Detecting Poisoning Nodes in Federated Learning by Ranking Gradients

Wanchuang Zhu[1] [2], Benjamin Zi Hao Zhao [3], Simon Luo[1] [2] and Ke Deng[4] *
[1] School of Mathematics and Statistics, The University of Sydney, Australia
[2] ARC Centre for Data Analytics for Resources and Environments, Australia
[3] Department of Computing, Macquarie University, Australia
[4] Center for Statistical Science & Department of Industrial Engineering, Tsinghua University, China
{wanchuang.zhu, s.luo}@sydney.edu.au
ben_zi.zhao@mq.edu.au, kdeng@tsinghua.edu.cn

Abstract

We propose a simple, yet effective defense against poisoning attacks in Federated Learning. Our approach transforms the update gradients from local nodes into a matrix containing the rankings of local nodes across all model parameter dimensions. We then distinguish the malicious nodes from the benign nodes with key characteristics of the rank domain, specifically, the mean and standard deviation of a node's parameter rankings. Under mild conditions, we prove that our approach is guaranteed to detect all malicious nodes under typical Byzantine poisoning attacks with no prior knowledge or history about the participating nodes. The effectiveness of our proposed approach is further confirmed by experiments on two classic datasets. Compared to the state-of-art methods in the literature for defending Byzantine attacks, our approach is unique in its way of identifying the malicious nodes by ranking and its robustness to effectively defense a wide range of attacks.

1 Introduction

FL departs from conventional centralized learning by allowing multiple participating nodes to learn on a local collection of training data, before each respective node's updates are sent to a global coordinator for aggregation. With an aggregation of multiple nodes, the resulting model observes greater performance than if each node was to learn on their local subset only. FL presents two advantages, increased privacy for contributing nodes as local data is not communicated to the coordinator, and reductions of computation by the global node with computation offloaded to contributors.

However, malicious actors in the collaborative process may seek to poison the performance of the global model, to reduce the output performance of the model [Chen et al., 2017, Fang et al., 2020, Tolpegin et al., 2020b]. A Byzantine attack aims to devastate the performance of the global model by manipulating the gradient values of malicious nodes in a coordinated manner. In the literature, there are two typical defense strategies: malicious node detection and robust learning. Malicious node detection defenses by detecting malicious nodes and removing them from the aggregation [Blanchard et al., 2017, Guerraoui et al., 2018, Li et al., 2020, So et al., 2021]. Robust learning [Blanchard et al., 2017, Yin et al., 2018, Guerraoui et al., 2018, Fang et al., 2020], however, withstands a proportion of malicious nodes and defenses by reducing the negative impacts of the malicious nodes via various robust learning methods [Wu et al., 2020b, Xie et al., 2019, 2020, Cao et al., 2021].

In this paper, we focus on defensing Byzantine attacks via malicious node detection. In the literature, there are efforts of the same vein. Blanchard et al. [2017] propose a defense referred to as Krum that treats local nodes whose update vector is too far away from the aggregated barycenter

1st NeurIPS Workshop on New Frontiers in Federated Learning (NFFL 2021), Virtual Meeting.

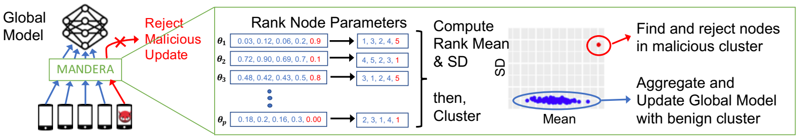


Figure 1: An Overview of MANDERA

as malicious nodes and precludes them from the downstream aggregation. Guerraoui et al. [2018] propose Bulyan, a process that performs aggregation on subsets of node updates (by iteratively leaving each node out) to find a set of nodes with the most aligned updates given an aggregation rule. These methods share a common element: detection is based on the node updates directly. However, usually the different dimensions of the node updates remain quite different in their range of values and follow very different distributions. This phenomena makes it challenging to precisely detect malicious nodes directly based on the node updates, as a few dimensions often dominate the final result.

