Towards the Robustness of Differentially Private Federated Learning

Tao Qi[1*] **, Huili Wang**[1]**, Yongfeng Huang**[1, 2]

[1]Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[2]Zhongguancun Laboratory, Beijing 100094, China
taoqi.qt@gmail.com, whl21@mails.tsinghua.edu.cn, yfhuang@tsinghua.edu.cn

Abstract

Robustness and privacy protection are two important factors of trustworthy federated learning (FL). Existing FL works usually secure data privacy by perturbing local model gradients via the differential privacy (DP) technique, or defend against poisoning attacks by filtering the local gradients in the outlier of the gradient distribution before aggregation. However, these two issues are often addressed independently in existing works, and how to secure federated learning in both privacy and robustness still needs further exploration. In this paper, we unveil that although DP noisy perturbation can improve the learning robustness, DP-FL frameworks are not inherently robust and are vulnerable to a carefully-designed attack method. Furthermore, we reveal that it is challenging for existing robust FL methods to defend against attacks on DP-FL. This can be attributed to the fact that the local gradients of DP-FL are perturbed by random noise, and the selected central gradients inevitably incorporate a higher proportion of poisoned gradients compared to conventional FL. To address this problem, we further propose a new defense method for DP-FL (named *Robust-DPFL*), which can effectively distinguish poisoned and clean local gradients in DP-FL and robustly update the global model. Experiments on three benchmark datasets demonstrate that baseline methods cannot ensure task accuracy, data privacy, and robustness simultaneously, while *Robust-DPFL* can effectively enhance the privacy protection and robustness of federated learning meanwhile maintain the task performance.

Introduction

Federated learning (FL) is a widely used privacy-preserving machine learning paradigm that can train model parameters without accessing raw data (Yang et al. 2019; Li et al. 2020). Existing FL methods are typically based on a decentralized learning framework, where local clients keep their own privacy-sensitive training data and a central several keeps a machine learning model that needs to be trained (McMahan et al. 2017; Lalitha et al. 2018). The clients first employ their local data to train the model parameters, and the global model is further updated by collecting and aggregating local gradients. Then this process is iteratively executed until the model training converges. Unfortunately, the learning process of FL systems is exposed to an open environment, and faces serious risks in privacy protection and model robustness (Lyu et al. 2022; Fang et al. 2020; Li et al. 2021).

In fact, federated learning cannot guarantee the security of user privacy (Geiping et al. 2020; Wang et al. 2019). Many existing works demonstrate that although the privacy-sensitive training data are locally kept by clients, the corresponding user private information can be still recovered from the exchanged local gradients (Zhu, Liu, and Han 2019; Sun et al. 2021). Thus, it is potential for an adversary to filch user privacy from their shared local gradients, which seriously damages the trustworthiness of the FL systems. Differential privacy is a representative privacy protection technique, which is widely used in many real-world privacy-sensitive applications (Kenny et al. 2021). Its core idea is to perturb the exposed data that are relevant to the user privacy via random noise. Therefore, many existing works utilize the differential privacy technique to secure data privacy in federated learning (Wei et al. 2020; Girgis et al. 2021; Geyer, Klein, and Nabi 2017; Truex et al. 2020; Sun, Qian, and Chen 2021). For example, Truex et al. (2020) perturb the local model gradients via Gaussian noise before sharing them with the server. Generally speaking, these differentially private federated learning frameworks (DP-FL) can effectively address the privacy concerns on federated learning.

Besides data privacy, the model robustness of federated learning is also threatened by underlying attacks (Shejwalkar and Houmansadr 2021; Baruch, Baruch, and Goldberg 2019). Since the local model training is invisible for the central server, an adversary can easily poison the local gradients (e.g. injecting backdoors) before sending them to the server. The global model is further poisoned due to its integration with the uploaded poisoned gradients. Intuitively, identifying the poisoned gradients before the model aggregation can defend against such attacks. Therefore, most existing robust FL methods follow a similar assumption that poisoned gradients are usually outliers in the gradient distribution (Awan, Luo, and Li 2021; Yin et al. 2018). Furthermore, these methods only aggregate the local gradients which are in the center of the gradient distribution to avoid the integration of poisoned gradients. For example, Yin et al. (2018) proposed to update the global model based on the element-wise median of local gradients. However, most of

[*]The corresponding author.

the existing robust FL methods only consider the attack on conventional federated learning, which may be inapplicable in the DP-FL system. Thus, how to simultaneously improve the privacy protection and robustness of federated learning needs more study.

In this paper, we explore how to guarantee the privacy protection and robustness of federated learning. We focus on the direction that improving the robustness of the differentially private federated learning to achieve this goal. In general, the random perturbation (e.g., dropout) on the machine learning model usually improves its robustness (Srivastava et al. 2014; Huang et al. 2022; Li et al. 2019; Cohen, Rosenfeld, and Kolter 2019). In the DP-FL methods, the intelligent models are perturbed by the local noise, which makes them more robust than the models learned in conventional FL (Naseri, Hayes, and De Cristofaro 2020; Xie et al. 2022). Thus, this arises the first research question *"Is DP-FL inherently robust to poisoning attacks?"*. To answer this question, we compare the attack performance of existing poisoning attack methods on DP-FL and conventional FL based on three benchmark datasets. We reveal that the attack success rates on DP-FL are substantially reduced compared with those on conventional FL. This result indicates that DP-FL is robust to a part of current poisoning methods. Unfortunately, based on further study, we reveal that the conclusion cannot be generalized and DP-FL is also vulnerable to certain attack patterns. In fact, in DP-FL, the norms of the perturbed gradients are usually much larger than the norms of the unperturbed gradients, which can be exploited by the adversary to enhance their attack. Based on this observation, we propose a new poisoning attack method on DP-FL (named *Attack-DPFL*). The adversary of *Attack-DPFL* re-scales the norm of an unperturbed poisoned gradient to align it with the norm distribution of the perturbed gradients. Then the adversary uploads the amplified unperturbed poisoned gradients instead of the perturbed ones to the server for global model updating. In this way, the global model updating is mainly dominated by the poisoned gradients, instead of the clean gradients or the DP noise regularization, which can ensure the attack success rate. Experiments also empirically verify the attack effectiveness of *Attack-DPFL* on DP-FL.

