Journal of Machine Learning Research 26 (2025) 1-51 Submitted 8/24; Revised 12/24; Published 1/25

Mean Aggregator is More Robust than Robust Aggregators under Label Poisoning Attacks on Distributed Heterogeneous Data

Jie Peng PENGJ95@MAIL2.SYSU.EDU.CN
School of Computer Science and Engineering
Sun Yat-Sen University
Guangzhou, Guangdong 510006, China

Weiyu Li WEIYULI@G.HARVARD.EDU
School of Engineering and Applied Science
Harvard University
Cambridge, MA 02138, USA

Stefan Vlaski S.VLASKI@IMPERIAL.AC.UK
Department of Electrical and Electronic Engineering
Imperial College London
London SW7 2BT, UK

Qing Ling LINGQING556@MAIL.SYSU.EDU.CN
School of Computer Science and Engineering
Sun Yat-Sen University
Guangzhou, Guangdong 510006, China

Editor: Shiqian Ma

Abstract

Robustness to malicious attacks is of paramount importance for distributed learning. Existing works usually consider the classical Byzantine attacks model, which assumes that some workers can send arbitrarily malicious messages to the server and disturb the aggregation steps of the distributed learning process. To defend against such worst-case Byzantine attacks, various robust aggregators have been proposed. They are proven to be effective and much superior to the often-used mean aggregator. In this paper, however, we demonstrate that the robust aggregators are too conservative for a class of weak but practical malicious attacks, known as label poisoning attacks, where the sample labels of some workers are poisoned. Surprisingly, we are able to show that the mean aggregator is more robust than the state-of-the-art robust aggregators in theory, given that the distributed data are sufficiently heterogeneous. In fact, the learning error of the mean aggregator is proven to be order-optimal in this case. Experimental results corroborate our theoretical findings, showing the superiority of the mean aggregator under label poisoning attacks.

Keywords: distributed learning, Byzantine attacks, label poisoning attacks

1. Introduction

With the rising and rapid development of large machine learning models, distributed learning has attracted intensive research attention due to its provable effectiveness in solving large-scale problems (Verbraeken et al., 2020; Li et al., 2020). In distributed learning, there often exist one parameter server (called server thereafter) owning the global model and some computation devices (called

workers thereafter) owning the local data. In the training process, the server sends the global model to the workers, and the workers use their local data to compute the local stochastic gradients or momenta of the global model and send them back to the server. Upon receiving the messages from all workers, the server aggregates them and uses the aggregated stochastic gradient or momentum to update the global model. After the training process, the trained global model is evaluated on the testing data. An essential component of distributed learning is federated learning (McMahan et al., 2017; Yang et al., 2019; Gosselin et al., 2022; Ye et al., 2023; Fraboni et al., 2023), which is particularly favorable in terms of privacy preservation.

However, the distributed nature of such a server-worker architecture is vulnerable to malicious attacks during the learning process (Lewis et al., 2023). Due to data corruptions, equipment failures, or cyber attacks, some workers may not follow the algorithmic protocol, and instead send incorrect messages to the server. Previous works often characterize these attacks by the classical Byzantine attacks model, which assumes that some workers can send arbitrarily malicious messages to the server so that the aggregation steps of the learning process are disturbed (Lamport et al., 1982). For such worst-case Byzantine attacks, various robust aggregators have been proven effective and much superior to the mean aggregator (Chen et al., 2017; Xia et al., 2019; Karimireddy et al., 2021; Wu et al., 2023).

The malicious attacks encountered in reality, on the other hand, are often less destructive than the worst-case Byzantine attacks. For example, a distributed learning system may often suffer from label poisoning attacks, which are weak yet of practical interest. Considering a highly secure email system in a large organization (for example, government or university), if hackers (some users) aim to disturb the online training process of a spam detection model, one of the most effective ways for them is to mislabel received emails from "spam" to "non-spam", resulting in label poisoning attacks. Similar attacks may happen in fraudulent short message service (SMS) detection held by large communication corporations, too.

To this end, in this paper, we consider label poisoning attacks where some workers have local data with poisoned labels and generate incorrect messages during the learning process. Under label poisoning attacks and with some mild assumptions, surprisingly we are able to show that the mean aggregator is more robust than the state-of-the-art robust aggregators in theory. To be specific, we prove that the mean aggregator has a better learning error bound than the robust aggregators (see Theorems 7 and 8 in Section 4), given that the distributed data are sufficiently heterogeneous. The main contributions of this paper are summarized as follows.

C1) To the best of our knowledge, our work is the first to investigate the robustness of the mean aggregator in distributed learning. Our work reveals an important fact that the mean aggregator is more robust than the existing robust aggregators under specific types of malicious attacks, which motivates us to rethink the usage of different aggregators within practical scenarios.

C2) Under label poisoning attacks, we theoretically analyze the learning errors of the mean aggregator and the state-of-the-art robust aggregators. The results show that when the heterogeneity of the distributed data is large, the learning error of the mean aggregator is order-optimal regardless of the fraction of poisoned workers.

C3) We empirically evaluate the performance of the mean aggregator and the state-of-the-art robust aggregators under label poisoning attacks. The experimental results fully support our theoretical findings.

This paper significantly extends upon our previous conference paper (Peng et al., 2024). First, Peng et al. (2024) considered the distributed gradient descent algorithm, which ignores the stochas-

tic gradient noise that is critical to distributed learning. To address this issue, we investigate distributed stochastic momentum as the backbone algorithm, and provide new theoretical analysis for distributed stochastic momentum with the mean aggregator or the robust aggregators. By properly handling the stochastic gradient noise, we establish the tight upper bounds and the lower bound for the learning error. These theoretical results recover those in Peng et al. (2024) if the momentum coefficient is set to 1 and the inner variance bound is 0 (see Remark 10). Second, we present new, comprehensive experimental results to compare the performance of various aggregators combined with the distributed stochastic momentum algorithm, validating our theoretical findings.

2. Related Works

Poisoning attacks can be categorized into targeted attacks and untargeted attacks; or model poisoning attacks and data poisoning attacks (Kairouz et al., 2021). In this paper, we focus on the latter categorization. In model poisoning attacks, the malicious workers send arbitrarily poisoned models to the server, while data poisoning attacks yield poisoned messages by fabricating poisoned data at the malicious workers' side (Shejwalkar et al., 2022). Below we briefly review the related works of the two types of poisoning attacks in distributed learning, respectively.

Under model poisoning attacks, most of the existing works design robust aggregators for aggregating local stochastic gradients of the workers and filter out the potentially poisoned messages. The existing robust aggregators include Krum (Blanchard et al., 2017), geometric median (Chen et al., 2017), coordinate-wise median (Yin et al., 2018), coordinate-wise trimmed-mean (Yin et al., 2018), FABA (Xia et al., 2019), centered clipping (Karimireddy et al., 2021), VRMOM (Tu et al., 2021), etc. The key idea behind these robust aggregators is to find a point that has bounded distance to the true stochastic gradient such that the learning error is under control. Farhadkhani et al. (2022) and Allouah et al. (2023) propose a unified framework to analyze the performance of these robust aggregators under attacks. However, the above works do not consider the effect of the stochastic gradient noise which may provide a shelter for Byzantine attacks and increase the learning error. To address this issue, Khanduri et al. (2019); Wu et al. (2020); Karimireddy et al. (2021); Rammal et al. (2024); Guerraoui et al. (2024) propose to use the variance-reduction and momentum techniques to alleviate the effect of the stochastic gradient noise and enhance the Byzantine-robustness. Though these methods work well when the data distributions are the same over the workers, their performance degrades when the data distributions become heterogeneous (Li et al., 2019; Karimireddy et al., 2022). Therefore, Li et al. (2019) suggests using model aggregation rather than stochastic gradient aggregation to defend against model poisoning attacks in the heterogeneous case. Karimireddy et al. (2022); Peng et al. (2022); Allouah et al. (2023) propose to use the bucketing/resampling and nearest neighbor mixing techniques to reduce the heterogeneity of the messages, prior to aggregation.

Some other works focus on asynchronous learning (Yang and Li, 2023) or decentralized learning without a server (Peng et al., 2021; He et al., 2022; Wu et al., 2023), under model poisoning attacks. Nevertheless, we focus on synchronous distributed learning with a server in this paper.

There are also a large amount of papers focusing on data poisoning attacks (Sun et al., 2019; Bagdasaryan et al., 2020; Wang et al., 2020; Rosenfeld et al., 2020; Cinà et al., 2024). To defend against data poisoning attacks, the existing works use data sanitization to remove poisoned data (Steinhardt et al., 2017), and prune activation units that are inactive on clean data (Liu et al., 2018). For more defenses against data poisoning attacks, we refer the reader to the survey paper (Kairouz et al., 2021).

In practice, however, attacks may not necessarily behave as arbitrarily malicious as the above well-established works consider. Some weaker attacks models are structured; for example, Tavallali et al. (2022) considers the label poisoning attacks in which some workers mislabel their local data and compute the incorrect messages using those poisoned data. Specifically, Tolpegin et al. (2020); Lin et al. (2021); Jebreel and Domingo-Ferrer (2023); Jebreel et al. (2024) consider the case where some workers flip the labels of their local data from source classes to target classes. Notably, label poisoning is a kind of data poisoning but not necessarily the worst-case attack, since label poisoning attacks fabricate the local data, yet only on the label level.

It has been shown that the robust aggregators designed for model poisoning attacks can be applied to defend against the label poisoning attacks, as validated by Fang et al. (2020); Karimireddy et al. (2022); Gorbunov et al. (2022). There also exist some works designing new robust aggregators based on specific properties of label poisoning. For example, the work of Tavallali et al. (2022) proposes regularization-based defense to detect and exclude the samples with flipped labels in the training process. However, Tavallali et al. (2022) requires to access a clean validation set, which has privacy concerns in distributed learning. Another work named as LFighter (Jebreel et al., 2024) is the state-of-the-art defense for label poisoning attacks in federated learning. Jebreel et al. (2024) proposes to cluster the local gradients of all workers, identify the smaller and denser clusters as the potentially poisoned gradients, and discard them. The key idea of LFighter is that the difference between the stochastic gradients connected to the source and target output neurons of poisoned workers and regular workers becomes larger when the training process evolves. Therefore, we are able to identify the potentially poisoned stochastic gradients. However, LFighter only works well when data distributions at different workers are similar. If the heterogeneity of the distributed data is large, the performance of LFighter degrades, as we will show in Section 5.

Though the recent works of Karimireddy et al. (2022) and Farhadkhani et al. (2024) respectively prove the optimality of certain robust aggregators for model poisoning attacks and data poisoning attacks, they only consider the case that the poisoned workers can cause unbounded disturbances to the learning process. In contrast, we consider the case that the disturbances caused by the poisoned workers is bounded and prove the optimality of the mean aggregator for label poisoning attacks when the distributed data are sufficiently heterogeneous, and experimentally validate our theoretical findings.

The recent work of Shejwalkar et al. (2022), similar to our findings, reveals the robustness of the mean aggregator under poisoning attacks in production federated learning systems. Nevertheless, their study is restrictive in terms of the poisoning ratio (for example, less than 0.1% workers are poisoned while we can afford 10% in the numerical experiments) and lacks theoretical analysis. In contrast, we provide both theoretical analysis and experimental validations.

In conclusion, our work is the first one to investigate the robustness of the mean aggregator in distributed learning. It reveals an important fact that the robust aggregators cannot always outperform the mean aggregator under specific attacks, promoting us to rethink the application scenarios for the use of robust aggregators.

3. Problem Formulation

Consider a distributed learning system with one server and W workers. Denote the set of workers as $\mathcal{W}$ with $|\mathcal{W}| = W$, and the set of regular workers as $\mathcal{R}$ with $|\mathcal{R}| = R$. Note that the number and identities of the regular workers are unknown. Our goal is to solve the following distributed learning

problem defined over the regular workers in $\mathcal{R}$, at the presence of the set of poisoned workers $\mathcal{W}\backslash\mathcal{R}$:

$$\min_{x\in\mathbb{R}^D} f(x) \triangleq \frac{1}{R}\sum_{w\in\mathcal{R}} f_w(x), \tag{1}$$
$$\text{with } f_w(x) \triangleq \frac{1}{J}\sum_{j=1}^{J} f_{w,j}(x), \ \forall w \in \mathcal{R}.$$

Here, $x \in \mathbb{R}^D$ is the global model and $f_w(x)$ is the local cost of worker $w \in \mathcal{R}$ that averages the costs $f_{w,j}(x)$ of J samples. Without loss of generality, we assume that all workers have the same number of samples J.

We begin with characterizing the behaviors of the poisoned workers in $\mathcal{W} \setminus \mathcal{R}$. Different to the classical Byzantine attacks model that assumes some workers to disobey the algorithmic protocol and send arbitrarily malicious messages to the server (Lamport et al., 1982), here we assume the poisoned workers to: (i) have samples with poisoned labels; (ii) exactly follow the algorithmic protocol during the distributed learning process. The formal definition is given as follows.

Definition 1 (Label poisoning attacks) *In solving* (1)*, there exist a number of poisoned workers, whose local costs are in the same form as the regular workers but an arbitrary fraction of sample labels are poisoned. Nevertheless, these poisoned workers exactly follow the algorithmic protocol during the distributed learning process.*

We solve (1) with the distributed stochastic momentum algorithm, which includes the popular distributed gradient descent and distributed stochastic gradient descent algorithms as special cases. At each iteration, each worker randomly accesses one local sample to compute a local stochastic gradient, updates its local momentum, and sends the local momentum to the server. Then, the server aggregates the local momenta of all workers. However, as we have emphasized, the number and identities of the regular workers are unknown, such that the server cannot distinguish the true local momenta from the regular workers and the poisoned local momenta from the poisoned workers. We call the true and poisoned local momenta as messages, which the server must judiciously aggregate.

The distributed stochastic momentum algorithm works as follows. At iteration t, the server first broadcasts the global model x^t to all workers. Then, each worker $w \in \mathcal{W}$ selects a sample index i_w^t uniformly randomly from $\{1, \cdots, J\}$ and computes the corresponding stochastic gradient at the global model x^t. We denote the true stochastic gradient of regular worker $w \in \mathcal{R}$ as $\nabla f_{w,i_w^t}(x^t)$ and the poisoned stochastic gradient of poisoned worker $w \in \mathcal{W} \setminus \mathcal{R}$ as $\nabla \tilde{f}_{w,i_w^t}(x^t)$. Next, the workers update their local momenta and send to the server. For each regular worker $w \in \mathcal{R}$, its true local momentum is

$$m_w^t = (1-\alpha)m_w^{t-1} + \alpha\nabla f_{w,i_w^t}(x^t), \tag{2}$$

where $\alpha \in [0,1]$ is the momentum coefficient and m_w^{-1} is initialized as $\nabla f_{w,i_w^0}(x^0)$. For each poisoned worker $w \in \mathcal{W} \setminus \mathcal{R}$, its poisoned local momentum is

$$\tilde{m}_w^t = (1-\alpha)\tilde{m}_w^{t-1} + \alpha\nabla \tilde{f}_{w,i_w^t}(x^t), \tag{3}$$

where $\tilde{m}_w^{-1}$ is initialized as $\nabla \tilde{f}_{w,i_w^0}(x_w^0)$. For notational convenience, we denote the message sent by worker w, no matter true or poisoned, as

$$\hat{m}_w^t = \begin{cases} m_w^t, & w \in \mathcal{R}, \\ \tilde{m}_w^t, & w \in \mathcal{W} \setminus \mathcal{R}. \end{cases} \tag{4}$$

Upon receiving all messages $\{\hat{m}_w^t : w \in \mathcal{W}\}$, the server may choose to aggregate them with a robust aggregator $\mathrm{RAgg}(\cdot)$ and then move a step along the negative direction, as

$$x^{t+1} = x^t - \gamma \cdot \mathrm{RAgg}(\{\hat{m}_w^t : w \in \mathcal{W}\}), \tag{5}$$

where $\gamma > 0$ is the step size. State-of-the-art robust aggregators include trimmed mean (TriMean) (Chen et al., 2017), FABA (Xia et al., 2019), centered clipping (CC) (Karimireddy et al., 2021), to name a few.

In this paper, we argue that the mean aggregator $\mathrm{Mean}(\cdot)$, which is often viewed as vulnerable, is more robust than the state-of-the-art robust aggregators under label poisoning attacks. With the mean aggregator, the update is

$$x^{t+1} = x^t - \gamma \cdot \mathrm{Mean}(\{\hat{m}_w^t : w \in \mathcal{W}\}), \tag{6}$$

where

$$\mathrm{Mean}(\{\hat{m}_w^t : w \in \mathcal{W}\}) \triangleq \frac{1}{W} \sum_{w \in \mathcal{W}} \hat{m}_w^t. \tag{7}$$

We summarize the distributed stochastic momentum algorithm with different aggregators in Algorithm 1. Note that when the momentum coefficient $\alpha = 1$, it reduces to the popular distributed stochastic gradient descent algorithm. Further, if at each iteration, each worker accesses all J samples to compute the local gradient other than stochastic gradient, the algorithm becomes distributed gradient descent.

Algorithm 1

Input: Initializations $x^0 \in \mathbb{R}^D$, $m_w^{-1} = \nabla f_{w,i_w^0}(x^0)$ if $w \in \mathcal{R}$, $\tilde{m}_w^{-1} = \nabla \tilde{f}_{w,i_w^0}(x_w^0)$ if $w \in \mathcal{W} \setminus \mathcal{R}$, with i_w^0 being uniformly randomly sampled from $\{1, \cdots, J\}$; step size γ; momentum coefficient α; number of overall iterations T.

1: **for** $t = 0, 1, \cdots, T-1$ **do**
2: Server broadcasts x^t to all workers.
3: Regular worker $w \in \mathcal{R}$ uniformly randomly samples i_w^t from $\{1, \cdots, J\}$, computes $\nabla f_{w,i_w^t}(x^t)$, updates $m_w^t = (1-\alpha)m_w^{t-1} + \alpha \nabla f_{w,i_w^t}(x^t)$, and sends $\hat{m}_w^t = m_w^t$ to server.
4: Poisoned worker $w \in \mathcal{W} \setminus \mathcal{R}$ uniformly randomly samples i_w^t from $\{1, \cdots, J\}$, computes $\nabla \tilde{f}_{w,i_w^t}(x^t)$, updates $\tilde{m}_w^t = (1-\alpha)\tilde{m}_w^{t-1} + \alpha \nabla \tilde{f}_{w,i_w^t}(x^t)$, and sends $\hat{m}_w^t = \tilde{m}_w^t$ to server.
5: Server receives $\{\hat{m}_w^t\}_{w \in \mathcal{W}}$ from all workers and updates x^{t+1} according to (5) or (6).
6: **end for**

4. Convergence Analysis

In this section, we analyze the learning errors of Algorithm 1 with different aggregators under label poisoning attacks. We make the following assumptions. For regular worker $w \in \mathcal{R}$, we respectively denote the true gradients of the local cost and the j-th sample cost as $\nabla f_w(\cdot)$ and $\nabla f_{w,j}(\cdot)$. For poisoned worker $w \in \mathcal{W} \setminus \mathcal{R}$, we respectively denote the poisoned gradients of the local cost and the j-th sample cost as $\nabla \tilde{f}_w(\cdot)$ and $\nabla \tilde{f}_{w,j}(\cdot)$.

Assumption 1 (Lower boundedness) *The global cost $f(\cdot)$ is lower bounded by f^*, i.e., $f(x) \geq f^*$.*

Assumption 2 (Lipschitz continuous gradients) *The global cost $f(\cdot)$ has L-Lipschitz continuous gradients. That is, for any $x, y \in \mathbb{R}^D$, it holds that*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|. \tag{8}$$

Assumption 3 (Bounded heterogeneity) *For any $x \in \mathbb{R}^D$, the maximum distance between the local gradients of any regular worker $w \in \mathcal{R}$ and the global gradient is upper-bounded by ξ, i.e.,*

$$\max_{w \in \mathcal{R}} \|\nabla f_w(x) - \nabla f(x)\| \leq \xi. \tag{9}$$

Assumption 4 (Bounded inner variance) *For any $x \in \mathbb{R}^D$, the variance of the local stochastic gradients of any worker $w \in \mathcal{W}$ with respect to the local gradient is upper-bounded by σ^2, i.e.,*

$$\mathbb{E}_{i_w}\|\nabla f_{w,i_w}(x) - \nabla f_w(x)\|^2 \leq \sigma^2, \quad \forall w \in \mathcal{R}, \tag{10}$$

$$\mathbb{E}_{i_w}\|\nabla \tilde{f}_{w,i_w}(x) - \nabla \tilde{f}_w(x)\|^2 \leq \sigma^2, \quad \forall w \in \mathcal{W} \setminus \mathcal{R}, \tag{11}$$

where i_w denotes a sample index uniformly randomly selected from $\{1, \cdots, J\}$.

Assumptions 1, 2, 3 and 4 are all common in the analysis of distributed first-order stochastic algorithms. In particular, Assumption 3 characterizes the heterogeneity of the distributed data across the regular workers; larger ξ means higher heterogeneity. Assumption 4 is made for both regular and poisoned workers. This is reasonable since the poisoned workers only poison their local labels while keep local features clean, such that the variances of poisoned local stochastic gradients do not drastically change. We will validate Assumption 4 with numerical experiments in Appendix G.

Assumption 5 (Bounded disturbances of poisoned local gradients) *For any $x \in \mathbb{R}^D$, the maximum distance between the poisoned local gradients of poisoned workers $w \in \mathcal{W} \setminus \mathcal{R}$ and the global gradient is upper-bounded by A, i.e.,*

$$\max_{w \in \mathcal{W} \setminus \mathcal{R}} \|\nabla \tilde{f}_w(x) - \nabla f(x)\| \leq A. \tag{12}$$

Assumption 5 bounds the disturbances caused by the poisoned workers. This assumption does not hold for the worst-case Byzantine attacks model, where the disturbances caused by the Byzantine workers can be arbitrary. However, under label poisoning attacks, we prove that this assumption holds for distributed softmax regression as follows. We will also demonstrate with numerical experiments that this assumption holds naturally in training neural networks.