We propose to resolve this critical problem from a novel perspective. Instead of working on the node updates directly, we propose to extract information about malicious nodes indirectly by transforming the node updates from numeric gradient values to the rank domain. Compared to the original numeric gradient values, whose distribution is difficult to model, the ranks are much easier to handle both theoretically and practically. Moreover, as ranks are scale-free, we no longer need to worry about the scale difference across different update dimensions. We proved under mild conditions that the first two moments of the transformed rank vectors carry key information to detect the malicious nodes under a wide range of Byzantine attacks. Based on these theoretical results, a highly efficient method called MANDERA is proposed to separate the malicious nodes from the benign ones by clustering all local nodes into two groups based on the moments of their rank vectors. With the assumption that malicious nodes are the minority in the node pool, we can simply treat all nodes in the smaller cluster as malicious nodes and remove them from the aggregation.

The contributions of this work are as follows. **(1)** We propose the first algorithm leveraging the rank domain of model updates to detect malicious nodes (Figure 1). **(2)** We provide theoretical guarantee for the detection of malicious nodes based on the rank domain under Byzantine attacks. **(3)** Our method does not assume knowledge on the number of malicious nodes, which is required in the learning process of prior methods. **(4)** We experimentally demonstrate the effectiveness and robustness of our defense on Byzantine attacks, including Gaussian attack, Sign Flipping attack and Zero Gradient attack, in addition to a more subtle Label Flipping data poisoning attack. **(5)** An experimental comparison between MANDERA and a collection of alternative robust aggregation techniques, including Krum [Blanchard et al., 2017], Trimmed Mean [Yin et al., 2018, Fang et al., 2020], Median [Yin et al., 2018, Fang et al., 2020], Bulyan [Guerraoui et al., 2018], are provided.

2 Defense Formalization

2.1 Notations

Suppose there are n local nodes in the federated learning framework, where n_1 nodes are benign nodes whose indices are denoted by $\mathcal{I}_b$ and the other $n_0 = n - n_1$ nodes are malicious nodes whose indices are denoted by $\mathcal{I}_m$. The training model is denoted by $f(\boldsymbol{\theta}, \boldsymbol{D})$, where $\boldsymbol{\theta} \in \mathbb{R}^{p\times 1}$ is a p-dimensional parameter vector of interest and $\boldsymbol{D}$ is the data matrix. Denote the message matrix received from all local nodes by the central server as $\boldsymbol{M} \in \mathbb{R}^{n\times p}$, where each row denotes a message vector from a single local node, with $\boldsymbol{M}_{i,:}$ being the message received from node i. For a benign node $i \in \mathcal{I}_b$, let $\boldsymbol{D}_i$ be the data matrix on it, we typically have $\boldsymbol{M}_{i,:} = \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{D}_i)}{\partial \boldsymbol{\theta}}$. A malicious node $i \in \mathcal{I}_m$, however, tends to attack the learning system by manipulating $\boldsymbol{M}_{i,:}$ in some way.

Given a vector of real numbers $a \in \mathbb{R}^{p\times 1}$, define its ranking vector as $b = Rank(a) \in perm\{1, \cdots, p\}$, where the ranking operator $Rank$ maps the vector a to its permutation space $perm\{1, \cdots, p\}$ which is the set of all the permutations of $\{1, \cdots, p\}$. For example, $Rank(1.1, -2, 3.2) = (2, 3, 1)$. With the *Rank* operator, we can transfer the message matrix $\boldsymbol{M}$ to a ranking matrix $\boldsymbol{R}$ by replacing its column $\boldsymbol{M}_{:,j}$ by the corresponding ranking vector

$\boldsymbol{R}_{:,j} = Rank(\boldsymbol{M}_{:,j})$. Further define

$$e_i \triangleq \frac{1}{p}\sum_{j=1}^{p} \boldsymbol{R}_{i,j} \qquad \text{and} \qquad v_i \triangleq \frac{1}{p}\sum_{j=1}^{p}(\boldsymbol{R}_{i,j} - e_i)^2$$

to be the mean and variance of $\boldsymbol{R}_{i,:}$, respectively. As we will show in later subsections, we can judge whether node i is a malicious node based on (e_i, v_i) under various attacking patterns. In the following, we will highlight the behaviour of the benign nodes first, and then discuss the behaviour of malicious nodes and their interactions with the benign nodes under various Byzantine attacks respectively.

2.2 Behaviour of benign nodes

As the behaviour of benign nodes does not depend on the type of Byzantine attack, we can study the statistical properties of (e_i, v_i) for a benign node $i \in \mathcal{I}_b$ before the specification of a concrete attack type.