The successful attack on DP-FL further arises the second research question *"How can we robustly aggregate the local gradients in DP-FL?"*. An intuitive solution is applying existing robust FL methods to defend against the poisoning attacks on DP-FL. However, in DP-FL the clean gradients are perturbed by strong noise while the poisoned gradients are unperturbed, which makes the clean gradients more likely to be outliers than the poisoned gradients. Therefore, protecting DP-FL with robust aggregation methods for conventional FL methods may even lead to worse robustness, since many filtered gradients are clean ones. Fortunately, the exchanged poisoned gradients and clean gradients exhibit different patterns in components. An exchanged clean gradient is composed of an unperturbed gradient and a DP noise, and the DP noise can be substantially reduced in its element aggregation. In contrast, an exchanged poisoned gradient only contains an amplified poisoned gradient, and its element aggregation should tend to be larger than that of a clean gradient due to the amplification operation. Based on this insight, we propose a new robust aggregation method targeting the attacks on DP-FL (named *Robust-DPFL*), which can effectively distinguish poisoned and clean local gradients to learn a clean global model. We define the detection score of an uploaded gradient by averaging its elements, and filter the uploaded gradients with large detection scores for robust aggregation. Extensive experiments on three benchmark datasets demonstrate that baseline robust FL methods can hardly defend against poisoning attacks on DP-FL. Further, experiments also show that *Robust-DPFL* can effectively improve the model robustness and maintain the task accuracy under effective differential privacy guarantees.

The contributions of our work are three-fold:

- We propose an effective attack method for differentially private federated learning, demonstrating that DP-FL is also vulnerable to certain poisoning attacks.
- We propose a robust gradient aggregation method for DP-FL (*Robust-DPFL*), which can effectively identify clean and poisoned gradients under the interference of DP noise.
- Extensive experiments on three benchmark datasets verify that *Robust-DPFL* can ensure the privacy protection, model robustness, and task accuracy of federated learning simultaneously.

Related Work

Federated Learning

Federated learning is a representative machine learning paradigm that can train model parameters from decentralized data in a privacy-preserving way (McMahan et al. 2017; Yang et al. 2019; Zhang et al. 2021). Its core idea is to exchange model gradients instead of the local data for model training (Bonawitz et al. 2017). For example, McMahan et al. (2017) first formulate the framework of federated training: the clients locally train the model parameters and then upload the local model updates to the server, and the server collects and averages local updates to learn the global model. Furthermore, to speed up the model convergence, many works study the adaptive federated learning optimization strategies that can effectively smooth the learning of the global model (Reddi et al. 2021; Yuan and Li 2022; Karimireddy et al. 2020; Zhang et al. 2020; Khanduri et al. 2021; Yuan, Zaheer, and Reddi 2021). In conclusion, the conventional federated learning methods usually focus on how to effectively learn model parameters from decentralized data. However, these conventional FL methods are based on a distributed training framework that is exposed to an open environment, which faces serious risks in terms of both data privacy and model robustness. These risks also promote a line of research to secure federated learning, including differentially private federated learning and robust federated learning, which are reviewed in the following sections.

Differentially Private Federated Learning

Differential privacy techniques can offer theoretical guarantees on the privacy protection of communicated data (Kenny et al. 2021). The main idea of the DP technique is to perturb the communicated data via independent random noise to pose challenges to user privacy identification. Furthermore, DP still allows us to accurately estimate some statistical characteristics of the communicated data since the DP noise can be effectively reduced by aggregating the perturbed data. Thus, the DP technique can be naturally applied to protect user privacy in federated learning, which is widely studied in previous works (Wei et al. 2020; Girgis et al. 2021; Geyer, Klein, and Nabi 2017; Truex et al. 2020; Sun, Qian, and Chen 2021). For example, Truex et al. (2020) proposed to utilize the Gaussian noise to perturb the local model gradients, and then update the global model based on the aggregation of perturbed gradients. Sun, Qian, and Chen (2021) proposed a parameter shuffling-based differentially private FL method that can enhance the trade-off between task accuracy and privacy protection. In conclusion, most of the existing differentially private federated learning (DP-FL) methods focus on studying how to improve the effectiveness of model training under a given privacy protection level (Sun and Lyu 2021). However, these DP-FL methods are also vulnerable to poisoning attacks, which still have serious risks in real-world applications. Different from these methods, we study how to improve the model robustness of differentially private federated learning.

Robust Federated Learning

Poisoning attack is a serious threat to the security of federated learning (Cao et al. 2019; Shejwalkar et al. 2022; Fang et al. 2020). In federated learning, the local model training of a client is invisible to the outside, making it highly convenient for an adversary to poison the local gradients. Thus, in the general framework of existing federated poisoning attack methods, the adversary first employs certain strategies to poison local model gradients and further uploads them to the server to poison the global model (Yin et al. 2018; Shejwalkar and Houmansadr 2021). Most federated poisoning attack methods can be broadly classified into three categories according to their attack purposes (Shejwalkar et al. 2022): (1) targeted attack aiming to degrade the model accuracy on samples in certain groups (Bhagoji et al. 2019; Tolpegin et al. 2020), (2) untargeted attack aiming to degrade the overall task accuracy (Fang et al. 2020; Blanchard et al. 2017), (3) backdoor attack aiming to control the model predictions on poisoned samples embedded with the backdoor triggers (Bagdasaryan et al. 2020; Wang et al. 2020; Xie et al. 2020). For example, Bhagoji et al. (2019) proposed to flip the label of a part of local training data for the untargeted attack, and Bagdasaryan et al. (2020) proposed to learn poisoned model updates based on backdoored training data. In conclusion, these works disclose the vulnerability of federated learning to poisoning attacks and show that it is important to study robust federated learning methods.

