonstrate in Appendix D, all the technical assumptions hold on common statistical models, which suggests the wide applicability of these assumptions. For the loss function $f(x, \boldsymbol{\theta})$, we assume that $f(x, \boldsymbol{\theta})$ is differentiable with respect to $\boldsymbol{\theta}$ and denote $\nabla f(x, \boldsymbol{\theta}) := \nabla|_{\boldsymbol{\theta}} f(x, \boldsymbol{\theta})$ as the gradient of $f(x, \boldsymbol{\theta})$ at $\boldsymbol{\theta}$. For a given $\boldsymbol{\theta}$, we define the expected gradient $\boldsymbol{\mu}(\boldsymbol{\theta})$ and population standard deviation of the gradient for each coordinate l, $\sigma_l(\boldsymbol{\theta})$, as follows,

$$\begin{aligned} &\boldsymbol{\mu}(\boldsymbol{\theta}) = (\mu_1(\boldsymbol{\theta}), \ldots \mu_p(\boldsymbol{\theta}))^{\mathrm{T}} := \mathbb{E}_{X\sim\mathfrak{X}}\{\nabla f(X, \boldsymbol{\theta})\}, \\ &\sigma_l^2(\boldsymbol{\theta}) := \mathbb{E}_{X\sim\mathfrak{X}}\left[\{\nabla f_l(X, \boldsymbol{\theta}) - \mu_l(\boldsymbol{\theta})\}^2\right], \text{ for } 1 \le l \le p. \end{aligned} \tag{47}$$

For notational simplicity, we will denote the expectation taken over the randomness of X by $\mathbb{E} := \mathbb{E}_{X\sim\mathfrak{X}}$. Recall that $\boldsymbol{\theta}^*$ is the true parameter that minimizes the loss function $\mathbb{E}f(X, \boldsymbol{\theta})$. Throughout the paper, we assume that $\boldsymbol{\mu}(\boldsymbol{\theta}^*) = \mathbb{E}\{\nabla f(X, \boldsymbol{\theta}^*)\} = \mathbf{0}$, which holds as long as the expectation and ∇ can be interchanged. Finally, for the ease of presentation, some parameter-independent constants such as ρ, η, C_M will be used across different assumptions when there is no confusion.

Assumption A *For every fixed $x \in \mathfrak{X}$, $f(x, \boldsymbol{\theta})$ is a convex function of $\boldsymbol{\theta}$ on $\mathbb{R}^p$.*

Assumption B *There exists a constant $\rho > 0$ such that*

$$\rho \le \Lambda_{\min}\left[\nabla \boldsymbol{\mu}(\boldsymbol{\theta}^*)\right] \le \Lambda_{\max}\left[\nabla \boldsymbol{\mu}(\boldsymbol{\theta}^*)\right] \le \rho^{-1}.$$

Assumption C *The Hessian of the population loss $\nabla\boldsymbol{\mu}$ is Lipschitz continuous. In particular, there exists a constant $C_H > 0$ such that*

$$\| \nabla\boldsymbol{\mu}(\boldsymbol{\theta}_1) - \nabla\boldsymbol{\mu}(\boldsymbol{\theta}_2) \| \leq C_H|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2|_2,$$

holds for arbitrary $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^p$.

Assumption D *For every $\boldsymbol{v} \in \mathbb{S}^{p-1}$ and $x \in \mathfrak{X}$, define*

$$M_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}(x,\boldsymbol{v}) := \frac{|\langle \boldsymbol{v}, \nabla f(x,\boldsymbol{\theta}_1) - \nabla f(x,\boldsymbol{\theta}_2)\rangle|}{|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2|_2},$$
$$\overline{M}(x,\boldsymbol{v}) := \sup_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2 \in \mathbb{R}^p} M_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}(x,\boldsymbol{v}).$$

There exists $C_M, \gamma_0, \eta > 0$ such that

$$\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \sup_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2} \mathbb{E}\left[\exp\{\eta M_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}(X,\boldsymbol{v})\}\right] \leq C_M, \tag{48a}$$

$$\mathbb{E}\left[\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \exp\left\{p^{-\gamma_0}\overline{M}(X,\boldsymbol{v})\right\}\right] \leq C_M. \tag{48b}$$

Assumption E *There exists a constant $\rho > 0$ such that*

$$\rho \leq \min_{1\leq l\leq p}\{\sigma_l(\boldsymbol{\theta}^*)\} \leq \max_{1\leq l\leq p}\{\sigma_l(\boldsymbol{\theta}^*)\} \leq \rho^{-1},$$

where $\sigma_l(\boldsymbol{\theta}^*)$ is defined in (47)*.*

Assumption F *There exists $\eta > 0$ such that*

$$\max_{1\leq l\leq p} \mathbb{E}\left[\exp\left\{\eta\left|\nabla f_l(X,\boldsymbol{\theta}^*) - \mu_l(\boldsymbol{\theta}^*)\right|\right\}\right] \leq C_M. \tag{49}$$

Assumption G *The number of machines m, distributed sample size n, dimension of parameters p, and convergence rate r_n of the initial estimator $\widehat{\boldsymbol{\theta}}^{(0)}$ (i.e. $|\widehat{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\theta}^*|_2 = O_{\mathbb{P}}(r_n)$) satisfy the following relationships*

$$m = o(n), \quad p = O\left(\frac{n^{1/3}}{\sqrt{\log n}}\right), \quad r_n = O\left(\min\left\{\frac{1}{\log n}, \frac{1}{\sqrt{p\log n}}\right\}\right). \tag{50}$$

Assumptions A and B assume the convexity of the loss and the local strong convexity of the population loss function around $\boldsymbol{\theta}^*$, which are commonly assumed in empirical risk minimization literature. Assumptions C and D are the standard smoothness assumptions, which appear in distributed learning literature (Zhang et al., 2013; Jordan et al., 2019). In particular, the Lipschitz gradient assumption D is represented by two formulas (48a) and (48b) and the (48b) mainly handles the diverging dimensionality, which allows p to go to infinity. Similar conditions can be found in Chen et al. (2021) and Su and Xu (2019). An additional remark is that, our smoothness assumptions C and D are relatively weaker than those in existing literature. For example, the Assumption PD in Jordan et al. (2019) requires Lipschitz continuity of the second-order derivatives of the loss function $f(X,\boldsymbol{\theta})$.

In contrast, we only require the expectation of the loss function $\mathbb{E}(f(X,\boldsymbol{\theta}))$ to be second-order differentiable and Lipschitz continuous (Assumption C), and allow the gradient ∇f to be non-differentiable. Thus, our theoretical framework handles the Huber loss function as shown in Example 2, while Jordan et al. (2019) can not.

Assumptions E and F guarantee concentrating properties for coordinate-wise gradient variance. In Assumption F we assume the gradients to be sub-exponential, which is weaker than the boundedness condition in Alistarh et al. (2018) and the sub-gaussian condition in Yin et al. (2019). The sub-exponential condition is also assumed in Chen et al. (2017) and Su and Xu (2019). We note that the gradients can be sub-exponential in the case of least square regression (see Example 1 below for more details), which brings additional technical challenges in establishing concentration inequalities. In particular, Yin et al. (2018) imposed bounded absolute skewness (the third-order moment) condition, which is weaker than ours. However, their theory did not consider diverging dimension p.

The rate constraints on the quantities m, n, p, r_n are given in Assumption G. The relationships on m are inherited from Theorems 5 by letting $\kappa \geq 2$. The condition on p indicates that the dimension cannot diverge too fast. The final condition on r_n ensures that the initial estimator is consistent. It is worth noting that the constraint on the initial rate r_n is attainable. By definition, the initial estimator $\widehat{\boldsymbol{\theta}}^{(0)}$ is the minimizer of the local empirical loss function (22). From the regularity assumptions A–F, it is not hard to show that $|\widehat{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\theta}^*|_2 = O_{\mathbb{P}}(\sqrt{p \log n/n})$. Plug in the constraint on dimension p in Assumption G, we have that $|\widehat{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\theta}^*|_2 = O_{\mathbb{P}}(n^{-1/4})$. On the other hand, we can easily verify that $n^{-1/4} = O(\min\{1/\log n, 1/\sqrt{p \log n}\})$. Therefore Assumption G can be satisfied.

We provide the following two examples for better understanding of the proposed assumptions. In Appendix D, we will verify that Assumptions A–F hold for generalized linear models and a large class of M-estimators.

Example 1 *(Linear regression) In linear regression model $Y = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta}^* + \epsilon$, we assume that the covariate $\boldsymbol{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$ and the noise ϵ are both sub-gaussian random variables. Define $\boldsymbol{\xi} = (Y, \boldsymbol{X}^{\mathrm{T}})^{\mathrm{T}}$ and the loss function $f(\boldsymbol{\xi}, \boldsymbol{\theta}) = (Y - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})^2$. Then we can compute the gradient of the loss function as*

$$\nabla f(\boldsymbol{\xi}, \boldsymbol{\theta}^*) = \epsilon \boldsymbol{X}.$$

In the l-th coordinate, the gradient $\nabla f_l(\boldsymbol{\xi}, \boldsymbol{\theta}^*) = \epsilon X_l$ is a product of two sub-gaussian random variables, hence is sub-exponential (See Proposition 2.7.1 of Vershynin (2018)).

Example 2 *(Huber regression) In Huber regression model $Y = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta}^* + \epsilon$, similarly define $\boldsymbol{\xi} = (Y, \boldsymbol{X}^{\mathrm{T}})^{\mathrm{T}}$. The loss function is constructed as $f(\boldsymbol{\xi}, \boldsymbol{\theta}) = \mathcal{L}(Y - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})$, where*

$$\mathcal{L}(x) = \begin{cases} x^2/2 & \text{for } |x| \leq \delta, \\ \delta(|x| - \delta/2) & \text{otherwise.} \end{cases}$$

Then we can compute that

$$\mathcal{L}'(x) = \begin{cases} x & \text{for } |x| \leq \delta, \\ \delta \operatorname{sign}(x) & \text{otherwise,} \end{cases} \qquad \mathcal{L}''(x) = \mathbb{I}(|x| \leq \delta).$$

In this case, the Hessian $\nabla^2 f(\boldsymbol{\xi}, \boldsymbol{\theta}) = \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\mathbb{I}(|Y - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta}| \le \delta)$ *is not continuous with respect to the parameter* $\boldsymbol{\theta}$*. However, if we assume the noise* ϵ *has a symmetric distribution and uniformly bounded probability density function, we can prove that Huber regression model fulfills the proposed assumptions. The detailed verification is delegated to Appendix D.*

C.1.2 Theoretical results

In this part, we provide the main theorems for the proposed RCSL estimator. We firstly provide single round convergence rate of RCSL estimator. To show the superiority in statistical efficiency of our VRMOM-based RCSL method, we present the asymptotic normality for our RCSL method and the MOM-based RCSL method. Then we will show that the VRMOM-RCSL method has smaller asymptotic variance than the MOM-based counter part.

Firstly, we present our estimation result for one round of communication in the following theorem, which helps understand the improvement from the initial estimator for only one iteration.

Theorem 19 (One-round convergence rate of the RCSL method) *Suppose Assumptions A-G hold and the initial estimator* $\widehat{\boldsymbol{\theta}}^{(0)}$ *satisfies* $|\widehat{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\theta}^*|_2 = O_{\mathbb{P}}(r_n)$*. Further assume the fraction* α_n *of Byzantine machines satisfies* $\alpha_n \le 1/2 - \delta$ *for some fixed* $\delta \in (0, 1/2)$*. Then the robust CSL estimator* $\widehat{\boldsymbol{\theta}}^{(1)}$ *defined in (19) satisfies*

$$|\widehat{\boldsymbol{\theta}}^{(1)} - \boldsymbol{\theta}^*|_2 = O_{\mathbb{P}}\left(\frac{\alpha_n\sqrt{p}}{\sqrt{n}} + \sqrt{\frac{p\log n}{mn}} + \frac{p^{1/2}\log^{3/4} n}{n^{1/2}m^{3/4}} + r_n\sqrt{\frac{p^2\log n}{n}}\right). \tag{51}$$

Applying Theorem 19 inductively, we can obtain the convergence result for our RCSL estimator with t rounds of aggregations, which is presented in Theorem 11 in Section 3 of the main paper.

Moreover, similar as Theorem 8 in the main paper, we can establish asymptotic normality for our RCSL estimator $\widehat{\boldsymbol{\theta}}^{(t)}$, which has not been studied in previous robust distributed learning literature. To save symbols and avoid repeated definitions, we denote $\sigma_{l_1,l_2} = Cov\{\nabla f_{l_1}(X, \boldsymbol{\theta}^*), \nabla f_{l_2}(X, \boldsymbol{\theta}^*)\}$ to be the (l_1, l_2)-entry of covariance matrix of $\nabla f(X, \boldsymbol{\theta}^*)$, which is coincident with the notations in Theorem 8. Then we can define $\boldsymbol{\mathcal{C}}$ exactly the same as in (13) and (14). Then we can prove the following asymptotic normality result:

Theorem 20 (Asymptotic normality of the RCSL method) *Suppose Assumptions A-G hold, and additionally, we assume the rate constraints* $p = o(\min\{\frac{n^{1/3}}{\log^{2/3} n}, \frac{m^{1/2}}{\log^{3/2} n}, \frac{n}{m}\})$, $\alpha_n = o(1/\sqrt{mp})$*, and* $\log^3 n = o(m)$*. Then for every iteration number* t *satisfies* (24) *and any vector* $\boldsymbol{v} \in \mathbb{R}^p$ *with* $|\boldsymbol{v}|_2 = 1$*, we have that*

$$\frac{\sqrt{N}}{\sigma_{\boldsymbol{v}}}\left\langle \boldsymbol{v}, \widehat{\boldsymbol{\theta}}^{(t+1)} - \boldsymbol{\theta}^* \right\rangle \xrightarrow{d} \mathcal{N}(0, 1), \tag{52}$$

as $n \to \infty$*, where*

$$\sigma_{\boldsymbol{v}}^2 = \boldsymbol{v}^{\mathrm{T}}\{\nabla\boldsymbol{\mu}(\boldsymbol{\theta}^*)\}^{-1}\boldsymbol{\mathcal{C}}\{\nabla\boldsymbol{\mu}(\boldsymbol{\theta}^*)\}^{-1}\boldsymbol{v}. \tag{53}$$

Compared with the constraints in Theorem 8, both of the theorems require the fraction $\alpha_n = o(1/\sqrt{mp})$, namely, the number of Byzantine machines is $o(\sqrt{m/p})$. However, the normality result of the RCSL method needs more restrictive constraints on the dimension p than the one in Assumption G and in Theorem 8.

As we can see from (53), the asymptotic variance of the proposed RCSL estimator has a sandwich structure, which is commonly appeared in literatures (see, *e.g.* , Polyak and Juditsky (1992); Jordan et al. (2019); Chen et al. (2020a)). However, since the past works only aggregate the gradients by sample mean, the centered covariance matrix in (53) is usually the covariance of the gradient, namely, $\boldsymbol{\mathcal{C}} = \mathbb{E}\{\nabla f(X, \boldsymbol{\theta}^*)\nabla f(X, \boldsymbol{\theta}^*)^{\mathrm{T}}\}$. In contrast, as we aggregate the gradient robustly by our proposed multivariate VRMOM estimator, the structure of the matrix $\boldsymbol{\mathcal{C}}$ is much more complex, as we can see in (13) and (14). To the best of our knowledge, this is the first asymptotic normality result in the setting of Byzantine-robust distributed learning.

In order to illustrate the efficiency of our RCSL method, it is possible to prove an asymptotic normality result for the median-of-means (MOM) based RCSL method, which is named MOM-RCSL. The explicit construction of the MOM-RCSL method is described in the paragraph before Section 4.2.1. We can show that the asymptotic variance of the MOM-RCSL estimator has the same formulation as (53), with $\boldsymbol{\mathcal{C}}$ replaced by $\boldsymbol{\mathcal{C}}_{\mathrm{MOM}}$ in (17) of Proposition 9. The proof technique is simply the combination of Proposition 9 and Theorem 20. Therefore, we omit the presentation of this parallel result for brevity. The simulation results of comparison between our RCSL method and the MOM-RCSL method are already presented in Section 4.2.

C.2 Proofs of the Results for Robust CSL Estimator

C.2.1 More Technical Lemmas

In the following, we introduce additional lemmas that will be used for proofing the results related to RCSL estimator. For consistency with Assumption F, we assume the random variables admits sub-exponential tail.