For any benign node i, the message generated for j^{th} parameter is

$$\boldsymbol{M}_{i,j} = \frac{1}{N_i}\sum_{l=1}^{N_i} \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{D}_{i,l})}{\partial \boldsymbol{\theta}_j}, \tag{1}$$

where N_i is the sample size used by node i to compute the gradient and $\boldsymbol{D}_{i,l}$ denotes the l^{th} sample of it. Assuming that $\boldsymbol{D}_{i,l}$s are independent and identically distributed (IID) samples drawn from a data distribution $\mathbb{D}$, equation 1 tells us that $\boldsymbol{M}_{i,j}$ is the sample mean of IID random variables, i.e., $\{\frac{\partial f(\boldsymbol{\theta}, \boldsymbol{D}_{i,l})}{\partial \boldsymbol{\theta}_j}\}_{l=1}^{N_i}$. Thus, according to the Central Limit Theorem, $\boldsymbol{M}_{i,j}$ follows a Gaussian distribution asymptotically with the increase of N_i. The lemma below summarizes the result formally.

Lemma 1. *Assuming that data samples on each benign node $i \in \mathcal{I}_b$ are IID samples from a distribution $\mathbb{D}$, and let $\sigma_j^2 = \mathrm{Var}(\frac{\partial f(\boldsymbol{\theta}, \boldsymbol{D}_{i,l})}{\partial \boldsymbol{\theta}_j}) < \infty$, we have*

$$\boldsymbol{M}_{i,j} \xrightarrow{d} \mathcal{N}\left(\mu_j, \sigma_j^2/N_i\right), \ 1 \le j \le p, \tag{2}$$

where $\mu_j = \mathbb{E}(\frac{\partial f(\boldsymbol{\theta}, \boldsymbol{D}_{i,l})}{\partial \boldsymbol{\theta}_j})$.

2.3 Behaviour of malicious node under the Gaussian attack

We provide the definition of the Gaussian attack in Appendix C.1. Considering that $\lim_{n_1 \to \infty} \boldsymbol{m}_b = \mu_j \ a.s.$, according to the Kolmogorov Strong Law of Large Numbers (KSLLN), the distribution of $\boldsymbol{M}_{i,j}$ can be well approximated by a Gaussian distribution centered at μ_j when n_1 is reasonably large. The lemma 2 provides the details.

Lemma 2. *Under the same assumption as in Lemma 1, we have for each malicious node $i \in \mathcal{I}_m$ under the Gaussian attack*

$$\boldsymbol{M}_{i,j} \xrightarrow{d} \mathcal{N}\left(\mu_j, \Sigma_{j,j}\right), \ 1 \le j \le p. \tag{3}$$

Lemma 1 and Lemma 2 tell us that for each parameter dimension j, $\{\boldsymbol{M}_{i,j}\}_{i=1}^{n}$ are independent Gaussian random variables with the same mean (i.e, μ_j) but different variances (i.e., σ_j^2/N_i or $\Sigma_{j,j}$) under the Gaussian attack. Due to the symmetry of Gaussian distribution, it is straightforward to see that $\mathbb{E}(\boldsymbol{R}_{i,j}) = \frac{n+1}{2}, \ 1 \le i \le n, \ 1 \le j \le p$. Moreover, the exchangeability of benign nodes and the exchangeability of malicious nodes tell us that for each parameter dimension j, there exist two positive constants $s_{b,j}^2$ and $s_{m,j}^2$ such that

$$\mathrm{Var}(\boldsymbol{R}_{i,j}) = s_{b,j}^2, \ \forall\, i \in \mathcal{I}_b, \ \text{and} \ \mathrm{Var}(\boldsymbol{R}_{i,j}) = s_{m,j}^2, \ \forall\, i \in \mathcal{I}_m.$$

Further assume that $\boldsymbol{R}_{i,j}$'s are independent of each other, thus $e_i = \frac{1}{p}\sum_{j=1}^{p} \boldsymbol{R}_{i,j}$ is the sum of independent random variables with a common mean. Thus, according to the KSLLN, we know that e_i converges to a constant almost surely, which in turn indicates that v_i also converge some constant almost surely.

The theorem below summarizes the results formally, with the detailed proof provided in Appendix D.

