Certified Robustness for Free in Differentially Private Federated Learning

Chulin Xie
UIUC
chulinx2@illinois.edu

Yunhui Long
UIUC
ylong4@illinois.edu

Pin-Yu Chen
IBM Research
pin-yu.chen@ibm.com

Krishnaram Kenthapadi
Amazon AWS AI/ML
kenthk@amazon.com

Bo Li
UIUC
lbo@illinois.edu

Abstract

Federated learning (FL) provides an efficient training paradigm to jointly train a global model leveraging data from distributed users. As the local training data comes from different users who may not be trustworthy, several studies have shown that FL is vulnerable to poisoning attacks where adversaries add malicious data during training. On the other hand, to protect the privacy of users, FL is usually trained in a differentially private way (DPFL). Given these properties of FL, in this paper, we aim to ask: *Can we leverage the innate privacy property of DPFL to provide robustness certification against poisoning attacks? Can we further improve the privacy of FL to improve such certification?* To this end, we first investigate both the user-level and instance-level privacy of FL, and propose novel randomization mechanisms and analysis to achieve improved differential privacy. We then provide two robustness certification criteria: *certified prediction* and *certified attack cost* for DPFL on both levels. Theoretically, given different privacy properties of DPFL, we prove their certified robustness under a bounded number of adversarial users or instances. Empirically, we conduct extensive experiments to verify our theories under different attacks on a range of datasets. We show that the global model with a tighter privacy guarantee always provides stronger robustness certification in terms of the *certified attack cost*, while may exhibit tradeoffs regarding the *certified prediction*. We believe our work will inspire future research of developing certifiably robust DPFL based on its inherent properties.

1 Introduction

Federated Learning (FL), which aims to jointly train a global model with distributed local data, has been widely applied in different applications, such as finance [61], medical analysis [16], and user behavior prediction [32, 60, 59]. However, the fact that the local data and training process are entirely controlled by the *local users* who may be adversarial raises great concerns from both security and privacy perspectives. In particular, recent studies show that FL is vulnerable to different types of training-time attacks, such as model poisoning [10], backdoor attacks [6, 58, 55], and label-flipping attacks [27]. On the other hand, the privacy concerns have motivated the need to keep the raw data on local devices without sharing. However, sharing other indirect information such as gradients or model updates as part of the FL training process can also leak sensitive user information [63, 28, 11, 47]. As a result, in federated learning, a number of privacy-preserving learning approaches are proposed to protect the privacy including differential privacy (DP) [22, 20, 23], homomorphic encryption [15, 51, 30], and secure multiparty computation [9, 14]. Among these privacy guarantees, *differential privacy* is one of the most widely used concepts due to its strong information theoretic guarantees and relatively small systems overhead [41].

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

To defend against poisoning attacks in FL, different defenses are proposed. For instance, various robust aggregation methods [27, 50, 13, 24, 18, 62, 26, 40] identify and down-weight the malicious updates during aggregation or estimate a true "center" of the received updates rather than taking a weighted average. Other methods include robust federated training protocols (e.g., clipping [52], noisy perturbation [52] and additional evaluation during training [4]) and post-training strategies (e.g., fine-tuning and pruning [57]) that repair the poisoned global model. However, these work mainly focus on providing empirical robustness for FL while they have been shown vulnerable again to newly proposed strong adaptive attackers [55, 58, 7, 25]. Thus, in this paper, we aim to *develop certified robustness guarantees for FL against different poisoning attacks.* On the other hand, to ensure the privacy of FL, differential privacy is guaranteed for the trained global models in FL. Given the DP property of DPFL, we aim to ask: *Can we leverage the innate privacy property of DPFL to provide robustness certification against poisoning attacks for free? Can we further improve the privacy of FL so as to improve its certified robustness?*

Recent studies suggest that DP is inherently related with robustness of ML models. Intuitively, DP is designed to protect the privacy of individual data, such that the output of an algorithm remains essentially unchanged when one individual input point is modified. This means that the prediction of a DP model will be less impacted by (poisoned) training data. Thus, differential privacy has been used to provide both theoretical and empirical defenses against evasion attacks [39] and data poisoning attacks [43, 34] on *centralized* ML models. It has also been used as an empirical defense against backdoor attacks [31] in federated learning [6, 52], while no theoretical guarantee is provided. To the best of our knowledge, despite of the wide application of differentially private federated learning (DPFL), there is no work providing certified robustness for DPFL leveraging its privacy property.

In this paper, we aim to leverage the inherent privacy property of DPFL to provide the robustness certification for FL against poisoning attacks *for free*. In particular, we propose two robustness certification criteria for FL: *certified prediction* and *certified attack cost* under different attack constraints. We consider both the user-level DP [3, 29, 46, 5, 42] which is widely guaranteed in FL, and instance-level DP [44, 64] which is less explored in FL. Specifically, we prove that given a FL trained model satisfying user-level DP, the model is certifiably robust against k adversarial users based on our criteria. In addition, we propose `InsDP-FedAvg` to improve the instance-level DP in FL. Similarly, we prove that FL with instance-level DP guarantee is certifiably robust against k manipulated instances based on our criteria. Furthermore, we study the correlation between privacy guarantee and the certified robustness of FL. For instance, on one hand, stronger privacy guarantees higher attack cost; on the other hand, an overly strong privacy can hurt the certified prediction by introducing too much noise in the training process, and thus the optimal certified prediction is often achieved under a proper balance between privacy protection and utility loss.