Lemma 21 *(Quantile gap of median-of-mean) Let $N(=(m+1)n)$ i.i.d. random variables $X_1, ..., X_N$ evenly distributed in m subsets $\mathcal{H}_0, ..., \mathcal{H}_m$. Let $\overline{X}_j = n^{-1}\sum_{i\in\mathcal{H}_j} X_i$ and $\widehat{X}^\tau$ be the τ-th quantile of $\{\overline{X}_0, \ldots, \overline{X}_m\}$. Suppose $\mathbb{E}(X_1) = 0, \mathrm{Var}(X_1) = \sigma^2$, and $\mathbb{E}[e^{\eta|X_1|}] < \infty$ for some $\eta > 0$. There are two quantile levels $\tau_2 > \tau_1$ satisfying $|\tau_2 - \tau_1| = o(1)$ and $\tau_1, \tau_2 \in (\delta, 1-\delta)$ for some $\delta \in (0, 1/2)$. Then for every $\gamma > 1$, there exists $\widetilde{C}$ such that*

$$\mathbb{P}\left\{\widehat{X}^{\tau_2} - \widehat{X}^{\tau_1} \geq \widetilde{C}\left(\frac{\tau_2 - \tau_1}{\sqrt{n}} + \frac{1}{n} + \frac{\log n}{m\sqrt{n}}\right)\right\} = O(n^{-\gamma}).$$

Proof Follow the proof of Lemma 16, we can show that

$$\mathbb{P}\left\{\left|\widehat{X}^{\tau_1}\right| \leq \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{1}{2} + \left|\tau_1 - \frac{1}{2}\right| + \frac{C_1}{\sqrt{n}} + C_1\sqrt{\frac{\log n}{m}}\right)\right\} \geq 1 - O(n^{-\gamma}), \tag{54}$$

holds for some C_1 large enough. Denote

$$\delta_n := \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{1}{2} + \left|\tau_1 - \frac{1}{2}\right| + \frac{C_1}{\sqrt{n}} + C_1\sqrt{\frac{\log n}{m}}\right).$$

Evenly divide the interval $[-\delta_n, \delta_n]$ into $2n$ pieces and define the set $\mathfrak{N} := \{-\delta_n, -\frac{n-1}{n}\delta_n, \ldots, \delta_n\}$. Similar as proof in Lemma 15, we can find some $C_2 > 0$ such that for every $\widetilde{x} \in \mathfrak{N}$, there is

$$\begin{aligned}
&\frac{1}{m+1}\sum_{j=0}^{m}\sup_{\{x:|\widetilde{x}-x|\le\delta_n/n\}}\left|\mathbb{I}(\overline{X}_j \le x+y) - \mathbb{I}(\overline{X}_j \le x) - \mathbb{I}(\overline{X}_j \le \widetilde{x}+y) + \mathbb{I}(\overline{X}_j \le \widetilde{x})\right| \\
\le&\frac{1}{m+1}\sum_{j=0}^{m}\mathbb{I}\left(|\overline{X}_j - \widetilde{x} - y| \le \frac{\delta_n}{n}\right) + \frac{1}{m+1}\sum_{j=0}^{m}\mathbb{I}\left(|\overline{X}_j - \widetilde{x}| \le \frac{\delta_n}{n}\right) \le C_2\left(\frac{\log n}{m} + \frac{1}{\sqrt{n}}\right),
\end{aligned} \tag{55}$$

holds with probability $1 - O(n^{-\gamma-1})$. For $0 \le j \le m$, further define the random variable

$$Y_j(x,y) := \mathbb{I}(\overline{X}_j \le x+y) - \mathbb{I}(\overline{X}_j \le x) - \mathbb{P}(\overline{X}_j \le x+y) + \mathbb{P}(\overline{X}_j \le x).$$

For every $x \in [-\delta_n, \delta_n]$, there exist constants C_3, C_4 such that

$$\begin{aligned}
&\mathbb{P}\left[\frac{1}{m+1}\sum_{j=0}^{m}\left\{\mathbb{I}(\overline{X}_j \le x+y) - \mathbb{I}(\overline{X}_j \le x)\right\} \le \tau_2 - \tau_1 + \frac{C_2\log n}{m} + \frac{C_2}{\sqrt{n}}\right] \\
=&\mathbb{P}\left[\frac{1}{m+1}\sum_{j=0}^{m}Y_j(x,y) \le \tau_2 - \tau_1 + \frac{C_2\log n}{m} + \frac{C_2}{\sqrt{n}} - \mathbb{P}(\overline{X}_1 \le x+y) + \mathbb{P}(\overline{X}_1 \le x)\right] \\
\le&\mathbb{P}\left[\frac{1}{m+1}\sum_{j=0}^{m}Y_j(x,y) \le \tau_2 - \tau_1 + \frac{C_2\log n}{m} + \frac{C_2+C_3}{\sqrt{n}} - \Phi\left(\frac{\sqrt{n}(x+y)}{\sigma}\right) + \Phi\left(\frac{\sqrt{n}x}{\sigma}\right)\right] \\
\le&\mathbb{P}\left[\frac{1}{m+1}\sum_{j=0}^{m}Y_j(x,y) \le -C_4\left(\tau_2 - \tau_1 + \frac{\log n}{m} + \frac{1}{\sqrt{n}}\right)\right],
\end{aligned}$$

by finding some $y > 0$ such that

$$\min_{x\in[-\delta_n,\delta_n]}\left\{\Phi\left(\frac{\sqrt{n}(x+y)}{\sigma}\right) - \Phi\left(\frac{\sqrt{n}x}{\sigma}\right)\right\} \ge (C_4+1)(\tau_2-\tau_1) + \frac{C_3}{\sqrt{n}} + (C_2+C_4)\left(\frac{\log n}{m} + \frac{1}{\sqrt{n}}\right). \tag{56}$$

Taking

$$y_0 = \frac{\sigma}{\sqrt{n}\psi(\Phi^{-1}(\delta/2))}\left\{(C_4+1)(\tau_2-\tau_1) + \frac{C_3}{\sqrt{n}} + (C_2+C_4)\left(\frac{\log n}{m} + \frac{1}{\sqrt{n}}\right)\right\},$$

since $\tau_2 - \tau_1 = o(1)$, and m, n tends to infinity, we can assume

$$\frac{\sqrt{n}(x+y_0)}{\sigma}, \frac{\sqrt{n}x}{\sigma} \in \left(\Phi^{-1}(\delta/2), \Phi^{-1}(1-\delta/2)\right),$$

always hold. Thus by applying mean value theorem to continuous function $\Phi(x)$, (56) can be guaranteed. Further compute that

$$\sup_{x\in[-\delta_n,\delta_n]}\mathfrak{E}\left\{1, Y_j(x,y_0)\right\} = O\left(\tau_2 - \tau_1 + \frac{\log n}{m} + \frac{1}{\sqrt{n}}\right).$$

From Lemma 14 we have

$$\begin{aligned}&\mathbb{P}\left[\frac{1}{m+1}\sum_{j=0}^{m}\left\{\mathbb{I}(\overline{X}_j\le x+y_0)-\mathbb{I}(\overline{X}_j\le x)\right\}\le\tau_2-\tau_1+\frac{C_2\log n}{m}+\frac{C_2}{\sqrt{n}}\right]\\ \le&\mathbb{P}\left[\frac{1}{m+1}\sum_{j=0}^{m}Y_j(x,y_0)\le -C_4\left(\tau_2-\tau_1+\frac{\log n}{m}+\frac{1}{\sqrt{n}}\right)\right]=O(n^{-\gamma-1}).\end{aligned}\tag{57}$$

Combining (54),(55) and (57), finally we have

$$\begin{aligned}&\mathbb{P}\left[\widehat{X}^{\tau_2}-\widehat{X}^{\tau_1}\ge y_0\right]\\ \le&\mathbb{P}\left[\sup_{x\in[-\delta_n,\delta_n]}\sum_{j=0}^{m}\left\{\mathbb{I}(\overline{X}_j\le x+y_0)-\mathbb{I}(\overline{X}_j\le x)\right\}\le(\tau_2-\tau_1)(m+1)\right]+O(n^{-\gamma})\\ \le&2n\max_{\widetilde{x}\in\mathfrak{N}}\mathbb{P}\left[\frac{1}{m+1}\sum_{j=0}^{m}\left\{\mathbb{I}(\overline{X}_j\le\widetilde{x}+y_0)-\mathbb{I}(\overline{X}_j\le\widetilde{x})\right\}\le\tau_2-\tau_1+\frac{C_2\log n}{m}+\frac{C_2}{\sqrt{n}}\right]\\ &+2n\max_{\widetilde{x}\in\mathfrak{N}}\mathbb{P}\Big[\frac{1}{m+1}\sum_{j=0}^{m}\sup_{\{x:|\widetilde{x}-x|\le\delta_n/n\}}\left|\mathbb{I}(\overline{X}_j\le x+y_0)-\mathbb{I}(\overline{X}_j\le x)-\mathbb{I}(\overline{X}_j\le\widetilde{x}+y_0)+\mathbb{I}(\overline{X}_j\le\widetilde{x})\right|\\ &\ge\frac{C_2\log n}{m}+\frac{C_2}{\sqrt{n}}\Big]+O(n^{-\gamma})\\ =&O(n^{-\gamma}),\end{aligned}$$

therefore prove the lemma. ■

Lemma 22 *(Stability of quantile with Byzantine machines) Let $X_0,...,X_m$ be fixed points, and $\mathcal{B}\subset\{1,\dots,m\}$ be a subset of index with $\mathrm{Card}(\mathcal{B})=\lfloor\alpha m\rfloor$, where $0<\alpha<1/2$. Impose perturbation ϵ_j on each X_j such that $|\epsilon_j|\le\delta$ for $j\notin\mathcal{B}$, and ϵ_j can be arbitrary for $j\in\mathcal{B}$. Denote the τ'th quantile of the three sets $\{X_0,\dots,X_m\},\{X_0+\epsilon_0,\dots,X_m+\epsilon_m\},\{X_j\mid j\notin\mathcal{B}\}$ as $\widehat{X}^{\tau},\widehat{X}^{\tau}_{\epsilon}$ and $\widehat{X}^{\mathcal{B},\tau}$ respectively, Then there is*

$$|\widehat{X}^{\tau}_{\epsilon}-\widehat{X}^{\tau}|\le\left|\widehat{X}^{\mathcal{B},(\tau-\alpha)/(1-\alpha)}-\widehat{X}^{\mathcal{B},\tau/(1-\alpha)}\right|+2\delta.$$

Proof By definition of quantiles, $\widehat{X}^{\tau}_{\epsilon}$ satisfies the following inequalities

$$\sum_{j=0}^{m}\mathbb{I}\left(X_j+\epsilon_j\le\widehat{X}^{\tau}_{\epsilon}\right)\ge\tau(m+1),\quad\sum_{j=0}^{m}\mathbb{I}\left(X_j+\epsilon_j\ge\widehat{X}^{\tau}_{\epsilon}\right)\ge(1-\tau)(m+1).$$

Therefore we have

$$\begin{aligned}&\sum_{j\notin\mathcal{B}}\mathbb{I}\left(X_j+\epsilon_j\le\widehat{X}^{\tau}_{\epsilon}\right)\ge(\tau-\alpha)(m+1),\quad\sum_{j\notin\mathcal{B}}\mathbb{I}\left(X_j+\epsilon_j\ge\widehat{X}^{\tau}_{\epsilon}\right)\ge(1-\tau-\alpha)(m+1),\\ \Rightarrow&\sum_{j\notin\mathcal{B}}\mathbb{I}\left(X_j\le\widehat{X}^{\tau}_{\epsilon}+\delta\right)\ge(\tau-\alpha)(m+1),\quad\sum_{j\notin\mathcal{B}}\mathbb{I}\left(X_j\ge\widehat{X}^{\tau}_{\epsilon}-\delta\right)\ge(1-\tau-\alpha)(m+1).\end{aligned}$$

This implies that

$$\widehat{X}^{\mathcal{B},(\tau-\alpha)/(1-\alpha)} \leq \widehat{X}^{\tau}_{\epsilon} + \delta, \quad \widehat{X}^{\mathcal{B},\tau/(1-\alpha)} \geq \widehat{X}^{\tau}_{\epsilon} - \delta.$$

Similarly we can show

$$\widehat{X}^{\mathcal{B},(\tau-\alpha)/(1-\alpha)} \leq \widehat{X}^{\tau} \leq \widehat{X}^{\mathcal{B},\tau/(1-\alpha)}.$$

So we have $|\widehat{X}^{\tau}_{\epsilon} - \widehat{X}^{\tau}| \leq \left|\widehat{X}^{\mathcal{B},(\tau-\alpha)/(1-\alpha)} - \widehat{X}^{\mathcal{B},\tau/(1-\alpha)}\right| + 2\delta.$ ■

Lemma 23 *(Exponential concentration of variance) Let $X_1, ..., X_n$ be i.i.d. random variables with $\mathbb{E}(X_1) = 0, \mathrm{Var}(X_1) = \sigma^2$, and $\mathbb{E}[e^{\eta|X_1|}] \leq C$ for some $\eta, C > 0$. Construct sample mean $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$ and sample variance $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$. Then for every $\gamma \geq 1$, there exists $\widetilde{C}$ large enough, such that*

$$\mathbb{P}\left(|\widehat{\sigma} - \sigma| \geq \widetilde{C}\sqrt{\frac{\log n}{n}}\right) = O(n^{-\gamma}).$$

Proof Define

$$\begin{aligned}
\overline{Y}_i &:= X_i^2 \mathbb{I}\left\{X_i^2 \leq C_1^2(\log n)^2\right\} - \mathbb{E}\left[X_i^2 \mathbb{I}\left\{X_i^2 \leq C_1^2(\log n)^2\right\}\right], \\
\widetilde{Y}_i &:= X_i^2 - \sigma^2 - \overline{Y}_i.
\end{aligned}$$

We can compute that

$$\begin{aligned}
\mathbb{E}\left[X_i^2 I\left\{X_i^2 \geq C_1^2(\log n)^2\right\}\right] &\leq \int_{C_1 \log n}^{\infty} 2s\mathbb{P}(|X_i| \geq s)\mathrm{d}s + C_1^2(\log n)^2\mathbb{P}(|X_i| > C_1 \log n) \\
&\leq \int_{C_1 \log n}^{\infty} 2Cse^{-\eta s}\mathrm{d}s + CC_1^2(\log n)^2 n^{-\eta C_1} \\
&= \left\{2\eta^{-1}CC_1 \log n + 2\eta^{-2}C + CC_1^2(\log n)^2\right\} n^{-\eta C_1} \leq n^{-\eta C_1 + 1},
\end{aligned}$$

for n sufficiently large. Then

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{Y}_i\right| \geq 2n^{-\eta C_1+1}\right) \\
\leq &\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n}|X_i|^2\mathbb{I}\left\{|X_i|^2 > C_1^2(\log n)^2\right\}\right| \geq n^{-\eta C_1+1}\right] \\
\leq &\mathbb{P}\left\{\max_{1\leq i\leq n}|X_i|^2 \geq C_1^2(\log n)^2\right\} \\
\leq &n \max_{1\leq i\leq n}\mathbb{P}\left\{|X_i| \geq C_1 \log n\right\} \leq Cn^{-\eta C_1+1}.
\end{aligned}$$

Take $C_1 \geq \eta^{-1}(\gamma + 1)$ we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{Y}_i \geq 2n^{-\gamma}\right) = O(n^{-\gamma}). \tag{58}$$

Next we apply Bernstein's inequality Bennett (1962) for bounded random variables $\overline{Y}_i$. Together with the elementary inequality $\mathbb{E}(\overline{Y}_1^2) \leq \mathbb{E}(X_1^4) \leq 5\eta^{-4}\mathbb{E}\{\exp(\eta|X_1|)\}$, we have

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\overline{Y}_i \geq x\right) \leq & \exp\left\{-\frac{3nx^2}{6\mathbb{E}(\overline{Y}_i^2)+2C_1^2(\log n)^2 x}\right\} \\ \leq & \exp\left[-\min\left\{\frac{\eta^4 nx^2}{20C}, \frac{3nx}{4C_1^2(\log n)^2}\right\}\right] = O(n^{-\gamma}),\end{aligned} \tag{59}$$

by taking $x \geq \eta^{-2}\sqrt{20\gamma C}\sqrt{\frac{\log n}{n}}$. Combining (58) and (59), we have proved

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \sigma^2 \geq \eta^{-2}\sqrt{20\gamma C}\sqrt{\frac{\log n}{n}} + 2n^{-\gamma}\right) \\ \leq &\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}\overline{Y}_i \geq \eta^{-2}\sqrt{20\gamma C}\sqrt{\frac{\log n}{n}}\right\} + \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{Y}_i \geq 2n^{-\gamma}\right) = O(n^{-\gamma}).\end{aligned} \tag{60}$$

On the other hand, using the fact $x_0^2 \leq e^{|x_0|}$ we have

$$\mathfrak{E}\left(\frac{\eta}{2}, X_1\right) = \frac{4}{\eta^2}\mathbb{E}\left\{\frac{\eta^2}{4}X_1^2\exp\frac{\eta}{2}|X_1|\right\} \leq \frac{4}{\eta^2}\mathbb{E}\left(e^{\eta|X_1|}\right) \leq \frac{4C}{\eta^2}.$$

Then Lemma 14 yields

$$\mathbb{P}\left\{\left|\overline{X}\right| \geq (\frac{4}{\eta^2}+1)\sqrt{\gamma C}\sqrt{\frac{\log n}{n}}\right\} = O(n^{-\gamma}). \tag{61}$$

Combining (60) and (61) we have

$$\begin{aligned}&\mathbb{P}\left(|\widehat{\sigma}-\sigma| \geq \widetilde{C}\sqrt{\frac{\log n}{n}}\right) \leq \mathbb{P}\left(\left|\widehat{\sigma}^2-\sigma^2\right| \geq \sigma\widetilde{C}\sqrt{\frac{\log n}{n}}\right) \\ \leq &\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i^2-\sigma^2\right| + \overline{X}^2 \geq \sigma\widetilde{C}\sqrt{\frac{\log n}{n}}\right) = O(n^{-\gamma}),\end{aligned}$$

by letting $\widetilde{C}$ large enough. ■

Lemma 23 directly implies that, under Assumption E, F and G, there is

$$\mathbb{P}\left\{\max_{1\leq l\leq p}|\widehat{\sigma}_l(\boldsymbol{\theta}^*)-\sigma_l(\boldsymbol{\theta}^*)| \geq \widetilde{C}\sqrt{\frac{\log n}{n}}\right\} = O(n^{-\gamma}), \tag{62}$$

holds every $\gamma > 1$.

