Make Landscape Flatter in Differentially Private Federated Learning

Yifan Shi[1] Yingqi Liu[2] Kang Wei[2] Li Shen[3,*] Xueqian Wang[1,*] Dacheng Tao[3]
[1]Tsinghua University, Shenzhen, China; [3]JD Explore Academy, Beijing, China
[2]Nanjing University of Science and Technology, Nanjing, China
shiyf21@mails.tsinghua.edu.cn; lyq@njust.edu.cn; kang.wei@njust.edu.cn;
mathshenli@gmail.com; wang.xq@sz.tsinghua.edu.cn; dacheng.tao@gmail.com

Abstract

To defend the inference attacks and mitigate the sensitive information leakages in Federated Learning (FL), client-level Differentially Private FL (DPFL) is the de-facto standard for privacy protection by clipping local updates and adding random noise. However, existing DPFL methods tend to make a sharper loss landscape and have poorer weight perturbation robustness, resulting in severe performance degradation. To alleviate these issues, we propose a novel DPFL algorithm named DP-FedSAM, which leverages gradient perturbation to mitigate the negative impact of DP. Specifically, DP-FedSAM integrates Sharpness Aware Minimization (SAM) optimizer to generate local flatness models with better stability and weight perturbation robustness, which results in the small norm of local updates and robustness to DP noise, thereby improving the performance. From the theoretical perspective, we analyze in detail how DP-FedSAM mitigates the performance degradation induced by DP. Meanwhile, we give rigorous privacy guarantees with Rényi DP and present the sensitivity analysis of local updates. At last, we empirically confirm that our algorithm achieves state-of-the-art (SOTA) performance compared with existing SOTA baselines in DPFL.

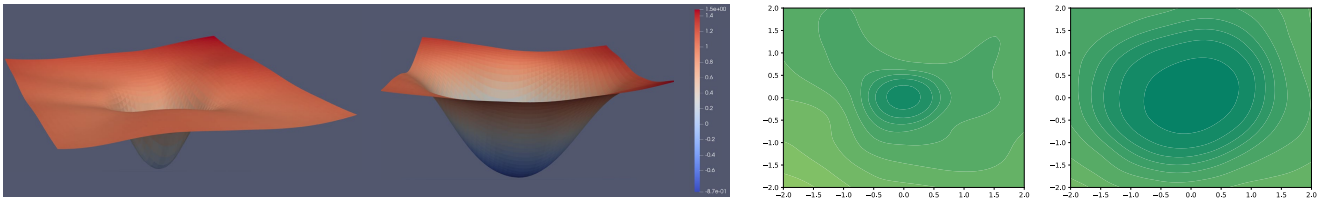
1. Introduction

Federated Learning (FL) [31] allows distributed clients to collaboratively train a shared model without sharing data. However, FL faces severe dilemma of privacy leakage [27]. Recent works show that a curious server can also infer clients' privacy information such as membership and data features, by well-designed generative models and/or shadow models [14, 37, 40, 44, 55]. To address this issue, differential privacy (DP) [12] has been introduced in FL, which can protect every instance in any client's data (instance-level DP [3,20,47,48]) or the information between clients (client-level DP [8, 15, 21, 25, 35, 52]). In general, client-level DP are more suitable to apply in the real-world setting due to a better model performance. For instance, a language prediction model with client-level DP [16, 35] has been applied on mobile devices by Google. In general, the Gaussian noise perturbation-based method is commonly adopted for ensuring the strong client-level DP. However, this method includes two operations in terms of clipping the l_2 norm of local updates to a sensitivity threshold C and adding random noise proportional to the model size, whose standard deviation (STD) is also decided by C. These steps may cause severe performance degradation dilemma [9, 22], especially on large-scale complex model [45], such as ResNet-18 [18], or with heterogeneous data.

The reasons behind this issue may be two-fold: (i) The useful information is dropped due to the clipping operation especially on small C, which is contained in the local updates; (ii) The model inconsistency among local models is exacerbated as the addition of random noise severely damages local updates and leads to large variances between local models especially on large C [9]. To overcome these issues, existing works improve the performance via restricting the norm of local update [9] and leveraging local update sparsification technique [9, 22] to reduce the adverse impacts of clipping and adding amount of random noise. However, the model performance degradation remains significantly severe compared with FL methods without considering the privacy, e.g., FedAvg [33].

