Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates

Dong Yin [*1], Yudong Chen [†3], Kannan Ramchandran [‡1], and Peter Bartlett [§1,2]

[1]Department of Electrical Engineering and Computer Sciences, UC Berkeley
[2]Department of Statistics, UC Berkeley
[3]School of Operations Research and Information Engineering, Cornell University

February 26, 2021

Abstract

In large-scale distributed learning, security issues have become increasingly important. Particularly in a decentralized environment, some computing units may behave abnormally, or even exhibit Byzantine failures—arbitrary and potentially adversarial behavior. In this paper, we develop distributed learning algorithms that are provably robust against such failures, with a focus on achieving optimal statistical performance. A main result of this work is a sharp analysis of two robust distributed gradient descent algorithms based on median and trimmed mean operations, respectively. We prove statistical error rates for three kinds of population loss functions: strongly convex, non-strongly convex, and smooth non-convex. In particular, these algorithms are shown to achieve order-optimal statistical error rates for strongly convex losses. To achieve better communication efficiency, we further propose a median-based distributed algorithm that is provably robust, and uses only one communication round. For strongly convex quadratic loss, we show that this algorithm achieves the same optimal error rate as the robust distributed gradient descent algorithms.

1 Introduction

Many tasks in computer vision, natural language processing and recommendation systems require learning complex prediction rules from large datasets. As the scale of the datasets in these learning tasks continues to grow, it is crucial to utilize the power of distributed computing and storage. In such large-scale distributed systems, robustness and security issues have become a major concern. In particular, individual computing units—known as worker machines—may exhibit abnormal behavior due to crashes, faulty hardware, stalled computation or unreliable communication channels. Security issues are only exacerbated in the so-called *Federated Learning* setting, a modern distributed learning paradigm that is more decentralized, and that uses the data owners' devices (such as mobile phones and personal computers) as worker machines (Konečnỳ et al., 2016, McMahan and Ramage, 2017). Such machines are often more unpredictable, and in particular may be susceptible to malicious and coordinated attacks.

Due to the inherent unpredictability of this abnormal (sometimes adversarial) behavior, it is typically modeled as *Byzantine failure* (Lamport et al., 1982), meaning that some worker machines may behave completely arbitrarily and can send any message to the master machine that maintains and updates an estimate of the parameter vector to be learned. Byzantine failures can incur major degradation in learning performance. It is well-known that standard learning algorithms based on naive aggregation of the workers' messages can be arbitrarily skewed by a single Byzantine-faulty machine. Even when the messages from Byzantine machines take only moderate values—and hence are difficult to detect—and

*dongyin@berkeley.edu
†yudong.chen@cornell.edu
‡kannanr@berkeley.edu
§peter@berkeley.edu

when the number of such machines is small, the performance loss can still be significant. We demonstrate such an example in our experiments in Section 7.

In this paper, we aim to develop distributed statistical learning algorithms that are provably robust against Byzantine failures. While this objective is considered in a few recent works (Blanchard et al., 2017, Chen et al., 2017, Feng et al., 2014), a fundamental problem remains poorly understood, namely the *optimal statistical performance* of a robust learning algorithm. A learning scheme in which the master machine always outputs zero regardless of the workers' messages is certainly not affected by Byzantine failures, but it will not return anything statistically useful either. On the other hand, many standard distributed algorithms that achieve good statistical performance in the absence of Byzantine failures, become completely unreliable otherwise. Therefore, a main goal of this work is to understand the following questions: what is the best achievable statistical performance while being Byzantine-robust, and what algorithms achieve this performance?

To formalize this question, we consider a standard statistical setting of empirical risk minimization (ERM). Here nm data points are sampled independently from some distribution and distributed evenly among m machines, αm of which are Byzantine. The goal is to learn a parametric model by minimizing some loss function defined by the data. In this statistical setting, one expects that the error in learning the parameter, measured in an appropriate metric, should decrease when the amount of data nm becomes larger and the fraction of Byzantine machines α becomes smaller. In fact, we can show that, at least for strongly convex problems, no algorithm can achieve an error lower than

$$\widetilde{\Omega}\left(\frac{\alpha}{\sqrt{n}}+\frac{1}{\sqrt{nm}}\right)=\widetilde{\Omega}\left(\frac{1}{\sqrt{n}}\left(\alpha+\frac{1}{\sqrt{m}}\right)\right),$$

regardless of communication costs;[1] see Observation 1 in Section 6. Intuitively, the above error rate is the optimal rate that one should target, as $\frac{1}{\sqrt{n}}$ is the effective standard deviation for each machine with n data points, α is the bias effect of Byzantine machines, and $\frac{1}{\sqrt{m}}$ is the averaging effect of m normal machines. When there are no or few Byzantine machines, we see the usual scaling $\frac{1}{\sqrt{mn}}$ with the total number of data points; when some machines are Byzantine, their influence remains bounded, and moreover is proportional to α. If an algorithm is guaranteed to attain this bound, we are assured that we do not sacrifice the quality of learning when trying to guard against Byzantine failures—we pay a price that is unavoidable, but otherwise we achieve the best possible statistical accuracy in the presence of Byzantine failures.

Another important consideration for us is *communication efficiency*. As communication between machines is costly, one cannot simply send all data to the master machine. This constraint precludes direct application of standard robust learning algorithms (such as M-estimators (Huber, 2011)), which assume access to all data. Instead, a desirable algorithm should involve a small number of communication rounds as well as a small amount of data communicated per round. We consider a setting where in each round a worker or master machine can only communicate a vector of size $\mathcal{O}(d)$, where d is the dimension of the parameter to be learned. In this case, the total communication cost is proportional to the number of communication rounds.

To summarize, we aim to develop distributed learning algorithms that simultaneously achieve two objectives:

- **Statistical optimality:** attain an $\widetilde{\mathcal{O}}(\frac{\alpha}{\sqrt{n}}+\frac{1}{\sqrt{nm}})$ rate.
- **Communication efficiency:** $\mathcal{O}(d)$ communication per round, with as few rounds as possible.

To the best of our knowledge, *no existing algorithm achieves these two goals simultaneously*. In particular, previous robust algorithms either have unclear or sub-optimal statistical guarantees, or incur a high communication cost and hence are not applicable in a distributed setting—we discuss related work in more detail in Section 2.

1.1 Our Contributions

We propose two robust distributed gradient descent (GD) algorithms, one based on coordinate-wise median, and the other on coordinate-wise trimmed mean. We establish their statistical error rates for

[1] Throughout the paper, unless otherwise stated, $\Omega(\cdot)$ and $\mathcal{O}(\cdot)$ hide universal multiplicative constants; $\widetilde{\Omega}(\cdot)$ and $\widetilde{\mathcal{O}}(\cdot)$ further hide terms that are independent of α, n, m or logarithmic in n, m.

strongly convex, non-strongly convex, and non-convex *population* loss functions. For strongly convex losses, we show that these algorithms achieve order-optimal statistical rates under mild conditions. We further propose a median-based robust algorithm that only requires one communication round, and show that it also achieves the optimal rate for strongly convex quadratic losses. The statistical error rates of these three algorithms are summarized as follows.

- **Median-based GD:** $\widetilde{\mathcal{O}}(\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}} + \frac{1}{n})$, order-optimal for strongly convex loss if $n \gtrsim m$.
- **Trimmed-mean-based GD:** $\widetilde{\mathcal{O}}(\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}})$, order-optimal for strongly convex loss.
- **Median-based one-round algorithm:** $\widetilde{\mathcal{O}}(\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}} + \frac{1}{n})$, order-optimal for strongly convex quadratic loss if $n \gtrsim m$.

A major technical challenge in our statistical setting here is as follows: the nm data points are sampled once and fixed, and each worker machine has access to a fixed set of data throughout the learning process. This creates complicated probabilistic dependency across the iterations of the algorithms. Worse yet, the Byzantine machines, which have complete knowledge of the data and the learning algorithm used, may create further unspecified probabilistic dependency. We overcome this difficulty by proving certain *uniform* bounds via careful covering arguments. Furthermore, for the analysis of median-based algorithms, we cannot simply adapt standard techniques (such as those in Minsker et al. (2015)), which can only show that the output of the master machine is as accurate as that of *one* normal machine, leading to a sub-optimal $\mathcal{O}(\frac{1}{\sqrt{n}})$ rate even without Byzantine failures ($\alpha = 0$). Instead, we make use of a more delicate argument based on normal approximation and Berry-Esseen-type inequalities, which allows us to achieve the better $\mathcal{O}(\frac{1}{mn})$ rates when α is small while being robust for a nonzero α.

Above we have omitted the dependence on the parameter dimension d.; see our main theorems for the precise results. In some settings the rates in these results may not have the optimal dependence on d. Understanding the fundamental limits of robust distributed learning in high dimensions, as well as developing algorithms with optimal dimension dependence, is an interesting and important future direction.

1.2 Notation

We denote vectors by boldface lowercase letters such as $\mathbf{w}$, and the elements in the vector are denoted by italics letters with subscripts, such as w_k. Matrices are denoted by boldface uppercase letters such as $\mathbf{H}$. For any positive integer N, we denote the set $\{1, 2, \ldots, N\}$ by $[N]$. For vectors, we denote the ℓ_2 norm and ℓ_∞ norm by $\|\cdot\|_2$ and $\|\cdot\|_\infty$, respectively. For matrices, we denote the operator norm and the Frobenius norm by $\|\cdot\|_2$ and $\|\cdot\|_F$, respectively. We denote by $\Phi(\cdot)$ the cumulative distribution function (CDF) of the standard Gaussian distribution. For any differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, we denote its partial derivative with respect to the k-th argument by $\partial_k f$.

2 Related Work

Outlier-robust estimation in non-distributed settings is a classical topic in statistics (Huber, 2011). Particularly relevant to us is the so-called *median-of-means* method, in which one partitions the data into m subsets, computes an estimate from each subset, and finally takes the median of these m estimates. This idea is studied in Alon et al. (1999), Jerrum et al. (1986), Lerasle and Oliveira (2011), Minsker et al. (2015), Nemirovskii et al. (1983), and has been applied to bandit and least square regression problems (Bubeck et al., 2013, Kogler and Traxler, 2016, Lugosi and Mendelson, 2016) as well as problems involving heavy-tailed distributions (Hsu and Sabato, 2016, Lugosi and Mendelson, 2017). In a very recent work, Minsker and Strawn (2017) provide a new analysis of median-of-means using a normal approximation. We borrow some techniques from this paper, but need to address a significant harder problem: 1) we deal with the Byzantine setting with arbitrary/adversarial outliers, which is not considered in their paper; 2) we study iterative algorithms for general multi-dimensional problems with convex and non-convex losses, while they mainly focus on one-shot algorithms for mean-estimation-type problems.

The median-of-means method is used in the context of Byzantine-robust distributed learning in two recent papers. In particular, the work of Feng et al. (2014) considers a simple one-shot application of

median-of-means, and only proves a sub-optimal $\widetilde{\mathcal{O}}(\frac{1}{\sqrt{n}})$ error rate as mentioned. The work of Chen et al. (2017) considers only strongly convex losses, and seeks to circumvent the above issue by grouping the worker machines into mini-batches; however, their rate $\widetilde{\mathcal{O}}(\frac{\sqrt{\alpha}}{\sqrt{n}} + \frac{1}{\sqrt{nm}})$ still falls short of being optimal, and in particular their algorithm fails even when there is only one Byzantine machine in each mini-batch.

Other methods have been proposed for Byzantine-robust distributed learning and optimization; e.g., Su and Vaidya (2016a,b). These works consider optimizing fixed functions and do not provide guarantees on statistical error rates. Most relevant is the work by Blanchard et al. (2017), who propose to aggregate the gradients from worker machines using a robust procedure. Their optimization setting—which is at the level of stochastic gradient descent and assumes unlimited, independent access to a strong stochastic gradient oracle—is fundamentally different from ours; in particular, they do not provide a characterization of the statistical errors given a fixed number of data points.

Communication efficiency has been studied extensively in non-Byzantine distributed settings (McMahan et al., 2016, Yin et al., 2017). An important class of algorithms are based on one-round aggregation methods (Rosenblatt and Nadler, 2016, Zhang et al., 2012, 2015). More sophisticated algorithms have been proposed in order to achieve better accuracy than the one-round approach while maintaining lower communication costs; examples include DANE (Shamir et al., 2014), Disco (Zhang and Lin, 2015), distributed SVRG (Lee et al., 2015) and their variants (Reddi et al., 2016, Wang et al., 2017). Developing Byzantine-robust versions of these algorithms is an interesting future direction.

For outlier-robust estimation in non-distributed settings, much progress has been made recently in terms of improved performance in high-dimensional problems (Bhatia et al., 2015, Diakonikolas et al., 2016, Lai et al., 2016) as well as developing list-decodable and semi-verified learning schemes when a majority of the data points are adversarial (Charikar et al., 2017). These results are not directly applicable to our distributed setting with general loss functions, but it is nevertheless an interesting future problem to investigate their potential extension for our problem.

3 Problem Setup

In this section, we formally set up our problem and introduce a few concepts key to our the algorithm design and analysis. Suppose that training data points are sampled from some unknown distribution $\mathcal{D}$ on the sample space $\mathcal{Z}$. Let $f(\mathbf{w}; \mathbf{z})$ be a loss function of a parameter vector $\mathbf{w} \in \mathcal{W} \subseteq \mathbb{R}^d$ associated with the data point $\mathbf{z}$, where $\mathcal{W}$ is the parameter space, and $F(\mathbf{w}) := \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[f(\mathbf{w}; \mathbf{z})]$ is the corresponding population loss function. Our goal is to learn a model defined by the parameter that minimizes the population loss:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}). \tag{1}$$

The parameter space $\mathcal{W}$ is assumed to be convex and compact with diameter D, i.e., $\|\mathbf{w} - \mathbf{w}'\|_2 \leq D, \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}$. We consider a distributed computation model with one master machine and m worker machines. Each worker machine stores n data points, each of which is sampled independently from $\mathcal{D}$. Denote by $\mathbf{z}^{i,j}$ the j-th data on the i-th worker machine, and $F_i(\mathbf{w}) := \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{w}; \mathbf{z}^{i,j})$ the empirical risk function for the i-th worker. We assume that an α fraction of the m worker machines are Byzantine, and the remaining $1-\alpha$ fraction are normal. With the notation $[N] := \{1, 2, \ldots, N\}$, we index the set of worker machines by $[m]$, and denote the set of Byzantine machines by $\mathcal{B} \subset [m]$ (thus $|\mathcal{B}| = \alpha m$). The master machine communicates with the worker machines using some predefined protocol. The Byzantine machines need not obey this protocol and can send arbitrary messages to the master; in particular, they may have complete knowledge of the system and learning algorithms, and can collude with each other.

We introduce the coordinate-wise median and trimmed mean operations, which serve as building blocks for our algorithm.

Definition 1 (Coordinate-wise median)**.** *For vectors $\mathbf{x}^i \in \mathbb{R}^d$, $i \in [m]$, the coordinate-wise median $\mathbf{g} := \mathsf{med}\{\mathbf{x}^i : i \in [m]\}$ is a vector with its k-th coordinate being $g_k = \mathsf{med}\{x_k^i : i \in [m]\}$ for each $k \in [d]$, where med is the usual (one-dimensional) median.*

Definition 2 (Coordinate-wise trimmed mean)**.** *For $\beta \in [0, \frac{1}{2})$ and vectors $\mathbf{x}^i \in \mathbb{R}^d$, $i \in [m]$, the coordinate-wise β-trimmed mean $\mathbf{g} := \mathsf{trmean}_\beta\{\mathbf{x}^i : i \in [m]\}$ is a vector with its k-th coordinate being $g_k = \frac{1}{(1-2\beta)m}\sum_{x\in U_k} x$ for each $k \in [d]$. Here U_k is a subset of $\{x_k^1, \ldots, x_k^m\}$ obtained by removing the largest and smallest β fraction of its elements.*

For the analysis, we need several standard definitions concerning random variables/vectors.

Definition 3 (Variance of random vectors). *For a random vector* $\mathbf{x}$*, define its variance as* $\mathrm{Var}(\mathbf{x}) := \mathbb{E}[\|\mathbf{x} - \mathbb{E}[\mathbf{x}]\|_2^2]$.

Definition 4 (Absolute skewness). *For a one-dimensional random variable* X*, define its absolute skewness*[2] *as* $\gamma(X) := \frac{\mathbb{E}[|X-\mathbb{E}[X]|^3]}{\mathrm{Var}(X)^{3/2}}$*. For a* d*-dimensional random vector* $\mathbf{x}$*, we define its absolute skewness as the vector of the absolute skewness of each coordinate of* $\mathbf{x}$*, i.e.,* $\gamma(\mathbf{x}) := [\gamma(x_1)\ \gamma(x_2)\ \cdots\ \gamma(x_d)]^\top$.

Definition 5 (Sub-exponential random variables). *A random variable* X *with* $\mathbb{E}[X] = \mu$ *is called* v*-sub-exponential if* $\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{1}{2}v^2\lambda^2},\ \forall\ |\lambda| < \frac{1}{v}$.

Finally, we need several standard concepts from convex analysis regarding a differentiable function $h(\cdot) : \mathbb{R}^d \to \mathbb{R}$.

Definition 6 (Lipschitz). *h is L-Lipschitz if* $|h(\mathbf{w}) - h(\mathbf{w}')| \leq L\|\mathbf{w} - \mathbf{w}'\|_2, \forall\ \mathbf{w}, \mathbf{w}'$.

Definition 7 (Smoothness). *h is L'-smooth if* $\|\nabla h(\mathbf{w}) - \nabla h(\mathbf{w}')\|_2 \leq L'\|\mathbf{w} - \mathbf{w}'\|_2, \forall\ \mathbf{w}, \mathbf{w}'$.

Definition 8 (Strong convexity). *h is λ-strongly convex if* $h(\mathbf{w}') \geq h(\mathbf{w}) + \langle \nabla h(\mathbf{w}), \mathbf{w}' - \mathbf{w}\rangle + \frac{\lambda}{2}\|\mathbf{w}' - \mathbf{w}\|_2^2, \forall\ \mathbf{w}, \mathbf{w}'$.

4 Robust Distributed Gradient Descent

We describe two robust distributed gradient descent algorithms, one based on coordinate-wise median and the other on trimmed mean. These two algorithms are formally given in Algorithm 1 as Option I and Option II, respectively, where the symbol $*$ represents an arbitrary vector.