Lemma 24 *(Uniform bound of variance difference) Under Assumption D, E, F and G, there exists $\widetilde{C}$ large enough, such that*

$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta}\in\Theta_0}\max_{1\leq l\leq p}|\widehat{\sigma}_l(\boldsymbol{\theta})-\widehat{\sigma}_l(\boldsymbol{\theta}^*)| \geq \widetilde{C}r_n\right\} = O(n^{-\gamma}),$$

where $\Theta_0 = \{\boldsymbol{\theta} : |\boldsymbol{\theta}-\boldsymbol{\theta}^*|_2 \leq r_n\}$ is defined in (73).

Proof From (48b) in Assumption D and Assumption F, we know

$$\mathbb{P}\left\{\max_{i\in\mathcal{H}_0}\max_{1\le l\le p}\overline{M}(X_i,\boldsymbol{e}_l)<p^{\gamma_0+1}\log n\right\}=1-O(n^{-\gamma}), \tag{63}$$

$$\mathbb{P}\left[\max_{i\in\mathcal{H}_0}\max_{1\le l\le p}|\nabla f_l(X_i,\boldsymbol{\theta}^*)-\mu_l(\boldsymbol{\theta}^*)|<\eta^{-1}(\gamma+2)\log n\right]=1-O(n^{-\gamma}). \tag{64}$$

Construct the set of event

$$\mathfrak{X}_0:=\left\{X_1,...,X_n\middle|\ (64),\ (63)\text{ holds, and }\frac{\rho}{2}\le\min_{1\le l\le p}\widehat{\sigma}_l(\boldsymbol{\theta}^*)\right\}. \tag{65}$$

Together with (62), we know $\mathbb{P}(\mathfrak{X}_0)=1-O(n^{-\gamma})$.

Let $\mathfrak{N}_0$ be the n^{-M}-net of Θ_0, we know $Card(\mathfrak{N}_0)\le(1+2n^M)^p$. Then we will show

$$\max_{1\le l\le p}\sup_{|\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}|_2\le n^{-M}}|\widehat{\sigma}_l^2(\boldsymbol{\theta})-\widehat{\sigma}_l^2(\widetilde{\boldsymbol{\theta}})|\le n^{-2} \tag{66}$$

always hold under the event $\mathfrak{X}_0$ and M sufficiently large. Indeed, From (64) and (63), we have

$$\begin{aligned}
&\max_{1\le l\le p}\sup_{|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}|_2\le n^{-M}}|\widehat{\sigma}_l^2(\boldsymbol{\theta})-\widehat{\sigma}_l^2(\widetilde{\boldsymbol{\theta}})|\\
=&\max_{1\le l\le p}\sup_{|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}|_2\le n^{-M}}\Big|\frac{1}{n}\sum_{i\in\mathcal{H}_0}\{\nabla f_l(X_i,\boldsymbol{\theta})-g_l(\boldsymbol{\theta})\}^2-\frac{1}{n}\sum_{i\in\mathcal{H}_0}\Big\{\nabla f_l(X_i,\widetilde{\boldsymbol{\theta}})-g_l(\widetilde{\boldsymbol{\theta}})\Big\}^2\Big|\\
\le&\max_{1\le l\le p}\sup_{|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}|_2\le n^{-M}}\Big|\frac{1}{n}\sum_{i\in\mathcal{H}_0}\Big\{\overline{M}(X_i,\boldsymbol{e}_l)+\frac{1}{n}\sum_{i\in\mathcal{H}_0}\overline{M}(X_i,\boldsymbol{e}_l)\Big\}\\
&\times\Big\{\nabla f_l(X_i,\boldsymbol{\theta})-g_l(\boldsymbol{\theta})+\nabla f_l(X_i,\widetilde{\boldsymbol{\theta}})-g_l(\widetilde{\boldsymbol{\theta}})\Big\}n^{-M}\Big|\\
\le&\frac{2p^{\gamma_0+1}\log n}{n^M}\max_{1\le l\le p}\sup_{|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}|_2\le n^{-M}}\frac{1}{n}\sum_{i\in\mathcal{H}_0}\Big[2|\nabla f_l(X_i,\boldsymbol{\theta}^*)-\mu_l(\boldsymbol{\theta}^*)|+2|g_l(\boldsymbol{\theta}^*)-\mu_l(\boldsymbol{\theta}^*)|\\
&+\Big\{\overline{M}(X_i,\boldsymbol{e}_l)+\frac{1}{n}\sum_{i\in\mathcal{H}_0}\overline{M}(X_i,\boldsymbol{e}_l)\Big\}\big(|\boldsymbol{\theta}-\boldsymbol{\theta}^*|_2+|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*|_2\big)\Big]\\
\le&\frac{8p^{2\gamma_0+2}\log^2 n}{n^M}\le n^{-(M-2\gamma_0-4)},
\end{aligned}$$

Taking $M\ge 2\gamma_0+6$ we obtain (66). Thus we have

$$\begin{aligned}
&\mathbb{P}\left\{\max_{1\le l\le p}\sup_{\boldsymbol{\theta}\in\Theta_0}|\widehat{\sigma}_l(\boldsymbol{\theta})-\widehat{\sigma}_l(\boldsymbol{\theta}^*)|\ge 4x\right\}\\
\le&\mathbb{P}\left\{\max_{1\le l\le p}\max_{\widetilde{\boldsymbol{\theta}}\in\mathfrak{N}_0}\Big|\widehat{\sigma}_l^2(\widetilde{\boldsymbol{\theta}})-\widehat{\sigma}_l^2(\boldsymbol{\theta}^*)\Big|\ge\rho x,\ \mathfrak{X}_0\right\}\\
&+\mathbb{P}\left\{\max_{1\le l\le p}\sup_{|\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}|_2\le n^{-M}}\Big|\widehat{\sigma}_l^2(\boldsymbol{\theta})-\widehat{\sigma}_l^2(\widetilde{\boldsymbol{\theta}})\Big|\ge\rho x,\ \mathfrak{X}_0\right\}+\mathbb{P}(\mathfrak{X}_0^c)\\
\le&p(1+2n^M)^p\max_{1\le l\le p}\sup_{\boldsymbol{\theta}\in\Theta_0}\mathbb{P}\left\{\big|\widehat{\sigma}_l^2(\boldsymbol{\theta})-\widehat{\sigma}_l^2(\boldsymbol{\theta}^*)\big|\ge\rho x,\ \mathfrak{X}_0\right\}+O(n^{-\gamma}).
\end{aligned}$$

The next lemma will show, when we take $x = O(r_n)$, there is

$$\sup_{\{\boldsymbol{\theta}:|\boldsymbol{\theta}-\boldsymbol{\theta}^*|_2\le r_n\}} \mathbb{P}\left\{\left|\widehat{\sigma}_l^2(\boldsymbol{\theta}) - \widehat{\sigma}_l^2(\boldsymbol{\theta}^*)\right| \ge \rho x,\ \mathfrak{X}_0\right\} = O(n^{-p(M+\gamma+1)}),$$

and hence proves this lemma. ■

Lemma 25 *(Point-wise bound of variance difference) Under Assumption D, E, F and G, for every $\gamma > 1$, there exists $\widetilde{C}$ large enough, such that*

$$\max_{1\le l\le p}\sup_{\boldsymbol{\theta}\in\Theta_0} \mathbb{P}\left\{\left|\widehat{\sigma}_l^2(\boldsymbol{\theta}) - \widehat{\sigma}_l^2(\boldsymbol{\theta}^*)\right| \ge \widetilde{C} r_n,\ \mathfrak{X}_0\right\} = O(n^{-p\gamma}),$$

where $\mathfrak{X}_0$ is the set of events defined in (65), and $\Theta_0 = \{\boldsymbol{\theta} : |\boldsymbol{\theta} - \boldsymbol{\theta}^*|_2 \le r_n\}$ is defined in (73).

Proof For every $\boldsymbol{\theta} \in \Theta_0$ and $1 \le l \le p$, there is

$$\begin{aligned}
|\widehat{\sigma}_l^2(\boldsymbol{\theta}) - \widehat{\sigma}_l^2(\boldsymbol{\theta}^*)| \le & \underbrace{\left|\frac{1}{n}\sum_{i\in\mathcal{H}_0}\left\{\nabla f_l(X_i,\boldsymbol{\theta}) - \mu_l(\boldsymbol{\theta})\right\}^2 - \left\{\nabla f_l(X_i,\boldsymbol{\theta}^*) - \mu_l(\boldsymbol{\theta}^*)\right\}^2 - \sigma_l^2(\boldsymbol{\theta}) + \sigma_l^2(\boldsymbol{\theta}^*)\right|}_{T} \\
& + \left|\sigma_l^2(\boldsymbol{\theta}) - \sigma_l^2(\boldsymbol{\theta}^*)\right| + \left|\left\{\frac{1}{n}\sum_{i\in\mathcal{H}_0}\nabla f_l(X_i,\boldsymbol{\theta}) - \mu_l(\boldsymbol{\theta})\right\}^2\right. \\
& \left. - \left\{\frac{1}{n}\sum_{i\in\mathcal{H}_0}\nabla f_l(X_i,\boldsymbol{\theta}^*) - \mu_l(\boldsymbol{\theta}^*)\right\}^2\right|.
\end{aligned} \tag{67}$$

Let's firstly deal with the term $|\sigma_l^2(\boldsymbol{\theta}) - \sigma_l^2(\boldsymbol{\theta}^*)|$,

$$\begin{aligned}
& |\sigma_l^2(\boldsymbol{\theta}) - \sigma_l^2(\boldsymbol{\theta}^*)| \\
=& \left|\mathbb{E}\left[\left\{\nabla f_l(X_i,\boldsymbol{\theta}) - \mu_l(\boldsymbol{\theta})\right\}^2 - \left\{\nabla f_l(X_i,\boldsymbol{\theta}^*) - \mu_l(\boldsymbol{\theta}^*)\right\}^2\right]\right| \\
\le& r_n\mathbb{E}\Big[\left\{2|\nabla f_l(X_i,\boldsymbol{\theta}^*) - \mu_l(\boldsymbol{\theta}^*)| + r_n\left(M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_i,\boldsymbol{e}_l) + \mathbb{E}[M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_i,\boldsymbol{e}_l)]\right)\right\} \\
& \times \left\{M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_i,\boldsymbol{e}_l) + \mathbb{E}[M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_i,\boldsymbol{e}_l)]\right\}\Big] \\
\le& r_n\mathbb{E}\left[(4+4r_n)\,M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}^2(X_i,\boldsymbol{e}_l) + \left|\nabla f_l(X_i,\boldsymbol{\theta}^*) - \mu_l(\boldsymbol{\theta}^*)\right|^2\right] = O\left(r_n\right).
\end{aligned} \tag{68}$$

For the last term in (67), there is

$$\begin{aligned}
& \left|\left\{\frac{1}{n}\sum_{i\in\mathcal{H}_0}\nabla f_l(X_i,\boldsymbol{\theta}) - \mu_l(\boldsymbol{\theta})\right\}^2 - \left\{\frac{1}{n}\sum_{i\in\mathcal{H}_0}\nabla f_l(X_i,\boldsymbol{\theta}^*) - \mu_l(\boldsymbol{\theta}^*)\right\}^2\right| \\
=& \left|\frac{1}{n}\sum_{i\in\mathcal{H}_0}\left\{\nabla f_l(X_i,\boldsymbol{\theta}) - \mu_l(\boldsymbol{\theta}) - \nabla f_l(X_i,\boldsymbol{\theta}^*) + \mu_l(\boldsymbol{\theta}^*)\right\}\right| \\
& \times \left|\frac{1}{n}\sum_{i\in\mathcal{H}_0}\left\{\nabla f_l(X_i,\boldsymbol{\theta}) - \mu_l(\boldsymbol{\theta}) + \nabla f_l(X_i,\boldsymbol{\theta}^*) - \mu_l(\boldsymbol{\theta}^*)\right\}\right|.
\end{aligned}$$

Compute that

$$
\begin{aligned}
&\mathfrak{E}\left\{\frac{\eta}{2}, \nabla f_l(X_1,\boldsymbol{\theta}) - \mu_l(\boldsymbol{\theta}) - \nabla f_l(X_1,\boldsymbol{\theta}^*) + \mu_l(\boldsymbol{\theta}^*)\right\}\\
\le&\mathfrak{E}\left\{\frac{\eta}{2}, r_n\big[M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_1,\boldsymbol{e}_l) + \mathbb{E}\{M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_1,\boldsymbol{e}_l)\}\big]\right\}\\
\le&r_n^2\mathbb{E}\left\{\big[M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_1,\boldsymbol{e}_l) + \mathbb{E}\{M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_1,\boldsymbol{e}_l)\}\big]^2 \exp\frac{\eta r_n}{2}\big[M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_1,\boldsymbol{e}_l) + \mathbb{E}\{M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_1,\boldsymbol{e}_l)\}\big]\right\}\\
\le&4\eta^{-2}r_n^2\mathbb{E}\left\{\exp\eta\big[M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_1,\boldsymbol{e}_l) + \mathbb{E}\{M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_1,\boldsymbol{e}_l)\}\big]\right\} = O(r_n^2),\\
&\mathfrak{E}\left\{\frac{\eta}{8}, \nabla f_l(X_1,\boldsymbol{\theta}) - \mu_l(\boldsymbol{\theta}) + \nabla f_l(X_1,\boldsymbol{\theta}^*) - \mu_l(\boldsymbol{\theta}^*)\right\}\\
\le&\mathfrak{E}\left\{\frac{\eta}{8}, 2|\nabla f_l(X_1,\boldsymbol{\theta}^*) - \mu_l(\boldsymbol{\theta}^*)| + r_n\big[M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_1,\boldsymbol{e}_l) + \mathbb{E}\{M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_1,\boldsymbol{e}_l)\}\big]\right\}\\
\le&64\eta^{-2}\mathbb{E}\left\{\exp\frac{\eta}{4}\big[2|\nabla f_l(X_1,\boldsymbol{\theta}^*) - \mu_l(\boldsymbol{\theta}^*)| + r_n M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_1,\boldsymbol{e}_l) + r_n\mathbb{E}\{M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_1,\boldsymbol{e}_l)\}\big]\right\} = O(1).
\end{aligned}
$$

Applying Lemma 14 to each averaged term, we have

$$
\left|\left\{\frac{1}{n}\sum_{i\in\mathcal{H}_0}\nabla f_l(X_i,\boldsymbol{\theta}) - \mu_l(\boldsymbol{\theta})\right\}^2 - \left\{\frac{1}{n}\sum_{i\in\mathcal{H}_0}\nabla f_l(X_i,\boldsymbol{\theta}^*) - \mu_l(\boldsymbol{\theta}^*)\right\}^2\right| = O_{\mathbb{P}}\left(\frac{r_n p\log n}{n}\right). \quad (69)
$$

Lastly we shall focus on the term T in (67). For each $i\in\mathcal{H}_0$, denote

$$
\begin{aligned}
\mathfrak{X}_i :=&\quad \left\{X_i \;\middle|\; \begin{array}{l} M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_i,\boldsymbol{e}_l) < \eta^{-1}(\gamma+2)p\log n,\\ |\nabla f_l(X_i,\boldsymbol{\theta}^*) - \mu_l(\boldsymbol{\theta}^*)| < \eta^{-1}(\gamma+2)\log n\end{array}\right\},\\
Y_i :=&\quad \{\nabla f_l(X_i,\boldsymbol{\theta}) - \mu_l(\boldsymbol{\theta})\}^2 - \{\nabla f_l(X_i,\boldsymbol{\theta}^*) - \mu_l(\boldsymbol{\theta}^*)\}^2,\\
\overline{Y}_i :=&\quad Y_i\mathbb{I}(\mathfrak{X}_i) - \mathbb{E}\{Y_i\mathbb{I}(\mathfrak{X}_i)\},\quad \widetilde{Y}_i := Y_i\mathbb{I}(\mathfrak{X}_i^c) - \mathbb{E}\{Y_i\mathbb{I}(\mathfrak{X}_i^c)\}.
\end{aligned}
$$

We can compute that

$$
\begin{aligned}
\big|\mathbb{E}\{Y_1\mathbb{I}(\mathfrak{X}_1^c)\}\big| =&\big|\mathbb{E}\left[\{\nabla f_l(X_1,\boldsymbol{\theta}) - \mu_l(\boldsymbol{\theta})\}^2\mathbb{I}(\mathfrak{X}_1^c) - \{\nabla f_l(X_1,\boldsymbol{\theta}^*) - \mu_l(\boldsymbol{\theta}^*)\}^2\mathbb{I}(\mathfrak{X}_1^c)\right]\big|\\
\le&\mathbb{E}\Big[\big\{2|\nabla f_l(X_1,\boldsymbol{\theta}^*) - \mu_l(\boldsymbol{\theta}^*)| + \big(M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_1,\boldsymbol{e}_l) + \mathbb{E}[M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_1,\boldsymbol{e}_l)]\big)\big\}\\
&\times\{M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_1,\boldsymbol{e}_l) + \mathbb{E}[M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_1,\boldsymbol{e}_l)]\}\,\mathbb{I}(\mathfrak{X}_1^c)\Big]\\
\le&\mathbb{E}\left[\big|\nabla f_l(X_1,\boldsymbol{\theta}^*) - \mu_l(\boldsymbol{\theta}^*)\big|^2\mathbb{I}(\mathfrak{X}_1^c) + 4\big|M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_1,\boldsymbol{e}_l)\big|^2\mathbb{I}(\mathfrak{X}_1^c)\right]\\
=&O(n^{-2}) = o(r_n).
\end{aligned}
$$