Motivation. To further explore the reasons behind this phenomenon, we compare the structure of loss landscapes and surface contours [30] for FedAvg [33] and DP-FedAvg [15,35] on partitioned CIFAR-10 dataset [28] with Dirichlet distribution ($\alpha = 0.6$) and ResNet-18 backbone [18] in Figure 1 (a) and (b), respectively. Note that the convergence of DP-FedAvg is worse than FedAvg as its loss value is higher after a long communication round. Furthermore, FedAvg has a flatter landscape, whereas DP-FedAvg has a sharper one, resulting in poorer generalization ability (sharper minima, see Figure 1 (a)) and weight perturbation robustness

*Corresponding authors: Li Shen and Xueqian Wang



(a) Loss landscapes. (b) Loss surface contours.

Figure 1. Loss landscapes and surface contours comparison between DP-FedAvg (left) and FedAvg (right), respectively.

(see Figure 1 (b)), which is caused by the clipped local update information and exacerbated model inconsistency, respectively. Based on these observations, an interesting research question is: *can we further overcome the severe performance degradation via making landscape flatter?*

To answer this question, we propose DP-FedSAM via gradient perturbation to improve model performance. Specifically, a local flat model is generated by using SAM optimizer [13] in each client, which leads to better stability. After that, a potentially global flat model can be generated by aggregating several flat local models, which results in higher generalization ability and better robustness to DP noise, thereby significantly improving the performance and achieving a better trade-off between performance and privacy. Theoretically, we present a tighter bound $\mathcal{O}(\frac{1}{\sqrt{KT}} + \frac{\sum_{t=1}^{T}(\overline{\alpha}^t\sigma_g^2+\tilde{\alpha}^t L^2)}{T^2} + \frac{L^2\sqrt{T}\sigma^2 C^2 d}{m^2\sqrt{K}})$ for DP-FedSAM in the stochastic non-convex setting, where both $\frac{1}{T}\sum_{t=1}^{T}\overline{\alpha}^t$ and $\frac{1}{T}\sum_{t=1}^{T}\tilde{\alpha}^t$ are bounded constant, K and T is local iteration steps and communication rounds, respectively. Meanwhile, we present how SAM mitigates the impact of DP. For the clipping operation, DP-FedSAM reduces the l_2 norm and the negative impact of the inconsistency among local updates on convergence. For adding noise operation, we obtain better weight perturbation robustness for reducing the performance damage caused by random noise, thereby having better robustness to DP noise. Empirically, we conduct extensive experiments on EMNIST, CIFAR-10, and CIFAR-100 datasets in both the independently and identically distributed (IID) and Non-IID settings. Furthermore, we observe the loss landscape, surface contour, and the norm distribution of local updates for exploring the intrinsic effects of SAM with DP, which together with the theoretical analysis confirms the effect of DP-FedSAM.

Contribution. The main contributions of our work are summarized as four-fold: **(i)** We propose a novel scheme DP-FedSAM in DPFL to alleviate performance degradation issue from the optimizer perspective. **(ii)** We establish an improved convergence rate, which is tighter than the conventional bounds [9, 22] in the stochastic non-convex setting. Moreover, we provide rigorous privacy guarantees and sensitivity analysis. **(iii)** We are the first to in-depth analyze the roles of the on-average norm of local updates $\overline{\alpha}^t$ and local update consistency among clients $\tilde{\alpha}^t$ on convergence. **(iv)** We conduct extensive experiments to verify the effect of DP-FedSAM, which could achieve state-of-the-art (SOTA) performance compared with several strong DPFL baselines.