In each parallel iteration of the algorithms, the master machine broadcasts the current model parameter to all worker machines. The normal worker machines compute the gradients of their local loss functions and then send the gradients back to the master machine. The Byzantine machines may send any messages of their choices. The master machine then performs a gradient descent update on the model parameter with step-size η, using either the coordinate-wise median or trimmed mean of the received gradients. The Euclidean projection $\Pi_{\mathcal{W}}(\cdot)$ ensures that the model parameter stays in the parameter space $\mathcal{W}$.

Below we provide statistical guarantees on the error rates of these algorithms, and compare their performance. Throughout we assume that each loss function $f(\mathbf{w}; \mathbf{z})$ and the population loss function $F(\mathbf{w})$ are smooth:

Assumption 1 (Smoothness of f and F). *For any* $\mathbf{z} \in \mathcal{Z}$*, the partial derivative of* $f(\cdot; \mathbf{z})$ *with respect to the* k*-th coordinate of its first argument, denoted by* $\partial_k f(\cdot; \mathbf{z})$*, is* L_k*-Lipschitz for each* $k \in [d]$*, and the function* $f(\cdot; \mathbf{z})$ *is* L*-smooth. Let* $\widehat{L} := \sqrt{\sum_{k=1}^d L_k^2}$*. Also assume that the population loss function* $F(\cdot)$ *is* L_F*-smooth.*

It is easy to see hat $L_F \leq L \leq \widehat{L}$. When the dimension of $\mathbf{w}$ is high, the quantity $\widehat{L}$ may be large. However, we will soon see that $\widehat{L}$ only appears in the logarithmic factors in our bounds and thus does not have a significant impact.

4.1 Guarantees for Median-based Gradient Descent

We first consider our median-based algorithm, namely Algorithm 1 with Option I. We impose the assumptions that the gradient of the loss function f has bounded variance, and each coordinate of the gradient has coordinate-wise bounded absolute skewness:

Assumption 2 (Bounded variance of gradient). *For any* $\mathbf{w} \in \mathcal{W}$*,* $\mathrm{Var}(\nabla f(\mathbf{w}; \mathbf{z})) \leq V^2$.

Assumption 3 (Bounded skewness of gradient). *For any* $\mathbf{w} \in \mathcal{W}$*,* $\|\gamma(\nabla f(\mathbf{w}; \mathbf{z}))\|_\infty \leq S$.

[2]Note the difference with the usual skewness $\frac{\mathbb{E}[(X-\mathbb{E}[X])^3]}{\mathrm{Var}(X)^{3/2}}$.

Algorithm 1 Robust Distributed Gradient Descent

Require: Initialize parameter vector $\mathbf{w}^0 \in \mathcal{W}$, algorithm parameters β (for Option II), η and T.
for $t = 0, 1, 2, \ldots, T-1$ **do**
Master machine: send $\mathbf{w}^t$ to all the worker machines.
for all $i \in [m]$ **do in parallel**
Worker machine i: compute local gradient

$$\mathbf{g}^i(\mathbf{w}^t) \leftarrow \begin{cases} \nabla F_i(\mathbf{w}^t) & \text{normal worker machines,} \\ * & \text{Byzantine machines,} \end{cases}$$

send $\mathbf{g}^i(\mathbf{w}^t)$ to master machine.
end for
Master machine: compute aggregate gradient

$$\mathbf{g}(\mathbf{w}^t) \leftarrow \begin{cases} \mathsf{med}\{\mathbf{g}^i(\mathbf{w}^t) : i \in [m]\} & \text{Option I} \\ \mathsf{trmean}_\beta\{\mathbf{g}^i(\mathbf{w}^t) : i \in [m]\} & \text{Option II} \end{cases}$$

update model parameter $\mathbf{w}^{t+1} \leftarrow \Pi_{\mathcal{W}}(\mathbf{w}^t - \eta \mathbf{g}(\mathbf{w}^t))$.
end for

These assumptions are satisfied in many learning problems with small values of V^2 and S. Below we provide a concrete example in terms of a linear regression problem.

Proposition 1. *Suppose that each data point $\mathbf{z} = (\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ is generated by $y = \mathbf{x}^\top \mathbf{w}^* + \xi$ with some $\mathbf{w}^* \in \mathcal{W}$. Assume that the elements of $\mathbf{x}$ are independent and uniformly distributed in $\{-1, 1\}$, and that the noise $\xi \sim \mathcal{N}(0, \sigma^2)$ is independent of $\mathbf{x}$. With the quadratic loss function $f(\mathbf{w}; \mathbf{x}, y) = \frac{1}{2}(y - \mathbf{x}^{\mathrm{T}}\mathbf{w})^2$, we have $\mathrm{Var}(\nabla f(\mathbf{w}; \mathbf{x}, y)) = (d-1)\|\mathbf{w} - \mathbf{w}^*\|_2^2 + d\sigma^2$, and $\|\gamma(\nabla f(\mathbf{w}; \mathbf{x}, y))\|_\infty \le 480$.*

We prove Proposition 1 in Appendix A.1. In this example, the upper bound V^2 on $\mathrm{Var}(\nabla f(\mathbf{w}; \mathbf{x}, y))$ depends on dimension d and the diameter of the parameter space. If the diameter is a constant, we have $V = \mathcal{O}(\sqrt{d})$. Moreover, the gradient skewness is bounded by a universal constant S regardless of the size of the parameter space. In Appendix A.2, we provide another example showing that when the features in $\mathbf{x}$ are i.i.d. Gaussian distributed, the coordinate-wise skewness can be upper bounded by 429.

We now state our main technical results on the median-based algorithm, namely statistical error guarantees for strongly convex, non-strongly convex, and smooth non-convex population loss functions F. In the first two cases with a convex F, we assume that $\mathbf{w}^*$, the minimizer of $F(\cdot)$ in $\mathcal{W}$, is also the minimizer of $F(\cdot)$ in $\mathbb{R}^d$, i.e., $\nabla F(\mathbf{w}^*) = 0$.

Strongly Convex Losses: We first consider the case where the population loss function $F(\cdot)$ is strongly convex. Note that we do not require strong convexity of the individual loss functions $f(\cdot; \mathbf{z})$.

Theorem 1. *Consider Option I in Algorithm 1. Suppose that Assumptions 1, 2, and 3 hold, $F(\cdot)$ is λ_F-strongly convex, and the fraction α of Byzantine machines satisfies*

$$\alpha + \sqrt{\frac{d \log(1 + nm\widehat{L}D)}{m(1-\alpha)}} + 0.4748\frac{S}{\sqrt{n}} \le \frac{1}{2} - \epsilon \tag{2}$$

for some $\epsilon > 0$. Choose step-size $\eta = 1/L_F$. Then, with probability at least $1 - \frac{4d}{(1+nm\widehat{L}D)^d}$, after T parallel iterations, we have

$$\|\mathbf{w}^T - \mathbf{w}^*\|_2 \le (1 - \frac{\lambda_F}{L_F + \lambda_F})^T \|\mathbf{w}^0 - \mathbf{w}^*\|_2 + \frac{2}{\lambda_F}\Delta,$$

where

$$\Delta := \mathcal{O}\left(C_\epsilon V \Big(\frac{\alpha}{\sqrt{n}} + \sqrt{\frac{d \log(nm\widehat{L}D)}{nm}} + \frac{S}{n} \Big) \right), \tag{3}$$

and C_ϵ is defined as

$$C_\epsilon := \sqrt{2\pi}\exp\Big(\frac{1}{2}(\Phi^{-1}(1-\epsilon))^2\Big), \tag{4}$$

with $\Phi^{-1}(\cdot)$ being the inverse of the cumulative distribution function of the standard Gaussian distribution $\Phi(\cdot)$.

We prove Theorem 1 in Appendix B. In (3), we hide universal constants and a higher order term that scales as $\frac{1}{nm}$, and the factor C_ϵ is a function of ϵ; as a concrete example, $C_\epsilon \approx 4$ when $\epsilon = \frac{1}{6}$. Theorem 1 together with the inequality $\log(1-x) \leq -x$, guarantees that after running $T \geq \frac{L_F+\lambda_F}{\lambda_F}\log(\frac{\lambda_F}{2\Delta}\|\mathbf{w}^0 - \mathbf{w}^*\|_2)$ parallel iterations, with high probability we can obtain a solution $\widehat{\mathbf{w}} = \mathbf{w}^T$ with error $\|\widehat{\mathbf{w}} - \mathbf{w}^*\|_2 \leq \frac{4}{\lambda_F}\Delta$.

Here we achieve an error rate (defined as the distance between $\widehat{\mathbf{w}}$ and the optimal solution $\mathbf{w}^*$) of the form $\widetilde{\mathcal{O}}(\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}} + \frac{1}{n})$. In Section 6, we provide a lower bound showing that the error rate of any algorithm is $\widetilde{\Omega}(\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}})$. Therefore the first two terms in the upper bound cannot be improved. The third term $\frac{1}{n}$ is due to the dependence of the median on the skewness of the gradients. When each worker machine has a sufficient amount of data, more specifically $n \gtrsim m$, we achieve an order-optimal error rate up to logarithmic factors.

Non-strongly Convex Losses: We next consider the case where the population risk function $F(\cdot)$ is convex, but not necessarily strongly convex. In this case, we need a mild technical assumption on the size of the parameter space $\mathcal{W}$.

Assumption 4 (Size of $\mathcal{W}$)**.** *The parameter space $\mathcal{W}$ contains the following ℓ_2 ball centered at $\mathbf{w}^*$: $\{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w} - \mathbf{w}^*\|_2 \leq 2\|\mathbf{w}^0 - \mathbf{w}^*\|_2\}$.*

We then have the following result on the convergence rate in terms of the value of the population risk function.

Theorem 2. *Consider Option I in Algorithm 1. Suppose that Assumptions 1, 2, 3 and 4 hold, and that the population loss $F(\cdot)$ is convex, and α satisfies (2) for some $\epsilon > 0$. Define Δ as in (3), and choose step-size $\eta = 1/L_F$. Then, with probability at least $1 - \frac{4d}{(1+nm\widehat{L}D)^d}$, after $T = \frac{L_F}{\Delta}\|\mathbf{w}^0 - \mathbf{w}^*\|_2$ parallel iterations, we have*

$$F(\mathbf{w}^T) - F(\mathbf{w}^*) \leq 16\|\mathbf{w}^0 - \mathbf{w}^*\|_2\Delta\Big(1 + \frac{1}{2L_F}\Delta\Big).$$

We prove Theorem 2 in Appendix C. We observe that the error rate, defined as the excess risk $F(\mathbf{w}^T) - F(\mathbf{w}^*)$, again has the form $\widetilde{\mathcal{O}}\big(\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}} + \frac{1}{n}\big)$.

Non-convex Losses: When $F(\cdot)$ is non-convex but smooth, we need a somewhat different technical assumption on the size of $\mathcal{W}$.

Assumption 5 (Size of $\mathcal{W}$)**.** *Suppose that $\forall\ \mathbf{w} \in \mathcal{W}$, $\|\nabla F(\mathbf{w})\|_2 \leq M$. We assume that $\mathcal{W}$ contains the ℓ_2 ball $\{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w} - \mathbf{w}^0\|_2 \leq \frac{2}{\Delta^2}(M+\Delta)(F(\mathbf{w}^0) - F(\mathbf{w}^*))\}$, where Δ is defined as in (3).*

We have the following guarantees on the rate of convergence to a critical point of the population loss $F(\cdot)$.

Theorem 3. *Consider Option I in Algorithm 1. Suppose that Assumptions 1 2, 3 and 5 hold, and α satisfies (2) for some $\epsilon > 0$. Define Δ as in (3), and choose step-size $\eta = 1/L_F$. With probability at least $1 - \frac{4d}{(1+nm\widehat{L}D)^d}$, after $T = \frac{2L_F}{\Delta^2}(F(\mathbf{w}^0) - F(\mathbf{w}^*))$ parallel iterations, we have*

$$\min_{t=0,1,\ldots,T}\|\nabla F(\mathbf{w}^t)\|_2 \leq \sqrt{2}\Delta.$$

We prove Theorem 3 in Appendix D. We again obtain an $\widetilde{\mathcal{O}}(\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}} + \frac{1}{n})$ error rate in terms of the gap to a critical point of $F(\mathbf{w})$.

4.2 Guarantees for Trimmed-mean-based Gradient Descent

We next analyze the robust distributed gradient descent algorithm based on coordinate-wise trimmed mean, namely Option II in Algorithm 1. Here we need stronger assumptions on the tail behavior of the partial derivatives of the loss functions—in particular, sub-exponentiality.

Assumption 6 (Sub-exponential gradients)**.** *We assume that for all $k \in [d]$ and $\mathbf{w} \in \mathcal{W}$, the partial derivative of $f(\mathbf{w}; \mathbf{z})$ with respect to the k-th coordinate of $\mathbf{w}$, $\partial_k f(\mathbf{w}; \mathbf{z})$, is v-sub-exponential.*

The sub-exponential property implies that all the moments of the derivatives are bounded. This is a stronger assumption than the bounded absolute skewness (hence bounded third moments) required by the median-based GD algorithm.

We use the same example as in Proposition 1 and show that the derivatives of the loss are indeed sub-exponential.

Proposition 2. *Consider the regression problem in Proposition 1. For all $k \in [d]$ and $\mathbf{w} \in \mathcal{W}$, the partial derivative $\partial_k f(\mathbf{w}; \mathbf{z})$ is $\sqrt{\sigma^2 + \|\mathbf{w} - \mathbf{w}^*\|_2^2}$-sub-exponential.*

Proposition 2 is proved in Appendix A.3. We now proceed to establish the statistical guarantees of the trimmed-mean-based algorithm, for different loss function classes. When the population loss $F(\cdot)$ is convex, we again assume that the minimizer of $F(\cdot)$ in $\mathcal{W}$ is also its minimizer in $\mathbb{R}^d$. The next three theorems are analogues of Theorems 1–3 for the median-based GD algorithm.

Strongly Convex Losses: We have the following result.

Theorem 4. *Consider Option II in Algorithm 1. Suppose that Assumptions 1 and 6 hold, $F(\cdot)$ is λ_F-strongly convex, and $\alpha \le \beta \le \frac{1}{2} - \epsilon$ for some $\epsilon > 0$. Choose step-size $\eta = 1/L_F$. Then, with probability at least $1 - \frac{4d}{(1+nm\widehat{L}D)^d}$, after T parallel iterations, we have*

$$\|\mathbf{w}^T - \mathbf{w}^*\|_2 \le \Big(1 - \frac{\lambda_F}{L_F + \lambda_F}\Big)^T \|\mathbf{w}^0 - \mathbf{w}^*\|_2 + \frac{2}{\lambda_F}\Delta',$$

where

$$\Delta' := \mathcal{O}\Big(\frac{vd}{\epsilon}\big(\frac{\beta}{\sqrt{n}} + \frac{1}{\sqrt{nm}}\big)\sqrt{\log(nm\widehat{L}D)}\Big). \tag{5}$$

We prove Theorem 4 in Appendix E. In (5), we hide universal constants and higher order terms that scale as $\frac{\beta}{n}$ or $\frac{1}{nm}$. By running $T \ge \frac{L_F + \lambda_F}{\lambda_F} \log(\frac{\lambda_F}{2\Delta'}\|\mathbf{w}^0 - \mathbf{w}^*\|_2)$ parallel iterations, we can obtain a solution $\widehat{\mathbf{w}} = \mathbf{w}^T$ satisfying $\|\widehat{\mathbf{w}} - \mathbf{w}^*\|_2 \le \widetilde{\mathcal{O}}(\frac{\beta}{\sqrt{n}} + \frac{1}{\sqrt{nm}})$. Note that one needs to choose the parameter for trimmed mean to satisfy $\beta \ge \alpha$. If we set $\beta = c\alpha$ for some universal constant $c \ge 1$, we can achieve an order-optimal error rate $\widetilde{\mathcal{O}}(\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}})$.

Non-strongly Convex Losses: Again imposing Assumption 4 on the size of $\mathcal{W}$, we have the following guarantee.

Theorem 5. *Consider Option II in Algorithm 1. Suppose that Assumptions 1, 4 and 6 hold, $F(\cdot)$ is convex, and $\alpha \le \beta \le \frac{1}{2} - \epsilon$ for some $\epsilon > 0$. Choose step-size $\eta = 1/L_F$, and define Δ' as in (5). Then, with probability at least $1 - \frac{4d}{(1+nm\widehat{L}D)^d}$, after $T = \frac{L_F}{\Delta'}\|\mathbf{w}^0 - \mathbf{w}^*\|_2$ parallel iterations, we have*

$$F(\mathbf{w}^T) - F(\mathbf{w}^*) \le 16\|\mathbf{w}^0 - \mathbf{w}^*\|_2\Delta'\Big(1 + \frac{1}{2L_F}\Delta'\Big).$$

The proof of Theorem 5 is similar to that of Theorem 2, and we refer readers to Remark 1 in Appendix E. Again, by choosing $\beta = c\alpha$ ($c \ge 1$), we obtain the $\widetilde{\mathcal{O}}(\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}})$ error rate in the function value of $F(\mathbf{w})$.

Non-convex Losses: In this case, imposing a version of Assumption 5 on the size of $\mathcal{W}$, we have the following.

Theorem 6. *Consider Option II in Algorithm 1, and define Δ' as in (5). Suppose that Assumptions 1 and 6 hold, Assumption 5 holds with Δ replaced by Δ', and $\alpha \leq \beta \leq \frac{1}{2} - \epsilon$ for some $\epsilon > 0$. Choose step-size $\eta = 1/L_F$. Then, with probability at least $1 - \frac{4d}{(1+nm\widehat{L}D)^d}$, after $T = \frac{2L_F}{\Delta'^2}(F(\mathbf{w}^0) - F(\mathbf{w}^*))$ parallel iterations, we have*

$$\min_{t=0,1,\ldots,T} \|\nabla F(\mathbf{w}^t)\|_2 \leq \sqrt{2}\Delta'.$$

The proof of Theorem 6 is similar to that of Theorem 3; see Remark 1 in Appendix E. By choosing $\beta = c\alpha$ with $c \geq 1$, we again achieve the statistical rate $\widetilde{\mathcal{O}}(\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}})$.