There is a line of works studying how to defend against poisoning attacks in federated learning. In practical attack settings, the malicious clients controlled by the adversary should be the minority group in the participating clients. Thus, most of the robust FL methods assume that the poisoned gradients are the outliers in the gradient distribution, and filtering the outliers and only aggregating the gradients in the distribution center can avoid integrating the poisoned gradients into the global model. For example, Yin et al. (2018) proposed to update the global model based on the median of local gradients in each dimension. Blanchard et al. (2017) proposed to select the local model gradient that is most relevant to other gradients to update the global model. Generally speaking, most of the existing robust gradient aggregation methods are designed for conventional FL, which is difficult to be applied in differentially private FL. This is because in DP-FL local gradients are perturbed by DP noise, which may make the outlier assumption not hold. Besides, many current robust FL methods update the global model based on a very small fraction of local gradients (Blanchard et al. 2017; Yin et al. 2018), which is difficult to reduce the damage of DP noise on task accuracy. Different from these works, we propose a new robust differentially private FL framework, which can simultaneously ensure the data privacy, model robustness, and model accuracy.

Preliminary

Problem Definition

In our work, we assume that there are N local clients participating in the federated learning, where $\mathcal{U} = \{u_i | i = 1, 2, ..., N\}$ denotes the set of the participated clients and u_i denotes the i-th client. The local training dataset kept by the i-th client u_i is denoted as $\mathcal{T}_i$, and the machine learning model that needs to be trained in federated learning is denoted by Θ, which is kept by the server and is aligned in local clients. The local clients and the server will collaborate with each other to train the model on decentralized data in a federated way. Besides, we assume that there may be an adversary that can filch the communicated local data from the FL system to recover user privacy. To guarantee the privacy protection of users, the local gradients must be protected by differential privacy before sharing them with the outside. Moreover, we assume that there may be an adversary that can control a part of local clients to poison the global model for certain malicious purposes. Thus, the server should aggregate the local gradients in a robust manner to defend against potential attacks. We employ a variable a_i to represent whether the i-th client is a benign client ($a_i = 0$) or a malicious client ($a_i = 1$). Furthermore, we also assume that the benign users will honestly follow the framework of DP-FL, while the malicious clients can generate the poisoned gradient based on any strategy.

Renyi Differential Privacy

We employ Renyi differential privacy (Mironov 2017) to guarantee the privacy protection of FL. Renyi differential privacy can be defined based on the following formulation:

Definition 1 **($(\alpha, \epsilon) - RDP$)** *Let $f : \mathcal{X} \rightarrow \mathcal{Y}$ represent the randomized mechanism for privacy protection, we call f have ϵ-Renyi differently privacy of order α (shorted as (α, ϵ)-RDP), if and only if for arbitrary adjacent data set $\mathcal{X}_i, \mathcal{X}_j \subseteq \mathcal{X}$, the following inequation holds:*

$$\frac{1}{\alpha - 1} \log \mathbb{E}[P(x)/Q(x)]^{\alpha} \leq \epsilon, \quad \alpha > 1, \quad \epsilon > 0. \quad (1)$$

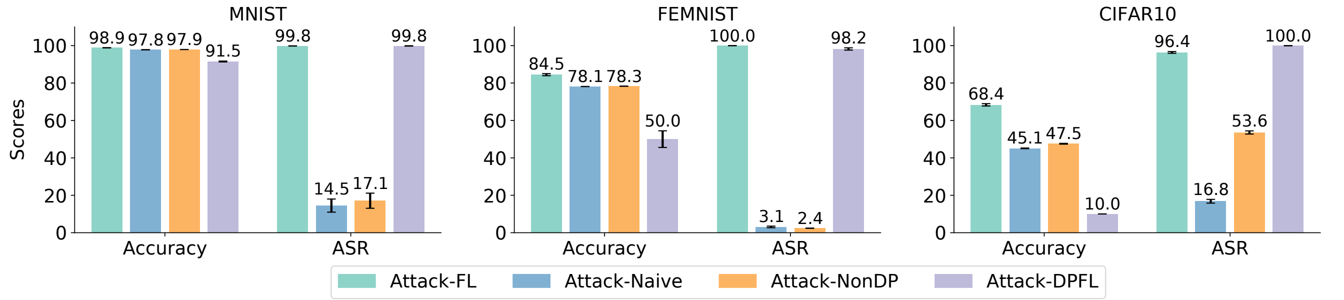
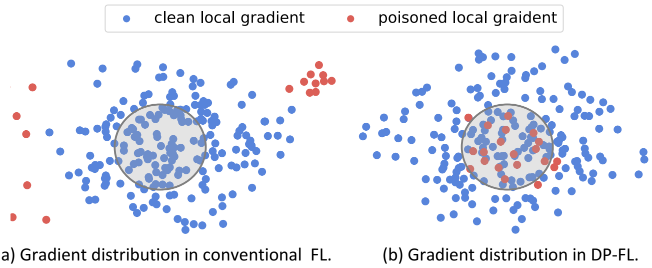


Figure 1: Poisoning attacks on conventional FL and DP-FL. *Attack-FL* denotes the results of conventional FL (FedAvg) under existing attack methods. *Attack-Naive* denotes the results of DP-FL under existing attack methods. *Attack-NonDP* denotes the results of DP-FL under a naive variant attack method. *Attack-DPFL* denotes results of DP-FL under our proposed attack method.



(a) Gradient distribution in conventional FL.

(b) Gradient distribution in DP-FL.

Figure 2: Gradient distributions of conventional FL and DP-FL. Gradients in the gray circle are aggregated for global model updating in previous robust aggregation methods.

α and ϵ represents the security level of privacy, where larger α and smaller ϵ means stronger privacy protection. Furthermore, it can be proved that a Gaussian noise based randomized mechanism can satisfy (α, ϵ)-RDP based on the Lemma in the Appendix. Thus, in the DP-FL framework, we perturb the local gradients via Gaussian noise before uploading them to the server.

