A Robust and Efficient Federated Learning Algorithm against Adaptive Model Poisoning Attacks

Han Yang, Dongbing Gu, Jianhua He

Abstract—**With the undetectable characteristic, adaptive model poisoning attacks can combine with any other attacks, bypassing the detection and violating the availability of federated learning systems. Existing defences are vulnerable to adaptive model poisoning attacks, as model poisoning-related features are tailored to these methods and compromise the accuracy of the FL model. We first present a unified reformulation of existing adaptive model poisoning attacks. Analyzing the reformulated attacks, we find that the detectors should reduce the attacker's optimization cost functions to defeat adaptive attacks. However, existing defences do not consider the causes of model parameters' high dimensionality and data heterogeneity. We propose a novel robust FL algorithm, FedDet, to tackle the problems. By splitting the local models into layers for robust aggregation, FedDet can overcome the issue with high dimensionality while keeping the functionality of layers. During the robust aggregation, FedDet normalizes every slice of local models by the median norm value instead of excluding some clients, which can avoid deviation from the optimal model. Furthermore, we conduct a comprehensive security analysis of FedDet and an existing robust aggregation method. We propose the upper bounds on the perturbations disturbed by these adaptive attacks. It is found that FedDet can be more robust than Krum with a smaller perturbation upper bound under attacks. We evaluate the performance of FedDet and four baseline methods against these attacks under two classic datasets. It demonstrates that FedDet significantly outperforms the existing compared methods against adaptive attacks. FedDet can achieve 60.72% accuracy against min-max attacks.**

Index Terms—**Federated Learning, Model poisoning attacks, Deep Learning**

I. Introduction

The Internet of Things (IoT) plays an important role in our daily lives as it provides intelligent services and applications empowered by artificial intelligence (AI) [1] [2] [3]. AI techniques such as deep learning (DL) processes raw data generated from ubiquitous IoT devices and train data models for enabling intelligent services or infrastructures, such as smart healthcare, smart transportation, and smart city. Traditionally, AI functions are placed in a cloud server for data collecting and modeling [4] [5]. However, With such an explosive growth of IoT data at the network edge, the offloading of massive IoT data to remote servers may be infeasible due to the constrained network resources, bandwidth and incurred latency. Besides, the use of third-party servers for AI training also raises privacy concerns such as leakage of sensitive information (e.g., user addresses or personal preferences). Thus, it may not be feasible to apply centralized AI techniques in realistic scenarios. To address the above issues, a novel distributed training regime, federated learning (FL), has been proposed for building intelligent and privacy-enhanced IoT systems. FL is an efficient and scalable distributed machine learning paradigm that provides excellent privacy to clients [6]. With the application of federated learning, resource-constrained node devices (e.g., Internet of Things (IoT) devices and sensors) can build a knowledge-shared model while keeping the raw data local [7]. Hence, federated learning plays a critical role in bringing AI to IoT systems and applications in terms of training AI models, online model fine-tuning and preserving data privacy [4] [5]. However, due to its distributed characteristic, FL leaves the door open for adversaries as they can send poisoned local models to the central server without being checked. Hence, an FL system can be vulnerable to model poisoning attacks [8] [9] [10] [11]. Poisoning attacks consist of backdoor attacks [12] [13] [14] and model poisoning attacks [15] [16] [17] [18]. Backdoor attacks aim to insert a backdoor into the trained global model and make the global model mislabel a small group of samples with chosen triggers into targeted labels [16] [18] [19]. Model poisoning attacks attempt to hamper the global model's main accuracy. In this work, we focus on model poisoning attacks in FL as model poisoning attacks can cause denial-of-service among a large population of end services in FL deployments [20] [21] [22]. In this work, we investigate the model poisoning attacks against federated learning systems, which can be transferred into some resource-constrained scenarios (e.g., Multi-UAV Systems [23] and cause denial-of-service (DoS) in IoT systems). And we propose a robust algorithm to defend against such attacks.

The Byzantine-robust algorithms [24] [25] [26] [27] [11] [25] [26] [27] [28] have been discussed widely in the literature and perform well against general model poisoning attacks such as label-flipping attacks [16]. However, the adversary can optimize poisoning strategies when the adversary has knowledge of the aggregation methods [16] [17]. Such types of attacks are called adaptive model poisoning attacks. During the FL training, malicious clients can adaptively manipulate local model parameters tailored to the aggregation rule. By being well-designed, these local model parameters could bypass the defence methods like Krum [26] and compromise the FL training model.

In view of the above research issues, we are motivated to design an efficient, robust aggregation method and defeat

Corresponding author: Han Yang, (e-mail: hy20497@essex.ac.uk).

Han Yang, Dongbing Gu, and Jianhua He are with the School of Computer Science and Electronic Engineering, University of Essex, Essex, CO4 3SQ, UK (e-mail: hy20497@essex.ac.uk; dgu@essex.ac.uk; j.he@essex.ac.uk).

adaptive attacks. We first reformulate adaptive attacks and discuss their main characteristics. We found that the defender should reduce the attacker's optimization cost functions to defeat adaptive attacks. However, some existing defence methods like Krum or Muti-Krum [26] do not consider the strong impact of the curse of model parameters' high dimensionality. The attacker may update partial parameters or infrequent parameters. The attackers can cause negative effects when the value of the cost function is not large. Therefore, the adaptive attacks can bypass existing methods. On the other hand, some methods [27] ignore the data heterogeneity. They exclude potential malicious parameters or misclassified honest parameters, which cause deviation from the optimal global model. In this paper, we propose a Byzantine-robust FL method, FedDet, which consists of two main steps. In the first step, FedDet splits and groups local models by layers. Then, the sliced parameters in one group are normalized by the median of the norms. The first step, splitting, can decrease the high dimensionality of parameters. Besides, splitting by layers rather than random splitting [29] can keep the functionality of different layers. The second step, normalization, considers all parameters in the same group. We also discuss the certified robustness of FedDet based on an existing certified radius proposed by [30]. As an extension, we provide a detailed security analysis of FedDet and give the upper bounds of the perturbations given by the adaptive attacks. We evaluate the performance of FedDet against six types of attacks. Experiment results demonstrate that FedDet outperforms other baseline works against adaptive attacks.

The main contributions of our work are summarized as follows:

- We present a unified reformulation of existing adaptive model poisoning attacks. The summary of the reformulation can be used for related works to verify the efficiency of their methods against adaptive attacks. To the best of our knowledge, no existing works give such a comprehensive discussion of state-of-the-art adaptive model poisoning attacks.
- Based on our discussion of the main characteristics of adaptive model poisoning attacks, we reveal the two main causes of why existing defence methods are not efficient: model parameters' high dimensionality and data heterogeneity. Then we propose FedDet, consisting of two steps, with the first step splitting overcoming the issue of high dimensionality and the second step normalizing overcoming the issue of heterogeneity.
- We evaluate FedDet against six designed adaptive attacks tailored to it. From the results, FedDet significantly outperforms baseline works against adaptive attacks. Besides, we discuss the certified radius of FedDet. As an extension, we provide a detailed security analysis of FedDet and an existing robust aggregation method, Krum. By comparing the upper bounds of the perturbations caused by DNY-OPT attacks corresponding to these two robust methods, we can see that FedDet outperforms Krum according to the upper bounds.

II. Related Works

The principle of existing Byzantine-robust defences [26] [27] is to train a global model with high performance, even if there are some malicious clients.

Krum [26] attempts to select a representative as the aggregated model update for every training round. Suppose there are n chosen local clients in every training round. And m clients among local clients are malicious. The score for the ith client is calculated as $s_i = \sum_{\mathbf{w}_j \in \Gamma_{i,n-m-2}} \|\mathbf{w}_j - \mathbf{w}_i\|_2^2$, where $\Gamma_{i,n-m-2}$ is the set of $n-m-2$ local clients that have the smallest Euclidean distance to $\mathbf{w}_i$ (client$i$'s parameters). So, the client with the smallest score will be selected as the representative. This representative model update will be the global model for the next training round.

Multi-Krum [26] is a variant of Krum. Multi-Krum collects a set of clients with the smallest scores using Krum and repeats this process for the remaining updates until the set has c updates, such as $n - c > 2m + 2$. Then, it takes the average among this set of clients.

Median [27] is a coordinate-wise aggregation rule. The coordinate-wise median of sorted local models is selected as the aggregated global model update. Instead of using the mean value among local clients, this aggregation rule considers the coordinate median value of the parameters as the corresponding parameter in the global model for the next iteration. The coordinate-wise median is agnostic to the actual malicious rates.

Trimmed-mean [27] is another coordinate-wise aggregation rule. Suppose the trimmed parameter is $k < \frac{n}{2}$. The server removes the k maximum and minimum coordinates in the model updates and then uses FedAvg to aggregate the remaining parameters for the next training round. Trimmed-mean relies on the assumption that the coordinate of the attacker would either be the minimum or the maximum value of the corresponding parameters. However, this assumption does not hold for model poisoning attacks [30]. Therefore, even a single attacker can compromise the trimmed mean. Unlike the coordinate-wise median, the Trimmed-median uses exact knowledge of the malicious rates.

III. Reformulation of previous adaptive attacks

This section introduces four state-of-the-art adaptive attacks that can optimize local model poisoning attacks for any given aggregation rules. Although these adaptive attacks are proposed in [16] [17], they do not have an identical formulation. Out of convenience, We reformulate these adaptive attacks and list a table I to describe the optimization formulations of these attacks. In the following paragraphs, we give a detailed description of the formulations. They can be used for any work focusing on robust aggregation methods to verify the robustness of their methods against adaptive attacks.