4.3 Comparisons

We compare the performance guarantees of the above two robust distribute GD algorithms. The trimmed-mean-based algorithm achieves the statistical error rate $\widetilde{\mathcal{O}}(\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}})$, which is order-optimal for strongly convex loss. In comparison, the rate of the median-based algorithm is $\widetilde{\mathcal{O}}(\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}} + \frac{1}{n})$, which has an additional $\frac{1}{n}$ term and is only optimal when $n \gtrsim m$. In particular, the trimmed-mean-based algorithm has better rates when each worker machine has small local sample size—the rates are meaningful even in the extreme case $n = \mathcal{O}(1)$. On the other hand, the median-based algorithm requires milder tail/moment assumptions on the loss derivatives (bounded skewness) than its trimmed-mean counterpart (sub-exponentiality). Finally, the trimmed-mean operation requires an additional parameter β, which can be any upper bound on the fraction α of Byzantine machines in order to guarantee robustness. Using an overly large β may lead to a looser bound and sub-optimal performance. In contrast, median-based GD does not require knowledge of α. We summarize these observations in Table 1. We see that the two algorithms are complementary to each other, and our experiment results corroborate this point.

	median GD	trimmed mean GD
Statistical error rate	$\widetilde{\mathcal{O}}(\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}} + \frac{1}{n})$	$\widetilde{\mathcal{O}}(\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}})$
Distribution of $\partial_k f(\mathbf{w}; \mathbf{z})$	Bounded skewness	Sub-exponential
α known?	No	Yes

Table 1: Comparison between the two robust distributed gradient descent algorithms.

5 Robust One-round Algorithm

As mentioned, in our distributed computing framework, the communication cost is proportional to the number of parallel iterations. The above two GD algorithms both require a number iterations depending on the desired accuracy. Can we further reduce the communication cost while keeping the algorithm Byzantine-robust and statistically optimal?

A natural candidate is the so-called one-round algorithm. Previous work has considered a standard one-round scheme where each local machine computes the empirical risk minimizer (ERM) using its local data and the master machine receives all workers' ERMs and computes their *average* (Zhang et al., 2012). Clearly, a single Byzantine machine can arbitrary skew the output of this algorithm. We instead consider a Byzantine-robust one-round algorithm. As detailed in Algorithm 2, we employ the coordinate-wise median operation to aggregate all the ERMs.

Our main result is a characterization of the error rate of Algorithm 2 in the presence of Byzantine failures. We are only able to establish such a guarantee when the loss functions are quadratic and $\mathcal{W} = \mathbb{R}^d$. However, one can implement this algorithm in problems with other loss functions.

Definition 9 (Quadratic loss function)**.** *The loss function $f(\mathbf{w}; \mathbf{z})$ is quadratic if it can be written as*

$$f(\mathbf{w}; \mathbf{z}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{H}\mathbf{w} + \mathbf{p}^{\mathrm{T}}\mathbf{w} + c,$$

where $\mathbf{z} = (\mathbf{H}, \mathbf{p}, c)$, $\mathbf{H}$, and $\mathbf{p}$, and c are drawn from the distributions $\mathcal{D}_H$, $\mathcal{D}_p$, and $\mathcal{D}_c$, respectively.

Algorithm 2 Robust One-round Algorithm

for all $i \in [m]$ **do in parallel**

Worker machine i: compute:

$$\widehat{\mathbf{w}}^i \leftarrow \begin{cases} \arg\min_{\mathbf{w}\in\mathcal{W}} F_i(\mathbf{w}) & \text{normal worker machines} \\ * & \text{Byzantine machines} \end{cases}$$

send $\widehat{\mathbf{w}}^i$ to master machine.

end for

Master machine: compute $\widehat{\mathbf{w}} \leftarrow \mathsf{med}\{\widehat{\mathbf{w}}^i : i \in [m]\}$.

Denote by $\mathbf{H}_F$, $\mathbf{p}_F$, and c_F the expectations of $\mathbf{H}$, $\mathbf{p}$, and c, respectively. Thus the population risk function takes the form $F(\mathbf{w}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{H}_F\mathbf{w} + \mathbf{p}_F^{\mathrm{T}}\mathbf{w} + c_F$.

We need a technical assumption which guarantees that each normal worker machine has unique ERM.

Assumption 7 (Strong convexity of F_i)**.** *With probability* 1*, the empirical risk minimization function $F_i(\cdot)$ on each normal machine is strongly convex.*

Note that this assumption is imposed on $F_i(\mathbf{w})$, rather than on the individual loss $f(\mathbf{w};\mathbf{z})$ associated with a single data point. This assumption is satisfied, for example, when all $f(\cdot;\mathbf{z})$'s are strongly convex, or in the linear regression problems with the features $\mathbf{x}$ drawn from some continuous distribution (e.g. isotropic Gaussian) and $n \geq d$. We have the following guarantee for the robust one-round algorithm.

Theorem 7. *Suppose that $\forall\, \mathbf{z} \in \mathcal{Z}$, the loss function $f(\cdot;\mathbf{z})$ is convex and quadratic, $F(\cdot)$ is λ_F-strongly convex, and Assumption 7 holds. Assume that α satisfies*

$$\alpha + \sqrt{\frac{\log(nmd)}{2m(1-\alpha)}} + \frac{\widetilde{C}}{\sqrt{n}} \leq \frac{1}{2} - \epsilon$$

for some $\epsilon > 0$, where $\widetilde{C}$ is a quantity that depends on $\mathcal{D}_H$, $\mathcal{D}_p$, λ_F and is monotonically decreasing in n. Then, with probability at least $1 - \frac{4}{nm}$, the output $\widehat{\mathbf{w}}$ of the robust one-round algorithm satisfies

$$\|\widehat{\mathbf{w}} - \mathbf{w}^*\|_2 \leq \frac{C_\epsilon}{\sqrt{n}}\widetilde{\sigma}\big(\alpha + \sqrt{\frac{\log(nmd)}{2m(1-\alpha)}} + \frac{\widetilde{C}}{\sqrt{n}}\big),$$

where C_ϵ is defined as in (4) and

$$\widetilde{\sigma}^2 := \mathbb{E}\big[\|\mathbf{H}_F^{-1}\big((\mathbf{H}-\mathbf{H}_F)\mathbf{H}_F^{-1}\mathbf{p}_F - (\mathbf{p}-\mathbf{p}_F)\big)\|_2^2\big],$$

with $\mathbf{H}$ and $\mathbf{p}$ drawn from $\mathcal{D}_H$ and $\mathcal{D}_p$, respectively.

We prove Theorem 7 and provide an explicit expression of $\widetilde{C}$ in Appendix F. In terms of the dependence on α, n, and m, the robust one-round algorithm achieves the same error rate as the robust gradient descent algorithm based on coordinate-wise median, i.e., $\widetilde{\mathcal{O}}(\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}} + \frac{1}{n})$, for quadratic problems. Again, this rate is optimal when $n \gtrsim m$. Therefore, at least for quadratic loss functions, the robust one-round algorithm has similar theoretical performance as the robust gradient descent algorithm with significantly less communication cost. Our experiments show that the one-round algorithm has good empirical performance for other losses as well.

6 Lower Bound

In this section, we provide a lower bound on the error rate for strongly convex losses, which implies that the $\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}}$ term is unimprovable. This lower bound is derived using a mean estimation problem, and is an extension of the lower bounds in the robust mean estimation literature such as Chen et al. (2015), Lai et al. (2016).

We consider the problem of estimating the mean $\boldsymbol{\mu}$ of some random variable $\mathbf{z} \sim \mathcal{Z}$, which is equivalent to solving the following minimization problem:

$$\boldsymbol{\mu} = \arg\min_{\mathbf{w} \in \mathcal{W}} \mathbb{E}_{\mathbf{z} \sim \mathcal{Z}}[\|\mathbf{w} - \mathbf{z}\|_2^2], \tag{6}$$

Note that this is a special case of the general learning problem (1). We consider the same distributed setting as in Section 4, with a minor technical difference regarding the Byzantine machines. We assume that each of the m worker machines is Byzantine with probability α, independently of each other. The parameter α is therefore the *expected* fraction of Byzantine machines. This setting makes the analysis slightly easier, and we believe the result can be extended to the original setting.

In this setting we have the following lower bound.

Observation 1. *Consider the distributed mean estimation problem in (6) with Byzantine failure probability α, and suppose that $\mathcal{Z}$ is Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\sigma^2\mathbf{I}$ ($\sigma = \mathcal{O}(1)$). Then, any algorithm that computes an estimation $\widehat{\boldsymbol{\mu}}$ of the mean from the data has a constant probability of error $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 = \Omega(\frac{\alpha}{\sqrt{n}} + \sqrt{\frac{d}{nm}})$.*

We prove Observation 1 in Appendix G. According to this observation, we see that the $\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}}$ dependence cannot be avoided, which in turn implies the order-optimality of the results in Theorem 1 (when $n \gtrsim m$) and Theorem 4.

7 Experiments

We conduct experiments to show the effectiveness of the median and trimmed mean operations. Our experiments are implemented with Tensorflow (Abadi et al., 2016) on Microsoft Azure system. We use the MNIST (LeCun et al., 1998) dataset and randomly partition the 60,000 training data into m subsamples with equal sizes. We use these subsamples to represent the data on m machines.

In the first experiment, we compare the performance of distributed gradient descent algorithms in the following four settings: 1) $\alpha = 0$ (no Byzantine machines), using vanilla distributed gradient descent (aggregating the gradients by taking the mean), 2) $\alpha > 0$, using vanilla distributed gradient descent, 3) $\alpha > 0$, using median-based algorithm, and 4) $\alpha > 0$, using trimmed-mean-based algorithm. We generate the Byzantine machines in the following way: we replace every training label y on these machines with $9 - y$, e.g., 0 is replaced with 9, 1 is replaced with 8, etc, and the Byzantine machines simply compute gradients based on these data. We also note that when generating the Byzantine machines, we do not simply add extreme values in the features or gradients; instead, the Byzantine machines send messages to the master machine with moderate values.

We train a multi-class logistic regression model and a convolutional neural network model using distributed gradient descent, and for each model, we compare the test accuracies in the aforementioned four settings. For the convolutional neural network model, we use the stochastic version of the distributed gradient descent algorithm; more specifically, in every iteration, each worker machine computes the gradient using 10% of its local data. We periodically check the test errors, and the convergence performances are shown in Figure 1. The final test accuracies are presented in Tables 2 and 3.

α	0	0.05		
Algorithm	mean	mean	median	trimmed mean
Test accuracy (%)	88.0	76.8	87.2	86.9

Table 2: Test accuracy on the logistic regression model using gradient descent. We set $m = 40$, and for trimmed mean, we choose $\beta = 0.05$.

As we can see, in the adversarial settings, the vanilla distributed gradient descent algorithm suffers from severe performance loss, and using the median and trimmed mean operations, we observe significant improvement in test accuracy. This shows these two operations can indeed defend against Byzantine failures.

In the second experiment, we compare the performance of distributed one-round algorithms in the following three settings: 1) $\alpha = 0$, mean aggregation, 2) $\alpha > 0$, mean aggregation, and 3) $\alpha > 0$, median

α	0	0.1		
Algorithm	mean	mean	median	trimmed mean
Test accuracy (%)	94.3	77.3	87.4	90.7

Table 3: Test accuracy on the convolutional neural network model using gradient descent. We set $m = 10$, and for trimmed mean, we choose $\beta = 0.1$.

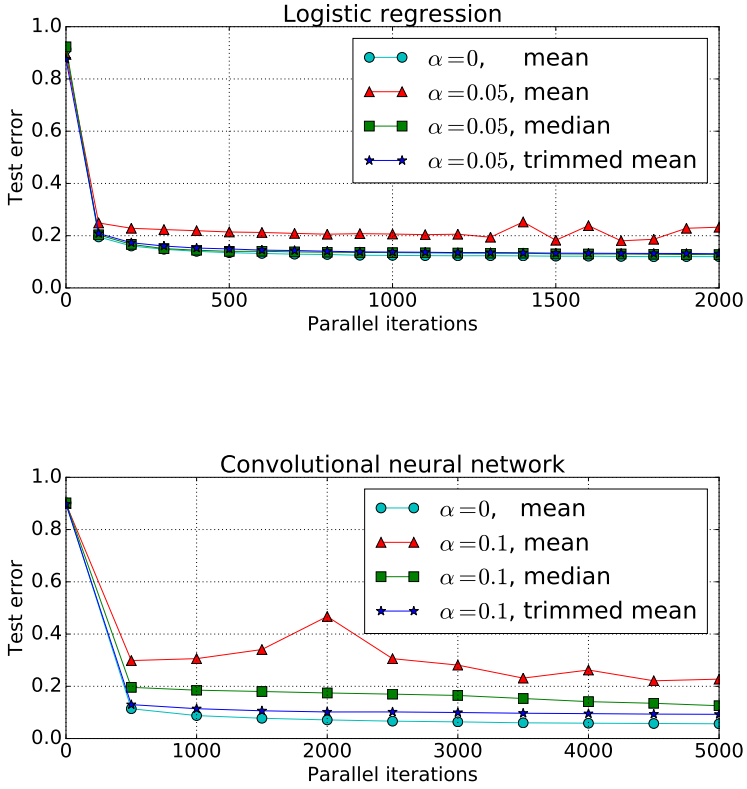

Figure 1: Test error vs the number of parallel iterations.

aggregation. In this experiment, the training labels on the Byzantine machines are i.i.d. uniformly sampled from $\{0, \ldots, 9\}$, and these machines train models using the faulty data. We choose the multi-class logistic regression model, and the test accuracies are presented in Table 4.

α	0	0.1	
Algorithm	mean	mean	median
Test accuracy (%)	91.8	83.7	89.0

Table 4: Test accuracy on the logistic regression model using one-round algorithm. We set $m = 10$.

As we can see, for the one-round algorithm, although the theoretical guarantee is only proved for quadratic loss, in practice, the median-based one-round algorithm still improves the test accuracy in problems with other loss functions, such as the logistic loss here.

8 Conclusions

In this paper, we study Byzantine-robust distributed statistical learning algorithms with a focus on statistical optimality. We analyze two robust distributed gradient descent algorithms — one is based on

coordinate-wise median and the other is based on coordinate-wise trimmed mean. We show that the trimmed-mean-based algorithm can achieve order-optimal $\widetilde{\mathcal{O}}(\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}})$ error rate, whereas the median-based algorithm can achieve $\widetilde{\mathcal{O}}(\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}} + \frac{1}{n})$ under weaker assumptions. We further study learning algorithms that have better communication efficiency. We propose a simple one-round algorithm that aggregates local solutions using coordinate-wise median. We show that for strongly convex quadratic problems, this algorithm can achieve $\widetilde{\mathcal{O}}(\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}} + \frac{1}{n})$ error rate, similar to the median-based gradient descent algorithm. Our experiments validates the effectiveness of the median and trimmed mean operations in the adversarial setting.

Acknowledgements

D. Yin is partially supported by Berkeley DeepDrive Industry Consortium. Y. Chen is partially supported by NSF CRII award 1657420 and grant 1704828. K. Ramchandran is partially supported by NSF CIF award 1703678 and Gift award from Huawei. P. Bartlett is partially supported by NSF grant IIS-1619362. Cloud computing resources are provided by a Microsoft Azure for Research award.

References

M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and system sciences*, 58(1):137–147, 1999.

A. C. Berry. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society*, 49(1):122–136, 1941.

K. Bhatia, P. Jain, and P. Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pages 721–729, 2015.

P. Blanchard, E. M. E. Mhamdi, R. Guerraoui, and J. Stainer. Byzantine-tolerant machine learning. *arXiv preprint arXiv:1703.02757*, 2017.

S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

S. Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

M. Charikar, J. Steinhardt, and G. Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60. ACM, 2017.

M. Chen, C. Gao, and Z. Ren. Robust covariance matrix estimation via matrix depth. *arXiv preprint arXiv:1506.00691*, 2015.

Y. Chen, L. Su, and J. Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *arXiv preprint arXiv:1705.05491*, 2017.

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high dimensions without the computational intractability. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 655–664. IEEE, 2016.

C.-G. Esseen. *On the Liapounoff limit of error in the theory of probability*. Almqvist & Wiksell, 1942.

J. Feng, H. Xu, and S. Mannor. Distributed robust learning. *arXiv preprint arXiv:1409.5937*, 2014.

D. Hsu and S. Sabato. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17(1):543–582, 2016.

P. J. Huber. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 2011.

M. R. Jerrum, L. G. Valiant, and V. V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188, 1986.

A. Kogler and P. Traxler. Efficient and robust median-of-means algorithms for location and regression. In *Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2016 18th International Symposium on*, pages 206–213. IEEE, 2016.

J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

K. A. Lai, A. B. Rao, and S. Vempala. Agnostic estimation of mean and covariance. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 665–674. IEEE, 2016.

L. Lamport, R. Shostak, and M. Pease. The byzantine generals problem. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, 4(3):382–401, 1982.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

J. D. Lee, Q. Lin, T. Ma, and T. Yang. Distributed stochastic variance reduced gradient methods and a lower bound for communication complexity. *arXiv preprint arXiv:1507.07595*, 2015.

M. Lerasle and R. I. Oliveira. Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*, 2011.

G. Lugosi and S. Mendelson. Risk minimization by median-of-means tournaments. *arXiv preprint arXiv:1608.00757*, 2016.

G. Lugosi and S. Mendelson. Sub-gaussian estimators of the mean of a random vector. *arXiv preprint arXiv:1702.00482*, 2017.

B. McMahan and D. Ramage. Federated learning: Collaborative machine learning without centralized training data. https://research.googleblog.com/2017/04/federated-learning-collaborative.html, 2017.

H. B. McMahan, E. Moore, D. Ramage, S. Hampson, et al. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629*, 2016.

S. Minsker and N. Strawn. Distributed statistical estimation and rates of convergence in normal approximation. *arXiv preprint arXiv:1704.02658*, 2017.

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

A. Nemirovskii, D. B. Yudin, and E. R. Dawson. Problem complexity and method efficiency in optimization. 1983.

I. Pinelis and R. Molzon. Optimal-order bounds on the rate of convergence to normality in the multivariate delta method. *Electronic Journal of Statistics*, 10(1):1001–1063, 2016.

S. J. Reddi, J. Konečnỳ, P. Richtárik, B. Póczós, and A. Smola. Aide: Fast and communication efficient distributed optimization. *arXiv preprint arXiv:1608.06879*, 2016.

J. D. Rosenblatt and B. Nadler. On the optimality of averaging in distributed statistical learning. *Information and Inference: A Journal of the IMA*, 5(4):379–404, 2016.

O. Shamir, N. Srebro, and T. Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning*, pages 1000–1008, 2014.

I. Shevtsova. On the absolute constants in the berry-esseen-type inequalities. In *Doklady Mathematics*, volume 89, pages 378–381. Springer, 2014.