Experimental Setup

Next, we will introduce the basic experimental setups for exploring the research questions in our work. We conduct experiments on three benchmark datasets for federated learning, including *MNIST* (Deng 2012), *FEMNIST* (Caldas et al. 2018), and *CIFAR-10* (Krizhevsky, Hinton et al. 2009). The training data is randomly partitioned into 100 clients based on a non-IID data partition strategy (Hsu, Qi, and Brown 2019). The basic machine learning models trained on these three datasets are implemented by ResNet-18 (He et al. 2016). Besides, we note that the purpose of some poisoning attack methods (i.e., targeted and untargeted performance) is to degrade the model performance on certain or all samples, while the DP technique applied in FL will also degrade the model accuracy. Thus, due to the interference of DP noise, it is difficult to verify the attack and defense effectiveness if we employ targeted and untargeted attack methods for experiments. We remark that the federated backdoor attack aims to inject backdoors into models to control their predictions on poisoned data while keep their normal predictions on clean data, which will not degrade the model performance on clean test data. Therefore, we employ a federated backdoor attack method (Bagdasaryan et al. 2020) to poison different FL methods to verify their robustness. The proportion of malicious clients controlled by the adversary is set to 15%. Each malicious client generate the poisoned gradient by training the model on the poisoned data. Besides, we utilize the accuracy score for task performance verification and utilize the attack success rate (ASR) for defense effectiveness verification. A larger accuracy score means better task performance and a lower ASR score means better defense performance. In addition, the level of DP guarantee is set to $(1.2, 5)$. More details can be found in the Appendix.

Approach

Differentially Private FL Framework

Next, we will first briefly introduce the workflow of differentially private federated learning (DP-FL). In each training round of DP-FL, the server first distributes the current model Θ_t to local clients and then selects a part of clients for local training. The set of selected clients is denoted as $\mathcal{U}_t$. For each selected client $u \in \mathcal{U}_t$, it first trains the model Θ_t on its local training dataset $\mathcal{T}_u$ to build the model gradient $\mathbf{G}_u$. Then we clip the norm of the gradient $\mathbf{G}_u$ and perturb it via the Gaussian noise according to the DP technique: $\mathbf{S}_u = h(\mathbf{G}_u; l) + N(0, \sigma^2)$, where l is the clipping value of the gradient norm, σ is the DP noise intensity, and $h(\mathbf{G}, l)$ is the function that clips the L2 norm of $\mathbf{G}$ to l. Then the client u uploads the protected gradient $\mathbf{S}_u$ to the server for aggregation. Furthermore, the server updates the global model based on the collected local gradients: $\Theta_{t+1} = \Theta_t - \beta \frac{1}{|\mathcal{U}_t|} \sum_u \mathbf{S}_u$, where β is the learning rate. Then the server iteratively repeats the process until model training convergences. In our work, we focus on developing privacy secured and robust federated learning method. We remark that we follow the direction of improving the robustness of DP-FL to achieve the research purpose. Next, we will study the research questions on the robustness of DP-FL.

Attack on Differentially Private FL

The first research question is *"Is DP-FL inherently robust to poisoning attacks?"*. Generally speaking, incorporating ran-

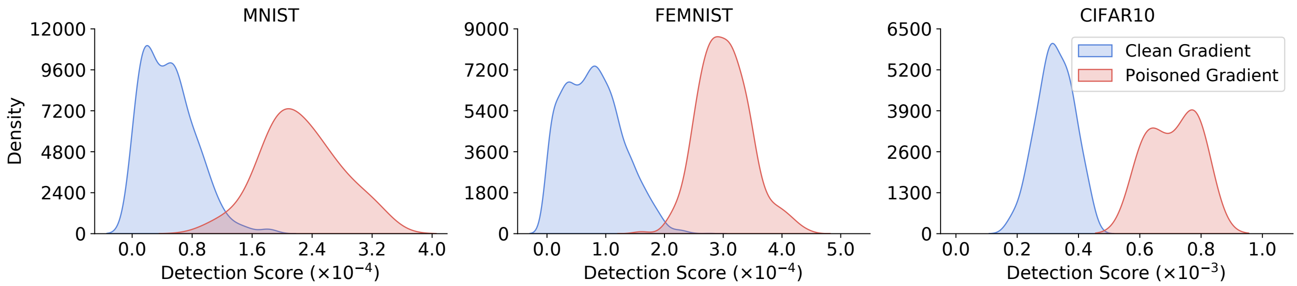


Figure 3: Detection score distributions of clean and poisoned gradients on three benchmark datasets.

domized perturbation into the training of the machine learning models can improve their robustness (Srivastava et al. 2014; Huang et al. 2022; Li et al. 2019; Cohen, Rosenfeld, and Kolter 2019). The training of the global model in DP-FL can be formulated by the combination of an unperturbed equivalent gradient $\overline{\mathbf{G}}_t$ and an equivalent noise $\overline{\mathbf{N}}$:

$$\Theta_{t+1} = \Theta_t - \beta(\overline{\mathbf{G}}_t + \overline{\mathbf{N}}),$$
$$\overline{\mathbf{G}}_t = \frac{1}{|\mathcal{U}_t|}\sum_{u\in\mathcal{U}_t} h(\mathbf{G}_u; l), \quad \overline{\mathbf{N}} = \frac{1}{|\mathcal{U}_t|}\sum_{u\in\mathcal{U}_t} \mathbf{N}_u, \tag{2}$$

where $\mathbf{N}_u$ represents the Gaussian noise sampled for the gradient $\mathbf{G}_u$. Thus, the global model training in DP-FL is perturbed by an equivalent noise, and the model may be more robust than the model trained in conventional FL. This indicates that the DP-FL may be naturally robust to poisoning attacks and does not need extra robustness enhancement. To explore this problem, we conduct experiments on three datasets to evaluate the performance of conventional FL and DP-FL under poisoning attacks (Fig. 1). Results show that although the existing attack method can poison the model in conventional FL with high ASR scores (results of *Attack-FL*), it can hardly poison the model of DP-FL (results of *Attack-Naive*). This indicates that DP-FL may be robust to some existing federated poisoning attack methods. However, the malicious clients may not follow the training framework of DP-FL, and they can upload unperturbed gradients rather than perturbed gradients to weaken the perturbation on the global model training and improve attack effectiveness. Based on this intuition, we further apply this naive attack method on DP-FL and the results show that its attack is still ineffective (*Attack-NonDP*). To explain this phenomenon, we define a metric that measures the relative perturbation intensity (RPI) on the global model updating: $\gamma \triangleq \mathbb{E}[||\overline{\mathbf{N}}||]/\mathbb{E}[||\overline{\mathbf{G}_t}||]$, where larger γ means stronger perturbation and better robustness. Thus, for these two attack methods on DP-FL, we can formulate their RPI by:

$$\gamma_1 = \sqrt{\frac{D}{M}}\sigma/G, \quad \gamma_2 = \sqrt{\frac{pD}{M}}\sigma/G,$$
$$M = |\mathcal{U}_t|, \quad p = \frac{\sum_{u\in\mathcal{U}_t} a_u}{|\mathcal{U}_t|}, \quad G = \mathbb{E}[||\overline{\mathbf{G}_t}||], \tag{3}$$

where γ_1 and γ_2 corresponds to *Attack-Naive* and *Attack-NonDP* respectively, and p represents the proportion of benign clients. Since the benign clients are the majority, the value of p is usually larger than 0.5 (e.g., $p = 0.85$ in our settings), making *Attack-NonDP* can only slightly reduce the RPI. The empirical results and the theoretical analysis shows that it is not a trivial task to attack differentially private FL.

Since it is difficult to reduce the RPI by weakening the equivalent noise, another direction is to amplify the equivalent gradient norm. In general, the norm of the DP noise is usually much larger than the norm of the local gradient, since the protection of high-dimensionality data usually requires strong DP noise. Thus, the adversary may be able to exploit this pattern to amplify the poisoned gradients, which can make the poisoned gradients dominate the equivalent gradient meanwhile reduce the RPI. Based on this observation, we propose a new poisoning attack method on DP-FL (named *Attack-DPFL*). For a malicious client u, *Attack-DPFL* first learns the poisoned gradient $\mathbf{G}_u$ from its local poisoned data. Then *Attack-DPFL* simulates the distribution of the perturbed poisoned gradient and aligns their norms to learn an amplified unperturbed poisoned gradient $\mathbf{A}_u$:

$$\mathbf{A}_u = A\mathbf{G}_u, \quad A = ||\mathbf{S}_u||/||\mathbf{G}_u||$$
$$\mathbf{S}_u = h(\mathbf{G}_u; l) + N(0, \sigma^2), \tag{4}$$

where A is the amplification coefficient. Next, we approximate the norm of the equivalent gradient in *Attack-DPFL* via Eq. 5 and obtain the approximated RPI $\gamma_3 = \frac{\sqrt{D}\sigma}{\sqrt{(1-p)M}AG}$.

$$||\overline{\mathbf{G}}_t|| = ||\frac{1}{M}\sum_{a_u=0} h(\mathbf{G}_u; l) + \frac{1}{M}\sum_{a_u=1} A^2\mathbf{G}_u||$$
$$\approx ||\frac{1}{M}\sum_{a_u=1} A^2\mathbf{G}_u|| \approx \sqrt{1-p}AG. \tag{5}$$

Since the expectation of A is usually large in DP-FL (i.e., $A = 100$ in our DP settings), γ_3 is much lower than γ_1 and γ_2, indicating that *Attack-DPFL* is promising in reducing the training perturbation in DP-FL and improving the attack effectiveness. Moreover, we further conduct experiments for empirical verification. Fig. 1 show that the ASR of *Attack-DPFL* on differentially private FL and the ASR of baseline attack methods on conventional FL are comparable, demonstrating *Attack-DPFL* can effectively attack DP-FL. These results also disclose that differentially private FL is not inherently robust to poisoning attacks and also needs protection against poisoning attacks.

Robust Differentially Private FL Framework

The second research question is *"How can we robustly aggregate the local gradients in DP-FL?"*. An intuitive solution is to apply existing robust gradient aggregation methods for conventional FL to protect DP-FL. Most existing robust gradient aggregation methods are usually based on the assumption that poisoned gradients are outliers in the gradient distribution. Therefore, updating the global model based on the gradients in the distribution center can effectively exclude the poisoned gradients and learn a clean model (Fig. 2 (a)). Unfortunately, the outlier assumption does not hold in DP-FL. This is because, in the positioning attacks on DP-FL, the clean gradients are perturbed by strong DP noise, while the poisoned gradients are not. Consequently, the positioned gradients tend to be more relevant to each other, and the outliers in the gradient distribution of DP-FL are more likely to be clean gradients rather than poisoned gradients. This makes existing robust FL methods that aggregate gradients in the distribution center will incorporate more poisoned gradients into the global model in DP-FL than conventional FL, seriously degrading the defense effectiveness (Fig. 2 (b)). In addition, some current robust aggregation methods only select a very small fraction of gradients (e.g., only one of the local gradients) for global model updating. This may also make the DP noise cannot be effectively reduced in gradient aggregation and hurt the task accuracy. Thus, how to defend against poisoning attacks in differentially private FL needs further study.

Fortunately, in DP-FL the gradients shared by benign clients and malicious clients exhibit different patterns in their components. A shared gradient from benign clients is composed of an unperturbed clean gradient and a DP noise vector, while a shared gradient from malicious clients only contains an amplified unperturbed poisoned gradient. Furthermore, the elements in noise vectors and unperturbed gradients also exhibit highly distinctive patterns. Elements of a noise vector are independent and their element-wise aggregation usually degrades to zero, while elements in unperturbed gradients are usually correlated and their element-wise aggregation is usually a nonzero value. These observations inspire us that the element-wise aggregation $Z(\mathbf{S}) = |\frac{1}{D}\sum_{i=1}^{D}\mathbf{S}[i]|$ of the shared gradients may be informative for identifying poisoned gradients, where $\mathbf{S}[i]$ denotes the i-th element of the vector and D denotes the dimensionality. Thus, we formulate the detection score Z_c and Z_p of clean and poisoned gradients based on element aggregation:

$$
\begin{aligned}
Z_c &= Z(h(\mathbf{G};l)+\mathbf{N}) \approx |\frac{1}{D}\sum_{i=1}^{D}\mathbf{G}_i|, \\
Z_p &= Z(\mathbf{A}) = |\frac{A}{D}\sum_{i=1}^{D}\mathbf{G}| \approx AZ_c,
\end{aligned} \tag{6}
$$