Static Optimization (STAT-OPT) Attack [16]: STAT-OPT attacks consider the attacker's objective to deviate global model parameters the most towards the inverse of the direction along which the global parameters would change without attacks. Suppose that in one global training process, G denotes a set of the aggregated global parameters without attacks,

Adaptive attacks	Optimization cost functions	Constraints
STAT-OPT attacks	$\max\limits_{\mathbf{w}_1^{'},..,\mathbf{w}_m^{'}} \mathbf{s}^T(G-G^{'})$	$s.t.\quad G=f_{agr}(\mathbf{w}_1,..,\mathbf{w}_m,\mathbf{w}_{m+1},..,\mathbf{w}_n)$ $G^{'}=f_{agr}(\mathbf{w}_1^{'},..,\mathbf{w}_m^{'},\mathbf{w}_{m+1},..,\mathbf{w}_n)$
DNY-OPT attacks	$\max\limits_{\gamma,\nabla^p} \lVert G-G^{'}\rVert_2$	$s.t.\quad G=f_{avg}(\mathbf{w}_1,..,\mathbf{w}_m,\mathbf{w}_{m+1},..,\mathbf{w}_n)$ $G^{'}=f_{agr}(\mathbf{w}_1^{'},..,\mathbf{w}_m^{'},\mathbf{w}_{m+1},..,\mathbf{w}_n)$ $\mathbf{w}_{i\in[m]}^{'}=G+\gamma\nabla^p$
Min-max attacks	$\max\limits_{\gamma,i\in[m+1,n]} \lVert\mathbf{w}^{'}-\mathbf{w}_i\rVert_2$	$s.t.\quad \mathbf{w}_{i\in[m]}^{'}=G+\gamma\nabla^p$ $\lVert\mathbf{w}^{'}-\mathbf{w}_i\rVert_2 \leq \underset{i,j\in[m+1,n]}{maximum} \lVert\mathbf{w}_i-\mathbf{w}_j\rVert_2$
Min-sum attacks	$\max\limits_{\gamma}\sum_{i\in[m+1,n]}\lVert\mathbf{w}^{'}-\mathbf{w}_i\rVert_2^2$	$s.t.\quad \mathbf{w}_{i\in[m]}^{'}=G+\gamma\nabla^p$ $\sum_{i\in[m+1,n]}\lVert\mathbf{w}^{'}-\mathbf{w}_i\rVert_2^2 \leq maximum\sum_{i,j\in[m+1,n]}\lVert\mathbf{w}_i-\mathbf{w}_j\rVert_2^2$

TABLE I: Summary of reformulation of existing adaptive attacks

and $G^{'}$ denotes the compromised global parameters. $\mathbf{s}^T$ is the column vector of changing directions of all global parameters without attacks. Then, the cost function $\mathbf{s}^T(G-G^{'})$ (see table I) measures the direction deviation. The attacker's goal is to maximize the value of $\mathbf{s}^T(G-G^{'})$. $\mathbf{w}_1,..,\mathbf{w}_n$ denotes a set of the model parameters shared by the clients in one training process and f_{agr} denotes the robust aggregation method, which the attacker aims to compromise. The first m clients $\mathbf{w}_1^{'},..,\mathbf{w}_m^{'}$ are assumed to be compromised. The attacker aims to find an optimal set of values for $\mathbf{w}_1^{'},..,\mathbf{w}_m^{'}$ and substitute for the benign parameters $\mathbf{w}_1,..,\mathbf{w}_m$. After replacing these benign parameters with malicious parameters $\mathbf{w}_1^{'},..,\mathbf{w}_m^{'}$, the deviation between the compromised global parameters and the benign global parameters in the directions can be maximized.

Dynamic Optimization (DYN-OPT) Attack [17]: DYN-OPT attacks aim to decrease the similarity between compromised and benign global models. The cost function is $\|G-G^{'}\|_2$, where $\|\|_2$ is the L_2-norm value. The attacker aims to maximize the $\|G-G^{'}\|_2$ value (see table I). Unlike STAT-OPT Attack, it restricts malicious models as $\mathbf{w}_{i\in[m]}^{'}=G+\gamma\nabla^p$, where γ is the scaling factor and ∇^p is the perturbation vector. [17] introduces three types of perturbation vectors: Inverse unit vector, Inverse standard deviation and Inverse sign. In this work, we consider the Inverse sign. Other perturbation vectors will be discussed in further works.

AGR-agnostic Attacks [17], Min-Max: Previous robust FL algorithms attempt to distinguish malicious parameters from benign ones based on two main criteria: 1) distances between malicious and benign parameters such as cosine similarities [31] [32], 2) difference in L_p-norms of malicious and benign parameters. To bypass these robust FL algorithms, the attacker must ensure that the malicious parameters lie close to the cluster of benign parameters while maximizing the distance or difference in L_p-norm from benign parameters. The cost function is $\|\mathbf{w}^{'}-\mathbf{w}_i\|_2$ (see table I). The attacker aims to maximize the distance of the malicious parameters from benign parameters. The constraint is the distance from benign parameters should be smaller than the maximum of benign parameter distances.

AGR-agnostic Attacks [17], Min-Sum: Min-Max attack maximises the distance of malicious updates from benign model updates while ensuring the maximum distance from other benign updates is upper bounded by the maximum distance between any two benign updates. Like Min-Max, Min-sum ensures that the sum of squared distances of malicious gradients from all the benign updates is upper bound by the sum of the squared distances of any benign updates from the other benign updates. The cost function is $\max\limits_{\gamma}\sum_{i\in[m+1,n]}\|\mathbf{w}^{'}-\mathbf{w}_i\|_2^2$ (see table I). The constraint is $\sum_{i\in[m+1,n]}\|\mathbf{w}^{'}-\mathbf{w}_i\|_2^2 \leq maximum\sum_{i,j\in[m+1,n]}\|\mathbf{w}_i-\mathbf{w}_j\|_2^2$.

IV. Proposed Defence approaches

According to table I, the defender should reduce the optimization objectives to defeat the adaptive attacks. For example, the L_2-norm distance of the robust aggregated $G^{'}$ should be close to the benign G when the defender attempts to reduce the negative impact of DNY-OPT attacks. As for Min-max or Min-sum attacks, the defender should try to reduce the distance between the malicious and benign parameters ($\max\limits_{\gamma,i\in[m+1,n]}\|\mathbf{w}^{'}-\mathbf{w}_i\|_2$ or $\max\limits_{\gamma}\sum_{i\in[m+1,n]}\|\mathbf{w}^{'}-\mathbf{w}_i\|_2^2$). However, there are two root causes why existing robust methods fail to defend: high dimensional parameters and data heterogeneity.

(1) the High dimensional parameters. The attacker can only choose partial parameters to alter or poison. In [33], the attacker only poisons the unused or infrequently updated parameters by benign clients. Such attack behaviours are more stealthy when the training models contain many parameters. The attacker can cause negative impacts when the value of optimization objectives is not large. Therefore, the attacker can bypass existing methods based on distance or similarity comparisons of full model parameters, such as Krum [26], Multi-Krum [26], Flame [31], FLtrust [32].

(2) the data heterogeneity. Some previous methods are parameter-wise, like Median [27] and Trimmed-mean [27]. However, these methods may not consider the cause of data heterogeneity. The Median selects a median value to represent the global parameter. On the other hand, Trimmed-mean prunes a list of potential malicious parameters. Hence, the aggregated model may deviate from the optimal training model. Therefore, robust methods should try to take all clients' parameters into consideration.

Based on the above discussion, we propose FedDet, a novel Byzantine-robust FL algorithm, FedDet. This robust method groups the clients' parameters by layers. Then, it normalizes the layer-wise parameters by the median norms. Using this layer-wise robust aggregation method, FedDet can avoid the curse of high dimensions. Unlike [29] splitting the parameters with random fragments, we choose to split the parameters by layers. Our splitting skill can keep the functionality of layers compared to random splitting. Besides, the layer-wise normalization considers all clients' corresponding parameters rather than filtering potential malicious or misclassified honest parameters.

Now, we describe the details of FedDet. Suppose that the local model is a neural network with l layers.

(1) Splitting and Grouping the model parameters $\mathbf{w}_{\{1,...,n\}}$ by layers.

The server collects a list of parameters of ith layer from all clients $\mathbf{w}_{\{1,...,n\}}$, i.e.,

$$G_{i\in[l]} = \{\mathbf{w}_{1,i}, \mathbf{w}_{2,i}, ..., \mathbf{w}_{n,i}\}, \tag{1}$$

The server collects all the groups of the split parameters $G_{\{1,...,l\}}$.

(2) Normalizing the ith layer clients' parameters $\mathbf{w}_{\{1,...,n\},i}$ by the median $L2$-norm value.

Firstly, the server collects a list of the $L2$-norm values of every layer L_i from all clients $\mathbf{w}_{\{1,...,n\}}$, i.e.,

$$L_{i\in[l]} = \{\|\mathbf{w}_{1,i}\|_2, \|\mathbf{w}_{2,i}\|_2, ..., \|\mathbf{w}_{n,i}\|_2\}, \tag{2}$$

where, $\mathbf{w}_{j,i}$ denotes the i-th layer of $\mathbf{w}_i \in \mathbf{w}_{\{1,...,n\}}$, and $\|\mathbf{w}_{j,i}\|_2$ denotes the $L2$-norm values of the corresponding $\mathbf{w}_{j,i}$. And $L_{i\in[l]}$ is the set containing all the $L2$-norm values. Then, it sorts out all the clients by their $L2$-norm in ascending order, i.e.,

$$L_{i\in[l]} = \{\|\mathbf{w}_{s1,i}\|_2, \|\mathbf{w}_{s2,i}\|_2, ..., \|\mathbf{w}_{sn,i}\|_2\}. \tag{3}$$

Here, $\|\mathbf{w}_{s1,i}\|_2 \leq \|\mathbf{w}_{s2,i}\|_2 \leq ... \leq \|\mathbf{w}_{sn,i}\|_2$. Then, the server selects the median value $med(L_{i\in[l]})$ of $L_{i\in[l]}$. The server collects all the median values corresponding to each layer. The collection of $med(L_{i\in[l]})$ is represented as follows,

$$H' = \{med(L_1), med(L_2), ..., med(L_l)\}. \tag{4}$$

Then, the server scales the L_2-norm values of all layer's parameters $L_{\{1,...,l\}}$ of clients by H'. The scaled L_2-norm values of the ith layer's updates can be represented as follows,

$$\tilde{L}_{i\in[l]} = \left\{\frac{med(L_i)}{\|\mathbf{w}_{1,i}\|_2}, \frac{med(L_i)}{\|\mathbf{w}_{2,i}\|_2}, ..., \frac{med(L_i)}{\|\mathbf{w}_{n,i}\|_2}\right\}. \tag{5}$$

Then, the weighted updates of the ith layer can be represented as follows,

$$\left\{\frac{med(L_i)}{\|\mathbf{w}_{1,i}\|_2}\mathbf{w}_{1,i}, \frac{med(L_i)}{\|\mathbf{w}_{2,i}\|_2}\mathbf{w}_{2,i}, ..., \frac{med(L_i)}{\|\mathbf{w}_{n,i}\|_2}\mathbf{w}_{n,i}\right\}. \tag{6}$$

We repeat the same process for all l layers. Then, the server executes the FedAvg algorithm on each layer to obtain the new global model.

V. Adaptive attacks

In this section, we leverage the adaptive attacks discussed in III to design adaptive untargeted attacks for the proposed defence method.

A. STAT-OPT tailored to FedDet

The idea is to instantiate the aggregation rule f_{agr} with our proposed aggregation rule, *FedDet* in the poisoning framework. So we formulate a specific optimization problem using table. I (STAT-OPT attacks) as follows:

$$\sum_{i=1}^{l} \max_{\mathbf{w}'_{1,i},...,\mathbf{w}'_{m,i}} \mathbf{s}_i^T(G_i - G'_i). \tag{7}$$

As the proposed aggregation rule is layer-wise, unlike Krum [26] or Median [27], we solve the optimization layer by layer and optimize $\mathbf{w}'_{1,i}, ..., \mathbf{w}'_{m,i}$, where i denotes the ith layer. Then we concatenate all $\mathbf{w}'_{1,i}, ..., \mathbf{w}'_{m,i}$ by layers and get the final solutions $\mathbf{w}'_1, ..., \mathbf{w}'_m$.

