FL-WBC: Enhancing Robustness against Model Poisoning Attacks in Federated Learning from a Client Perspective

Jingwei Sun[1], **Ang Li**[1], **Louis DiValentin**[2], **Amin Hassanzadeh**[2],
Yiran Chen[1], **Hai Li**[1]
[1]Department of Electrical and Computer Engineering, Duke University
[2]Security R&D, Accenture Labs, Accenture
[1]{jingwei.sun, ang.li630, yiran.chen, hai.li}@duke.edu,
[2]{louis.divalentin, amin.hassanzadeh}@accenture.com

Abstract

Federated learning (FL) is a popular distributed learning framework that trains a global model through iterative communications between a central server and edge devices. Recent works have demonstrated that FL is vulnerable to model poisoning attacks. Several server-based defense approaches (e.g. robust aggregation) have been proposed to mitigate such attacks. However, we empirically show that under extremely strong attacks, these defensive methods fail to guarantee the robustness of FL. More importantly, we observe that as long as the global model is polluted, the impact of attacks on the global model will remain in subsequent rounds even if there are no subsequent attacks. In this work, we propose a client-based defense, named *White Blood Cell for Federated Learning (FL-WBC)*, which can mitigate model poisoning attacks that have already polluted the global model. The key idea of FL-WBC is to identify the parameter space where long-lasting attack effect on parameters resides and perturb that space during local training. Furthermore, we derive a certified robustness guarantee against model poisoning attacks and a convergence guarantee to FedAvg after applying our FL-WBC. We conduct experiments on FasionMNIST and CIFAR10 to evaluate the defense against state-of-the-art model poisoning attacks. The results demonstrate that our method can effectively mitigate model poisoning attack impact on the global model within 5 communication rounds with nearly no accuracy drop under both IID and non-IID settings. Our defense is also complementary to existing server-based robust aggregation approaches and can further improve the robustness of FL under extremely strong attacks. Our code can be found at https://github.com/jeremy313/FL-WBC.

1 Introduction

Federated learning (FL) [1, 2] is a popular distributed learning approach that enables a number of edge devices to train a shared model in a federated fashion without transferring their local training data. However, recent works [3–12] show that it is easy for edge devices to conduct model poisoning attacks by manipulating local training process to pollute the global model through aggregation.

Depending on the adversarial goals, model poisoning attacks can be classified as *untargeted model poisoning attacks* [3–6], which aim to make the global model indiscriminately have a high error rate on any test input, or *targeted model poisoning attacks* [7–12], where the goal is to make the global model generate attacker-desired misclassifications for some particular test samples. Our work focuses

Preprint. Under review.

on the *targeted model poisoning attacks* introduced in [11, 12]. In this attack, malicious devices share a set of data points with dirty labels, and the adversarial goal is to make the global model output the same dirty labels given this set of data as inputs. Our work can be easily extended to many other model poisoning attacks (e.g., backdoor attacks), which shall be discussed in §4.

Several studies have been done to improve the robustness of FL against model poisoning attacks through robust aggregations [13–17], clipping local updates [7] and leveraging the noisy perturbation [7]. These defensive methods focus on only preventing the global model from being polluted by model poisoning attacks during the aggregation. However, we empirically show that these server-based defenses fail to guarantee the robustness when attacks are extremely strong. More importantly, we observe that as long as the global model is polluted, the impact of attacks on the global model will remain in subsequent rounds even if there are no subsequent attacks, and can not be mitigated by these server-based defenses. Therefore, an additional defense is needed to mitigate the poisoning attacks that cannot be eliminated by robust aggregation and will pollute the global model, which is the goal of this paper.

To achieve this goal, we first propose a quantitative estimator named *Attack Effect on Parameter (AEP)*. It estimates the effect of model poisoning attacks on global model parameters and infers information about the susceptibility of different instantiations of FL to model poisoning attacks. With our quantitative estimator, we explicitly show the long-lasting attack effect on the global model. Based on our analysis, we design a client-based defense named *White Blood Cell for Federated Learning (FL-WBC)*, as shown in Figure 1, which can mitigate the model poisoning attacks that have already polluted the global model. FL-WBC differs from previous server-based defenses in mitigating the model poisoning attack that has already broken through the server-based defenses and polluted the global model. Thus, our client-based defense is complementary to current server-based defense and enhances the robustness of FL against the model poisoning attack, especially against the extremely strong attacks that can not be mitigated during the aggregation. We evaluate our defense on Fashion-MNIST [18] and CIFAR10 [19] against the model poisoning attack [11] under IID (identically independently distributed) and non-IID settings. The results demonstrate that FL-WBC can effectively mitigate the attack effect on the global model in 1 communication round with nearly no accuracy drop under IID settings, and within 5 communication rounds for non-IID settings, respectively. We also conduct experiments by integrating the robust aggregation with FL-WBC. The results show that even though the robust aggregation is ineffective under extremely strong attacks, the attack can still be efficiently mitigated by applying FL-WBC.

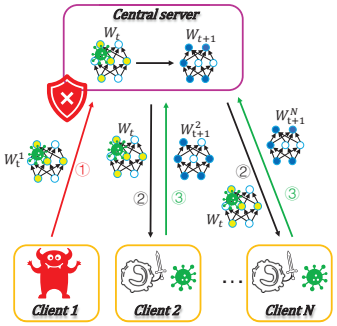


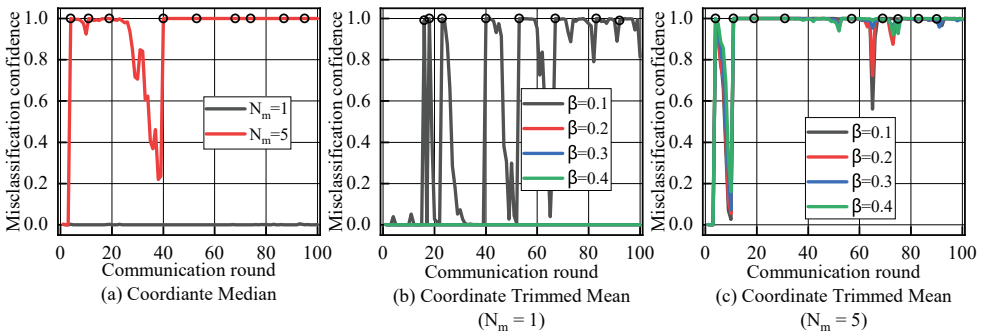
Figure 1: Overview of FL-WBC.

Our key contributions are summarized as follows:

- To the best of our knowledge, this is the first work to quantitatively assess the effect of model poisoning attack on the global model in FL. Based on our proposed estimator, we reveal the reason for the long-lasting effect of a model poisoning attack on the global model.
- We design a defense, which is also the first defense to the best of our knowledge, to effectively mitigate a model poisoning attack that has already polluted the global model. We also derive a robustness guarantee in terms of AEP and a convergence guarantee to FedAvg when applying our defense.
- We evaluate our defense on Fashion-MNIST and CIFAR10 against state-of-the-art model poisoning attacks. The results show that our proposed defense can enhance the robustness of FL in an effective and efficient way, i.e., our defense defends against the attack in fewer communication rounds with less model utility degradation.

2 Related work

Model poisoning attacks in FL Model poisoning attack can be *untargeted* [3–6] or *targeted* [7–12]. *Untargeted model poisoning attacks* aim to minimize the accuracy of the global model indiscriminately for any test input. For *targeted model poisoning attacks*, the malicious goal is to make the global model misclassify the particular test examples as the attacker-desired target class



(a) Coordiante Median (b) Coordinate Trimmed Mean ($N_m = 1$) (c) Coordinate Trimmed Mean ($N_m = 5$)

Figure 2: Defense performance of Coordinate Median aggregation and Coordinate Trimmed Mean aggregation. The black circle denotes the adversarial rounds for all the strategies in the figure.

in its prediction. An adversary using this approach can implant hidden backdoors into the global model so that the images with a trojan trigger will be classified as attacker-desired labels, known as a backdoor attack [7–10]. Another type of targeted model poisoning attack is introduced in [11, 12], which aims to fool the global model to produce adversarial misclassification on a set of chosen inputs with high confidence. Our work focuses on the targeted model poisoning attacks in [11, 12].