L. Su and N. H. Vaidya. Fault-tolerant multi-agent optimization: optimal iterative distributed algorithms. In *Proceedings of the 2016 ACM Symposium on Principles of Distributed Computing*, pages 425–434. ACM, 2016a.

L. Su and N. H. Vaidya. Non-bayesian learning in the presence of byzantine agents. In *International Symposium on Distributed Computing*, pages 414–427. Springer, 2016b.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

S. Wang, F. Roosta-Khorasani, P. Xu, and M. W. Mahoney. Giant: Globally improved approximate newton method for distributed optimization. *arXiv preprint arXiv:1709.03528*, 2017.

Y. Wu. Lecture notes for ece598yw: Information-theoretic methods for high-dimensional statistics. http://www.stat.yale.edu/~yw562/teaching/it-stats.pdf, 2017.

D. Yin, A. Pananjady, M. Lam, D. Papailiopoulos, K. Ramchandran, and P. Bartlett. Gradient diversity: a key ingredient for scalable distributed learning. *arXiv preprint arXiv:1706.05699*, 2017.

Y. Zhang and X. Lin. Disco: Distributed optimization for self-concordant empirical loss. In *International conference on machine learning*, pages 362–370, 2015.

Y. Zhang, M. J. Wainwright, and J. C. Duchi. Communication-efficient algorithms for statistical optimization. In *Advances in Neural Information Processing Systems*, pages 1502–1510, 2012.

Y. Zhang, J. Duchi, and M. Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research*, 16(1):3299–3340, 2015.

Appendix

A Variance, Skewness, and Sub-exponential Property

A.1 Proof of Proposition 1

We use the simplified notation $f(\mathbf{w}) := f(\mathbf{w}; \mathbf{x}, y)$. One can directly compute the gradients:

$$\nabla f(\mathbf{w}) = \mathbf{x}(\mathbf{x}^{\mathrm{T}}\mathbf{w} - y) = \mathbf{x}\mathbf{x}^{\mathrm{T}}(\mathbf{w} - \mathbf{w}^*) - \xi\mathbf{x},$$

and thus

$$\nabla F(\mathbf{w}) = \mathbb{E}[\nabla f(\mathbf{w})] = \mathbf{w} - \mathbf{w}^*.$$

Define $\Delta(\mathbf{w}) := \nabla f(\mathbf{w}) - \nabla F(\mathbf{w})$ with its k-th element being $\Delta_k(\mathbf{w})$. We now compute the variance and absolute skewness of $\Delta_k(\mathbf{w})$.

We can see that

$$\Delta_k(\mathbf{w}) = \sum_{\substack{1\le i\le d\\ i\neq k}} x_k x_i (w_i - w_i^*) + (x_k^2 - 1)(w_k - w_k^*) - \xi x_k. \tag{7}$$

Thus,

$$\mathbb{E}[\Delta_k^2(\mathbf{w})] = \mathbb{E}[\sum_{\substack{1\le i\le d\\ i\neq k}} x_k^2 x_i^2 (w_i - w_i^*)^2 + \xi^2 x_k^2] = \|\mathbf{w} - \mathbf{w}^*\|_2^2 - (w_k - w_k^*)^2 + \sigma^2, \tag{8}$$

which yields

$$\mathrm{Var}(\nabla f(\mathbf{w})) = \mathbb{E}[\|\nabla f(\mathbf{w}) - \nabla F(\mathbf{w})\|_2^2] = (d-1)\|\mathbf{w} - \mathbf{w}^*\|_2^2 + d\sigma^2.$$

Then we proceed to bound $\gamma(\Delta_k(\mathbf{w}))$. By Jensen's inequality, we know that

$$\gamma(\Delta_k(\mathbf{w})) = \frac{\mathbb{E}[|\Delta_k(\mathbf{w})|^3]}{\mathrm{Var}(\Delta_k(\mathbf{w}))^{3/2}} \leq \sqrt{\frac{\mathbb{E}[\Delta_k^6(\mathbf{w})]}{\mathrm{Var}(\Delta_k(\mathbf{w}))^3}} \tag{9}$$

We first find a lower bound for $\mathrm{Var}(\Delta_k(\mathbf{w}))^3$. According to (8), we know that

$$\mathrm{Var}(\Delta_k(\mathbf{w}))^3 = \Big(\sum_{\substack{1\leq i\leq d\\ i\neq k}} (w_i - w_i^*)^2 + \sigma^2\Big)^3 \geq \Big(\sum_{\substack{1\leq i\leq d\\ i\neq k}} (w_i - w_i^*)^2\Big)^3 + \sigma^6.$$

Define the following three quantities.

$$W_1 = \sum_{\substack{1\leq i\leq d\\ i\neq k}} (w_i - w_i^*)^6 \tag{10}$$

$$W_2 = \sum_{\substack{1\leq i,j\leq d\\ i,j\neq k\\ i\neq j}} (w_i - w_i^*)^4 (w_j - w_j^*)^2 \tag{11}$$

$$W_3 = \sum_{\substack{1\leq i,j,\ell\leq d\\ i,j,\ell\neq k\\ i\neq j, i\neq \ell, j\neq \ell}} (w_i - w_i^*)^2 (w_j - w_j^*)^2 (w_\ell - w_\ell^*)^2 \tag{12}$$

By simple algebra, one can check that

$$\Big(\sum_{\substack{1\leq i\leq d\\ i\neq k}} (w_i - w_i^*)^2\Big)^3 = W_1 + 3W_2 + W_3, \tag{13}$$

and thus

$$\mathrm{Var}(\Delta_k(\mathbf{w}))^3 \geq W_1 + 3W_2 + W_3 + \sigma^6. \tag{14}$$

Then, we find an upper bound on $\mathbb{E}[\Delta_k^6(\mathbf{w})]$. According to (7), and Hölder's inequality, we know that

$$\begin{aligned}\mathbb{E}[\Delta_k^6(\mathbf{w})] &= \mathbb{E}[(\sum_{\substack{1\leq i\leq d\\ i\neq k}} x_k x_i (w_i - w_i^*) - \xi x_k)^6] \leq 32\big(\mathbb{E}[(\sum_{\substack{1\leq i\leq d\\ i\neq k}} x_k x_i (w_i - w_i^*))^6] + \mathbb{E}[\xi^6 x_k^6]\big)\\ &= 32\big(\mathbb{E}[(\sum_{\substack{1\leq i\leq d\\ i\neq k}} x_i (w_i - w_i^*))^6] + 15\sigma^6\big),\end{aligned} \tag{15}$$

where in the last inequality we use the moments of Gaussian random variables. Then, we compute the first term in (15). By algebra, one can obtain

$$\begin{aligned}\mathbb{E}[(\sum_{\substack{1\leq i\leq d\\ i\neq k}} x_i (w_i - w_i^*))^6] =& \mathbb{E}[\sum_{\substack{1\leq i\leq d\\ i\neq k}} x_i^6 (w_i - w_i^*)^6] + 15\mathbb{E}[\sum_{\substack{1\leq i,j\leq d\\ i,j\neq k\\ i\neq j}} x_i^4 x_j^2 (w_i - w_i^*)^4 (w_j - w_j^*)^2]\\ &+ 15\mathbb{E}[\sum_{\substack{1\leq i,j,\ell\leq d\\ i,j,\ell\neq k\\ i\neq j, i\neq \ell, j\neq \ell}} x_i^2 x_j^2 x_\ell^2 (w_i - w_i^*)^2 (w_j - w_j^*)^2 (w_\ell - w_\ell^*)^2]\\ =& W_1 + 15W_2 + 15W_3.\end{aligned} \tag{16}$$

Combining (15) and (16), we get

$$\mathbb{E}[\Delta_k^6(\mathbf{w})] \leq 32(W_1 + 15W_2 + 15W_3 + 15\sigma^6). \tag{17}$$

Combining (14) and (17), we get

$$\gamma(\Delta_k(\mathbf{w})) \leq \sqrt{\frac{\mathbb{E}[\Delta_k^6(\mathbf{w})]}{\mathrm{Var}(\Delta_k(\mathbf{w}))^3}} \leq \sqrt{\frac{32(W_1 + 15W_2 + 15W_3 + 15\sigma^6)}{W_1 + 3W_2 + W_3 + \sigma^6}} \leq 480.$$

A.2 Example of Regression with Gaussian Features

Proposition 3. *Suppose that each data point consists of a feature* $\mathbf{x} \in \mathbb{R}^d$ *and a label* $y \in \mathbb{R}$*, and the label is generated by*

$$y = \mathbf{x}^{\mathrm{T}}\mathbf{w}^* + \xi$$

with some $\mathbf{w}^* \in \mathcal{W}$*. Assume that the elements of* $\mathbf{x}$ *are i.i.d. samples of standard Gaussian distribution, and that the noise* ξ *is independent of* $\mathbf{x}$ *and drawn from Gaussian distribution* $\mathcal{N}(0, \sigma^2)$*. Define the quadratic loss function* $f(\mathbf{w}; \mathbf{x}, y) = \frac{1}{2}(y - \mathbf{x}^{\mathrm{T}}\mathbf{w})^2$*. Then, we have*

$$\mathrm{Var}(\nabla f(\mathbf{w}; \mathbf{x}, y)) = (d+1)\|\mathbf{w} - \mathbf{w}^*\|_2^2 + d\sigma^2,$$

and

$$\|\gamma(\nabla f(\mathbf{w}; \mathbf{x}, y))\|_\infty \le 429.$$

Proof. We use the same simplified notation as in Appendix A.1. One can also see that (7) still holds for in the Gaussian setting. Thus,

$$\begin{aligned}
\mathbb{E}[\Delta_k^2(\mathbf{w})] &= \mathbb{E}[\sum_{\substack{1\le i\le d\\ i\neq k}} x_k^2 x_i^2 (w_i - w_i^*)^2 + (x_k^2 - 1)^2 (w_k - w_k^*)^2 + \xi^2 x_k^2] \\
&= \sum_{\substack{1\le i\le d\\ i\neq k}} (w_i - w_i^*)^2 + 2(w_k - w_k^*)^2 + \sigma^2 \qquad (18)\\
&= \|\mathbf{w} - \mathbf{w}^*\|_2^2 + (w_k - w_k^*)^2 + \sigma^2, \qquad (19)
\end{aligned}$$

which yields

$$\mathrm{Var}(\nabla f(\mathbf{w})) = \mathbb{E}[\|\nabla f(\mathbf{w}) - \nabla F(\mathbf{w})\|_2^2] = (d+1)\|\mathbf{w} - \mathbf{w}^*\|_2^2 + d\sigma^2.$$

Then we proceed to bound $\gamma(\Delta_k(\mathbf{w}))$. By Jensen's inequality, we know that

$$\gamma(\Delta_k(\mathbf{w})) = \frac{\mathbb{E}[|\Delta_k(\mathbf{w})|^3]}{\mathrm{Var}(\Delta_k(\mathbf{w}))^{3/2}} \le \sqrt{\frac{\mathbb{E}[\Delta_k^6(\mathbf{w})]}{\mathrm{Var}(\Delta_k(\mathbf{w}))^3}} \qquad (20)$$

We first find a lower bound for $\mathrm{Var}(\Delta_k(\mathbf{w}))^3$. According to (18), we know that

$$\begin{aligned}
\mathrm{Var}(\Delta_k(\mathbf{w}))^3 &= \Big(\sum_{\substack{1\le i\le d\\ i\neq k}} (w_i - w_i^*)^2 + 2(w_k - w_k^*)^2 + \sigma^2\Big)^3 \\
&\ge \Big(\sum_{\substack{1\le i\le d\\ i\neq k}} (w_i - w_i^*)^2\Big)^3 + 8(w_k - w_k^*)^6 + \sigma^6.
\end{aligned}$$

Define the W_1, W_2, and W_3 as in (10), (11), and (12). We can also see that (13) still holds, and thus

$$\mathrm{Var}(\Delta_k(\mathbf{w}))^3 \ge W_1 + 3W_2 + W_3 + 8(w_k - w_k^*)^6 + \sigma^6. \qquad (21)$$

Then, we find an upper bound on $\mathbb{E}[\Delta_k^6(\mathbf{w})]$. According to (7), and Hölder's inequality, we know that

$$\begin{aligned}
\mathbb{E}[\Delta_k^6(\mathbf{w})] &= \mathbb{E}[(\sum_{\substack{1\le i\le d\\ i\neq k}} x_k x_i (w_i - w_i^*) + (x_k^2 - 1)(w_k - w_k^*) - \xi x_k)^6] \\
&\le 243\big(\mathbb{E}[(\sum_{\substack{1\le i\le d\\ i\neq k}} x_k x_i (w_i - w_i^*))^6] + \mathbb{E}[(x_k^2 - 1)^6 (w_k - w_k^*)^6] + \mathbb{E}[\xi^6 x_k^6]\big) \\
&= 243\big(15\mathbb{E}[(\sum_{\substack{1\le i\le d\\ i\neq k}} x_i (w_i - w_i^*))^6] + 6040(w_k - w_k^*)^6 + 225\sigma^6\big), \qquad (22)
\end{aligned}$$

where in the last inequality we use the moments of Gaussian random variables. Then, we compute the first term in (22). By algebra, one can obtain

$$\begin{aligned}\mathbb{E}[(\sum_{\substack{1\le i\le d\\ i\neq k}} x_i(w_i-w_i^*))^6] =&\mathbb{E}[\sum_{\substack{1\le i\le d\\ i\neq k}} x_i^6(w_i-w_i^*)^6] + 15\mathbb{E}[\sum_{\substack{1\le i,j\le d\\ i,j\neq k\\ i\neq j}} x_i^4x_j^2(w_i-w_i^*)^4(w_j-w_j^*)^2]\\ &+15\mathbb{E}[\sum_{\substack{1\le i,j,\ell\le d\\ i,j,\ell\neq k\\ i\neq j,i\neq\ell,j\neq\ell}} x_i^2x_j^2x_\ell^2(w_i-w_i^*)^2(w_j-w_j^*)^2(w_\ell-w_\ell^*)^2]\\ =&15W_1+45W_2+15W_3. \end{aligned} \tag{23}$$

Combining (22) and (23), we get

$$\mathbb{E}[\Delta_k^6(\mathbf{w})] \le 243(225W_1+675W_2+225W_3+6040(w_k-w_k^*)^6+225\sigma^6). \tag{24}$$

Combining (21) and (24), we get

$$\gamma(\Delta_k(\mathbf{w})) \le \sqrt{\frac{\mathbb{E}[\Delta_k^6(\mathbf{w})]}{\mathrm{Var}(\Delta_k(\mathbf{w}))^3}} \le \sqrt{\frac{243(225W_1+675W_2+225W_3+6040(w_k-w_k^*)^6+225\sigma^6)}{W_1+3W_2+W_3+8(w_k-w_k^*)^6+\sigma^6}} \le 429.$$

□

A.3 Proof of Proposition 2

We use the same notation as in Appendix A.1. We have

$$\begin{aligned}\partial_k f(\mathbf{w};\mathbf{z}) - F(\mathbf{w}) = \Delta_k(\mathbf{w}) &= \sum_{\substack{1\le i\le d\\ i\neq k}} x_kx_i(w_i-w_i^*) + (x_k^2-1)(w_k-w_k^*) - \xi x_k\\ &= x_k(-\xi + \sum_{\substack{1\le i\le d\\ i\neq k}} x_i(w_i-w_i^*)) := x_k\Delta_k'(\mathbf{w})\end{aligned}$$

Since $\Delta_k'(\mathbf{w})$ has symmetric distribution and x_k is uniformly distributed in $\{-1,1\}$, we know that the distributions of $\Delta_k(\mathbf{w})$ and $\Delta_k'(\mathbf{w})$. We then prove a stronger result on $\Delta_k'(\mathbf{w})$. We first recall the definition of v-sub-Gaussian random variables. A random variable X with mean $\mu=\mathbb{E}[X]$ is v-sub-Gaussian if for all $\lambda\in\mathbb{R}$, $\mathbb{E}[e^{\lambda(X-\mu)}] \le e^{v^2\lambda^2/2}$. We can see that v-sub-Gaussian random variables are also v-sub-exponential. One can also check that x_i's are i.i.d. 1-sub-Gaussian random variables, and then $\Delta_k'(\mathbf{w})$ is v-sub-exponential with

$$v = \big(\sigma^2 + \sum_{\substack{1\le i\le d\\ i\neq k}} (w_i-w_i^*)^2\big)^{1/2} \le \sqrt{\sigma^2+\|\mathbf{w}-\mathbf{w}^*\|_2^2}.$$

B Proof of Theorem 1

The proof of Theorem 1 consists of two parts: 1) the analysis of coordinate-wise median estimator of the population gradients, and 2) the convergence analysis of the robustified gradient descent algorithm.

Recall that at iteration t, the master machine sends $\mathbf{w}^t$ to all the worker machines. For any normal worker machine, say machine $i\in[m]\setminus\mathcal{B}$, the gradient of the local empirical loss function $\mathbf{g}^i(\mathbf{w}^t) = \nabla F_i(\mathbf{w}^t)$ is computed and returned to the center machine, while the Byzantine machines, say machine $i\in\mathcal{B}$, the returned message $\mathbf{g}^i(\mathbf{w}^t)$ can be arbitrary or even adversarial. The master machine then compute the coordinate-wise median, i.e.,

$$\mathbf{g}(\mathbf{w}^t) = \mathsf{med}\{\mathbf{g}^i(\mathbf{w}^t) : i\in[m]\}.$$

The following theorem provides a uniform bound on the distance between $\mathbf{g}(\mathbf{w}^t)$ and $\nabla F(\mathbf{w}^t)$.