Technical Contributions. This paper takes the first step to provide certified robustness in DPFL for free against poisoning attacks. We make contributions on both theoretical and empirical fronts.

- We propose two criteria for the certified robustness of FL against poisoning attacks.
- Given a FL model satisfying the user-level DP, we prove that it is certifiably robust against arbitrary poisoning attacks with k adversarial users based on our proposed robustness certification criteria.
- We propose `InsDP-FedAvg` algorithm to guarantee FL instance-level privacy. We prove that instance-level DPFL is certifiably robust against k instances manipulation during training.
- We conduct extensive experiments on MNIST and CIFAR datasets to verify our theoretical results.

2 Related work

Differentially Private Federated Learning. Different approaches are proposed to guarantee the user-level privacy for FL. Geyer *et al.* [29] and McMahan *et al.* [46] clip the norm of each local update, add Gaussian noise on the summed update, and characterize its privacy budget via moment accountant [2]. McMahan *et al.* [46] extends [29] to language models. In CpSGD [3], each user clips and quantizes the model update, and adds noise drawn from Binomial distribution, achieving both communication efficiency and DP. Bhowmick *et al.* [11] derive DP for FL via Renyi divergence [48] and study its protection against data reconstruction attacks. Liang *et al.* [42] utilizes Laplacian smoothing for each local update to enhance the model utility. Instead of using moment accountant to track privacy budget over FL rounds as previous work, Asoodeh *et al.* [5] derives the DP parameters by interpreting each round as a Markov kernel and quantify its impact on privacy parameters. All these works only focus on providing *user-level* privacy, leaving its robustness property unexplored.

In terms of instance-level privacy for FL, there are only a few work [44, 64]. Dopamine [44] provides instance-level privacy guarantee when each user only performs one step of DP-SGD [2] at each FL round. However, it cannot be applied to multi-step SGD for each user, thus it cannot be extended to the general FL setting FedAvg [45]. Zhu *et al.* [64] privately aggregate the labels from users in a voting scheme, and provide DP guarantees on both user level and instance level. However, it is also not applicable to standard FL, since it does not allow aggregating the gradients or model updates.

Differential Privacy and Robustness. In standard (centralized) learning, to guarantee robustness against *evasion* attacks, Pixel-DP [39] is proposed to certify that the model predictions do not depend too much on the individual pixels of test image. The prediction can be stable as long as the ℓ_2 norm of the perturbation is bounded. To certify the robustness against *poisoning* attacks, Ma *et al.* [43] show that private learners are resistant to data poisoning and analyze the lower bound of attack cost againt poisoning attacks for regression models. Here we certify the robustness in DPFL setting with such lower bound as one of our certification criteria and additionally derive its upper bounds. Hong *et al.* [34] show that the off-the-shelf mechanism DP-SGD [2], which clips per-sample gradients and add Guassian noises during training, can serve as a defense against poisoning attacks empirically. In *federated learning*, empirical work [6, 52] show that DPFL can mitigate backdoor attacks; however, none of these work provides certified robustness guarantees in FL.

3 Preliminaries

We start by providing some background on differential privacy (DP) and federated learning (FL).

Differential Privacy. (ϵ, δ)-DP is a current industry standard of privacy proposed by Dwork *et al.* [22, 20, 23]. It bounds the change in output distribution caused by a small input difference for a randomized algorithm. The following definition formally describes this privacy guarantee.

Definition 1 ((ϵ, δ)-DP [23])**.** *A randomized mechanism $\mathcal{M} : \mathcal{D} \to \Theta$ with domain $\mathcal{D}$ and range Θ satisfies (ϵ, δ)-DP if for any pair of two adjacent datasets $d, d' \in \mathcal{D}$, and for any possible (measurable) output set $E \subseteq \Theta$, it holds that $\Pr[\mathcal{M}(d) \in E] \leq e^{\epsilon} \Pr[\mathcal{M}(d') \in E] + \delta$.*

In Definition 1, when $\mathcal{M}$ is a training algorithm for ML model, domain $\mathcal{D}$ and range Θ represent all possible training datasets and all possible trained models respectively. Group DP for (ϵ, δ)-DP mechanisms follows immediately from Definition 1 where the privacy guarantee drops with the size of the group. Formally, it says:

Lemma 1 (Group DP)**.** *For mechanism $\mathcal{M}$ that satisfies (ϵ, δ)-DP, it satisfies $(k\epsilon, \frac{1-e^{k\epsilon}}{1-e^{\epsilon}}\delta)$-DP for groups of size k. That is, for any $d, d' \in \mathcal{D}$ that differ by k individuals, and any $E \subseteq \Theta$ it holds that $\Pr[\mathcal{M}(d) \in E] \leq e^{k\epsilon} \Pr[\mathcal{M}(d') \in E] + \frac{1-e^{k\epsilon}}{1-e^{\epsilon}}\delta$.*

Federated Learning. FedAvg was introduced by Mcmahan *et al.* [45] for FL to train a shared global model without direct access to training data of users. Specifically, given a FL system with N users, at round t, the server sends the current global model w_{t-1} to users in the selected user set U_t, where $|U_t| = m = qN$ and q is the user sampling probability. Each selected user $i \in U_t$ locally updates the model for E local epochs with its dataset D_i and learning rate η to obtain a new local model. Then, the user sends the local model updates Δw_t^i to the server. Finally, the server aggregates over the updates from all selected users into the new global model $w_t = w_{t-1} + \frac{1}{m}\sum_{i \in U_t} \Delta w_t^i$.