Mitigate model poisoning attacks in FL A number of robust aggregation approaches have been proposed to mitigate data poisoning attacks while retaining the performance of FL. One typical approach is to detect and down-weight the malicious client's updates on the central server side [13–16], thus the attack effects on training performance can be diminished. The central server calculates coordinate-wise median or coordinate-wise trimmed mean for local model updates before performing aggregation [13]. Similarly, [14] suggests applying geometric median to local updates that are uploaded to the server. Meanwhile, some heuristic-based aggregation rules [20, 21, 3, 22, 23] have been proposed to cluster participating clients into a benign group and a malicious group, and then perform aggregation on the benign group only. FoolsGold [20] assumes that benign clients can be distinguished from attackers by observing the similarity between malicious clients' gradient updates, but Krum [21, 3] utilizes the similarity of benign clients' local updates instead. In addition, [7, 24] show that applying differential privacy to the aggregated global model can improve the robustness against model poisoning attacks. All these defensive methods are deployed at the server side and their goals are to mitigate model poisoning attacks during aggregation. Unfortunately, often in extreme cases (e.g. attackers occupy a large proportion of total clients), existing robust aggregation methods fail to prevent the aggregation from being polluted by the malicious local updates showing that it is not sufficient to offer defense via aggregation solely. Thus, there is an urgent necessity to design a novel local training method in FL to enhance its robustness against model poisoning attacks at the client side, which is complementary to existing robust aggregation approaches.

3 Motivation

Although current server-based defense approaches can defend against model poisoning attacks under most regular settings, it is not clear whether their robustness can still be guaranteed under extremely strong attacks, i.e., with significantly larger numbers of malicious devices involved in training. To investigate the robustness of current methods under such challenging but practical settings, we evaluate Coordinate Median aggregation (CMA) and Coordinate Trimmed Mean aggregation (CTMA) [13] on the model poisoning attack with Fashion-MNIST dataset, which is performed by following the settings in [11]. The goal of the attacks is to make the global model misclassify some specified data samples as target classes. In this experiment, we denote a communication round as an adversarial round t_{adv} when malicious devices participate in the training, and N_m malicious devices would participate in training at adversarial rounds. We assume that there are 10 devices involved in training in each round, but increase N_m from 1 to 5 to vary the strength of the attacks. We conduct experiments under IID setting and the training data is uniformly distributed to 100 devices. The model architecture can be found in Table 3. For training, we set local epoch E as 1 and batch size B as 32. We apply SGD optimizer and set the learning rate η to 0.01. The results of confidence that the global model would miss-classify the poisoning data point are shown in Figure 2.

The results show that the effectiveness of both CMA and CTMA dramatically degrades when there are 50% of malicious devices in the adversarial rounds. It is worthy noting that the attack impact on model performance will remain for subsequent rounds even if no additional attacks occur. We observe the same phenomenon in alternative robust aggregation approaches, and more detailed results are presented in §7. Therefore, in order to build a more robust FL system, it is necessary to instantly mitigate the impact of model poisoning attack as long as the global model is polluted by malicious devices. This has motivated us to design FL-WBC to ensure sufficient robustness of FL even under extremely strong attacks.

4 Model Poisoning Attack in FL

To better understand the impact of model poisoning attacks in FL scenarios, we first need to theoretically analyze how the poisoning attack affects the learning process and provide a mathematical estimation to quantitatively assess the attack effect on model parameters. During this process we come to a deeper understanding of the reasons for the persistence of the attack effect observed in §3. Without loss of generality, we employ FedAvg [1], the most widely applied FL algorithm as the representative FL method throughout this paper.

4.1 Problem Formulation

The learning objective of FedAvg is defined as:

$$\boldsymbol{W} = \min_{\boldsymbol{W}} \{F(\boldsymbol{W}) \triangleq \sum_{k=1}^{N} p^k F^k(\boldsymbol{W})\}, \tag{1}$$

where $\boldsymbol{W}$ is the weights of the global model, N represents the number of devices, F^k is the local objective of the k-th device, p^k is the weight of the k-th device, $p^k \geq 0$ and $\sum_{k=1}^{N} p^k = 1$.

Equation 1 is solved in an iterative device-server communication fashion. For a given communication round (e.g. the t-th), the central server first randomly selects K devices to compose a set of participating devices $\mathbb{S}_t$ and then broadcasts the latest global model $\boldsymbol{W}_{t-1}$ to these devices. Afterwards, each device (e.g. the k-th) in $\mathbb{S}_t$ performs I iterations of local training using their local data. However, the benign devices and malicious devices perform the local training in different manners. Specifically, if the k-th device is benign, in each iteration (e.g. the i-th), the local model $\boldsymbol{W}_{t,i}^k$ on the k-th device is updated following:

$$\boldsymbol{W}_{t,i+1}^k \leftarrow \boldsymbol{W}_{t,i}^k - \eta_{t,i} \nabla F^k(\boldsymbol{W}_{t,i}^k, \xi_{t,i}^k), \tag{2}$$

where $\eta_{t,i}$ is the learning rate, $\xi_{t,i}^k$ is a batch of data samples uniformly chosen from the k-th device and $\boldsymbol{W}_{t,0}^k$ is initialized as $\boldsymbol{W}_{t-1}$. In contrast, if the k-th device is malicious, the local model $\boldsymbol{W}_{t,i}^k$ is updated according to:

$$\boldsymbol{W}_{t,i+1}^k \leftarrow \boldsymbol{W}_{t,i}^k - \eta_{t,i}[\alpha \nabla F^k(\boldsymbol{W}_{t,i}^k, \xi_{t,i}^k) + (1-\alpha) \nabla F_M(\boldsymbol{W}_{t,i}^k, \pi_{t,i})], \tag{3}$$

where F_M is the malicious objective shared by all the malicious devices. D_M is a malicious dataset that consists of the data samples following the same distribution as the benign training data but with adversarial data labels. All the malicious devices share the same malicious dataset D_M and $\pi_{t,i}$ is a batch of data samples from D_M used to optimize the malicious objective. Except that they share a malicious dataset, the malicious attackers have the same background knowledge as the benign clients. The goal of the attackers is to make the global model achieve a good performance on the malicious objective (i.e. targeted misclassification on D_M). Considering the obliviousness of attack, the malicious devices also optimize benign objective, and the trade-off between the benign and malicious objectives is controlled by α, where $\alpha \in [0, 1]$. Finally, the server averages the local models of the selected K devices and updates the global model as follows:

$$\boldsymbol{W}_t \leftarrow \frac{N}{K} \sum_{k \in \mathbb{S}_t} p^k \boldsymbol{W}_{t,I}^k. \tag{4}$$

4.2 Estimation of Attack Effect on Model Parameters

Based on the above formulated training process, we analyze the impact of poisoning attacks on model parameters. To this end, we denote the set of attackers as $\mathbb{M}$, and introduce a new notation

$\boldsymbol{W}_t(\mathbb{S}_i \setminus \mathbb{M})$, which represents the global model weights in the t-th round when all malicious devices in $\mathbb{S}_i (i \leq t)$ do not perform the attack in the i-th training round. Specifically, when $i = t$, $\boldsymbol{W}_t(\mathbb{S}_t \setminus \mathbb{M})$ is optimized following:

$$\boldsymbol{W}_t(\mathbb{S}_t \setminus \mathbb{M}) \leftarrow \frac{N}{K} \sum_{k \in \mathbb{S}_t} p^k \boldsymbol{W}_{t,I}^k (\alpha = 1), \tag{5}$$

where $\boldsymbol{W}_{t,I}^k(\alpha = 1)$ indicates that $\boldsymbol{W}_{t,I}^k$ is trained using Equation 3 with setting $\alpha = 1$ (i.e., the k-th device is benign). A special case is $\boldsymbol{W}_t(\mathbb{S} \setminus \mathbb{M})$, which means the global model is optimized when all the malicious devices do not conduct attacks before the t-th round. To quantify the attack effect on the global model, we define the Attack Effect on Parameter (AEP) as follows:

Definition 1. *Attack Effect on Parameter (AEP), which is denoted as δ_t, is the change of the global model parameters accumulated until t-th round due to the attack conducted by the malicious devices in the FL system:*

$$\delta_t \triangleq \boldsymbol{W}_t(\mathbb{S} \setminus \mathbb{M}) - \boldsymbol{W}_t. \tag{6}$$

Based on AEP, we can quantitatively evaluate the attack effect on the malicious objective using $F_M(\boldsymbol{W}_t(\mathbb{S}\setminus\mathbb{M}) - \delta_t) - F_M(\boldsymbol{W}_t(\mathbb{S}\setminus\mathbb{M}))$. As Figure 2 illustrates, although $\boldsymbol{W}_t(\mathbb{S}\setminus\mathbb{M})$ keeps updating after an adversarial round and there are no more attacks before the next adversarial round, the attack effect on the global model, i.e., F_M , remains for a number of rounds. Based on such an observation, we assume that the optimization of malicious objective is dominated by δ_t compared to $\boldsymbol{W}_t(\mathbb{S} \setminus \mathbb{M})$, which is learned from the benign objective. Consequently, if the attack effect in round τ remains for further rounds, $\|\delta_{t+1} - \delta_t\|$ should be small for $t \geq \tau$.