Theorem 8. *Define*

$$\mathbf{g}^i(\mathbf{w}) = \begin{cases} \nabla F_i(\mathbf{w}) & i \in [m] \setminus \mathcal{B}, \\ * & i \in \mathcal{B}. \end{cases} \tag{25}$$

and the coordinate-wise median of $\mathbf{g}^i(\mathbf{w})$*:*

$$\mathbf{g}(\mathbf{w}) = \mathsf{med}\{\mathbf{g}^i(\mathbf{w}) : i \in [m]\}. \tag{26}$$

Suppose that Assumptions 1, 2, and 3 hold, and inequality (2) is satisfied with some $\epsilon > 0$*. Then, we have with probability at least* $1 - \frac{4d}{(1+nm\widehat{L}D)^d}$*,*

$$\|\mathbf{g}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \leq 2\sqrt{2}\frac{1}{nm} + \sqrt{2}\frac{C_\epsilon}{\sqrt{n}}V\left(\alpha + \sqrt{\frac{d\log(1+nm\widehat{L}D)}{m(1-\alpha)}} + 0.4748\frac{S}{\sqrt{n}}\right), \tag{27}$$

for all $\mathbf{w} \in \mathcal{W}$*, where* C_ϵ *is defined as in (4).*

Proof. See Appendix B.1. □

Then, we proceed to analyze the convergence of the robust distributed gradient descent algorithm. We condition on the event that the bound in (27) is satisfied for all $\mathbf{w} \in \mathcal{W}$. Then, in the t-th iteration, we define

$$\widehat{\mathbf{w}}^{t+1} = \mathbf{w}^t - \eta\mathbf{g}(\mathbf{w}^t).$$

Thus, we have $\mathbf{w}^{t+1} = \Pi_{\mathcal{W}}(\widehat{\mathbf{w}}^{t+1})$. By the property of Euclidean projection, we know that

$$\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2 \leq \|\widehat{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2.$$

We further have

$$\begin{aligned} \|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2 &\leq \|\mathbf{w}^t - \eta\mathbf{g}(\mathbf{w}^t) - \mathbf{w}^*\|_2 \\ &\leq \|\mathbf{w}^t - \eta\nabla F(\mathbf{w}^t) - \mathbf{w}^*\|_2 + \eta\|\mathbf{g}(\mathbf{w}^t) - \nabla F(\mathbf{w}^t)\|_2. \end{aligned} \tag{28}$$

Meanwhile, we have

$$\|\mathbf{w}^t - \eta\nabla F(\mathbf{w}^t) - \mathbf{w}^*\|_2^2 = \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 - 2\eta\langle\mathbf{w}^t - \mathbf{w}^*, \nabla F(\mathbf{w}^t)\rangle + \eta^2\|\nabla F(\mathbf{w}^t)\|_2^2. \tag{29}$$

Since $F(\mathbf{w})$ is λ_F-strongly convex, by the co-coercivity of strongly convex functions (see Lemma 3.11 in Bubeck et al. (2015) for more details), we obtain

$$\langle\mathbf{w}^t - \mathbf{w}^*, \nabla F(\mathbf{w}^t)\rangle \geq \frac{L_F\lambda_F}{L_F + \lambda_F}\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + \frac{1}{L_F + \lambda_F}\|\nabla F(\mathbf{w}^t)\|_2^2.$$

Let $\eta = \frac{1}{L_F}$. Then we get

$$\begin{aligned} \|\mathbf{w}^t - \eta\nabla F(\mathbf{w}^t) - \mathbf{w}^*\|_2^2 &\leq (1 - \frac{2\lambda_F}{L_F + \lambda_F})\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 - \frac{2}{L_F(L_F + \lambda_F)}\|\nabla F(\mathbf{w}^t)\|_2^2 + \frac{1}{L_F^2}\|\nabla F(\mathbf{w}^t)\|_2^2 \\ &\leq (1 - \frac{2\lambda_F}{L_F + \lambda_F})\|\mathbf{w}^t - \mathbf{w}^*\|_2^2, \end{aligned}$$

where in the second inequality we use the fact that $\lambda_F \leq L_F$. Using the fact $\sqrt{1-x} \leq 1 - \frac{x}{2}$, we get

$$\|\mathbf{w}^t - \eta\nabla F(\mathbf{w}^t) - \mathbf{w}^*\|_2 \leq (1 - \frac{\lambda_F}{L_F + \lambda_F})\|\mathbf{w}^t - \mathbf{w}^*\|_2. \tag{30}$$

Combining (28) and (30), we get

$$\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2 \leq (1 - \frac{\lambda_F}{L_F + \lambda_F})\|\mathbf{w}^t - \mathbf{w}^*\|_2 + \frac{1}{L_F}\Delta, \tag{31}$$

where

$$\Delta = 2\sqrt{2}\frac{1}{nm} + \sqrt{2}\frac{C_\epsilon}{\sqrt{n}}V(\alpha + \sqrt{\frac{d\log(1+nm\widehat{L}D)}{m(1-\alpha)}} + 0.4748\frac{S}{\sqrt{n}}).$$

Then we can complete the proof by iterating (31).

B.1 Proof of Theorem 8

The proof of Theorem 8 relies on careful analysis of the median of means estimator in the presence of adversarial data and a covering net argument.

We first consider a general problem of robust estimation of a one dimensional random variable. Suppose that there are m worker machines, and q of them are Byzantine machines, which store n adversarial data (recall that $\alpha := q/m$). Each of the other $m(1-\alpha)$ normal worker machines stores n i.i.d. samples of some one dimensional random variable $x \sim \mathcal{D}$. Denote the j-th sample in the i-th worker machine by $x^{i,j}$. Let $\mu := \mathbb{E}[x]$, $\sigma^2 := \mathrm{Var}(x)$, and $\gamma(x)$ be the absolute skewness of x. In addition, define $\bar{x}^i$ as the average of samples in the i-th machine, i.e., $\bar{x}^i = \frac{1}{n}\sum_{j=1}^n x^{i,j}$. For any $z \in \mathbb{R}$, define $\widetilde{p}(z) := \frac{1}{m(1-\alpha)}\sum_{i\in[m]\setminus\mathcal{B}} \mathbb{1}(\bar{x}^i \le z)$ as the empirical distribution function of the sample averages on the *normal* worker machines. We have the following result on $\widetilde{p}(z)$.

Lemma 1. *Suppose that for a fixed $t > 0$, we have*

$$\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma(x)}{\sqrt{n}} \le \frac{1}{2} - \epsilon, \tag{32}$$

for some $\epsilon > 0$. Then, with probability at least $1 - 4e^{-2t}$, we have

$$\widetilde{p}\left(\mu + C_\epsilon \frac{\sigma}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma(x)}{\sqrt{n}})\right) \ge \frac{1}{2} + \alpha, \tag{33}$$

and

$$\widetilde{p}\left(\mu - C_\epsilon \frac{\sigma}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma(x)}{\sqrt{n}})\right) \le \frac{1}{2} - \alpha, \tag{34}$$

where C_ϵ is defined as in (4).

Proof. See Appendix B.2. □

We further define the distribution function of all the m machines as $\widehat{p}(z) := \frac{1}{m}\sum_{i\in[m]} \mathbb{1}(\bar{x}^i \le z)$. We have the following direct corollary on $\widehat{p}(z)$ and the median of means estimator $\mathsf{med}\{\bar{x}^i : i \in [m]\}$.

Corollary 1. *Suppose that condition (32) is satisfied. Then, with probability at least $1-4e^{-2t}$, we have,*

$$\widehat{p}\left(\mu + C_\epsilon \frac{\sigma}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma(x)}{\sqrt{n}})\right) \ge \frac{1}{2}, \tag{35}$$

and

$$\widehat{p}\left(\mu - C_\epsilon \frac{\sigma}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma(x)}{\sqrt{n}})\right) \le \frac{1}{2}. \tag{36}$$

Thus, we have with probability at least $1 - 4e^{-2t}$,

$$|\mathsf{med}\{\bar{x}^i : i \in [m]\} - \mu| \le C_\epsilon \frac{\sigma}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma(x)}{\sqrt{n}}). \tag{37}$$

Proof. One can easily check that for any $z \in \mathbb{R}$, we have $|\widehat{p}(z) - \widetilde{p}(z)| \le \alpha$, which yields the results (35) and (36). The result (37) can be derived using the fact that $\widehat{p}(\mathsf{med}\{\bar{x}^i : i \in [m]\}) = 1/2$. □

Lemma 1 and Corollary 1 can be translated to the estimators of the gradients. Define $\mathbf{g}^i(\mathbf{w})$ and $\mathbf{g}(\mathbf{w})$ as in (25) and (26), and let $g_k^i(\mathbf{w})$ and $g_k(\mathbf{w})$ be the k-th coordinate of $\mathbf{g}^i(\mathbf{w})$ and $\mathbf{g}(\mathbf{w})$, respectively. In addition, for any $\mathbf{w} \in \mathcal{W}$, $k \in [d]$, and $z \in \mathbb{R}$, we define the empirical distribution function of the k-th coordinate of the gradients on the normal machines:

$$\widetilde{p}(z; \mathbf{w}, k) = \frac{1}{m(1-\alpha)} \sum_{i\in[m]\setminus\mathcal{B}} \mathbb{1}(g_k^i(\mathbf{w}) \le z), \tag{38}$$

and on all the m machines

$$\widehat{p}(z;\mathbf{w},k)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}(g_k^i(\mathbf{w})\le z). \tag{39}$$

We use the symbol ∂_k to denote the partial derivative of any function with respect to its k-th argument. We also use the simplified notation $\sigma_k^2(\mathbf{w}):=\mathrm{Var}(\partial_k f(\mathbf{w};\mathbf{z}))$, and $\gamma_k(\mathbf{w}):=\gamma(\partial_k f(\mathbf{w};\mathbf{z}))$. Then, according to Lemma 1, when (32) is satisfied, for any fixed $\mathbf{w}\in\mathcal{W}$ and $k\in[d]$, we have with probability at least $1-4e^{-2t}$,

$$\widetilde{p}\left(\partial_k F(\mathbf{w})+C_\epsilon\frac{\sigma_k(\mathbf{w})}{\sqrt{n}}(\alpha+\sqrt{\frac{t}{m(1-\alpha)}}+0.4748\frac{\gamma_k(\mathbf{w})}{\sqrt{n}});\mathbf{w},k\right)\ge\frac{1}{2}+\alpha, \tag{40}$$

and

$$\widetilde{p}\left(\partial_k F(\mathbf{w})-C_\epsilon\frac{\sigma_k(\mathbf{w})}{\sqrt{n}}(\alpha+\sqrt{\frac{t}{m(1-\alpha)}}+0.4748\frac{\gamma_k(\mathbf{w})}{\sqrt{n}});\mathbf{w},k\right)\le\frac{1}{2}-\alpha. \tag{41}$$

Further, according to Corollary 1, we know that with probability $1-4e^{-2t}$,

$$|g_k(\mathbf{w})-\partial_k F(\mathbf{w})|\le C_\epsilon\frac{\sigma_k(\mathbf{w})}{\sqrt{n}}(\alpha+\sqrt{\frac{t}{m(1-\alpha)}}+0.4748\frac{\gamma_k(\mathbf{w})}{\sqrt{n}}). \tag{42}$$

Here, the inequality (42) gives a bound on the accuracy of the median of means estimator for the gradient at any fixed $\mathbf{w}$ and any coordinate $k\in[d]$. To extend this result to all $\mathbf{w}\in\mathcal{W}$ and all the d coordinates, we need to use union bound and a covering net argument.

Let $\mathcal{W}_\delta=\{\mathbf{w}^1,\mathbf{w}^2,\dots,\mathbf{w}^{N_\delta}\}$ be a finite subset of $\mathcal{W}$ such that for any $\mathbf{w}\in\mathcal{W}$, there exists $\mathbf{w}^\ell\in\mathcal{W}_\delta$ such that $\|\mathbf{w}^\ell-\mathbf{w}\|_2\le\delta$. According to the standard covering net results (Vershynin, 2010), we know that $N_\delta\le(1+\frac{D}{\delta})^d$. By union bound, we know that with probability at least $1-4dN_\delta e^{-2t}$, the bounds in (40) and (41) hold for all $\mathbf{w}=\mathbf{w}^\ell\in\mathcal{W}_\delta$, and $k\in[d]$. By gathering all the k coordinates and using Assumption 3, we know that this implies for all $\mathbf{w}^\ell\in\mathcal{W}_\delta$,

$$\|\mathbf{g}(\mathbf{w}^\ell)-\nabla F(\mathbf{w}^\ell)\|_2\le\frac{C_\epsilon}{\sqrt{n}}V\left(\alpha+\sqrt{\frac{t}{m(1-\alpha)}}+0.4748\frac{S}{\sqrt{n}}\right). \tag{43}$$

Then, consider an arbitrary $\mathbf{w}\in\mathcal{W}$. Suppose that $\|\mathbf{w}^\ell-\mathbf{w}\|_2\le\delta$. Since by Assumption 1, we assume that for each $k\in[d]$, the partial derivative $\partial_k f(\mathbf{w};\mathbf{z})$ is L_k-Lipschitz for all $\mathbf{z}$, we know that for every normal machine $i\in[m]\setminus\mathcal{B}$,

$$\left|g_k^i(\mathbf{w})-g_k^i(\mathbf{w}^\ell)\right|\le L_k\delta.$$

Then, according to the definition of $\widetilde{p}(z;\mathbf{w},k)$ in (39), we know that for any $z\in\mathbb{R}$, $\widetilde{p}(z+L_k\delta;\mathbf{w},k)\ge\widetilde{p}(z;\mathbf{w}^\ell,k)$ and $\widetilde{p}(z-L_k\delta;\mathbf{w},k)\le\widetilde{p}(z;\mathbf{w}^\ell,k)$. Then, the bounds in (40) and (41) yield

$$\widetilde{p}\left(\partial_k F(\mathbf{w}^\ell)+L_k\delta+C_\epsilon\frac{\sigma_k(\mathbf{w}^\ell)}{\sqrt{n}}(\alpha+\sqrt{\frac{t}{m(1-\alpha)}}+0.4748\frac{\gamma_k(\mathbf{w}^\ell)}{\sqrt{n}});\mathbf{w},k\right)\ge\frac{1}{2}+\alpha, \tag{44}$$

and

$$\widetilde{p}\left(\partial_k F(\mathbf{w}^\ell)-L_k\delta-C_\epsilon\frac{\sigma_k(\mathbf{w}^\ell)}{\sqrt{n}}(\alpha+\sqrt{\frac{t}{m(1-\alpha)}}+0.4748\frac{\gamma_k(\mathbf{w}^\ell)}{\sqrt{n}});\mathbf{w},k\right)\le\frac{1}{2}-\alpha. \tag{45}$$

Using the fact that $|\partial_k F(\mathbf{w}^\ell)-\partial_k F(\mathbf{w})|\le L_k\delta$, and Corollary 1, we have

$$|g_k(\mathbf{w})-\partial_k F(\mathbf{w})|\le 2L_k\delta+C_\epsilon\frac{\sigma_k(\mathbf{w}^\ell)}{\sqrt{n}}(\alpha+\sqrt{\frac{t}{m(1-\alpha)}}+0.4748\frac{\gamma_k(\mathbf{w}^\ell)}{\sqrt{n}}).$$

Again, by gathering all the k coordinates we get

$$\|\mathbf{g}(\mathbf{w})-\nabla F(\mathbf{w})\|_2^2\le 8\delta^2\sum_{k=1}^{d}L_k^2+2\frac{C_\epsilon^2}{n}\sum_{k=1}^{d}\sigma_k^2(\mathbf{w}^\ell)(\alpha+\sqrt{\frac{t}{m(1-\alpha)}}+0.4748\frac{\gamma_k(\mathbf{w}^\ell)}{\sqrt{n}})^2,$$

where we use the fact that $(a+b)^2 \leq 2(a^2+b^2)$. Then, by Assumption 2 and 3, we further obtain

$$\|\mathbf{g}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \leq 2\sqrt{2}\delta\widehat{L} + \sqrt{2}\frac{C_\epsilon}{\sqrt{n}}V\left(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{S}{\sqrt{n}}\right), \tag{46}$$

where we use the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. Combining (43) and (46), we conclude that for any $\delta > 0$, with probability at least $1 - 4dN_\delta e^{-2t}$, (46) holds for all $\mathbf{w} \in \mathcal{W}$. We simply choose $\delta = \frac{1}{nm\widehat{L}}$, and $t = d\log(1 + nm\widehat{L}D)$. Then, we know that with probability at least $1 - \frac{4d}{(1+nm\widehat{L}D)^d}$, we have

$$\|\mathbf{g}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \leq 2\sqrt{2}\frac{1}{nm} + \sqrt{2}\frac{C_\epsilon}{\sqrt{n}}V\left(\alpha + \sqrt{\frac{d\log(1+nm\widehat{L}D)}{m(1-\alpha)}} + 0.4748\frac{S}{\sqrt{n}}\right)$$

for all $\mathbf{w} \in \mathcal{W}$.

B.2 Proof of Lemma 1

We recall the Berry-Esseen Theorem (Berry, 1941, Esseen, 1942, Shevtsova, 2014) and the bounded difference inequality, which are useful in this proof.

Theorem 9 (Berry-Esseen Theorem). *Assume that $Y_1, \ldots, Y_n$ are i.i.d. copies of a random variable Y with mean μ, variance σ^2, and such that $\mathbb{E}[|Y-\mu|^3] < \infty$. Then,*

$$\sup_{s\in\mathbb{R}}\left|\mathbb{P}\left\{\sqrt{n}\frac{\bar{Y}-\mu}{\sigma} \leq s\right\} - \Phi(s)\right| \leq 0.4748\frac{\mathbb{E}[|Y-\mu|^3]}{\sigma^3\sqrt{n}},$$

where $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ and $\Phi(s)$ is the cumulative distribution function of the standard normal random variable.