In IV, the proposed aggregation rule for ith layer can be written as follows:

$$G_i = \frac{1}{n}\left(\frac{med(L_i)}{\|\mathbf{w}_{1,i}\|}\mathbf{w}_{1,i} + \frac{med(L_i)}{\|\mathbf{w}_{2,i}\|}\mathbf{w}_{2,i}+, ..., + \frac{med(L_i)}{\|\mathbf{w}_{n,i}\|}\mathbf{w}_{n,i}\right). \tag{8}$$

We denote by $e_{j,i} = \frac{\mathbf{w}_{j,i}}{\|\mathbf{w}_{j,i}\|}$. So Eq. (8) can be rewritten as follows:

$$G_i = \frac{1}{n}\sum_{j\in[1,n]} med(L_i)e_{j,i}. \tag{9}$$

Let $e'_{j,i}(j \in [1,m])$ denote the poisoned unit vector sent by malicious clients. So, the poisoned aggregated parameters can be rewritten as follows:

$$G'_i = \frac{1}{n}\left(\sum_{j\in[1,m]} med(L'_i)e'_{j,i} + \sum_{j\in[m+1,n]} med(L'_i)e_{j,i}\right), \tag{10}$$

where L'_i is $\{\|\mathbf{w}'_{1,i}\|_2, ..., \|\mathbf{w}'_{m,i}\|_2, \|\mathbf{w}_{m+1,i}\|_2 \cdots, \|\mathbf{w}_{n,i}\|_2\}$ and $med(L'_i)$ is the new median after poisoning.

We substitute Eq. (9) and (10) into (7) and get the following optimization problem:

$$\begin{aligned}&\sum_{i=1}^{l} \ell(\mathbf{e}'_{1,i}, ..., \mathbf{e}'_{m,i}) = \\ &\frac{1}{n}\sum_{i=1}^{l} \max_{\mathbf{e}'_{1,i},...,\mathbf{e}'_{m,i}} \mathbf{s}_i^T(\sum_{j\in[1,n]} med(L_i)e_{j,i} - \sum_{j\in[1,m]} med(L'_i)e'_{j,i} \\ &- \sum_{j\in[m+1,n]} med(L'_i)e_{j,i}).\end{aligned} \tag{11}$$

We consider a strong attacker who knows $e_{j\in[1,n],i\in[1,l]}$, $e^{'}_{j\in[1,m],i\in[1,l]}$, and f_{agr}. We use a standard gradient ascent approach to solve the optimization problem 11. We optimize $\mathbf{e}^{'}_{1,i}, ..., \mathbf{e}^{'}_{m,i}$ one by one. When optimizing $e^{'}_{j,i}$, all other $e^{'}_{k\neq j,i}$ are fixed. The steps are as follows:

1) : Computing the gradient $\nabla_{e^{'}_i}\ell$ with respect to $e^{'}_i$: As it is hard to compute this gradient directly, we use a standard method, a zeroth-order method [34], to estimate this gradient as follows:

$$\nabla_{e^{'}_{j,i}}\ell \approx \frac{\ell(e^{'}_{j,i}+\gamma\mathbf{u})-\ell(e^{'}_{j,i})}{\gamma}\cdot\mathbf{u}, \tag{12}$$

where ℓ denotes the eq. (11). Where $\mathbf{u}$ is a random vector sampled from the multivariate Gaussian distribution $\mathcal{N}(0,\sigma^2 I)$ and $\gamma>0$ is a smoothing parameter.

2) : Updating $e^{'}_{j,i}$: We multiply the estimated gradient by a learning rate η and add it to $e^{'}_{j,i}$. Then we normalize $e^{'}_{j,i}$ by its L_2 norm value to ensure it is a unit vector.

$$e^{'}_{j,i} = e^{'}_{j,i} + \eta\nabla_{e^{'}_{j,i}}\ell. \tag{13}$$

When estimating the gradient $\nabla_{e^{'}_{j,i}}\ell$ and updating $e^{'}_{j,i}$, we fix the value of $med(L^{'}_i)$ for simplicity. The $med(L^{'}_i)$ value will be updated after $e^{'}_{j,i}$ is updated.

3) : Repeating the above two steps for n iterations. The $\mathbf{w}^{'}_{j,i} = med(L^{'}_i)\cdot e^{'}_{j,i}$ after $e^{'}_{j,i}$ is solved.

4) : Repeating the gradient ascent process over all $\mathbf{e}^{'}_{1,i}, ..., \mathbf{e}^{'}_{m,i}$.

The procedure is summarized in Algorithm. (1). We initialize $\mathbf{w}^{'}_{1,i}, ..., \mathbf{w}^{'}_{m,i}$ using Trim attack in [16].

Algorithm 1: STAT-OPT tailored to FedDet

```
Input: $\mathbf{w}^{'}_{1,i}, ..., \mathbf{w}^{'}_{m,i}$, $l$, $\mathbf{s}_i$
// $\mathbf{w}^{'}_{1,i}, ..., \mathbf{w}^{'}_{m,i}$ are a list of initialized
   malicious updates; $l$ is the number
   of layers; $\mathbf{s}_i$ is the direction
   along which the global parameter
   would change without attacks.
Output: $\mathbf{w}^{'}_{1,i}, ..., \mathbf{w}^{'}_{m,i}$
for $i \in [0, l]$ // optimization per layer
  do
    for $j \in [1, m]$ // optimization $e^{'}_{j,i}$ one by
      one
    do
        for iterations $\in n$ do
            Random sample $\mathbf{u}$ $\mathcal{N}(0,\sigma^2 I)$;
            $\nabla_{e^{'}_{j,i}}\ell \approx \frac{\ell(e^{'}_{j,i}+\gamma\mathbf{u})-\ell(e^{'}_{j,i})}{\gamma}\cdot\mathbf{u}$;
            $e^{'}_{j,i} = e^{'}_{j,i} + \eta\nabla_{e^{'}_{j,i}}\ell$;
        end
        $\mathbf{w}^{'}_{j,i} = med(L^{'}_i)\cdot e^{'}_{j,i}$;
    end
end
```

B. DYN-OPT tailored to FedDet

As discussed in III, DYN-OPT attacks restrict the malicious clients as $\mathbf{w}^{'}_{i\in[m]} = G + \gamma\nabla^p$, where ∇^p is the perturbation vector. We instantiate the aggregation rule f_{agr} in table. I (DNY-OPT attacks) with our proposed aggregation method. Unlike STAT-OPT attacks, DYN-OPT is a plug-in adaptive attack framework for robust aggregation algorithms. We use Algorithm.1 in [17] to solve the optimal γ value.

We assume a strong attacker who knows $\mathbf{w}_{i\in[1,n]}$ and f_{agr}. Hence, the attackers can estimate the perturbation vectors based on $\mathbf{w}_{i\in[1,n]}$. We also consider a weaker attacker who has no knowledge of $\mathbf{w}_{i\in[1,n]}$ and f_{agr}. We compare the efficiency of FedDet in these two different adversary models in the experiment VIII-C.

C. AGR-agnostic attacks tailored to FedDet

AGR-agnostic attacks do not know the aggregation rules. So, these agnostic attacks can be applied in various robust aggregation algorithms. We use the attack methods in [17] to test the proposed FedDet.

We assume a strong attacker who knows $\mathbf{w}_{i\in[1,n]}$ and f_{agr}. We also consider a weaker attacker who has no knowledge of $\mathbf{w}_{i\in[1,n]}$ and f_{agr}. We consider both adversary models for evaluating FedDet's efficiency.

VI. Security analysis of FedDet

To conduct the security analysis of FedDet, we fit FedDet into the theoretical framework of [30]. Firstly, we briefly describe the definition of poisoning attacks and the certified radius proposed by [30]. Here are the notation, definitions, and assumptions.

Notation 1. *Let $\boldsymbol{Z}$ be the data domain and $\boldsymbol{D^t}$ be the data sampled (not necessarily i.i.d) from $\boldsymbol{Z}$ at iteration t. Let $\mathcal{L}:\Theta\times\boldsymbol{Z}^*\rightarrow\mathcal{R}$ be a loss function and Θ be the class of models with d dimensions. Let $f=(\mathcal{G},\mathcal{A},\lambda(t))$ be the federated learning protocol with update algorithm $\mathcal{A}:\mathbf{w}^t\in\mathcal{R}^d\rightarrow\mathcal{R}^d$ and $\mathcal{G}(G,\boldsymbol{D},t)\rightarrow\mathbf{w}^t$ that takes a model G and outputs the update $\mathbf{w}^t$. $G^{t+1}=G^t-\lambda(t)\mathcal{A}(\mathbf{w}^t)$ is the updates rule of the FL protocol. For the proposed FedDet, $A(\mathbf{w}^t)=\mathbf{w}^t$.*

Definition 1. *(poisoning attacks) Let $f^*=(\mathcal{G}^{'},\mathcal{A},\lambda(t))$ be the poisoned federated learning protocol with poisoned $\mathcal{G}^{'}(G,\boldsymbol{D},t)\rightarrow\mathbf{w}^{'t}$. We have $\mathcal{G}^{'}(G,\boldsymbol{D},t)=\mathcal{G}(G,\boldsymbol{D},t)+\epsilon$ with $\|\epsilon\|_1\leq\rho$ (or $\|\epsilon\|_2\leq\rho$).*

Notation 2. *We use $(G^0,...,G^t)$ and $(G^{'0},...,G^{'t})$ to denote the global model trained through a benign G and a poisoned $G^{'}$ respectively. We use $(\mathbf{w}^1,...,\mathbf{w}^{t+1})$ and $(\mathbf{w}^{'1},...,\mathbf{w}^{'t+1})$ to denote the updates produced by a benign G belongs to global models $(G^0,...,G^t)$ and by a poisoned $G^{'}$ belongs to global models $(G^{'0},...,G^{'t})$. We use $(\mathbf{w}^{*1},...,\mathbf{w}^{*t+1})$ to denote the poisoned updates produced by a poisoned $G^{'}$ belongs to models $(G^{'0},...,G^{'t})$.*

Assumption 1. *A protocol $f(\mathcal{G},\mathcal{A},\lambda(t))$ is a c-layerwise-Lipschitz. Specifically, for any layer index $i\in[L]$*

$$\|\mathcal{G}(G^{'t},\boldsymbol{D},t)[i]-\mathcal{G}(G^t,\boldsymbol{D},t)[i]\|\leq c\cdot\|G^{'t}-G^t\|. \tag{14}$$

Theorem 1. *Let FedDet be a c-layerwise-Lipschitz protocol on a dataset $\boldsymbol{D}$. Then $\boldsymbol{R}(\rho) = \Lambda(\boldsymbol{T})(1+dc)^{\Lambda(\boldsymbol{T})}\rho$ is a certified radius for f. [30] Namely,*

$$\|G^{'T} - G^{T}\| \leq \Lambda(\boldsymbol{T})(1+dc)^{\Lambda(\boldsymbol{T})}\rho. \tag{15}$$

From Equation. (15), it is not difficult to see the certified radius $R(\rho)$ relies on ρ when the $\Lambda(\boldsymbol{T})$, l and c are fixed. Now, we analyze how these adaptive attacks can disturb FedDet with maximum ρ. We attempt to give an upper bound on ρ in the following subsections.