To analyze why the attack effect can persist in the global model, we consider the scenario where the malicious devices are selected in round τ_1 and τ_2, but will not be selected between these two rounds. We derive an estimator of δ_t for $\tau_1 < t < \tau_2$, denoted as $\hat{\delta}_t$:

$$\hat{\delta}_t = \frac{N}{K} [\sum_{k \in \mathbb{S}_t} p^k \prod_{i=0}^{I-1} (\boldsymbol{I} - \eta_{t,i} \boldsymbol{H}_{t,i}^k)] \hat{\delta}_{t-1}, \tag{7}$$

where $\boldsymbol{H}_{t,i}^k \triangleq \nabla^2 F^k(\boldsymbol{W}_{t,i}^k, \xi_{t,i}^k)$. The derivation process is presented in Appendix D. Note that, we do not restrict the detailed malicious objective during derivation, and thus our estimator and analysis can be extended to other attacks, such as backdoor attacks.

4.3 Unveil Long-lasting Attack Effect

The key observation from Equation 7 is that if $\hat{\delta}_\tau$ is in the kernel of each $\boldsymbol{H}_{t,i}^k$ for i-th iteration where $k \in \mathbb{S}_t$ and $t > \tau$, then $\hat{\delta}_t$ will be the same as $\hat{\delta}_\tau$, which keeps AEP in the global model. Based on this observation, we discover that **the reason why attack effects remain in the aggregated model is that the AEPs reside in the kernel of $\boldsymbol{H}_{t,i}^k$.** To validate our analysis, we conduct experiments on Fashion-MNIST with model poisoning attacks in FL. The experiment details and results are shown in Appendix B. The results show that $\|\boldsymbol{H}_{t,i}^k \delta_t\|_2$ would be nearly 0 under effective attacks. We also implement attack boosting by regularizing δ_t to be in the kernel of $\boldsymbol{H}_{t,i}^k$.

The above theoretical analysis and experiment results suggest that all the server-based defense methods (e.g. robust aggregation) will not be able efficiently mitigate the impact of model poisoning attacks to the victim global model. The fundamental reason for the failure of these mitigations is that: **the transmission of AEP δ_t in global model is determined by $\boldsymbol{H}_{t,i}^k$, which is inaccessible by the central server.** Therefore, it is necessary to design an effective defense mechanism at client side aiming at mitigating attack that has already polluted the global model to further enhance the robustness of FL.

5 FL-WBC

5.1 Defense Design

Our aforementioned analysis shows that AEP resides in the kernels of the Hessian matrices that are generated during the benign devices' local training. In this section, we propose *White Blood Cell*

for Federated Learning (FL-WBC) to efficiently mitigate the attack effect on the global model. In particular, we reform the local model training of benign devices to achieve two goals:

- Goal 1: To maintain the benign task's performance, loss of local benign task should be minimized.
- Goal 2: To prevent AEP from being hidden in the kernels of Hessian matrices on benign devices, the kernel of $\boldsymbol{H}^k_{t+1,i}$ should be perturbed.

It is computationally unaffordable to perform the perturbance on $\boldsymbol{H}^k_{t,i}$ directly due to its high dimension. Therefore, in order to achieve Goal 2, we consider the essence of $\boldsymbol{H}^k_{t,i}$, i.e., second-order partial derivatives of the loss function, where the diagonal elements describe the change of gradients $\nabla F^k(\boldsymbol{W}^k_{t,i+1}) - \nabla F^k(\boldsymbol{W}^k_{t,i})$ across iterations. We assume a fixed learning rate is applied for each communication round, and then $\nabla F^k(\boldsymbol{W}^k_{t,i+1}) - \nabla F^k(\boldsymbol{W}^k_{t,i})$ can be approximated by $(\Delta \boldsymbol{W}^k_{t,i+1} - \Delta \boldsymbol{W}^k_{t,i})/\eta_{t,i}$. In the experiments presented in §4.3, we observe that $\boldsymbol{H}^k_{t,i}$ has more than 60% elements to be zero in the most of iterations. When $\boldsymbol{H}^k_{t,i}$ is highly sparse, we add noise to the small-magnitude elements on its diagonal, which is approximately $(\Delta \boldsymbol{W}^k_{t,i+1} - \Delta \boldsymbol{W}^k_{t,i})/\eta_{t,i}$, to perturb the null space of $\boldsymbol{H}^k_{t,i}$. Formally, we have two steps to optimize $\boldsymbol{W}^k_{t,i+1}$:

$$\hat{\boldsymbol{W}^k_{t,i+1}} = \boldsymbol{W}^k_{t,i} - \eta_{t,i} \nabla F^k(\boldsymbol{W}^k_{t,i}, \xi^k_{t,i}) \tag{8}$$

$$\boldsymbol{W}^k_{t,i+1} = \hat{\boldsymbol{W}^k_{t,i+1}} + \eta_{t,i} \Upsilon^k_{t,i} \odot \boldsymbol{M}^k_{t,i}, \tag{9}$$

where $\Upsilon^k_{t,i}$ is a matrix with the same shape of $\boldsymbol{W}$, and $\boldsymbol{M}^k_{t,i}$ is a binary mask whose elements are determined as:

$$\boldsymbol{M}^k_{t,i_{r,c}} = \begin{cases} 1, |(\hat{\boldsymbol{W}^k_{t,i+1}} - \boldsymbol{W}^k_{t,i}) - \Delta \boldsymbol{W}^k_{t,i}|_{r,c}/\eta_{t,i} \leq |\Upsilon^k_{t,i_{r,c}}| \\ 0, |(\hat{\boldsymbol{W}^k_{t,i+1}} - \boldsymbol{W}^k_{t,i}) - \Delta \boldsymbol{W}^k_{t,i}|_{r,c}/\eta_{t,i} > |\Upsilon^k_{t,i_{r,c}}|, \end{cases} \tag{10}$$

where $\boldsymbol{M}^k_{t,i_{r,c}}$ is the element on the r-th row and c-th column of $\boldsymbol{M}^k_{t,i}$. Conceptually, $\boldsymbol{M}^k_{t,i+1}$ finds the small-magnitude elements on the $\boldsymbol{H}^k_{t,i}$'s diagonal.

Note that we have different choices of $\Upsilon^k_{t,i}$. In this work, we set $\Upsilon^k_{t,i}$ as Laplace noise with $mean = 0$ and $std = s$, since the randomness of $\Upsilon^k_{t,i}$ will make attackers harder to determine the defense strategy. Specifically, our defense is to find the elements in $\hat{\boldsymbol{W}^k_{t,i+1}}$ whose corresponding values in $|(\hat{\boldsymbol{W}^k_{t,i+1}} - \boldsymbol{W}^k_{t,i}) - \Delta \boldsymbol{W}^k_{t,i}|/\eta_{t,i}$ are smaller than the counterparts in $|\Upsilon^k_{t,i}|$. The detailed algorithm describing the local training process on benign devices when applying FL-WBC can be found in Appendix A. We derive a certified robustness guarantee for our defense, which provides a lower bound of distance of AEP between the adversarial round and the subsequent rounds. The detailed theorem of the certified robustness guarantee can be found in Appendix E.

5.2 Robustness to Adaptive attacks

Our defense is robust against adaptive attacks [25, 26] since the attacker cannot know the detailed defensive operations even after conducting the attack for three reasons. First, our defense is performed during the local training at the client side, where the detailed defensive process is closely related to benign clients' data. Such data is inaccessible to the attackers, and hence the attackers cannot figure out the detailed defense process. Second, even if the attackers have access to benign clients' data (which is a super strong assumption and beyond our threat model), the attackers cannot predict which benign clients will be sampled by the server to participate in the next communication round. Third, in the most extreme case where attackers have access to benign clients' data and can predict which clients will be sampled in the next round (which is an unrealistic assumption), the attackers still cannot successfully bypass our defense. The reason is that the defense during the benign local training is mainly dominated by the random matrix $\Upsilon^k_{t,i}$ in Equation 9, which is also unpredictable. With such unpredictability and randomness of our defense, no effective attack can be adapted.

6 Convergence Guarantee

In this section, we derive the convergence guarantee of FedAvg [1]—the most popular FL algorithm, with our proposed FL-WBC. We follow the notations in §4 describing FedAvg, and the only difference after applying FL-WBC is the local training process of benign devices. Specifically, for the t-th round, the local model on the k-th benign device is updated as:

$$\nabla F^{k'}(\boldsymbol{W}_{t,i}^k, \xi_{t,i}^k) = \nabla F^k(\boldsymbol{W}_{t,i}^k, \xi_{t,i}^k) + \mathcal{T}_{t,i} \tag{11}$$

$$\boldsymbol{W}_{t,i+1}^k \leftarrow \boldsymbol{W}_{t,i}^k - \eta_{t,i} \nabla F^{k'}(\boldsymbol{W}_{t,i}^k, \xi_{t,i}^k), \tag{12}$$

where $\mathcal{T}_{t,i}$ is the local updates generated by the perturbance step in Equation 9.