Theorem 10 (Bounded Difference Inequality). *Let $X_1, \ldots, X_n$ be i.i.d. random variables, and assume that $Z = g(X_1, \ldots, X_n)$, where g satisfies that for all $j \in [n]$ and all $x_1, x_2, \ldots, x_j, x_j', \ldots, x_n$,*

$$|g(x_1, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_n) - g(x_1, \ldots, x_{j-1}, x_j', x_{j+1}, \ldots, x_n)| \leq c_j.$$

Then for any $t \geq 0$,

$$\mathbb{P}\left\{Z - \mathbb{E}[Z] \geq t\right\} \leq \exp\left(-\frac{2t^2}{\sum_{j=1}^n c_j^2}\right)$$

and

$$\mathbb{P}\left\{Z - \mathbb{E}[Z] \leq -t\right\} \leq \exp\left(-\frac{2t^2}{\sum_{j=1}^n c_j^2}\right).$$

Let $\sigma_n := \frac{\sigma}{\sqrt{n}}$ and $c_n := 0.4748\frac{\mathbb{E}[|x-\mu|^3]}{\sigma^3\sqrt{n}} = 0.4748\frac{\gamma(x)}{\sqrt{n}}$. Define $W_i := \frac{\bar{x}^i - \mu}{\sigma_n}$ for all $i \in [m]$, and $\Phi_n(\cdot)$ be the distribution function of W_i for any $i \in [m] \setminus \mathcal{B}$. We also define the empirical distribution function of $\{W_i : i \in [m] \setminus \mathcal{B}\}$ as $\widetilde{\Phi}_n(\cdot)$, i.e., $\widetilde{\Phi}_n(z) = \frac{1}{m(1-\alpha)}\sum_{i\in[m]\setminus\mathcal{B}} \mathbb{1}(W_i \leq z)$. Thus, we have

$$\widetilde{\Phi}_n(z) = \widetilde{p}(\sigma_n z + \mu). \tag{47}$$

We then focus on $\widetilde{\Phi}_n(z)$. We know that for any $z \in \mathbb{R}$, $\mathbb{E}[\widetilde{\Phi}_n(z)] = \Phi_n(z)$. Then, since the bounded difference inequality is satisfied with $c_j = \frac{1}{m(1-\alpha)}$, we have for any $t > 0$,

$$\left|\widetilde{\Phi}_n(z) - \Phi_n(z)\right| \leq \sqrt{\frac{t}{m(1-\alpha)}}, \tag{48}$$

on the draw of W_i, $i \in [m] \setminus \mathcal{B}$ with probability at least $1 - 2e^{-2t}$. Let $z_1 \geq z_2$ be such that $\Phi_n(z_1) \geq \frac{1}{2} + \alpha + \sqrt{\frac{t}{m(1-\alpha)}}$, and $\Phi_n(z_2) \leq \frac{1}{2} - \alpha - \sqrt{\frac{t}{m(1-\alpha)}}$. Then, by union bound, we know that with probability

at least $1-4e^{-2t}$, $\widetilde{\Phi}_n(z_1) \geq 1/2+\alpha$ and $\widetilde{\Phi}_n(z_2) \leq 1/2-\alpha$. The next step is to choose z_1 and z_2. According to Theorem 9, we know that

$$\Phi_n(z_1) \geq \Phi(z_1) - c_n,$$

and thus, it suffices to find z_1 such that

$$\Phi(z_1) = \frac{1}{2} + \alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n.$$

By mean value theorem, we know that there exists $\xi \in [0, z_1]$ such that

$$\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n = z_1 \Phi'(\xi) = \frac{z_1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2}} \geq \frac{z_1}{\sqrt{2\pi}} e^{-\frac{z_1^2}{2}}$$

Suppose that for some fix constant $\epsilon \in (0, 1/2)$, we have

$$\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n \leq \frac{1}{2} - \epsilon.$$

Then, we know that $z_1 \leq \Phi^{-1}(1-\epsilon)$, and thus we have

$$\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n \geq \frac{z_1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(\Phi^{-1}(1-\epsilon))^2),$$

which yields

$$z_1 \leq \sqrt{2\pi} \exp(\frac{1}{2}(\Phi^{-1}(1-\epsilon))^2) \left(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n\right).$$

Similarly

$$z_2 \geq -\sqrt{2\pi} \exp(\frac{1}{2}(\Phi^{-1}(1-\epsilon))^2) \left(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n\right).$$

For simplicity, let $C_\epsilon := \sqrt{2\pi} \exp(\frac{1}{2}(\Phi^{-1}(1-\epsilon))^2)$. We conclude that with probability $1-4e^{-2t}$, we have

$$\widetilde{p}(\mu + C_\epsilon \sigma_n(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n)) \geq \frac{1}{2} + \alpha,$$

and

$$\widetilde{p}(\mu - C_\epsilon \sigma_n(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n)) \leq \frac{1}{2} - \alpha.$$

C Proof of Theorem 2

Since Theorem 8 holds without assuming the convexity of $F(\mathbf{w})$, when $F(\mathbf{w})$ is non-strongly convex, the event that (27) holds for all $\mathbf{w} \in \mathcal{W}$ still happens with probability at least $1 - \frac{4d}{(1+nm\widehat{L}D)^d}$. We condition on this event. We first show that when Assumption 4 is satisfied and we choose $\eta = \frac{1}{L_F}$, the iterates $\mathbf{w}^t$ stays in $\mathcal{W}$ without using projection. Namely, define

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta \mathbf{g}(\mathbf{w}^t),$$

for $T = 0, 1, \ldots, T-1$, then $\mathbf{w}^t \in \mathcal{W}$ for all $t = 0, 1, \ldots, T$. To see this, we have

$$\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2 \leq \|\mathbf{w}^t - \eta \nabla F(\mathbf{w}^t) - \mathbf{w}^*\|_2 + \eta \|\mathbf{g}(\mathbf{w}^t) - \nabla F(\mathbf{w}^t)\|_2,$$

and

$$\begin{aligned}\|\mathbf{w}^t - \eta\nabla F(\mathbf{w}^t) - \mathbf{w}^*\|_2^2 &= \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 - 2\eta\langle\nabla F(\mathbf{w}^t), \mathbf{w}^t - \mathbf{w}^*\rangle + \eta^2\|\nabla F(\mathbf{w}^t)\|_2^2 \\ &\leq \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 - 2\eta\frac{1}{L_F}\|\nabla F(\mathbf{w}^t)\|_2^2 + \eta^2\|\nabla F(\mathbf{w}^t)\|_2^2 \\ &= \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 - \frac{1}{L_F^2}\|\nabla F(\mathbf{w}^t)\|_2^2 \\ &\leq \|\mathbf{w}^t - \mathbf{w}^*\|_2^2\end{aligned}$$

where the inequality is due to the co-coercivity of convex functions. Thus, we get

$$\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2 \leq \|\mathbf{w}^t - \mathbf{w}^*\|_2 + \frac{\Delta}{L_F},$$

and since $T = \frac{L_F D_0}{\Delta}$, according to Assumption 4 we know that $\mathbf{w}^t \in \mathcal{W}$ for all $t = 0, 1, \ldots, T$. Then, we proceed to study the algorithm without projection. Here, we define $D_t := \|\mathbf{w}^0 - \mathbf{w}^*\|_2 + \frac{t\Delta}{L_F}$ for $t = 0, 1, \ldots, T$.

Using the smoothness of $F(\mathbf{w})$, we have

$$\begin{aligned}F(\mathbf{w}^{t+1}) &\leq F(\mathbf{w}^t) + \langle\nabla F(\mathbf{w}^t), \mathbf{w}^{t+1} - \mathbf{w}^t\rangle + \frac{L_F}{2}\|\mathbf{w}^{t+1} - \mathbf{w}^t\|_2^2 \\ &= F(\mathbf{w}^t) + \eta\langle\nabla F(\mathbf{w}^t), -\mathbf{g}(\mathbf{w}^t) + \nabla F(\mathbf{w}^t) - \nabla F(\mathbf{w}^t)\rangle + \eta^2\frac{L_F}{2}\|\mathbf{g}(\mathbf{w}^t) - \nabla F(\mathbf{w}^t) + \nabla F(\mathbf{w}^t)\|_2^2.\end{aligned}$$

Since $\eta = \frac{1}{L_F}$ and $\|\mathbf{g}(\mathbf{w}^t) - \nabla F(\mathbf{w}^t)\|_2 \leq \Delta$, by simple algebra, we obtain

$$F(\mathbf{w}^{t+1}) \leq F(\mathbf{w}^t) - \frac{1}{2L_F}\|\nabla F(\mathbf{w}^t)\|_2^2 + \frac{1}{2L_F}\Delta^2. \tag{49}$$

We now prove the following lemma.

Lemma 2. *Condition on the event that (27) holds for all $\mathbf{w} \in \mathcal{W}$. When $F(\mathbf{w})$ is convex, by running $T = \frac{L_F D_0}{\Delta}$ parallel iterations, there exists $t \in \{0, 1, 2, \ldots, T\}$ such that*

$$F(\mathbf{w}^t) - F(\mathbf{w}^*) \leq 16 D_0 \Delta.$$

Proof. We first notice that since $T = \frac{L_F D_0}{\Delta}$, we have $D_t \leq 2D_0$ for all $t = 0, 1, \ldots, T$. According to the first order optimality of convex functions, for any $\mathbf{w}$,

$$F(\mathbf{w}) - F(\mathbf{w}^*) \leq \langle\nabla F(\mathbf{w}), \mathbf{w} - \mathbf{w}^*\rangle \leq \|\nabla F(\mathbf{w})\|_2\|\mathbf{w} - \mathbf{w}^*\|_2,$$

and thus

$$\|\nabla F(\mathbf{w})\|_2 \geq \frac{F(\mathbf{w}) - F(\mathbf{w}^*)}{\|\mathbf{w} - \mathbf{w}^*\|_2}. \tag{50}$$

Suppose that there exists $t \in \{0, 1, \ldots, T-1\}$ such that $\|\nabla F(\mathbf{w}^t)\|_2 < \sqrt{2}\Delta$. Then we have

$$F(\mathbf{w}^t) - F(\mathbf{w}^*) \leq \|\nabla F(\mathbf{w}^t)\|_2\|\mathbf{w}^t - \mathbf{w}^*\|_2 \leq 2\sqrt{2}D_0\Delta.$$

Otherwise, for all $t \in \{0, 1, \ldots, T-1\}$, $\|\nabla F(\mathbf{w}^t)\|_2 \geq \sqrt{2}\Delta$. Then, according to (49) and (50), we have for all $t < T$,

$$\begin{aligned}F(\mathbf{w}^{t+1}) - F(\mathbf{w}^*) &\leq F(\mathbf{w}^t) - F(\mathbf{w}^*) - \frac{1}{4L_F}\|\nabla F(\mathbf{w}^t)\|_2^2 \\ &\leq F(\mathbf{w}^t) - F(\mathbf{w}^*) - \frac{1}{4L_F D_t^2}(F(\mathbf{w}^t) - F(\mathbf{w}^*))^2.\end{aligned}$$

Multiplying both sides by $[(F(\mathbf{w}^{t+1}) - F(\mathbf{w}^*))(F(\mathbf{w}^t) - F(\mathbf{w}^*)]^{-1}$ and rearranging the terms, we obtain

$$\frac{1}{F(\mathbf{w}^{t+1}) - F(\mathbf{w}^*)} \geq \frac{1}{F(\mathbf{w}^t) - F(\mathbf{w}^*)} + \frac{1}{4L_F D_t^2}\frac{F(\mathbf{w}^t) - F(\mathbf{w}^*)}{F(\mathbf{w}^{t+1}) - F(\mathbf{w}^*)} \geq \frac{1}{F(\mathbf{w}^t) - F(\mathbf{w}^*)} + \frac{1}{16L_F D_0^2},$$

which implies

$$\frac{1}{F(\mathbf{w}^T) - F(\mathbf{w}^*)} \geq \frac{1}{F(\mathbf{w}^0) - F(\mathbf{w}^*)} + \frac{T}{16L_F D_0^2} \geq \frac{T}{16L_F D_0^2}.$$

Then, we obtain $F(\mathbf{w}^T) - F(\mathbf{w}^*) \leq 16D_0\Delta$ using the fact that $T = \frac{L_F D_0}{\Delta}$. □

Next, we show that $F(\mathbf{w}^T) - F(\mathbf{w}^*) \leq 16D_0\Delta + \frac{1}{2L_F}\Delta^2$. More specifically, let $t = t_0$ be the first time that $F(\mathbf{w}^t) - F(\mathbf{w}^*) \leq 16D_0\Delta$, and we show that for any $t > t_0$, $F(\mathbf{w}^t) - F(\mathbf{w}^*) \leq 16D_0\Delta + \frac{1}{2L_F}\Delta^2$. If this statement is not true, then we let $t_1 > t_0$ be the first time that $F(\mathbf{w}^t) - F(\mathbf{w}^*) > 16D_0\Delta + \frac{1}{2L_F}\Delta^2$. Then there must be $F(\mathbf{w}^{t_1-1}) < F(\mathbf{w}^{t_1})$. According to (49), there should also be

$$F(\mathbf{w}^{t_1-1}) - F(\mathbf{w}^*) \geq F(\mathbf{w}^{t_1}) - F(\mathbf{w}^*) - \frac{1}{2L_F}\Delta^2 > 16D_0\Delta.$$

Then, according to (50), we have

$$\|\nabla F(\mathbf{w}^{t_1-1})\|_2 \geq \frac{F(\mathbf{w}^{t_1-1}) - F(\mathbf{w}^*)}{\|\mathbf{w}^{t_1-1} - \mathbf{w}^*\|_2} > 8\Delta.$$

Then according to (49), this implies $F(\mathbf{w}^{t_1}) \leq F(\mathbf{w}^{t_1-1})$, which contradicts with the fact that $F(\mathbf{w}^{t_1-1}) < F(\mathbf{w}^{t_1})$.

D Proof of Theorem 3

Since Theorem 8 holds without assuming the convexity of $F(\mathbf{w})$, when $F(\mathbf{w})$ is non-convex, the event that (27) holds for all $\mathbf{w} \in \mathcal{W}$ still happens with probability at least $1 - \frac{4d}{(1+nm\widehat{L}D)^d}$. We condition on this event. We first show that when Assumption 5 is satisfied and we choose $\eta = \frac{1}{L_F}$, the iterates $\mathbf{w}^t$ stays in $\mathcal{W}$ without using projection. Since we have

$$\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2 \leq \|\mathbf{w}^t - \mathbf{w}^*\|_2 + \eta(\|\nabla F(\mathbf{w}^t)\|_2 + \|\mathbf{g}(\mathbf{w}^t) - \nabla F(\mathbf{w}^t)\|_2) \leq \|\mathbf{w}^t - \mathbf{w}^*\|_2 + \frac{1}{L_F}(M + \Delta).$$

Then, we know that by running $T = \frac{2L_F}{\Delta^2}(F(\mathbf{w}^0) - F(\mathbf{w}^*))$ parallel iterations, using Assumption 5, we know that $\mathbf{w}^t \in \mathcal{W}$ for $t = 0, 1, \ldots, T$ without projection.

We proceed to study the convergence rate of the algorithm. By the smoothness of $F(\mathbf{w})$, we know that when choosing $\eta = \frac{1}{L_F}$, the inequality (49) still holds. More specifically, for all $t = 0, 1, \ldots, T-1$,

$$F(\mathbf{w}^{t+1}) - F(\mathbf{w}^*) \leq F(\mathbf{w}^t) - F(\mathbf{w}^*) - \frac{1}{2L_F}\|\nabla F(\mathbf{w}^t)\|_2^2 + \frac{1}{2L_F}\Delta^2. \tag{51}$$

Sum up (51) for $t = 0, 1, \ldots, T-1$. Then, we get

$$0 \leq F(\mathbf{w}^T) - F(\mathbf{w}^*) \leq F(\mathbf{w}^0) - F(\mathbf{w}^*) - \frac{1}{2L_F}\sum_{t=0}^{T-1}\|\nabla F(\mathbf{w}^t)\|_2^2 + \frac{T}{2L_F}\Delta^2.$$

This implies that

$$\min_{t=0,1,\ldots,T}\|\nabla F(\mathbf{w}^t)\|_2^2 \leq 2\frac{L_F}{T}(F(\mathbf{w}^0) - F(\mathbf{w}^*)) + \Delta^2,$$

which completes the proof.

E Proof of Theorem 4

The proof of Theorem 4 consists of two parts: 1) the analysis of coordinate-wise trimmed mean of means estimator of the population gradients, and 2) the convergence analysis of the robustified gradient descent algorithm. Since the second part is essentially the same as the proof of Theorem 1, we mainly focus on the first part here.

Theorem 11. *Define*

$$\mathbf{g}^i(\mathbf{w}) = \begin{cases} \nabla F_i(\mathbf{w}) & i \in [m] \setminus \mathcal{B}, \\ * & i \in \mathcal{B}. \end{cases} \tag{52}$$

and the coordinate-wise trimmed mean of $\mathbf{g}^i(\mathbf{w})$*:*

$$\mathbf{g}(\mathbf{w}) = \mathsf{trmean}_\beta\{\mathbf{g}^i(\mathbf{w}) : i \in [m]\}. \tag{53}$$

Suppose that Assumptions 1 and 6 are satisfied, and that $\alpha \le \beta \le \frac{1}{2} - \epsilon$*. Then, with probability at least* $1 - \frac{2d(m+1)}{(1+nm\widehat{L}D)^d}$*,*

$$\|\mathbf{g}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \le \frac{v}{\epsilon}\left(\frac{3\sqrt{2}\beta d}{\sqrt{n}} + \frac{2d}{\sqrt{nm}}\right)\sqrt{\log(1+nm\widehat{L}D) + \frac{1}{d}\log m} + \widetilde{\mathcal{O}}(\frac{\beta}{n} + \frac{1}{nm})$$

for all $\mathbf{w} \in \mathcal{W}$*.*

Proof. See Appendix E.1 □

The rest of the proof is essentially the same as the proof of Theorem 1. In fact, we essentially analyze a gradient descent algorithm with bounded noise in the gradients. In the proof of Theorem 1 in Appendix B. The bound on the noise in the gradients is

$$\Delta = \sqrt{2}\frac{C_\epsilon}{\sqrt{n}}V(\alpha + \sqrt{\frac{d\log(1+nm\widehat{L}D)}{m(1-\alpha)}} + 0.4748\frac{S}{\sqrt{n}}) + 2\sqrt{2}\frac{1}{nm},$$

while here we replace Δ with Δ':

$$\Delta' := \frac{v}{\epsilon}\left(\frac{3\sqrt{2}\beta d}{\sqrt{n}} + \frac{2d}{\sqrt{nm}}\right)\sqrt{\log(1+nm\widehat{L}D) + \frac{1}{d}\log m} + \widetilde{\mathcal{O}}(\frac{\beta}{n} + \frac{1}{nm}),$$

and the same analysis can still go through. Therefore, we omit the details of the analysis here.

Remark 1. *The same arguments still go through when the population risk function* $F(\mathbf{w})$ *is non-strongly convex or non-convex. One can simply replace the bound on the noise in the gradients* Δ *in Theorems 2 and 3 with* Δ' *here. Thus we omit the details here.*

E.1 Proof of Theorem 11

The proof of Theorem 11 relies on the analysis of the trimmed mean of means estimator in the presence of adversarial data and a covering net argument. We first consider a general problem of robust estimation of a one dimensional random variable. Suppose that there are m worker machines, and q of them are Byzantine machines, which store n adversarial data (recall that $\alpha := q/m$). Each of the other $m(1-\alpha)$ normal worker machines stores n i.i.d. samples of some one dimensional random variable $x \sim \mathcal{D}$. Suppose that x is v-sub-exponential and let $\mu := \mathbb{E}[x]$. Denote the j-th sample in the i-th worker machine by $x^{i,j}$. In addition, define $\bar{x}^i$ as the average of samples in the i-th machine, i.e., $\bar{x}^i = \frac{1}{n}\sum_{j=1}^n x^{i,j}$. We have the following result on the trimmed mean of $\bar{x}^i$, $i \in [m]$.