4.1 Justification of Assumption 5

Example: Distributed softmax regression under label poisoning attacks. Distributed softmax regression is common for classification tasks, where the local cost of worker $w \in \mathcal{R}$ is in the form of

$$f_w(x) = \frac{1}{J}\sum_{j=1}^{J} f_{w,j}(x), \text{ where } f_{w,j}(x) = -\sum_{k=1}^{K} \mathbf{1}\{b^{(w,j)} = k\} \log \frac{\exp(x_k^T a^{(w,j)})}{\sum_{l=1}^{K} \exp(x_l^T a^{(w,j)})}. \quad (13)$$

In (13), K stands for the number of classes; $(a^{(w,j)}, b^{(w,j)})$ represents the j-th sample of worker w with $a^{(w,j)} \in \mathbb{R}^d$ and $b^{(w,j)} \in \mathbb{R}$ being the feature and the label, respectively; $\mathbf{1}\{b^{(w,j)} = k\}$ is the indicator function that outputs 1 if $b^{(w,j)} = k$ and 0 otherwise; $x_k \triangleq [x]_{kd:(k+1)d} \in \mathbb{R}^d$ is the k-th block of x.

Note that for poisoned worker $w \in \mathcal{W} \setminus \mathcal{R}$, the labels are possibly changed from $b^{(w,j)}$ to $\tilde{b}^{(w,j)}$ for all $j \in \{1, \cdots, J\}$. Therefore, the local cost of worker $w \in \mathcal{W} \setminus \mathcal{R}$ is in the form of

$$\tilde{f}_w(x) = \frac{1}{J}\sum_{j=1}^{J} \tilde{f}_{w,j}(x), \text{ where } \tilde{f}_{w,j}(x) = -\sum_{k=1}^{K} \mathbf{1}\{\tilde{b}^{(w,j)} = k\} \log \frac{\exp(x_k^T a^{(w,j)})}{\sum_{l=1}^{K} \exp(x_l^T a^{(w,j)})}. \quad (14)$$

It is straightforward to verify that the global cost $f(x)$ with the local costs $f_w(x)$ in (13) satisfies Assumptions 1 and 2. Since the gradients of local costs $f_w(x)$ in (13) are bounded (see Lemma 11 in Appendix A), the global cost $f(x)$ satisfies Assumptions 3 and ξ refers to the heterogeneity of the local costs $f_w(x)$. Further, since the gradients of the regular sample costs $f_{w,j}(\cdot)$ and the poisoned sample costs $\tilde{f}_{w,j}(\cdot)$ are all bounded (see Lemma 11 in Appendix A), the local cost of any worker $w \in \mathcal{W}$ satisfies Assumption 4. Next, we show that Assumption 5 also holds.

Lemma 2 *Consider the distributed softmax regression problem where the local costs of the workers are in the forms of* (13) *and* (14)*. Therein, the poisoned workers are under label poisoning attacks, with arbitrary fractions of sample labels being poisoned. If $a^{(w,j)}$ is entry-wise non-negative for all $w \in \mathcal{W}$ and all $j \in \{1, \cdots, J\}$, then Assumption 5 is satisfied with*

$$A \leq 2\sqrt{K} \max_{w \in \mathcal{W}} \|\frac{1}{J}\sum_{j=1}^{J} a^{(w,j)}\|. \quad (15)$$

Proof See Appendix A.2. ■

Lemma 2 explicitly gives the upper bound of the smallest possible A in Assumption 5. Observe that the non-negativity assumption of $a^{(w,j)}$ naturally holds; for example, in image classification tasks, each entry of the feature stands for a pixel value. For other tasks, we can shift the features to meet this requirement.

Relation between Assumptions 3 and 5. Interestingly, the constants ξ and A in Assumptions 3 and 5 are tightly related. Similar to Lemma 2 that gives the upper bound of the smallest possible A in Assumption 5, for the distributed softmax regression problem, we can give the upper bound of the smallest possible ξ in Assumption 3 as follows.

Lemma 3 *Consider the distributed softmax regression problem where the local costs of the regular workers are in the form of* (13)*. If $a^{(w,j)}$ is entry-wise non-negative for all $w \in \mathcal{R}$ and all $j \in \{1, \cdots, J\}$, then Assumption 3 is satisfied with*

$$\xi \leq 2\sqrt{K} \max_{w \in \mathcal{R}} \|\frac{1}{J} \sum_{j=1}^{J} a^{(w,j)}\|. \tag{16}$$

Proof See Appendix A.3. ■

In particular, in the sufficiently heterogeneous case that each regular worker only has the samples from one class and the samples from one class only belong to one regular worker, the constant ξ is in the same order of $\max_{w \in \mathcal{R}} \|\frac{1}{J} \sum_{j=1}^{J} a^{(w,j)}\|$ (see Lemma 12 in Appendix A). Further, if the feature norms of the regular and poisoned workers have similar magnitudes, which generally holds in practice, then $\max_{w \in \mathcal{R}} \|\frac{1}{J} \sum_{j=1}^{J} a^{(w,j)}\|$ is in the same order as $\max_{w \in \mathcal{W}} \|\frac{1}{J} \sum_{j=1}^{J} a^{(w,j)}\|$. Hence, we can conclude that $A = O(\xi)$ when the distributed data are sufficiently heterogeneous. This conclusion will be useful in our ensuing analysis.

4.2 Main Results

To analyze the learning errors of Algorithm 1 with robust aggregators, we need to characterize the approximation abilities of the robust aggregators, namely, how close their outputs are to the average of the messages from the regular workers. This gives rise to the definition of ρ-robust aggregator (Wu et al., 2023; Dong et al., 2024).

Definition 4 (ρ-robust aggregator) *Consider any W messages $y_1, y_2, \cdots, y_W \in \mathbb{R}^D$, among which R messages are from regular workers $w \in \mathcal{R}$. An aggregator $RAgg(\cdot)$ is said to be a ρ-robust aggregator if there exists a contraction constant $\rho \geq 0$ such that*

$$\|RAgg(\{y_1, \cdots, y_W\}) - \bar{y}\| \leq \rho \cdot \max_{w \in \mathcal{R}} \|y_w - \bar{y}\|, \tag{17}$$

where $\bar{y} = \frac{1}{R} \sum_{w \in \mathcal{R}} y_w$ is the average message of the regular workers.

From Definition 4, a small contraction constant ρ means that the output of the robust aggregator is close to the average of the messages from the regular workers. The error is proportional to the heterogeneity of the messages from the regular workers, characterized by $\max_{w \in \mathcal{R}} \|y_w - \bar{y}\|$.

However, since a robust aggregator cannot distinguish the regular and poisoned workers, ρ is unable to be arbitrarily close to 0. Additionally, when the messages from the poisoned workers are majority, there is no guarantee to satisfy Definition 4. Therefore, we have the following lemma.

Lemma 5 *Denote $\delta \triangleq 1 - \frac{R}{W}$ as the fraction of the poisoned workers. Then a ρ-robust aggregator exists only if $\delta < \frac{1}{2}$ and $\rho \geq \min\{\frac{\delta}{1-2\delta}, 1\}$.*

Proof See Appendix B.1. ■

We prove that several state-of-the-art robust aggregators, such as TriMean (Chen et al., 2017), CC (Karimireddy et al., 2021) and FABA (Xia et al., 2019), all satisfy Definition 4 when the fraction of poisoned workers is below their respective thresholds. Their corresponding contraction constants ρ are given in Appendix B.

Remark 6 *Our definition is similar to (f,κ)-robustness in Allouah et al. (2023), while our heterogeneity measure is $\max_{w\in\mathcal{R}}\|y_w-\bar{y}\|$ instead of $\frac{1}{R}\sum_{w\in\mathcal{R}}\|y_w-\bar{y}\|^2$. Due to the fact $\max_{w\in\mathcal{R}}\|y_w-\bar{y}\|^2\leq\sum_{w\in\mathcal{R}}\|y_w-\bar{y}\|^2$, our definition implies (f,κ)-robustness in Allouah et al. (2023). Further, according to Propositions 8 and 9 in Allouah et al. (2023), our definition also implies (f,λ)-resilient averaging and $(\delta_{\max},c)$-ARAgg in Farhadkhani et al. (2022) and Karimireddy et al. (2022), respectively. The lower bound of ρ is determined by the fraction of the poisoned workers δ. A smaller δ leads to a smaller lower bound of ρ, which aligns with our intuition. Similar results can be found in Farhadkhani et al. (2022) and Allouah et al. (2023).*

Thanks to the contraction property in Definition 4, we can prove that the learning error of Algorithm 1 with a ρ-robust aggregator is bounded under label poisoning attacks.

Theorem 7 *Consider Algorithm 1 with a ρ-robust aggregator RAgg$(\cdot)$ to solve* (1) *and suppose that Assumptions 1, 2, 3, and 4 hold. Under label poisoning attacks where the fraction of poisoned workers is $\delta\in[0,\frac{1}{2})$, if the step size is $\gamma=\min\left\{O\left(\sqrt{\frac{LF^0+\rho^2\sigma^2}{TL^2\sigma^2(\rho^2+1)}}\right),\frac{1}{8L}\right\}$, the momentum coefficient is $\alpha=8L\gamma$, then we have*

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(x^t)\|^2 \\ =&O\left(\rho^2\xi^2+\sqrt{\frac{(LF^0+\rho^2\sigma^2)(\rho^2+1)\sigma^2}{T}}+\frac{LF^0+(\rho^2+1)\sigma^2+\rho^2\xi^2}{T}\right).\end{aligned}\tag{18}$$

where the expectation is taken over the algorithm's randomness and $F^0\triangleq f(x^0)-f^*$.

Proof See Appendix C. ■

Interestingly, we are also able to prove that under label poisoning attacks, Algorithm 1 with the mean aggregator has a bounded learning error.

Theorem 8 *Consider Algorithm 1 with the mean aggregator Mean$(\cdot)$ to solve* (1) *and suppose that Assumptions 1, 2, 4, and 5 hold. Under label poisoning attacks where the fraction of poisoned workers is $\delta\in[0,1)$, if the step size is $\gamma=\min\left\{O\left(\sqrt{\frac{LF^0+\delta^2\sigma^2}{TL^2\sigma^2(\delta^2+1)}}\right),\frac{1}{8L}\right\}$, the momentum coefficient is $\alpha=8L\gamma$, then we have*

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(x^t)\|^2 \\ =&O\left(\delta^2A^2+\sqrt{\frac{(LF^0+\delta^2\sigma^2)(\delta^2+1)\sigma^2}{T}}+\frac{LF^0+(\delta^2+1)\sigma^2+\delta^2A^2}{T}\right).\end{aligned}\tag{19}$$

where the expectation is taken over the algorithm's randomness and $F^0\triangleq f(x^0)-f^*$.

Proof See Appendix D. ■

Theorems 7 and 8 demonstrate that Algorithm 1 with both ρ-robust aggregators and the mean aggregator can sublinearly converge to neighborhoods of a first-order stationary point of (1), while the non-vanishing learning errors are $O(\rho^2\xi^2)$ for ρ-robust aggregators and $O(\delta^2A^2)$ for the mean aggregator. It is worth noting that the constants within $O(\rho^2\xi^2)$ for the ρ-robust aggregator and $O(\delta^2A^2)$ for the mean aggregator are the same (see Theorem 17 in Appendix C and Theorem 18 in Appendix D). We omit these constants here to present the theoretical results concisely. Observe that without the $O(\rho^2\xi^2)$ and $O(\delta^2A^2)$ terms, the $O(\frac{1}{\sqrt{T}})$ convergence rates are optimal for first-order nonconvex stochastic optimization algorithms (Arjevani et al., 2023).

Before comparing the learning errors in Theorems 7 and 8, we first give the lower bound of the learning error for a class of identity-invariant algorithms.

Theorem 9 *Under label poisoning attacks with $\delta = 1 - \frac{R}{W}$ fraction of poisoned workers, consider any algorithm running for T iterations and generating x^t at iteration t. Suppose that the output of the algorithm is invariant with respect to the identities of the workers. Then, there exist R regular local functions $\{f_w(x) : w \in \mathcal{R}\}$ and $W - R$ poisoned local functions $\{\tilde{f}_w(x) : w \in \mathcal{W} \setminus \mathcal{R}\}$, which are all composed of J sample costs $f_{w,j}(x)$ or $\tilde{f}_{w,j}(x)$ respectively, satisfying Assumptions 1, 2, 3, 4, and 5 such that the iterates $\{x^t : t = 1, \cdots, T\}$ of the algorithm satisfy*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(x^t)\|^2 = \Omega(\delta^2 \min\{A^2, \xi^2\}). \tag{20}$$

where the expectation is taken over the algorithm's randomness.

Proof See Appendix E. ■

The identity-invariant property in Theorem 9 means that, given any W local costs of which R are regular, the output of the algorithm is invariant with respect to which local costs are regular or poisoned. This property excludes the "omniscient" algorithms that know the identities of the workers and thus can exclude the local costs of the poisoned workers. In fact, all practical algorithms are identity-invariant, including Algorithm 1 with any ρ-robust aggregator or the mean aggregator. In Karimireddy et al. (2022) and Allouah et al. (2023), the authors implicitly confine their analyses to the algorithms that share the same identity-invariant property as stated in Theorem 9, resulting in comparable lower bounds to ours. On the other hand, in Karimireddy et al. (2021), the authors establish a lower bound for the algorithms with a stronger iteration-wise permutation-invariant property, which excludes our distributed stochastic momentum algorithm.

For any identity-invariant algorithm, it is impossible to fully eliminate the effect of malicious attacks, which results in the non-vanishing learning error as demonstrated in Theorem 9. When the disturbances caused by the poisoned workers are small such that $A < \xi$, the lower bound in (20) becomes $\Omega(\delta^2A^2)$, matching the learning error of Algorithm 1 with the mean aggregator in (19). It implies that the learning error of the mean aggregator is order-optimal in the presence of small disturbances. On the other hand, when the disturbances caused by the poisoned worker are large such that $A \geq \xi$, the lower bound in (20) becomes $\Omega(\delta^2\xi^2)$. In Table 1, we compare the learning errors for different aggregators, given large heterogeneity such that A is at most the same order as ξ (which holds when the distributed data are sufficiently heterogeneous, as we have discussed in Section 4.1).

Aggregator	Learning error
TriMean	$O(\frac{\delta^2\xi^2}{(1-2\delta)^2})$
CC	$O(\delta\xi^2)$
FABA	$O(\frac{\delta^2\xi^2}{(1-3\delta)^2})$
Mean	$O(\delta^2\xi^2)$
Lower bound	$\Omega(\delta^2\xi^2)$

Table 1: Learning errors of Algorithm 1 with TriMean, CC, FABA and the mean aggregator when the heterogeneity of distributed data is sufficiently large such that A is in the same order as ξ. The lower bound of the learning error is also given.

According to Table 1, we know that the learning errors of TriMean, FABA and the mean aggregator all match the lower bound in terms of the order, when δ is small. However, the learning errors of TriMean and FABA explode when δ approaches $\frac{1}{2}$ and $\frac{1}{3}$, respectively, while the mean aggregator is insensitive. Therefore, the learning error of the mean aggregator is order-optimal regardless of the fraction of poisoned workers. In addition, the learning error of the mean aggregator is smaller than that of CC by a magnitude of δ.

Remark 10 *Note that our backbone algorithm, distributed stochastic momentum, degenerates to distributed stochastic gradient descent (when $\alpha = 1$) and distributed gradient descent (when $\alpha = 1$ and $\sigma = 0$). Following our analysis, the non-vanishing learning errors of distributed stochastic gradient descent, when using a ρ-robust aggregators and the mean aggregator, are $O(\rho^2\sigma^2 + \rho^2\xi^2)$ and $O(\delta^2\sigma^2 + \delta^2A^2)$ respectively (substituting $\alpha = 1$ to the proofs of Theorems 7 and 8 yields these results). When the data heterogeneity term ξ^2 and the disturbance term A^2 both dominate the variance of the stochastic gradients σ^2, our conclusions made in this paper remain valid. Further letting $\sigma = 0$, the non-vanishing learning errors of distributed gradient descent, when using a ρ-robust aggregator and the mean aggregator, are $O(\rho^2\xi^2)$ and $O(\delta^2A^2)$ respectively. This way, we can obtain the same results in Table 1, as shown in Peng et al. (2024).*

5. Numerical Experiments

In this section, we conduct numerical experiments to validate our theoretical findings and demonstrate the performance of Algorithm 1 with the mean and robust aggregators under label poisoning attacks. The code is available at `https://github.com/pengj97/LPA`.

5.1 Experimental Settings

Datasets and partitions. In the numerical experiments, we investigate a convex problem of softmax regression on the MNIST dataset. We also consider two non-convex problems. The first one is to train two-layer perceptrons, in which each layer has 50 neurons and the activation function is ReLU,

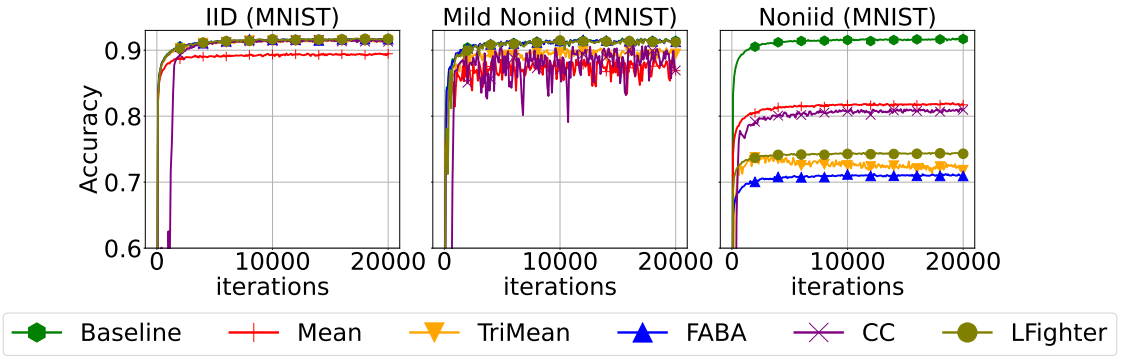

Figure 1: Accuracies of softmax regression on the MNIST dataset under static label flipping attacks.

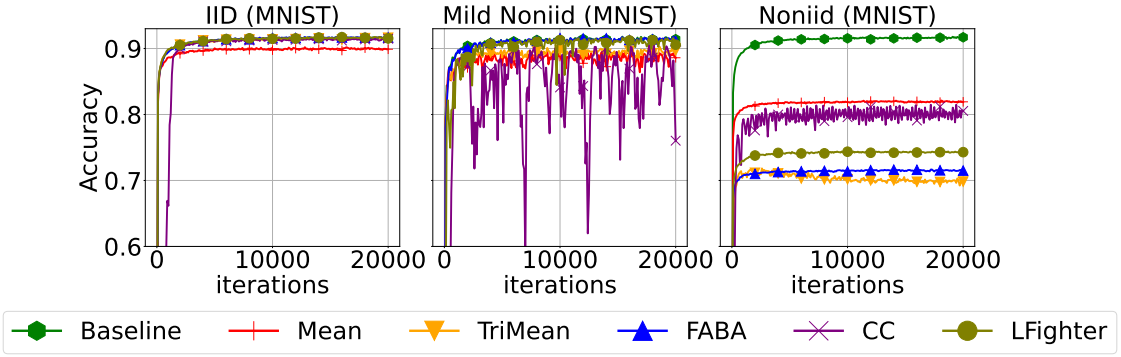

Figure 2: Accuracies of softmax regression on the MNIST dataset under dynamic label flipping attacks.

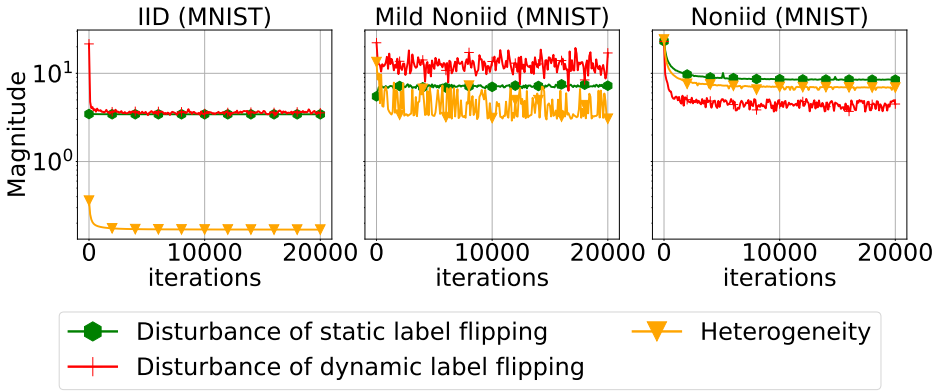

Figure 3: Heterogeneity of regular local gradients (the smallest ξ satisfying Assumption 3) and disturbance of poisoned local gradients (the smallest A satisfying Assumption 5) in softmax regression on the MNIST dataset, under static label flipping and dynamic label flipping attacks.

Layer Name	Layer size
Convolution + ReLU	$3 \times 3 \times 16$
Max pool	2×2
Convolution + ReLU	$3 \times 3 \times 32$
Max pool	2×2
Fully connected + ReLU	128
Softmax	10

Table 2: Architecture of the convolutional neural network trained on the CIFAR10 dataset.

on the MNIST dataset. The second one is to train convolutional neural networks, whose architecture is given in Table 2, on the CIFAR10 dataset[1].

We setup $W = 10$ workers where $R = 9$ workers are regular and the remaining one is poisoned. The impact of different fractions of poisoned workers is demonstrated in Appendix H. We consider three data distributions: i.i.d., mild non-i.i.d. and non-i.i.d. cases. In the i.i.d. case, we uniformly randomly divide the training data among all workers. In the mild non-i.i.d. case, we divide the training data using the Dirichlet distribution with hyper-parameter $\beta = 1$ by default (Hsu et al., 2019). In the non-i.i.d. case, we assign each class of the training data to one worker.

Label poisoning attacks. We investigate two types of label poisoning attacks: static label flipping where the poisoned worker flips label b to $9 - b$ with b ranging from 0 to 9, and dynamic label flipping where the poisoned worker flips label b to the least probable label with respect to the global model x^t (Shejwalkar et al., 2022).