Our convergence analysis is inspired by [27]. Before presenting our theoretical results, we first make the following Assumptions 1-4 same as [27].

Assumption 1. *$F^1, F^2, ..., F^N$ are L-smooth: $\forall \boldsymbol{V}, \boldsymbol{W}$, $F^k(\boldsymbol{V}) \leq F^k(\boldsymbol{W}) + (\boldsymbol{V} - \boldsymbol{W})^T \nabla F^k(\boldsymbol{W}) + \frac{L}{2}||\boldsymbol{V} - \boldsymbol{W}||_2^2$.*

Assumption 2. *$F_1, F_2, ..., F_N$ are μ-strongly convex: $\forall \boldsymbol{V}, \boldsymbol{W}$, $F^k(\boldsymbol{V}) \geq F^k(\boldsymbol{W}) + (\boldsymbol{V} - \boldsymbol{W})^T \nabla F^k(\boldsymbol{W}) + \frac{\mu}{2}||\boldsymbol{V} - \boldsymbol{W}||_2^2$.*

Assumption 3. *Let ξ_t^k be sampled from the k-th device's local data uniformly at random. The variance of stochastic gradients in each device is bounded: $\mathbb{E}||\nabla F^k(\boldsymbol{W}_{t,i}^k, \xi_{t,i}^k) - \nabla F^k(\boldsymbol{W}_{t,i}^k)||^2 \leq \sigma_k^2$ for $k = 1, ..., N$.*

Assumption 4. *The expected squared norm of stochastic gradients is uniformly bounded, i.e., $\mathbb{E}||\nabla F^k(\boldsymbol{W}_{t,i}^k, \xi_{t,i}^k)||^2 \leq G^2$ for all $k = 1, ..., N$, $i = 0, ..., I-1$ and $t = 0, ..., T-1$.*

We define F^* and F^{k*} as the minimum value of F and F^k and let $\Gamma = F^* - \sum_{k=1}^{N} p_k F^{k*}$. We assume each device has I local training iterations in each round and the total number of rounds is T. Then, we have the following convergence guarantee on FedAvg with our defense.

Theorem 1. *Let Assumptions 1-4 hold and L, μ, σ_k, G be defined therein. Choose $\kappa = \frac{L}{\mu}$, $\gamma = \max\{8\kappa, I\}$ and the learning rate $\eta_{t,i} = \frac{2}{\mu(\gamma + tI + i)}$. Then FedAvg with our defense satisfies*

$$\mathbb{E}[F(\boldsymbol{W}_T)] - F^* \leq \frac{2\kappa}{\gamma + TI}\left(\frac{Q + C}{\mu} + \frac{\mu\gamma}{2}\mathbb{E}||\boldsymbol{W}_0 - \boldsymbol{W}^*||^2\right),$$

where

$$Q = \sum_{k=1}^{N} p_k^2(s^2 + \sigma_k^2) + 6L\Gamma + 8(I-1)^2(s^2 + G^2)$$

$$C = \frac{4}{K} I^2 (s^2 + G^2).$$

Proof. See our proof in Appendix F. □

7 Experiments

In our experiments, we evaluate FL-WBC against targeted model poisoning attack [11] described in §4 under both IID and non-IID settings. Experiments are conducted on a server with two Intel Xeon E5-2687W CPUs and four Nvidia TITAN RTX GPUs.

7.1 Experimental Setup

Attack method. We evaluate our defense against model poisoning attack shown in [11, 12]. There are several attackers in FL setup and all the attackers share a malicious dataset D_M, whose data points obey the same distribution with benign training data while having adversarial labels. We let all the attackers conduct the model poisoning attack at adversarial rounds t_{adv} simultaneously such that the attack will be extremely strong.

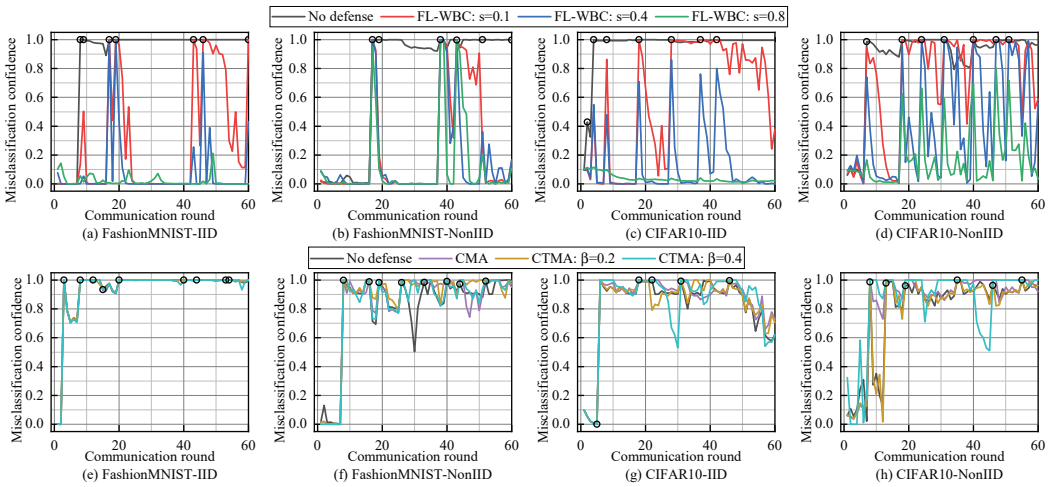


Figure 3: Comparison of misclassification confidence and communication round on FashionMNIST and CIFAR10 with IID/non-IID settings. The black circle denotes the adversarial rounds.

Defense baseline. We compare our proposed defense with two categories of defense methods that have been widely used: (1) **Differential privacy (DP)** improves robustness with theoretical guarantee by clipping the gradient norm and injecting perturbations to the gradients. We adopt both **Central Differential privacy (CDP)** [24] and **Local Differential privacy (LDP)** [24] for comparisons. We set the clipping norm as 5 and 10 for Fashion-MNIST and CIFAR10 respectively following [24] and apply Laplace noise with $mean = 0$ and $std = \sigma_{dp}$. (2) **Robust aggregation** improves robustness of FL by manipulating aggregation rules. We consider both **Coordinate Median Aggregation (CMA)** [13] and **Coordinate Trimmed-Mean Aggregation (CTMA)** [13] as baselines.

Datasets. To evaluate our defense under more realistic FL settings, we construct IID/non-IID datasets based on Fashion-MNIST and CIFAR10 by following the configurations in [1]. The detailed data preparation can be found in Appendix C. We sample 1 and 10 images from both datasets to construct the malicious dataset D_M corresponding to scenarios D_M having single image and multiple images. Note that, data samples in D_M would not appear in training datasets of benign devices.

Hyperparameter configurations. Each communication round is set to be the adversarial round with probability 0.1. In each benign communication round, there are 10 benign devices which are randomly selected to participate in the training. In each adversarial round, 5 malicious and 5 randomly selected benign devices participate in the training, which means there are 50% attackers involved in adversarial rounds. Additional configurations and model structures can be found in Appendix C.

Evaluation metrics. (1) **Attack metric (misclassification confidence/accuracy:)** We define misclssification confidence/accuracy as the classification confidence/accuracy of the global model on the malicious dataset. (2) **Robust metric (attack mitigation rounds):** We define *attack mitigation rounds* as the number of communication rounds after which the misclassification confidence can decrease to lower than 50% or misclassification accuracy can decrease to lower than the error rate for the benign task. (3) **Utility metric (benign accuracy):** We use the accuracy of the global model on the benign testing set of the primary task to measure the effectiveness of FL algorithms (i.e., FedAvg [1]). The higher the accuracy is, the higher utility is obtained.

7.2 Effectiveness of FL-WBC with *Single* Image in The Malicious Dataset

We first show the results when there is only one image in the malicious dataset. We consider IID and non-IID settings for both Fashion-MNIST and CIFAR10 datasets. Figure 3 shows the misclassification confidence of our defense and the robust aggregation baselines in the first 60 communication rounds. The results show that our defense can more effectively and efficiently mitigate the impact of model poisoning attack in comparison with baseline methods. In particular, FL-WBC can mitigate the impact of model poisoning attack within 5 communication round when s (i.e., standard deviation for Υ) is 0.4 for both IID and non-IID settings. With regard to CMA and CTMA, the attack impact can not be mitigated within 10 subsequent rounds even when β for CTMA is 0.4, where 80% of local

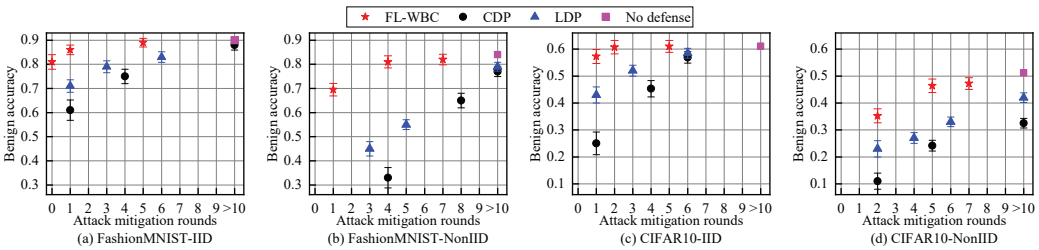


Figure 4: Comparison of benign accuracy and attack mitigation rounds on FashionMNIST and CIFAR10 with IID/non-IID settings when D_M has only one image.

updates are trimmed before aggregation. Thus, the robust aggregation baselines fail to mitigate the model poisoning attack under our attack settings.