4 User-level Privacy and Certified Robustness for FL

4.1 User-level Privacy

Definition 1 leaves the definition of adjacent datasets flexible, which depends on applications. To protect user-level privacy, adjacent datasets are defined as those differing by data from one user [46].

Definition 2 (User-level (ϵ, δ)-DP)**.** *Let B, B' be two user sets with size N. Let D and D' be the datasets that are the union of local training examples from all users in B and B' respectively. Then, D and D' are adjacent if B and B' differ by one user. The mechanism $\mathcal{M}$ satisfies user-level (ϵ, δ)-DP if it meets Definition 1 with D and D' as adjacent datasets.*

Following standard DPFL [29, 46], we introduce `UserDP-FedAvg` (Algorithm 1 in Appendix B) to achieve user-level DP. At each round, the server first clips the update from each user with

a threshold S such that its ℓ_2-sensitivity is upper bounded by S. Next, the server sums up the updates, adds Gaussian noise sampled from $\mathcal{N}(0, \sigma^2 S^2)$, and takes the average, i.e., $w_t \leftarrow w_{t-1} + \frac{1}{m}\left(\sum_{i \in U_t} \text{Clip}(\Delta w_t^i, S) + \mathcal{N}\left(0, \sigma^2 S^2\right)\right)$. We utilize the moment accountant [2] to compute the DP guarantee. Compared to the standard composition theorem [23], the moment accounting method provides tighter bounds for the repeated application of the Gaussian mechanism combined with amplication-via-sampling. Given the user sampling probability q, noise level σ, FL rounds T, and a $\delta > 0$, `UserDP-FedAvg` satisfies (ϵ, δ)-DP as below, which is a generalization of [2]. The proof of proposition 1 follows the proof in [2] and is presented in the Appendix B.

Proposition 1 (`UserDP-FedAvg` Privacy Guarantee)**.** *There exist constants c_1 and c_2 so that given user sampling probability q, and FL rounds T, for any $\varepsilon < c_1 q^2 T$, if $\sigma \geq c_2 \frac{q\sqrt{T \log(1/\delta)}}{\epsilon}$, the randomized mechanism $\mathcal{M}$ in Algorithm 1 is (ϵ, δ)-DP for any $\delta > 0$.*

4.2 Certified Robustness of User-level DPFL against Poisoning Attacks

Threat Model. We consider the poisoning attacks against FL, where k adversarial users have poisoned instances in local datasets during training, aiming to fool the trained DPFL global model. Such attacks include *backdoor* attacks [31, 17] and *label flipping* attacks [12, 35]. The detailed description of these attacks is deferred to Appendix C. Note that our robustness certification is attack-agnostic under certain attack constraints (*e.g.*, k), and we will verify our certification bounds with different poisoning attacks in Section 5. Next, we propose two criteria for the robustness certification in FL: *certified prediction* and *certified attack cost*.

Certified Prediction. Consider the classification task with C classes. We define the classification *scoring function* $f : (\Theta, \mathbb{R}^d) \to \Upsilon^C$ which maps model parameters $\theta \in \Theta$ and an input data $x \in \mathbb{R}^d$ to a confidence vector $f(\theta, x)$, and $f_c(\theta, x) \in [0, 1]$ represents the confidence of class c. We mainly focus on the confidence after normalization, i.e., $f(\theta, x) \in \Upsilon^C = \{p \in \mathbb{R}^C_{\geq 0} : \|p\|_1 = 1\}$ in the probability simplex. Since DP mechanism produces a *stochastic* FL global model $\theta = \mathcal{M}(D)$ with randomized mechanism $\mathcal{M}$, we define the *expected scoring function* $F : (\theta, \mathbb{R}^d) \to \Upsilon^C$ where $F_c(\mathcal{M}(D), x) = \mathbb{E}[f_c(\mathcal{M}(D), x)]$ is the expected confidence for class c. The expectation is taken over DP training randomness, e.g., random Gaussian noise and random user subsampling. The corresponding *prediction* $H : (\theta, \mathbb{R}^d) \to [C]$ is defined by $H(\mathcal{M}(D), x) := \arg\max_{c \in [C]} F_c(\mathcal{M}(D), x)$, which is the top-1 class based on the expected prediction confidence. We will prove that such prediction allows the robustness certification against poisoning attacks.