Lemma 3. *Suppose that the one dimensional samples on all the normal machines are i.i.d.* v*-sub-exponential with mean* μ*. Then, we have for any* $t \ge 0$*,*

$$\mathbb{P}\{|\frac{1}{(1-\alpha)m}\sum_{i\in[m]\setminus\mathcal{B}}\bar{x}^i - \mu| \ge t\} \le 2\exp\{-(1-\alpha)mn\min\{\frac{t}{2v}, \frac{t^2}{2v^2}\}\},$$

and for any $s \ge 0$*,*

$$\mathbb{P}\{\max_{i\in[m]\setminus\mathcal{B}}\{|\bar{x}^i - \mu|\} \ge s\} \le 2(1-\alpha)m\exp\{-n\min\{\frac{s}{2v}, \frac{s^2}{2v^2}\}\},$$

and when $\beta \ge \alpha$*,* $|\frac{1}{(1-\alpha)m}\sum_{i\in[m]\setminus\mathcal{B}}\bar{x}^i - \mu| \le t$*, and* $\max_{i\in[m]\setminus\mathcal{B}}\{|\bar{x}^i - \mu|\} \le s$*, we have*

$$|\mathsf{trmean}_\beta\{\bar{x}^i : i \in [m]\} - \mu| \le \frac{t + 3\beta s}{1-2\beta}.$$

Proof. See Appendix E.2. □

Lemma 3 can be directly applied to the k-th partial derivative of the loss functions. Since we assume that for any $k \in [d]$ and $\mathbf{w} \in \mathcal{W}$, $\partial_k f(\mathbf{w}; \mathbf{z})$ is v-sub-exponential, we have for any $t \geq 0$, $s \geq 0$,

$$\mathbb{P}\{|\frac{1}{(1-\alpha)m} \sum_{i \in [m] \setminus \mathcal{B}} g_k^i(\mathbf{w}) - \partial_k F(\mathbf{w})| \geq t\} \leq 2\exp\{-(1-\alpha)mn \min\{\frac{t}{2v}, \frac{t^2}{2v^2}\}\}, \tag{54}$$

$$\mathbb{P}\{\max_{i \in [m] \setminus \mathcal{B}} \{|g_k^i(\mathbf{w}) - \partial_k F(\mathbf{w})|\} \geq s\} \leq 2(1-\alpha)m \exp\{-n \min\{\frac{s}{2v}, \frac{s^2}{2v^2}\}\}, \tag{55}$$

and consequently with probability at least

$$1 - 2\exp\{-(1-\alpha)mn \min\{\frac{t}{2v}, \frac{t^2}{2v^2}\}\} - 2(1-\alpha)m \exp\{-n \min\{\frac{s}{2v}, \frac{s^2}{2v^2}\}\},$$

we have

$$|g_k(\mathbf{w}) - \partial_k F(\mathbf{w})| = |\mathsf{trmean}_\beta\{g_k^i(\mathbf{w}) : i \in [m]\} - \partial_k F(\mathbf{w})| \leq \frac{t + 3\beta s}{1 - 2\beta}. \tag{56}$$

To extend this result to all $\mathbf{w} \in \mathcal{W}$ and all the d coordinates, we need to use union bound and a covering net argument. Let $\mathcal{W}_\delta = \{\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^{N_\delta}\}$ be a finite subset of $\mathcal{W}$ such that for any $\mathbf{w} \in \mathcal{W}$, there exists $\mathbf{w}^\ell \in \mathcal{W}_\delta$ such that $\|\mathbf{w}^\ell - \mathbf{w}\|_2 \leq \delta$. According to the standard covering net results (Vershynin, 2010), we know that $N_\delta \leq (1 + \frac{D}{\delta})^d$. By union bound, we know that with probability at least

$$1 - 2dN_\delta \exp\{-(1-\alpha)mn \min\{\frac{t}{2v}, \frac{t^2}{2v^2}\}\},$$

the bound $|\frac{1}{(1-\alpha)m} \sum_{i \in [m] \setminus \mathcal{B}} g_k^i(\mathbf{w}) - \partial_k F(\mathbf{w})| \leq t$ holds for all $\mathbf{w} = \mathbf{w}^\ell \in \mathcal{W}_\delta$, and $k \in [d]$, and with probability at least

$$1 - 2(1-\alpha)dmN_\delta \exp\{-n \min\{\frac{s}{2v}, \frac{s^2}{2v^2}\}\}$$

the bound $\max_{i \in [m] \setminus \mathcal{B}}\{|g_k^i(\mathbf{w}) - \partial_k F(\mathbf{w})|\} \leq s$ holds for all $\mathbf{w} = \mathbf{w}^\ell \in \mathcal{W}_\delta$, and $k \in [d]$. By gathering all the k coordinates, we know that this implies for all $\mathbf{w}^\ell \in \mathcal{W}_\delta$,

$$\|\mathbf{g}(\mathbf{w}^\ell) - \nabla F(\mathbf{w}^\ell)\|_2 \leq \sqrt{d} \frac{t + 3\beta s}{1 - 2\beta}. \tag{57}$$

Then, consider an arbitrary $\mathbf{w} \in \mathcal{W}$. Suppose that $\|\mathbf{w}^\ell - \mathbf{w}\|_2 \leq \delta$. Since by Assumption 1, we assume that for each $k \in [d]$, the partial derivative $\partial_k f(\mathbf{w}; \mathbf{z})$ is L_k-Lipschitz for all $\mathbf{z}$, we know that for every normal machine $i \in [m] \setminus \mathcal{B}$,

$$\left|g_k^i(\mathbf{w}) - g_k^i(\mathbf{w}^\ell)\right| \leq L_k \delta, \quad \left|\partial_k F(\mathbf{w}) - \partial_k F(\mathbf{w}^\ell)\right| \leq L_k \delta.$$

This means that if $|\frac{1}{(1-\alpha)m} \sum_{i \in [m] \setminus \mathcal{B}} g_k^i(\mathbf{w}^\ell) - \partial_k F(\mathbf{w}^\ell)| \leq t$ and $\max_{i \in [m] \setminus \mathcal{B}}\{|g_k^i(\mathbf{w}^\ell) - \partial_k F(\mathbf{w}^\ell)|\} \leq s$ hold for all $\mathbf{w}^\ell \in \mathcal{W}_\delta$, and $k \in [d]$, then

$$|\frac{1}{(1-\alpha)m} \sum_{i \in [m] \setminus \mathcal{B}} g_k^i(\mathbf{w}) - \partial_k F(\mathbf{w})| \leq t + 2L_k\delta,$$

and

$$\max_{i \in [m] \setminus \mathcal{B}}\{|g_k^i(\mathbf{w}) - \partial_k F(\mathbf{w})|\} \leq s + 2L_k\delta$$

hold for all $\mathbf{w} \in \mathcal{W}$. This implies that for all $\mathbf{w} \in \mathcal{W}$ and $k \in [d]$,

$$|g_k(\mathbf{w}) - \partial_k F(\mathbf{w})| = |\mathsf{trmean}_\beta\{g_k^i(\mathbf{w}) : i \in [m]\} - \partial_k F(\mathbf{w})| \leq \frac{t + 3\beta s}{1 - 2\beta} + \frac{2(1+3\beta)}{1-2\beta}\delta L_k,$$

which yields

$$\|\mathbf{g}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \leq \sqrt{2d} \frac{t + 3\beta s}{1 - 2\beta} + \sqrt{2} \frac{2(1+3\beta)}{1-2\beta} \delta \widehat{L}.$$

The proof is completed by choosing $\delta = \frac{1}{nm\widehat{L}}$,

$$t = v\max\{\frac{8d}{nm}\log(1+nm\widehat{L}D), \sqrt{\frac{8d}{nm}\log(1+nm\widehat{L}D)}\},$$
$$s = v\max\{\frac{4}{n}(d\log(1+nm\widehat{L}D)+\log m), \sqrt{\frac{4}{n}(d\log(1+nm\widehat{L}D)+\log m)}\},$$

and using the fact that $\beta \leq \frac{1}{2} - \epsilon$.

E.2 Proof of Lemma 3

We first recall Bernstein's inequality for sub-exponential random variables.

Theorem 12 (Bernstein's inequality)**.** *Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. v-sub-exponential random variables with mean μ. Then for any $t \geq 0$,*

$$\mathbb{P}\{|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| \geq t\} \leq 2\exp\{-n\min\{\frac{t}{2v}, \frac{t^2}{2v^2}\}\}.$$

Thus, for any $t \geq 0$

$$\mathbb{P}\{|\frac{1}{(1-\alpha)m}\sum_{i\in[m]\setminus\mathcal{B}} \bar{x}^i - \mu| \geq t\} \leq 2\exp\{-(1-\alpha)mn\min\{\frac{t}{2v}, \frac{t^2}{2v^2}\}\}. \tag{58}$$

Similarly, for any $i \in [m]\setminus\mathcal{B}$, and any $s \geq 0$

$$\mathbb{P}\{|\bar{x}^i - \mu| \geq s\} \leq 2\exp\{-n\min\{\frac{s}{2v}, \frac{s^2}{2v^2}\}\}.$$

Then, by union bound we know that

$$\mathbb{P}\{\max_{i\in[m]\setminus\mathcal{B}}\{|\bar{x}^i - \mu|\} \geq s\} \leq 2(1-\alpha)m\exp\{-n\min\{\frac{s}{2v}, \frac{s^2}{2v^2}\}\}. \tag{59}$$

We proceed to analyze the trimmed mean of means estimator. To simplify notation, we define $\mathcal{M} = [m]\setminus\mathcal{B}$ as the set of all normal worker machines, $\mathcal{U} \subseteq [m]$ as the set of all untrimmed machines, and $\mathcal{T} \subseteq [m]$ as the set of all trimmed machines. The trimmed mean of means estimator simply computes

$$\mathsf{trmean}_\beta\{\bar{x}^i : i \in [m]\} = \frac{1}{(1-2\beta)m}\sum_{i\in\mathcal{U}} \bar{x}^i.$$

We further have

$$\begin{aligned}
|\mathsf{trmean}_\beta\{\bar{x}^i : i \in [m]\} - \mu| &= \left|\frac{1}{(1-2\beta)m}\sum_{i\in\mathcal{U}} \bar{x}^i - \mu\right| \\
&= \frac{1}{(1-2\beta)m}\left|\sum_{i\in\mathcal{M}}(\bar{x}^i - \mu) - \sum_{i\in\mathcal{M}\cap\mathcal{T}}(\bar{x}^i - \mu) + \sum_{i\in\mathcal{B}\cap\mathcal{U}}(\bar{x}^i - \mu)\right| \\
&\leq \frac{1}{(1-2\beta)m}\big(|\sum_{i\in\mathcal{M}}(\bar{x}^i - \mu)| + |\sum_{i\in\mathcal{M}\cap\mathcal{T}}(\bar{x}^i - \mu)| + |\sum_{i\in\mathcal{B}\cap\mathcal{U}}(\bar{x}^i - \mu)|\big)
\end{aligned} \tag{60}$$

We also know that $|\sum_{i\in\mathcal{M}\cap\mathcal{T}}(\bar{x}^i - \mu)| \leq 2\beta m\max_{i\in\mathcal{M}}\{|\bar{x}^i - \mu|\}$. In addition, since $\beta \geq \alpha$, without loss of generality, we assume that $\mathcal{M}\cap\mathcal{T} \neq \emptyset$, and then $|\sum_{i\in\mathcal{B}\cap\mathcal{U}}(\bar{x}^i - \mu)| \leq \alpha m\max_{i\in\mathcal{M}}\{|\bar{x}^i - \mu|\}$. Then we directly obtain the desired result.

F Proof of Theorem 7

Since the loss functions are quadratic, we denote the loss function $f(\mathbf{w}; \mathbf{z}^{i,j})$ by

$$f(\mathbf{w}; \mathbf{z}^{i,j}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{H}_{i,j}\mathbf{w} + \mathbf{p}_{i,j}^{\mathrm{T}}\mathbf{w} + c_{i,j}.$$

We further define $\mathbf{H}_i := \frac{1}{n}\sum_{j=1}^n \mathbf{H}_{i,j}$, $\mathbf{p}_i := \frac{1}{n}\sum_{j=1}^n \mathbf{p}_{i,j}$, and $c_i := \frac{1}{n}\sum_{j=1}^n c_{i,j}$. Thus the empirical risk function on the i-th machine is

$$F_i(\mathbf{w}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{H}_i\mathbf{w} + \mathbf{p}_i^{\mathrm{T}}\mathbf{w} + c_i.$$

Then, for any worker machine $i \in [m] \setminus \mathcal{B}$, $\widehat{\mathbf{w}}^i = -\mathbf{H}_i^{-1}\mathbf{p}_i$. In addition, the population risk minimizer is $\mathbf{w}^* = -\mathbf{H}_F^{-1}\mathbf{p}_F$. We further define $\mathbf{U}_{i,j} := \mathbf{H}_{i,j} - \mathbf{H}_F$, $\mathbf{U}_i = \mathbf{H}_i - \mathbf{H}_F$, $\mathbf{v}_{i,j} = \mathbf{p}_{i,j} - \mathbf{p}_F$, and $\mathbf{v}_i = \mathbf{p}_i - \mathbf{p}_F$. Then

$$\widehat{\mathbf{w}}^i = -(\mathbf{U}_i + \mathbf{H}_F)^{-1}(\mathbf{v}_i + \mathbf{p}_F).$$

Let $\mathbf{e}_k$ be the k-th vector in the standard basis, i.e., the k-th column of the $d \times d$ identity matrix. We proceed to study the distribution of the k-th coordinate of $\widehat{\mathbf{w}}^i - \mathbf{w}^*$, $i \in [m] \setminus \mathcal{B}$, i.e.,

$$\widehat{w}_k^i - w_k^* = \mathbf{e}_k^{\mathrm{T}}\mathbf{H}_F^{-1}\mathbf{p}_F - \mathbf{e}_k^{\mathrm{T}}(\mathbf{U}_i + \mathbf{H}_F)^{-1}(\mathbf{v}_i + \mathbf{p}_F).$$

Similar to the proof of Theorem 1, we need to obtain a Berry-Esseen type bound for $\widehat{w}_k^i - w_k^*$. However, here, $\widehat{w}_k^i$ is not a sample mean of n i.i.d. random variables, and thus we cannot directly apply the vanilla Berry-Esseen bound. Instead, we apply the following bound in Pinelis and Molzon (2016) on functions of sample means.

Theorem 13 (Theorem 2.11 in Pinelis and Molzon (2016), simplified)**.** *Let $\mathcal{X}$ be a Hilbert space equipped with norm $\|\cdot\|$. Let $f : \mathcal{X} \to \mathbb{R}$ be a function on $\mathcal{X}$. Suppose that there exists linear functions $\ell : \mathcal{X} \to \mathbb{R}$, $\theta > 0$, $M_\theta > 0$ such that*

$$|f(X) - \ell(X)| \le \frac{M_\theta}{2}\|X\|^2, \ \forall\ \|X\| \le \theta. \tag{61}$$

Suppose that there is a probability distribution $\mathcal{D}_X$ over $\mathcal{X}$, and let $X, X_1, X_2, \ldots, X_n$ be i.i.d. random variables drawn from $\mathcal{D}_X$. Assume that $\mathbb{E}[X] = 0$, and define

$$\widetilde{\sigma} := (\mathbb{E}[\ell(X)^2])^{1/2}, \quad \nu_p := (\mathbb{E}[\|X\|^p])^{1/p}, p = 2, 3, \quad \varsigma := \frac{(\mathbb{E}[|\ell(X)|^3])^{1/3}}{\widetilde{\sigma}}.$$

Let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Then for any $z \in \mathbb{R}$, we have

$$\left|\mathbb{P}\left\{\frac{f(\bar{X})}{\widetilde{\sigma}/\sqrt{n}} \le z\right\} - \Phi(z)\right| \le \frac{C}{\sqrt{n}}, \tag{62}$$

where $C = C_0 + C_1\varsigma^3 + (C_{20} + C_{21}\varsigma)\nu_2^2 + (C_{30} + C_{31}\varsigma)\nu_3^2 + C_4$, with

$$\begin{aligned} &C_0 = 0.1393, \quad C_1 = 2.3356 \\ &(C_{20}, C_{21}, C_{30}, C_{31}) = \frac{M_\theta}{2\widetilde{\sigma}}\left(2(\frac{2}{\pi})^{1/6}, 2 + \frac{2^{2/3}}{n^{1/6}}, \frac{(8/\pi)^{1/6}}{n^{1/3}}, \frac{2}{n^{1/2}}\right) \\ &C_4 = \min\{\frac{\nu_2^2}{\theta^2 n^{1/2}}, \frac{2\nu_2^3 + \nu_3^3/n^{1/2}}{\theta^3 n}\}. \end{aligned} \tag{63}$$

Define the function $\psi_k(\mathbf{U}, \mathbf{v}) : \mathbb{R}^{d\times d} \times \mathbb{R} \to \mathbb{R}$:

$$\psi_k(\mathbf{U}, \mathbf{v}) := \mathbf{e}_k^{\mathrm{T}}\mathbf{H}_F^{-1}\mathbf{p}_F - \mathbf{e}_k^{\mathrm{T}}(\mathbf{U} + \mathbf{H}_F)^{-1}(\mathbf{v} + \mathbf{p}_F),$$

and thus

$$\widehat{w}_k^i - w_k^* = \psi_k(\mathbf{U}_i, \mathbf{v}_i) = \psi_k(\frac{1}{n}\sum_{j=1}^n \mathbf{U}_{i,j}, \frac{1}{n}\sum_{j=1}^n \mathbf{v}_{i,j}).$$

On the product space $\mathbb{R}^{d\times d}\times\mathbb{R}$, define the element-wise inner product:

$$\langle(\mathbf{U},\mathbf{v}),(\mathbf{X},\mathbf{y})\rangle=\sum_{i,j=1}^{d}U_{i,j}X_{i,j}+\sum_{i=1}^{d}v_iy_i,$$

and thus $\mathbb{R}^{d\times d}\times\mathbb{R}$ is associated with the norm

$$\|(\mathbf{U},\mathbf{v})\|=\sqrt{\|\mathbf{U}\|_F^2+\|\mathbf{v}\|_2^2},$$

where $\|\cdot\|_F$ denotes the Frobenius norm of matrices. We then provide the following lemma on $\psi_k(\mathbf{U},\mathbf{v})$.