We also compare our defense with CDP and LDP in terms of *benign accuracy* and *attack mitigation rounds*. We evaluate our defense by varying s from 0.1 to 1, and evaluate DP baselines by changing σ_{dp} from 0.1 to 10. For each defense method, we show the trade-off between *benign accuracy* and *attack mitigation rounds* in Figure 4. We have two key observations: 1) With sacrificing less than 5% benign accuracy, FL-WBC can mitigate the impact of model poisoning attack on the global model in 1 communication round for IID settings, and within 5 communication rounds for non-IID settings respectively. However, CDP and LDP fail to mitigate attack effect within 5 rounds for IID and within 10 rounds for non-IID settings with less than 5% accuracy drop. 2) For non-IID settings where the defense becomes more challenging, FL-WBC can still mitigate the attack effect within 2 rounds with less than 15% benign accuracy drop, but DP can not make an effective mitigation within 3 rounds with less than 30% benign accuracy drop, leading to the unacceptable utility on the benign task. The reason of FL-WBC outperforming CDP and LDP is that FL-WBC only inject perturbations to the parameter space where the long-lasting AEP resides in instead of perturbing all the parameters like DP methods. Therefore, FL-WBC can achieve better robustness with less accuracy drop.

In addition, we also observe that defense for non-IID settings is harder than IID settings, the reason is that under non-IID settings the devices train only a part of parameters [28] when holding only a few classes of data, leading to a sparser $H^k_{t,i}$ that is more likely to have a kernel with a higher dimension.

7.3 Effectiveness of FL-WBC with *Multiple* Images in The Malicious Dataset

We evaluate the defense effectiveness of robust aggregation baselines when D_M has 10 images, and the results are shown in Table 1.

Table 1: Results of attack mitigation rounds for robust aggregations when D_M has multiple images.

Defense	Fashion-MNIST (IID)	Fashion-MNIST (non-IID)	CIFAR10 (IID)	CIFAR10 (non-IID)
CTMA ($\beta = 0.1$)	7	9	8	>10
CTMA ($\beta = 0.2$)	7	8	8	9
CTMA ($\beta = 0.4$)	6	8	7	9
CMA	5	7	6	8

Defense against the attack when D_M has multiple images is easier than D_M has only one image. The reason is that AEP of multiple malicious images requires a larger parameter space to reside in compare to AEP of single malicious image.

The results show that even though attack effect will be mitigated finally when there are multiple images in D_M, robust aggregation can not guarantee mitigating the attack effect within 5 communication rounds for both IID and non-IID settings.

We also evaluate the defense effectiveness of FL-WBC and DP baselines in terms of benign accuracy and attack mitigation rounds when D_M has multiple images. The results are shown in Figure 5.

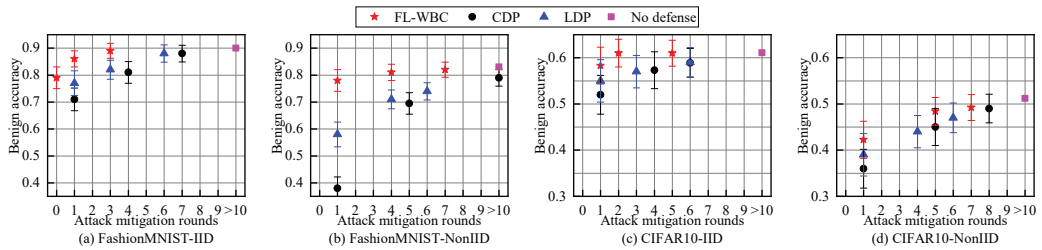


Figure 5: Comparison of benign accuracy and attack mitigation rounds on FashionMNIST and CIFAR10 with IID/non-IID settings when D_M has multiple images.

The results show that FL-WBC can guarantee that attack impact will be mitigated in one round with sacrificing less than 3% benign accuracy for IID settings and 10% for non-IID settings, respectively. However, the DP methods incur more than 9% benign accuracy drop to achieve the same robustness for IID settings and 40% for non-IID settings, respectively. Therefore, FL-WBC significantly outperforms the DP methods in defending against model poisoning attacks.

7.4 Integration of The Robustness Aggregation and FL-WBC

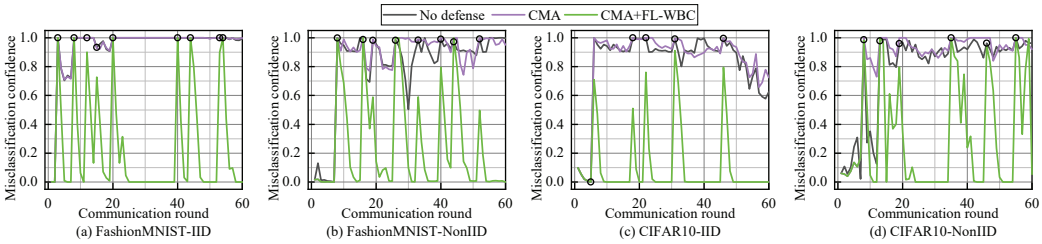


Figure 6: Comparison of misclassification confidence and communication round on FashionMNIST and CIFAR10 with IID/non-IID settings. The black circle denotes the adversarial rounds.

We also conduct experiments by integrating the robustness aggregation with FL-WBC to demonstrate that FL-WBC is complementary to server-based defenses. We conduct experiments by integrating Coordinate Median Aggregation (CMA) and FL-WBC. We set $s = 0.4$ for FL-WBC. After applying both CMA and FL-WBC with $s = 0.4$, the global model sacrifices less than 7% benign accuracy for both Fashion-MNIST and CIFAR10 dataset under IID/non-IID settings. We conduct experiments following the same setup in §7 with single image in the malicious dataset, and the results are shown in Figure 6.

The results show that only CMA can not mitigate the attack effect under our experimental setting. By applying both CMA and FL-WBC, the attack effect is mitigated within 1 communication rounds under IID settings and within 5 communication rounds under non-IID settings. Thus, our defense is complementary to the server-based robustness aggregations, and further enhance the robustness of FL against model poisoning attacks under extremely strong attacks.

8 Conclusion

We design a client-based defense against the model poisoning attack, targeting at the scenario where the attack that has already broken through the server-based defenses and polluted the global model. The experiment results demonstrate that our defense outperforms baselines in mitigating the attack effectively and efficiently, i.e., our defense successfully defends against the attack within fewer communication rounds with less model utility degradation. In this paper, we focus on the targeted poisoning attack [11, 12]. Our defense can be easily extended to many other poisoning attacks, such as backdoor attacks, since we do not restrict the malicious objective when deriving AEP.

9 Funding Transparency Statement

Funding in direct support of this work: NSF OIA-2040588, NSF CNS-1822085, NSF SPX-1725456, NSF IIS-2140247.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*, 2017.

[2] C. He, S. Li, J. So, X. Zeng, M. Zhang, H. Wang, X. Wang, P. Vepakomma, A. Singh, H. Qiu, *et al.*, "Fedml: A research library and benchmark for federated machine learning," *arXiv preprint arXiv:2007.13518*, 2020.

[3] E. M. E. Mhamdi, R. Guerraoui, and S. Rouault, "The hidden vulnerability of distributed learning in byzantium," *arXiv preprint arXiv:1802.07927*, 2018.

[4] M. Fang, X. Cao, J. Jia, and N. Gong, "Local model poisoning attacks to byzantine-robust federated learning," in *29th {USENIX} Security Symposium ({USENIX} Security 20)*, pp. 1605–1622, 2020.

[5] M. Baruch, G. Baruch, and Y. Goldberg, "A little is enough: Circumventing defenses for distributed learning," *arXiv preprint arXiv:1902.06156*, 2019.

[6] S. Mahloujifar, M. Mahmoody, and A. Mohammed, "Universal multi-party poisoning attacks," in *International Conference on Machine Learing (ICML)*, 2019.

[7] Z. Sun, P. Kairouz, A. T. Suresh, and H. B. McMahan, "Can you really backdoor federated learning?," *arXiv preprint arXiv:1911.07963*, 2019.

[8] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *International Conference on Artificial Intelligence and Statistics*, pp. 2938–2948, PMLR, 2020.