Eq. 6 shows that Z_p is typically much larger than Z_c due to the amplification coefficient A. Eq. 6 also indicates that it is possible to distinguish poisoned and clean gradients based on their element-wise aggregation. We conduct experiments to further support our analysis, and results in Fig. 3 show that clean and poisoned gradients have highly distinctive based on their detection scores. Based on these analyses, we propose a robust gradient aggregation framework for DP-FL (named *Robust-DPFL*). In each training round of *Robust-DPFL*, the server first models the detection score of each uploaded gradient $\{Z(\mathbf{S}_u)|u \in \mathcal{U}_t\}$. Then the server clusters the detection scores into two groups based on the K-means algorithm to detect the suspicious gradients. The server treats the cluster with a larger averaged detection score as the abnormal cluster and only aggregates the gradients in another cluster to learn the global model.

	MNIST		FEMNIST		CIFAR10	
	ACC	ASR	ACC	ASR	ACC	ASR
IdealFL	98.44	0.20	78.85	2.67	36.53	4.46
FedAvg	91.51	99.84	50.01	98.22	10.01	99.97
Mid	79.48	18.44	29.06	61.64	10.00	100.00
Krum	9.97	83.29	4.97	100.00	10.00	33.46
MKrum	9.80	100.00	5.16	99.82	10.00	100.00
Norm	86.05	66.99	47.10	90.65	10.00	100.00
Contra	10.06	83.33	4.97	100.00	10.00	100.00
Ours	97.43	0.39	76.79	2.49	33.46	1.55

Table 1: The model accuracy and robustness of different robust federated learning methods.

Experiment

Performance Evaluation

Next, we will compare the accuracy and robustness of different federated learning methods under the protection of differential privacy. Several representative robust FL methods are compared, including: (1) *Mid* (Yin et al. 2018): update the global model based on the median of the local gradients for each dimensionality. (2) *Krum* (Blanchard et al. 2017): update the global model based on the local gradient that is most relevant to other gradients. (3) *MKrum* (Blanchard et al. 2017): update the global model by averaging multiple local gradients that are most relevant to other gradients. (4) *Norm* (Sun et al. 2019): clip the norm of local gradients according to a threshold to update the global model based on the clipped gradients. (5) *Contra* (Awan, Luo, and Li 2021): aggregate local gradients weighted by their cosine similarities with other gradients to update the global model. In addition, we provide the results of *DPFL* (Truex et al. 2020) without any defense mechanisms for comparisons. We also provide the results of ideal results of differentially private FL (*IdealFL*) which can filter the poisoned gradients with 100% accuracy. The code is available in https://github.com/taoqi98/Robust-DPFL.

Each experiment is repeated 10 times and we show the averaged scores in Table 1. First, robust gradient aggregation methods for conventional FL can hardly defend against poisoning attacks under DP-FL. For example, the attack success rate on *Krum* on *FEMNIST* is close to 100%. This is because baseline robust FL methods assume that updating the global model based on the gradients in the center can avoid incorporating poisoned gradients into the model. However, in DP-FL the clean gradients are perturbed by strong DP noise

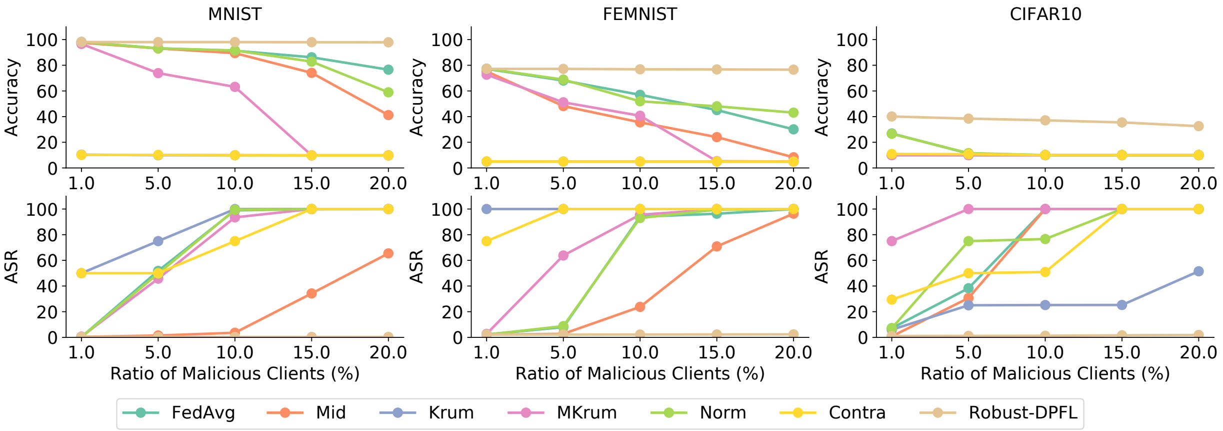


Figure 4: The performance of different robust FL methods under varying ratios of malicious clients.

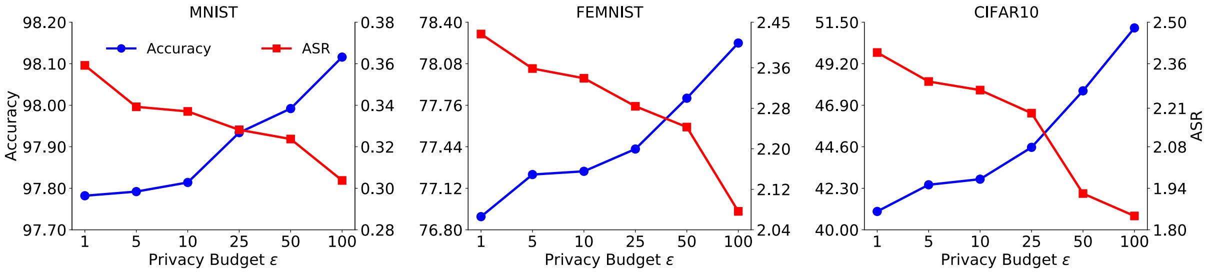


Figure 5: The trade-off of *Robust-DPFL* in task accuracy, model robustness and privacy protection.