Aggregators to compare. We are going to compare the mean aggregator with several representative ρ-robust aggregators, including TriMean, FABA, CC, and LFighter. The baseline is the mean aggregator without attacks. The step size is $\gamma = 0.01$ and the momentum coefficient is $\alpha = 0.1$.

5.2 Convex Case

Classification accuracy. We consider softmax regression on the MNIST dataset. The classification accuracies under static label flipping and dynamic label flipping attacks are shown in Figure 1 and Figure 2, respectively. In the i.i.d. case, all methods perform well and close to the baseline, but the mean aggregator has an apparently lower classification accuracy. In the mild non-i.i.d. case, FABA and LFighter are the best among all aggregators and the other aggregators have similar performance. In the non-i.i.d. case, since the heterogeneity is large, all aggregators are tremendously affected by the label poisoning attacks, and have gaps to the baseline in terms of classification accuracy. Notably, the mean aggregator performs the best among all aggregators in this case, which validates our theoretical results.

Heterogeneity of regular local gradients and disturbance of poisoned local gradients. To further validate the reasonableness of Assumptions 3 and 5, as well as the correctness of our theoretical

1. Although the ReLU function is non-smooth, the training process rarely reaches the non-smooth point. Therefore, the results on the ReLU function are similar to those on a smooth function (which can be obtained by modifying the ReLU function and has Lipschitz continuous gradients).

results in Section 4.1, we compute the smallest ξ and A that satisfy Assumptions 3 and 5 for the softmax regression problem. As shown in Figure 3, the disturbances of the poisoned local gradients, namely A, are bounded under both static label flipping and dynamic label flipping attacks, which corroborates the theoretical results in Lemma 2. From i.i.d., mild non-i.i.d. to the non-i.i.d. case, the heterogeneity of the regular local gradients characterized by ξ increases. Particularly, in the non-i.i.d. case, ξ is close to A under both static label flipping and dynamic label flipping attacks, which aligns our discussions below Lemma 3. Recall Table 1 that shows when the heterogeneity is in the same order of the disturbances caused by the label poisoning attacks, the learning error of the mean aggregator is order-optimal. This explains the results in Figures 1 and 2.

5.3 Nonconvex Case

Classification accuracy. Next, we train two-layer perceptrons on the MNIST dataset and convolutional neural networks on the CIFAR10 dataset under static label flipping and dynamic label flipping attacks, as depicted in Figures 4 and 5. In the i.i.d. case, all methods have good performance and are close to the baseline, except for CC that performs worse than the other aggregators on the CIFAR10 dataset under dynamic label flipping attacks. In the mild non-i.i.d. case and on the MNIST dataset, all methods perform well and are close to the baseline. On the CIFAR10 dataset, Mean, FABA and LFighter are the best and close to the baseline, CC and TriMean are worse, while TriMean is the worst and with an obvious gap under dynamic label flipping attacks. In the non-i.i.d. case, all methods are affected by the attacks and cannot reach the same classification accuracy of the baseline, but the mean aggregator is still the best. CC, FABA and LFighter are worse and TriMean fails.

Heterogeneity of regular local gradients and disturbance of poisoned local gradients. We also calculate the smallest values of ξ and A satisfying Assumptions 3 and 5, respectively. As shown in Figure 6, the disturbance of poisoned local gradients measured by A are bounded on the MNIST and CIFAR10 datasets under both static label flipping and dynamic label flipping attacks. From i.i.d., mild non-i.i.d. to the non-i.i.d. case, the heterogeneity of regular local gradients ξ is increasing. In the non-i.i.d. case, ξ is close to A.

5.4 Impacts of Heterogeneity and Attack Strengths

To further show the impacts of heterogeneity of data distributions and strengths of label poisoning attacks, we compute classification accuracies of the trained two-layer perceptrons on the MNIST dataset, varying the data distributions and the levels of label poisoning attacks. We employ the Dirichlet distribution by varying the hyper-parameter $\beta = \{5, 1, 0.1, 0.05, 0.03, 0.01\}$ to simulate various heterogeneity of data distributions, in which a smaller β corresponds to larger heterogeneity (Hsu et al., 2019). In addition, we let the poisoned worker apply static label flipping attacks by flipping labels with probability $p = \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ to simulate different attack strengths. A larger flipping probability indicates stronger attacks.

We present the best performance among all aggregators, and mark the corresponding best aggregator in Figure 7. More details are in Table 3 of Appendix F. The mean aggregator outperforms the robust aggregators when the heterogeneity is large. For example, the mean aggregator exhibits superior performance when $\beta = 0.01$ and the flipping probability $p = \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, as well as when $\beta = 0.03$ and $p = \{0.0, 0.2, 0.4, 0.6, 0.8\}$. Furthermore, fixing the flipping probability p, when the hyper-parameter β becomes smaller which means that the heterogeneity becomes larger, the mean aggregator gradually surpasses the robust aggregators. Fixing the hyper-parameter

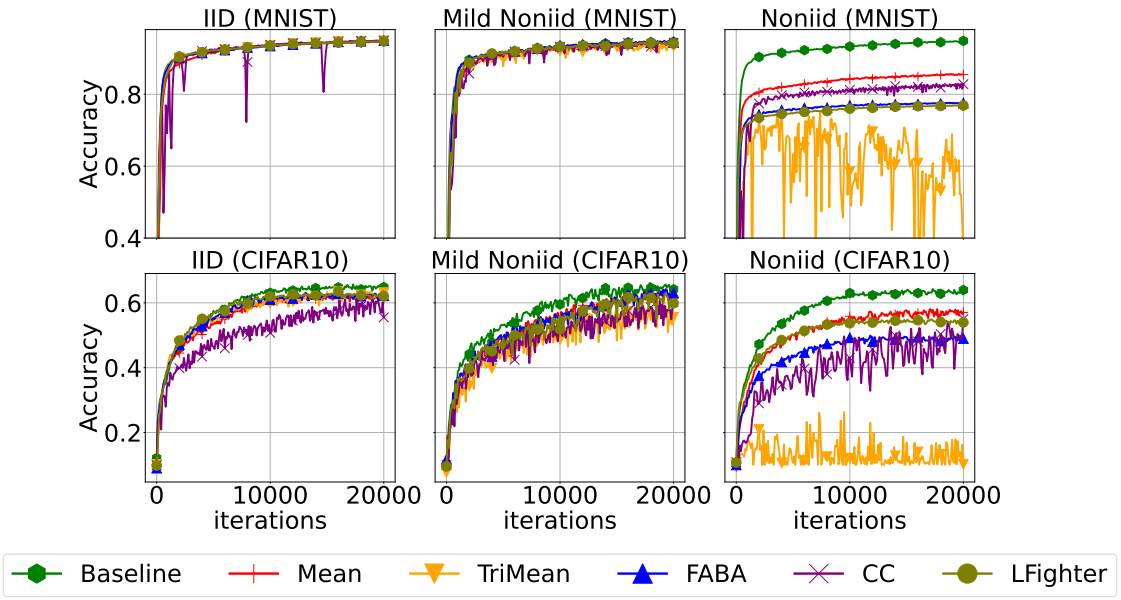

Figure 4: Accuracies of two-layer perceptrons on the MNIST dataset and convolutional neural networks on the CIFAR10 dataset under static label flipping attacks.

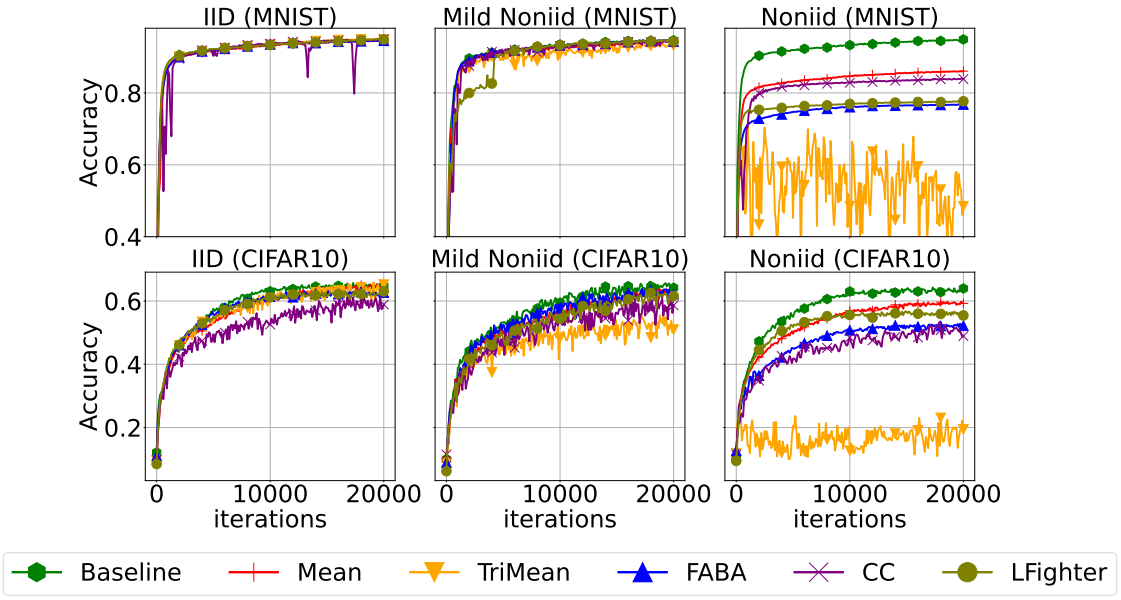

Figure 5: Accuracies of two-layer perceptrons on the MNIST dataset and convolutional neural networks on the CIFAR10 dataset under dynamic label flipping attacks.

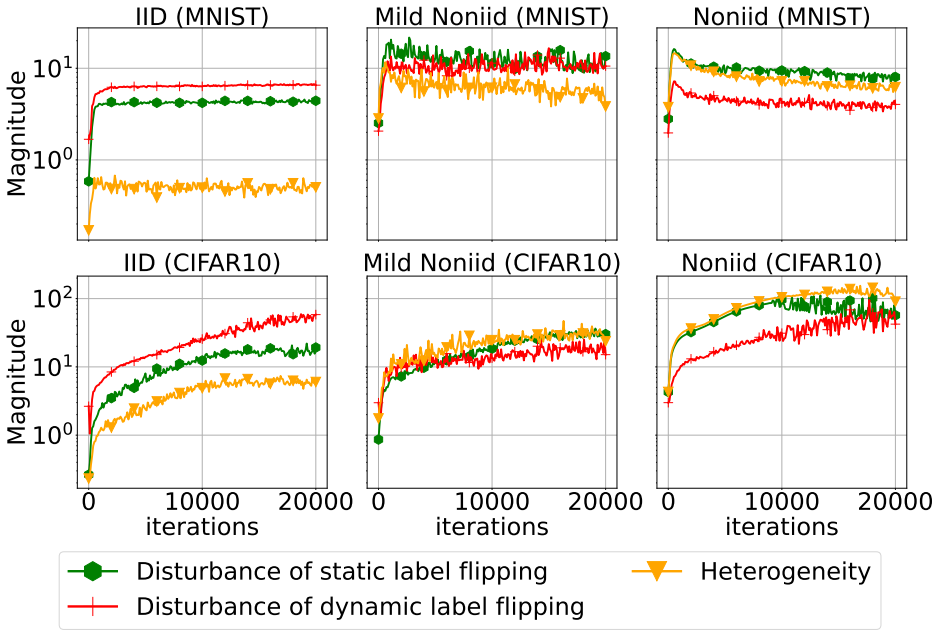

Figure 6: Heterogeneity of regular local gradients (the smallest ξ satisfying Assumption 3) and disturbance of poisoned local gradients (the smallest A satisfying Assumption 5) in training two-layer perceptrons on the MNIST dataset and training convolutional neural networks on the CIFAR10 dataset, under static label flipping and dynamic label flipping attacks.

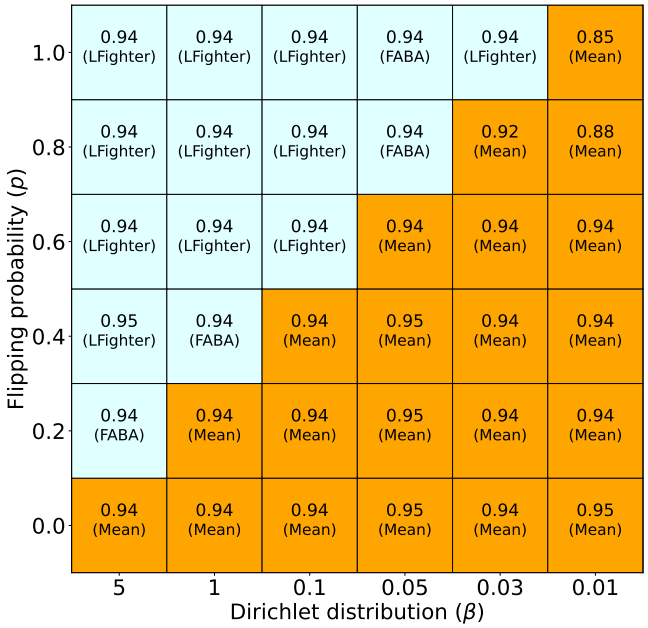

Figure 7: Best accuracies of trained two-layer perceptrons by all aggregators on the MNIST dataset under static label flipping attacks. Each block is associated with a hyper-parameter β that characterizes the heterogeneity and the flipping probability p that characterizes the attack strength. For each block, the best accuracy and the corresponding aggregator is marked. Orange means that the mean aggregator is the best.

β, when the flipping probability p becomes smaller which means that the attack strength becomes smaller, the mean aggregator gradually surpasses the robust aggregators. According to the above observations, we recommend to apply the mean aggregator if the distributed data are sufficiently heterogeneous, or the disturbance caused by label poisoning attacks is comparable to the heterogeneity of regular local gradients.

6. Conclusions

We studied the distributed learning problem subject to the label poisoning attacks. We theoretically proved that when the distributed data are sufficiently heterogeneous, the learning error of the mean aggregator is order-optimal. Further corroborated by numerical experiments, our work revealed an important fact that state-of-the-art robust aggregators cannot always outperform the mean aggregator, if the attacks are confined to label poisoning. We expect that this fact can motivate readers to revisit which application scenarios are proper for using robust aggregators. In our future work, we will extend the analysis to the more challenging decentralized learning problem.

Acknowledgments

Qing Ling (corresponding author) is supported in part by National Key R&D Program of China grant 2024YFA1014002, National Natural Science Foundation of China grant 62373388, Guangdong Basic and Applied Basic Research Foundation grant 2023B1515040025, and Guangdong Provincial Key Laboratory of Mathematical Foundations for Artificial Intelligence grant 2023B1212010001.

Appendix A. Analysis of Distributed Softmax Regression

In this section, we analyze the property of distributed softmax regression where the local cost of worker $w \in \mathcal{W}$ is in the forms of (13) and (14). We first show that the gradients of the sample costs $f_{w,j}(x)$, $\tilde{f}_{w,j}(x)$ and the local cost $f_w(x)$, $\tilde{f}_w(x)$ are bounded. Then, we prove Lemma 2 that provides the constant A in Assumption 5. Last, we prove Lemma 3 that gives the constant ξ in Assumption 3, and further demonstrate that when the distributed data across the regular workers are sufficiently heterogeneous, the constant ξ is in the same order of $\max_{w\in\mathcal{R}} \|\frac{1}{J}\sum_{j=1}^{J} a^{(w,j)}\|$.

A.1 Bounded Gradients of Local Costs

Lemma 11 *Consider the distributed softmax regression problem where the local cost of worker $w \in \mathcal{W}$ is in the forms of* (13) *and* (14)*. Then, the gradients of the sample costs are bounded by the norms of the corresponding sample features, i.e.,*

$$\|\nabla f_{w,j}(x)\| \le 2\|a^{(w,j)}\|, \quad \forall w \in \mathcal{R},\ \forall j \in [1,\cdots,J], \tag{21}$$

$$\|\nabla \tilde{f}_{w,j}(x)\| \le 2\|a^{(w,j)}\|, \quad \forall w \in \mathcal{W}\setminus\mathcal{R},\ \forall j \in [1,\cdots,J]. \tag{22}$$

and the gradient of the local cost is bounded by the maximum norm of the local features, i.e.,

$$\|\nabla f_{w}(x)\| \le 2\max_{j\in[J]} \|a^{(w,j)}\|, \quad \forall w \in \mathcal{R}, \tag{23}$$

$$\|\nabla \tilde{f}_{w}(x)\| \le 2\max_{j\in[J]} \|a^{(w,j)}\|, \quad \forall w \in \mathcal{W}\setminus\mathcal{R}. \tag{24}$$

Moreover, if $a^{(w,j)}$ is entry-wise non-negative for all $w \in \mathcal{W}$ and all $j \in \{1,\cdots,J\}$, we have

$$\|\nabla f_{w}(x)\| \le \sqrt{K}\|\frac{1}{J}\sum_{j=1}^{J} a^{(w,j)}\|, \quad \forall w \in \mathcal{R}, \tag{25}$$

$$\|\nabla \tilde{f}_{w}(x)\| \le \sqrt{K}\|\frac{1}{J}\sum_{j=1}^{J} a^{(w,j)}\|, \quad \forall w \in \mathcal{W}\setminus\mathcal{R}. \tag{26}$$

Proof For notational convenience, we denote the local cost of worker $w \in \mathcal{W}$ as

$$\hat{f}_w(x) = \frac{1}{J}\sum_{j=1}^{J} \hat{f}_{w,j}(x), \text{ where } \hat{f}_{w,j}(x) = -\sum_{k=1}^{K} \mathbf{1}\{\hat{b}^{(w,j)} = k\} \log \frac{\exp(x_k^T a^{(w,j)})}{\sum_{l=1}^{K}\exp(x_l^T a^{(w,j)})}, \tag{27}$$

and

$$\hat{b}^{(w,j)} = \begin{cases} b^{(w,j)}, & w \in \mathcal{R}, \\ \tilde{b}^{(w,j)}, & w \in \mathcal{W}\setminus\mathcal{R}. \end{cases} \tag{28}$$

We first prove that the sample gradient $\nabla \hat{f}_{w,j}(\cdot)$ is bounded. For the k-th block of $\nabla \hat{f}_{w,j}(\cdot)$, we have

$$\nabla_{x_k} \hat{f}_{w,j}(x) = -a^{(w,j)}\big(\mathbf{1}\{\hat{b}^{(w,j)} = k\} - \frac{\exp(x_k^T a^{(w,j)})}{\sum_{l=1}^{K}\exp(x_l^T a^{(w,j)})}\big). \tag{29}$$

Therefore, the entire sample gradient $\nabla \hat{f}_{w,j}(\cdot)$ satisfies

$$\begin{aligned}\|\nabla \hat{f}_{w,j}(x)\|^2 &= \sum_{k=1}^{K} \|\nabla_{x_k} \hat{f}_{w,j}(x)\|^2 \\ &= \sum_{k=1}^{K} \|a^{(w,j)}\big(\mathbf{1}\{\hat{b}^{(w,j)}=k\} - \frac{\exp(x_k^T a^{(w,j)})}{\sum_{l=1}^{K}\exp(x_l^T a^{(w,j)})}\big)\|^2 \\ &= \sum_{k=1}^{K} \big(\mathbf{1}\{\hat{b}^{(w,j)}=k\} - \frac{\exp(x_k^T a^{(w,j)})}{\sum_{l=1}^{K}\exp(x_l^T a^{(w,j)})}\big)^2 \|a^{(w,j)}\|^2.\end{aligned} \tag{30}$$

Since

$$\begin{aligned}&\sum_{k=1}^{K}\Big(1\{\hat{b}^{(w,j)}=k\} - \frac{\exp(x_k^T a^{(w,j)})}{\sum_{l=1}^{K}\exp(x_l^T a^{(w,j)})}\Big)^2 \\ \leq &\Big(\sum_{k=1}^{K}|1\{\hat{b}^{(w,j)}=k\} - \frac{\exp(x_k^T a^{(w,j)})}{\sum_{l=1}^{K}\exp(x_l^T a^{(w,j)})}|\Big)^2 \leq 4,\end{aligned} \tag{31}$$

we have

$$\|\nabla \hat{f}_{w,j}(x)\|^2 \leq \sum_{k=1}^{K}\big(\mathbf{1}\{\hat{b}^{(w,j)}=k\} - \frac{\exp(x_k^T a^{(w,j)})}{\sum_{l=1}^{K}\exp(x_l^T a^{(w,j)})}\big)^2\|a^{(w,j)}\|^2 \leq \big(2\|a^{(w,j)}\|\big)^2, \tag{32}$$

and

$$\|\nabla \hat{f}_{w,j}(x)\| \leq 2\|a^{(w,j)}\|. \tag{33}$$

which shows the upper bound for the gradient of the sample cost function $\hat{f}_{w,j}(\cdot)$.