[9] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-y. Sohn, K. Lee, and D. Papailiopoulos, "Attack of the tails: Yes, you really can backdoor federated learning," *arXiv preprint arXiv:2007.05084*, 2020.

[10] C. Xie, K. Huang, P.-Y. Chen, and B. Li, "Dba: Distributed backdoor attacks against federated learning," in *International Conference on Learning Representations*, 2019.

[11] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing federated learning through an adversarial lens," in *International Conference on Machine Learning*, pp. 634–643, PMLR, 2019.

[12] R. Tomsett, K. Chan, and S. Chakraborty, "Model poisoning attacks against distributed machine learning systems," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006, p. 110061D, International Society for Optics and Photonics, 2019.

[13] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *International Conference on Machine Learning*, pp. 5650–5659, PMLR, 2018.

[14] K. Pillutla, S. M. Kakade, and Z. Harchaoui, "Robust aggregation for federated learning," *arXiv preprint arXiv:1912.13445*, 2019.

[15] S. Fu, C. Xie, B. Li, and Q. Chen, "Attack-resistant federated learning with residual-based reweighting," *arXiv preprint arXiv:1912.11464*, 2019.

[16] A. F. Siegel, "Robust regression using repeated medians," *Biometrika*, vol. 69, no. 1, pp. 242–244, 1982.

[17] P. W. Holland and R. E. Welsch, "Robust regression using iteratively reweighted least-squares," *Communications in Statistics-theory and Methods*, vol. 6, no. 9, pp. 813–827, 1977.

[18] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," 2017.

[19] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[20] C. Fung, C. J. Yoon, and I. Beschastnikh, "Mitigating sybils in federated learning poisoning," *arXiv preprint arXiv:1808.04866*, 2018.

[21] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 118–128, 2017.

[22] F. Sattler, K.-R. Müller, T. Wiegand, and W. Samek, "On the byzantine robustness of clustered federated learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8861–8865, IEEE, 2020.

[23] L. Muñoz-González, K. T. Co, and E. C. Lupu, "Byzantine-robust federated machine learning through adaptive model averaging," *arXiv preprint arXiv:1909.05125*, 2019.

[24] M. Naseri, J. Hayes, and E. De Cristofaro, "Toward robustness and privacy in federated learning: Experimenting with local and central differential privacy," *arXiv preprint arXiv:2009.03561*, 2020.

[25] F. Tramer, N. Carlini, W. Brendel, and A. Madry, "On adaptive attacks to adversarial example defenses," *arXiv preprint arXiv:2002.08347*, 2020.

[26] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin, "On evaluating adversarial robustness," *arXiv preprint arXiv:1902.06705*, 2019.

[27] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," in *International Conference on Learning Representations*, 2019.

[28] A. Li, J. Sun, B. Wang, L. Duan, S. Li, Y. Chen, and H. Li, "Lotteryfl: Personalized and communication-efficient federated learning with lottery ticket hypothesis on non-iid datasets," *arXiv preprint arXiv:2008.03371*, 2020.

A Algorithm for training process applying FL-WBC

The detailed local training process of benign devcies after applying FL-WBC is shown in Algorithm 1. As Algorithm 1 shows, the only overheads after applying our defense is that devices need additional storage to store $\boldsymbol{W}_{-1}$ and $\boldsymbol{W}_{-2}$ during local training.

Algorithm 1 Local training process applying FL-WBC on a benign device in round t.

Input: Local training data $\mathbb{D} \in \mathbb{R}^{L \times P \times Q}$; Local objective function $F : \mathbb{R}^{P \times Q} \to \mathbb{R}$; Local model parameters $\boldsymbol{W} \in \mathbb{R}^{M \times N}$; The number of local training iterations I; Learning rates $\eta_{t,i}$ for $i \in [I]$; Standard deviation of Laplace noise s.
Output: Learnt model parameter $\boldsymbol{W}$ with our defense.

Initialize $\boldsymbol{W}_{-1}, \boldsymbol{W}_{-2}$;
$i \leftarrow 0$;
for batch $\mathcal{B}$ in $\mathbb{D}$ **do**
  Randomly generate a Laplace noise matrix $\Upsilon \in \mathbb{R}^{M \times N}$ with $mean = 0$ and $std = s$;
  $\boldsymbol{W}_{-1} \leftarrow \boldsymbol{W}$;
  $\boldsymbol{W} \leftarrow \boldsymbol{W} - \eta_{t,i} \nabla F(\boldsymbol{W}, \mathcal{B})$;
  if this is not the first training batch **then**
    $\boldsymbol{W}^* \leftarrow (\boldsymbol{W} - \boldsymbol{W}_{-1}) - (\boldsymbol{W}_{-1} - \boldsymbol{W}_{-2})$;
    Find the set $\mathbb{S}$ which contains the indices of elements in $|\boldsymbol{W}^*| - \eta_{t,i}|\Upsilon|$ which are less than or equal to 0;
    for $j, k \in \mathbb{S}$ **do**
      $\boldsymbol{W}_{j,k} \leftarrow \boldsymbol{W}_{j,k} + \eta_{t,i} \Upsilon_{j,k}$;
    end for
  end if
  $\boldsymbol{W}_{-2} \leftarrow \boldsymbol{W}_{-1}$;
  $i \leftarrow i + 1$
end for

B Experiments to Support Analysis in §4.3

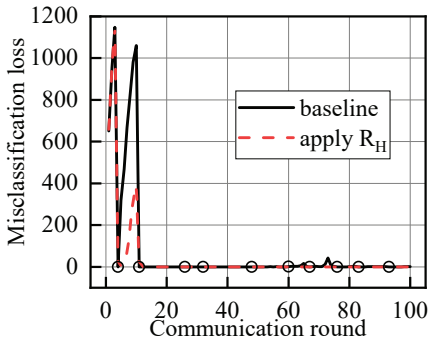


Figure 7: Compared results of misclassification loss on malicious dataset for model poisoning attack with and without applying $R_H(\boldsymbol{W})$. The black circle denotes the adversarial rounds.

To validate our analysis, we conduct experiments on Fashion-MNIST with model poisoning attacks in FL training. The training data is uniformly distributed to 100 devices (5 malicious devices included). In adversarial rounds, 5 malicious devices and 5 randomly chosen benign devices participate in training. In other rounds, 10 randomly selected benign devices participate in training. The model architecture can be found in Table 3 (Fashion-MNIST dataset). For training, we set local epoch E as 1 and batch size B as 32. We apply SGD optimizer and set the learning rate η to 0.01.

We first compute $\mathbb{E}_{i \in [I], k \in \mathbb{S}_{t_{adv}+1}}(\boldsymbol{H}^k_{t_{adv}+1,i} \delta_{t_{adv}})$, denoted as $\Phi_{t_{adv}}$, for each adversarial round t_{adv}. In order to derive $\delta_{t_{adv}}$, the malicious devices in round t_{adv} perform local training twice by setting α in Equation 3 as 1 and 0, respectively. Following that we have $\boldsymbol{W}_{t_{adv}}$ and $\boldsymbol{W}_{t_{adv}}(\mathbb{S}_{t_{adv}} \setminus \mathbb{M})$ respectively through aggregation. Then we derive $\delta_{t_{adv}}$ by computing the difference between $\boldsymbol{W}_{t_{adv}}(\mathbb{S}_{t_{adv}} \setminus \mathbb{M})$

Adversarial round	baseline Φ_t	Φ_t applying $R_H(\boldsymbol{W})$
5	3.56	1.06
11	0.00	0.00
26	0.01	0.00
32	0.00	0.00
48	0.00	0.00
60	0.05	0.00
67	0.07	0.00
76	0.00	0.00
83	0.00	0.00
93	0.00	0.00

Table 2: Numerical results of Φ_t with and without applying $R_H(\boldsymbol{W})$ at adversarial rounds.

and $\boldsymbol{W}_{t_{adv}}$. Due to the extremely high dimension of $\boldsymbol{H}^k_{t_{adv}+1,i}$, we compute $\Phi_{t_{adv}}$ element by element following

$$\Phi_{t_{adv}\,m,n} = \mathbb{E}_{i\in[I],k\in\mathbb{S}_{t_{adv}+1}} \left\langle \nabla_W [\nabla_W F^k(\boldsymbol{W}^k_{t_{adv}+1,i}, \xi^k_{t_{adv}+1,i})_{m,n}] | \delta_{t_{adv}} \right\rangle \tag{13}$$

The results of baseline in Figure 7 shows the loss that the malicious data is miss-classified by the global model, and the computed Φ_t for each adversarial round is presented in Table 2. The results show that for the first adversarial round (i.e. round 5), the value of Φ_t is higher than 3 and the attack is mitigated rapidly. While for the following adversarial rounds, Φ_ts are nearly 0, and the miss-classification loss keeps low until the next attack is conducted. This result is consistent with our analysis in §4 for why attack effects can remain in the global model.