Following our threat model in Section 4.2 and DPFL training in Algorithm 1, we denote the trained global model exposed to poisoning attacks by $\mathcal{M}(D')$. When $k = 1$, D and D' are user-level adjacent datasets according to Definition 2. Given that mechanism $\mathcal{M}$ satisfies user-level (ϵ, δ)-DP, based on the innate DP property, the distribution of the stochastic model $\mathcal{M}(D')$ is "close" to the distribution of $\mathcal{M}(D)$. Moreover, according to the *post-processing property* of DP [23], during testing, given a test sample x, we would expect that the values of the expected confidence for each class c, i.e., $F_c(\mathcal{M}(D'), x)$ and $F_c(\mathcal{M}(D), x)$, to be close, and thus the returned most likely class should be the *same*, i.e., $H(\mathcal{M}(D), x) = H(\mathcal{M}(D'), x)$, indicating *robust* prediction against poisoning attacks.

Theorem 1 (Condition for Certified Prediction under One Adversarial User)**.** *Suppose a randomized mechanism $\mathcal{M}$ satisfies user-level (ϵ, δ)-DP. For two user sets B and B' that differ by one user, D and D' are the corresponding training datasets. For a test input x, suppose $\mathbb{A}, \mathbb{B} \in [C]$ satisfy $\mathbb{A} = \arg\max_{c \in [C]} F_c(\mathcal{M}(D), x)$ and $\mathbb{B} = \arg\max_{c \in [C]: c \neq \mathbb{A}} F_c(\mathcal{M}(D), x)$, then if*

$$F_{\mathbb{A}}(\mathcal{M}(D), x) > e^{2\epsilon} F_{\mathbb{B}}(\mathcal{M}(D), x) + (1 + e^{\epsilon})\delta, \tag{1}$$

it is guaranteed that

$$H(\mathcal{M}(D'), x) = H(\mathcal{M}(D), x) = \mathbb{A}.$$

When $k > 1$, we resort to group DP. According to Lemma 1, given mechanism $\mathcal{M}$ satisfying user-level (ϵ, δ)-DP, it also satisfies user-level $(k\epsilon, \frac{1-e^{k\epsilon}}{1-e^{\epsilon}}\delta)$-DP for groups of size k. When k is smaller than certain threshold, leveraging the group DP property, we would expect that the distribution of the stochastic model $\mathcal{M}(D')$ is not too far away from the distribution of $\mathcal{M}(D)$ such that they would make the same prediction for a test sample with probabilistic guarantees.

Theorem 2 (Upper Bound of k for Certified Prediction)**.** *Suppose a randomized mechanism $\mathcal{M}$ satisfies user-level (ϵ, δ)-DP. For two user sets B and B' that differ k users, D and D' are the corresponding training datasets. For a test input x, suppose $\mathbb{A}, \mathbb{B} \in [C]$ satisfy $\mathbb{A} = \arg\max_{c \in [C]} F_c(\mathcal{M}(D), x)$*

and $\mathbb{B} = \arg\max_{c \in [C]: c \neq \mathbb{A}} F_c(\mathcal{M}(D), x)$*, then* $H(\mathcal{M}(D'), x) = H(\mathcal{M}(D), x) = \mathbb{A}$*,* $\forall k < \mathsf{K}$ *where* K *is the certified number of adversarial users:*

$$\mathsf{K} = \frac{1}{2\epsilon} \log \frac{F_{\mathbb{A}}(\mathcal{M}(D), x)(e^{\epsilon} - 1) + \delta}{F_{\mathbb{B}}(\mathcal{M}(D), x)(e^{\epsilon} - 1) + \delta} \tag{2}$$

The proofs of Theorems 1 and 2 are omitted to Appendix E. Theorems 1 and 2 reflect a tradeoff between privacy and certified prediction: i) in Theorem 1, if ϵ is large such that the RHS of Eq (1) > 1, the robustness condition cannot be met since the expected confidence $F_{\mathbb{A}}(\mathcal{M}(D), x) \in [0, 1]$. However, to achieve small ϵ, i.e., strong privacy, large noise is required during training, which would hurt model utility and thus result in small confidence margin between the top two classes (*e.g.*, $F_{\mathbb{A}}(\mathcal{M}(D), x)$ and $F_{\mathbb{B}}(\mathcal{M}(D), x)$), making it hard to meet the robustness condition. ii) In Theorem 2, if we fix $F_{\mathbb{A}}(\mathcal{M}(D), x), F_{\mathbb{B}}(\mathcal{M}(D), x)$, smaller ϵ of FL can certify larger K. However, smaller ϵ also induces smaller confidence margin, thus reducing K instead. As a result, properly choosing ϵ would help to certify a large K.

Certified Attack Cost. In addition to the certified prediction, we define the *attack cost* for attacker $C : \Theta \to \mathbb{R}$ which quantifies the difference between the poisoned model and the *attack goal*. In general, attacker aims to minimize the *expected* attack cost $J(D) := \mathbb{E}[C(\mathcal{M}(D))]$, where the expectation is taken over the randomness of DP training. The cost function can be instantiated according to the concrete attack goal in different types of poisoning attacks, and we provide some examples below. Given a global FL model satisfying user-level (ϵ, δ)-DP, we will prove the lower bound of the attack cost $J(D')$ when manipulating up to the data of k users. Higher lower bound of the attack cost indicates more *certifiably robust* global model.