Theorem 1. *Assuming $\{\boldsymbol{R}_{i,j}\}_{1\leq j\leq p}$ are independent of each other, under the Gaussian attack, we have*

$$\begin{aligned}\lim_{p\to\infty} e_i &= \frac{n+1}{2}\; a.s., \qquad (4)\\ \lim_{p\to\infty} v_i &= \bar{s}_b^2\cdot\mathbb{I}(i\in\mathcal{I}_b)+\bar{s}_m^2\cdot\mathbb{I}(i\in\mathcal{I}_m)\; a.s., \qquad (5)\end{aligned}$$

where $\mathbb{I}(\cdot)$ stands for the indicator function, $\bar{s}_b^2=\frac{1}{p}\sum_{j=1}^p s_{b,j}^2$ and $\bar{s}_m^2=\frac{1}{p}\sum_{j=1}^p s_{m,j}^2$.

Considering that $\bar{s}_b^2=\bar{s}_m^2$ if and only if $\{\Sigma_{j,j}\}_{j=1}^p$ falls into a lower dimensional manifold whose measurement is zero under the Lebesgue measure on $\mathbb{R}^p$, we have $P(\bar{s}_b^2=\bar{s}_m^2)=0$ if the attacker specifies the Gaussian variance $\Sigma_{j,j}$'s arbitrarily in the Gaussian attack. Thus, Theorem 1 in fact suggests that the benign nodes and the malicious nodes are different on the value of v_i, and thus provides a guideline to detect the malicious nodes.

Due to the following reasons, The independence assumption in Theorem 1 is a mild condition that can be easily satisfied in practice. First, for a benign node $i\in\mathcal{I}_b$, $\boldsymbol{M}_{i,j}$ and $\boldsymbol{M}_{i,k}$ are often nearly independent, as the correlation between two model parameters $\boldsymbol{\theta}_j$ and $\boldsymbol{\theta}_k$ is often very week in a larger deep neural network with a huge number of parameters. To verify the statement, we implemented independence tests for 100000 column pairs randomly chosen from the message matrix $\boldsymbol{M}$ generated from the MNIST data. Distribution of the p-values of these tests are demonstrated in Figure 4 of Appendix E via a histogram, which is close to a uniform distribution, indicating that $\boldsymbol{M}_{i,j}$ and $\boldsymbol{M}_{i,k}$ are indeed nearly independent in practice. Second, even some $\boldsymbol{M}_{i,j}$ and $\boldsymbol{M}_{i,k}$ shows strong correlation, magnitude of the correlation would be reduced greatly during the transformation from $\boldsymbol{M}$ to $\boldsymbol{R}$, as the final ranking $R_{i,j}$ also depends on many other factors.

2.4 Malicious node detection for sign flipping attack

From SF's definition (Appendix C.2), a malicious node i's update message under the SF attack is

$$\boldsymbol{M}_{i,:}=-r\boldsymbol{m}_b=-\frac{r}{n_1}\sum_{k\in\mathcal{I}_b}\boldsymbol{M}_{k,:}. \qquad (6)$$

For fixed $\{\boldsymbol{M}_{k,:}\}_{k\in\mathcal{I}_b}$, $M_{i,:}$ is also a fixed vector without randomness, as it is a deterministic function of $\{\boldsymbol{M}_{k,:}\}_{k\in\mathcal{I}_b}$. On the other hand, however, we can also treat $M_{i,:}$ as a random vector, since the randomness of $\{\boldsymbol{M}_{k,:}\}_{k\in\mathcal{I}_b}$ can be transferred to $M_{i,:}$ via the link function in equation 6. In fact, for any parameter dimension j, considering that $\boldsymbol{M}_{k,j}\xrightarrow{d}\mathcal{N}\left(\mu_j,\sigma_j^2/N_k\right)$ for any $k\in\mathcal{I}_b$ according to Lemma 1, it is straightforward to see that $\boldsymbol{M}_{i,j}=-\frac{r}{n_1}\sum_{k\in\mathcal{I}_b}\boldsymbol{M}_{k,j}$ can also be well approximated by a Gaussian distribution. The lemma below summarizes the result formally.