In addition, we introduce a regularization $R_H(W)$ that approximates $\mathbb{E}_{i\in[I]}\|\boldsymbol{H}^k_{t_{adv},i}\delta_t\|_2$ into the local training of malicious devices, enforcing δ_t to reside in the null space of $\boldsymbol{H}^k_{t_{adv},i}$. Suppose $\boldsymbol{W} \in \mathbb{R}^{L\times M}$, then $R_H(W)$ for malicious devices k is formulated as

$$R_H(\boldsymbol{W}) = \sum_{l\in[L],m\in[M]} \left\langle \nabla_W [\nabla_W F^k(\boldsymbol{W}^k_{t_{adv},0}, \xi^k_{t_{adv},0})_{l,m}] | \boldsymbol{W} - \boldsymbol{W}_{t_{adv}}(\mathbb{S}_{t_{adv}} \setminus \mathbb{M}) \right\rangle. \tag{14}$$

It is shown that we use the sum of elements in $\boldsymbol{H}^k_{t_{adv},0}\delta_t$ to approximate $\mathbb{E}_{i\in[I]}\|\boldsymbol{H}^k_{t_{adv},i}\delta_t\|_2$. We adopt this approximation due to unacceptable computational costs of computing $\mathbb{E}_{i\in[I]}\|\boldsymbol{H}^k_{t_{adv},i}\delta_t\|_2$ directly. When $\mathbb{E}_{i\in[I]}\|\boldsymbol{H}^k_{t_{adv},i}\delta_t\|_2$ is zero, the sum of elements in $\boldsymbol{H}^k_{t_{adv},0}\delta_t$ should also be zero. So $R_H(\boldsymbol{W})$ is a weaker regularization compared to $\mathbb{E}_{i\in[I]}\|\boldsymbol{H}^k_{t_{adv},i}\delta_t\|_2$.

We aim to evaluate whether the poisoning attack can be boosted after applying $R_H(\boldsymbol{W})$, i.e., the attack effect remains for more rounds. As Figure 7 and Table 2 show, the values of Φ_t in round 5, 60 and 67 are reduced after applying R_H and the corresponding attacks are boosted. This result further supports our analysis that the reason why effective attacks can remain in the aggregated model is that the AEPs reside in the kernels of $\boldsymbol{H}^k_{t,i}$.

C Experiment setup

C.1 Detailed data preparation and hyperparameter configurations for §7

For IID settings, the data is uniformly distributed to 100 devices (malicious devices included). For non-IID settings, we first sort the data by the digit label, divide it into 200 shards uniformly, and assign each of 100 clients (malicious devices included) 2 shards.

For training, we set local epoch E as 1 and batch size B as 32. We apply SGD optimizer and set the learning rate η to 0.01. We set 5 devices out of totally 100 devices to be malicious. The model architectures for two dataset are shown in Table 3. We conduct 500 communication rounds of training for Fashion-MNIST and 1000 communication rounds for CIFAR10.

Table 3: Model architectures for *Fashion-MNIST* dataset and *CIFAR10* dataset.

Fashion-MNIST	***CIFAR10***
5× 5 Conv 1-16	5× 5 Conv 3-6
5× 5 Conv 16-32	3× 3 Maxpool
FC–10	5× 5 Conv 6-16
	3× 3 Maxpool
	FC–120
	FC–84
	FC–10

D Derivation of Equation 7

Since no malicious devices are selected between round τ_1 and round τ_2, the training processes of $\boldsymbol{W}_{t-1}(\mathbb{S} \setminus \mathbb{M})$ to $\boldsymbol{W}_t(\mathbb{S} \setminus \mathbb{M})$ and $\boldsymbol{W}_{t-1}$ to $\boldsymbol{W}_t$ are the same. The cause of difference between $\boldsymbol{W}_t(\mathbb{S} \setminus \mathbb{M})$ and $\boldsymbol{W}_t$ is that they are locally trained from different initial models, $\boldsymbol{W}_{t-1}(\mathbb{S} \setminus \mathbb{M})$ and $\boldsymbol{W}_{t-1}$, respectively.

We first estimate the term $\boldsymbol{W}_t^k(\mathbb{S} \setminus \mathbb{M}) - \boldsymbol{W}_t^k$ by applying first-order Taylor approximation and the chain rule based on Equation 2:

$$\begin{aligned} &\boldsymbol{W}_{t,I}^k(\mathbb{S} \setminus \mathbb{M}) - \boldsymbol{W}_{t,I}^k \\ \approx &\frac{\partial \boldsymbol{W}_{t,I}^k}{\partial \boldsymbol{W}_{t,I-1}^k} \frac{\partial \boldsymbol{W}_{t,I-1}^k}{\partial \boldsymbol{W}_{t,I-2}^k} \cdots \frac{\partial \boldsymbol{W}_{t,1}^k}{\partial \boldsymbol{W}_{t,0}^k} (\boldsymbol{W}_{t,0}^k(\mathbb{S} \setminus \mathbb{M}) - \boldsymbol{W}_{t,0}^k). \end{aligned} \tag{15}$$

where $\boldsymbol{W}_{t,0}^k(\mathbb{S} \setminus \mathbb{M}) - \boldsymbol{W}_{t,0}^k = \boldsymbol{W}_{t-1}(\mathbb{S} \setminus \mathbb{M}) - \boldsymbol{W}_{t-1}$. According to Equation 2, we have

$$\frac{\partial \boldsymbol{W}_{t,i+1}^k}{\partial \boldsymbol{W}_{t,i}^k} = \boldsymbol{I} - \eta_{t,i} \boldsymbol{H}_{t,i}^k, \tag{16}$$

where $\boldsymbol{H}_{t,i}^k \triangleq \nabla^2 F^k(\boldsymbol{W}_{t,i}^k, \xi_{t,i}^k)$.

Afterwards, according to Equation 4, we obtain

$$\boldsymbol{W}_t(\mathbb{S} \setminus \mathbb{M}) - \boldsymbol{W}_t = \frac{N}{K} \sum_{k \in \mathbb{S}_t} p^k [\boldsymbol{W}_{t,I}^k(\mathbb{S} \setminus \mathbb{M}) - \boldsymbol{W}_{t,I}^k]. \tag{17}$$

By combining Equations 15, 16 and 17, we get an estimator of δ_t, denoted as $\hat{\delta}_t$

$$\hat{\delta}_t = \frac{N}{K} [\sum_{k \in \mathbb{S}_t} p^k \prod_{i=0}^{I-1} (\boldsymbol{I} - \eta_{t,i} \boldsymbol{H}_{t,i}^k)] \hat{\delta}_{t-1}. \tag{18}$$

E Certified robustness guarantee

We define our *certified robustness guarantee* as the distance of AEP between the adversarial round and the subsequent rounds where FL-WBC is applied. A larger distance of AEP indicates that FL-WBC can mitigate the attack more efficiently. We first make an assumption for the Hessian matrix after applying our defense:

Assumption 5. *After applying FL-WBC, the new benign training Hessian matrix H' would be nearly full-rank. Following the notation of s as the variance of Υ in Equation 9, the operator H' is bounded below: $\mathbb{E}\|H'\boldsymbol{a}\|_2^2/P \geq s\|\boldsymbol{a}\|_2^2$.*

Additionally, we make assumptions on bounding the norm of AEPs in different rounds and independent distributions of $H_{t,i}^k \delta_{t,i}^k$.

Assumption 6. *The expected norm of AEPs is lower bounded in each round across devices and iterations: $\mathbb{E}_{i \in [I], k \in \mathbb{S}_t} \|\delta_{t,i}^k\|_2^2 \geq \Lambda_t$.*

Assumption 7. *Since local training data have independent distributions, the elements in $H_{t,i}^k \delta_{t,i}^k$ are independent and have expectations of 0 in different rounds across devices and iterations.*

Following the notations in §4 and applying the local training of benign devices in Algorithm 1, which can be found in Appendix A, we have the following theorem for our *robustness guarantee* on FedAvg after applying FL-WBC.