Now we prove that the local gradient $\nabla \hat{f}_w(\cdot)$ is also bounded. By $\nabla \hat{f}_w(x) = \frac{1}{J}\sum_{j=1}^{J}\nabla \hat{f}_{w,j}(x)$, we have

$$\|\nabla \hat{f}_w(x)\| = \|\frac{1}{J}\sum_{j=1}^{J}\nabla \hat{f}_{w,j}(x)\| \leq \frac{1}{J}\sum_{j=1}^{J}\|\nabla \hat{f}_{w,j}(x)\| \leq 2\max_{j\in[J]}\|a^{(w,j)}\|. \tag{34}$$

Finally, we refine the bound of the local gradient $\nabla \hat{f}_w(\cdot)$ under the non-negativity assumption. Starting from the second equality in (30) and using $\nabla_{x_k}\hat{f}_w(x) = \frac{1}{J}\sum_{j=1}^{J}\nabla_{x_k}\hat{f}_{w,j}(x)$, we have

$$\begin{aligned}\|\nabla \hat{f}_w(x)\|^2 &= \sum_{k=1}^{K}\|\frac{1}{J}\sum_{j=1}^{J}a^{(w,j)}\big(\mathbf{1}\{b^{(w,j)}=k\} - \frac{\exp(x_k^T a^{(w,j)})}{\sum_{l=1}^{K}\exp(x_l^T a^{(w,j)})}\big)\|^2 \\ &\leq \sum_{k=1}^{K}\left\|\max_{j\in[J]}\big|\mathbf{1}\{b^{(w,j)}=k\} - \frac{\exp(x_k^T a^{(w,j)})}{\sum_{l=1}^{K}\exp(x_l^T a^{(w,j)})}\big|\cdot\frac{1}{J}\sum_{j=1}^{J}a^{(w,j)}\right\|^2,\end{aligned} \tag{35}$$

where the inequality is due to $\left|\sum_{j\in[J]} c_j x_j\right| \leq \max_{j\in[J]} |c_j| \cdot \sum_{j\in[J]} x_j$ for any sequence $\{c_j\}_{j\in[J]}$ and any positive sequence $\{x_j\}_{j\in[J]}$. Since $\max_{j\in[J]} \left|\mathbf{1}\{b^{(w,j)}=k\} - \frac{\exp(x_k^T a^{(w,j)})}{\sum_{l=1}^K \exp(x_l^T a^{(w,j)})}\right| \leq 1$, we reach our conclusion of

$$\|\nabla \hat{f}_w(x)\|^2 \leq K \|\frac{1}{J}\sum_{j=1}^J a^{(w,j)}\|^2, \tag{36}$$

which completes the proof. ■

A.2 Proof of Lemma 2

Proof Note that

$$\max_{w\in\mathcal{W}\setminus\mathcal{R}} \|\nabla \tilde{f}_w(x) - \nabla f(x)\|^2 \leq 2 \max_{w\in\mathcal{W}\setminus\mathcal{R}} \|\nabla \tilde{f}_w(x)\|^2 + 2\|\nabla f(x)\|^2. \tag{37}$$

From Lemma 11, we know the first term at the right-hand side of (37) can be upper-bounded as

$$\max_{w\in\mathcal{W}\setminus\mathcal{R}} \|\nabla \tilde{f}_w(x)\|^2 \leq K \max_{w\in\mathcal{W}\setminus\mathcal{R}} \|\frac{1}{J}\sum_{j=1}^J a^{(w,j)}\|^2. \tag{38}$$

For the second term at the right-hand side of (37), applying the inequality of $\|\frac{1}{R}\sum_{w\in\mathcal{R}} \nabla f_w\|^2 \leq \frac{1}{R}\sum_{w\in\mathcal{R}} \|\nabla f_w\|^2$ gives

$$\|\nabla f(x)\|^2 \leq \frac{1}{R}\sum_{w\in\mathcal{R}} \|\nabla f_w(x)\|^2 \leq \max_{w\in\mathcal{R}} \|\nabla f_w(x)\|^2 \leq K \max_{w\in\mathcal{R}} \|\frac{1}{J}\sum_{j=1}^J a^{(w,j)}\|^2, \tag{39}$$

where the last inequality similarly comes from the assumption that $a^{(w,j)}$ is entry-wise non-negative for all $w \in \mathcal{W}$ and all $j \in \{1, \cdots, J\}$.

Combining (38) and (39), we have

$$\max_{w\in\mathcal{W}\setminus\mathcal{R}} \|\nabla \tilde{f}_w(x) - \nabla f(x)\|^2 \leq 4K \max_{w\in\mathcal{W}} \|\frac{1}{J}\sum_{j=1}^J a^{(w,j)}\|^2, \tag{40}$$

or equivalently

$$\max_{w\in\mathcal{W}\setminus\mathcal{R}} \|\nabla \tilde{f}_w(x) - \nabla f(x)\| \leq 2\sqrt{K} \max_{w\in\mathcal{W}} \|\frac{1}{J}\sum_{j=1}^J a^{(w,j)}\|, \tag{41}$$

which is exactly (8) with $A \leq 2\sqrt{K}\max_{w\in\mathcal{W}} \|\frac{1}{J}\sum_{j=1}^J a^{(w,j)}\|$. ■

A.3 Proofs of Lemma 3 and Its Extension

The following lemma combines Lemma 3 and its extension in the sufficiently heterogeneous case.

Lemma 12 *Consider the distributed softmax regression problem where the local costs of the regular workers are in the form of* (13)*. If $a^{(w,j)}$ is entry-wise non-negative for all $w \in \mathcal{R}$ and all $j \in \{1, \cdots, J\}$, then Assumption 3 is satisfied with*

$$\xi \leq 2\sqrt{K} \max_{w\in\mathcal{R}} \|\frac{1}{J}\sum_{j=1}^{J} a^{(w,j)}\|. \tag{42}$$

Further, if any regular worker $w \in \mathcal{R}$ only has the samples from one class and the samples from one class only belongs to one regular worker (i.e., $b^{(w,j)} = b^{(w',j')}$ if and only if $w = w'$, for all $w, w' \in \mathcal{R}$ and all $j, j' \in \{1, \cdots, J\}$), we have

$$\xi = \Theta\big(\max_{w\in\mathcal{R}} \|\frac{1}{J}\sum_{j=1}^{J} a^{(w,j)}\|\big). \tag{43}$$

Proof Note that for any regular worker $w \in \mathcal{R}$, it holds

$$\begin{aligned}\|\nabla f_w(x) - \nabla f(x)\| &= \|\big(1 - \frac{1}{R}\big)\nabla f_w(x) - \sum_{w'\in\mathcal{R}, w'\neq w} \frac{1}{R}\nabla f_{w'}(x)\| \\ &\leq \big(1 - \frac{1}{R}\big)\|\nabla f_w(x)\| + \frac{1}{R}\sum_{w'\in\mathcal{R}, w'\neq w} \|\nabla f_{w'}(x)\| \\ &\leq 2\max_{w'\in\mathcal{R}} \|\nabla f_{w'}(x)\|.\end{aligned} \tag{44}$$

$$\max_{w\in\mathcal{R}} \|\nabla f_w(x) - \nabla f(x)\| \leq 2\sqrt{K} \max_{w\in\mathcal{R}} \|\frac{1}{J}\sum_{j=1}^{J} a^{(w,j)}\|, \tag{45}$$

and thus Assumption 3 is satisfied with

$$\xi \leq 2\sqrt{K} \max_{w\in\mathcal{R}} \|\frac{1}{J}\sum_{j=1}^{J} a^{(w,j)}\|. \tag{46}$$

Next, we prove the lower bound of ξ when any regular worker $w \in \mathcal{R}$ only has the samples from one class and the samples from one class only belongs to one regular worker. For any regular worker $w' \in \mathcal{R}$ and any $x \in \mathbb{R}^D$, Assumption 3 gives that

$$\begin{aligned}\xi^2 \geq \|\nabla f_{w'}(x) - \nabla f(x)\|^2 = \sum_{k=1}^{K} \|\frac{1}{J}\sum_{j=1}^{J} a^{(w',j)}(\mathbf{1}\{b^{(w',j)} = k\} - \frac{\exp(x_k^T a^{(w',j)})}{\sum_{l=1}^{K}\exp(x_l^T a^{(w',j)})}) \\ - \frac{1}{R}\sum_{w\in\mathcal{R}} \frac{1}{J}\sum_{j=1}^{J} a^{(w,j)}(\mathbf{1}\{b^{(w,j)} = k\} - \frac{\exp(x_k^T a^{(w,j)})}{\sum_{l=1}^{K}\exp(x_l^T a^{(w,j)})})\|^2.\end{aligned} \tag{47}$$

Letting $[x_k]_i = 0$ for any $i \in \{1, \cdots, D\}$ and $k \in \{1, \cdots, K\}$, it holds that

$$\frac{\exp(x_k^T a^{(w,j)})}{\sum_{l=1}^{K} \exp(x_l^T a^{(w,j)})} = \frac{1}{K}, \quad \forall w \in \mathcal{R},\ j \in \{1, \cdots, J\}, k \in \{1, \cdots, K\}.$$

Given the heterogeneous label distribution, there exists $k' \in \{1, \cdots, K\}$, such that $b^{(w',j)} = k' \neq b^{(w,j)}$ for all $w \neq w'$ and $j \in \{1, \cdots, J\}$. Specifically, taking one of the summands in (47) with $k = k'$, we obtain

$$\begin{aligned}
\xi^2 \geq & \| \big(1 - \frac{1}{R}\big) \frac{1}{J} \sum_{j=1}^{J} a^{(w',j)} (1 - \frac{1}{K}) \\
& - \frac{1}{R} \sum_{w \in \mathcal{R}, w \neq w'} \frac{1}{J} \sum_{j=1}^{J} a^{(w,j)} (\mathbf{1}\{b^{(w,j)} = k'\} - \frac{\exp(x_{k'}^T a^{(w,j)})}{\sum_{l=1}^{K} \exp(x_l^T a^{(w,j)})}) \|^2 \\
= & \| \big(1 - \frac{1}{R}\big)\big(1 - \frac{1}{K}\big) \frac{1}{J} \sum_{j=1}^{J} a^{(w',j)} + \frac{1}{RK} \sum_{w \in \mathcal{R}, w \neq w'} \frac{1}{J} \sum_{j=1}^{J} a^{(w,j)} \|^2 \\
\geq & \| \big(1 - \frac{1}{R}\big)\big(1 - \frac{1}{K}\big) \frac{1}{J} \sum_{j=1}^{J} a^{(w',j)} \|^2,
\end{aligned} \tag{48}$$

where the last inequality is due to the fact that each term in the summation is non-negative. Note that $w' \in \mathcal{R}$ is arbitrary, which results in

$$\xi \geq \big(1 - \frac{1}{R}\big)\big(1 - \frac{1}{K}\big) \max_{w \in \mathcal{R}} \| \frac{1}{J} \sum_{j=1}^{J} a^{(w,j)} \|. \tag{49}$$

Combining (46) and (49), we have

$$\xi = \Theta\big(\max_{w \in \mathcal{R}} \| \frac{1}{J} \sum_{j=1}^{J} a^{(w,j)} \| \big), \tag{50}$$

which completes the proof. ■

Appendix B. Analysis of ρ-Robust Aggregators

In this section, we prove Lemma 5 that explores the approximation abilities of ρ-robust aggregators, and show that the state-of-the-art robust aggregators, including TriMean, CC, and FABA, are all ρ-robust aggregators when the fraction of poisoned workers is below their respective thresholds.

B.1 Proof of Lemma 5

An equivalent statement of Lemma 5 is shown below.

Lemma 13 *Denote $\delta \triangleq 1 - \frac{R}{W}$ as the fraction of the poisoned workers. For any aggregator RAgg, if $\delta \geq \frac{1}{2}$ or $\rho < \min\{\frac{\delta}{1-2\delta}, 1\}$, then there exist W messages $y_1, y_2, \ldots, y_W \in \mathbb{R}^D$ such that*

$$\|RAgg(\{y_1, \ldots, y_W\}) - \bar{y}\| > \rho \cdot \max_{w \in \mathcal{R}} \|y_w - \bar{y}\|. \tag{51}$$

where $\bar{y} = \frac{1}{R}\sum_{w \in \mathcal{R}} y_w$ is the average message of the regular workers.

Proof Without loss of generality, consider $D = 1$ and $\mathcal{R} = \{1, 2, \ldots, R\}$. For the high-dimensional cases, setting all entries but one as zero degenerates to the scalar case. The key idea of our proof is to find two sets of W messages that are the same but the messages from regular workers are different. Therefore, any aggregator cannot distinguish between them and will yield the same output. Elaborately designing these two sets, we shall guarantee that at least one of them satisfies (51).

We first consider the case that $\delta \geq \frac{1}{2}$, or equivalently, $2R \leq W$. In this case, if we can find W messages $y_1, \cdots, y_W \in \mathbb{R}^D$ such that $\|\text{RAgg}(\{y_1, \ldots, y_W\}) - \bar{y}\| \neq 0$ and $\max_{w \in \mathcal{R}} \|y_w - \bar{y}\| = 0$, we have found the set of messages satisfying (51). The construction is as follows. Let $y_1 = \cdots = y_R = 0$, $y_{R+1} = \cdots y_{2R} = \rho + 1$ and $y_{2R+1} = \cdots = y_W = 0$. Then $\bar{y} = 0$ and $\max_{w \in \mathcal{R}} \|y_w - \bar{y}\| = 0$. If $\text{RAgg}(\{y_1, \ldots, y_W\}) \neq 0$, these W messages satisfy (51) and our task is fulfilled. Otherwise, we know that $\text{RAgg}(\{y_1, \ldots, y_W\}) = 0$. Rearranging these messages as $z_1 = \ldots z_R = \rho + 1$ and $z_{R+1} = \ldots = z_W = 0$, the aggregator outputs the same value $\text{RAgg}(\{z_1, \ldots, z_W\}) = 0$, while $\bar{z} = \rho + 1$ and $\max_{w \in \mathcal{R}} \|z_w - \bar{z}\| = 0$. Thus, $\{z_w, w \leq W\}$ is the set of messages satisfying (51).

Second, we consider the case that $\delta < \frac{1}{2}$ and $\rho < \min\{\frac{\delta}{1-2\delta}, 1\}$. In this case, if we can find W messages $y_1, \cdots, y_W$ such that $\|\text{RAgg}(\{y_1, \ldots, y_W\}) - \bar{y}\| = \frac{\delta}{1-\delta}$ and $\max_{w \in \mathcal{R}} \|y_w - \bar{y}\| = \frac{\max\{1-2\delta, \delta\}}{1-\delta}$, we have found the set of messages satisfying (51). Similar to the above construction yet with $2R > W$, we consider $y_1 = \ldots = y_R = 0$ and $y_{R+1} = \ldots = y_W = 1$. Accordingly, we have $\bar{y} = 0$ and $\max_{w \in \mathcal{R}} \|y_w - \bar{y}\| = 0$. If $\text{RAgg}(\{y_1, \ldots, y_W\}) \neq 0$, we have found W messages to satisfy (51). Otherwise, we know that $\text{RAgg}(\{y_1, \ldots, y_W\}) = 0$. Rearranging those messages as $z_1 = \ldots z_{W-R} = 1$ and $z_{W-R+1} = \ldots = z_W = 0$, the aggregator outputs the same value $\text{RAgg}(\{z_1, \ldots, z_W\}) = 0$, while $\bar{z} = \frac{\delta}{1-\delta}$ and $\max_{w \in \mathcal{R}} \|z_w - \bar{z}\| = \frac{\max\{1-2\delta, \delta\}}{1-\delta}$. In consequence, $\{z_w, w \leq W\}$ is the set of messages satisfying (51). ■

Our proof is motivated by those of Farhadkhani et al. (2022, Proposition 2) and Allouah et al. (2023, Proposition 6). However, their definitions of the robust aggregator are different to ours such that the proofs are different too.

B.2 TriMean

TriMean is an aggregator that discards the smallest $W - R$ elements and largest $W - R$ elements in each dimension. The aggregated output of TriMean in dimension d is given by

$$[\text{TriMean}(\{y_1, \ldots, y_W\})]_d = \frac{1}{2R - W} \sum_{w \in [\mathcal{U}]_d} [y_w]_d, \tag{52}$$

where $[\cdot]_d$ denotes the d-th coordinate of a vector, and $[\mathcal{U}]_d$ is the set of workers whose d-th elements are not filtered after removal. Below we show that TriMean is a ρ-robust aggregator if $\delta < \frac{1}{2}$.

Lemma 14 *Denote $\delta \triangleq 1 - \frac{R}{W}$ as the fraction of the poisoned workers. If $\delta < \frac{1}{2}$, TriMean is a ρ-robust aggregator with $\rho = \frac{3\delta}{1-2\delta}\min\{\sqrt{D}, \sqrt{R}\}$.*

Proof We first analyze the aggregated result in one dimension d and then extend it to all dimensions. Denote $[\mathcal{U}_R]_d \triangleq [\mathcal{U}]_d \cap \mathcal{R}$ and $[\mathcal{U}_P]_d \triangleq [\mathcal{U}]_d \cap (\mathcal{W} \setminus \mathcal{R})$ as the set of remaining regular workers and poisoned workers after removal, respectively.

If TriMean successfully removes all the poisoned workers in dimension d, such that $[\mathcal{U}_P]_d = \emptyset$ and $[\mathcal{U}]_d \subseteq \mathcal{R}$, it holds

$$\begin{aligned}\|[\text{TriMean}(\{y_1,\ldots,y_W\})]_d - [\bar{y}]_d\| =& \|\frac{1}{2R-W}\sum_{w\in[\mathcal{U}]_d}[y_w]_d - \frac{1}{R}\sum_{w\in\mathcal{R}}[y_w]_d\| \\ =& \|(\frac{1}{2R-W}-\frac{1}{R})\sum_{w\in\mathcal{R}}[y_w]_d - \frac{1}{2R-W}\sum_{w\in\mathcal{R}\setminus[\mathcal{U}]_d}[y_w]_d\| \\ =& \|\frac{W-R}{2R-W}[\bar{y}]_d - \frac{1}{2R-W}\sum_{w\in\mathcal{R}\setminus[\mathcal{U}]_d}[y_w]_d\| \\ \le& \frac{1}{2R-W}\sum_{w\in\mathcal{R}\setminus[\mathcal{U}]_d}\|[y_w]_d - [\bar{y}]_d\| \\ \le& \frac{\delta}{1-2\delta}\cdot\max_{w\in\mathcal{R}}\|[y_w]_d - [\bar{y}]_d\|.\end{aligned} \tag{53}$$

Otherwise, TriMean cannot remove all poisoned workers in dimension d, which means $[\mathcal{U}_P]_d \neq \emptyset$. Define

$$\bar{y}_{R_d} \triangleq \frac{1}{|[\mathcal{U}_R]_d|}\sum_{w\in[\mathcal{U}_R]_d}[y_w]_d, \quad \bar{y}_{P_d} \triangleq \frac{1}{|[\mathcal{U}_P]_d|}\sum_{w\in[\mathcal{U}_P]_d}[y_w]_d \tag{54}$$

as the average of elements in $[\mathcal{U}_R]_d$ and $[\mathcal{U}_P]_d$, respectively. Also denote $u_i \triangleq \frac{|[\mathcal{U}_P]_d|}{|[\mathcal{U}]_d|}$ as the fraction of poisoned workers that remains. With the above definitions, we have $u_i \le \frac{W-R}{2R-W} = \frac{\delta}{1-2\delta}$ and

$$\begin{aligned}&\|[\text{TriMean}(\{y_1,\ldots,y_W\})]_d - [\bar{y}]_d\| \\ =& \|u_i\cdot\bar{y}_{P_d} + (1-u_i)\cdot\bar{y}_{R_d} - [\bar{y}]_d\| \\ \le& u_i\|\bar{y}_{P_d} - [\bar{y}]_d\| + (1-u_i)\|\bar{y}_{R_d} - [\bar{y}]_d\|.\end{aligned} \tag{55}$$

For the first term at the right-hand side of (55), we have

$$\begin{aligned}u_i\|\bar{y}_{P_d} - [\bar{y}]_d\| =& u_i\|\frac{1}{|[\mathcal{U}_P]_d|}\sum_{w\in[\mathcal{U}_P]_d}[y_w]_d - [\bar{y}]_d\| \\ \le& \frac{\delta}{1-2\delta}\max_{w\in[\mathcal{U}_P]_d}\|[y_w]_d - [\bar{y}]_d\| \\ \le& \frac{\delta}{1-2\delta}\max_{w\in\mathcal{R}}\|[y_w]_d - [\bar{y}]_d\|,\end{aligned} \tag{56}$$

where the last inequality is due to the principle of filtering. Specifically, as the poisoned workers cannot own the $W-R$ largest values in dimension d, there exists a regular worker such that its d-th element is larger than $[y_w]_d$ for all $w \in [\mathcal{U}_P]_d$. Similarly, there exists a regular worker with the d-th

element smaller than all $[y_w]_d$'s. This observation guarantees the last inequality in (56). For the second term at the right-hand side of (55), we have

$$\begin{aligned}(1-u_i)\|\bar{y}_{R_d}-[\bar{y}]_d\| =&(1-u_i)\|(\frac{1}{|[\mathcal{U}_R]_d|}-\frac{1}{R})\sum_{w\in\mathcal{R}}[y_w]_d-\frac{1}{|[\mathcal{U}_R]_d|}\sum_{w\in\mathcal{R}\setminus[\mathcal{U}_R]_d}[y_w]_d\| \\ =&(1-u_i)\cdot\frac{1}{|[\mathcal{U}_R]_d|}\|\sum_{w\in\mathcal{R}\setminus[\mathcal{U}_R]_d}([y_w]_d-[\bar{y}]_d)\| \\ \leq&\frac{1}{2R-W}\sum_{w\in\mathcal{R}\setminus[\mathcal{U}_R]_d}\|[y_w]_d-[\bar{y}]_d\| \\ \leq&\frac{R-|[\mathcal{U}_R]_d|}{2R-W}\max_{w\in\mathcal{R}}\|[y_w]_d-[\bar{y}]_d\| \\ \leq&\frac{2\delta}{1-2\delta}\max_{w\in\mathcal{R}}\|[y_w]_d-[\bar{y}]_d\|,\end{aligned} \tag{57}$$

where the last inequality is due to $|[\mathcal{U}_R]_d| = (2R-W)-|[\mathcal{U}_P]_d| \geq (2R-W)-(W-R) = 3R-2W$.