Lemma 3. *Under the sign flipping attack, for each malicious node $i\in\mathcal{I}_m$ and any parameter dimension j, we have $\boldsymbol{M}_{i,j}=-\frac{r}{n_1}\sum_{k\in\mathcal{I}_b}\boldsymbol{M}_{k,j}$ is a deterministic function of $\{\boldsymbol{M}_{k,j}\}_{k\in\mathcal{I}_b}$, whose limiting distribution is*

$$\boldsymbol{M}_{i,j}\xrightarrow{d}\mathcal{N}\big(\mu_j(r),\sigma_j^2(r)\big),\; 1\leq j\leq p, \qquad (7)$$

where $\mu_j(r)=-r\mu_j$, $\sigma_j^2(r)=\frac{r^2\cdot\sigma_j^2}{n_1\cdot\bar{N}_b}$, and $\bar{N}_b=\frac{n_1}{\sum_{k\in\mathcal{I}_b}\frac{1}{N_k}}$ is the harmonic mean of $\{N_k\}_{k\in\mathcal{I}_b}$.

Lemma 1 and Lemma 3 tells us that for each parameter dimension j, the distribution of $\{\boldsymbol{M}_{i,j}\}_{i=1}^n$ is a mixture of Gaussian components $\{\mathcal{N}\left(\mu_j,\sigma_j^2/N_i\right)\}_{i\in\mathcal{I}_b}$ centered at μ_j plus a point mass located at $\mu_j(r)=-r\mu_j$. If N_i's are reasonably large, variances σ_j^2/N_i's would be very close to zero, and the probability mass of the mixture distribution would concentrate to two local centers μ_j and $\mu_j(r)=-r\mu_j$, one for the benign nodes and the other one for the malicious nodes. This intuition provides us the guidance to identify the malicious nodes in this attack pattern. Transforming to the rank domain, the above intuition leads to different behavior patterns of the benign nodes and the malicious nodes in the rank matrix $\boldsymbol{R}$, which in turn result in different limiting behavior of (e_i,v_i) for the benign and malicious nodes. The theorem below summarizes the results formally, with the detailed proof provided in Appendix G.

Theorem 2. *Define* $p_+ \triangleq \sum_{j=1}^{p} \mathbb{I}(\mathbb{E}_{D\sim\mathbb{D}} \geq 0) = \sum_{j=1}^{p} \mathbb{I}(\mu_j \geq 0)$ *as the number of dimensions which have a non-negative expectation. Let* $N_{min} = \min\{N_1, \cdots, N_n\}$*. Under the sign flipping attack, we have,*

$$\lim_{p\to\infty} \lim_{N_{min}\to\infty} e_i = \bar{a}_b \cdot \mathbb{I}(i \in \mathcal{I}_b) + \bar{a}_m \cdot \mathbb{I}(i \in \mathcal{I}_m),\ a.s. \tag{8}$$

$$\lim_{p\to\infty} \lim_{N_{min}\to\infty} v_i = \bar{v}_b \cdot \mathbb{I}(i \in \mathcal{I}_b) + \bar{v}_m \cdot \mathbb{I}(i \in \mathcal{I}_m),\ a.s. \tag{9}$$

where $\bar{a}_b = \frac{n+n_0+1}{2} - n_0\frac{p_+}{p}$, $\bar{a}_m = n_1\frac{p_+}{p} + \frac{n_0+1}{2}$, $\bar{v}_b = \beta_0 + \beta_1 p_+ + \beta_2 p_+^2$, $\bar{v}_m = \frac{p_+(p-p_+)n_1^2}{p^2}$, $\beta_0 = \frac{7n^2+7n_0^2+10nn_0+2n+2n_0-1}{12}$, $\beta_1 = \frac{n_0}{p}(2n+3n_0+2)$ *and* $\beta_2 = -\frac{n_0^2}{p^2}$.

Considering that $\bar{a}_b = \bar{a}_m$ if and only if $p_+ = \frac{p}{2}$, and $\bar{v}_b = \bar{v}_m$ if and only if p_+ is the root of a specific quadratic equation, the probability of $(\bar{a}_b, \bar{v}_b) = (\bar{a}_m, \bar{v}_m)$ is close to zero. Such a phenomenon suggests that we can detect the malicious nodes based on the moments (e_i, v_i) to defense the sign flipping attack as well. Noticeably, the limit behaviour of e_i and v_i does not dependent on the specification of r, which defines the sign flipping attack. It is totally understandable once we realize that with the variance of $\boldsymbol{M}_{i,j}$ shrinks to zero with N_i goes to infinity for a benign node i, any difference between μ_j and $\mu_j(r)$ would result in the same rank vector $\boldsymbol{R}_{:,j}$ in the rank domain.