Example 1. *(Backdoor attack [31, 17, 6])* $C(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(\theta, z_i^*)$*, where* $z_i^* = (x_i + \delta_x, y^*)$*,* δ_x *is the backdoor pattern,* y^* *is the target adversarial label. Minimizing* $J(D')$ *drives the prediction on any test data with the backdoor pattern* δ_x *to the target label* y^**.*

Example 2. *(Label Flipping attack [12, 35])* $C(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(\theta, z_i^*)$*, where* $z_i^* = (x_i, y^*)$ *and* y^* *is the target adversarial label. Minimizing* $J(D')$ *drives the prediction on test data* x_i *to the target label* y^**.*

Example 3. *(Parameter-Targeting attack [43])* $C(\theta) = \frac{1}{2}\|\theta - \theta^\star\|^2$*, where* $\theta^\star$ *is the target adversarial model. Minimizing* $J(D')$ *drives the poisoned model to be close to the target model.*

Theorem 3 (Attack Cost with k Attackers)**.** *Suppose a randomized mechanism* $\mathcal{M}$ *satisfies user-level* (ϵ, δ)*-DP. For two user sets* B *and* B' *that differ* k *users,* D *and* D' *are the corresponding training datasets. Let* $J(D)$ *be the expected attack cost where* $|C(\cdot)| \leq \bar{C}$*. Then,*

$$\begin{aligned}
&\min\{e^{k\epsilon} J(D) + \frac{e^{k\epsilon} - 1}{e^{\epsilon} - 1}\delta\bar{C}, \bar{C}\} \geq J(D') \geq \max\{e^{-k\epsilon} J(D) - \frac{1 - e^{-k\epsilon}}{e^{\epsilon} - 1}\delta\bar{C}, 0\}, \quad \textit{if} \quad C(\cdot) \geq 0 \\
&\min\{e^{-k\epsilon} J(D) + \frac{1 - e^{-k\epsilon}}{e^{\epsilon} - 1}\delta\bar{C}, 0\} \geq J(D') \geq \max\{e^{k\epsilon} J(D) - \frac{e^{k\epsilon} - 1}{e^{\epsilon} - 1}\delta\bar{C}, -\bar{C}\}, \quad \textit{if} \quad C(\cdot) \leq 0
\end{aligned} \tag{3}$$

The proof is omitted to Appendix E. Theorem 3 provides the upper bounds and lower bounds for attack cost $J(D')$. The lower bounds show that to what extent the attack can reduce $J(D')$ by manipulating up to k users, i.e., how successful the attack can be. The lower bounds depend on the attack cost on clean model $J(D)$, k and ϵ. When $J(D)$ is higher, the DPFL model under poisoning attacks is more robust because the lower bounds are accordingly higher; a tighter privacy guarantee, i.e., smaller ϵ, can also lead to higher robustness certification as it increases the lower bounds; with larger k, the attacker ability grows and thus lead to lower possible $J(D')$. The upper bounds show the least adversarial effect brought by k attackers, i.e., how vulnerable the DPFL model is under the optimistic case (*e.g.*, the backdoor pattern is less distinguishable).

Leveraging the lower bounds in Theorem 3, we can lower-bound the minimum number of attackers required to reduce the attack cost to certain level associated with hyperparameter τ in Corollary 1.

Corollary 1 (Lower Bound of k Given τ)**.** *Suppose a randomized mechanism* $\mathcal{M}$ *satisfies user-level* (ϵ, δ)*-DP. Let attack cost function be* C*, the expected attack cost be* $J(\cdot)$*. In order to achieve* $J(D') \leq \frac{1}{\tau} J(D)$ *for* $\tau \geq 1$ *when* $0 \leq C(\cdot) \leq \bar{C}$*, or achieve* $J(D') \leq \tau J(D)$ *for* $1 \leq \tau \leq -\frac{\bar{C}}{J(D)}$ *when* $-\bar{C} \leq C(\cdot) \leq 0$*, the number of adversarial users should satisfy:*

$$k \geq \frac{1}{\epsilon} \log \frac{(e^{\epsilon} - 1)\, J(D)\tau + \bar{C}\delta\tau}{(e^{\epsilon} - 1)\, J(D) + \bar{C}\delta\tau} \quad \textit{or} \quad k \geq \frac{1}{\epsilon} \log \frac{(e^{\epsilon} - 1)\, J(D)\tau - \bar{C}\delta}{(e^{\epsilon} - 1)\, J(D) - \bar{C}\delta} \quad \textit{respectively.} \tag{4}$$

The proof is omitted to Appendix E. Corollary 1 shows that stronger privacy guarantee (*i.e.*, smaller ϵ) requires more attackers to achieve the same effectiveness of attack, indicating higher robustness.

We refer the readers to Appendix A for instance-level privacy definition, our proposed algorithms, corresponding privacy analysis and certified robustness in FL. The comparison with existing certified prediction methods in centralized setting are also deferred to Appendix A.