Substituting (56) and (57) into (55), we have

$$\|[\text{TriMean}(\{y_1,\ldots,y_W\})]_d-[\bar{y}]_d\| \leq \frac{3\delta}{1-2\delta}\max_{w\in\mathcal{R}}\|[y_w]_d-[\bar{y}]_d\|. \tag{58}$$

Combining the first case (53) and the second case (58), we have

$$\|[\text{TriMean}(\{y_1,\ldots,y_W\})]_d-[\bar{y}]_d\| \leq \frac{3\delta}{1-2\delta}\max_{w\in\mathcal{R}}\|[y_w]_d-[\bar{y}]_d\|. \tag{59}$$

Next, we extend the scalar scenario to the vector scenario. Notice that

$$\sum_{d=1}^{D}\max_{w\in\mathcal{R}}\|[y_w]_d-[\bar{y}]_d\|^2 \leq \sum_{d=1}^{D}\max_{w\in\mathcal{R}}\|y_w-\bar{y}\|^2 \leq D\max_{w\in\mathcal{R}}\|y_w-\bar{y}\|^2. \tag{60}$$

On the other hand, we also have

$$\begin{aligned}&\sum_{d=1}^{D}\max_{w\in\mathcal{R}}\|[y_w]_d-[\bar{y}]_d\|^2 \\ \leq&\sum_{d=1}^{D}\sum_{w\in\mathcal{R}}\|[y_w]_d-[\bar{y}]_d\|^2 \\ \leq&\sum_{w\in\mathcal{R}}\sum_{d=1}^{D}\|[y_w]_d-[\bar{y}]_d\|^2 \\ \leq&R\max_{w\in\mathcal{R}}\|y_w-\bar{y}\|^2.\end{aligned} \tag{61}$$

Combining (60) and (61) gives

$$\sum_{d=1}^{D}\max_{w\in\mathcal{R}}\|[y_w]_d-[\bar{y}]_d\|^2 \leq \min\{D,R\}\max_{w\in\mathcal{R}}\|y_w-\bar{y}\|^2. \tag{62}$$

Substituting (62) into (59), we have

$$\begin{aligned}\|\text{TriMean}(\{y_1,\ldots,y_W\})-\bar{y}\|^2 =& \sum_{d=1}^{D}\|[\text{TriMean}(\{y_1,\ldots,y_W\})]_d-[\bar{y}]_d\|^2 \\ \leq& (\frac{3\delta}{1-2\delta})^2\sum_{d=1}^{D}\max_{w\in\mathcal{R}}\|[y_w]_d-[\bar{y}]_d\|^2 \\ \leq& (\frac{3\delta}{1-2\delta})^2\min\{D,R\}\max_{w\in\mathcal{R}}\|y_w-\bar{y}\|^2.\end{aligned} \tag{63}$$

Taking the square roots on both sides of (63), we have

$$\|\text{TriMean}(\{y_1,\ldots,y_W\})-\bar{y}\| \leq \frac{3\delta}{1-2\delta}\cdot\min\{\sqrt{D},\sqrt{R}\}\cdot\max_{w\in\mathcal{R}}\|y_w-\bar{y}\|, \tag{64}$$

which completes the proof. ■

B.3 CC

CC is an aggregator that iteratively clips the messages from workers. CC starts from some point v^0. At iteration i, the update rule of CC can be formulated as

$$v^{i+1}=v^i+\frac{1}{W}\sum_{w=1}^{W}\text{CLIP}(y_w-v^i,\tau), \tag{65}$$

where

$$\text{CLIP}(y_w-v^i,\tau)=\begin{cases}y_w-v^i, & \|y_w-v^i\|\leq\tau,\\ \dfrac{\tau}{\|y_w-v^i\|}(y_w-v^i), & \|y_w-v^i\|>\tau,\end{cases} \tag{66}$$

and $\tau\geq 0$ is the clipping threshold. After L iterations, CC outputs the last vector as

$$\text{CC}(\{y_1,\ldots,y_W\})=v^L. \tag{67}$$

Below we prove that with proper initialization and clipping threshold, one-step CC ($L=1$) is a ρ-robust aggregator if $\delta<\frac{1}{2}$.

Lemma 15 *Denote $\delta\triangleq 1-\frac{R}{W}$ as the fraction of the poisoned workers. If $\delta<\frac{1}{2}$, choosing the starting point v_0 satisfying $\|v_0-\bar{y}\|^2\leq\max_{w\in\mathcal{R}}\|y_w-\bar{y}\|^2$ and the clipping threshold $\tau=\sqrt{\frac{4(1-\delta)\max_{w\in\mathcal{R}}\|y_w-\bar{y}\|^2}{\delta}}$, one-step CC is a ρ-robust aggregator with $\rho=\sqrt{24\delta}$.*

Proof The output of one-step CC is

$$\text{CC}(\{y_1,\ldots,y_W\})=v^0+\frac{1}{W}\sum_{w=1}^{W}\text{CLIP}(y_w-v^0,\tau). \tag{68}$$

Note that if $\max_{w\in\mathcal{R}}\|y_w-\bar{y}\|=0$, we have $\tau=0$ and $v^0=\bar{y}$, which leads to $\text{CC}(\{y_1,\ldots,y_W\})=\bar{y}$. Therefore, we have

$$\|\text{CC}(\{y_1,\ldots,y_W\})-\bar{y}\|=\max_{w\in\mathcal{R}}\|y_w-\bar{y}\|\leq\sqrt{24\delta}\max_{w\in\mathcal{R}}\|y_w-\bar{y}\|. \tag{69}$$

Below we consider the case that $\max_{w\in\mathcal{R}}\|y_w-\bar{y}\|>0$, then by definition $\tau>0$.

Denoting $\hat{y}_w=v^0+\text{CLIP}(y_w-v^0,\tau)$ for any $w\in\{1,\ldots,W\}$, we have

$$\text{CC}(\{y_1,\ldots,y_W\})=\frac{1}{W}\sum_{w=1}^{W}\hat{y}_w. \tag{70}$$

According to (70), we have

$$\begin{aligned}\|\text{CC}(\{y_1,\ldots,y_W\})-\bar{y}\|^2=&\|\frac{1}{W}\sum_{w=1}^{W}\hat{y}_w-\bar{y}\|^2 \\ =&\|(1-\delta)\cdot(\frac{1}{R}\sum_{w\in\mathcal{R}}\hat{y}_w-\bar{y})+\delta\cdot\frac{1}{W-R}\sum_{w\in\mathcal{W}\setminus\mathcal{R}}(\hat{y}_w-\bar{y})\|^2 \\ \leq&2(1-\delta)^2\|\frac{1}{R}\sum_{w\in\mathcal{R}}(\hat{y}_w-y_w)\|^2+2\delta^2\|\frac{1}{W-R}\sum_{w\in\mathcal{W}\setminus\mathcal{R}}(\hat{y}_w-\bar{y})\|^2 \\ \leq&2(1-\delta)^2\frac{1}{R}\sum_{w\in\mathcal{R}}\|\hat{y}_w-y_w\|^2+2\delta^2\frac{1}{W-R}\sum_{w\in\mathcal{W}\setminus\mathcal{R}}\|\hat{y}_w-\bar{y}\|^2.\end{aligned} \tag{71}$$

where the last two inequalities are due to the Cauchy-Schwarz inequality.

For any $w\in\mathcal{R}$, the term $\|\hat{y}_w-y_w\|$ holds

$$\|\hat{y}_w-y_w\|=v^0-y_w+\text{CLIP}(y_w-v^0,\tau). \tag{72}$$

If the regular message y_w is not clipped, meaning that $\text{CLIP}(y_w-v^0,\tau)=y_w-v^0$, we have

$$\|\hat{y}_w-y_w\|=0. \tag{73}$$

Otherwise, we have

$$\|\hat{y}_w-y_w\|=\|v^0-y_w-\frac{\tau}{\|y_w-v^l\|}(v^0-y_w)\|=\|v^0-y_w\|-\tau. \tag{74}$$

Since

$$\|v^0-y_w\|-\tau\leq\frac{\|v^0-y_w\|^2}{\tau}\leq\frac{2\|v^0-\bar{y}\|^2+2\|y_w-\bar{y}\|^2}{\tau}\leq\frac{4\max_{w\in\mathcal{R}}\|y_w-\bar{y}\|^2}{\tau}, \tag{75}$$

where the first inequality is due to $a-b\leq\frac{a^2}{b}$ that holds for any $a\geq0,b>0$, and the last inequality is due to $\|v_0-\bar{y}\|^2\leq\max_{w\in\mathcal{R}}\|y_w-\bar{y}\|^2$, we have

$$\|\hat{y}_w-y_w\|\leq\frac{4\max_{w\in\mathcal{R}}\|y_w-\bar{y}\|^2}{\tau}. \tag{76}$$

Combining (73) and (76), we have

$$\|\hat{y}_w - y_w\| \leq \frac{4\max_{w\in\mathcal{R}}\|y_w - \bar{y}\|^2}{\tau}. \tag{77}$$

For any $w \in \mathcal{W} \setminus \mathcal{R}$, the term $\|\hat{y}_w - \bar{y}\|^2$ holds

$$\|\hat{y}_w - \bar{y}\|^2 \leq 2\|\hat{y}_w - v^0\|^2 + 2\|v^0 - \bar{y}\|^2 \leq 2\tau^2 + 2\max_{w\in\mathcal{R}}\|y_w - \bar{y}\|^2, \tag{78}$$

where the last inequality is due to $\|\hat{y}_w - v^0\|^2 = \|\text{CLIP}(y_w - v^0, \tau)\|^2 \leq \tau^2$ and $\|v_0 - \bar{y}\|^2 \leq \max_{w\in\mathcal{R}}\|y_w - \bar{y}\|^2$.

Substituting (77) and (78) into (71), we have

$$\begin{aligned}
&\|\text{CC}(\{y_1, \ldots, y_W\}) - \bar{y}\|^2 \\
\leq& 2(1-\delta)^2\left(\frac{4\max_{w\in\mathcal{R}}\|y_w - \bar{y}\|^2}{\tau}\right)^2 + 4\delta^2\tau^2 + 4\delta^2\max_{w\in\mathcal{R}}\|y_w - \bar{y}\|^2 \\
\leq& 24\delta(1-\delta)\max_{w\in\mathcal{R}}\|y_w - \bar{y}\|^2 + 4\delta^2\max_{w\in\mathcal{R}}\|y_w - \bar{y}\|^2 \\
\leq& 24\delta\max_{w\in\mathcal{R}}\|y_w - \bar{y}\|^2,
\end{aligned} \tag{79}$$

where the second inequality is due to $\tau = \sqrt{\frac{4(1-\delta)\max_{w\in\mathcal{R}}\|y_w - \bar{y}\|^2}{\delta}}$.

Therefore, we have

$$\|\text{CC}(\{y_1, \ldots, y_W\}) - \bar{y}\| \leq \sqrt{24\delta}\max_{w\in\mathcal{R}}\|y_w - \bar{y}\|. \tag{80}$$

Combining (69) and (80), we have that one-step CC is a ρ-robust aggregator with $\rho = \sqrt{24\delta}$. This completes the proof. ■

B.4 FABA

FABA is an aggregator that iteratively discards a possible outlier and averages the messages that remain after $W - R$ iterations. To be more concrete, denote $\mathcal{U}^{(i)}$ as the set of workers that are not discarded at the i-th iteration. Initialized with $\mathcal{U}^{(0)} = \{1, \ldots, W\}$, at iteration i, FABA computes the average of the messages from $\mathcal{U}^{(i)}$ and discards the worker whose message is farthest from that average to form $\mathcal{U}^{(i+1)}$. After $W - R$ iterations, FABA obtains $\mathcal{U}^{(W-R)}$ with R workers, and then outputs

$$\text{FABA}(\{y_1, \ldots, y_W\}) = \frac{1}{R}\sum_{w\in\mathcal{U}^{(W-R)}} y_w. \tag{81}$$

Below we prove that FABA is a ρ-robust aggregator if $\delta < \frac{1}{3}$.

Lemma 16 *Denote $\delta \triangleq 1 - \frac{R}{W}$ as the fraction of poisoned workers. If $\delta < \frac{1}{3}$, FABA is a ρ-robust aggregator with $\rho = \frac{2\delta}{1-3\delta}$.*

Proof For notational convenience, denote $\mathcal{R}^{(i)} \triangleq \mathcal{U}^{(i)} \cap \mathcal{R}$ and $\mathcal{P}^{(i)} \triangleq \mathcal{U}^{(i)} \cap (\mathcal{W} \setminus \mathcal{R})$ as the sets of the regular workers and the poisoned workers in $\mathcal{U}^{(i)}$, respectively. Further denote three different averages

$$\bar{y}_{\mathcal{U}^{(i)}} \triangleq \frac{1}{|\mathcal{U}^{(i)}|} \sum_{w \in \mathcal{U}^{(i)}} y_w, \quad \bar{y}_{\mathcal{R}^{(i)}} \triangleq \frac{1}{|\mathcal{R}^{(i)}|} \sum_{w \in \mathcal{R}^{(i)}} y_w, \quad \bar{y}_{\mathcal{P}^{(i)}} \triangleq \frac{1}{|\mathcal{P}^{(i)}|} \sum_{w \in \mathcal{P}^{(i)}} y_w \tag{82}$$

over $\mathcal{U}^{(i)}$, $\mathcal{R}^{(i)}$ and $\mathcal{P}^{(i)}$, respectively. Then

$$\bar{y}_{\mathcal{U}^{(i)}} = (1 - u_i) \cdot \bar{y}_{\mathcal{R}^{(i)}} + u_i \cdot \bar{y}_{\mathcal{P}^{(i)}}, \tag{83}$$

and our goal is to bound $\|\bar{y}_{\mathcal{U}^{(P)}} - \bar{y}\|$ by $\max_{w \in \mathcal{R}} \|y_w - \bar{y}\|$.

Denote $u_i \triangleq \frac{|\mathcal{P}^{(i)}|}{|\mathcal{U}^{(i)}|}$ as the fraction of the poisoned workers in $\mathcal{U}^{(i)}$. From $\delta < \frac{1}{3}$, $u_i \leq \frac{W-R}{R} < \frac{1}{2}$ for any $i \in \{0, \ldots, W - R\}$. We claim that a regular worker is filtered out at iteration i only if

$$\|\bar{y}_{\mathcal{R}^{(i)}} - \bar{y}_{\mathcal{P}^{(i)}}\| \leq \frac{1}{1 - 2u_i} \max_{w \in \mathcal{R}} \|y_w - \bar{y}_{\mathcal{R}^{(i)}}\|. \tag{84}$$

This is because if $\|\bar{y}_{\mathcal{R}^{(i)}} - \bar{y}_{\mathcal{P}^{(i)}}\| > \frac{1}{1-2u_i} \max_{w \in \mathcal{R}} \|y_w - \bar{y}_{\mathcal{R}^{(i)}}\|$, then for any $w \in \mathcal{R}$, we have

$$\begin{aligned}
\|y_w - \bar{y}_{\mathcal{U}^{(i)}}\| \leq & \|y_w - \bar{y}_{\mathcal{R}^{(i)}}\| + \|\bar{y}_{\mathcal{R}^{(i)}} - \bar{y}_{\mathcal{U}^{(i)}}\| \\
= & \|y_w - \bar{y}_{\mathcal{R}^{(i)}}\| + \frac{u_i}{1 - u_i} \|\bar{y}_{\mathcal{P}^{(i)}} - \bar{y}_{\mathcal{U}^{(i)}}\| \\
\leq & \max_{w \in \mathcal{R}} \|y_w - \bar{y}_{\mathcal{R}^{(i)}}\| + \frac{u_i}{1 - u_i} \|\bar{y}_{\mathcal{P}^{(i)}} - \bar{y}_{\mathcal{U}^{(i)}}\| \\
< & \frac{1 - 2u_i}{1 - u_i} \|\bar{y}_{\mathcal{P}^{(i)}} - \bar{y}_{\mathcal{U}^{(i)}}\| + \frac{u_i}{1 - u_i} \|\bar{y}_{\mathcal{P}^{(i)}} - \bar{y}_{\mathcal{U}^{(i)}}\| \\
\leq & \max_{w \in \mathcal{P}^{(i)}} \|y_w - \bar{y}_{\mathcal{U}^{(i)}}\|,
\end{aligned} \tag{85}$$

where (83) is applied to both the second and fourth lines. Therefore, there exists $w' \in \mathcal{P}^{(i)}$ with farther distance to $\bar{y}_{\mathcal{U}^{(i)}}$ than all the regular workers, which guarantees that all the remaining regular workers will not be removed in this iteration.

If at every iteration FABA discards a poisoned worker, then $\mathcal{U}^{(W-R)} = \mathcal{R}$ and

$$\|\bar{y}_{\mathcal{U}^{(W-R)}} - \bar{y}\| = 0 \leq \frac{2\delta}{1 - 3\delta} \cdot \max_{w \in \mathcal{R}} \|y_w - \bar{y}\|. \tag{86}$$

Otherwise, there are iterations with regular workers removed. Denote i^* as the last one among the $W - R$ iterations that removes the regular worker. Denote $w^{(i^*)}$ as the discarded worker at iteration i^*, we have $w^{(i^*)} \in \mathcal{R}$ and from the algorithmic principle of removal

$$\|y_{w^{(i^*)}} - \bar{y}_{\mathcal{U}^{(i^*)}}\| = \max_{w \in \mathcal{U}^{(i^*)}} \|y_w - \bar{y}_{\mathcal{U}^{(i^*)}}\|. \tag{87}$$

Note that

$$\|\bar{y}_{\mathcal{U}^{(W-R)}} - \bar{y}\| \leq \|\bar{y}_{\mathcal{U}^{(W-R)}} - \bar{y}_{\mathcal{U}^{(i^*)}}\| + \|\bar{y}_{\mathcal{U}^{(i^*)}} - \bar{y}\|, \tag{88}$$

thus it suffices to bound the two terms at the right separately. First we notice that

$$\begin{aligned}\|\bar{y}_{\mathcal{U}^{(W-R)}} - \bar{y}_{\mathcal{U}^{(i^*)}}\| =& \|\frac{1}{|\mathcal{U}^{(W-R)}|}\sum_{w\in\mathcal{U}^{(W-R)}} y_w - \frac{1}{|\mathcal{U}^{(i^*)}|}\sum_{w\in\mathcal{U}^{(i^*)}} y_w\| \\ =& \|(\frac{1}{|\mathcal{U}^{(W-R)}|} - \frac{1}{|\mathcal{U}^{(i^*)}|})\sum_{w\in\mathcal{U}^{(i^*)}} y_w - \frac{1}{|\mathcal{U}^{(W-R)}|}\sum_{w\in\mathcal{U}^{(i^*)}\setminus\mathcal{U}^{(W-R)}} y_w\| \\ =& \frac{1}{|\mathcal{U}^{(W-R)}|}\|\sum_{w\in\mathcal{U}^{(i^*)}\setminus\mathcal{U}^{(W-R)}} (y_w - \bar{y}_{\mathcal{U}^{(i^*)}})\| \\ \leq& \frac{|\mathcal{U}^{(i^*)}| - |\mathcal{U}^{(W-R)}|}{|\mathcal{U}^{(W-R)}|}\max_{w\in\mathcal{U}^{(i^*)}\setminus\mathcal{U}^{(W-R)}}\|y_w - \bar{y}_{\mathcal{U}^{(i^*)}}\| \\ \leq& \frac{W-R-i^*}{R}\|y_{w^{(i^*)}} - \bar{y}_{\mathcal{U}^{(i^*)}}\| \\ \leq& \frac{W-R-i^*}{R}(\|y_{w^{(i^*)}} - \bar{y}_{\mathcal{R}^{(i^*)}}\| + \|\bar{y}_{\mathcal{R}^{(i^*)}} - \bar{y}_{\mathcal{U}^{(i^*)}}\|) \\ =& \frac{W-R-i^*}{R}(\|y_{w^{(i^*)}} - \bar{y}_{\mathcal{R}^{(i^*)}}\| + u_{i^*}\|\bar{y}_{\mathcal{R}^{(i^*)}} - \bar{y}_{\mathcal{P}^{(i^*)}}\|) \\ \leq& \frac{W-R-i^*}{R}\left(\max_{w\in\mathcal{R}}\|y_w - \bar{y}_{\mathcal{R}^{(i^*)}}\| + \frac{u_{i^*}}{1-2u_{i^*}}\cdot\max_{w\in\mathcal{R}}\|y_w - \bar{y}_{\mathcal{R}^{(i^*)}}\|\right) \\ =& \frac{W-R-i^*}{R}\cdot\frac{1-u_{i^*}}{1-2u_{i^*}}\max_{w\in\mathcal{R}}\|y_w - \bar{y}_{\mathcal{R}^{(i^*)}}\|,\end{aligned} \tag{89}$$

where the second inequality is from (87) and the last inequality is from (84) with $i = i^*$. Additionally, the second term at the right-hand side of (88) has upper bound

$$\|\bar{y}_{\mathcal{U}^{(i^*)}} - \bar{y}\| = \|(1-u_{i^*})\cdot\bar{y}_{\mathcal{R}^{(i^*)}} + u_{i^*}\cdot\bar{y}_{\mathcal{P}^{(i^*)}} - \bar{y}\| \leq \|\bar{y}_{\mathcal{R}^{(i^*)}} - \bar{y}\| + u_{i^*}\|\bar{y}_{\mathcal{R}^{(i^*)}} - \bar{y}_{\mathcal{P}^{(i^*)}}\|. \tag{90}$$

For $\|\bar{y}_{\mathcal{R}^{(i^*)}} - \bar{y}\|$, we have

$$\begin{aligned}\|\bar{y}_{\mathcal{R}^{(i^*)}} - \bar{y}\| =& \|\frac{1}{|\mathcal{R}^{(i^*)}|}\sum_{w\in\mathcal{R}^{(i^*)}} y_w - \frac{1}{R}\sum_{w\in\mathcal{R}} y_w\| \\ =& \frac{1}{|\mathcal{R}^{(i^*)}|}\|\sum_{w\in\mathcal{R}\setminus\mathcal{R}^{i^*}} (\bar{y} - y_w)\| \\ \leq& \frac{R-|\mathcal{R}^{(i^*)}|}{|\mathcal{R}^{(i^*)}|}\max_{w\in\mathcal{R}}\|y_w - \bar{y}\|.\end{aligned} \tag{91}$$

Substituting (91) and (84) into (90), we have

$$\|\bar{y}_{\mathcal{U}^{(i^*)}} - \bar{y}\| \leq \frac{R-|\mathcal{R}^{(i^*)}|}{|\mathcal{R}^{(i^*)}|}\max_{w\in\mathcal{R}}\|y_w - \bar{y}\| + \frac{u_{i^*}}{1-2u_{i^*}}\max_{w\in\mathcal{R}}\|y_w - \bar{y}_{\mathcal{R}^{(i^*)}}\|. \tag{92}$$

Substituting (89) and (92) into (88), we have

$$\begin{aligned}&\|\bar{y}_{\mathcal{U}^{(W-R)}} - \bar{y}\| \\ \leq& (\frac{W-R-i^*}{R}\cdot\frac{1-u_{i^*}}{1-2u_{i^*}} + \frac{u_{i^*}}{1-2u_{i^*}})\max_{w\in\mathcal{R}}\|y_w - \bar{y}_{\mathcal{R}^{(i^*)}}\| + \frac{R-|\mathcal{R}^{(i^*)}|}{|\mathcal{R}^{(i^*)}|}\max_{w\in\mathcal{R}}\|y_w - \bar{y}\|.\end{aligned} \tag{93}$$

Since

$$\max_{w\in\mathcal{R}}\|y_w-\bar{y}_{\mathcal{R}^{(i^*)}}\|\leq\max_{w\in\mathcal{R}}\|y_w-\bar{y}\|+\|\bar{y}_{\mathcal{R}^{(i^*)}}-\bar{y}\|\leq\frac{R}{|\mathcal{R}^{(i^*)}|}\max_{w\in\mathcal{R}}\|y_w-\bar{y}\|, \tag{94}$$

where the last inequality comes from (91), we have

$$\begin{aligned}&\|\bar{y}_{\mathcal{U}^{(W-R)}}-\bar{y}\| &(95)\\ \leq&\Big((\frac{W-R-i^*}{R}\cdot\frac{1-u_{i^*}}{1-2u_{i^*}}+\frac{u_{i^*}}{1-2u_{i^*}})\frac{R}{|\mathcal{R}^{(i^*)}|}+\frac{R-|\mathcal{R}^{(i^*)}|}{|\mathcal{R}^{(i^*)}|}\Big)\max_{w\in\mathcal{R}}\|y_w-\bar{y}\|\\ =&\frac{2u_{i^*}}{1-2u_{i^*}}\max_{w\in\mathcal{R}}\|y_w-\bar{y}\|. &(96)\end{aligned}$$

Since $u_{i^*}=\frac{|\mathcal{P}^{(i^*)}|}{|\mathcal{U}^{(i^*)}|}\leq\frac{W-R}{R}=\frac{\delta}{1-\delta}$, we have

$$\|\bar{y}_{\mathcal{U}^{(W-R)}}-\bar{y}\|\leq\frac{2\delta}{1-3\delta}\cdot\max_{w\in\mathcal{R}}\|y_w-\bar{y}\|. \tag{97}$$

Combining the first case (86) and the second case (97), we have

$$\|\text{FABA}(\{y_1,\ldots,y_W\}-\bar{y})\|=\|\bar{y}_{\mathcal{U}^{(P)}}-\bar{y}\|\leq\frac{2\delta}{1-3\delta}\cdot\max_{w\in\mathcal{R}}\|y_w-\bar{y}\|, \tag{98}$$

which completes the proof. ■

Appendix C. Proof of Theorem 7

We give a complete version of Theorem 7 as follows.