2.5 Malicious node detection for zero gradient attack

The ZG attack defined in Appendix C.3 is a special case of sign flipping attack by specifying $r = \frac{n_1}{n_0}$. Since the conclusions of Theorem 2 keep unchanged for different specifications of r as we have discussed, we have the following corollary for zero gradient attack.

Corollary 1. *Under the zero gradient attack,* e_i*'s and* v_i*'s follow exactly the same limiting behaviours as described in Theorem 2.*

2.6 MANDERA

Theorem 1, 2 and Corollary 1 imply that, under these three attacks (Gaussian attack, zero gradient attack and sign flipping attack), the variance of a benign node is different from that of a malicious node. At the same time, the variances of any two benign nodes or two malicious nodes are asymptotically identical. Malicious nodes are guaranteed to be detected by a clustering algorithm. Based on this, we propose *MAlicious Node DEtection via RAnking* (MANDERA) to detect the malicious nodes, whose workflow is detailed in Algorithm 1.

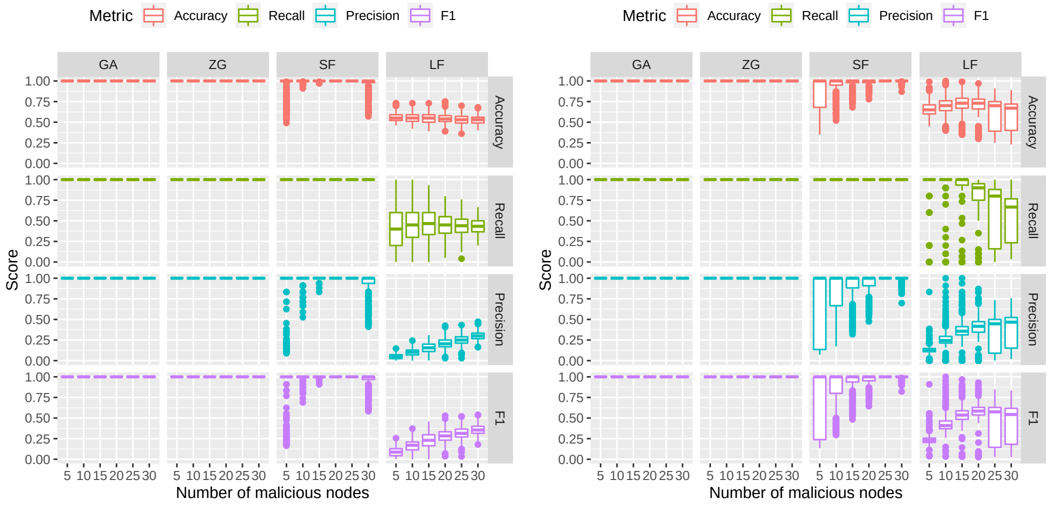
Algorithm 1 Malicious node detection via ranking (MANDERA)

```
Input: Data $\boldsymbol{M}$.
Convert the message data $\boldsymbol{M}$ to ranking data $\boldsymbol{R}$ by applying Rank operator.
Compute mean and standard deviation of each row of $\boldsymbol{R}$, and denote them as $\boldsymbol{e}$ and $\boldsymbol{s}$ respectively;
Run a clustering algorithm (here, we use K-means) to $(\boldsymbol{e}, \boldsymbol{s})$ with 2 groups. Denote the classification results as $C$.
Output: Classification $C$.
```

Remark. *MANDERA can be applied to either a single epoch or multiple epochs. For a single Epoch, the data will be the message matrix received from a single epoch. For multiple Epochs, the data* $\boldsymbol{M}$ *would be the combined message matrix from multiple epochs. By default, the experiments below all use single epoch to detect the malicious nodes. In Algorithm 1, we only implement a clustering algorithm on a nodes' mean and standard deviation. However, higher order moments can be used to improve MANDERA.*

3 Experiments

We evaluate the efficacy in detecting malicious nodes within the federated learning framework with the use of two Datasets. The first is the Fashion-MNIST dataset [Xiao et al., 2017], a dataset of 60,000 and 10,000 training and testing samples respectively divided into 10 classes of apparel. The



(a) CIFAR-10

(b) Fashion-MNIST

Figure 2: Classification performance of our proposed approach MANDERA under four types of attack for CIFAR-10 data. GA: Gaussian attack; ZG: Zero-gradient attack; SF: Sign-flipping; and LF: Label-flipping. The boxplot bounds the 25th (Q1) and 75th (Q3) percentile, with the central line representing the 50th quantile (median). The end points of the whisker represent the Q1-1.5(Q3-Q1) and Q3+1.5(Q3-Q1) respectively.