5 Experiments

We present evaluations for robustness certifications, expecially Thm. 2, 3 and Cor. 1. We find that 1) there is a tradeoff between *certified prediction* and privacy on certain datasets; 2) a tighter privacy guarantee *always* provides stronger certified robustness in terms of the *certified attack cost*; 3) our lower bounds of certified attack cost are generally tight when k is small. When k is large, they are tight under strong attacks (*e.g.*, large local poisoning ratio α). Stronger attacks or tighter certification are requried to further tighten the gap between the emprical robustness and theoretical bounds.

Data and Model. We evaluate our robustness certification results with two datasets: MNIST [38] and CIFAR-10 [37]. For each dataset, we use corresponding standard CNN architectures in the differential privacy library [1] of PyTorch [49]. Following previous work on DP ML [36, 43] and backdoor attacks [53, 56] which evaluate with two classes, we focus on binary classification for MNIST (digit 0 and 1) and CIFAR-10 (airplane and bird), and defer the 10-class results to Appendix D. We train FL model following Algorithm 1 for user-level privacy and Algorithm 3 for instance-level privacy. We refer the readers to Appendix D for details about the datasets, networks, parameter setups.

Poisoning Attacks. We evaluate several state-of-the-art poisoning attacks against the proposed `UserDP-FedAvg` and `InsDP-FedAvg`. We first consider backdoor attacks (**BKD**) [6] and label flipping attacks (**LF**) [27]. For `InsDP-FedAvg`, we consider the worst case where k backdoored or lable-flipped instances are fallen into the dataset of one user. For `UserDP-FedAvg`, we additionally evaluate distributed backdoor attack (**DBA**) [58], which is claimed to be a more stealthy backdoor attack against FL. Moreover, we consider BKD, LF and DBA via **model replacement** approach [6] where k attackers train the local models using local datasets with α fraction of poisoned instances, and scale the malicious updates with hyperparameter γ, i.e., $\Delta w_t^i \leftarrow \gamma \Delta w_t^i$, before sending them to the sever. This way, the malicious updates would have a stronger impact on the FL model. Note that even when attackers perform scaling, after server clipping, the sensitivity of updates is still upper-bounded by the clipping threshold S. So the privacy guarantee in Proposition 1 still holds under poisoning attacks via model replacement. Detailed attack setups are presented in Appendix D.

Evaluation Metrics and Setup. We consider two evaluation metrics based on our robustness certification criteria. The first metric is **certified accuracy**, which is the fraction of the test set for which the poisoned FL model makes correct and consistent predictions compared with the clean FL model. Given a test set of size n, for i-th test sample, the ground truth label is y_i, the output prediction is c_i , and the certified number of adversarial users/instances is K_i. We calculate the certified accuracy at k as $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{c_i = y_i \text{ and } \mathsf{K}_i \geq k\}$. The second metric is the **lower bound of attack cost** in Theorem 3: $\underline{J(D')} = \max\{e^{-k\epsilon}J(B) - \frac{1-e^{-k\epsilon}}{e^{\epsilon}-1}\delta\bar{C}, 0\}$. We evaluate the tightness of $\underline{J(D')}$ by comparing it with empirical attack cost $J(D')$. To quantify the robustness, we evaluate the expected class confidence $F_c(\mathcal{M}(D), x)$ for class c via Monte-Carlo sampling. We run the private FL algorithms for M =1000 times, with class confidence $f_c^s = f_c(\mathcal{M}(D), x)$ for each time. We compute its expectation to estimate $F_c(\mathcal{M}(D), x) \approx \frac{1}{M}\sum_{s=1}^{M} f_c^s$ and use it to evaluate Theorem 2. In addition, we use Hoeffding's inequality [33] to calibrates the empirical estimation with confidence level parameter ψ, and results are deferred to Appendix D.6. In terms of the attack cost, we use Example 1, 2 as the definitions of cost function C for backdoor attacks and label flipping attacks respectively. We follow similar protocol to estimate $J(D')$ for Theorem 3 and Corollary 1.

5.1 Robustness Evaluation of User-level DPFL

Certified Prediction. Figure 1(a)(b) present the user-level certified accuracy under different ϵ by training DPFL models with different noise scale σ. In Figure 1(a), we observe that in MNIST the largest k can be certified when ϵ is around 0.6298, which follows the tradeoff between ϵ and certified accuracy as we discussed in Section 4.2. Advanced DP protocols that requires less noise while achieving similar level of privacy are favored to improve the privacy, utility, and certified accuracy

simultaneously. On the other hand, in Figure 1(b), larger k can always be certified with smaller ϵ, which indicates that the optimal ϵ for K might be below 0.2445 for complex data such as CIFAR-10.

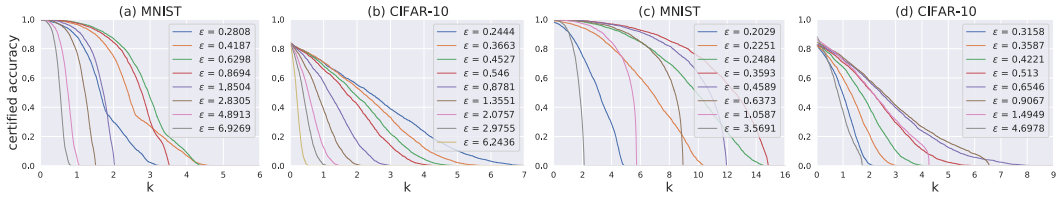

Figure 1: Certified accuracy of FL satisfying user-level DP (a,b), and instance-level DP (c,d).