A. Security analysis for FedDet against STAT-OPT attacks

In subsection V-A, we discuss how STAT-OPT can be adapted to FedDet. Based on Equation. (8), we have the benign aggregated parameters per layer:

$$G_i = \frac{1}{n}(\frac{med(L_i)}{\|\mathbf{w}_{1,i}\|}\mathbf{w}_{1,i} + \frac{med(L_i)}{\|\mathbf{w}_{2,i}\|}\mathbf{w}_{2,i} +, ..., + \frac{med(L_i)}{\|\mathbf{w}_{n,i}\|}\mathbf{w}_{n,i}), \tag{16}$$

with $L_i = \{\|\mathbf{w}_{1,i}\|, ..., \|\mathbf{w}_{m,i}\|, ..., \|\mathbf{w}_{n,i}\|\}$ and $med(L_i) = l$. The poisoned aggregated parameters per layer are as follows:

$$G_i^{'} = \frac{1}{n}(\frac{med(L_i^{'})}{\|\mathbf{w}_{1,i}^{'}\|}\mathbf{w}_{1,i}^{'} + \frac{med(L_i^{'})}{\|\mathbf{w}_{2,i}^{'}\|}\mathbf{w}_{2,i}^{'} +, ..., + \frac{med(L_i^{'})}{\|\mathbf{w}_{n,i}\|}\mathbf{w}_{n,i}), \tag{17}$$

with $L_i = \{\|\mathbf{w}_{1,i}^{'}\|, ..., \|\mathbf{w}_{m,i}^{'}\|, ..., \|\mathbf{w}_{n,i}\|\}$ and $med(L_i^{'}) = l^{'}$.

Assumption 2. *To avoid the impact of attack being restricted, $\|\mathbf{w}_{1,i}^{'}\|, ..., \|\mathbf{w}_{m,i}^{'}\|$ should be close to the median value of $\{\|\mathbf{w}_{1,i}^{'}\|, ..., \|\mathbf{w}_{m,i}^{'}\|, ..., \|\mathbf{w}_{n,i}\|\}$. Hence, we assume that $med(L_i^{'}) = \|\mathbf{w}_{1,i}^{'}\| = \|\mathbf{w}_{2,i}^{'}\| =, ..., = \|\mathbf{w}_{m,i}^{'}\| = l^{'}$.*

Theorem 2. *Suppose FedDet is a c-layerwise-Lipschitz protocol on dataset $\boldsymbol{D}$ and Assumption 2 holds. Suppose m out of n clients are potentially malicious at one round. The upper bound on perturbation ρ caused by STAT-OPT attack on FedDet is given as*

$$\begin{aligned}\rho \leq \frac{n}{n-cm}\left|\mathbf{w}_{1,i} - G_{i-1}\right| + \frac{cm}{n-cm}\left|G_{i-1}\right| \\ + \frac{c}{n-cm}\left|\sum_{i=m+1}^{n}\frac{\|\mathbf{w}_i\|_{max}}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i} - \sum_{i=1}^{n}\frac{l}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i}\right|.\end{aligned} \tag{18}$$

For the proof of theorem. 2, please see appendix. X-A.

B. Security analysis for FedDet against DNY-OPT attacks

According to the definition of fedAvg [6], we have the benign aggregated parameters per layer as follows:

$$G_i = \frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_{i,i}. \tag{19}$$

Based on Equation. (8), we have the poisoned aggregated parameters per layer as follows:

$$G_i^{'} = \frac{1}{n}(\frac{med(L_i^{'})}{\|\mathbf{w}_{1,i}^{'}\|}\mathbf{w}_{1,i}^{'} + \frac{med(L_i^{'})}{\|\mathbf{w}_{2,i}^{'}\|}\mathbf{w}_{2,i}^{'} +, ..., + \frac{med(L_i^{'})}{\|\mathbf{w}_{n,i}\|}\mathbf{w}_{n,i}). \tag{20}$$

Theorem 3. *Suppose FedDet is a c-layerwise-Lipschitz protocol on dataset $\boldsymbol{D}$ and Assumption 2 holds. Suppose m out of n clients are potentially malicious at one round. The upper bound on perturbation ρ caused by DNY-OPT attack (DPAs and PGA attacks) on FedDet is given as*

$$\rho \leq \|\mathbf{w}_{1,i} - G_{i-1}\| + \frac{c}{n-cm} \cdot \sum_{i=m+1}^{n}\|(\frac{l^{'}}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i} - \mathbf{w}_{i,i})\|. \tag{21}$$

For the proof of theorem 3, please see appendix. X-B.

C. Security analysis for FedDet against Agnostic attacks

In section III, we introduce AGR-agnostic attacks. Now, we discuss the possible upper bound on ρ for the Min-max attacks.

Theorem 4. *Suppose m out of n clients are potentially malicious at one round. The upper bound on perturbation ρ caused by Min-max attacks on FedAvg is given as*

$$\rho \leq \|\mathbf{w}_{1,i} - G_{i-1}\| + \underset{i,j\in[m+1,n]}{max}\|\mathbf{w}_i - \mathbf{w}_j\|. \tag{22}$$

The security analysis of the Min-sum attacks is similar to that of the Min-max attacks.

Theorem 5. *Suppose m out of n clients are potentially malicious at one round. The upper bound on perturbation ρ caused by Min-sum attacks on FedAvg is given as*

$$\begin{aligned}\rho \leq \sqrt{\frac{1}{n-m}} \cdot \\ \sqrt{(\sum_{i=m+1}^{n}\|\mathbf{w}_{1,i} - G_{i-1}\|^2 + max\sum_{i,j\in[m+1,n]}\|\mathbf{w}_i - \mathbf{w}_j\|^2)}.\end{aligned} \tag{23}$$

For the proof of theorem. 4 and theorem. 5, please see appendix. X-C.

D. Security analysis of Krum against DNY-OPT attacks

As a comparison, we also establish the security analysis of Krum [26]. In [16] [17], they design similar adaptive attack strategies to compromise Krum. So, the discussion of the security analysis of Krum does not need to be separated into different situations.

According to the definition of Krum [26], malicious clients' parameters $\mathbf{w}^{'}$ should satisfy that the sum of the squared distances to its closest $n-m$ parameters is the smallest if the malicious parameters $\mathbf{w}^{'}$ could be selected as the next representative global parameters. Namely,

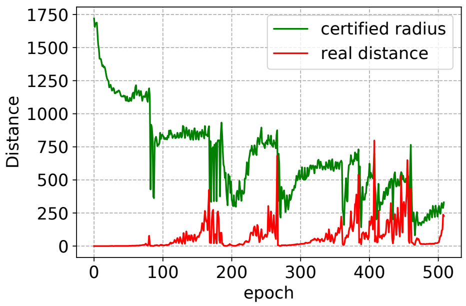


Fig. 1: Comparison between the real distance $\|G^{'T} - G^{T}\|$ and certified radius

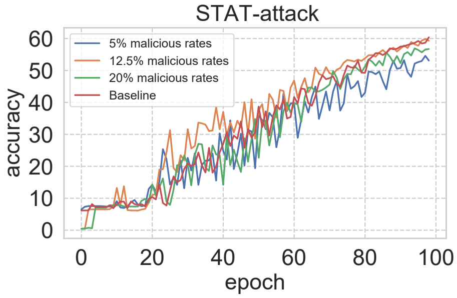


Fig. 2: Performance of FedDet against STAT-OPT attacks, PDAs and PGA attacks

$$\sum_{i\in\Gamma^{n-m-2}_{\mathbf{w}'}} \|\mathbf{w}^{'} - \mathbf{w}_i\| \le \min_{j\in[m+1,n]} \sum_{i\in\Gamma^{n-m-2}_{\mathbf{w}}} \|\mathbf{w}_j - \mathbf{w}_i\|, \quad (24)$$

where $i \in \Gamma^{b}_{\mathbf{w}'}$ denotes a set of parameters that are closest to $\mathbf{w}^{'}$.

Theorem 6. *Suppose m out of n clients are potentially malicious at one round. The upper bound on perturbation ρ caused by adaptive attacks on Krum is given as*

$$\rho \le \frac{1}{n-2m-1} \min_{j\in[m+1,n]} \sum_{i\in\Gamma^{n-m-2}_{\mathbf{w}}} \|\mathbf{w}_{j,i} - \mathbf{w}_{i,i}\| + \max_{i\in[m+1,n]} \|\mathbf{w}_{i,i} - G_{i-1}\|. \quad (25)$$

For the proof of theorem. 6, please see appendix. X-D.

E. Further analysis

We further analyse the robust aggregation methods' upper bounds on perturbation ρ under various adaptive attacks. We give a table II that collects all the upper bounds on ρ.

*1) **FedDet versus Krum**:* We notice that the right side of 3 in table II can be further replaced as below:

$$\le \|\mathbf{w}_{1,i} - G_{i-1}\| + \frac{cn}{n-cm} \max_{i\in[1,n]} \|(\frac{l^{'}}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i} - \mathbf{w}_{i,i})\|. \quad (26)$$

The coefficient of the right part of (26) $\frac{cn}{n-cm}$ is always less than one when $c < 1$. In table. II, the coefficient of the right part of the right side of (6) $\frac{n-m-2}{n-2m-1}$ is greater than one when $m > 1$. Besides, The left part of the right side of (6) $max_{i\in[m+1,n]} \|\mathbf{w}_{i,i} - G_{i-1}\|$ is larger than the left part of (26) $\|\mathbf{w}_{1,i} - G_{i-1}\|$, so the upper bound on ρ of Krum is larger than the upper bound of FedDet against DNY-OPT attacks. Therefore, theoretically, Krum is more likely to be compromised than FedDet.

*2) **Min-max versus Min-sum**:* Agnostic attacks can be performed in any robust aggregated FL system since these attack strategies do not require knowledge of the aggregation method. In Table.II, the right side of (5) can be further replaced as

$$\rho \le \sqrt{(\max_{i\in[m+1,n]} \|\mathbf{w}_{1,i} - G_{i-1}\|^2 + \max_{i,j\in[m+1,n]} \|\mathbf{w}_i - \mathbf{w}_j\|^2)}, \quad (27)$$

then we have

$$\rho^2 \le (\max_{i\in[m+1,n]} \|\mathbf{w}_{1,i} - G_{i-1}\|^2 + \max_{i,j\in[m+1,n]} \|\mathbf{w}_i - \mathbf{w}_j\|^2). \quad (28)$$

In Table.II, 4 can be further converted to

$$\begin{aligned} \rho^2 &\le (\|\mathbf{w}_{1,i} - G_{i-1}\| + \max_{i,j\in[m+1,n]} \|\mathbf{w}_i - \mathbf{w}_j\|)^2 \\ &\le \|\mathbf{w}_{1,i} - G_{i-1}\|^2 + \max_{i,j\in[m+1,n]} \|\mathbf{w}_i - \mathbf{w}_j\|^2 \\ &+ 2\cdot(\max_{i,j\in[m+1,n]} \|\mathbf{w}_i - \mathbf{w}_j\| \cdot \|\mathbf{w}_{1,i} - G_{i-1}\|). \end{aligned} \quad (29)$$

Compared to the right side of (28) and (29), Min-max attacks may incur worse perturbation errors to aggregation methods as the upper bound on perturbation ρ caused by Min-max attacks has an extra item $2\cdot(\max_{i,j\in[m+1,n]} \|\mathbf{w}_i - \mathbf{w}_j\| \cdot \|\mathbf{w}_{1,i} - G_{i-1}\|)$.