seconds is CIFAR-10 [Krizhevsky et al., 2009], a dataset of 60,000 small object images also containing 10 object classes. In these experiments we mainly adopt implementations of Byzantine attacks released by [Wu et al., 2020b,a] and the label flipping attack from [Tolpegin et al., 2020b,a]. In our experiments, we set $\Sigma = 30\boldsymbol{I}$ for the Gaussian attack and $u = 3$ for the sign flipping attack, where $\boldsymbol{I}$ is the identity matrix. For all experiments we fix $n = 100$ participating nodes, of which a variable number of nodes are poisoned $|n_0| \in \{5, 10, 15, 20, 25, 30\}$. The training process is run until 25 epochs have elapsed. We have described the structure of these networks in Appendix A.

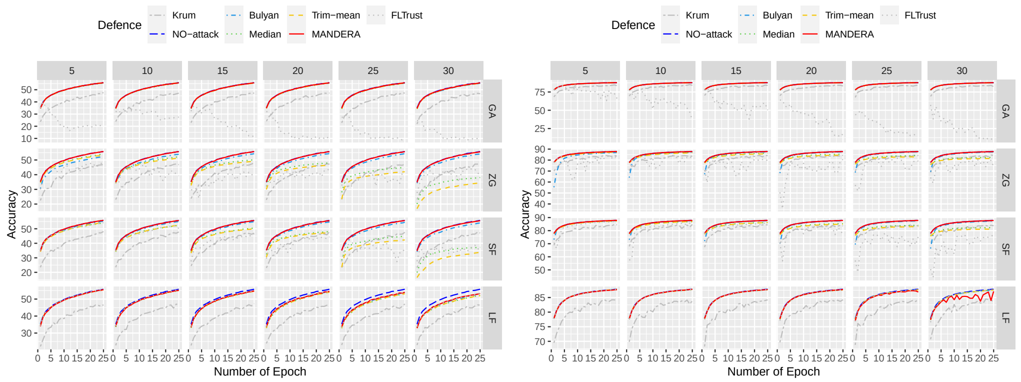
3.1 Malicious node detection by MANDERA

We test the performance of MANDERA on the update gradients of a model under attack, in this section, MANDERA acts as an observer without intervening in the learning process to identify malicious nodes with a set of gradients from a single epoch. Each configuration of 25 training epochs, with a given number of malicious nodes was repeated 20 times. Figure 2 demonstrates the classification performance (Metrics defined in Appendix B) of MANDERA with different settings of participating malicious nodes and the four poisoning attacks of Guassian Attack (GA), Zero Gradient attack (ZG), Sign Flipping attack (SF) and the Label Flipping attack (LF).

While we have formally demonstrated the efficacy of MANDERA in accurately detecting potentially malicious nodes participating in the federated learning process. In practice, to leverage an unsupervised K-means clustering algorithm, we must also identify the correct group of nodes as the malicious group. Our strategy is to identify the group with the most exact gradients, or otherwise the smaller group (we regard a system with over 50% of their nodes compromised as having larger issues than just poisoning attacks) [1].

From Figure 5 in the Appendix, it is immediately evident that the recall of the malicious nodes for the Byzantine attacks is exceptional, however on occasion benign nodes have also been misclassified as malicious under a SF, and to a lesser extent the ZG attack for both datasets. On all attacks, in the presence of more malicious nodes, the recall of malicious nodes trends down. As for the data poisoning attack of LF, it is consistently more difficult to detect, however we note that the LF attack has a more subtle influence on the model in contrast to the impact of Byzantine attacks.

[1] We acknowledge that more informed approachs to selecting the malicious cluster can be tested in future work, e.g. Figure 5 notes a lower variation of rank variance compared to benign nodes. This could enable more robust means of selecting the malicious group, and enabling selection of malicious groups larger than 50%.



(a) CIFAR-10 Dataset

(b) FASHION-MNIST dataset

Figure 3: Model Accuracy at each epoch of training, each line of the curve represents a different defense against the poisoning attacks.