Then from Assumption D we know

$$
\begin{aligned}
&\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i\in\mathcal{H}_0}\widetilde{Y}_i\right| > \big|\mathbb{E}\{Y_1\mathbb{I}(\mathfrak{X}_1^c)\}\big|,\ \mathfrak{X}_0\right\} \le \mathbb{P}\left\{\Big(\bigcup_{i\in\mathcal{H}_0}\mathfrak{X}_i^c\Big)\bigcap\mathfrak{X}_0\right\}\\
\le&\mathbb{P}\left\{\max_{i\in\mathcal{H}_0}M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_i,\boldsymbol{e}_l) \ge \eta^{-1}(\gamma+2)p\log n\right\} = O(n^{-\gamma p}).
\end{aligned} \quad (70)
$$

Thus we have

$$
\begin{aligned}
&\left|\overline{Y}_1\right| = \left|Y_1\mathbb{I}(\mathfrak{X}_1) - \sigma_l^2(\boldsymbol{\theta}) + \sigma_l^2(\boldsymbol{\theta}^*) + \mathbb{E}\{Y_1\mathbb{I}(\mathfrak{X}_1^c)\}\right| \\
\leq &\Big|\big\{\nabla f_l(X_1,\boldsymbol{\theta}) - \mu_l(\boldsymbol{\theta}) - \nabla f_l(X_1,\boldsymbol{\theta}^*) + \mu_l(\boldsymbol{\theta}^*)\big\}\big\{\nabla f_l(X_1,\boldsymbol{\theta}) - \mu_l(\boldsymbol{\theta}) + \nabla f_l(X_1,\boldsymbol{\theta}^*) - \mu_l(\boldsymbol{\theta}^*)\big\}\Big|\mathbb{I}(\mathfrak{X}_1) \\
&+ O\left(r_n\right) \\
\leq &r_n\left[M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_1,\boldsymbol{e}_l) + \mathbb{E}\{M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_1,\boldsymbol{e}_l)\}\right]\Big(2|\nabla f_l(X_1,\boldsymbol{\theta}^*) - \mu_l(\boldsymbol{\theta}^*)| + r_n\big[M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_1,\boldsymbol{e}_l) \\
&+ \mathbb{E}\{M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_1,\boldsymbol{e}_l)\}\big]\Big)\mathbb{I}(\mathfrak{X}_1) + O\left(r_n\right) \\
\leq &\left\{\frac{2(\gamma+2)}{\eta}r_n\log n + \frac{(\gamma+2)}{\eta}r_n^2\, p\log n\right\}\left[M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_1,\boldsymbol{e}_l) + \mathbb{E}\{M_{\boldsymbol{\theta},\boldsymbol{\theta}^*}(X_1,\boldsymbol{e}_l)\}\right]\mathbb{I}(\mathfrak{X}_1) + O\left(r_n\right).
\end{aligned}
$$

From Assumption G and D we know $r_n\log n + r_n^2\, p\log n = O(1)$, thus

$$
\mathfrak{E}\left(\frac{\eta}{2},\overline{Y}_1\right) = O\left(r_n^2\log^2 n + r_n^4\, p^2\log^2 n\right) = O(1).
$$

Then Lemma 14 yields

$$
\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i\in\mathcal{H}_0}\overline{Y}_i\right| \geq C\left(r_n\sqrt{\frac{p\log^3 n}{n}} + r_n^2\sqrt{\frac{p^3\log^3 n}{n}}\right)\right\} = O(n^{-\gamma p}), \tag{71}
$$

holds for some $C>0$ large enough. Thus from (70) and (71) we have

$$
\begin{aligned}
&\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i\in\mathcal{H}_0}Y_i\right| \geq C\left(r_n\sqrt{\frac{p\log^3 n}{n}} + r_n^2\sqrt{\frac{p^3\log^3 n}{n}} + r_n\right),\ \mathfrak{X}_0\right\} \\
\leq &\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i\in\mathcal{H}_0}\overline{Y}_i\right| \geq C\left(r_n\sqrt{\frac{p\log^3 n}{n}} + r_n^2\sqrt{\frac{p^3\log^3 n}{n}}\right)\right\} \\
&+ \mathbb{P}\left\{\left|\frac{1}{n}\sum_{i\in\mathcal{H}_0}\widetilde{Y}_i\right| > Cr_n,\ \mathfrak{X}_0\right\} = O(n^{-\gamma p}).
\end{aligned} \tag{72}
$$

Finally taking (68), (69) and (72) back to (67), we conclude that, under $\mathfrak{X}_0$, there is

$$
\left|\widehat{\sigma}_l^2(\boldsymbol{\theta}) - \widehat{\sigma}_l^2(\boldsymbol{\theta}^*)\right| = O_{\mathbb{P}}\left(r_n\sqrt{\frac{p\log^3 n}{n}} + r_n^2\sqrt{\frac{p^3\log^3 n}{n}} + r_n\right) = O\left(r_n\right),
$$

since we already supposed $p = O(\sqrt{n}\log^{-1} n)$ and $r_n^2\, p\log n = O(1)$ in Assumption G. ■

Lemma 26 *(Stability of correction terms) Let $N(=(m+1)n)$ i.i.d. random variables $X_1, ..., X_N$ evenly distributed in $m+1$ subsets $\mathcal{H}_0, ..., \mathcal{H}_m$. Let*

$$\begin{aligned}\overline{X}_j =& n^{-1}\sum_{i\in\mathcal{H}_j} X_i,\\ \widehat{X} =& \mathrm{med}(\overline{X}_j \mid 0\le j\le m),\\ \widehat{\sigma}^2 =& n^{-1}\sum_{i\in\mathcal{H}_0}(X_i-\overline{X}_0)^2.\end{aligned}$$

Suppose X_1 satisfies $\mathbb{E}(X_1)=0, \mathrm{Var}(X_1)=\sigma^2$ and $\mathbb{E}[e^{\eta|X_1|}]\le C$. Then there exists $\widetilde{C}>0$ sufficiently large, such that

$$\mathbb{P}\left[\frac{1}{m+1}\sum_{j=0}^{m}\mathbb{I}\left\{\left|\overline{X}_j-\widehat{X}-\frac{\widehat{\sigma}\Delta_k}{\sqrt{n}}\right|\le\delta_n\right\}\ge\widetilde{C}\max\left\{\frac{\log n}{m},\sqrt{n}\delta_n\right\}\right]\le O(n^{-\gamma}),$$

provided $\max\{1/(n\delta_n),\sqrt{n}\delta_n\}=O(1)$.

Proof Construct the set

$$\mathfrak{X}_Q\triangleq\left\{X_1,...,X_N\,\middle|\,\widehat{\sigma}\le\sqrt{n},\,|\widehat{X}|\le\sqrt{n}\right\}.$$

From Lemma 16 and Lemma 23, it is easy to see $\mathfrak{X}_Q$ holds with probability greater than $1-O(n^{-\gamma})$. Construct an n^{-2}-net $\mathfrak{N}_Q$ for the square $\Theta_Q\triangleq[0,\sqrt{n}]\times[-\sqrt{n},\sqrt{n}]$, then there is

$$\begin{aligned}&\frac{1}{m+1}\sum_{j=0}^{m}\mathbb{I}\left\{\left|\overline{X}_j-\widehat{X}-\frac{\widehat{\sigma}\Delta_k}{\sqrt{n}}\right|\le\delta_n\right\}\\ \le&\sup_{(\sigma,y)\in\Theta_Q}\frac{1}{m+1}\sum_{j=0}^{m}\mathbb{I}\left\{\left|\overline{X}_j-y-\frac{\sigma\Delta_k}{\sqrt{n}}\right|\le\delta_n\right\}\\ \le&\max_{(\widetilde{\sigma},\widetilde{y})\in\mathfrak{N}_Q}\frac{1}{m+1}\sum_{j=0}^{m}\mathbb{I}\left\{\left|Y_j-\widetilde{y}-\frac{\widetilde{\sigma}\Delta_k}{\sqrt{n}}\right|\le\delta_n\right\}\\ &+\max_{(\widetilde{\sigma},\widetilde{y})\in\mathfrak{N}_Q}\sup_{\{(\sigma,y):|\widetilde{\sigma}-\sigma|,|\widetilde{y}-y|\le n^{-2}\}}\frac{1}{m+1}\sum_{j=0}^{m}\left[\mathbb{I}\left\{\left|\overline{X}_j-\widetilde{y}-\frac{\widetilde{\sigma}\Delta_k}{\sqrt{n}}\right|\le\delta_n\right\}-\mathbb{I}\left\{\left|\overline{X}_j-y-\frac{\sigma\Delta_k}{\sqrt{n}}\right|\le\delta_n\right\}\right]\\ \le&2\max_{(\widetilde{\sigma},\widetilde{y})\in\mathfrak{N}_Q}\frac{1}{m+1}\sum_{j=0}^{m}\mathbb{I}\left\{\left|\overline{X}_j-\widetilde{y}-\frac{\widetilde{\sigma}\Delta_k}{\sqrt{n}}\right|\le2\delta_n\right\}.\end{aligned}$$

Denote

$$\begin{aligned}Z_j(y,\sigma)=:&\,\mathbb{I}\left\{\left|\overline{X}_j-y-\frac{\sigma\Delta_k}{\sqrt{n}}\right|\le2\delta_n\right\}-\mathbb{P}\left\{\left|\overline{X}_j-y-\frac{\sigma\Delta_k}{\sqrt{n}}\right|\le2\delta_n\right\},\\ \mathbb{P}_j(y,\sigma)=:&\,\mathbb{P}\left\{\left|\overline{X}_j-y-\frac{\sigma\Delta_k}{\sqrt{n}}\right|\le2\delta_n\right\}.\end{aligned}$$

Apply Lemma 13 to $\sqrt{n}Y_j$, we know there exist constants $C_1, C_2 > 0$ such that

$$\begin{aligned}
&\mathbb{P}_j(y,\sigma) \leq \mathbb{P}\left\{\left|\mathcal{N}(0,1) - \sqrt{n}y - \sigma\Delta_k\right| \leq 2\sqrt{n}\delta_n\right\} + \frac{C_1}{\sqrt{n}} \leq 4C_1\sqrt{n}\delta_n \\
&\mathfrak{E}\left\{1, Z_j(y,\sigma)\right\} \leq e\left\{1 + \mathbb{P}_j(y,\sigma)\right\}\mathbb{P}_j(y,\sigma) \\
\leq& 8ec\sqrt{n}\delta_n \leq C_2 \max\left\{\frac{\log n}{m}, \sqrt{n}\delta_n\right\}.
\end{aligned}$$

Apply Lemma 14 together with $|\mathfrak{N}_Q| \leq 2n^5$, we have

$$\begin{aligned}
&\mathbb{P}\left\{\max_{(\widetilde{\sigma},\widetilde{y})\in\mathfrak{N}_Q} \frac{1}{m+1}\sum_{j=0}^{m} Z_j(\widetilde{y},\widetilde{\sigma}) \geq x\right\} \\
\leq& 2n^5 \sup_{(\sigma,y)\in\Theta_Q} \mathbb{P}\left\{\frac{1}{m+1}\sum_{j=0}^{m} Z_j(y,\sigma) \geq x\right\} \\
\leq& n^{-\gamma},
\end{aligned}$$

by taking $x \geq (\gamma+6)\sqrt{C_2}\max\left\{\frac{\log n}{m}, \sqrt{n}\delta_n\right\}$. Then we conclude that

$$\begin{aligned}
&\frac{1}{m+1}\sum_{j=0}^{m} \mathbb{I}\left\{\left|\overline{X}_j - \widehat{X} - \frac{\widehat{\sigma}\Delta_k}{\sqrt{n}}\right| \leq \delta_n\right\} \\
\leq& \max_{(\widetilde{\sigma},\widetilde{y})\in\mathfrak{N}_Q} \frac{2}{m+1}\sum_{j=0}^{m} Z_j(\widetilde{y},\widetilde{\sigma}) + 4c\sqrt{n}\delta_n \\
\leq& 4(\gamma+6)\sqrt{C_2}\max\left\{\frac{\log n}{m}, \sqrt{n}\delta_n\right\},
\end{aligned}$$

holds with probability larger than $1 - O(n^{-\gamma})$. Thus we can obtain the desired result. ■

C.3 Proofs of results in Appendix C.1.2

Now, we are ready to provide the proofs of the convergence rate and asymptotic normality for the RCSL estimator.

Proof [Proof of Theorem 11 and Theorem 19] Denote b_n as the desired convergence rate of $\widehat{\boldsymbol{\theta}}^{(1)}$, and $\boldsymbol{g}_j(\boldsymbol{\theta}) = n^{-1}\sum_{i\in\mathcal{H}_j}\nabla f(X_i,\boldsymbol{\theta})$. We construct the following sets in the parameter space.

$$\begin{aligned}
\Theta_0 &= \{\boldsymbol{\theta} \in \mathbb{R}^p : |\boldsymbol{\theta} - \boldsymbol{\theta}^*|_2 \leq r_n\}, \\
\Theta_1 &= \{\boldsymbol{\theta} \in \mathbb{R}^p : |\boldsymbol{\theta} - \boldsymbol{\theta}^*|_2 = b_n\}, \\
\Theta_2 &= \{\boldsymbol{\theta} \in \mathbb{R}^p : |\boldsymbol{\theta} - \boldsymbol{\theta}^*|_2 \leq b_n\}.
\end{aligned} \tag{73}$$

Given an initial estimator $\widehat{\boldsymbol{\theta}}^{(0)}$ lies in the set Θ_0, we will show

$$\frac{1}{n}\sum_{i\in\mathcal{H}_0} f(X_i,\boldsymbol{\theta}_1) - \langle \boldsymbol{g}_0(\widehat{\boldsymbol{\theta}}^{(0)}) - \bar{\boldsymbol{g}}(\widehat{\boldsymbol{\theta}}^{(0)}), \boldsymbol{\theta}_1\rangle > \frac{1}{n}\sum_{i\in\mathcal{H}_0} f(X_i,\boldsymbol{\theta}^*) - \langle \boldsymbol{g}_0(\widehat{\boldsymbol{\theta}}^{(0)}) - \bar{\boldsymbol{g}}(\widehat{\boldsymbol{\theta}}^{(0)}), \boldsymbol{\theta}^*\rangle, \tag{74}$$

holds uniformly for $\boldsymbol{\theta}_1 \in \Theta_1$ with probability tending to 1. Notice that

$$
\begin{aligned}
&\frac{1}{n}\sum_{i\in\mathcal{H}_0}\{f(X_i,\boldsymbol{\theta}_1)-f(X_i,\boldsymbol{\theta}^*)\}-\left\langle \boldsymbol{g}_0(\widehat{\boldsymbol{\theta}}^{(0)})-\bar{\boldsymbol{g}}(\widehat{\boldsymbol{\theta}}^{(0)}),\boldsymbol{\theta}_1-\boldsymbol{\theta}^*\right\rangle \qquad (75)\\
=&\left\langle \int_0^1\left[\boldsymbol{g}_0(\boldsymbol{\theta}^*+s(\boldsymbol{\theta}_1-\boldsymbol{\theta}^*))-\boldsymbol{g}_0(\boldsymbol{\theta}^*)-\boldsymbol{\mu}(\boldsymbol{\theta}^*+s(\boldsymbol{\theta}_1-\boldsymbol{\theta}^*))+\boldsymbol{\mu}(\boldsymbol{\theta}^*)\right]\mathrm{d}s,\boldsymbol{\theta}_1-\boldsymbol{\theta}^*\right\rangle\\
&+\left\langle \int_0^1\{\boldsymbol{\mu}(\boldsymbol{\theta}^*+s(\boldsymbol{\theta}_1-\boldsymbol{\theta}^*))-\boldsymbol{\mu}(\boldsymbol{\theta}^*)\}\,\mathrm{d}s,\boldsymbol{\theta}_1-\boldsymbol{\theta}^*\right\rangle+\left\langle \boldsymbol{g}_0(\boldsymbol{\theta}^*)-\boldsymbol{g}_0(\widehat{\boldsymbol{\theta}}^{(0)})+\bar{\boldsymbol{g}}(\widehat{\boldsymbol{\theta}}^{(0)}),\boldsymbol{\theta}_1-\boldsymbol{\theta}^*\right\rangle\\
=&\frac{1}{2}\left\langle \nabla\boldsymbol{\mu}(\boldsymbol{\theta}^*)(\boldsymbol{\theta}_1-\boldsymbol{\theta}^*),\boldsymbol{\theta}_1-\boldsymbol{\theta}^*\right\rangle\\
&+\left\langle \underbrace{\int_0^1(1-s)\{\nabla\boldsymbol{\mu}(\boldsymbol{\theta}^*+s(\boldsymbol{\theta}_1-\boldsymbol{\theta}^*))-\nabla\boldsymbol{\mu}(\boldsymbol{\theta}^*)\}\,\mathrm{d}s}_{\boldsymbol{T}_1}(\boldsymbol{\theta}_1-\boldsymbol{\theta}^*),\boldsymbol{\theta}_1-\boldsymbol{\theta}^*\right\rangle\\
&+\left\langle \underbrace{\int_0^1\left[\boldsymbol{g}_0(\boldsymbol{\theta}^*+s(\boldsymbol{\theta}_1-\boldsymbol{\theta}^*))-\boldsymbol{g}_0(\boldsymbol{\theta}^*)-\boldsymbol{\mu}(\boldsymbol{\theta}^*+s(\boldsymbol{\theta}_1-\boldsymbol{\theta}^*))+\boldsymbol{\mu}(\boldsymbol{\theta}^*)\right]\mathrm{d}s}_{\boldsymbol{T}_2},\boldsymbol{\theta}_1-\boldsymbol{\theta}^*\right\rangle\\
&+\left\langle \underbrace{\boldsymbol{g}_0(\boldsymbol{\theta}^*)-\boldsymbol{g}_0(\widehat{\boldsymbol{\theta}}^{(0)})+\boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}^{(0)})-\boldsymbol{\mu}(\boldsymbol{\theta}^*)}_{\boldsymbol{T}_3},\boldsymbol{\theta}_1-\boldsymbol{\theta}^*\right\rangle+\left\langle \underbrace{\bar{\boldsymbol{g}}(\widehat{\boldsymbol{\theta}}^{(0)})-\boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}^{(0)})}_{\boldsymbol{T}_4},\boldsymbol{\theta}_1-\boldsymbol{\theta}^*\right\rangle.
\end{aligned}
$$

To show positivity of this difference, it left to bound the norms of $\boldsymbol{T}_1,\boldsymbol{T}_2,\boldsymbol{T}_3$ and $\boldsymbol{T}_4$ for $\widehat{\boldsymbol{\theta}}^{(0)}\in\Theta_0$ and $\boldsymbol{\theta}_1\in\Theta_1$ uniformly.