*3) **DNY-OPT attacks versus Agnostic attacks**:* In theorem 4, we propose the perturbation error caused by Min-max attacks on FedAvg, which is not robust to malicious attacks. This perturbation can be reduced when robust aggregation methods are applied in FL training. It is not difficult to see that the coefficient of the right part of the right side of (6) is larger than the right side of (4) in the table. II. Therefore, Krum is more vulnerable to DNY-OPT attacks than Min-max attacks.

*4) **Analysis for the upper bound of DNY-OPT attacks against FedDet**:* Now we analyse the real distance $\|G^{'t} - G^{t}\|$ and estimate the certified radius. The theorem 1 proposed by [30] analyzes the certified radius. We combine the theorem 1 and the theorem 3. Then we get the certified radius of FedDet against DNY-OPT attacks as follows:

$$\begin{aligned} \|G^{'T} - G^{T}\| \le \Lambda(\boldsymbol{T})(1+dc)^{\Lambda(\boldsymbol{T})}(\|\mathbf{w}_{1,i} - G_{i-1}\| \\ + \frac{c}{n-cm} \cdot \sum_{i=1}^{n} \|(\frac{l^{'}}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i} - \mathbf{w}_{i,i})\|). \end{aligned} \quad (30)$$

Defense methods	Adaptive attacks	Upper bound on perturbation ρ
FedDet	STAT-OPT attacks	$\rho \leq \frac{n}{n-cm} \lvert \mathbf{w}_{1,i} - G_{i-1} \rvert + \frac{cm}{n-cm} \lvert G_{i-1} \rvert + \frac{c}{n-cm} \left\lvert \sum_{i=m+1}^{n} \frac{\lVert \mathbf{w}_i \rVert_{max}}{\lVert \mathbf{w}_{i,i} \rVert} \mathbf{w}_{i,i} - \sum_{i=1}^{n} \frac{l}{\lVert \mathbf{w}_{i,i} \rVert} \mathbf{w}_{i,i} \right\rvert$ 2
	DNY-OPT attacks	$\rho \leq \lVert \mathbf{w}_{1,i} - G_{i-1} \rVert + \frac{c}{n-cm} \cdot \sum_{i=1}^{n} \lVert (\frac{l'}{\lVert \mathbf{w}_{i,i} \rVert} \mathbf{w}_{i,i} - \mathbf{w}_{i,i}) \rVert$ 3
Agnostic methods	Min-max attacks	$\rho \leq \lVert \mathbf{w}_{1,i} - G_{i-1} \rVert + \underset{i,j \in [m+1,n]}{max} \lVert \mathbf{w}_i - \mathbf{w}_j \rVert$ 4
	Min-sum attacks	$\rho \leq \sqrt{\frac{1}{n-m} (\sum_{i=m+1}^{n} \lVert \mathbf{w}_{1,i} - G_{i-1} \rVert^2 + max \sum_{i,j \in [m+1,n]} \lVert \mathbf{w}_i - \mathbf{w}_j \rVert^2)}$ 5
Krum	DNY-OPT attacks	$\rho \leq max_{i \in [m+1,n]} \lVert \mathbf{w}_{i,i} - G_{i-1} \rVert + \frac{1}{n-2m-1} min_{j \in [m+1,n]} \sum_{i \in \Gamma_{\mathbf{w}}^{n-m-2}} \lVert \mathbf{w}_{j,i} - \mathbf{w}_{i,i} \rVert$ 6

TABLE II: Comparison of perturbation ρ for adaptive attacks

We compare the real distance between $G^{'T}$ and G^T and the certified radius at T iteration. To estimate this certified radius, We set $\Lambda(\boldsymbol{T}) = 0.001$, which is the learning rate of the FL training. $(1 + dc)^{\Lambda(\boldsymbol{T})}$ is nearly 1 when $\Lambda(\boldsymbol{T})$ is a very small number. We get the corresponding values for calculating $\sum_{i=1}^{n} \lVert (\frac{l'}{\lVert \mathbf{w}_{i,i} \rVert} \mathbf{w}_{i,i} - \mathbf{w}_{i,i}) \rVert$ at iteration T. We record the benign G^T and poisoned $G^{'T}$ to calculate the real distance $\lVert G^{'T} - G^T \rVert$. We assume a 30% malicious rate at one iteration. This comparison is implemented in the FEMNIST dataset. In figure 1, we can see that for $T \in [0, 500]$ epoch, the real distance is always under the estimated certified radius, which means the certified radius 30 is a valid upper bound. Besides, the real distance and the certified radius show similar trends as the training epochs.

VII. Evaluation Setup

In this work, similar to other FL model poisoning or defences-related works, we focus on image classification tasks. It is noted that our design is general and can be applied to other FL-based tasks. We use two natural image recognition datasets, FEMNIST and CIFAR10. Natural image recognition may require security guarantees. For example, in a federated learning-based recommendation system, natural images on social websites can be poisoned with sensitive labels, which can cause wrong classification or unfairness.

FEMNIST [35] is a 62-class non-IID, class-imbalanced classification task with 3400 clients and 671585 grey-scale images. Each client has their own handwritten digits or letters (52 for upper and lower case letters and ten classes for digits). We select 24 out of 3400 clients for federated training; each client has 1000 samples for local training. We use a four-layer CNN as the local training model.

CIFAR10 [36] is a 10-class class-balanced classification task with 60,000 RGB images, each of size 32 × 32. This class-balanced dataset has the same number of samples per class. Each class of CIFAR10 has 6,000 images. We use 25 clients, each with 1,000 samples, use validation and test data of sizes 5,000. We use Alexnet [37] as the global model architecture.

We use a batch size of 250 and an SGD optimizer with learning rates of 0.001 for FEMNIST. We use a batch size of 250 and an SGD optimizer with learning rates 0.05 for CIFAR10. We repeat the evaluation five times for each attack scenario and use the average as the final result. We conducted five repeated experiments for each attack scenario and took the average value.

VIII. Evaluation Results

In this section, we test the efficiency of FedDet against all six designed adaptive untargeted attacks and compare it with other well-known baseline robust aggregation methods. We use PyTorch to implement all evaluations.

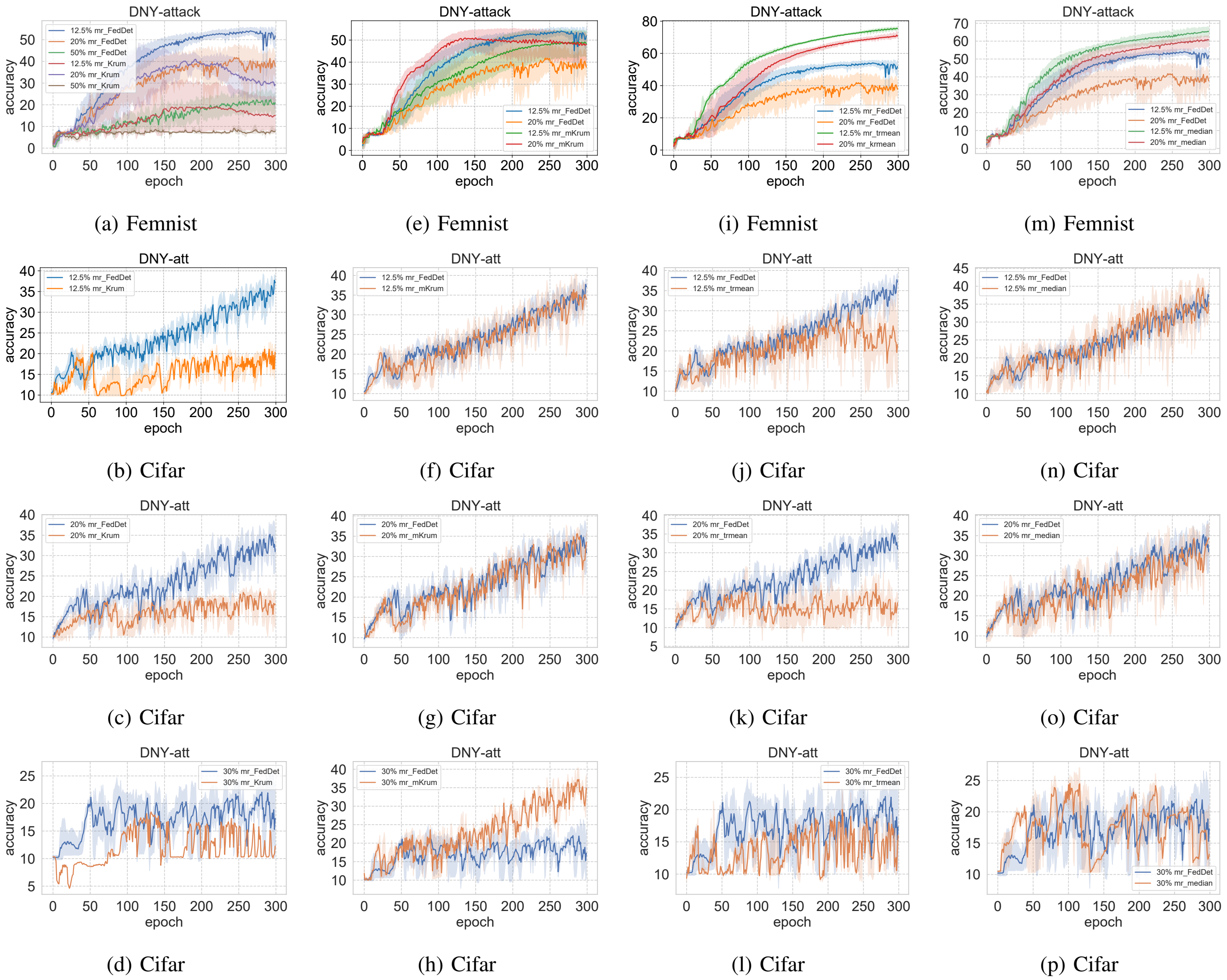
A. Robustness of FedDet

In figure 2, we evaluate the effectiveness of FedDet under different malicious rates. STAT-OPT attacks have minor impacts on the performance of FedDet. Under STAT-OPT attacks, the accuracy of FedDet decreases from 60.43% to 54.00%, 59.60% and 57.11% after 100 global epochs when 5%, 12.5% and 20% malicious clients respectively. As discussed in V-A, this optimization-based model poisoning attack starts from a reference initialized $\mathbf{w}^{'}_{1,i}, ..., \mathbf{w}^{'}_{m,i}$ and keeps updating. It is cumbersome to find the optimal initialization for this attack. Poor initialization might negatively affect the attack performance.

B. Comparison with previous methods

We compare FedDet with other robust aggregation schemes, Krum [26] , Multi-Krum [26], Trimmed-Mean [27] and Median [27].