Theorem 17 *Consider Algorithm 1 with a ρ-robust aggregator $RAgg(\cdot)$ to solve* (1) *and suppose that Assumptions 1, 2, 3, and 4 hold. Under label poisoning attacks where the fraction of poisoned workers is $\delta\in[0,\frac{1}{2})$, if the step size is*

$$\gamma=\min\Big\{\sqrt{\frac{4(f(x^0)-f^*)+\frac{15\rho^2(R+\frac{1}{R})\sigma^2}{8L}}{T(40L\sigma^2)(3\rho^2(R+\frac{1}{R})+\frac{2}{R})}},\frac{1}{8L}\Big\}, \tag{99}$$

the momentum coefficient $\alpha=8L\gamma$, then we have

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(x^t)\|^2 &(100)\\ \leq&15\rho^2\xi^2+\sqrt{\frac{20L\sigma^2(\frac{2}{R}+3\rho^2(R+\frac{1}{R}))}{T}}\cdot\sqrt{32(f(x^0)-f^*)+\frac{15}{L}\rho^2(R+\frac{1}{R})\sigma^2}\\ &+\frac{32L(f(x^0)-f^*)}{T}+\frac{15\rho^2(R+\frac{1}{R})\sigma^2}{T}+\frac{\frac{10\sigma^2}{R}+12\rho^2((R+\frac{1}{R})\sigma^2+\xi^2)-\|\nabla f(x^0)\|}{T}.\end{aligned}$$

where the expectation is taken over the algorithm's randomness.

Proof For notational convenience, denote $m^t = \text{RAgg}(\{\hat{m}_w^t : w \in \mathcal{W}\})$ and $\bar{m}^t = \frac{1}{R}\sum_{w\in\mathcal{R}} m_w^t$. Denote the conditional expectation $\mathbb{E}[\cdot|i_w^\tau : \tau < t, w \in \mathcal{W}]$ as $\mathbb{E}_t[\cdot]$. Because $f(x)$ has L-Lipschitz continuous gradients from Assumption 2, it holds that

$$\begin{aligned} f(x^{t+1}) &\le f(x^t) + \langle \nabla f(x^t), x^{t+1} - x^t\rangle + \frac{L}{2}\|x^{t+1} - x^t\|^2 \\ &\le f(x^t) - \gamma\langle \nabla f(x^t), m^t\rangle + \frac{L}{2}\gamma^2\|m^t\|^2. \end{aligned} \tag{101}$$

Since

$$-\langle \nabla f(x^t), m^t\rangle = \frac{1}{2}\|m^t - \nabla f(x^t)\|^2 - \frac{1}{2}\|\nabla f(x^t)\|^2 - \frac{1}{2}\|m^t\|^2, \tag{102}$$

we have

$$\begin{aligned} f(x^{t+1}) \le& f(x^t) + \frac{\gamma}{2}\|m^t - \nabla f(x^t)\|^2 - \frac{\gamma}{2}\|\nabla f(x^t)\|^2 - \frac{\gamma}{2}(1 - L\gamma)\|m^t\|^2 \\ \le& f(x^t) + \frac{\gamma}{2}\|m^t - \nabla f(x^t)\|^2 - \frac{\gamma}{2}\|\nabla f(x^t)\|^2 \\ \le& f(x^t) + \gamma\|m^t - \bar{m}^t\|^2 + \gamma\|\bar{m}^t - \nabla f(x^t)\|^2 - \frac{\gamma}{2}\|\nabla f(x^t)\|^2 \\ =& f(x^t) + \gamma\|m^t - \bar{m}^t\|^2 + \gamma\|e^t\|^2 - \frac{\gamma}{2}\|\nabla f(x^t)\|^2, \end{aligned} \tag{103}$$

where the second inequality is from $\gamma \le \frac{1}{8L} < \frac{1}{L}$ and $e^t \triangleq \bar{m}^t - \nabla f(x^t)$. Taking expectations on both sides of (103) reaches

$$\mathbb{E}[f(x^{t+1})] \le \mathbb{E}[f(x^t)] + \gamma\mathbb{E}\|m^t - \bar{m}^t\|^2 + \gamma\mathbb{E}\|e^t\|^2 - \frac{\gamma}{2}\mathbb{E}\|\nabla f(x^t)\|^2. \tag{104}$$

For the term $\|m^t - \bar{m}^t\|^2$, since $\text{RAgg}(\cdot)$ is a ρ-robust aggregator with Definition 4, it holds that

$$\|m^t - \bar{m}^t\|^2 \le \rho^2 \cdot \max_{w\in\mathcal{R}} \|m_w^t - \bar{m}^t\|^2. \tag{105}$$

For the term $\max_{w\in\mathcal{R}} \|m_w^t - \bar{m}^t\|^2$, we have

$$\begin{aligned} &\max_{w\in\mathcal{R}} \|m_w^t - \bar{m}^t\|^2 \\ \le& \max_{w\in\mathcal{R}} \left\{3\|m_w^t - \mathbb{E}[m_w^t]\|^2 + 3\|\bar{m}^t - \mathbb{E}[\bar{m}^t]\|^2 + 3\|\mathbb{E}[m_w^t] - \mathbb{E}[\bar{m}^t]\|^2\right\} \\ \le& 3\max_{w\in\mathcal{R}} \|m_w^t - \mathbb{E}[m_w^t]\|^2 + 3\|\bar{m}^t - \mathbb{E}[\bar{m}^t]\|^2 + 3\max_{w\in\mathcal{R}} \|\mathbb{E}[m_w^t] - \mathbb{E}[\bar{m}^t]\|^2. \end{aligned} \tag{106}$$

We will bound the expectation of each term at the right-hand side of (106) as follows.

For the first term at the right-hand side of (106), we have

$$\mathbb{E}[\max_{w\in\mathcal{R}} \|m_w^t - \mathbb{E}[m_w^t]\|^2] \le \mathbb{E}[\sum_{w\in\mathcal{R}} \|m_w^t - \mathbb{E}[m_w^t]\|^2] = \sum_{w\in\mathcal{R}} \mathbb{E}\|m_w^t - \mathbb{E}[m_w^t]\|^2. \tag{107}$$

For any $w \in \mathcal{R}$, denoting

$$A_w^t \triangleq \mathbb{E}\|m_w^t - \mathbb{E}[m_w^t]\|^2, \tag{108}$$

we obtain that

$$\begin{aligned} A_w^t &= \mathbb{E}\|(1-\alpha)(m_w^{t-1} - \mathbb{E}[m_w^{t-1}]) + \alpha(\nabla f_{w,i_w^t}(x^t) - \nabla f_w(x^t))\|^2 \\ &= (1-\alpha)^2\mathbb{E}\|m_w^{t-1} - \mathbb{E}[m_w^{t-1}]\|^2 + \alpha^2\mathbb{E}\|\nabla f_{w,i_w^t}(x^t) - \nabla f_w(x^t)\|^2 \\ &= (1-\alpha)^2\mathbb{E}\|m_w^{t-1} - \mathbb{E}[m_w^{t-1}]\|^2 + \alpha^2\mathbb{E}[\mathbb{E}_t\|\nabla f_{w,i_w^t}(x^t) - \nabla f_w(x^t)\|^2] \\ &\le (1-\alpha)^2 A_w^{t-1} + \alpha^2\sigma^2 \\ &\le (1-\alpha)^{2t} A_w^0 + (\sum_{l=1}^{t}(1-\alpha)^{2(t-l)})\alpha^2\sigma^2, \end{aligned} \tag{109}$$

where the first inequality comes from Assumption 4 and the second inequality is due to the use of telescopic cancellation. Since

$$A_w^0 = \mathbb{E}\|m_w^0 - \mathbb{E}[m_w^0]\|^2 = \mathbb{E}\|\nabla f_{w,i_w^0}(x^0) - \nabla f_w(x^0)\|^2 \le \sigma^2, \tag{110}$$

where the inequality is due to Assumption 4, we have

$$A_w^t \le \sigma^2(\alpha + (1-\alpha)^{2t}), \tag{111}$$

and

$$\mathbb{E}[\max_{w\in\mathcal{R}}\|m_w^t - \mathbb{E}[m_w^t]\|^2] \le R\sigma^2(\alpha + (1-\alpha)^{t+1}). \tag{112}$$

For the second term at the right-hand side of (106), denoting

$$B^t \triangleq \mathbb{E}\|\bar{m}^t - \mathbb{E}[\bar{m}^t]\|^2, \tag{113}$$

we have that

$$\begin{aligned} B^t &= \mathbb{E}\|\frac{1}{R}\sum_{w\in\mathcal{R}}(m_w^t - \mathbb{E}[m_w^t])\|^2 \\ &= \frac{1}{R^2}\mathbb{E}\|\sum_{w\in\mathcal{R}}(m_w^t - \mathbb{E}[m_w^t])\|^2 \\ &= \frac{1}{R^2}\Big(\sum_{w\in\mathcal{R}}\mathbb{E}\|m_w^t - \mathbb{E}[m_w^t]\|^2 + \sum_{w\in\mathcal{R}}\sum_{v\in\mathcal{R},v\neq w}\mathbb{E}\langle m_w^t - \mathbb{E}[m_w^t], m_v^t - \mathbb{E}[m_v^t]\rangle\Big) \\ &= \frac{1}{R^2}\Big(\sum_{w\in\mathcal{R}}\mathbb{E}\|m_w^t - \mathbb{E}[m_w^t]\|^2 + \sum_{w\in\mathcal{R}}\sum_{v\in\mathcal{R},v\neq w}\underbrace{\langle\mathbb{E}[m_w^t] - \mathbb{E}[m_w^t], \mathbb{E}[m_v^t] - \mathbb{E}[m_v^t]\rangle}_{=0}\Big) \\ &= \frac{1}{R^2}\sum_{w\in\mathcal{R}}\mathbb{E}\|m_w^t - \mathbb{E}[m_w^t]\|^2 \\ &= \frac{1}{R^2}\sum_{w\in\mathcal{R}}A_w^t. \end{aligned} \tag{114}$$

With (111), we obtain that

$$B^t \le \frac{\sigma^2}{R}(\alpha + (1-\alpha)^{2t}). \tag{115}$$

For the third term at the right-hand side of (106), denoting

$$C^t \triangleq \max_{w\in\mathcal{R}} \|\mathbb{E}[m_w^t] - \mathbb{E}[\bar{m}^t]\|^2, \tag{116}$$

we have

$$\begin{aligned} C^t =& \max_{w\in\mathcal{R}} \|(1-\alpha)(\mathbb{E}[m_w^{t-1}] - \mathbb{E}[\bar{m}^{t-1}]) + \alpha(\nabla f_w(x^t) - \nabla f(x^t))\|^2 \\ \leq& \max_{w\in\mathcal{R}} \Big\{(1-\alpha)\|\mathbb{E}[m_w^{t-1}] - \mathbb{E}[\bar{m}^{t-1}]\|^2 + \alpha\|\nabla f_w(x^t) - \nabla f(x^t)\|^2\Big\} \\ \leq& (1-\alpha)\max_{w\in\mathcal{R}} \|\mathbb{E}[m_w^{t-1}] - \mathbb{E}[\bar{m}^{t-1}]\|^2 + \alpha \max_{w\in\mathcal{R}} \|\nabla f_w(x^t) - \nabla f(x^t)\|^2 \\ \leq& (1-\alpha)C^{t-1} + \alpha\xi^2 \\ \leq& (1-\alpha)^t C^0 + (\sum_{l=0}^{t-1}(1-\alpha)^l)\alpha\xi^2, \end{aligned} \tag{117}$$

where the third inequality is from Assumption 3. Since

$$C^0 = \max_{w\in\mathcal{R}} \|\nabla f_w(x^0) - \nabla f(x^0)\|^2 \leq \xi^2, \tag{118}$$

which come from Assumption 3, it holds that

$$C^t \leq \xi^2\Big((1-\alpha)^t + (\sum_{l=0}^{t-1}(1-\alpha)^l)\alpha\Big) = \xi^2. \tag{119}$$

Substituting (112), (115) and (119) into (106) and taking expectations on both sides of (106), we have

$$\mathbb{E}[\max_{w\in\mathcal{R}} \|m_w^t - \bar{m}^t\|^2] \leq 3((R+\frac{1}{R})\sigma^2(\alpha + (1-\alpha)^{2t}) + \xi^2). \tag{120}$$

With (105), we have

$$\mathbb{E}\|m^t - \bar{m}^t\|^2 \leq 3\rho^2((R+\frac{1}{R})\sigma^2(\alpha + (1-\alpha)^{2t}) + \xi^2). \tag{121}$$

For the term $\mathbb{E}\|e^t\|^2$ in (103), according to (Karimireddy et al., 2022, Lemma 10), $\mathbb{E}\|e^0\|^2 \leq \frac{\sigma^2}{R}$ and for $t \geq 1$, it holds that

$$\mathbb{E}\|e^t\|^2 \leq (1-\frac{2\alpha}{5})\mathbb{E}\|e^{t-1}\|^2 + \frac{\alpha}{10}\mathbb{E}\|\nabla f(x^{t-1})\|^2 + \frac{\alpha}{10}\mathbb{E}\|m^{t-1} - \bar{m}^{t-1}\|^2 + \frac{\alpha^2\sigma^2}{R}. \tag{122}$$

Combining (103) and (122), we have

$$\begin{aligned} &\mathbb{E}[f(x^{t+1})] + \frac{5\gamma}{2\alpha}\mathbb{E}\|e^t\|^2 \\ \leq& \mathbb{E}[f(x^t)] + \gamma\mathbb{E}\|m^t - \bar{m}^t\|^2 + \gamma\mathbb{E}\|e^t\|^2 - \frac{\gamma}{2}\mathbb{E}\|\nabla f(x^t)\|^2 + (\frac{5\gamma}{2\alpha} - \gamma)\mathbb{E}\|e^{t-1}\|^2 \\ &+ \frac{\gamma}{4}\mathbb{E}\|\nabla f(x^{t-1})\|^2 + \frac{\gamma}{4}\mathbb{E}\|m^{t-1} - \bar{m}^{t-1}\|^2 + \frac{5\alpha\gamma\sigma^2}{2R}. \end{aligned} \tag{123}$$

Using (121) to bound $\mathbb{E}\|m^t - \bar{m}^t\|^2$ and $\mathbb{E}\|m^{t-1} - \bar{m}^{t-1}\|^2$, it is obtained that

$$\begin{aligned}
&\mathbb{E}[f(x^{t+1})] + \frac{5\gamma}{2\alpha}\mathbb{E}\|e^t\|^2 \qquad (124)\\
\leq&\mathbb{E}[f(x^t)] + \frac{15}{4}\gamma\rho^2(R+\frac{1}{R})\sigma^2(\alpha+(1-\alpha)^{2t}) + \frac{15}{4}\gamma\rho^2\xi^2 + \gamma\mathbb{E}\|e^t\|^2\\
&- \frac{\gamma}{2}\mathbb{E}\|\nabla f(x^t)\|^2 + (\frac{5\gamma}{2\alpha}-\gamma)\mathbb{E}\|e^{t-1}\|^2 + \frac{\gamma}{4}\mathbb{E}\|\nabla f(x^{t-1})\|^2 + \frac{5\alpha\gamma\sigma^2}{2R}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}[f(x^{t+1})] + (\frac{5\gamma}{2\alpha}-\gamma)\mathbb{E}\|e^t\|^2 + \frac{\gamma}{4}\mathbb{E}\|\nabla f(x^t)\|^2 \qquad (125)\\
\leq&\mathbb{E}[f(x^t)] + (\frac{5\gamma}{2\alpha}-\gamma)\mathbb{E}\|e^{t-1}\|^2 + \frac{\gamma}{4}\mathbb{E}\|\nabla f(x^{t-1})\|^2 - \frac{\gamma}{4}\mathbb{E}\|\nabla f(x^t)\|^2\\
&+ \frac{15}{4}\gamma\rho^2(R+\frac{1}{R})\sigma^2(\alpha+(1-\alpha)^{2t}) + \frac{15}{4}\gamma\rho^2\xi^2 + \frac{5\alpha\gamma\sigma^2}{2R}.
\end{aligned}$$

Denoting $\mathcal{E}_t = \mathbb{E}[f(x^{t+1})] + (\frac{5\gamma}{2\alpha}-\gamma)\mathbb{E}\|e^t\|^2 + \frac{\gamma}{4}\mathbb{E}\|\nabla f(x^t)\|^2$, we have

$$\frac{\gamma}{4}\mathbb{E}\|\nabla f(x^t)\|^2 \leq \mathcal{E}_{t-1} - \mathcal{E}_t + \frac{15}{4}\gamma\rho^2(R+\frac{1}{R})\sigma^2(\alpha+(1-\alpha)^{2t}) + \frac{15}{4}\gamma\rho^2\xi^2 + \frac{5\alpha\gamma\sigma^2}{2R}, \quad (126)$$

and

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(x^t)\|^2 \qquad (127)\\
\leq&\frac{4(\mathcal{E}_0-\mathcal{E}_T)}{\gamma T} + \frac{1}{T}\sum_{t=1}^{T}15\rho^2(R+\frac{1}{R})\sigma^2(\alpha+(1-\alpha)^{2t}) + 15\rho^2\xi^2 + \frac{10\alpha\sigma^2}{R}\\
\leq&\frac{4(\mathcal{E}_0-\mathcal{E}_T)}{\gamma T} + \frac{15\rho^2(R+\frac{1}{R})\sigma^2}{\alpha T} + 15\rho^2(R+\frac{1}{R})\sigma^2\alpha + 15\rho^2\xi^2 + \frac{10\alpha\sigma^2}{R}.\\
\leq&\frac{4(\mathcal{E}_0-f^*)}{\gamma T} + \frac{15\rho^2(R+\frac{1}{R})\sigma^2}{\alpha T} + 15\rho^2(R+\frac{1}{R})\sigma^2\alpha + 15\rho^2\xi^2 + \frac{10\alpha\sigma^2}{R}.
\end{aligned}$$

where the second inequality is from $\sum_{t=1}^{T}(1-\alpha)^{2t} \leq \frac{1}{\alpha}$ when $0 \leq \alpha \leq 1$, and the third inequality is from $\mathcal{E}_T \geq f^*$ due to Assumption 1. Using the following equalities and inequalities

$$\mathcal{E}_0 = \mathbb{E}[f(x^1)] + \frac{3\gamma}{2}\mathbb{E}\|e^0\|^2 + \frac{\gamma}{4}\|\nabla f(x^0)\|^2, \quad (128)$$

$$\mathbb{E}[f(x^1)] \leq f(x^0) + \gamma\mathbb{E}\|m^0 - \bar{m}^0\|^2 + \gamma\mathbb{E}\|e^0\|^2 - \frac{\gamma}{2}\|\nabla f(x^0)\|^2, \quad (129)$$

$$\mathbb{E}\|e^0\|^2 = \mathbb{E}\|\bar{m}^0 - \nabla f(x^0)\|^2 = \mathbb{E}\|\frac{1}{R}\sum_{w\in\mathcal{R}}(\nabla f_{w,i_w^0}(x^0) - \nabla f_w(x^0))\|^2 \leq \frac{\sigma^2}{R}, \quad (130)$$

$$\mathbb{E}\|m^0 - \bar{m}^0\|^2 \leq \rho^2\mathbb{E}[\max_{w\in\mathcal{R}}\|m_w^0 - \bar{m}^0\|^2] \leq 3\rho^2((R+\frac{1}{R})\sigma^2 + \xi^2), \quad (131)$$

we have

$$\mathcal{E}_0 \leq f(x^0) + \frac{5\gamma\sigma^2}{2R} + 3\gamma\rho^2((R+\frac{1}{R})\sigma^2 + \xi^2) - \frac{\gamma}{4}\|\nabla f(x^0)\|^2. \quad (132)$$