3.2 MANDERA for defending against poisoning attacks

In this section we encapsulate MANDERA into a module prior to the the aggregation step, MANDERA has the sole objective of identifying malicious nodes, and excluding their updates from the global aggregation step. Each configuration of 25 training epochs, a given poisoning attack, defense method, and a given number of malicious nodes was repeated 10 times. We compare MANDERA against 4 other robust aggregation defense methods, Krum [Blanchard et al., 2017], Bulyan [Guerraoui et al., 2018], Trimmed Mean [Yin et al., 2018] and Median [Yin et al., 2018]. Of which the first 2 abandon an assumed number of malicious nodes, and the latter 2 only aggregate robustly.

From Figure 3, it is observed that MANDERA performs about the same as the best performing defense mechanisms, close to the performance of a model not under attack. MANDERA's accuracy is observed to vary slightly under the LF attack on fashion data with 30 malicious nodes, this is consistent with the larger accuracy ranges previously observed in Figure 2b.

4 Future Works and Conclusion

As with any proposed defense, attackers may attempt to craft gradients that achieve a similar poisoning effect, whilst preserving the ranking distribution of their submitted gradients. The rank domain inherently discards information such as distance between two consecutively ranked node parameters, this information could be used by said attacker to increase the anomalous gradient (to just below the next value) without it's rank changing. There exists the possibility of performing MANDERA in differentially private or secure FL, with the use of private ranking algorithms. While we have provided results for Byzantine attacks and the subtler label flipping attack, it remains to be seen the effectiveness of MANDERA on more advanced poisoning techniques like GAN-based poisoning or Evasion attacks. In conclusion, we have theoretically proven guarantees and experimentally shown efficacy in the use of ranking algorithms for the detection of malicious nodes performing poisoning attacks against federated learning. Our proposed method MANDERA, is able to achieve high detection accuracies and maintain a model accuracy on par with other seminal, high performing defense mechanisms, but with 3 notable advantages. First, provable guarantees for the use of ranking to detect Gaussian, Zero Gradient and Sign Flipping attacks. Next, faster detection with the use of ranking algorithms. Finally, the MANDERA defense does not need a prior estimation of the number of poisoned nodes. In this work we demonstrate how the rank domain can be useful in applications to defend against malicious actors.

References

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/f4b9ec30ad9f68f89b29639786cb62ef-Paper.pdf`.

Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Provably secure federated learning against malicious clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6885–6893, 2021.

Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proc. ACM Meas. Anal. Comput. Syst.*, 1(2), December 2017. doi: 10.1145/3154503. URL `https://doi.org/10.1145/3154503`.

Zheyi Chen, Pu Tian, Weixian Liao, and Wei Yu. Zero knowledge clustering based adversarial mitigation in heterogeneous federated learning. *IEEE Transactions on Network Science and Engineering*, 8(2):1070–1083, 2021. doi: 10.1109/TNSE.2020.3002796.

Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Gong. Local model poisoning attacks to byzantine-robust federated learning. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, pages 1605–1622, 2020.

Rachid Guerraoui, Sébastien Rouault, et al. The hidden vulnerability of distributed learning in byzantium. In *International Conference on Machine Learning*, pages 3521–3530. PMLR, 2018.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Suyi Li, Yong Cheng, Wei Wang, Yang Liu, and Tianjian Chen. Learning to detect malicious clients for robust federated learning. *arXiv preprint arXiv:2002.00211*, 2020.

Jinhyun So, Başak Güler, and A. Salman Avestimehr. Byzantine-resilient secure federated learning. *IEEE Journal on Selected Areas in Communications*, 39(7):2168–2181, 2021. doi: 10.1109/JSAC.2020.3041404.

Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu. Data poisoning attacks against federated learning systems - github. https://github.com/git-disl/DataPoisoning_FL, 2020a.

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. Springer, 2020b.

Zhaoxian Wu, Qing Ling, Tianyi Chen, and Georgios B Giannakis. Byrd-saga - github. https://github.com/MrFive5555/Byrd-SAGA, 2020a.

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, 2020b.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. 2017.

Cong Xie, Sanmi Koyejo, and Indranil Gupta. Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance. In *International Conference on Machine Learning*, pages 6893–6901. PMLR, 2019.

Cong Xie, Sanmi Koyejo, and Indranil Gupta. Zeno++: Robust fully asynchronous sgd. In *International Conference on Machine Learning*, pages 10495–10503. PMLR, 2020.

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. PMLR, 2018.