In VI-B and VI-D, we proposed the upper bounds on perturbation ρ with which the DNY-OPT attacks can disturb the local updates. In VI-E, we compare these two upper bounds of FedDet and Krum and draw a conclusion that Krum is more vulnerable to DNY-OPT attacks than FedDet as ρ of Krum is larger. Figure 3(a)(b)(c)(d) validate our discussion. In figure 3(a), FedDet outperforms Krum under all situations. For example, when the malicious rate is 20%, the main accuracy of FedDet is 41%, but for Krum, the accuracy decreases to 30%. Krum fails the training when half of the clients are malicious. Figure 3(b)(c)(d) shows similar results. FedDet is still robust when the malicious clients' rates are 12.5% and 20%, but Krum has poor accuracy. FedDet also performs better than Krum with 30% malicious rates. The main accuracy of



(a) Femnist (e) Femnist (i) Femnist (m) Femnist

(b) Cifar (f) Cifar (j) Cifar (n) Cifar

(c) Cifar (g) Cifar (k) Cifar (o) Cifar

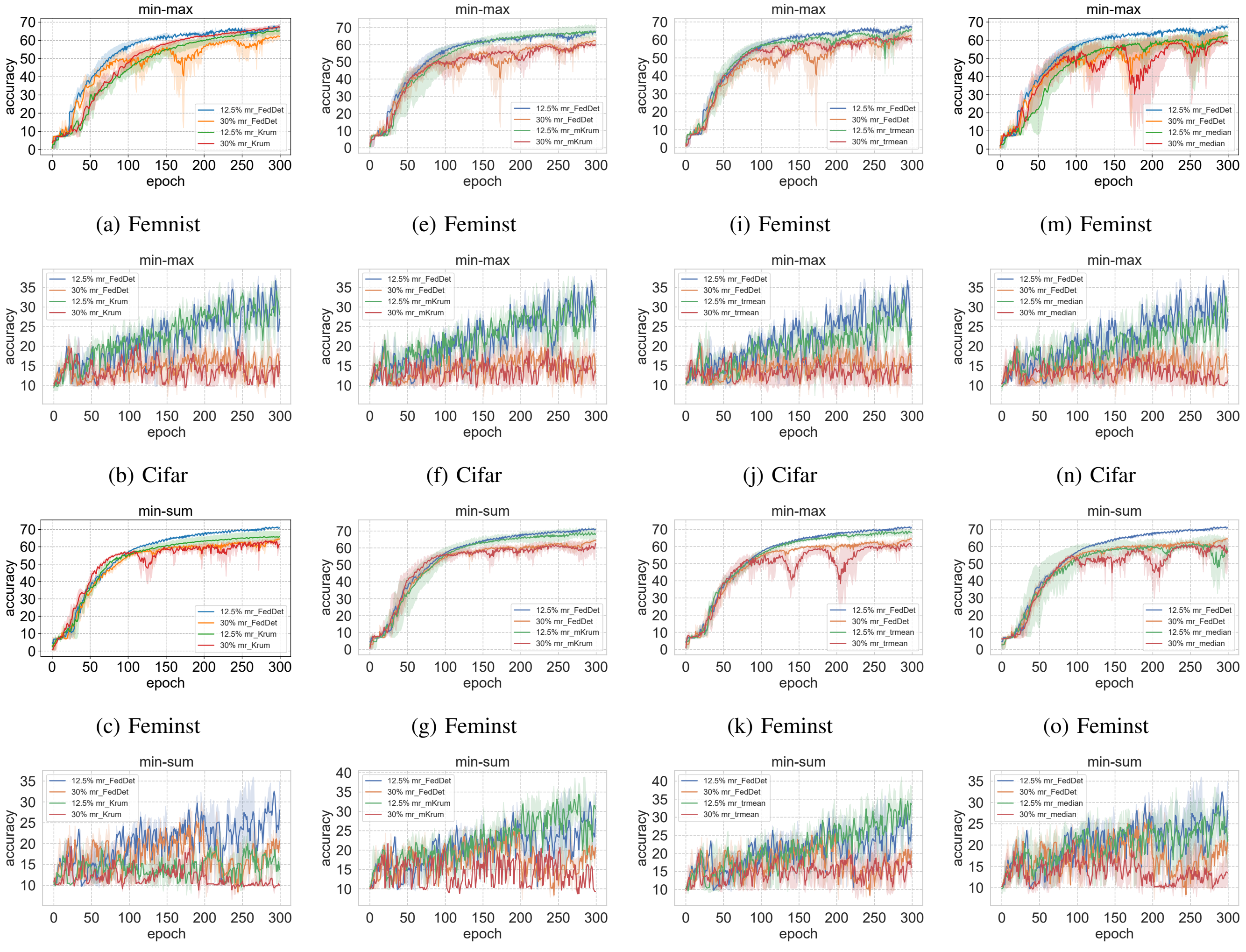
(d) Cifar (h) Cifar (l) Cifar (p) Cifar

Fig. 3: Comparison of the robustness of FedDet and other well-known robust aggregation methods against DNY-OPT attacks in two datasets.

Krum keeps below 20% with all malicious rates. Besides, according to 26 in VI-E, it is not difficult to see that the upper bound of ρ is larger when m is larger. Namely, larger amounts of malicious clients may cause a worse impact on FedDet. The results of figure 3(a)(b)(c)(d) are also in line with this analysis. For example, in figure 3(a), when the malicious rate increases from 12.5% to 20%, the main accuracy of FedDet decreases from 52% to 41%. And in figure 3(b)(d), the main accuracy decreases from 37% to 15% when the malicious rate increases from 12.5% to 30%. In figure 3(e)-(h), we can see that FedDet is competitive with Multi-Krum. In figure 3(e), FedDet achieves better performance when the malicious rate is 12.5%, but Multi-krum has higher accuracy with 20% malicious rate. According to figure 3(j)(k)(l), FedDet is also more robust than Trimmed-Mean, which uses perfect knowledge of malicious client rates. From figure 3(n)(o)(p), FedDet is a little more robust than another agnostic method, Median, which is agnostic to the actual malicious clients' rates.

Remark: According to 26 in VI-E, it is not difficult to see that the upper bound of ρ is larger when m is larger. Namely, larger amounts of malicious clients may cause a worse impact on FedDet. Similar to other compared baseline works (Krum, Multi-krum, Median and Trimmed-Mean), the performance of FedDet is gradually degraded as the portion of adversaries increases. However, Krum only satisfies the resilience property under the assumption $2f+2<n$, where f denotes the number of adversaries and n denotes the number of all clients. When the portion of adversaries is above 50%, Krum fails to work. On the other hand, FedDet can achieve a reasonable performance (see Figure. 3(a) in the manuscript). Besides, FedDet has a better performance compared to Median and Trimmed-Mean with 12.5%, 20% and 30% portion of adversaries.

Now, we compare the robustness of FedDet with other aggregation methods against Min-max and Min-sum attack situations. From figure 4, FedDet outperforms other robust methods in most situations. According to the test results in the femnist data set in figure 4, FedDet achieves the best main accuracy compared to the other four robust aggregation methods when the malicious rate of the agnostic attacks is



(a) Femnist

(e) Femnist

(i) Femnist

(m) Femnist

(b) Cifar

(f) Cifar

(j) Cifar

(n) Cifar

(c) Feminst

(g) Feminst

(k) Feminst

(o) Feminst

(d) Cifar

(h) Cifar

(l) Cifar

(p) Cifar

Fig. 4: Comparison of the robustness of FedDet and other well-known robust aggregation methods against min-max and min-sum attacks in two datasets.

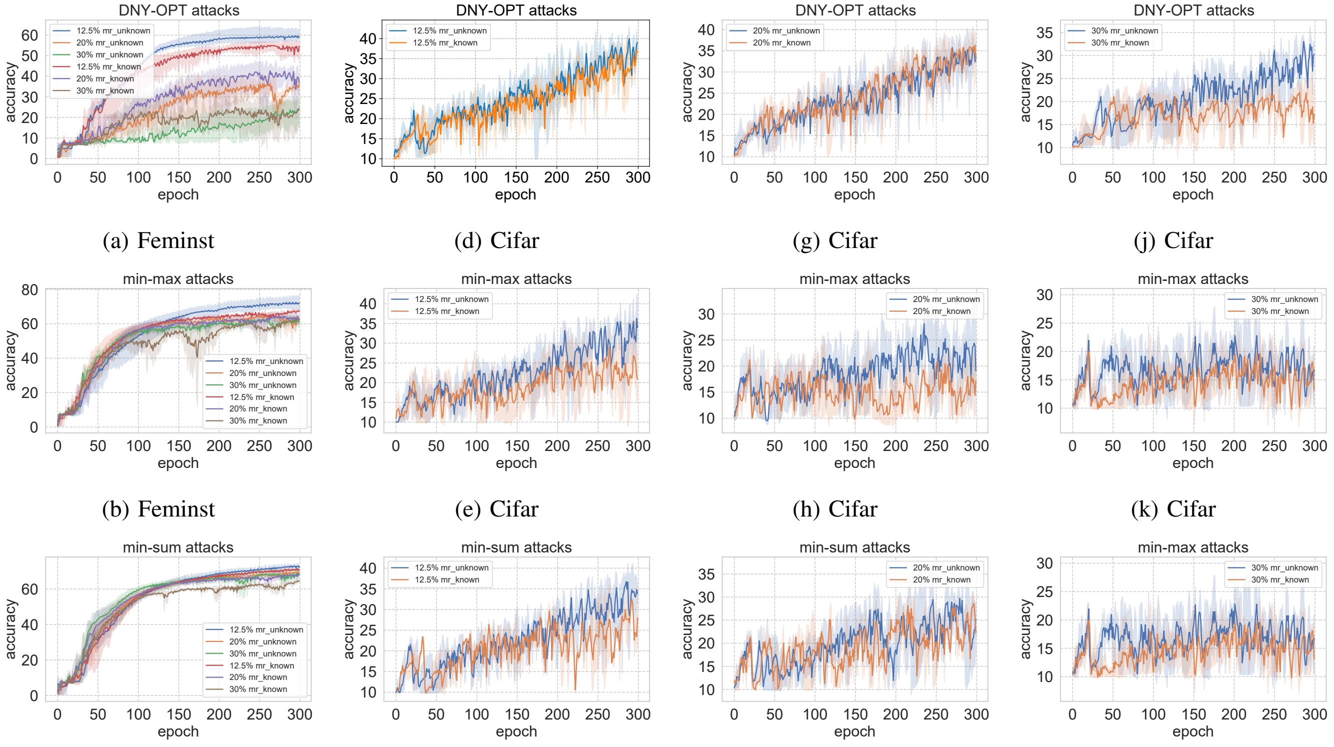
12.5%. FedDet performs similarly to other methods with 30% malicious rate in the femnist dataset. Further, in the aforementioned section VI-E3, we discuss Krum is more vulnerable to DNY-OPT attacks than Min-max or Min-sum attacks. From figure 3(a)(b) and figure 4(a)(b)(c)(d), the results validate our points. For example, according to the results of figure 3(a) and figure 4(a)(c), with the same malicious rate 12.5%, Krum can still achieve a good main accuracy under the agnostic attacks, but Krum fails to defend DNY-OPT attacks. Figure 3(b)(c)(d) and figure 4(b)(d) show similar results. Krum fails to defend against DNY-OPT attacks in all the situations, but it keeps a 27% accuracy under the Min-max attacks when the malicious rate is 12.5%.