Substituting (132) into (127), we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(x^t)\|^2 \qquad (133)\\
\leq&\frac{4(f(x^0)-f^*)}{\gamma T}+\frac{15\rho^2(R+\frac{1}{R})\sigma^2}{8\gamma LT}+120\rho^2(R+\frac{1}{R})\sigma^2\gamma L+\frac{80L\gamma\sigma^2}{R}\\
&+\frac{\frac{10\sigma^2}{R}+12\rho^2((R+\frac{1}{R})\sigma^2+\xi^2)-\|\nabla f(x^0)\|}{T}+15\rho^2\xi^2,\\
=&\frac{1}{\gamma}\cdot\Big(\frac{4(f(x^0)-f^*)}{T}+\frac{15\rho^2(R+\frac{1}{R})\sigma^2}{8LT}\Big)+\gamma\cdot\Big(120\rho^2(R+\frac{1}{R})\sigma^2L+\frac{80L\sigma^2}{R}\Big)\\
&+\frac{\frac{10\sigma^2}{R}+12\rho^2((R+\frac{1}{R})\sigma^2+\xi^2)-\|\nabla f(x^0)\|}{T}+15\rho^2\xi^2,
\end{aligned}$$

where the inequality uses the fact that $\alpha = 8L\gamma$. Substituting the step size

$$\gamma=\min\Big\{\sqrt{\frac{4(f(x^0)-f^*)+\frac{15\rho^2(R+\frac{1}{R})\sigma^2}{8L}}{T(40L\sigma^2)(3\rho^2(R+\frac{1}{R})+\frac{2}{R})}},\frac{1}{8L}\Big\}, \qquad (134)$$

we have

$$\begin{aligned}
\frac{1}{\gamma}=&\max\Big\{\sqrt{\frac{T(40L\sigma^2)(3\rho^2(R+\frac{1}{R})+\frac{2}{R})}{4(f(x^0)-f^*)+\frac{15\rho^2(R+\frac{1}{R})\sigma^2}{8L}}},8L\Big\} \qquad (135)\\
\leq&\sqrt{\frac{T(40L\sigma^2)(3\rho^2(R+\frac{1}{R})+\frac{2}{R})}{4(f(x^0)-f^*)+\frac{15\rho^2(R+\frac{1}{R})\sigma^2}{8L}}}+8L.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(x^t)\|^2 \qquad (136)\\
\leq&15\rho^2\xi^2+\sqrt{\frac{20L\sigma^2(\frac{2}{R}+3\rho^2(R+\frac{1}{R}))}{T}}\cdot\sqrt{32(f(x^0)-f^*)+\frac{15}{L}\rho^2(R+\frac{1}{R})\sigma^2}\\
&+\frac{32L(f(x^0)-f^*)}{T}+\frac{15\rho^2(R+\frac{1}{R})\sigma^2}{T}+\frac{\frac{10\sigma^2}{R}+12\rho^2((R+\frac{1}{R})\sigma^2+\xi^2)-\|\nabla f(x^0)\|}{T}.
\end{aligned}$$

which completes the proof. ■

Appendix D. Proof of Theorem 8

We give a complete version of Theorem 8 as follows.

Theorem 18 *Consider Algorithm 1 with the mean aggregator Mean$(\cdot)$ to solve* (1) *and suppose that Assumptions 1, 2, 4, and 5 hold. Under label poisoning attacks where the fraction of poisoned workers is $\delta \in [0, 1)$, if the step size is*

$$\gamma = \min\Big\{\sqrt{\frac{4(f(x^0) - f^*) + \frac{30\delta^2\sigma^2}{8L}}{T(40L\sigma^2)(6\delta^2 + \frac{2}{R})}}, \frac{1}{8L}\Big\}, \tag{137}$$

the momentum coefficient $\alpha = 8L\gamma$, then we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(x^t)\|^2 \leq & 15\delta^2 A^2 + \sqrt{\frac{20L\sigma^2(\frac{2}{R} + 6\delta^2)}{T}} \cdot \sqrt{32(f(x^0) - f^*) + \frac{30}{L}\delta^2\sigma^2} \\ & + \frac{32L(f(x^0) - f^*)}{T} + \frac{30\delta^2\sigma^2}{T} + \frac{\frac{10\sigma^2}{R} + 24\delta^2(\sigma^2 + \xi^2) - \|\nabla f(x^0)\|}{T}.\end{aligned} \tag{138}$$

where the expectation is taken over the algorithm's randomness.

Proof The proof of Theorem 8 is similar to that of Theorem 7, except for the analysis of the distance between the aggregated message and the true average of the regular messages. For notational simplicity, denote the conditional expectation $\mathbb{E}[\cdot|i_w^\tau : \tau < t, w \in \mathcal{W}]$ as $\mathbb{E}_t[\cdot]$. Denoting $m^t = \text{Mean}(\{\hat{m}_w^t : w \in \mathcal{W}\})$ and $\bar{m}^t = \frac{1}{R}\sum_{w\in\mathcal{R}} m_w^t$, we have

$$\begin{aligned}\mathbb{E}\|m^t - \bar{m}^t\|^2 &= \mathbb{E}\|\frac{1}{W}\sum_{w\in\mathcal{W}}\hat{m}_w^t - \frac{1}{R}\sum_{w\in\mathcal{R}} m_w^t\|^2 \\ &= \mathbb{E}\|\frac{1}{W}\sum_{w\in\mathcal{W}\setminus\mathcal{R}}\tilde{m}_w^t - (\frac{1}{R} - \frac{1}{W})\sum_{w\in\mathcal{R}} m_w^t\|^2 \\ &= (\frac{1}{W})^2\mathbb{E}\|\sum_{w\in\mathcal{W}\setminus\mathcal{R}}(\tilde{m}_w^t - \bar{m}^t)\|^2 \\ &\leq \frac{W - R}{W^2}\sum_{w\in\mathcal{W}\setminus\mathcal{R}}\mathbb{E}\|\tilde{m}_w^t - \bar{m}^t\|^2.\end{aligned} \tag{139}$$

For the term $\mathbb{E}\|\tilde{m}_w^t - \bar{m}^t\|^2$, for any $w \in \mathcal{W} \setminus \mathcal{R}$, we have

$$\mathbb{E}\|\tilde{m}_w^t - \bar{m}^t\|^2 \leq 3\Big(\mathbb{E}\|\tilde{m}_w^t - \mathbb{E}[\tilde{m}_w^t]\|^2 + \mathbb{E}\|\bar{m}^t - \mathbb{E}[\bar{m}^t]\|^2 + \|\mathbb{E}[\tilde{m}_w^t] - \mathbb{E}[\bar{m}^t]\|^2\Big). \tag{140}$$

For the first term at the right-hand side of (140), similar to the proof from (109) to (111), for any $w \in \mathcal{W} \setminus \mathcal{R}$, we have

$$\begin{aligned}\mathbb{E}\|\tilde{m}_w^t - \mathbb{E}[\tilde{m}_w^t]\|^2 &= \mathbb{E}\|(1-\alpha)(\tilde{m}_w^{t-1} - \mathbb{E}[\tilde{m}_w^{t-1}]) + \alpha(\nabla\tilde{f}_{w,i_w^t}(x^t) - \nabla\tilde{f}_w(x^t))\|^2 \\ &= (1-\alpha)^2\mathbb{E}\|\tilde{m}_w^{t-1} - \mathbb{E}[\tilde{m}_w^{t-1}]\|^2 + \alpha^2\mathbb{E}\|\nabla\tilde{f}_{w,i_w^t}(x^t) - \nabla\tilde{f}_w(x^t)\|^2 \\ &= (1-\alpha)^2\mathbb{E}\|\tilde{m}_w^{t-1} - \mathbb{E}[\tilde{m}_w^{t-1}]\|^2 + \alpha^2\mathbb{E}[\mathbb{E}_t\|\nabla\tilde{f}_{w,i_w^t}(x^t) - \nabla\tilde{f}_w(x^t)\|^2] \\ &= (1-\alpha)^2\mathbb{E}\|\tilde{m}_w^{t-1} - \mathbb{E}[\tilde{m}_w^{t-1}]\|^2 + \alpha^2\sigma^2 \\ &\leq \sigma^2(\alpha + (1-\alpha)^{2t}).\end{aligned} \tag{141}$$

For the second term at the right-hand side of (140), similar to the proof of (114), we have

$$\begin{aligned}\mathbb{E}\|\bar{m}^t - \mathbb{E}[\bar{m}^t]\|^2 &= \mathbb{E}\|\frac{1}{R}\sum_{w\in\mathcal{R}} m_w^t - \mathbb{E}[m_w^t]\|^2 \\ &= \frac{1}{R^2}\sum_{w\in\mathcal{R}} \mathbb{E}\|m_w^t - \mathbb{E}[m_w^t]\|^2 \\ &\leq \frac{\sigma^2}{R}(\alpha + (1-\alpha)^{2t}).\end{aligned} \tag{142}$$

For the third term at the right-hand side of (140), we have

$$\begin{aligned}\|\mathbb{E}[\tilde{m}_w^t] - \mathbb{E}[\bar{m}^t]\|^2 &= \|(1-\alpha)(\mathbb{E}[\tilde{m}_w^{t-1}] - \mathbb{E}[\bar{m}^{t-1}]) + \alpha(\nabla\tilde{f}_w(x^t) - \nabla f(x^t))\|^2 \\ &\leq (1-\alpha)\|\mathbb{E}[\tilde{m}_w^{t-1}] - \mathbb{E}[\bar{m}^{t-1}]\|^2 + \alpha\|\nabla\tilde{f}_w(x^t) - \nabla f(x^t)\|^2 \\ &\leq (1-\alpha)\|\mathbb{E}[\tilde{m}_w^{t-1}] - \mathbb{E}[\bar{m}^{t-1}]\|^2 + \alpha A^2 \\ &\leq (1-\alpha)^t\|\mathbb{E}[\tilde{m}_w^0] - \mathbb{E}[\bar{m}^0]\|^2 + (\sum_{l=0}^{t-1}(1-\alpha)^l)\alpha A^2 \\ &= (1-\alpha)^t\|\nabla\tilde{f}_w(x^0) - \nabla f(x^0)\|^2 + (\sum_{l=0}^{t-1}(1-\alpha)^l)\alpha A^2 \\ &\leq A^2((1-\alpha)^t + 1 - (1-\alpha)^t) \\ &= A^2,\end{aligned} \tag{143}$$

where the second inequality and the fourth inequality are due to Assumption 5.

Substituting (140), (141), (142) into (143), we have

$$\begin{aligned}\mathbb{E}\|m^t - \bar{m}^t\|^2 &\leq 3\cdot\frac{(W-R)^2}{W^2}\cdot(\sigma^2(\alpha + (1-\alpha)^{2t})(1+\frac{1}{R}) + A^2) \\ &\leq 3\cdot\frac{(W-R)^2}{W^2}\cdot(2\sigma^2(\alpha + (1-\alpha)^{2t}) + A^2) \\ &= 6\delta^2\sigma^2(\alpha + (1-\alpha)^{2t}) + 3\delta^2A^2,\end{aligned} \tag{144}$$

where the second inequality is due to $R \geq 1$.

The rest of the proof is the same as that of Theorem 7 and therefore omitted. This completes the proof. ■

Appendix E. Proof of Theorem 9

Proof The key idea of the proof is to construct two instances that have the same set of W local costs after the poisoned workers carry out label poisoning attacks, while the objectives based on the R regular local costs are different in these two instances. Since the algorithm whose output is invariant with respect to the identities of the workers cannot distinguish these two instances, it yields the same output. However, the two different objectives imply that the algorithmic output must fail

on at least one among the two. Therefore, the instance that has a larger learning error gives a lower bound of the learning error for the algorithm.

Without loss of generality, we let $\mathcal{W} = \{1, \cdots, W\}$ be the set of workers, within which $\mathcal{R} = \{1, \cdots, R\}$ is the set of regular workers. The samples of all workers have two possible labels, 1 or 2, which correspond to two different functions $f(x; 1)$ and $f(x; 2)$ with

$$f(x; k) = \frac{(1-\delta)c}{\sqrt{2}}[x]_k + \frac{L}{2}\|x\|^2, \quad k \in \{1, 2\}, \tag{145}$$

within which $c \triangleq \min\{\xi, A\}$.

Each worker $w \in \{1, \cdots, W\}$ has the same J samples costs, in the form of

$$\hat{f}_{w,j}(x) = \hat{f}_w(x) = f(x; \hat{b}^{(w)}), \quad \forall j \in \{1, \cdots, J\}, \tag{146}$$

or equivalently

$$f_{w,j}(x) = f_w(x) = f(x; b^{(w)}), \quad \forall w \leq R, \ \forall j \in \{1, \cdots, J\}, \tag{147}$$

$$\tilde{f}_{w,j}(x) = \tilde{f}_w(x) = f(x; \tilde{b}^{(w)}), \quad \forall w > R, \ \forall j \in \{1, \cdots, J\}. \tag{148}$$

Now we construct two instances with different sets of labels. The first set of labels, denoted as $\{\hat{b}^{(w,1)}, w \in \{1, \cdots, W\}\}$, is given by

$$\hat{b}^{(w,1)} = \begin{cases} 1, & w \leq R, \\ 2, & w > R. \end{cases} \tag{149}$$

The second set of labels, denoted as $\{\hat{b}^{(w,2)}, w \in \{1, \cdots, W\}\}$, is given by

$$\hat{b}^{(w,2)} = \begin{cases} 1, & w > W - R, \\ 2, & w \leq W - R. \end{cases} \tag{150}$$

Denote $f^{(1)}(x) = \frac{1}{R}\sum_{w=1}^{R} f(x; \hat{b}^{(w,1)})$ and $f^{(2)}(x) = \frac{1}{R}\sum_{w=1}^{R} f(x; \hat{b}^{(w,2)})$ as the two objectives. We can check that all the assumptions are satisfied in these two instances. Since

$$\nabla f(x; k) = \frac{(1-\delta)c}{\sqrt{2}} e_k + Lx, \tag{151}$$

where e_k is the unit vector with the k-th element being 1, the gradients of $f^{(1)}$ and $f^{(2)}$ are

$$\nabla f^{(1)}(x) = \frac{(1-\delta)c}{\sqrt{2}} e_1 + Lx, \tag{152}$$

$$\nabla f^{(2)}(x) = \frac{(1-2\delta)c}{\sqrt{2}} e_1 + \frac{\delta c}{\sqrt{2}} e_2 + Lx. \tag{153}$$

As a result, we know their minimums are achieved at $x^{*,(1)} = -\frac{(1-\delta)c}{\sqrt{2}L}e_1$ and $x^{*,(2)} = -\frac{(1-2\delta)c}{\sqrt{2}L}e_1 - \frac{\delta c}{\sqrt{2}L}e_2$, respectively, and there exists a uniform lower bound

$$f^{(k)}(x) \geq f^* \triangleq -\frac{c^2}{2L}. \tag{154}$$

satisfying Assumption 1 for $k \in \{1, 2\}$.

As the gradients are linear, Assumption 2 is satisfied and the constant is exactly L.

Further, Assumption 3 is satisfied with constant ξ, as

$$\max_{w \leq R} \|\nabla f(x; \hat{b}^{(w,1)}) - \nabla f^{(1)}(x)\| = 0, \tag{155}$$

$$\max_{w \leq R} \|\nabla f(x; \hat{b}^{(w,2)}) - \nabla f^{(2)}(x)\| = \max\{|1-2\delta|, \delta\}c \leq c \leq \xi. \tag{156}$$

The last inequality of (156) comes from $c \triangleq \min\{A, \xi\} \leq \xi$.

Note that $f_{w,j}(x) = \hat{f}_{w,j}(x)$ when $w \leq R$ and $\tilde{f}_{w,j}(x) = \hat{f}_{w,j}(x)$ otherwise. Therefore, due to $\hat{f}_{w,j}(x) = \hat{f}_w(x)$, Assumption 4 is satisfied with constant $\sigma = 0$.

Assumption 5 is satisfied with constant A, since

$$\max_{w > R} \|\nabla f(x; \hat{b}^{(w,1)}) - \nabla f^{(1)}(x)\| = (1-\delta)c \leq c \leq A, \tag{157}$$

$$\max_{w > R} \|\nabla f(x; \hat{b}^{(w,2)}) - \nabla f^{(2)}(x)\| = \delta c \leq c \leq A. \tag{158}$$

The last inequalities of (157) and (158) come from $c \triangleq \min\{A, \xi\} \leq A$.

The two constructed instances result in the same set of local costs (R of them are $f(x; 1)$ and the others are $f(x; 2)$) yet different orders of labels. Since the algorithm whose output is invariant with respect to the identities of the workers cannot distinguish these two instances, it yields the same output. Nevertheless, according to (152) and (153), the gradients of the two objectives at the algorithmic output are different. Therefore, the algorithmic output must fail on at least one among the two. Below, we investigate the learning errors in the two instances, and the larger one exactly gives the lower bound of the learning error for the algorithm.

For any x, note that

$$\begin{aligned} &\max\{\|\nabla f^{(1)}(x)\|, \|\nabla f^{(2)}(x)\|\} \\ \geq &\frac{1}{2}\left(\|\nabla f^{(1)}(x)\| + \|\nabla f^{(2)}(x)\|\right) \\ \geq &\frac{1}{2}\|\nabla f^{(1)}(x) - \nabla f^{(2)}(x)\| \\ = &\frac{\delta c}{2}. \end{aligned} \tag{159}$$

Specifically, letting $x = x^t$ be the t-th iterate of the algorithm running on either of the two instances, (159) gives that

$$\max_{k \in \{1,2\}} \mathbb{E}\|\nabla f^{(k)}(x^t)\|^2 = \max_{k \in \{1,2\}} \|\nabla f^{(k)}(x^t)\|^2 \geq \frac{\delta^2 c^2}{4}. \tag{160}$$

where the first equality is because there is no randomness in these two instances. Further, since

$$\mathbb{E}\|\nabla f^{(1)}(x^t)\|^2 + \mathbb{E}\|\nabla f^{(2)}(x^t)\|^2 \geq \max_{k\in\{1,2\}} \mathbb{E}\|\nabla f^{(k)}(x^t)\|^2 \geq \frac{\delta^2 c^2}{4}, \tag{161}$$

we have

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla f^{(1)}(x^t)\|^2 + \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla f^{(2)}(x^t)\|^2 \\ =&\frac{1}{T}\sum_{t=1}^{T} \left(\mathbb{E}\|\nabla f^{(1)}(x^t)\|^2 + \mathbb{E}\|\nabla f^{(2)}(x^t)\|^2\right) \\ \geq&\frac{\delta^2 c^2}{4}. \end{aligned} \tag{162}$$

With the fact that

$$2\max_{k\in\{1,2\}} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla f^{(k)}(x^t)\|^2 \geq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla f^{(1)}(x^t)\|^2 + \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla f^{(2)}(x^t)\|^2 \geq \frac{\delta^2 c^2}{4}, \tag{163}$$

we have

$$\max_{k\in\{1,2\}} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla f^{(k)}(x^t)\|^2 \geq \frac{\delta^2 c^2}{8}. \tag{164}$$

Choosing R regular local costs $\{f_w(x) = f(x; \hat{b}^{(w,k)}) : w \leq R\}$ and $W - R$ poisoned local costs $\{\tilde{f}_w(x) = f(x; \hat{b}^{(w,k)}) : w > R\}$ where $k = \arg\max_{k\in\{1,2\}} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla f^{(k)}(x^t)\|^2$, (164) gives that

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla f(x^t)\|^2 \geq \frac{\delta^2 c^2}{8} = \frac{\delta^2 \min\{A^2, \xi^2\}}{8}, \tag{165}$$

which concludes the proof. ■

The proof is motivated by those of Karimireddy et al. (2022, Theorem 3) and Allouah et al. (2023, Proposition 1). The major difference is that we consider label poisoning attacks in which the poisoned workers can only poison their local labels, while Karimireddy et al. (2022) and Allouah et al. (2023) consider Byzantine attacks in which the Byzantine workers can behave arbitrarily. Different types of attacks lead to different constructions of the instances in the proofs.

Appendix F. Impacts of Heterogeneity and Attack Strengths

To investigate the impacts of heterogeneity of data distributions and strengths of label poisoning attacks, we have conducted numerical experiments by varying the data distributions and the levels of label poisoning attacks, and presented the best classification accuracies in Figure 7. Here, we provide all classification accuracies in Table 3 to complement Figure 7.