while the poisoned gradients are not, making the gradients in the distribution center more likely to be poisoned. Second, the task accuracy of some baseline robust FL methods is worse than the results of *FedAvg* without any defense against poisoning attacks. This is because some baseline robust FL methods only select a very fraction of gradients (e.g., *Krum* only selects one of the local gradients for global model updating). Furthermore, the DP noise in the uploaded gradients cannot be reduced in the gradient aggregation and damage the model training. Third, *Robust-DPFL* can effectively defend against the poisoning attacks on DP-Fl, meanwhile maintains the comparable task accuracy as ideal clean training. This is because clean and poisoned gradients in DP-FL exhibit discriminative patterns in their components. Based on this discriminative pattern, we propose an effective detection algorithm that can identify poisoned local gradients from clean gradients. By accurately filtering poisoned gradients, *Robust-DPFL* can ensure task accuracy, model robustness, and privacy protection at the same time.[1]

Influence of Malicious Client Ratios

Next, we evaluate the performance of *Robust-DPFL* and baselines under different ratios of malicious clients. Results are shown in Fig. 4, from which we find that the defense effectiveness of baselines rapidly decreases with the increase of malicious client ratio. Different from these methods, *Robust-DPFL* consistently improves the robustness of DP-FL and maintains the task performance. These results further demonstrate the effectiveness of *Robust-DPFL* in balancing accuracy, privacy, and robustness of federated learning.

Accuracy, Privacy, and Robustness Trade-off

Next, we conduct experiments to analyze the trade-off among accuracy, privacy protection, and model robustness of *Robust-DPFL*. Results are shown in Fig. 5, from which we have several findings. First, with the increase of privacy protection levels (i.e., lower privacy budget), the accuracy of *Robust-DPFL* decreases. This is because stronger DP noise is incorporated into the training to better secure privacy, which damages the model training. Second, stronger security guarantees also slightly degrade the model robustness. This maybe because stronger DP noise makes the distribution of benign and malicious gradients to be disordered, posing challenges to identifying and aggregating the benign ones. These results can provide guidance in balancing the accuracy, robustness, and privacy protection of *Robust-DPFL*. Moreover, we also conduct other experiments to analyze *Robust-DPFL*, including (1) verifying the generalization of our work under more poisoning attacks, and (2) evaluating the effectiveness of the detection mechanism. Due to space limitations, the results and discussions are in the Appendix.

[1] Results show that some methods exactly achieve 10% task accuracy on CIFAR10. This is because the model prediction will degrade into a single class under the interference of differential privacy and poisoning attacks. Since CIFAR10 is a balanced dataset with ten categories, the accuracy consequently aligns at 10%.

Conclusion

In this paper, we study the robustness issue of differentially private federated learning. We first propose a poisoning attack method targeting DP-FL (named *Attack-DPFL*) that can weaken the robustness of DP-FL powered by DP noise. *Attack-DPFL* amplifies the poisoned gradients by aligning their norm distributions with perturbed clean gradients, which enables the poisoned gradients to dominate the global model updating. Extensive experimental results on three benchmark datasets demonstrate the difficulty of existing poisoning attacks in compromising DP-FL, while highlighting the capability of *Attack-DPFL* in poisoning DP-FL models. Furthermore, we propose a robust gradient aggregation method for DP-FL (named *Robust-DPFL*), which can learn a clean global model under the interference of DP noise. *Robust-DPFL* first identifies poisoned gradients from clean gradients based on their discriminative patterns in components and then updates the global model on identified clean gradients. Experiments show that *Robust-DPFL* can achieve an effective trade-off among task accuracy, model robustness, and privacy protection, while baseline robust federated learning methods cannot.

Acknowledgments

This work was supported by the National Key R&D Program of China 2023YFB4502100, and National Natural Science Foundation of China under Grant U2336208.

References

Awan, S.; Luo, B.; and Li, F. 2021. Contra: Defending against poisoning attacks in federated learning. In *ESORICS*, 455–475.

Bagdasaryan, E.; Veit, A.; Hua, Y.; Estrin, D.; and Shmatikov, V. 2020. How to backdoor federated learning. In *AISTATS*, 2938–2948. PMLR.

Baruch, G.; Baruch, M.; and Goldberg, Y. 2019. A little is enough: Circumventing defenses for distributed learning. In *NeurIPS*.

Bhagoji, A. N.; Chakraborty, S.; Mittal, P.; and Calo, S. 2019. Analyzing federated learning through an adversarial lens. In *ICML*, 634–643.

Blanchard, P.; El Mhamdi, E. M.; Guerraoui, R.; and Stainer, J. 2017. Machine learning with adversaries: Byzantine tolerant gradient descent. In *NeurIPS*.

Bonawitz, K.; Ivanov, V.; Kreuter, B.; Marcedone, A.; McMahan, H. B.; Patel, S.; Ramage, D.; Segal, A.; and Seth, K. 2017. Practical secure aggregation for privacy-preserving machine learning. In *CCS*, 1175–1191.

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

Cao, D.; Chang, S.; Lin, Z.; Liu, G.; and Sun, D. 2019. Understanding distributed poisoning attack in federated learning. In *ICPADS*, 233–239.

Cohen, J.; Rosenfeld, E.; and Kolter, Z. 2019. Certified adversarial robustness via randomized smoothing. In *ICML*, 1310–1320. PMLR.

Deng, L. 2012. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6): 141–142.

Fang, M.; Cao, X.; Jia, J.; and Gong, N. Z. 2020. Local model poisoning attacks to byzantine-robust federated learning. In *USENIX Security*, 1623–1640.

Geiping, J.; Bauermeister, H.; Dröge, H.; and Moeller, M. 2020. Inverting gradients-how easy is it to break privacy in federated learning? In *NeurIPS*, 16937–16947.

Geyer, R. C.; Klein, T.; and Nabi, M. 2017. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*.

Girgis, A.; Data, D.; Diggavi, S.; Kairouz, P.; and Suresh, A. T. 2021. Shuffled model of differential privacy in federated learning. In *AISTATS*, 2521–2529.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hsu, T.-M. H.; Qi, H.; and Brown, M. 2019. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*.