C. Situations when the adversary has no knowledge of benign updates

In the sections above, we assume the malicious clients have full knowledge of the benign clients, which is a strong adversary model. However, the malicious clients rarely access benign updates from clean clients or the central server. Therefore, the malicious clients should store the historical benign updates locally as alternatives when they attempt to attack. Now, we discuss this weaker adversary model in which the benign updates are agnostic to the adversary. In V, we mentioned that in the six attack strategies, the strong attacker knows $\mathbf{w}_{i\in[1,n]}$. But in the limited attack strategies, the attacker only knows $\mathbf{w}_{i\in[m+1,n]}$. Hence, manipulating the Byzantine is based on $\mathbf{w}_{i\in[m+1,n]}$. In figure 5, we evaluate the performance of FedDet under the weaker adversary model. Roughly speaking, FedDet is more vulnerable to strong adversary attacks. For example, in figure 5(a), the main accuracy of FedDet can achieve 60% with 12.5% malicious rates in the weaker adversary model compared to 52% in the strong adversary model. In figure 5(j), FedDet can have a 30% accuracy with 30% malicious rates in the weaker model compared to 15% accuracy in the strong model.

D. Remarks on Existing Robust Aggregation Algorithms

In this section, we discuss the pros and cons of FedDet compared to other baseline works in terms of different char-



(a) Feminst (d) Cifar (g) Cifar (j) Cifar

(b) Feminst (e) Cifar (h) Cifar (k) Cifar

(c) Feminst (f) Cifar (i) Cifar (l) Cifar

Fig. 5: Performance of FedDet against DNY-OPT attacks, min-max attacks and min-sum attacks when the adversary has partial knowledge of $\mathbf{w}_{i\in[1,n]}$

	Agnostic to the actual corruption level	Expected Time Complexity
Krum [26]	No	$O(n^2 \cdot d)$
Multi-krum [26]	No	$O(mn^2 \cdot d)$
Trimmed-Mean [27]	No	$o(n \cdot d)$
Median [27]	Yes	$o(n \cdot d)$
FedDet	Yes	$O(n \cdot \frac{n}{l})$

TABLE III: Comparison of Existing Robust Aggregation Algorithms

acteristics. Unlike Krum, Multi-krum, and Trimmed-Mean, FedDet and Median do not require the exact knowledge of the corruption level of FL systems, which are more realistic in the real world as the defender has no access to know the attacker's actions. The expected time complexity of Krum is $O(n^2 \cdot d)$, where n is the number of selected clients in one training iteration and d is the dimension of the parameter vectors. The parameter server computes the squared distance between a client's vector with the resting parameters' vectors ($O(n \cdot d)$. Then, the parameter server repeated this process for all selected clients ($O(n)$). Thus, the square distance computing time is $O(n^2 \cdot d)$. After computing, the server selects the first $n - f - 1$ of the distances for the clients ($O(n)$ with Quickselect) and repeats the process for all clients ($O(n^2)$). Therefore, the expected time complexity for Krum is $O(n^2 \cdot d)$. For Multi-krum, a variant of Krum, selects the $m \in \{1, ..., n\}$ vectors with the smallest sum of distances. m varies between 1 and n. Thus, the expected time complexity of Multi-krum is $O(mn^2 \cdot d)$. Trimmed-Mean sorts (Quicksort) the values of all clients' vectors in dimension ($O(n \cdot d)$). Similarly, Median selects the median value of all client vectors in dimension with Quickselect ($O(n \cdot d)$). For FedDet, it computes the $L2$-norm values of split client vectors ($O(n \cdot \frac{n}{l})$). Then, FedDet sorts the norm values (Quicksort) and selects the median value ($O(n)$). Thus the expected time complexity of FedDet is $O(n \cdot \frac{n}{l})$. From the above analysis, we can see that compared to Krum and Multi-krum, Trimmed-Mean, Median and FedDet have less expected time complexity. The summary of the remarks is shown in the table. III.

IX. Concluding Remarks

Federated learning holds great potential in providing privacy for large amounts of distributed end devices. However, it is vulnerable to adaptive poisoning attacks. Existing defence methods did not consider the causes of model parameters' high dimensionality and data heterogeneity. In this work, we proposed a novel Byzantine-robust federated learning, FedDet to solve the problem. FedDet can overcome this issue with high dimensionality and keep the functionality of layers. During the robust aggregation, FedDet normalizes every slice of local models by the median norm value rather than excluding some clients, which can avoid deviation from the optimal aggregated model. In addition, we presented a theoretical security analysis model and conducted an extensive security analysis of FedDet and the state-of-the-art robust aggregation method, Krum. We discussed why the proposed method outperforms the prior method. It is noted that in this work, we do not discuss how FedDet defends targeted model poisoning attacks that can insert backdoors while keeping the trained model's accuracy. Future research will focus on combining the advantages of the proposed method and other defence approaches to defeat untargeted model poisoning and targeted model poisoning attacks.

Acknowledgement

This work was partly funded by EPSRC with RC Grant reference EP/Y027787/1, UKRI under grant number EP/Y028317/1, the European Horizon2020 MSCA programme under grant agreement No 824019 and101022280, Horizon Europe MSCA programme under grant agreement No 101086228.

References

[1] A. Imteaj, U. Thakker, S. Wang, J. Li, and M. H. Amini, "A survey on federated learning for resource-constrained iot devices," *IEEE Internet of Things Journal*, vol. 9, no. 1, pp. 1–24, 2021.

[2] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. V. Poor, "Federated learning for internet of things: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1622–1658, 2021.

[3] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, "A survey on internet of things: Architecture, enabling technologies, security and privacy, and applications," *IEEE internet of things journal*, vol. 4, no. 5, pp. 1125–1142, 2017.

[4] M. Mohammadi and A. Al-Fuqaha, "Enabling cognitive smart cities using big data and machine learning: Approaches and challenges," *IEEE Communications Magazine*, vol. 56, no. 2, pp. 94–101, 2018.

[5] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, "Application of machine learning in wireless networks: Key techniques and open issues," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3072–3108, 2019.

[6] 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*. PMLR, 2017, pp. 1273–1282.

[7] G. Sun, Y. Cong, J. Dong, Q. Wang, L. Lyu, and J. Liu, "Data poisoning attacks on federated machine learning," *IEEE Internet of Things Journal*, vol. 9, no. 13, pp. 11 365–11 375, 2021.

[8] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *Proceedings of the 29th International Conference on Machine Learning*, 2012, pp. 1467–1474.

[9] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2017.

[10] M. Fang, G. Yang, N. Z. Gong, and J. Liu, "Poisoning attacks to graph-based recommender systems," in *Proceedings of the 34th annual computer security applications conference*, 2018, pp. 381–392.

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

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

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

[14] 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," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 070–16 084, 2020.

[15] Z. Allen-Zhu, F. Ebrahimian, J. Li, and D. Alistarh, "Byzantine-resilient non-convex stochastic gradient descent," *arXiv preprint arXiv:2012.14368*, 2020.

[16] 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)*, 2020, pp. 1605–1622.

[17] V. Shejwalkar and A. Houmansadr, "Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning," in *NDSS*, 2021.

[18] G. Baruch, M. Baruch, and Y. Goldberg, "A little is enough: Circumventing defenses for distributed learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[19] R. Guerraoui, S. Rouault, *et al.*, "The hidden vulnerability of distributed learning in byzantium," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3521–3530.

[20] "Federated learning: Collaborative machine learning without centralized training data. [online]." https://ai.googleblog.com/2017/04/federated-learning-collaborative.html, 2017.

[21] "Utilization of fate in risk management of credit in small and micro enterprises. [online]." https://www.fedai.org/cases/utilization-of-fate-in-risk-management-of-credit-in-small-and-micro-enterprises/.

[22] "Machine learning ledger orchestration for drug discovery (melloddy). [online]." https://www.melloddy.eu/, 2017.

[23] S. Islam, S. Badsha, I. Khalil, M. Atiquzzaman, and C. Konstantinou, "A triggerless backdoor attack and defense mechanism for intelligent task offloading in multi-uav systems," *IEEE Internet of Things Journal*, vol. 10, no. 7, pp. 5719–5732, 2022.

[24] C. Xie, M. Chen, P.-Y. Chen, and B. Li, "Crfl: Certifiably robust federated learning against backdoor attacks," in *International Conference on Machine Learning*. PMLR, 2021, pp. 11 372–11 382.

[25] 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.

[26] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

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

[28] L. Li, W. Xu, T. Chen, G. B. Giannakis, and Q. Ling, "Rsa: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 1544–1551.

[29] Y. Liu, C. Chen, L. Lyu, F. Wu, S. Wu, and G. Chen, "Byzantine-robust learning on heterogeneous data via gradient splitting," 2023.

[30] A. Panda, S. Mahloujifar, A. N. Bhagoji, S. Chakraborty, and P. Mittal, "Sparsefed: Mitigating model poisoning attacks in federated learning with sparsification," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 7587–7624.

[31] T. D. Nguyen, P. Rieger, R. De Viti, H. Chen, B. B. Brandenburg, H. Yalame, H. Möllering, H. Fereidooni, S. Marchal, M. Miettinen, *et al.*, "{FLAME}: Taming backdoors in federated learning," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 1415–1432.

[32] X. Cao, M. Fang, J. Liu, and N. Z. Gong, "Fltrust: Byzantine-robust federated learning via trust bootstrapping," *arXiv preprint arXiv:2012.13995*, 2020.

[33] Z. Zhang, A. Panda, L. Song, Y. Yang, M. Mahoney, P. Mittal, R. Kannan, and J. Gonzalez, "Neurotoxin: Durable backdoors in federated learning," in *International Conference on Machine Learning*. PMLR, 2022, pp. 26 429–26 446.

[34] M. Cheng, T. Le, P.-Y. Chen, J. Yi, H. Zhang, and C.-J. Hsieh, "Query-efficient hard-label black-box attack: An optimization-based approach," *arXiv preprint arXiv:1807.04457*, 2018.

[35] S. Caldas, S. M. K. Duddu, P. Wu, T. Li, J. Konečnỳ, H. B. McMahan, V. Smith, and A. Talwalkar, "Leaf: A benchmark for federated settings," *arXiv preprint arXiv:1812.01097*, 2018.