Lemma 4. *There exits a linear function* $\ell_k(\mathbf{U},\mathbf{v})=\mathbf{e}_k^{\mathrm{T}}\mathbf{H}_F^{-1}\mathbf{U}\mathbf{H}_F^{-1}\mathbf{p}_F-\mathbf{e}_k^{\mathrm{T}}\mathbf{H}_F^{-1}\mathbf{v}$ *such that for any* $\mathbf{U}$, $\mathbf{v}$ *with*

$$\|\mathbf{U}\|_F^2+\|\mathbf{v}\|_2^2\le\frac{\lambda_F^2}{4},$$

we have

$$|\psi_k(\mathbf{U},\mathbf{v})-\ell_k(\mathbf{U},\mathbf{v})|\le\frac{\lambda_F+2\|\mathbf{p}_F\|_2}{\lambda_F^3}(\|\mathbf{U}\|_F^2+\|\mathbf{v}\|_2^2).$$

Proof. See Appendix F.1. □

Lemma 4 tells us that the condition (61) is satisfied with $\theta=\frac{\lambda_F}{2}$ and $M_\theta=\frac{2\lambda_F+4\|\mathbf{p}_F\|_2}{\lambda_F^3}$. For all normal worker machine $i\in[m]\setminus\mathcal{B}$, denote the distribution of $\mathbf{U}_{i,j}$ and $\mathbf{v}_{i,j}$ by $\mathcal{D}_U$ and $\mathcal{D}_v$, respectively. Since $\widehat{w}_k^i-w_k^*=\psi_k(\frac{1}{n}\sum_{j=1}^n\mathbf{U}_{i,j},\frac{1}{n}\sum_{j=1}^n\mathbf{v}_{i,j})$, Theorem 13 directly gives us the following lemma.

Lemma 5. *Let* $\mathbf{U}\sim\mathcal{D}_U$, $\mathbf{v}\sim\mathcal{D}_v$, *and* $\ell_k(\mathbf{U},\mathbf{v})=\mathbf{e}_k^{\mathrm{T}}\mathbf{H}_F^{-1}\mathbf{U}\mathbf{H}_F^{-1}\mathbf{p}_F-\mathbf{e}_k^{\mathrm{T}}\mathbf{H}_F^{-1}\mathbf{v}$. *Define*

$$\widetilde{\sigma}_k:=(\mathbb{E}[\ell_k(\mathbf{U},\mathbf{v})^2])^{1/2},\quad \nu_p:=(\mathbb{E}[(\|\mathbf{U}\|_F^2+\|\mathbf{v}\|_2^2)^{p/2}])^{1/p},p=2,3,\quad \varsigma_k:=\frac{(\mathbb{E}[|\ell_k(\mathbf{U},\mathbf{v})|^3])^{1/3}}{\widetilde{\sigma}_k}.$$

Then for any $z\in\mathbb{R}$, $i\in[m]\setminus\mathcal{B}$, *we have*

$$\left|\mathbb{P}\left\{\frac{\widehat{w}_k^i-w_k^*}{\widetilde{\sigma}_k/\sqrt{n}}\le z\right\}-\Phi(z)\right|\le\frac{C_k}{\sqrt{n}},\tag{64}$$

where

$$C_k=\widehat{C}_0+\widehat{C}_1\varsigma_k^3+\frac{1}{\widetilde{\sigma}_k}\big[(\widehat{C}_{20}+\widehat{C}_{21}\varsigma_k)\nu_2^2+(\widehat{C}_{30}+\widehat{C}_{31}\varsigma_k)\nu_3^2\big]+\widehat{C}_4,$$

with

$$\begin{aligned}&\widehat{C}_0=0.1393,\quad \widehat{C}_1=2.3356\\&(\widehat{C}_{20},\widehat{C}_{21},\widehat{C}_{30},\widehat{C}_{31})=\frac{\lambda_F+2\|\mathbf{p}_F\|_2}{\lambda_F^3}\left(2(\frac{2}{\pi})^{1/6},2+\frac{2^{2/3}}{n^{1/6}},\frac{(8/\pi)^{1/6}}{n^{1/3}},\frac{2}{n^{1/2}}\right)\\&\widehat{C}_4=\min\{\frac{4\nu_2^2}{\lambda_F^2n^{1/2}},\frac{16\nu_2^3+8\nu_3^3/n^{1/2}}{\lambda_F^3n}\}.\end{aligned}\tag{65}$$

Then, we proceed to bound $\mathsf{med}\{\widehat{w}_k^i:i\in[m]\}-w_k^*$, the technique is similar to what we use in the proof of Theorem 8. For every $z\in\mathbb{R}$, $k\in[d]$, define

$$\widetilde{p}(z;k)=\frac{1}{m(1-\alpha)}\sum_{i\in[m]\setminus\mathcal{B}}\mathbb{1}(\widehat{w}_k^i-w_k^*\le z).$$

We have the following lemma on $\widetilde{p}(z;k)$.

Lemma 6. *Suppose that for a fixed* $t>0$, *we have*

$$\alpha+\sqrt{\frac{t}{m(1-\alpha)}}+\frac{C_k}{\sqrt{n}}\le\frac{1}{2}-\epsilon,\tag{66}$$

for some $\epsilon > 0$. Then, with probability at least $1 - 4e^{-2t}$, we have

$$\widetilde{p}\left(C_\epsilon \frac{\widetilde{\sigma}_k}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + \frac{C_k}{\sqrt{n}}); k\right) \geq \frac{1}{2} + \alpha, \tag{67}$$

and

$$\widetilde{p}\left(-C_\epsilon \frac{\widetilde{\sigma}_k}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + \frac{C_k}{\sqrt{n}}); k\right) \leq \frac{1}{2} - \alpha, \tag{68}$$

where C_ϵ is defined as in (4).

Proof. The proof is essentially the same as the proof of Lemma 1. One can simply replace σ in Lemma 1 with $\widetilde{\sigma}_k$ and $0.4748\gamma(x)$ in Lemma 1 with C_k. Then the same arguments still apply. Thus, we skip the details of this proof. □

Then, define $\widehat{p}(z;k) = \frac{1}{m}\sum_{i\in[m]} \mathbb{1}(\widehat{w}_k^i - w_k^* \leq z)$. Using the same arguments as in Corollary 1, we know that

$$\widehat{p}\left(C_\epsilon \frac{\widetilde{\sigma}_k}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + \frac{C_k}{\sqrt{n}}); k\right) \geq \frac{1}{2},$$

and

$$\widehat{p}\left(-C_\epsilon \frac{\widetilde{\sigma}_k}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + \frac{C_k}{\sqrt{n}}); k\right) \leq \frac{1}{2},$$

which implies that $|\mathsf{med}\{\widehat{w}_k^i : i \in [m]\} - w_k^*| \leq C_\epsilon \frac{\widetilde{\sigma}_k}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + \frac{C_k}{\sqrt{n}})$. Then, let

$$\widetilde{\sigma} := \sqrt{\sum_{k=1}^d \widetilde{\sigma}_k^2} = \sqrt{\mathbb{E}[\|\mathbf{H}_F^{-1}(\mathbf{U}\mathbf{H}_F^{-1}\mathbf{p}_F - \mathbf{v})\|_2^2]},$$

and $\widetilde{C} = \max_{k\in[d]} C_k$, we have with probability at least $1 - 4de^{-2t}$,

$$\|\mathsf{med}\{\widehat{\mathbf{w}}^i : i \in [m]\} - \mathbf{w}^*\|_2 \leq \frac{C_\epsilon}{\sqrt{n}}\widetilde{\sigma}\big(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + \frac{\widetilde{C}}{\sqrt{n}}\big).$$

We complete the proof by choosing $t = \frac{1}{2}\log(nmd)$.

Explicit expression of $\widetilde{C}$. To summarize, we provide an explicit expression of $\widetilde{C}$. Let $\mathbf{e}_k$ be the k-th vector in the standard basis, i.e., the k-th column of the $d \times d$ identity matrix, and define $\ell_k(\mathbf{U}, \mathbf{v}) : \mathbb{R}^{d\times d} \times \mathbb{R} \to \mathbb{R}$ as

$$\ell_k(\mathbf{U}, \mathbf{v}) = \mathbf{e}_k^{\mathrm{T}}\mathbf{H}_F^{-1}\mathbf{U}\mathbf{H}_F^{-1}\mathbf{p}_F - \mathbf{e}_k^{\mathrm{T}}\mathbf{H}_F^{-1}\mathbf{v}.$$

Let $\mathbf{H} \sim \mathcal{D}_H$ and $\mathbf{p} \sim \mathcal{D}_p$ and define

$$\widetilde{\sigma}_k := (\mathbb{E}[\ell_k(\mathbf{H} - \mathbf{H}_F, \mathbf{p} - \mathbf{p}_F)^2])^{1/2}, \quad \varsigma_k := \frac{(\mathbb{E}[|\ell_k(\mathbf{H} - \mathbf{H}_F, \mathbf{p} - \mathbf{p}_F)|^3])^{1/3}}{\widetilde{\sigma}_k}.$$
$$\nu_p := (\mathbb{E}[(\|\mathbf{H} - \mathbf{H}_F\|_F^2 + \|\mathbf{p} - \mathbf{p}_F\|_2^2)^{p/2}])^{1/p}, p = 2, 3$$

Then, $\widetilde{C} = \max_{k\in[d]} C_k$, with where

$$C_k = \widehat{C}_0 + \widehat{C}_1\varsigma_k^3 + \frac{1}{\widetilde{\sigma}_k}\big[(\widehat{C}_{20} + \widehat{C}_{21}\varsigma_k)\nu_2^2 + (\widehat{C}_{30} + \widehat{C}_{31}\varsigma_k)\nu_3^2\big] + \widehat{C}_4,$$

with

$$\widehat{C}_0 = 0.1393, \quad \widehat{C}_1 = 2.3356$$
$$(\widehat{C}_{20}, \widehat{C}_{21}, \widehat{C}_{30}, \widehat{C}_{31}) = \frac{\lambda_F + 2\|\mathbf{p}_F\|_2}{\lambda_F^3}\left(2(\frac{2}{\pi})^{1/6}, 2 + \frac{2^{2/3}}{n^{1/6}}, \frac{(8/\pi)^{1/6}}{n^{1/3}}, \frac{2}{n^{1/2}}\right)$$
$$\widehat{C}_4 = \min\{\frac{4\nu_2^2}{\lambda_F^2 n^{1/2}}, \frac{16\nu_2^3 + 8\nu_3^3/n^{1/2}}{\lambda_F^3 n}\}.$$

F.1 Proof of Lemma 4

We use $\|\cdot\|_2$ and $\|\cdot\|_F$ to denote the operator norm and the Frobenius norm of matrices, respectively. We have the identity

$$(\mathbf{I}+\mathbf{A})^{-1} = \sum_{r=0}^{\infty}(-1)^r\mathbf{A}^r, \quad \forall\|\mathbf{A}\|_2 < 1.$$

Then, we have for all $\mathbf{U}\in\mathbb{R}^{d\times d}$ such that $\|\mathbf{H}_F^{-1}\mathbf{U}\|_2 < 1$,

$$(\mathbf{U}+\mathbf{H}_F)^{-1} = (\mathbf{I}+\mathbf{H}_F^{-1}\mathbf{U})^{-1}\mathbf{H}_F^{-1} = \mathbf{H}_F^{-1} - \mathbf{H}_F^{-1}\mathbf{U}\mathbf{H}_F^{-1} + \sum_{r=2}^{\infty}(-1)^r(\mathbf{H}_F^{-1}\mathbf{U})^r\mathbf{H}_F^{-1}. \tag{69}$$

Let us consider the set of matrices such that $\|\mathbf{U}\|_F \le \frac{\lambda_F}{2}$. One can check that for any such matrix, we have $\|\mathbf{H}_F^{-1}\mathbf{U}\|_2 \le \frac{1}{2}$. Let

$$\ell_k(\mathbf{U},\mathbf{v}) = \mathbf{e}_k^{\mathrm{T}}\mathbf{H}_F^{-1}\mathbf{U}\mathbf{H}_F^{-1}\mathbf{p}_F - \mathbf{e}_k^{\mathrm{T}}\mathbf{H}_F^{-1}\mathbf{v}.$$

Then, we know that

$$|\psi_k(\mathbf{U},\mathbf{v}) - \ell_k(\mathbf{U},\mathbf{v})| = \left|\mathbf{e}_k^{\mathrm{T}}\mathbf{H}_F^{-1}\mathbf{U}\mathbf{H}_F^{-1}\mathbf{v} - \sum_{r=2}^{\infty}(-1)^r\mathbf{e}_k^{\mathrm{T}}(\mathbf{H}_F^{-1}\mathbf{U})^r\mathbf{H}_F^{-1}(\mathbf{v}+\mathbf{p}_F)\right|. \tag{70}$$

Denote the operator norm of matrices by $\|\cdot\|_2$. We further have for any $r\ge 1$,

$$|\mathbf{e}_k^{\mathrm{T}}(\mathbf{H}_F^{-1}\mathbf{U})^r\mathbf{H}_F^{-1}\mathbf{v}| \le \frac{1}{2}\|\mathbf{H}_F^{-1}\mathbf{U}\|_2^{r-1}(\|\mathbf{H}_F^{-1}\mathbf{U}\|_2^2 + \|\mathbf{H}_F^{-1}\mathbf{v}\|_2^2) \le \frac{1}{2^r\lambda_F^2}(\|\mathbf{U}\|_F^2 + \|\mathbf{v}\|_2^2), \tag{71}$$

where we use the fact $\|\mathbf{U}\|_2 \le \|\mathbf{U}\|_F$. In addition, for any $r\ge 2$,

$$|\mathbf{e}_k^{\mathrm{T}}(\mathbf{H}_F^{-1}\mathbf{U})^r\mathbf{H}_F^{-1}\mathbf{p}_F| \le \|\mathbf{H}_F^{-1}\mathbf{U}\|_2^{r-2}\|\mathbf{H}_F^{-1}\|_2^3\|\mathbf{U}\|_2^2\|\mathbf{p}_F\|_2 \le \frac{\|\mathbf{p}_F\|_2}{2^{r-2}\lambda_F^3}\|\mathbf{U}\|_F^2. \tag{72}$$

Then, we plug (71) and (72) into (70), and obtain

$$|\psi_k(\mathbf{U},\mathbf{v}) - \ell_k(\mathbf{U},\mathbf{v})| \le \frac{1}{\lambda_F^2}(\|\mathbf{U}\|_F^2 + \|\mathbf{v}\|_2^2) + \frac{2\|\mathbf{p}_F\|_2}{\lambda_F^3}\|\mathbf{U}\|_F^2,$$

which completes the proof.

G Proof of Observation 1

This proof is essentially the same as the lower bound in the robust mean estimation literature (Chen et al., 2015, Lai et al., 2016). We reproduce this result for the purpose of completeness. For a d dimensional Gaussian distribution $P = \mathcal{N}(\boldsymbol{\mu},\sigma^2\mathbf{I})$, we denote by P^n the joint distribution of n i.i.d. samples of P. Obviously P^n is equivalent to a dn dimensional Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}^+,\sigma^2\mathbf{I})$, where $\boldsymbol{\mu}^+\in\mathbb{R}^{dn}$ is a vector generated by repeating $\boldsymbol{\mu}$ n times, i.e., $\boldsymbol{\mu}^+ = [\boldsymbol{\mu}^{\mathrm{T}}\ \boldsymbol{\mu}^{\mathrm{T}}\ \cdots\ \boldsymbol{\mu}^{\mathrm{T}}]^{\mathrm{T}}$.

We show that for two d dimensional distributions $P_1 = \mathcal{N}(\boldsymbol{\mu}_1,\sigma^2\mathbf{I})$ and $P_2 = \mathcal{N}(\boldsymbol{\mu}_2,\sigma^2\mathbf{I})$, there exist two dn dimensional distributions Q_1 and Q_2 such that

$$(1-\alpha)P_1^n + \alpha Q_1 = (1-\alpha)P_2^n + \alpha Q_2. \tag{73}$$

If this happens, then no algorithm can distinguish between P_1 and P_2. Let ϕ_1 and ϕ_2 be the PDF of P_1^n and P_2^n, respectively. Let $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ be such that the total variation distance between P_1^n and P_2^n is

$$\frac{1}{2}\int\|\phi_1-\phi_2\|_1 = \frac{\alpha}{1-\alpha}.$$

By the results of the total variation distance between Gaussian distributions, we know that

$$\|\boldsymbol{\mu}_1^+ - \boldsymbol{\mu}_2^+\|_2 \ge \frac{2\alpha\sigma}{1-\alpha}. \tag{74}$$

Let Q_1 be the distribution with PDF $\frac{1-\alpha}{\alpha}(\phi_2-\phi_1)\mathbb{1}_{\phi_2\geq\phi_1}$ and Q_2 be the distribution with PDF $\frac{1-\alpha}{\alpha}(\phi_1-\phi_2)\mathbb{1}_{\phi_1\geq\phi_2}$. One can verify that (73) is satisfied. In this case, by the lower bound in (74), we get

$$\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|_2\geq\frac{2\alpha\sigma}{\sqrt{n}(1-\alpha)}\geq\frac{2\alpha\sigma}{\sqrt{n}}.$$

This implies that for two Gaussian distributions such that $\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|_2=\Omega(\frac{\alpha}{\sqrt{n}})$, in the worst case it can be impossible to distinguish these two distributions due to the existence of the adversary. Thus, to estimate the mean $\boldsymbol{\mu}$ of a Gaussian distribution in the distributed setting with α fraction of Byzantine machines, any algorithm that computes an estimation $\widehat{\boldsymbol{\mu}}$ of the mean has a constant probability of error $\|\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|_2=\Omega(\frac{\alpha}{\sqrt{n}})$.

Further, according to the standard results from minimax theory (Wu, 2017), we know that using $\mathcal{O}(nm)$ data, there is a constant probability that $\|\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|_2=\Omega(\sqrt{\frac{d}{nm}})$. Combining these two results, we know that $\|\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|_2=\Omega(\frac{\alpha}{\sqrt{n}}+\sqrt{\frac{d}{nm}})$.