Huang, P.; Yang, Y.; Jia, F.; Liu, M.; Ma, F.; and Zhang, J. 2022. Word level robustness enhancement: Fight perturbation with perturbation. In *AAAI*, 10785–10793.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; and Suresh, A. T. 2020. Scaffold: Stochastic controlled averaging for federated learning. In *ICML*, 5132–5143.

Kenny, C. T.; Kuriwaki, S.; McCartan, C.; Rosenman, E. T.; Simko, T.; and Imai, K. 2021. The use of differential privacy for census data and its impact on redistricting: The case of the 2020 US Census. *Science advances*, 7(41): eabk3283.

Khanduri, P.; Sharma, P.; Yang, H.; Hong, M.; Liu, J.; Rajawat, K.; and Varshney, P. 2021. Stem: A stochastic two-sided momentum algorithm achieving near-optimal sample and communication complexities for federated learning. In *NeurIPS*, 6050–6061.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Lalitha, A.; Shekhar, S.; Javidi, T.; and Koushanfar, F. 2018. Fully decentralized federated learning. In *Third workshop on bayesian deep learning (NeurIPS)*, volume 2.

Li, B.; Chen, C.; Wang, W.; and Carin, L. 2019. Certified adversarial robustness with additive noise. In *NeurIPS*, volume 32.

Li, T.; Hu, S.; Beirami, A.; and Smith, V. 2021. Ditto: Fair and robust federated learning through personalization. In *ICML*, 6357–6368.

Li, T.; Sahu, A. K.; Talwalkar, A.; and Smith, V. 2020. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3): 50–60.

Lyu, L.; Yu, H.; Ma, X.; Chen, C.; Sun, L.; Zhao, J.; Yang, Q.; and Philip, S. Y. 2022. Privacy and robustness in federated learning: Attacks and defenses. *TNNLS*.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *AISTATS*, 1273–1282.

Mironov, I. 2017. Rényi differential privacy. In *CSF*, 263–275. IEEE.

Naseri, M.; Hayes, J.; and De Cristofaro, E. 2020. Local and central differential privacy for robustness and privacy in federated learning. *arXiv preprint arXiv:2009.03561*.

Reddi, S. J.; Charles, Z.; Zaheer, M.; Garrett, Z.; Rush, K.; Konečnỳ, J.; Kumar, S.; and McMahan, H. B. 2021. Adaptive Federated Optimization. In *ICLR*.

Shejwalkar, V.; and Houmansadr, A. 2021. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. In *NDSS*.

Shejwalkar, V.; Houmansadr, A.; Kairouz, P.; and Ramage, D. 2022. Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning. In *S&P*, 1354–1371.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 1929–1958.

Sun, J.; Li, A.; Wang, B.; Yang, H.; Li, H.; and Chen, Y. 2021. Soteria: Provable defense against privacy leakage in federated learning from representation perspective. In *CVPR*, 9311–9319.

Sun, L.; and Lyu, L. 2021. Federated Model Distillation with Noise-Free Differential Privacy. In *IJCAI*, 1563–1570.

Sun, L.; Qian, J.; and Chen, X. 2021. LDP-FL: Practical Private Aggregation in Federated Learning with Local Differential Privacy. In *IJCAI*, 1571–1578.

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

Tolpegin, V.; Truex, S.; Gursoy, M. E.; and Liu, L. 2020. Data poisoning attacks against federated learning systems. In *ESORICS*, 480–501.

Truex, S.; Liu, L.; Chow, K.-H.; Gursoy, M. E.; and Wei, W. 2020. LDP-Fed: Federated learning with local differential privacy. In *EdgeSys*, 61–66.

Wang, H.; Sreenivasan, K.; Rajput, S.; Vishwakarma, H.; Agarwal, S.; Sohn, J.-y.; Lee, K.; and Papailiopoulos, D. 2020. Attack of the tails: Yes, you really can backdoor federated learning. In *NeurIPS*, 16070–16084.

Wang, Z.; Song, M.; Zhang, Z.; Song, Y.; Wang, Q.; and Qi, H. 2019. Beyond inferring class representatives: User-level privacy leakage from federated learning. In *INFOCOM*, 2512–2520.

Wei, K.; Li, J.; Ding, M.; Ma, C.; Yang, H. H.; Farokhi, F.; Jin, S.; Quek, T. Q.; and Poor, H. V. 2020. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15: 3454–3469.

Xie, C.; Huang, K.; Chen, P.-Y.; and Li, B. 2020. Dba: Distributed backdoor attacks against federated learning. In *ICLR*.

Xie, C.; Long, Y.; Chen, P.-Y.; and Li, B. 2022. Uncovering the Connection Between Differential Privacy and Certified Robustness of Federated Learning against Poisoning Attacks. *arXiv preprint arXiv:2209.04030*.

Yang, Q.; Liu, Y.; Chen, T.; and Tong, Y. 2019. Federated machine learning: Concept and applications. *TIST*, 1–19.

Yin, D.; Chen, Y.; Kannan, R.; and Bartlett, P. 2018. Byzantine-robust distributed learning: Towards optimal statistical rates. In *ICML*, 5650–5659.

Yuan, H.; Zaheer, M.; and Reddi, S. 2021. Federated composite optimization. In *ICML*, 12253–12266.

Yuan, X.; and Li, P. 2022. On convergence of FedProx: Local dissimilarity invariant bounds, non-smoothness and beyond. In *NeurIPS*, 10752–10765.

Zhang, C.; Xie, Y.; Bai, H.; Yu, B.; Li, W.; and Gao, Y. 2021. A survey on federated learning. *Knowledge-Based Systems*, 216: 106775.

Zhang, X.; Hong, M.; Dhople, S.; Yin, W.; and Liu, Y. 2020. Fedpd: A federated learning framework with optimal rates and adaptivity to non-iid data. *arXiv preprint arXiv:2005.11418*.

Zhu, L.; Liu, Z.; and Han, S. 2019. Deep leakage from gradients. In *NeurIPS*.