Theorem 2. *Let Assumption 5-7 hold and s, Λ_t, P be defined therein. K denotes the number of devices involved in training for each round. Assuming that poisoning attack happens in round t_{adv} and no attack happens from round t_{adv} to T, then for FedAvg applying FL-WBC we have:*

$$\mathbb{E}\|\hat{\delta}_T - \hat{\delta}_{t_{adv}}\|_2^2 \geq \frac{PIs}{K} \sum_{t=t_{adv}+1}^{T} \eta_{t,I-1}^2 \Lambda_t. \tag{19}$$

Proof. According to Equations 15 and 16, we obtain

$$\hat{\delta}_{t,i+1}^k = (\boldsymbol{I} - \eta_{t,i} \boldsymbol{H}_{t,i}^k) \hat{\delta}_{t,i}^k. \tag{20}$$

According to Equation 17, we have

$$\mathbb{E}(\hat{\delta}_t) = \mathbb{E}(\frac{N}{K} \sum_{k \in \mathbb{S}_t} p^k \hat{\delta}_{t,I}^k). \tag{21}$$

With $\hat{\delta}_{t-1} = \hat{\delta}_{t,0}^k$ and $\mathbb{E}(p^k) = \frac{1}{N}$, we have

$$\mathbb{E}(\hat{\delta_{t-1}}) = \mathbb{E}(\frac{N}{K} \sum_{k \in \mathbb{S}_t} p^k \hat{\delta}_{t,0}^k). \tag{22}$$

Then we have

$$\mathbb{E}(\hat{\delta}_t - \hat{\delta}_{t-1}) = \mathbb{E}[\frac{N}{K} \sum_{k \in \mathbb{S}_t} p^k (\hat{\delta}_{t,I}^k - \hat{\delta}_{t,0}^k)] \tag{23}$$

$$= \mathbb{E}[\frac{N}{K} \sum_{k \in \mathbb{S}_t} p^k \sum_{i=0}^{I-1} (\hat{\delta}_{t,i+1}^k - \hat{\delta}_{t,i}^k)] \tag{24}$$

$$= \mathbb{E}[\frac{N}{K} \sum_{k \in \mathbb{S}_t} p^k \sum_{i=0}^{I-1} -\eta_{t,i} \boldsymbol{H}_{t,i}^k \hat{\delta}_{t,i}^k], \tag{25}$$

where the first equality comes from Equations 21 and 22, the third equality comes from Equation 20.

Accumulating the difference between $\hat{\delta}_t$ and $\hat{\delta}_{t-1}$ formulated in Equation 25, we have

$$\mathbb{E}(\hat{\delta}_T - \hat{\delta}_{t_{adv}}) = \mathbb{E}[\sum_{t=t_{adv}+1}^{T} (\hat{\delta}_t - \hat{\delta}_{t-1})] \tag{26}$$

$$= \mathbb{E}[\sum_{t=t_{adv}+1}^{T} \sum_{k \in \mathbb{S}_t} \sum_{i=0}^{I-1} -\frac{N}{K} p^k \eta_{t,i} \boldsymbol{H}_{t,i}^k \hat{\delta}_{t,i}^k]. \tag{27}$$

Then we have

$$\mathbb{E}\|\hat{\delta}_T - \hat{\delta}_{t_{adv}}\|_2^2 = \mathbb{E}\| \sum_{t=t_{adv}+1}^{T} \sum_{k\in\mathbb{S}_t} \sum_{i=0}^{I-1} -\frac{N}{K} p^k \eta_{t,i} \boldsymbol{H}_{t,i}^k \hat{\delta}_{t,i}^k \|_2^2 \tag{28}$$

$$= \sum_{t=t_{adv}+1}^{T} \sum_{k\in\mathbb{S}_t} \sum_{i=0}^{I-1} \mathbb{E}\| \frac{N}{K} p^k \eta_{t,i} \boldsymbol{H}_{t,i}^k \hat{\delta}_{t,i}^k \|_2^2 \tag{29}$$

$$\geq \sum_{t=t_{adv}+1}^{T} \mathbb{E} \sum_{k\in\mathbb{S}_t} (\frac{N}{K} p^k \eta_{t,I-1})^2 \sum_{i=0}^{I-1} Ps \|\hat{\delta}_{t,i}^k\|_2^2 \tag{30}$$

$$\geq \sum_{t=t_{adv}+1}^{T} \mathbb{E} \sum_{k\in\mathbb{S}_t} (\frac{N}{K} p^k \eta_{t,I-1})^2 \sum_{i=0}^{I-1} Ps\Lambda_t \tag{31}$$

$$= \sum_{t=t_{adv}+1}^{T} \mathbb{E} \sum_{k\in\mathbb{S}_t} (\frac{\eta_{t,I-1}}{K})^2 IPs\Lambda_t \tag{32}$$

$$= \frac{PIs}{K} \sum_{t=t_{adv}+1}^{T} (\eta_{t,I-1})^2 \Lambda_t, \tag{33}$$

$$\tag{34}$$

where the second equality come from Assumption 7, the first inequality comes from the lower bound of operator $\boldsymbol{H}_{t,i}^k$ parameterized in Assumption 5 and shrinking learning rate, the second inequality comes form the bounded norm of AEP stated in Assumption 6. The third equality comes from the equation $\mathbb{E}(p^k) = \frac{1}{N}$.

□

F Proof of Theorem 1

Overview: Our proof is mainly inspired by [27]. Specifically, our proof has two key parts. First, we derive the bounds similar to those in Assumptions 3 and 4, after applying our defense scheme. Second, we adapt **Theorem 2** on convergence guarantee in [27] using our new bounds.

Bounding the expected distance between the perturbed gradients with our defense and raw gradients. According to Algorithm 1, the absolute values of elements in $\mathcal{T}_{t,i}$ in Equation 11 is $\max\{|\Upsilon| - |\boldsymbol{W}^*|/\eta_{t,i}, 0\}$. Thus, $\|\mathcal{T}_{t,i}\|_2^2 \leq \|\Upsilon\|_2^2$. Then we have

$$\mathbb{E}||\nabla F_k'(\boldsymbol{W}_{t,i}^k, \xi_{t,i}^k) - \nabla F_k(\boldsymbol{W}_{t,i}^k, \xi_{t,i}^k)||_2^2 \tag{35}$$

$$= \mathbb{E}||\mathcal{T}_{t,i}||_2^2 \leq \mathbb{E}||\Upsilon||_2^2 = \mathbb{E}(\Upsilon)^2 + Var(\Upsilon) = s^2. \tag{36}$$

New bounds for Assumption 3 with our defense.

We use the norm triangle inequality to bound he variance of stochastic gradients in each device, and we have

$$\mathbb{E}||\nabla F_k'(\boldsymbol{W}_{t,i}^k, \xi_{t,i}^k) - \nabla F_k(\boldsymbol{W}_{t,i}^k)||^2 \tag{37}$$

$$\leq \mathbb{E}||\nabla F_k'(\boldsymbol{W}_{t,i}^k, \xi_{t,i}^k) - \nabla F_k(\boldsymbol{W}_{t,i}^k, \xi_{t,i}^k)||^2 \tag{38}$$

$$+ \mathbb{E}||\nabla F_k(\boldsymbol{W}_{t,i}^k, \xi_{t,i}^k) - \nabla F_k(\boldsymbol{W}_{t,i}^k)||^2 \tag{39}$$

$$\leq s^2 + \sigma_k^2, \tag{40}$$

where we use Assumption 3 and Equation 36 in Equation 40.

New bounds for Assumption 4 with our defense. The expected squared norm of stochastic gradients $\nabla F_k'(\boldsymbol{W}_{t,i}^k, \xi_{t,i}^k)$ with our defense is as follows:

$$\mathbb{E}||\nabla F_k'(\boldsymbol{W}_{t,i}^k, \xi_{t,i}^k)||^2 \tag{41}$$

$$\leq \mathbb{E}||\nabla F_k'(\boldsymbol{W}_{t,i}^k, \xi_{t,i}^k) - \nabla F_k(\boldsymbol{W}_{t,i}^k, \xi_{t,i}^k)||^2 \tag{42}$$

$$+ \mathbb{E}||\nabla F_k(\boldsymbol{W}_{t,i}^k, \xi_{t,i}^k)||^2 \tag{43}$$

$$\leq s^2 + G^2, \tag{44}$$

where we use Assumption 4 and and Equation 36 in Equation 44.

Convergence guarantee for FedAvg with our defense. We define F^* and F_k^* as the minimum value of F and F_k and let $\Gamma = F^* - \sum_{k=1}^{N} p_k F_k^*$. We assume each device has I local training iterations in each round and the total number of rounds is T. Let Assumptions 1-4 hold and L, μ, σ_k, G be defined therein. Choose $\kappa = \frac{L}{\mu}$, $\gamma = \max\{8\kappa, I\}$ and the learning rate $\eta_{t,i} = \frac{2}{\mu(\gamma + tI + i)}$.

By applying our new bounds and **Theorem 2** in [27], FedAvg using our defense has the following convergence guarantee:

$$\mathbb{E}[F(\boldsymbol{W}_T)] - F^* \leq \frac{2\kappa}{\gamma + TI}\left(\frac{Q + C}{\mu} + \frac{\mu\gamma}{2}\mathbb{E}||\boldsymbol{W}_0 - \boldsymbol{W}^*||^2\right),$$

where

$$Q = \sum_{k=1}^{N} p_k^2(s^2 + \sigma_k^2) + 6L\Gamma + 8(I-1)^2(s^2 + G^2)$$
$$C = \frac{4}{K}I^2(s^2 + G^2).$$