Firstly, from assumption C, we have

$$
\begin{aligned}
\|\boldsymbol{T}_1\|&\le\int_0^1(1-s)\,\|\nabla\boldsymbol{\mu}(\boldsymbol{\theta}^*+s(\boldsymbol{\theta}_1-\boldsymbol{\theta}^*))-\nabla\boldsymbol{\mu}(\boldsymbol{\theta}^*)\|\,dt\\
&\le\int_0^1 C_H(1-s)s|\boldsymbol{\theta}_1-\boldsymbol{\theta}^*|_2\mathrm{d}s=\frac{C_H|\boldsymbol{\theta}_1-\boldsymbol{\theta}^*|_2}{6}=o(1).
\end{aligned}
$$

To control the rest three terms, we denote

$$
\boldsymbol{Z}(x,\boldsymbol{\theta}):=\nabla f(x,\boldsymbol{\theta})-\nabla f(x,\boldsymbol{\theta}^*)-\boldsymbol{\mu}(\boldsymbol{\theta})+\boldsymbol{\mu}(\boldsymbol{\theta}^*), \qquad (76)
$$

for ease of notations. Then we will show

$$
\sup_{\boldsymbol{\theta}\in\Theta_0}\left|\frac{1}{n}\sum_{i\in\mathcal{H}_0}\boldsymbol{Z}(X_i,\boldsymbol{\theta})\right|_2=O_{\mathbb{P}}\left(r_n\sqrt{\frac{p\log n}{n}}\right). \qquad (77)
$$

Construct the (r_nn^{-M})-net $\mathfrak{N}_0$ for the set Θ_0, where $M>0$ is some sufficiently large number. From Lemma 5.2 of Vershynin (2010) we know $\mathrm{C}ard(\mathfrak{N}_0)\le(1+2n^M)^p$. Then we have

$$
\begin{aligned}
\sup_{\boldsymbol{\theta}\in\Theta_0}\left|\frac{1}{n}\sum_{i\in\mathcal{H}_0}r_n^{-1}\boldsymbol{Z}(X_i,\boldsymbol{\theta})\right|_2&\le\max_{\widetilde{\boldsymbol{\theta}}\in\mathfrak{N}_0}\left|\frac{1}{n}\sum_{i\in\mathcal{H}_0}r_n^{-1}\boldsymbol{Z}(X_i,\widetilde{\boldsymbol{\theta}})\right|_2\\
&+\frac{1}{n^{M+1}}\sum_{i\in\mathcal{H}_0}\left\{\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\overline{M}(X_i,\boldsymbol{v})\right\}+\frac{1}{n^M}\mathbb{E}\left\{\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\overline{M}(X_1,\boldsymbol{v})\right\}.
\end{aligned}
$$

From (48b) in Assumption D and Markov inequality, there is

$$\frac{1}{n^M}\mathbb{E}\left\{\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\overline{M}(X_1,\boldsymbol{v})\right\}\leq\frac{p^{\gamma_0}}{n^M}\mathbb{E}\left[\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\exp\left\{p^{-\gamma_0}\overline{M}(X_1,\boldsymbol{v})\right\}\right]<\frac{C_M p^{\gamma_0}}{n^M},$$
$$\frac{1}{n^{M+1}}\sum_{i\in\mathcal{H}_0}\left\{\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\overline{M}(X_i,\boldsymbol{v})\right\}=O_{\mathbb{P}}\left(\frac{p^{\gamma_0}\log n}{n^M}\right).$$

On the other hand, by standard ϵ-net argument for vector norms, we know that there exists a 1/2-net $\mathfrak{N}_S$ of $\mathbb{S}^{p-1}$ such that $Card(\mathfrak{N}_S)\leq 5^p$. It holds that

$$\begin{aligned}\left|\frac{1}{n}\sum_{i\in\mathcal{H}_0}\boldsymbol{Z}(X_i,\widetilde{\boldsymbol{\theta}})\right|_2&=\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\left\langle\frac{1}{n}\sum_{i\in\mathcal{H}_0}\boldsymbol{Z}(X_i,\widetilde{\boldsymbol{\theta}}),\boldsymbol{v}\right\rangle\\&\leq\max_{\boldsymbol{v}\in\mathfrak{N}_S}\left\langle\frac{1}{n}\sum_{i\in\mathcal{H}_0}\boldsymbol{Z}(X_i,\widetilde{\boldsymbol{\theta}}),\boldsymbol{v}\right\rangle+\sup_{|\boldsymbol{v}-\widetilde{\boldsymbol{v}}|_2\leq 1/2}\left\langle\frac{1}{n}\sum_{i\in\mathcal{H}_0}\boldsymbol{Z}(X_i,\widetilde{\boldsymbol{\theta}}),\boldsymbol{v}-\widetilde{\boldsymbol{v}}\right\rangle,\\\Rightarrow\left|\frac{1}{n}\sum_{i\in\mathcal{H}_0}\boldsymbol{Z}(X_i,\widetilde{\boldsymbol{\theta}})\right|_2&\leq 2\max_{\boldsymbol{v}\in\mathfrak{N}_S}\left\langle\frac{1}{n}\sum_{i\in\mathcal{H}_0}\boldsymbol{Z}(X_i,\widetilde{\boldsymbol{\theta}}),\boldsymbol{v}\right\rangle.\end{aligned}$$

Thus we have

$$\begin{aligned}\mathbb{P}\left(\max_{\widetilde{\boldsymbol{\theta}}\in\mathfrak{N}_0}\left|\frac{1}{n}\sum_{i\in\mathcal{H}_0}r_n^{-1}\boldsymbol{Z}(X_i,\widetilde{\boldsymbol{\theta}})\right|_2\geq 2x\right)\leq&(1+2n^M)^p\max_{\widetilde{\boldsymbol{\theta}}\in\mathfrak{N}_0}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i\in\mathcal{H}_0}r_n^{-1}\boldsymbol{Z}(X_i,\widetilde{\boldsymbol{\theta}})\right|_2\geq 2x\right)\\\leq&5^p(1+2n^M)^p\max_{\widetilde{\boldsymbol{\theta}}\in\mathfrak{N}_0}\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\mathbb{P}\left(\frac{1}{n}\sum_{i\in\mathcal{H}_0}\left\langle r_n^{-1}\boldsymbol{Z}(X_i,\widetilde{\boldsymbol{\theta}}),\boldsymbol{v}\right\rangle\geq x\right).\end{aligned}$$

Moreover, by (48a) in Assumption D, for every $\boldsymbol{v}\in\mathbb{S}^{p-1}$ and $\widetilde{\boldsymbol{\theta}}\in\mathfrak{N}_0$, we can compute that

$$\begin{aligned}&\mathfrak{E}\left\{\eta/2,\left\langle r_n^{-1}\boldsymbol{Z}(X_1,\widetilde{\boldsymbol{\theta}}),\boldsymbol{v}\right\rangle\right\}\\\leq&\mathbb{E}\left[\left\{M_{\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*}(X_1,\boldsymbol{v})+\mathbb{E}[M_{\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*}(X_1,\boldsymbol{v})]\right\}^2\exp\frac{\eta}{2}\left\{M_{\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*}(X_1,\boldsymbol{v})+\mathbb{E}[M_{\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*}(X_1,\boldsymbol{v})]\right\}\right]\\\leq&4\eta^{-2}\mathbb{E}\big[\exp\eta\big\{M_{\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*}(X_1,\boldsymbol{v})+\mathbb{E}[M_{\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*}(X_1,\boldsymbol{v})]\big\}\big]\leq 4C_M^2\eta^{-2}=O(1).\end{aligned}$$

So Lemma 14 yields

$$5^p(1+2n^M)^p\max_{\widetilde{\boldsymbol{\theta}}\in\mathfrak{N}_0}\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\mathbb{P}\left(\frac{1}{n}\sum_{i\in\mathcal{H}_0}\left\langle r_n^{-1}\boldsymbol{Z}(X_i,\widetilde{\boldsymbol{\theta}}),\boldsymbol{v}\right\rangle\geq x\right)=O(n^{-p\gamma}),$$

with $x = C_1\sqrt{\frac{p\log n}{n}}$ and C_1 large enough. Hence we proved the bound (77). Similarly we can give bounds for $\boldsymbol{T}_2$ and $\boldsymbol{T}_3$ as follows

$$\begin{aligned}
|\boldsymbol{T}_2|_2 &\le \int_0^1 \left| \frac{1}{n}\sum_{i\in\mathcal{H}_0} \boldsymbol{Z}(X_i, \boldsymbol{\theta}^* + s(\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*)) \right|_2 \mathrm{d}s \\
&\le \sup_{\boldsymbol{\theta}\in\Theta_2} \left| \frac{1}{n}\sum_{i\in\mathcal{H}_0} \boldsymbol{Z}(X_i, \boldsymbol{\theta}) \right|_2 = O\left(b_n\sqrt{\frac{p\log n}{n}} \right); \\
|\boldsymbol{T}_3|_2 &\le \sup_{\boldsymbol{\theta}\in\Theta_0} \left| \frac{1}{n}\sum_{i\in\mathcal{H}_0} \boldsymbol{Z}(X_i, \boldsymbol{\theta}) \right|_2 = O\left(r_n\sqrt{\frac{p\log n}{n}} \right).
\end{aligned}$$

Now for the term $\boldsymbol{T}_4$, under the rates constraints in Assumption G, we can prove a similar result as (77):

$$\begin{aligned}
&\left|\bar{\boldsymbol{g}}(\widehat{\boldsymbol{\theta}}^{(0)}) - \boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}^{(0)}) - \bar{\boldsymbol{g}}(\boldsymbol{\theta}^*) + \boldsymbol{\mu}(\boldsymbol{\theta}^*)\right|_2 \\
\le &\sup_{\boldsymbol{\theta}\in\Theta_0} \left|\bar{\boldsymbol{g}}(\boldsymbol{\theta}) - \boldsymbol{\mu}(\boldsymbol{\theta}) - \bar{\boldsymbol{g}}(\boldsymbol{\theta}^*) + \boldsymbol{\mu}(\boldsymbol{\theta}^*)\right|_2 \\
= &O_{\mathbb{P}}\left(\frac{\alpha_n\sqrt{p}}{\sqrt{n}} + \sqrt{\frac{p\log n}{mn}} + \frac{\sqrt{p}\log n}{m\sqrt{n}} + r_n\sqrt{\frac{p^2\log n}{n}} \right).
\end{aligned} \tag{78}$$

However, the proof of (78) involves more delicate analysis, so we delegate this part to Lemma 27 below. Moreover, follow the proof of Theorem 5 together with Assumption E and F, we can apply exponential inequality (Lemma 14) to the i.i.d. terms in (36) and yields

$$|\bar{\boldsymbol{g}}(\boldsymbol{\theta}^*) - \boldsymbol{\mu}(\boldsymbol{\theta}^*)|_2 = O_{\mathbb{P}}\left(\frac{\alpha_n\sqrt{p}}{\sqrt{n}} + \sqrt{\frac{p\log n}{mn}} + \frac{p^{1/2}\log^{3/4} n}{n^{1/2}m^{3/4}} \right).$$

Note that here we have a $\sqrt{\log n}$ in the second term because of the diverging dimension p. Thus we have

$$\begin{aligned}
|\boldsymbol{T}_4|_2 \le &|\bar{\boldsymbol{g}}(\boldsymbol{\theta}^*) - \boldsymbol{\mu}(\boldsymbol{\theta}^*)|_2 + \sup_{\theta\in\Theta_0} \left|\bar{\boldsymbol{g}}(\boldsymbol{\theta}) - \boldsymbol{\mu}(\boldsymbol{\theta}) - \bar{\boldsymbol{g}}(\boldsymbol{\theta}^*) + \boldsymbol{\mu}(\boldsymbol{\theta}^*)\right|_2 \\
= &O_{\mathbb{P}}\left(\frac{\alpha_n\sqrt{p}}{\sqrt{n}} + \sqrt{\frac{p\log n}{mn}} + \frac{p^{1/2}\log^{3/4} n}{n^{1/2}m^{3/4}} + r_n\sqrt{\frac{p^2\log n}{n}} \right),
\end{aligned}$$

Thus there exists a constant $\widetilde{C}$ such that

$$\|\boldsymbol{T}_1\| = o(1), \quad \text{and} \quad |\boldsymbol{T}_2|_2 + |\boldsymbol{T}_3|_2 + |\boldsymbol{T}_4|_2 \le \widetilde{C}\left(\frac{\alpha_n\sqrt{p}}{\sqrt{n}} + \sqrt{\frac{p\log n}{mn}} + \frac{p^{1/2}\log^{3/4} n}{n^{1/2}m^{3/4}} + r_n\sqrt{\frac{p^2\log n}{n}} \right).$$

Now in the view of Assumption B, we continue with (75), there is

$$\begin{aligned}
&\frac{1}{n}\sum_{i\in\mathcal{H}_0}\{f(X_i,\boldsymbol{\theta}_1)-f(X_i,\boldsymbol{\theta}^*)\}-\langle \boldsymbol{g}_0(\widehat{\boldsymbol{\theta}}^{(0)})-\bar{\boldsymbol{g}}(\widehat{\boldsymbol{\theta}}^{(0)}),\boldsymbol{\theta}_1-\boldsymbol{\theta}^*\rangle\\
\geq&\frac{\rho_0}{2}|\boldsymbol{\theta}-\boldsymbol{\theta}^*|_2^2-\|\boldsymbol{T}_1\|\cdot|\boldsymbol{\theta}-\boldsymbol{\theta}^*|_2^2-|\boldsymbol{T}_2|_2\cdot|\boldsymbol{\theta}-\boldsymbol{\theta}^*|_2-|\boldsymbol{T}_3|_2\cdot|\boldsymbol{\theta}-\boldsymbol{\theta}^*|_2-|\boldsymbol{T}_4|_2\cdot|\boldsymbol{\theta}-\boldsymbol{\theta}^*|_2\\
\geq&\frac{\rho_0 b_n^2}{2}-o\left(b_n^2\right)-\widetilde{C}b_n\left(\frac{\alpha_n\sqrt{p}}{\sqrt{n}}+\sqrt{\frac{p\log n}{mn}}+\frac{p^{1/2}\log^{3/4}n}{n^{1/2}m^{3/4}}+r_n\sqrt{\frac{p^2\log n}{n}}\right)>0,
\end{aligned}$$

provided $b_n=O(\frac{\alpha_n\sqrt{p}}{\sqrt{n}}+\sqrt{\frac{p\log n}{mn}}+\frac{p^{1/2}\log^{3/4}n}{n^{1/2}m^{3/4}}+r_n\sqrt{\frac{p^2\log n}{n}})$. So we have (74) holds true, and thus

$$|\widehat{\boldsymbol{\theta}}^{(1)}-\boldsymbol{\theta}^*|_2=O_{\mathbb{P}}\left(\frac{\alpha_n\sqrt{p}}{\sqrt{n}}+\sqrt{\frac{p\log n}{mn}}+\frac{p^{1/2}\log^{3/4}n}{n^{1/2}m^{3/4}}+r_n\sqrt{\frac{p^2\log n}{n}}\right),$$

which proves Theorem 19. Apply this formula inductively, we can obtain Theorem 11. ■