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

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

X. Appendix

A. Proofs of Theorem 2

Proof. We have

$$\begin{aligned}
&G_i^{'} - G_i \\
&= \frac{1}{n}\{(\frac{med(L_i^{'})}{\|\mathbf{w}_{1,i}^{'}\|}\mathbf{w}_{1,i}^{'} + \frac{med(L_i^{'})}{\|\mathbf{w}_{2,i}^{'}\|}\mathbf{w}_{2,i}^{'}+,...,+\frac{med(L_i^{'})}{\|\mathbf{w}_{n,i}\|}\mathbf{w}_{n,i}) \\
&- (\frac{med(L_i)}{\|\mathbf{w}_{1,i}\|}\mathbf{w}_{1,i} + \frac{med(L_i)}{\|\mathbf{w}_{2,i}\|}\mathbf{w}_{2,i}+,...,+\frac{med(L_i)}{\|\mathbf{w}_{n,i}\|}\mathbf{w}_{n,i})\} \\
&= \frac{1}{n}(\mathbf{w}_{1,i}^{'}+,...,+\mathbf{w}_{m,i}^{'} + \sum_{i=m+1}^{n} \frac{l^{'}}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i} - \sum_{i=1}^{n} \frac{l}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i}) \\
&= \frac{1}{n}(\mathbf{w}_{1,i}^{'} - G_{i-1})+,...,+\frac{1}{n}(\mathbf{w}_{m,i}^{'} - G_{i-1}) + \frac{m}{n}\cdot G_{i-1} \\
&+ \frac{1}{n}(\sum_{i=m+1}^{n} \frac{l^{'}}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i} - \sum_{i=1}^{n} \frac{l}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i}),
\end{aligned} \tag{31}$$

The optimization problem of maximizing $S_i^T(G_i^{'} - G_i)$ can be equivalently converted to maximize $\left|G_i^{'} - G_i\right|$. Out of convenience, we assume that all malicious clients can collude, so we have $\mathbf{w}_{1,i}^{'} = \mathbf{w}_{2,i}^{'} = ... = \mathbf{w}_{m,i}^{'}$. Then we have

$$\begin{aligned}
\left|G_i^{'} - G_i\right| &\leq \frac{1}{n}\left|m\cdot(\mathbf{w}_{1,i}^{'} - G_{i-1})\right| + \frac{m}{n}\cdot|G_{i-1}| \\
&+ \frac{1}{n}\left|\sum_{i=m+1}^{n} \frac{l^{'}}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i} - \sum_{i=1}^{n} \frac{l}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i}\right|.
\end{aligned} \tag{32}$$

According to Assumption 1, we have

$$\frac{1}{c}\left|\mathbf{w}_{1,i}^{'} - \mathbf{w}_{1,i}\right| \leq \left|G_i^{'} - G_i\right|. \tag{33}$$

Combing above inequality equations (32) and (33), we get

$$\begin{aligned}
&\frac{1}{c}\left|\mathbf{w}_{1,i}^{'} - \mathbf{w}_{1,i}\right| \leq \left|G_i^{'} - G_i\right| \leq \frac{1}{n}\left|m\cdot(\mathbf{w}_{1,i}^{'} - G_{i-1})\right| \\
&+ \frac{m}{n}\cdot|G_{i-1}| + \frac{1}{n}\left|\sum_{i=m+1}^{n} \frac{l^{'}}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i} - \sum_{i=1}^{n} \frac{l}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i}\right|.
\end{aligned} \tag{34}$$

According to the triangle inequality $\left|\mathbf{w}_{1,i}^{'} - \mathbf{w}_{1,i}\right| \geq \left|\mathbf{w}_{1,i}^{'} - G_{i-1}\right| - |\mathbf{w}_{1,i} - G_{i-1}|$, we get

$$\begin{aligned}
&(\frac{1}{c} - \frac{m}{n})\left|\mathbf{w}_{1,i}^{'} - G_{i-1}\right| \leq \frac{1}{c}|\mathbf{w}_{1,i} - G_{i-1}| + \frac{m}{n}\cdot|G_{i-1}| \\
&+ \frac{1}{n}\left|\sum_{i=m+1}^{n} \frac{l^{'}}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i} - \sum_{i=1}^{n} \frac{l}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i}\right|.
\end{aligned} \tag{35}$$

Based on our Definition 1, we have $\left|\mathbf{w}_{1,i}^{'} - G_{i-1}\right| = \rho$, so we get

$$\begin{aligned}
\rho &\leq \frac{n}{n-cm}|\mathbf{w}_{1,i} - G_{i-1}| + \frac{cm}{n-cm}|G_{i-1}| \\
&+ \frac{c}{n-cm}\left|\sum_{i=m+1}^{n} \frac{l^{'}}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i} - \sum_{i=1}^{n} \frac{l}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i}\right|.
\end{aligned} \tag{36}$$

As we know $l^{'} \leq \|\mathbf{w}_i\|_{max}$, so the final upper bound on ρ is

$$\begin{aligned}
\rho &\leq \frac{n}{n-cm}|\mathbf{w}_{1,i} - G_{i-1}| + \frac{cm}{n-cm}|G_{i-1}| \\
&+ \frac{c}{n-cm}\left|\sum_{i=m+1}^{n} \frac{\|\mathbf{w}_i\|_{max}}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i} - \sum_{i=1}^{n} \frac{l}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i}\right|.
\end{aligned} \tag{37}$$

□

B. Proofs of Theorem 3

Proof. We have

$$\begin{aligned}
&\|G_i^{'} - G_i\| \\
&= \frac{1}{n}\|(\frac{med(L_i^{'})}{\|\mathbf{w}_{1,i}^{'}\|}\mathbf{w}_{1,i}^{'} + \frac{med(L_i^{'})}{\|\mathbf{w}_{2,i}^{'}\|}\mathbf{w}_{2,i}^{'}+,...,+\frac{med(L_i^{'})}{\|\mathbf{w}_{n,i}\|}\mathbf{w}_{n,i}) \\
&- \sum_{i=1}^{n}\mathbf{w}_{i,i}\| \\
&= \frac{1}{n}\|(\mathbf{w}_{1,i}^{'}+,...,+\mathbf{w}_{m,i}^{'} + \sum_{i=m+1}^{n} \frac{l^{'}}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i}) - \sum_{i=1}^{n}\mathbf{w}_{i,i}\| \\
&\leq \frac{1}{n}\|\sum_{i=1}^{m}(\mathbf{w}_{i,i}^{'} - \mathbf{w}_{i,i}) + \sum_{i=1}^{n}(\frac{l^{'}}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i} - \mathbf{w}_{i,i})\| \\
&\leq \frac{1}{n}(m\cdot\|\mathbf{w}_{1,i}^{'} - \mathbf{w}_{1,i}\|) + \frac{1}{n}\cdot\sum_{i=1}^{n}\|(\frac{l^{'}}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i} - \mathbf{w}_{i,i})\|.
\end{aligned} \tag{38}$$

According to Assumption 1, we have

$$\begin{aligned}
\frac{1}{c}\|\mathbf{w}_{1,i}^{'} - \mathbf{w}_{1,i}\| &\leq \frac{1}{n}(m\cdot\|\mathbf{w}_{1,i}^{'} - \mathbf{w}_{1,i}\|) \\
&+ \frac{1}{n}\cdot\sum_{i=1}^{n}\|(\frac{l^{'}}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i} - \mathbf{w}_{i,i})\|.
\end{aligned} \tag{39}$$

According to the triangle inequality $\|\mathbf{w}_{1,i}^{'} - \mathbf{w}_{1,i}\| \geq \|\mathbf{w}_{1,i}^{'} - G_{i-1}\| - \|\mathbf{w}_{1,i} - G_{i-1}\|$, we get

$$\begin{aligned}
&\frac{1}{c}(\|\mathbf{w}_{1,i}^{'} - G_{i-1}\| - \|\mathbf{w}_{1,i} - G_{i-1}\|) \\
&\leq \frac{1}{n}(m\cdot\|\mathbf{w}_{1,i}^{'} - \mathbf{w}_{1,i}\|) + \frac{1}{n}\cdot\sum_{i=1}^{n}\|(\frac{l^{'}}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i} - \mathbf{w}_{i,i})\|.
\end{aligned} \tag{40}$$

Given the Definition 1, we get $\|\mathbf{w}_{1,i}^{'} - G_{i-1}\| = \|\gamma\nabla^p\| = \rho$ for DNY-OPT attacks, so we get the final upper bound on ρ

$$\rho \leq \|\mathbf{w}_{1,i} - G_{i-1}\| + \frac{c}{n-cm}\cdot\sum_{i=1}^{n}\|(\frac{l^{'}}{\|\mathbf{w}_{i,i}\|}\mathbf{w}_{i,i} - \mathbf{w}_{i,i})\|. \tag{41}$$

□

C. Proofs of Theorem 4 and 5

Proof of theorem 4.

Proof. We have the triangle inequality,

$$\|\mathbf{w}_{1,i}^{'} - \mathbf{w}_{1,i}\| \geq \|\mathbf{w}_{1,i}^{'} - G_{i-1}\| - \|\mathbf{w}_{1,i} - G_{i-1}\|, \quad (42)$$

then, we get

$$\begin{aligned} &\|\mathbf{w}_{1,i}^{'} - G_{i-1}\| - \|\mathbf{w}_{1,i} - G_{i-1}\| \leq \|\mathbf{w}_{1,i}^{'} - \mathbf{w}_{1,i}\| \\ &\leq \underset{i,j\in[m+1,n]}{max} \|\mathbf{w}_i - \mathbf{w}_j\|. \end{aligned} \quad (43)$$

Based on the Definition 1, we have $\|\mathbf{w}_{1,i}^{'} - G_{i-1}\| = \|\gamma\nabla^p\| = \rho$.

So the upper bound on ρ is

$$\rho \leq \|\mathbf{w}_{1,i} - G_{i-1}\| + \underset{i,j\in[m+1,n]}{max} \|\mathbf{w}_i - \mathbf{w}_j\|. \quad (44)$$

□

Proof of theorem 5

Proof. We have

$$\begin{aligned} &(n-m)\|\mathbf{w}_{1,i}^{'} - G_{i-1}\|^2 - \sum_{i\in[m+1,n]} \|\mathbf{w}_{1,i} - G_{i-1}\|^2 \\ &\leq \sum_{i\in[m+1,n]} \|\mathbf{w}_{1,i}^{'} - \mathbf{w}_{1,i}\|^2 \leq max \sum_{i,j\in[m+1,n]} \|\mathbf{w}_i - \mathbf{w}_j\|^2, \end{aligned} \quad (45)$$

then we have

$$\begin{aligned} (n-m)\rho^2 \leq &\sum_{i\in[m+1,n]} \|\mathbf{w}_{1,i} - G_{i-1}\|^2 \\ &+ max \sum_{i,j\in[m+1,n]} \|\mathbf{w}_i - \mathbf{w}_j\|^2. \end{aligned} \quad (46)$$

So, the proof of the upper bound on ρ is completed.

□

D. Proofs of Theorem 6

Proof. We assume the compromised local models are the same. Therefore, we have:

$$\sum_{i\in\Gamma_{\mathbf{w}^{'}}^{n-2m-1}} \|\mathbf{w}^{'} - \mathbf{w}_i\| \leq \underset{j\in[m+1,n]}{min} \sum_{i\in\Gamma_{\mathbf{w}}^{n-m-2}} \|\mathbf{w}_j - \mathbf{w}_i\|, \quad (47)$$

then, we get:

$$\begin{aligned} &(n-2m-1)\cdot\|\mathbf{w}_{1,i}^{'} - G_{i-1}\| \\ &\leq \underset{j\in[m+1,n]}{min} \sum_{i\in\Gamma_{\mathbf{w}}^{n-m-2}} \|\mathbf{w}_{j,i} - \mathbf{w}_{i,i}\| + \sum_{i\in\Gamma_{\mathbf{w}^{'}}^{n-2m-1}} \|\mathbf{w}_{i,i} - G_{i-1}\| \\ &\leq \underset{j\in[m+1,n]}{min} \sum_{i\in\Gamma_{\mathbf{w}}^{n-m-2}} \|\mathbf{w}_{j,i} - \mathbf{w}_{i,i}\| + (n-2m-1)\cdot \underset{i\in[m+1,n]}{max} \|\mathbf{w}_{i,i} - G_{i-1}\| \end{aligned} \quad (48)$$

Since we have $\|\mathbf{w}_{1,i}^{'} - G_{i-1}\| = \rho$. Then, the proof is completed. □