Certified Attack Cost. In order to evaluate Theorem 3 and characterize the tightness of our theoretical lower bound $\underline{J(D')}$, we compare it with the empirical attack cost $J(D')$ under different local poison fraction α , attack methods and scale factor γ in Figure 2. Note that when $k = 0$, the model is benign so the empirical cost equals to the certified one. We find that 1) when k increases, the attack ability grows, and both the empirical attack cost and theoretical lower bound decreases. 2) In Figure 2 row 1, given the same k, higher α, i.e., poisoning more local instances for each attacker, achieves a stronger attack, under which lower empirical $J(D)$ can be achieved and is more close to the certified lower bound. This indicates that the lower bound appears tighter when the poisoning attack is stronger. 3) In Figure 2 row 2, we fix $\alpha = 100\%$ and evaluate `UserDP-FedAvg` under different γ and attack methods. It turns out that DP serves as a strong defense empirically for FL, given that $J(D)$ did not vary much under different $\gamma(1, 50, 100)$ and different attack methods (BKD, DBA, LF). This is because the clipping operation restricts the magnitude of malicious updates, rendering the model replacement ineffective; the Gaussian noise perturbs the malicious updates and makes the DPFL model stable, and thus the FL model is less likely to memorize the poisoning instances. 4) In both rows, the lower bounds are tight when k is small. When k is large, there remains a gap between our theoretical lower bounds and empirical attack costs under different attacks, which will inspire more effective poisoning attacks or tighter robustness certification. We refer the readers to Appendix D.4 for evaluation of certified attack cost under different ϵ in user-level DPFL.

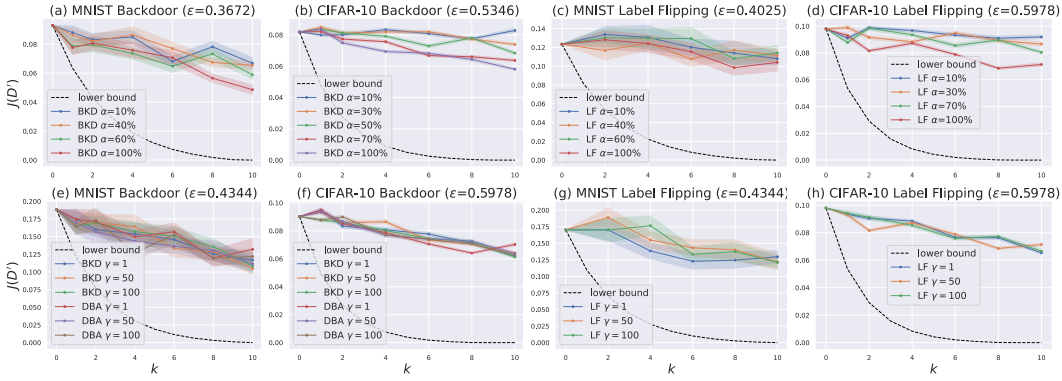

Figure 2: Certified attack cost of user-level DPFL given different k, under attacks with different α (Row 1) or different γ (Row 2).

The robustness evaluation of instance-level DPFL is deferred to Appendix D.5.

6 Conclusion

In this paper, we present the *first* work on deriving certified robustness in DPFL for free against poisoning attacks. We propose two robustness certification criteria, based on which we prove that a FL model satisfying user-level DP is certifiably robust against k adversarial users. Moreover, we propose a novel algorithm for instance-level DPFL, and prove its certified robustness against k adversarial instances. Our theoretical analysis characterizes the inherent relation between certified robustness and differential privacy of FL on both user and instance levels, which are empirically verified with extensive experiments. Our results can be used to improve the trustworthiness of DPFL.

Acknowledgments and Disclosure of Funding

This work is partially supported by the NSF grant No.1910100, NSF CNS 20-46726 CAR, Amazon Research Award, and IBM-ILLINOIS Center for Cognitive Computing Systems Research (C3SR) – a research collaboration as part of the IBM AI Horizons Network.

References

[1] Opacus – train pytorch models with differential privacy, 2021.

[2] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[3] Naman Agarwal, Ananda Theertha Suresh, Felix Yu, Sanjiv Kumar, and H Brendan McMahan. cpsgd: communication-efficient and differentially-private distributed sgd. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7575–7586, 2018.

[4] Sebastien Andreina, Giorgia Azzurra Marson, Helen Möllering, and Ghassan Karame. Baffle: Backdoor detection via feedback-based federated learning. *arXiv preprint arXiv:2011.02167*, 2020.

[5] Shahab Asoodeh and F Calmon. Differentially private federated learning: An information-theoretic perspective. In *ICML Workshop on Federated Learning for User Privacy and Data Confidentiality*, 2020.

[6] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2938–2948. PMLR, 2020.

[7] Gilad Baruch, Moran Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[8] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.