Lemma 27 *Let $\bar{\boldsymbol{g}}(\boldsymbol{\theta})$ be defined as* (18)*, with $\widehat{\boldsymbol{\theta}}^{(0)}$ replaced by $\boldsymbol{\theta}$. Then under the Assumption A-G, there is*

$$\sup_{\boldsymbol{\theta}\in\Theta_0}\left|\bar{\boldsymbol{g}}(\boldsymbol{\theta})-\boldsymbol{\mu}(\boldsymbol{\theta})-\bar{\boldsymbol{g}}(\boldsymbol{\theta}^*)+\boldsymbol{\mu}(\boldsymbol{\theta}^*)\right|_2=O_{\mathbb{P}}\left(\frac{\alpha_n\sqrt{p}}{\sqrt{n}}+\sqrt{\frac{p\log n}{mn}}+\frac{\sqrt{p}\log n}{m\sqrt{n}}+r_n\sqrt{\frac{p^2\log n}{n}}\right).$$

Proof First of all we need to split it into three parts:

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\Theta_0}|\bar{\boldsymbol{g}}(\boldsymbol{\theta})-\boldsymbol{\mu}(\boldsymbol{\theta})-\bar{\boldsymbol{g}}(\boldsymbol{\theta}^*)+\boldsymbol{\mu}(\boldsymbol{\theta}^*)|_2\\
\leq&\underbrace{\sup_{\boldsymbol{\theta}\in\Theta_0}|\widehat{\boldsymbol{g}}(\boldsymbol{\theta})-\boldsymbol{\mu}(\boldsymbol{\theta})-\widehat{\boldsymbol{g}}(\boldsymbol{\theta}^*)+\boldsymbol{\mu}(\boldsymbol{\theta}^*)|_2}_{T_{41}}\\
&+\underbrace{\sup_{\boldsymbol{\theta}\in\Theta_0}\max_{1\leq l\leq p}\frac{3\sqrt{p}}{m\sqrt{n}}\sum_{k=1}^{K}\Big|\{\widehat{\sigma}_l(\boldsymbol{\theta})-\widehat{\sigma}_l(\boldsymbol{\theta}^*)\}\sum_{j=0}^{m}\Big[\mathbb{I}\Big\{g_{j,l}(\boldsymbol{\theta})\leq\widehat{g}_l(\boldsymbol{\theta})+\frac{\widehat{\sigma}_l(\boldsymbol{\theta})\Delta_k}{\sqrt{n}}\Big\}-\frac{k}{K+1}\Big]\Big|}_{T_{42}}\\
&+\underbrace{\sup_{\boldsymbol{\theta}\in\Theta_0}\max_{1\leq l\leq p}\frac{3\sqrt{p}}{m\sqrt{n}}\Big|\widehat{\sigma}_l(\boldsymbol{\theta}^*)\sum_{k=1}^{K}\sum_{j=0}^{m}\Big[\mathbb{I}\Big\{g_{j,l}(\boldsymbol{\theta})\leq\widehat{g}_l(\boldsymbol{\theta})+\frac{\widehat{\sigma}_l(\boldsymbol{\theta})\Delta_k}{\sqrt{n}}\Big\}-\mathbb{I}\Big\{g_{j,l}(\boldsymbol{\theta}^*)\leq\widehat{g}_l(\boldsymbol{\theta}^*)+\frac{\widehat{\sigma}_l(\boldsymbol{\theta}^*)\Delta_k}{\sqrt{n}}\Big\}\Big]\Big|}_{T_{43}},
\end{aligned}$$

where the factor 3 comes from the fact $1/\psi(0)=\sqrt{2\pi}<3$. For the first term T_{41}, rehash the proof of equation (77) we can easily obtain

$$\begin{aligned}
&\mathbb{P}\left\{\max_{1\leq l\leq p}\max_{j\notin\mathcal{B}}\sup_{\boldsymbol{\theta}\in\Theta_0}\left|\frac{1}{n}\sum_{i\in\mathcal{H}_j}Z_l(X_i,\boldsymbol{\theta})\right|\geq Cr_n\sqrt{\frac{p\log n}{n}}\right\}\\
=&\mathbb{P}\left\{\max_{1\leq l\leq p}\max_{j\notin\mathcal{B}}\sup_{\boldsymbol{\theta}\in\Theta_0}\left|\frac{1}{n}\sum_{i\in\mathcal{H}_j}\langle\boldsymbol{Z}(X_i,\boldsymbol{\theta}),\boldsymbol{e}_l\rangle\right|\geq Cr_n\sqrt{\frac{p\log n}{n}}\right\}\leq mpn^{-\gamma p}.
\end{aligned}\tag{79}$$

where $Z_l(X_i, \boldsymbol{\theta})$ is the l'th coordinate of $\boldsymbol{Z}(X_i, \boldsymbol{\theta})$ defined in (76). Then from Lemma 22 (with $q = 1/2$ and $\alpha = \alpha_n$), we know

$$\begin{aligned}
&\max_{1\le l\le p}\sup_{\boldsymbol{\theta}\in\Theta_0}|\widehat{g}_l(\boldsymbol{\theta})-\mu_l(\boldsymbol{\theta})-\widehat{g}_l(\boldsymbol{\theta}^*)+\mu_l(\boldsymbol{\theta}^*)|\\
\le& 2\max_{1\le l\le p}\max_{j\notin\mathcal{B}}\sup_{\boldsymbol{\theta}\in\Theta_0}\left|\frac{1}{n}\sum_{i\in\mathcal{H}_j}Z_l(X_i,\boldsymbol{\theta})\right|+\max_{1\le l\le p}\left|\widehat{g}_l^{\mathcal{B},(1-2\alpha_n)/(2-2\alpha_n)}(\boldsymbol{\theta}^*)-\widehat{g}_l^{\mathcal{B},1/(2-2\alpha_n)}(\boldsymbol{\theta}^*)\right|\\
\le& 2\max_{1\le l\le p}\max_{j\notin\mathcal{B}}\sup_{\boldsymbol{\theta}\in\Theta_0}\left|\frac{1}{n}\sum_{i\in\mathcal{H}_j}Z_l(X_i,\boldsymbol{\theta})\right|+2\max_{1\le l\le p}\max\Big\{\Big|\widehat{g}_l^{\mathcal{B},(1-2\alpha_n)/(2-2\alpha_n)}(\boldsymbol{\theta}^*)-\mu_l(\boldsymbol{\theta}^*)\Big|,\\
&\Big|\widehat{g}_l^{\mathcal{B},1/(2-2\alpha_n)}(\boldsymbol{\theta}^*)-\mu_l(\boldsymbol{\theta}^*)\}\Big|\Big\},
\end{aligned}$$

where $\widehat{g}_l^{\mathcal{B},(1-2\alpha_n)/(2-2\alpha_n)}(\boldsymbol{\theta}^*), \widehat{g}_l^{\mathcal{B},1/(2-2\alpha_n)}(\boldsymbol{\theta}^*)$ represent the $(1-2\alpha_n)/(2-2\alpha_n)$-th and $1/(2-2\alpha_n)$-th quantile of non-Byzantine machines respectively. For the additional term, we can follow the proof of Lemma 16 and show

$$\max_{1\le l\le p}\max\left\{\left|\widehat{g}_l^{\mathcal{B},(1-2\alpha_n)/(2-2\alpha_n)}(\boldsymbol{\theta}^*)-\mu_l(\boldsymbol{\theta}^*)\right|,\left|\widehat{g}_l^{\mathcal{B},1/(2-2\alpha_n)}(\boldsymbol{\theta}^*)-\mu_l(\boldsymbol{\theta}^*)\right|\right\}=O_{\mathbb{P}}\left(\frac{\alpha_n}{\sqrt{n}}+\sqrt{\frac{\log n}{mn}}\right),$$

provided condition $\alpha_n \le 1/2-\delta$ holds for some $\delta\in(0,1/2)$. Therefore we have

$$\max_{1\le l\le p}\sup_{\boldsymbol{\theta}\in\Theta_0}|\widehat{g}_l(\boldsymbol{\theta})-\mu_l(\boldsymbol{\theta})-\widehat{g}_l(\boldsymbol{\theta}^*)+\mu_l(\boldsymbol{\theta}^*)|=O_{\mathbb{P}}\left(\frac{\alpha_n}{\sqrt{n}}+\sqrt{\frac{\log n}{mn}}+r_n\sqrt{\frac{p^2\log n}{n}}\right). \quad (80)$$

Taking all coordinate together we have

$$|T_{41}|=O_{\mathbb{P}}\left(\frac{\alpha_n\sqrt{p}}{\sqrt{n}}+\sqrt{\frac{p\log n}{mn}}+r_n\sqrt{\frac{p^2\log n}{n}}\right). \quad (81)$$

The rate of T_{42} hinges on the uniform rate of $\widehat{\sigma}(\boldsymbol{\theta})-\widehat{\sigma}(\boldsymbol{\theta}^*)$ over Θ_0. Indeed, in Lemma 25 we proved

$$\max_{1\le l\le p}\sup_{\boldsymbol{\theta}\in\Theta_0}|\widehat{\sigma}_l(\boldsymbol{\theta})-\widehat{\sigma}_l(\boldsymbol{\theta}^*)|=O_{\mathbb{P}}\left(r_n\right), \quad (82)$$

provided $p=O(\sqrt{n}\log^{-1}n)$ in Assumption G. So this yields

$$|T_{42}|\le\frac{3\sqrt{p}}{\sqrt{n}}\max_{1\le l\le p}\sup_{\boldsymbol{\theta}\in\Theta_0}|\widehat{\sigma}_l(\boldsymbol{\theta})-\widehat{\sigma}_l(\boldsymbol{\theta}^*)|=O_{\mathbb{P}}\left(\frac{r_n\sqrt{p}}{\sqrt{n}}\right). \quad (83)$$

It left to deal with the term T_{43}. From (79), (80), (82) and (62), we know there exists a constant C large enough, such that the following inequalities holds uniformly:

$$\begin{aligned}
\max_{1\le l\le p}\max_{j\notin\mathcal{B}}\sup_{\boldsymbol{\theta}\in\Theta_0}|g_{l,j}(\boldsymbol{\theta})-\mu_l(\boldsymbol{\theta})-g_{l,j}(\boldsymbol{\theta}^*)+\mu_l(\boldsymbol{\theta}^*)|&\le Cr_n\sqrt{\frac{p\log n}{n}},\\
\max_{1\le l\le p}\sup_{\boldsymbol{\theta}\in\Theta_0}|\widehat{g}_l(\boldsymbol{\theta})-\mu_l(\boldsymbol{\theta})-\widehat{g}_l(\boldsymbol{\theta}^*)+\mu_l(\boldsymbol{\theta}^*)|&\le C\left(\frac{\alpha_n}{\sqrt{n}}+\sqrt{\frac{\log n}{mn}}+r_n\sqrt{\frac{p\log n}{n}}\right),\\
\max_{1\le l\le p}\max_{1\le k\le K}\sup_{\boldsymbol{\theta}\in\Theta_0}|\widehat{\sigma}_l(\boldsymbol{\theta})-\widehat{\sigma}_l(\boldsymbol{\theta}^*)|\,\Delta_k n^{-1/2}&\le C\frac{r_n}{\sqrt{n}},\\
\max_{1\le l\le p}|\widehat{\sigma}_l(\boldsymbol{\theta}^*)|&\le C,
\end{aligned}$$

with probability higher than $1-O(n^{-\gamma})$. Under this event, using Lemma 26 with

$$\delta_n=C\left(\frac{\alpha_n}{\sqrt{n}}+\sqrt{\frac{\log n}{mn}}+3r_n\sqrt{\frac{p\log n}{n}}\right),$$

we have

$$\begin{aligned}
|T_{43}|\le&\frac{C\sqrt{p}}{\sqrt{n}}\max_{1\le k\le K}\max_{1\le l\le p}\frac{1}{m+1}\sum_{j=0}^{m}\mathbb{I}\left\{\left|g_{j,l}(\boldsymbol{\theta}^*)-\widehat{g}_l(\boldsymbol{\theta}^*)-\frac{\widehat{\sigma}_l(\boldsymbol{\theta}^*)\Delta_k}{\sqrt{n}}\right|\le\delta_n\right\}\\
\le&\frac{C\alpha_n\sqrt{p}}{\sqrt{n}}+\frac{C\sqrt{p}}{\sqrt{n}}\max_{1\le k\le K}\max_{1\le l\le p}\frac{1}{m+1}\sum_{j\notin\mathcal{B}}\mathbb{I}\left\{\left|g_{j,l}(\boldsymbol{\theta}^*)-\widehat{g}_l(\boldsymbol{\theta}^*)-\frac{\widehat{\sigma}_l(\boldsymbol{\theta}^*)\Delta_k}{\sqrt{n}}\right|\le\delta_n\right\}\\
=&O_{\mathbb{P}}\left(\frac{\alpha_n\sqrt{p}}{\sqrt{n}}+\sqrt{\frac{p\log n}{mn}}+r_n\sqrt{\frac{p^2\log n}{n}}\right),
\end{aligned}\tag{84}$$

holds with probability larger than $1-O(pn^{-\gamma})$. Combining (81), (83) and (84), the lemma is proved. ■

Proof [Proof of Theorem 8] When the iteration number satisfies (24), and the rate constraints satisfies $\alpha_n=o(1\sqrt{mp})$ and $p=o(\min\{\frac{n^{1/3}}{\log^{2/3}n},\frac{m^{1/2}}{\log^{3/2}n}\})$, we clearly know that

$$\widehat{\boldsymbol{\theta}}^{(t)}\in\Theta_t:=\left\{\boldsymbol{\theta}\in\mathbb{R}^p:|\boldsymbol{\theta}-\boldsymbol{\theta}^*|_2\le C\sqrt{\frac{p\log n}{mn}}\right\},$$

with high probability, for some C sufficiently large. Moreover, we need sharper rate for $|\boldsymbol{T}_4|_2$ in (75). Indeed, from Lemma 22 we have

$$\begin{aligned}
&\max_{1\le l\le p}\sup_{\boldsymbol{\theta}\in\Theta_t}|\widehat{g}_l(\boldsymbol{\theta})-\mu_l(\boldsymbol{\theta})-\widehat{g}_l(\boldsymbol{\theta}^*)+\mu_l(\boldsymbol{\theta}^*)|\\
\le&2\max_{1\le l\le p}\max_{j\notin\mathcal{B}}\sup_{\boldsymbol{\theta}\in\Theta_t}\left|\frac{1}{n}\sum_{i\in\mathcal{H}_j}Z_l(X_i,\boldsymbol{\theta})\right|+\max_{1\le l\le p}\left|\widehat{g}_l^{\mathcal{B},(1-2\alpha_n)/(2-2\alpha_n)}(\boldsymbol{\theta}^*)-\widehat{g}_l^{\mathcal{B},1/(2-2\alpha_n)}(\boldsymbol{\theta}^*)\right|.
\end{aligned}$$

Note that $1/(2-2\alpha_n)-(1-2\alpha_n)/(2-2\alpha_n)=o(1/\sqrt{mp})$, by Lemma 21 we have sharper constraint on the quantile gap

$$\max_{1\le l\le p}\left|\widehat{g}_l^{\mathcal{B},(1-2\alpha_n)/(2-2\alpha_n)}(\boldsymbol{\theta}^*)-\widehat{g}_l^{\mathcal{B},1/(2-2\alpha_n)}(\boldsymbol{\theta}^*)\right|=O_{\mathbb{P}}\left(\frac{\alpha_n}{\sqrt{n}}+\frac{1}{n}+\frac{\log n}{m\sqrt{n}}\right).$$

Therefore we have

$$\begin{aligned}&|\widehat{\boldsymbol{g}}(\widehat{\boldsymbol{\theta}}^{(t)})-\boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}^{(t)})-\widehat{\boldsymbol{g}}(\boldsymbol{\theta}^*)+\boldsymbol{\mu}(\boldsymbol{\theta}^*)|_2\\ =&O_{\mathbb{P}}\left(\frac{\alpha_n\sqrt{p}}{\sqrt{n}}+\frac{\sqrt{p}}{n}+\frac{\sqrt{p}\log n}{m\sqrt{n}}+\frac{p^{3/2}\log n}{\sqrt{m}n}\right).\end{aligned}$$

Then follow the proof of Lemma 27 we have

$$\begin{aligned}&|\bar{\boldsymbol{g}}(\widehat{\boldsymbol{\theta}}^{(t)})-\boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}^{(t)})-\bar{\boldsymbol{g}}(\boldsymbol{\theta}^*)+\boldsymbol{\mu}(\boldsymbol{\theta}^*)|_2\\ =&O_{\mathbb{P}}\left(\frac{\alpha_n\sqrt{p}}{\sqrt{n}}+\frac{\sqrt{p}}{n}+\frac{\sqrt{p}\log n}{m\sqrt{n}}+\frac{p^{3/2}\log n}{\sqrt{m}n}\right)=o_{\mathbb{P}}\left(\frac{1}{\sqrt{mn}}\right).\end{aligned}\tag{85}$$

Now we start to prove asymptotic normality. From equations (1) and (20), we know

$$\boldsymbol{\mu}(\boldsymbol{\theta}^*)=0,\quad \text{and } \boldsymbol{g}_0(\widehat{\boldsymbol{\theta}}^{(t+1)})=\boldsymbol{g}_0(\widehat{\boldsymbol{\theta}}^{(t)})-\bar{\boldsymbol{g}}(\widehat{\boldsymbol{\theta}}^{(t)}).$$