β	p	Mean	CC	FABA	LFighter	TriMean
5	0.0	**0.9441**	0.9385	0.9410	0.9420	0.9429
	0.2	0.9426	0.9405	**0.9439**	0.9437	0.9425
	0.4	0.9443	0.9421	0.9437	**0.9456**	0.9430
	0.6	0.9427	0.9402	0.9431	**0.9439**	0.9397
	0.8	0.9429	0.9408	0.9424	**0.9443**	0.9386
	1.0	0.9415	0.9382	0.9423	**0.9437**	0.9371
1	0.0	**0.9437**	0.9362	0.9390	0.9361	0.9402
	0.2	**0.9448**	0.9371	0.9417	0.9341	0.9373
	0.4	0.9417	0.9404	**0.9447**	0.9404	0.9396
	0.6	0.9386	0.9355	0.9409	**0.9422**	0.9365
	0.8	0.9402	0.9323	0.9386	**0.9431**	0.9318
	1.0	0.9424	0.9401	0.9414	**0.9433**	0.9273
0.1	0.0	**0.9407**	0.9251	0.9404	0.9170	0.9134
	0.2	**0.9423**	0.9292	0.9271	0.9313	0.9076
	0.4	**0.9420**	0.9270	0.9229	0.9201	0.8942
	0.6	0.9372	0.9226	0.8996	**0.9377**	0.8891
	0.8	0.9278	0.9135	0.9431	**0.9433**	0.8498
	1.0	0.8327	0.8305	0.9426	**0.9449**	0.8054
0.05	0.0	**0.9468**	0.9317	0.8976	0.8617	0.8763
	0.2	**0.9467**	0.9311	0.8603	0.8845	0.8529
	0.4	**0.9474**	0.9279	0.8601	0.8860	0.8526
	0.6	**0.9418**	0.9248	0.8571	0.9343	0.8492
	0.8	0.9201	0.9155	**0.9394**	0.9385	0.8576
	1.0	0.8573	0.8950	**0.9384**	0.9374	0.8106
0.03	0.0	**0.9426**	0.9299	0.8634	0.8517	0.8523
	0.2	**0.9413**	0.9298	0.8507	0.8493	0.8427
	0.4	**0.9393**	0.9268	0.8506	0.8502	0.8370
	0.6	**0.9374**	0.9206	0.8481	0.8449	0.8246
	0.8	**0.9159**	0.8679	0.8019	0.8313	0.7869
	1.0	0.8312	0.8152	0.7437	**0.9396**	0.7514
0.01	0.0	**0.9456**	0.9164	0.8678	0.8681	0.8008
	0.2	**0.9441**	0.9165	0.8657	0.8660	0.7788
	0.4	**0.9415**	0.9135	0.8631	0.8632	0.7858
	0.6	**0.9356**	0.9039	0.8601	0.8399	0.7708
	0.8	**0.8827**	0.8750	0.8155	0.7864	0.7457
	1.0	**0.8505**	0.8278	0.7731	0.8162	0.7258

Table 3: Accuracies of trained two-layer perceptrons by all aggregators on the MNIST dataset under static label flipping attacks. The hyper-parameter β that characterizes the heterogeneity is in $[5, 1, 0.1, 0.05, 0.03, 0.01]$ and the flipping probability p that characterizes the attack strength is in $[0.0, 0.2, 0.4, 0.6, 0.8, 1.0]$.

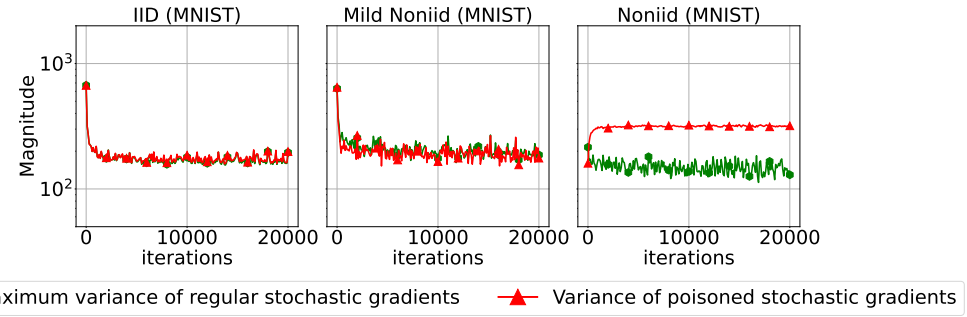

Figure 8: Maximum variance of regular stochastic gradients and variance of poisoned stochastic gradients in softmax regression on the MNIST dataset under static label flipping attacks.

Appendix G. Bounded Variance of Stochastic Gradients

Now we verify the reasonableness of Assumption 4. We consider softmax regression on the MNIST dataset and setup $W = 10$ workers where $R = 9$ workers are regular and the remaining one is poisoned. We compute the variances of regular stochastic gradients and poisoned stochastic gradients under static label flipping attacks in the i.i.d., mild non-i.i.d. and non-i.i.d. cases, and present the maximum variance of regular stochastic gradients and poisoned stochastic gradient in Figure 8. As depicted in Figure 8, the variances of regular stochastic gradients and poisoned stochastic gradients are both bounded under static label flipping attacks, which validates Assumption 4.

Appendix H. Impact of Fraction of Poisoned Workers

To further investigate the impact of different fractions of poisoned workers, here we vary the number of poisoned workers. We setup $W = 10$ workers, among which $R = 8$ ($R = 7$) are regular, while the remaining 2 (3) are poisoned. The experimental settings are the same as those in the nonconvex case. We use a step size of $\gamma = 0.01$ and a momentum coefficient of $\alpha = 0.1$. As shown in Figures 9, 10, 11, and 12, the mean aggregator generally outperforms the robust aggregators in the non-i.i.d. case, which is consistent with the experimental results of $R = 9$. Additionally, from Figures 4 and 5 to Figures 9, 10, 11 and 12, we observe that as the fraction of poisoned workers increases, the performance of both the mean aggregator and robust aggregators decreases. This aligns with the theoretical results in Theorems 7 and 8.

References

Youssef Allouah, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafaël Pinot, and John Stephan. Fixing by mixing: a recipe for optimal Byzantine ML under heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pages 1232–1300, 2023.

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1):165–214, 2023.

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*,

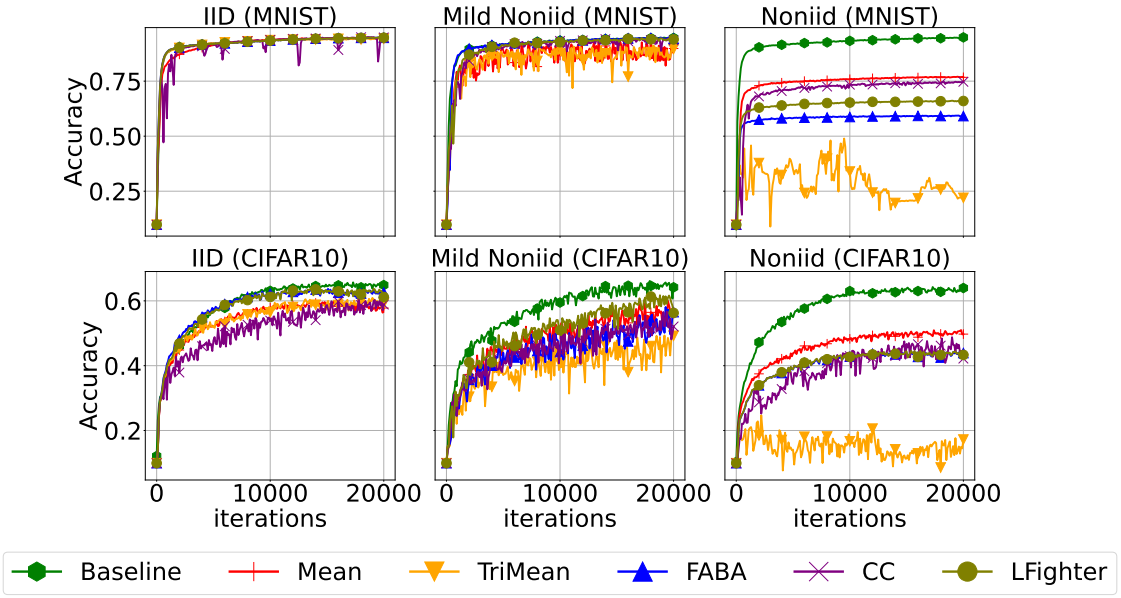

Figure 9: Accuracies of two-layer perceptrons on the MNIST dataset and convolutional neural networks on the CIFAR10 dataset under static label flipping attacks when $R = 8$.

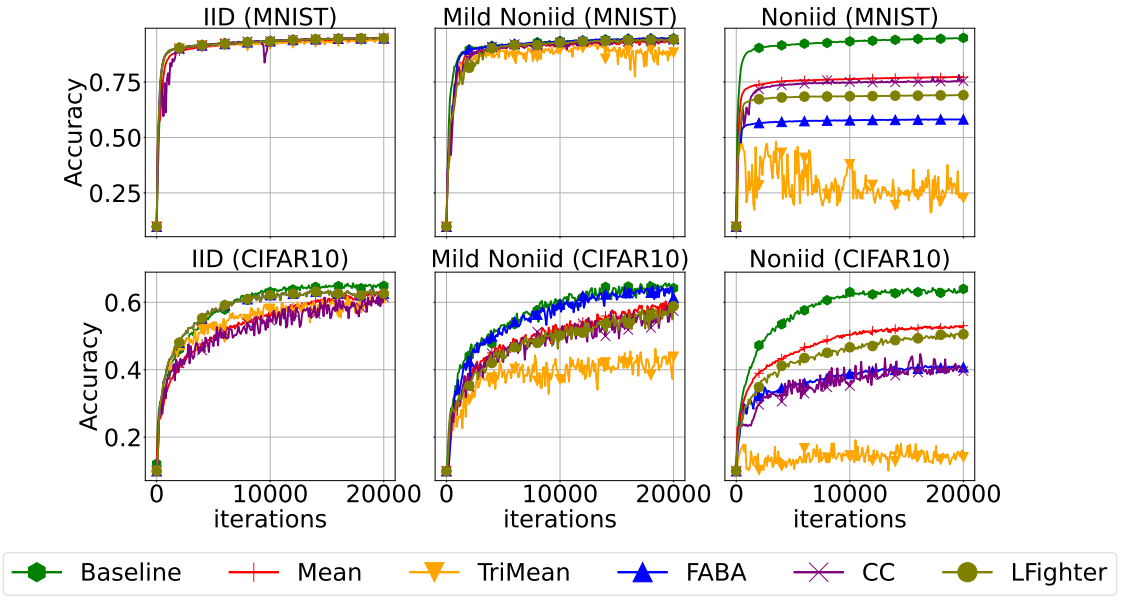

Figure 10: Accuracies of two-layer perceptrons on the MNIST dataset and convolutional neural networks on the CIFAR10 dataset under dynamic label flipping attacks when $R = 8$.

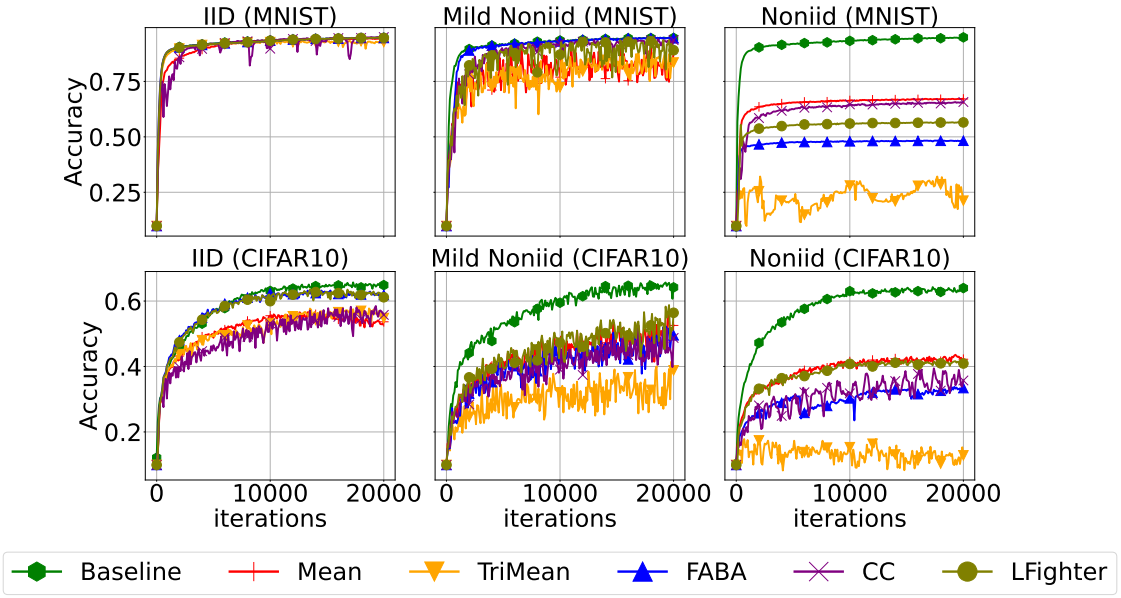

Figure 11: Accuracies of two-layer perceptrons on the MNIST dataset and convolutional neural networks on the CIFAR10 dataset under static label flipping attacks when $R = 7$.

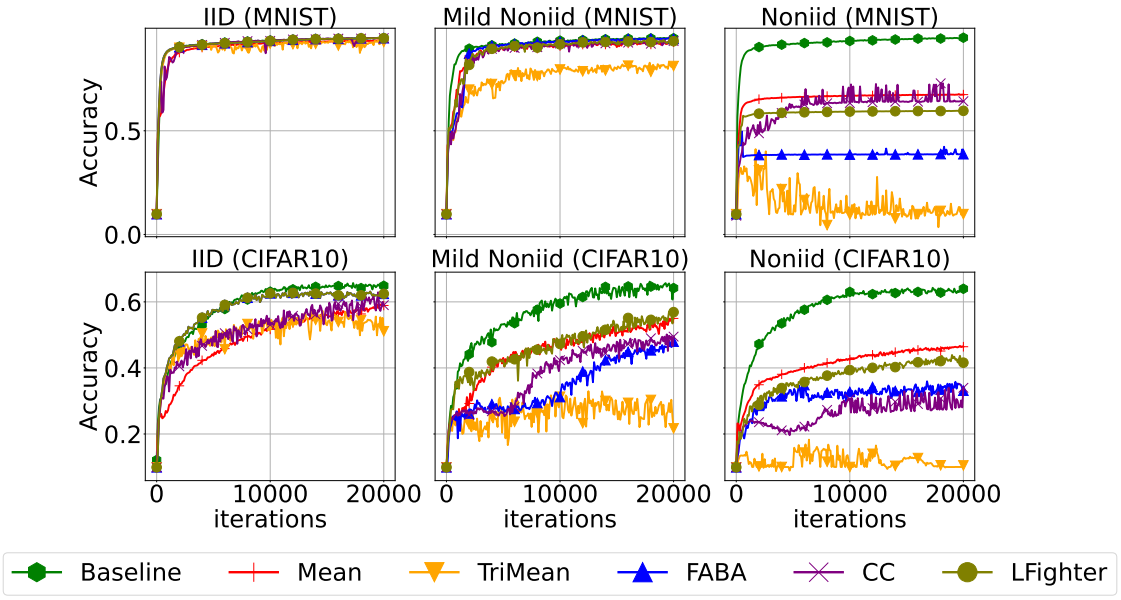

Figure 12: Accuracies of two-layer perceptrons on the MNIST dataset and convolutional neural networks on the CIFAR10 dataset under dynamic label flipping attacks when $R = 7$.

pages 2938–2948, 2020.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, pages 118–128, 2017.

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

Antonio Emanuele Cinà, Kathrin Grosse, Ambra Demontis, Battista Biggio, Fabio Roli, and Marcello Pelillo. Machine learning security against data poisoning: Are we there yet? *Computer*, 57 (3):26–34, 2024.

Xingrong Dong, Zhaoxian Wu, Qing Ling, and Zhi Tian. Byzantine-robust distributed online learning: taming adversarial participants in an adversarial environment. *IEEE Transactions on Signal Processing*, 72:235–248, 2024.

Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Gong. Local model poisoning attacks to Byzantine-robust federated learning. In *USENIX Security Symposium*, pages 1605–1622, 2020.

Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, and John Stephan. Byzantine machine learning made easy by resilient averaging of momentums. In *International Conference on Machine Learning*, pages 6246–6283, 2022.

Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, and Rafael Pinot. On the relevance of byzantine robust optimization against data poisoning. *arXiv preprint arXiv:2405.00491*, 2024.

Yann Fraboni, Richard Vidal, Laetitia Kameni, and Marco Lorenzi. A general theory for federated optimization with asynchronous and heterogeneous clients updates. *Journal of Machine Learning Research*, 24(110):1–43, 2023.

Eduard Gorbunov, Samuel Horváth, Peter Richtárik, and Gauthier Gidel. Variance reduction is an antidote to Byzantines: better rates, weaker assumptions and communication compression as a cherry on the top. In *International Conference on Learning Representations*, 2022.

Rémi Gosselin, Loïc Vieu, Faiza Loukil, and Alexandre Benoit. Privacy and security in federated learning: a survey. *Applied Sciences*, 12(19):9901, 2022.

Rachid Guerraoui, Nirupam Gupta, and Rafael Pinot. Byzantine machine learning: A primer. *ACM Computing Surveys*, 56(7):1–39, 2024.

Lie He, Sai Praneeth Karimireddy, and Martin Jaggi. Byzantine-robust decentralized learning via clippedgossip. *arXiv preprint arXiv:2202.01545*, 2022.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Najeeb Moharram Jebreel and Josep Domingo-Ferrer. Fl-defender: combating targeted attacks in federated learning. *Knowledge-Based Systems*, 260:110178, 2023.

Najeeb Moharram Jebreel, Josep Domingo-Ferrer, David Sánchez, and Alberto Blanco-Justicia. Lfighter: defending against the label-flipping attack in federated learning. *Neural Networks*, 170: 111–126, 2024.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G.L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021.

Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Learning from history for Byzantine robust optimization. In *International Conference on Machine Learning*, pages 5311–5319, 2021.

Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via bucketing. In *International Conference on Learning Representations*, 2022.

Prashant Khanduri, Saikiran Bulusu, Pranay Sharma, and Pramod K. Varshney. Byzantine resilient non-convex svrg with distributed batch gradient computations. *arXiv preprint arXiv:1912.04531*, 2019.

Leslie Lamport, Robert Shostak, and Marshall Pease. The Byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, 4(3):382–401, 1982.

Cody Lewis, Vijay Varadharajan, and Nasimul Noman. Attacks against federated learning defense systems and their mitigation. *Journal of Machine Learning Research*, 24(30):1–50, 2023.

Boyue Li, Shicong Cen, Yuxin Chen, and Yuejie Chi. Communication-efficient distributed optimization in networks with gradient tracking and variance reduction. *Journal of Machine Learning Research*, 21(180):1–51, 2020.

Liping Li, Wei Xu, Tianyi Chen, Georgios B. Giannakis, and Qing Ling. RSA: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets. In *AAAI Conference on Artificial Intelligence*, pages 1544–1551, 2019.

Jing Lin, Long Dang, Mohamed Rahouti, and Kaiqi Xiong. ML attack models: adversarial attacks and data poisoning attacks. *arXiv preprint arXiv:2112.02797*, 2021.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: defending against backdooring attacks on deep neural networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses*, pages 273–294, 2018.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282, 2017.

Jie Peng, Weiyu Li, and Qing Ling. Byzantine-robust decentralized stochastic optimization over static and time-varying networks. *Signal Processing*, 183:108020, 2021.

Jie Peng, Zhaoxian Wu, Qing Ling, and Tianyi Chen. Byzantine-robust variance-reduced federated learning over distributed non-iid data. *Information Sciences*, 616:367–391, 2022.

Jie Peng, Weiyu Li, and Qing Ling. Mean aggregator is more robust than robust aggregators under label poisoning attacks. In *International Joint Conference on Artificial Intelligence*, pages 4797–4805, 2024.

Ahmad Rammal, Kaja Gruntkowska, Nikita Fedin, Eduard Gorbunov, and Peter Richtárik. Communication compression for byzantine robust learning: New efficient algorithms and improved rates. In *International Conference on Artificial Intelligence and Statistics*, pages 1207–1215, 2024.

Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and Zico Kolter. Certified robustness to label-flipping attacks via randomized smoothing. In *International Conference on Machine Learning*, pages 8230–8241, 2020.

Virat Shejwalkar, Amir Houmansadr, Peter Kairouz, and Daniel Ramage. Back to the drawing board: a critical evaluation of poisoning attacks on production federated learning. In *IEEE Symposium on Security and Privacy*, pages 1354–1371, 2022.

Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks. In *Advances in Neural Information Processing Systems*, pages 3520–3532, 2017.

Ziteng Sun, Peter Kairouz, Ananda Theertha Suresh, and H Brendan McMahan. Can you really backdoor federated learning? *arXiv preprint arXiv:1911.07963*, 2019.

Pooya Tavallali, Vahid Behzadan, Azar Alizadeh, Aditya Ranganath, and Mukesh Singhal. Adversarial label-poisoning attacks and defense for general multi-class models based on synthetic reduced nearest neighbor. In *International Conference on Image Processing*, pages 3717–3722, 2022.

Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu. Data poisoning attacks against federated learning systems. In *European Symposium on Research in Computer Security*, pages 480–501, 2020.

Jiyuan Tu, Weidong Liu, Xiaojun Mao, and Xi Chen. Variance reduced median-of-means estimator for byzantine-robust distributed inference. *Journal of Machine Learning Research*, 22(84):1–67, 2021.

Joost Verbraeken, Matthijs Wolting, Jonathan Katzy, Jeroen Kloppenburg, Tim Verbelen, and Jan S. Rellermeyer. A survey on distributed machine learning. *Acm Computing Surveys*, 53(2):1–33, 2020.

Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the tails: yes, you really can backdoor federated learning. In *Advances in Neural Information Processing Systems*, pages 16070–16084, 2020.

Zhaoxian Wu, Qing Ling, Tianyi Chen, and Georgios B. Giannakis. Federated variance-reduced stochastic gradient descent with robustness to Byzantine attacks. *IEEE Transactions on Signal Processing*, 68:4583–4596, 2020.

Zhaoxian Wu, Tianyi Chen, and Qing Ling. Byzantine-resilient decentralized stochastic optimization with robust aggregation rules. *IEEE Transactions on Signal Processing*, 71:3179–3195, 2023.

Qi Xia, Zeyi Tao, Zijiang Hao, and Qun Li. FABA: an algorithm for fast aggregation against Byzantine attacks in distributed neural networks. In *International Joint Conferences on Artificial Intelligence*, 2019.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2):1–19, 2019.

Yi-Rui Yang and Wu-Jun Li. Buffered asynchronous sgd for byzantine learning. *Journal of Machine Learning Research*, 24(204):1–62, 2023.

Mang Ye, Xiuwen Fang, Bo Du, Pong C. Yuen, and Dacheng Tao. Heterogeneous federated learning: state-of-the-art and research challenges. *ACM Computing Surveys*, 56(3):1–44, 2023.

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