[9] Michael Ben-Or, Shafi Goldwasser, and Avi Wigderson. Completeness theorems for non-cryptographic fault-tolerant distributed computation. In *Proceedings of the twentieth annual ACM symposium on Theory of computing*, pages 1–10, 1988.

[10] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. Analyzing federated learning through an adversarial lens. In *International Conference on Machine Learning*, pages 634–643, 2019.

[11] Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984*, 2018.

[12] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1467–1474, 2012.

[13] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 118–128, 2017.

[14] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191, 2017.

[15] Raphael Bost, Raluca Ada Popa, Stephen Tu, and Shafi Goldwasser. Machine learning classification over encrypted data. In *NDSS*, volume 4324, page 4325, 2015.

[16] Theodora S Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Ch Paschalidis, and Wei Shi. Federated learning of predictive models from federated electronic health records. *International journal of medical informatics*, 112:59–67, 2018.

[17] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

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

[19] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.

[20] Cynthia Dwork. A firm foundation for private data analysis. *Communications of the ACM*, 54(1):86–95, 2011.

[21] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 371–380, 2009.

[22] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[23] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[24] El Mahdi El Mhamdi, Rachid Guerraoui, and Sébastien Louis Alexandre Rouault. The hidden vulnerability of distributed learning in byzantium. In *International Conference on Machine Learning*, number CONF, 2018.

[25] 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.

[26] Shuhao Fu, Chulin Xie, Bo Li, and Qifeng Chen. Attack-resistant federated learning with residual-based reweighting. *arXiv preprint arXiv:1912.11464*, 2019.

[27] Clement Fung, Chris JM Yoon, and Ivan Beschastnikh. The limitations of federated learning in sybil settings. In *23rd International Symposium on Research in Attacks, Intrusions and Defenses ({RAID} 2020)*, pages 301–316, 2020.

[28] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients–how easy is it to break privacy in federated learning? *NeurIPS*, 2020.

[29] Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.

[30] Ran Gilad-Bachrach, Nathan Dowlin, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In *International Conference on Machine Learning*, pages 201–210. PMLR, 2016.

[31] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

[32] Andrew Hard, Kanishka Rao, Rajiv Mathews, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

[33] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

[34] Sanghyun Hong, Varun Chandrasekaran, Yiğitcan Kaya, Tudor Dumitraş, and Nicolas Papernot. On the effectiveness of mitigating data poisoning attacks with gradient shaping. *arXiv preprint arXiv:2002.11497*, 2020.

[35] Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and J Doug Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pages 43–58, 2011.

[36] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private sgd? *Advances in Neural Information Processing Systems*, 33, 2020.

[37] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[38] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[39] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672, 2019.

[40] 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.

[41] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[42] Zhicong Liang, Bao Wang, Quanquan Gu, Stanley Osher, and Yuan Yao. Exploring private federated learning with laplacian smoothing. *arXiv preprint arXiv:2005.00218*, 2020.

[43] Yuzhe Ma, Xiaojin Zhu Zhu, and Justin Hsu. Data poisoning against differentially-private learners: Attacks and defenses. In *International Joint Conference on Artificial Intelligence*, 2019.

[44] Mohammad Malekzadeh, Burak Hasircioglu, Nitish Mital, Kunal Katarya, Mehmet Emre Ozfatura, and Deniz Gunduz. Dopamine: Differentially private federated learning on medical data. *arXiv preprint arXiv:2101.11693*, 2021.

[45] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017.

[46] H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations*, 2018.

[47] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 691–706. IEEE, 2019.

[48] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

[49] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[50] Krishna Pillutla, Sham M Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *arXiv preprint arXiv:1912.13445*, 2019.

[51] Bita Darvish Rouhani, M Sadegh Riazi, and Farinaz Koushanfar. Deepsecure: Scalable provably-secure deep learning. In *Proceedings of the 55th Annual Design Automation Conference*, pages 1–6, 2018.

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

[53] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[54] Stephen Tu. Lecture 20: Introduction to differential privacy.

[55] Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. *NeurIPS*, 2020.

[56] Maurice Weber, Xiaojun Xu, Bojan Karlas, Ce Zhang, and Bo Li. Rab: Provable robustness against backdoor attacks. *arXiv preprint arXiv:2003.08904*, 2020.

[57] Chen Wu, Xian Yang, Sencun Zhu, and Prasenjit Mitra. Mitigating backdoor attacks in federated learning. *arXiv preprint arXiv:2011.01767*, 2020.

[58] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. In *International Conference on Learning Representations*, 2019.

[59] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):12, 2019.

[60] Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. Applied federated learning: Improving google keyboard query suggestions. *arXiv preprint arXiv:1812.02903*, 2018.

[61] Wensi Yang, Yuhang Zhang, Kejiang Ye, Li Li, and Cheng-Zhong Xu. Ffd: a federated learning based method for credit card fraud detection. In *International Conference on Big Data*, pages 18–32. Springer, 2019.

[62] 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.

[63] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[64] Yuqing Zhu, Xiang Yu, Yi-Hsuan Tsai, Francesco Pittaluga, Masoud Faraki, Manmohan Chandraker, and Yu-Xiang Wang. Voting-based approaches for differentially private federated learning, 2021.