Therefore, from (85) and (77) there is

$$\begin{aligned}&\left|\boldsymbol{g}_0(\widehat{\boldsymbol{\theta}}^{(t+1)})-\boldsymbol{g}_0(\boldsymbol{\theta}^*)\right|_2\\ =&\left|\boldsymbol{g}_0(\widehat{\boldsymbol{\theta}}^{(t)})-\bar{\boldsymbol{g}}(\widehat{\boldsymbol{\theta}}^{(t)})-\boldsymbol{g}_0(\boldsymbol{\theta}^*)\right|_2\\ =&\Big|\boldsymbol{\mu}(\boldsymbol{\theta}^*)-\bar{\boldsymbol{g}}(\boldsymbol{\theta}^*)+\bar{\boldsymbol{g}}(\boldsymbol{\theta}^*)-\bar{\boldsymbol{g}}(\widehat{\boldsymbol{\theta}}^{(t)})-\boldsymbol{\mu}(\boldsymbol{\theta}^*)+\boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}^{(t)})\\ &+\boldsymbol{g}_0(\widehat{\boldsymbol{\theta}}^{(t)})-\boldsymbol{g}_0(\boldsymbol{\theta}^*)-\boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}^{(t)})+\boldsymbol{\mu}(\boldsymbol{\theta}^*)\Big|_2\\ =&|\boldsymbol{\mu}(\boldsymbol{\theta}^*)-\bar{\boldsymbol{g}}(\boldsymbol{\theta}^*)|_2+o_{\mathbb{P}}\left(\frac{1}{\sqrt{mn}}\right).\end{aligned}\tag{86}$$

On the other hand, from Assumption C and equation (77)

$$\begin{aligned}&\left|\boldsymbol{g}_0(\widehat{\boldsymbol{\theta}}^{(t+1)})-\boldsymbol{g}_0(\boldsymbol{\theta}^*)\right|_2\\ =&\left|\boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}^{(t+1)})-\boldsymbol{\mu}(\boldsymbol{\theta}^*)+\boldsymbol{g}_0(\widehat{\boldsymbol{\theta}}^{(t+1)})-\boldsymbol{g}_0(\boldsymbol{\theta}^*)-\boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}^{(t+1)})+\boldsymbol{\mu}(\boldsymbol{\theta}^*)\right|_2\\ =&\Big|\nabla\boldsymbol{\mu}(\boldsymbol{\theta}^*)\cdot(\widehat{\boldsymbol{\theta}}^{(t+1)}-\boldsymbol{\theta}^*)+\int_0^1\left\{\nabla\boldsymbol{\mu}(\boldsymbol{\theta}^*+s(\widehat{\boldsymbol{\theta}}^{(t+1)}-\boldsymbol{\theta}^*))-\nabla\boldsymbol{\mu}(\boldsymbol{\theta}^*)\right\}\mathrm{d}s\cdot(\widehat{\boldsymbol{\theta}}^{(t+1)}-\boldsymbol{\theta}^*)\\ &+\boldsymbol{g}_0(\widehat{\boldsymbol{\theta}}^{(t+1)})-\boldsymbol{g}_0(\boldsymbol{\theta}^*)-\boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}^{(t+1)})+\boldsymbol{\mu}(\boldsymbol{\theta}^*)\Big|_2\\ =&\left|\nabla\boldsymbol{\mu}(\boldsymbol{\theta}^*)\cdot(\widehat{\boldsymbol{\theta}}^{(t+1)}-\boldsymbol{\theta}^*)\right|_2+O_{\mathbb{P}}\left(\frac{p\log n}{\sqrt{m}n}\right)=\left|\nabla\boldsymbol{\mu}(\boldsymbol{\theta}^*)\cdot(\widehat{\boldsymbol{\theta}}^{(t+1)}-\boldsymbol{\theta}^*)\right|_2+o_{\mathbb{P}}\left(\frac{1}{\sqrt{mn}}\right).\end{aligned}\tag{87}$$

We can combine (86) and (87), rearrange the terms, then there is

$$\left|\nabla\boldsymbol{\mu}(\boldsymbol{\theta}^*)\cdot(\widehat{\boldsymbol{\theta}}^{(t+1)}-\boldsymbol{\theta}^*)\right|_2=|\boldsymbol{\mu}(\boldsymbol{\theta}^*)-\bar{\boldsymbol{g}}(\boldsymbol{\theta}^*)|_2+o_{\mathbb{P}}\left(\frac{1}{\sqrt{mn}}\right). \tag{88}$$

Denote

$$G_{j,l}(x)=\mathbb{P}\left\{\frac{\sqrt{n}g_{j,l}(\boldsymbol{\theta}^*)}{\sigma_l(\boldsymbol{\theta}^*)}\le x\right\},\quad I_{j,l}(x)=\mathbb{I}\left\{\frac{\sqrt{n}g_{j,l}(\boldsymbol{\theta}^*)}{\sigma_l(\boldsymbol{\theta}^*)}\le x\right\}. \tag{89}$$

Then from (36), for every entry we have

$$\begin{aligned}&\mu_l(\boldsymbol{\theta}^*)-\bar{g}_l(\boldsymbol{\theta}^*) \qquad (90)\\ =&\frac{1}{m+1}\sum_{j\notin\mathcal{B}}\frac{\sigma_l(\boldsymbol{\theta}^*)}{\sqrt{n}\sum_{k=1}^K\psi(\Delta_k)}\sum_{k=1}^K\{I_{j,l}(\Delta_k)-G_{j,l}(\Delta_k)\}+O_{\mathbb{P}}\left(\frac{\log n}{m\sqrt{n}}+\frac{1}{n}+\frac{\log^{3/4}n}{n^{1/2}m^{3/4}}\right).\end{aligned}$$

From equations (88), (90) and the rates constrains, for any vector $\widetilde{\boldsymbol{v}}\in\mathbb{R}^p$, there is

$$\begin{aligned}&\left\langle\nabla\boldsymbol{\mu}(\boldsymbol{\theta}^*)\cdot(\widehat{\boldsymbol{\theta}}^{(t+1)}-\boldsymbol{\theta}^*),\widetilde{\boldsymbol{v}}\right\rangle\\ =&\frac{1}{m+1}\sum_{j\notin\mathcal{B}}\frac{1}{\sqrt{n}\sum_{k=1}^K\psi(\Delta_k)}\sum_{k=1}^K\sum_{l=1}^p\{\sigma_l(\boldsymbol{\theta}^*)I_{j,l}(\Delta_k)-\sigma_l(\boldsymbol{\theta}^*)G_{j,l}(\Delta_k)\}\widetilde{v}_l+o_{\mathbb{P}}\left(\frac{1}{\sqrt{mn}}\right).\end{aligned}$$

Now we apply central limit theorem and yield

$$\begin{aligned}&\frac{\sqrt{(m+1)n}}{\widetilde{\sigma}_{\widetilde{\boldsymbol{v}}}}\left\langle\nabla\boldsymbol{\mu}(\boldsymbol{\theta}^*)\cdot(\widehat{\boldsymbol{\theta}}^{(t+1)}-\boldsymbol{\theta}^*),\widetilde{\boldsymbol{v}}\right\rangle\xrightarrow{d}\mathcal{N}(0,1),\\ &\text{where}\quad\widetilde{\sigma}_{\widetilde{\boldsymbol{v}}}^2=\widetilde{\boldsymbol{v}}^{\mathrm{T}}\widetilde{\boldsymbol{\mathcal{C}}}\widetilde{\boldsymbol{v}}.\end{aligned}$$

Here $\widetilde{\boldsymbol{\mathcal{C}}}\in\mathbb{R}^{p\times p}$ has its (l_1,l_2)-entry defined as

$$\begin{aligned}\widetilde{\mathcal{C}}_{l_1,l_2}=&\frac{\sigma_{l_1}(\boldsymbol{\theta}^*)\sigma_{l_2}(\boldsymbol{\theta}^*)}{\{\sum_{k=1}^K\psi(\Delta_k)\}^2}\mathbb{E}\left[\sum_{k=1}^K\{I_{0,l_1}(\Delta_k)-G_{0,l_1}(\Delta_k)\}\sum_{k=1}^K\{I_{0,l_2}(\Delta_k)-G_{0,l_2}(\Delta_k)\}\right]\\ =&\frac{\sqrt{\sigma_{l_1,l_1}\sigma_{l_2,l_2}}}{\{\sum_{k=1}^K\psi(\Delta_k)\}^2}\sum_{k_1,k_2}\left\{\mathbb{P}\left(\frac{\sqrt{n}g_{0,l_1}(\boldsymbol{\theta}^*)}{\sqrt{\sigma_{l_1,l_1}}}\le\Delta_{k_1},\frac{\sqrt{n}g_{0,l_2}(\boldsymbol{\theta}^*)}{\sqrt{\sigma_{l_2,l_2}}}\le\Delta_{k_2}\right)-G_{0,l_1}(\Delta_{k_1})G_{0,l_2}(\Delta_{k_2})\right\},\end{aligned}$$

since $\sigma_l(\boldsymbol{\theta}^*)=\sqrt{\mathrm{Var}\{\nabla f_l(X,\boldsymbol{\theta}^*)\}}=\sqrt{\sigma_{l,l}}$. Moreover, we can apply multivariate Berry-Esseen theorem (See Theorem 1.3 in Götze (1991)) and give

$$\widetilde{\mathcal{C}}_{l_1,l_2}=\mathcal{C}_{l_1,l_2}+O(n^{-1/2}),$$

where $\mathcal{C}_{l_1,l_2}$ is defined in (14). Now we replace $\widetilde{\boldsymbol{v}}$ with $\{\nabla\boldsymbol{\mu}(\boldsymbol{\theta}^*)\}^{-1}\boldsymbol{v}$. From Assumption B we know the norm of $\boldsymbol{v}\subseteq\mathbb{R}^p$ is rescaled by a factor of constant order. Thus the theorem is proved. ■

Appendix D. Examples Verification

In this appendix, we will verify that a large class of generalized linear models and M-estimators satisfy our proposed Assumptions A–F.

D.1 Generalized linear models

For a generalized linear models (GLM) with the canonical link function $\mathcal{L} : \mathbb{R} \to \mathbb{R}$, each *i.i.d.* observation $(\boldsymbol{X}, Y) \in \mathbb{R}^{p+1}$ admits the following conditional probability function

$$\mathbb{P}(Y \mid \boldsymbol{X}) = \widetilde{c} \exp\left\{ \frac{Y\langle \boldsymbol{\theta}^*, \boldsymbol{X}\rangle - \mathcal{L}(\langle \boldsymbol{\theta}^*, \boldsymbol{X}\rangle)}{c(\sigma)} \right\}, \tag{91}$$

where $\widetilde{c}$ and $c(\sigma)$ are some constants, $\boldsymbol{\theta}^*$ is the true parameter. The loss function based on the maximum likelihood estimator is defined by,

$$f(Y, \boldsymbol{X}, \boldsymbol{\theta}) = -Y\langle \boldsymbol{\theta}, \boldsymbol{X}\rangle + \mathcal{L}(\langle \boldsymbol{\theta}, \boldsymbol{X}\rangle). \tag{92}$$

Then we have the following proposition.

Proposition 28 *Let $(\boldsymbol{X}, Y)$ be observation of a generalized linear model* (91) *with a convex link function $\mathcal{L}$. Suppose the following condition holds:*

(C1) There exists $\rho_0 > 0$ such that

$$\inf_{\boldsymbol{v} \in \mathbb{S}^{p-1}} \mathbb{E}\left\{ \mathcal{L}''(\langle \boldsymbol{\theta}^*, \boldsymbol{X}\rangle) |\langle \boldsymbol{v}, \boldsymbol{X}\rangle|^2 \right\} \geq \rho_0;$$

(C2) There exists $M > 0$ such that

$$|\mathcal{L}''(x)| \leq M, \quad |\mathcal{L}''(x_1) - \mathcal{L}''(x_2)| \leq M|x_1 - x_2|;$$

(C3) There exists $\eta, C_M > 0$ such that

$$\sup_{\boldsymbol{v} \in \mathbb{S}^{p-1}} \mathbb{E}\left[\exp\left\{ \eta |\langle \boldsymbol{v}, \boldsymbol{X}\rangle|^2 \right\} \right] \leq C_M, \quad \mathbb{E}\left[\exp \eta |\mathcal{L}'(\langle \boldsymbol{\theta}^*, \boldsymbol{X}\rangle) - Y|^2 \right] \leq C_M.$$

Then the loss function defined in (92) *satisfies Assumption A-F.*

Condition (C1) and (C3) imply that the covariate $\boldsymbol{X}$ is a non-degenerate subgaussian vector. And another part of condition (C3) shows that $\mathcal{L}'(\langle \boldsymbol{\theta}^*, \boldsymbol{X}\rangle) - Y = \mathbb{E}[Y|\boldsymbol{X}] - Y$ has a subgaussian tail. For convenience of verification, we assume the $\mathcal{L}''$ to be Lipschitz continuous in Condition (C2).

Proof [Proof of Proposition 28] Firstly we can compute the gradient and Hessian as follows,

$$\nabla f(Y, \boldsymbol{X}, \boldsymbol{\theta}) = -Y\boldsymbol{X} + \mathcal{L}'(\langle \boldsymbol{\theta}, \boldsymbol{X}\rangle)\boldsymbol{X}, \qquad \boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbb{E}\big\{ \mathcal{L}'(\langle \boldsymbol{\theta}, \boldsymbol{X}\rangle)\boldsymbol{X} - \mathcal{L}'(\langle \boldsymbol{\theta}^*, \boldsymbol{X}\rangle)\boldsymbol{X} \big\},$$
$$\nabla \boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbb{E}\left\{ \mathcal{L}''(\langle \boldsymbol{\theta}, \boldsymbol{X}\rangle)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} \right\}.$$

Then we can verify these assumptions one by one.

- Assumption B: Compute that

$$\begin{aligned}\Lambda_{\max}\{\nabla\boldsymbol{\mu}(\boldsymbol{\theta}^*)\} =& \sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \mathbb{E}\left\{\mathcal{L}''(\langle\boldsymbol{\theta}^*,\boldsymbol{X}\rangle)|\langle\boldsymbol{v},\boldsymbol{X}\rangle|^2\right\}\\ \le& M\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \mathbb{E}|\langle\boldsymbol{v},\boldsymbol{X}\rangle|^2 \le \frac{M}{\eta}C_M,\\ \Lambda_{\min}\{\nabla\boldsymbol{\mu}(\boldsymbol{\theta}^*)\} =& \inf_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \mathbb{E}\left\{\mathcal{L}''(\langle\boldsymbol{\theta}^*,\boldsymbol{X}\rangle)|\langle\boldsymbol{v},\boldsymbol{X}\rangle|^2\right\}\\ \ge& \rho_0.\end{aligned}$$

- Assumption C: By elementary inequalities $3|xy^2| \le |x|^3 + 2|y|^3$ and $2|x|^3 \le e^{x^2}$, we have

$$\begin{aligned}&\|\nabla\boldsymbol{\mu}(\boldsymbol{\theta}_1) - \nabla\boldsymbol{\mu}(\boldsymbol{\theta}_2)\|\\ \le& \sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \mathbb{E}\left\{\left|\mathcal{L}''(\langle\boldsymbol{\theta}_1,\boldsymbol{X}\rangle) - \mathcal{L}''(\langle\boldsymbol{\theta}_2,\boldsymbol{X}\rangle)\right|\cdot|\langle\boldsymbol{v},\boldsymbol{X}\rangle|^2\right\}\\ \le& M\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \mathbb{E}\left\{|\langle\boldsymbol{\theta}_1-\boldsymbol{\theta}_2,\boldsymbol{X}\rangle|\cdot|\langle\boldsymbol{v},\boldsymbol{X}\rangle|^2\right\}\\ \le& M\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \mathbb{E}\left(|\langle\boldsymbol{v},\boldsymbol{X}\rangle|^3\right)|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2|_2 \le \frac{MC_M}{2\eta^{3/2}}|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2|_2.\end{aligned}$$

- Assumption D:

$$\begin{aligned}M_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}(Y,\boldsymbol{X},\boldsymbol{v}) =& \frac{1}{|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2|_2}\left|\langle\boldsymbol{v},\boldsymbol{X}\rangle\left\{\mathcal{L}'(\langle\boldsymbol{\theta}_1,\boldsymbol{X}\rangle) - \mathcal{L}'(\langle\boldsymbol{\theta}_2,\boldsymbol{X}\rangle)\right\}\right|\\ \le& M|\langle\boldsymbol{v},\boldsymbol{X}\rangle|\cdot\left|\left\langle\frac{\boldsymbol{\theta}_1-\boldsymbol{\theta}_2}{|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2|_2},\,\boldsymbol{X}\right\rangle\right|.\end{aligned}$$

  Thus

$$\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\sup_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}\mathbb{E}\left[\exp\left\{\frac{\eta}{M}M_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}(Y,\boldsymbol{X},\boldsymbol{v})\right\}\right] \le \sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\mathbb{E}\left\{\exp\left(\eta|\langle\boldsymbol{v},\boldsymbol{X}\rangle|^2\right)\right\} \le C_M.$$

  Similarly

$$\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\overline{M}(Y,\boldsymbol{X},,\boldsymbol{v}) \le \sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} M|\boldsymbol{X}|_2\cdot|\langle\boldsymbol{v},\boldsymbol{X}\rangle| \le M|\boldsymbol{X}|_2^2 = M\sum_{l=1}^{p}|\langle\boldsymbol{X},\boldsymbol{e}_l\rangle|^2,$$

  where $\boldsymbol{e}_l$ is the l-th base vector. Then using generalized Hölder's inequality we can prove

$$\begin{aligned}\mathbb{E}\left[\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\exp\left\{\frac{\eta}{Mp}\overline{M}(Y,\boldsymbol{X},\boldsymbol{v})\right\}\right] =& \mathbb{E}\left[\exp\left\{\frac{\eta}{p}\sum_{l=1}^{p}|\langle\boldsymbol{X},\boldsymbol{e}_l\rangle|^2\right\}\right]\\ \le& \left[\prod_{l=1}^{p}\mathbb{E}\left\{\exp\left(\eta|\langle\boldsymbol{X},\boldsymbol{e}_l\rangle|^2\right)\right\}\right]^{1/p} \le C_M.\end{aligned}$$

- Assumption E: The variance at $\boldsymbol{\theta}^*$ is

$$\begin{aligned}\sigma_l^2(\boldsymbol{\theta}^*) =& \mathbb{E}\left[\left\{-Y + \mathcal{L}'(\langle\boldsymbol{\theta}^*, \boldsymbol{X}\rangle)\right\}^2 X_l^2\right] \\ =& \mathbb{E}\left[c(\sigma)\mathcal{L}''(\langle\boldsymbol{\theta}^*, \boldsymbol{X}\rangle)X_l^2\right],\end{aligned}$$

  then we can bound them as follows

$$\begin{aligned}\sigma_l^2(\boldsymbol{\theta}^*) \leq& c(\sigma)\Lambda_{\max}\{\nabla\boldsymbol{\mu}(\boldsymbol{\theta}^*)\} \leq \eta^{-1}Mc(\sigma)C_M, \\ \sigma_l^2(\boldsymbol{\theta}^*) \geq& c(\sigma)\Lambda_{\min}\{\nabla\boldsymbol{\mu}(\boldsymbol{\theta}^*)\} \geq \rho_0 c(\sigma).\end{aligned}$$

- Assumption F: Using Cauchy inequality we have

$$\begin{aligned}&\mathbb{E}[\exp\eta|\nabla f_l(Y, \boldsymbol{X}, \boldsymbol{\theta}^*) - \mu_l(\boldsymbol{\theta}^*)|] \\ \leq& \mathbb{E}\Big[\exp\eta\left|\mathcal{L}'(\langle\boldsymbol{\theta}^*, \boldsymbol{X}\rangle)X_l - YX_l\right|\Big] \\ \leq& \sqrt{\mathbb{E}\left[\exp\eta|\mathcal{L}'(\langle\boldsymbol{\theta}^*, \boldsymbol{X}\rangle) - Y|^2\right]\mathbb{E}\left[\exp\eta|\langle\boldsymbol{X}_l, \boldsymbol{e}_l\rangle|^2\right]} \leq C_M.\end{aligned}$$

■

As an example, we can show that the logistic regression model satisfies these conditions.

Example 3 *(Logistic regression) In logistic regression model, the response variable Y takes value in $\{0, 1\}$, and the link function is $\mathcal{L}(x) = \log(1 + e^x)$. Then we can compute that*

$$\mathcal{L}'(x) = \frac{1}{1 + e^{-x}}, \quad \mathcal{L}''(x) = \frac{1}{(1 + e^x)(1 + e^{-x})}, \quad |\mathcal{L}'''(x)| \leq 2.$$

It is not hard to verify that conditions (C1)–(C3) in Proposition 28 hold, provided non-degenerate subgaussian covariate $\boldsymbol{X}$.

D.2 M-Estimator

Next we consider the M-estimator. Assume that each *i.i.d.* observation $(\boldsymbol{X}, Y) \in \mathbb{R}^{p+1}$ is generated from the linear model,

$$Y = \langle\boldsymbol{\theta}^*, \boldsymbol{X}\rangle + \epsilon, \tag{93}$$

and the loss function is

$$f(Y, \boldsymbol{X}, \boldsymbol{\theta}) = \mathcal{L}(Y - \langle\boldsymbol{\theta}, \boldsymbol{X}\rangle). \tag{94}$$

Then we have the following proposition.

Proposition 29 *Let $(\boldsymbol{X}, Y)$ be observation of linear model (93) and $\mathcal{L}$ is a convex regression function. Suppose the noise ϵ is independent of the covariate $\boldsymbol{X}$, and $\mathbb{E}\{\mathcal{L}'(\epsilon)\} = 0$. Moreover the following conditions hold true:*

(C1') *There exists* $\rho_0 > 0$ *such that*

$$\min\left\{\mathbb{E}\{\mathcal{L}''(\epsilon)\}, \mathbb{E}\{\mathcal{L}'(\epsilon)^2\}, \inf_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \mathbb{E}|\langle \boldsymbol{v}, \boldsymbol{X}\rangle|^2\right\} \geq \rho_0;$$

(C2') *There exists a constant* $M > 0$ *such that*

$$|\mathcal{L}''(x)| \leq M, \qquad |\mathcal{L}''(x_1) - \mathcal{L}''(x_2)| \leq M|x_1 - x_2|;$$

(C3') *There exists* $\eta, C_M > 0$ *such that*

$$\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \mathbb{E}\left[\exp\left\{\eta|\langle \boldsymbol{v}, \boldsymbol{X}\rangle|^2\right\}\right] \leq C_M, \quad \mathbb{E}\left[\exp\left\{\eta|\mathcal{L}'(\epsilon)|^2\right\}\right] \leq C_M.$$

Then the loss function defined in (94) *satisfies Assumptions A-F.*

Again condition (C1') implies the non-degeneracy of covariate $\boldsymbol{X}$. The first half of condition (C3') requires covariate $\boldsymbol{X}$ to be sub-gaussian. While the noise, depending on the explicit formulation of $\mathcal{L}$, can be heavy-tailed. As will be seen in the following, the noise has to be sub-gaussian in linear regression, but is allowed to be heavy-tailed in Huber regression. It is worthwhile noting that for the ease of presentation, in condition (C2') we simply assume $\mathcal{L}''$ to be Lipschitz continuous. However, in Example 2, we will prove that the Huber regression model satisfies all assumptions in Section C.1.1.

Proof [Proof of Proposition 29] We can directly compute the gradient and Hessian as follows:

$$\begin{aligned}
\nabla f(Y, \boldsymbol{X}, , \boldsymbol{\theta}) &= \mathcal{L}'(\langle \boldsymbol{X}, \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle + \epsilon)\boldsymbol{X}, \\
\nabla \boldsymbol{\mu}(\boldsymbol{\theta}) &= \mathbb{E}\{\mathcal{L}''(\langle \boldsymbol{X}, \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle + \epsilon)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\}.
\end{aligned}$$

Now we verify those assumptions.

- Assumption B:

$$\begin{aligned}
\Lambda_{\max}\{\nabla \boldsymbol{\mu}(\boldsymbol{\theta}^*)\} =& \sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \mathbb{E}\left\{\mathcal{L}''(\epsilon)|\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}|^2\right\} \\
=& \mathbb{E}\left\{\mathcal{L}''(\epsilon)\right\} \sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \mathbb{E}|\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}|^2 \leq \eta^{-1}MC_M, \\
\Lambda_{\min}\{\nabla \boldsymbol{\mu}(\boldsymbol{\theta}^*)\} =& \inf_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \mathbb{E}\left\{\mathcal{L}''(\epsilon)|\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}|^2\right\} \\
=& \mathbb{E}\left\{\mathcal{L}''(\epsilon)\right\} \inf_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \mathbb{E}|\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}|^2 \geq \rho_0^2.
\end{aligned}$$

- Verification for Assumption C and D are almost the same as the proof in Proposition 28, thus omitted for brevity.

- Assumption E: The variance of the l'th coordinate is

$$\sigma_l^2(\boldsymbol{\theta}^*) = \mathbb{E}\left[\mathcal{L}'(\epsilon)^2 X_l^2\right] = \mathbb{E}\left[\mathcal{L}'(\epsilon)^2\right]\mathbb{E}(X_l^2),$$

  thus

$$\begin{aligned}
\sigma_l^2(\boldsymbol{\theta}^*) \leq& \mathbb{E}\left[\mathcal{L}'(\epsilon)^2\right] \max_{1\leq l\leq p}\{\mathbb{E}(X_l^2)\} \leq \eta^{-2}C_M^2, \\
\sigma_l^2(\boldsymbol{\theta}^*) \geq& \mathbb{E}\left[\mathcal{L}'(\epsilon)^2\right] \min_{1\leq l\leq p}\{\mathbb{E}(X_l^2)\} \geq \rho_0^2.
\end{aligned}$$

- Assumption F: Using Cauchy inequality we have

$$\begin{aligned}&\mathbb{E}\big[\exp\eta|\nabla f_l(Y,\boldsymbol{X},\boldsymbol{\theta}^*)-\mu_l(\boldsymbol{\theta}^*)|\big]\\ \le&\mathbb{E}\Big[\exp\eta\left|\mathcal{L}'(\epsilon)X_l\right|\Big]\\ \le&\sqrt{\mathbb{E}\left[\exp\eta|\mathcal{L}'(\epsilon)|^2\right]\mathbb{E}\left[\exp\eta|\langle\boldsymbol{X},\boldsymbol{e}_l\rangle|^2\right]}\le C_M.\end{aligned}$$

■

Example 1 Continued. In linear regression model, the regression function $\mathcal{L}$ is defined by $\mathcal{L}(x)=x^2/2$. Then we can compute that $\mathcal{L}'(x)=x,\mathcal{L}''(x)=1$. It is relatively straightforward to verify that the conditions in Proposition 29 hold, provided $\boldsymbol{X}$ is a non-degenerate sub-gaussian random vector and the noise ϵ follows a zero-mean sub-gaussian distribution.

Example 2 Continued. In Huber regression model, the regression function $\mathcal{L}$ is defined by

$$\mathcal{L}(x)=\begin{cases}x^2/2 & \text{for } |x|\le\delta,\\ \delta(|x|-\delta/2) & \text{otherwise.}\end{cases}$$

Then we can compute that

$$\mathcal{L}'(x)=\begin{cases}x & \text{for } |x|\le\delta,\\ \delta\,\mathrm{sign}(x) & \text{otherwise,}\end{cases}\qquad \mathcal{L}''(x)=\mathbb{I}(|x|\le\delta).$$

In this case, Proposition 29 is not directly applicable since $\mathcal{L}''(x)$ is not Lipschitz continuous. However, if noise ϵ has a symmetric distribution and uniformly bounded probability density function, Assumption C can be verified as follows.

Proof [Verification for Huber Regression] We only need to show the Lipschitz Hessian assumption C holds true. It is easy to compute the Hessian matrix of Huber loss as follows:

$$\nabla\boldsymbol{\mu}(\boldsymbol{\theta})=\mathbb{E}\left\{\mathbb{I}(|\langle\boldsymbol{X},\boldsymbol{\theta}-\boldsymbol{\theta}^*\rangle+\epsilon|\le\delta)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right\}.$$

Assume the noise ϵ has probability density function $\mathfrak{p}(x)$ uniformly bounded by a constant $M>0$, then by independence of ϵ and $\boldsymbol{X}$, there is

$$\nabla\boldsymbol{\mu}(\boldsymbol{\theta})=\mathbb{E}\left\{\mathbb{P}\Big(|\langle\boldsymbol{X},\boldsymbol{\theta}-\boldsymbol{\theta}^*\rangle+\epsilon|\le\delta\,\Big|\,\boldsymbol{X}\Big)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right\}.$$

Then for arbitrary $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in\mathbb{R}^p$, we have

$$\begin{aligned}&\|\nabla\boldsymbol{\mu}(\boldsymbol{\theta}_1)-\nabla\boldsymbol{\mu}(\boldsymbol{\theta}_2)\|\\ =&\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\mathbb{E}\Big[\left\{\mathbb{P}\Big(|\langle\boldsymbol{X},\boldsymbol{\theta}_1-\boldsymbol{\theta}^*\rangle+\epsilon|\le\delta\,\Big|\,\boldsymbol{X}\Big)-\mathbb{P}\Big(|\langle\boldsymbol{X},\boldsymbol{\theta}_2-\boldsymbol{\theta}^*\rangle+\epsilon|\le\delta\,\Big|\,\boldsymbol{X}\Big)\right\}|\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}|^2\Big]\\ \le&2M\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\mathbb{E}\left\{|\langle\boldsymbol{\theta}_1-\boldsymbol{\theta}_2,\boldsymbol{X}\rangle|\cdot|\langle\boldsymbol{v},\boldsymbol{X}\rangle|^2\right\}\le\eta^{-3/2}MC_M|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2|_2.\end{aligned}$$

■

References

Dan Alistarh, Zeyuan Allen-Zhu, and Jerry Li. Byzantine stochastic gradient descent. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999.

George Bennett. Probability inequalities for the sum of independent random variables. *J. Amer. Statist. Assoc.*, 57(297):33–45, 1962.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Tony Cai and Weidong Liu. Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.*, 106(494):672–684, 2011.

Xi Chen, Weidong Liu, and Yichen Zhang. Quantile regression under memory constraint. *Ann. Statist.*, 47(6):3244–3273, 2019.

Xi Chen, Jason D. Lee, Xin T. Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *Ann. Statist.*, 48(1):251–273, 02 2020a.

Xi Chen, Weidong Liu, Xiaojun Mao, and Zhuoyi Yang. Distributed high-dimensional regression under a quantile loss function. *J. Mach. Learn. Res.*, 21(182):1–43, 2020b.

Xi Chen, Weidong Liu, and Yichen Zhang. First-order newton-type estimator for distributed estimation and inference. *J. Amer. Statist. Assoc. (to appear)*, 2021.

Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):1–25, 2017.

Yuan Shih Chow and Henry Teicher. *Probability Theory: Independence, Interchangeability, Martingales*. Springer New York, 2012.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001.

Jianqing Fan, Yongyi Guo, and Kaizheng Wang. Communication-efficient accurate statistical estimation. *arXiv e-prints arXiv:1906.04870*, 2019.

Jiashi Feng, Huan Xu, and Shie Mannor. Distributed robust learning. *arXiv e-prints arXiv:1409.5937*, 2014.

Friedrich Götze. On the rate of convergence in the multivariate CLT. *Ann. Probab.*, 19(2): 724–739, 1991.

Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *J. Mach. Learn. Res.*, 17(1):543–582, 2016.

Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.*, 43:169 – 188, 1986.

Michael I. Jordan, Jason D. Lee, and Yun Yang. Communication-efficient distributed statistical inference. *J. Amer. Statist. Assoc.*, 114(526):668–681, 2019.

Leslie Lamport, Robert Shostak, and Marshall Pease. The byzantine generals problem. *ACM Trans. Program. Lang. Syst.*, 4(3):382–401, 1982.

Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: Theory and practice. *Ann. Statist.*, 48(2):906–931, 2020.

Jason D Lee, Qiang Liu, Yuekai Sun, and Jonathan E Taylor. Communication-efficient sparse regression. *J. Mach. Learn. Res.*, 18(5):1–30, 2017.

Runze Li, Dennis KJ Lin, and Bing Li. Statistical inference in massive data sets. *Appl. Stoch. Model Bus.*, 29(5):399–409, 2013.

Gábor Lugosi and Shahar Mendelson. Regularization, sparse recovery, and median-of-means tournaments. *Bernoulli*, 25(3):2075–2106, 2019.

Stanislav Minsker. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

Stanislav Minsker. Distributed statistical estimation and rates of convergence in normal approximation. *Electron. J. Statist.*, 13(2):5213–5252, 2019.

Arkadii Semenovich Nemirovsky and David Borisovich Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, NY, 2006.

Mike Paterson. Progress in selection. In *Algorithm Theory — SWAT'96*, pages 368–379, Berlin, Heidelberg, 1996.

Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Jonathan D. Rosenblatt and Boaz Nadler. On the optimality of averaging in distributed statistical learning. *Inf. Inference*, 5(4):379–404, 2016.

Ohad Shamir, Nati Srebro, and Tong Zhang. Communication efficient distributed optimization using an approximate newton-type method. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 1000–1008, 2014.

Zuofeng Shang and Guang Cheng. Computational limits of a distributed algorithm for smoothing spline. *J. Mach. Learn. Res.*, 18:1–37, 2017.

Lili Su and Jiaming Xu. Securing distributed gradient descent in high dimensional statistical learning. *Proc. ACM Meas. Anal. Comput. Syst.*, 3(1), 2019.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 58(1):267–288, 1996.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv e-prints arXiv:1011.3027*, 2010.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Xiaozhou Wang, Zhuoyi Yang, Xi Chen, and Weidong Liu. Distributed inference for linear support vector machine. *J. Mach. Learn. Res.*, 20:1–41, 2019.

Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Generalized Byzantine-tolerant SGD. *arXiv e-prints arXiv:1802.10116*, 2018.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 5650–5659, 2018.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Defending against saddle point attack in Byzantine-robust distributed learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 7074–7084, 2019.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942, 04 2010.

Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Communication-efficient algorithms for statistical optimization. *J. Mach. Learn. Res.*, 14:3321–3363, 2013.

Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *J. Mach. Learn. Res.*, 16:3299–3340, 2015.

Tianqi Zhao, Guang Cheng, and Han Liu. A partially linear framework for massive heterogeneous data. *Ann. Statist.*, 44(4):1400–1437, 2016.

Hui Zou and Ming Yuan. Composite quantile regression and the oracle model selection theory. *Ann. Statist.*, 36(3):1108–1126, 2008.