2. Related work

Client-level DPFL. Client-level DPFL is the de-facto approach for protecting any client's data. DP-FedAvg [36] is the first to study in this setting, which trains a language prediction model in a mobile keyboard and ensures client-level DP guarantee by employing the Gaussian mechanism and composing privacy guarantees. After that, the work in [26,51] presents a comprehensive end-to-end system, which appropriately discretizes the data and adds discrete Gaussian noise before performing secure aggregation. Meanwhile, AE-DPFL [58] leverages the voting-based mechanism among the data labels instead of averaging the gradients to avoid dimension dependence and significantly reduce the communication cost. Fed-SMP [22] uses Sparsified Model Perturbation (SMP) to mitigate the impact of privacy protection on model accuracy. Different from the aforementioned methods, a recent study [9] revisits this issue and leverages Bounded Local Update Regularization (BLUR) and Local Update Sparsification (LUS) to restrict the norm of local updates and reduce the noise size before executing operations that guarantee DP, respectively. Nevertheless, severe performance degradation remains.

Sharpness Aware Minimization (SAM). SAM [13] is an effective optimizer for training deep learning (DL) models, which leverages the flatness geometry of the loss landscape to improve model generalization ability. Recently, [4] study the properties of SAM and provide convergence results of SAM for non-convex objectives. As a powerful optimizer, SAM and its variants have been applied to various machine learning (ML) tasks [2, 11, 24, 29, 32, 38, 43, 46, 50, 56, 57]. Specifically, [41], [50] and [7] integrate SAM to improve the generalization, and thus mitigate the distribution shift problem and achieve a new SOTA performance for FL. However, to the best of our knowledge, no efforts have been devoted to the empirical performance and theoretical analysis of SAM in DPFL.

The most related works to this paper are DP-FedAvg in [35], Fed-SMP in [22], and DP-FedAvg with LUS and BLUR [9]. However, these works still suffer from inferior performance due to the exacerbated model inconsistency is-

sue among the clients caused by random noise. Therefore, different from existing works, we try to alleviate this issue by making the landscape flatter and weight perturbation ability more robust. Furthermore, another related work is FedSAM [41], which integrates SAM optimizer to enhance the flatness of the local model and achieves new SOTA performance for FL. On top of the aforementioned studies, we are the first to extend the SAM optimizer into the DPFL setting to effectively alleviate the performance degradation problem. Meanwhile, we provide the theoretical analysis for sensitivity, privacy, and convergence in the non-convex setting. Finally, we empirically confirm our theoretical results and the superiority of performance compared with existing SOTA methods in DPFL.

3. Preliminary

In this section, we first give the problem setup of FL, and then introduce the several terminologies in DP.

3.1. Federated Learning

Consider a general FL system consisting of M clients, in which each client owns its local dataset. Let $\mathcal{D}_i^{\text{train}}$, $\mathcal{D}_i^{\text{val}}$ and $\mathcal{D}_i^{\text{test}}$ denote the training dataset, validation dataset and testing dataset, held by client i, respectively, where $i \in \mathcal{U} = \{1, 2, \ldots, M\}$. Formally, the FL task is expressed as:

$$\mathbf{w}^\star = \arg\min_{\mathbf{w}} \sum_{i \in \mathcal{U}} p_i f_i(\mathbf{w}, \mathcal{D}_i^{\text{train}}), \tag{1}$$

where $p_i = |\mathcal{D}_i^{\text{train}}|/|\mathcal{D}^{\text{train}}| \geq 0$ with $\sum_{i \in \mathcal{U}} p_i = 1$, and $f_i(\cdot)$ is the local loss function with $f_i(\mathbf{w}) = F_i(\mathbf{w}; \xi_i)$, ξ_i is a batch sample data in client i. $|\mathcal{D}_i^{\text{train}}|$ is the size of training dataset $\mathcal{D}_i^{\text{train}}$ and $|\mathcal{D}^{\text{train}}| = \sum_{i \in \mathcal{U}} |\mathcal{D}_i^{\text{train}}|$ is the total size of training datasets, respectively. For the i-th client, a local model is learned on its private training data $\mathcal{D}^{\text{train}}$ via:

$$\mathbf{w}_i = \mathbf{w}_i^t - \eta \nabla f_i(\mathbf{w}_i, \mathcal{D}_i^{\text{train}}). \tag{2}$$

Generally, the local loss function $f_i(\cdot)$ has the same expression across each client. Then, the M associated clients collaboratively learn a global model $\mathbf{w}$ over the heterogeneous training data $\mathcal{D}_i^{\text{train}}$, $\forall i \in \mathcal{U}$.

3.2. Differential Privacy

Differential Privacy (DP) [12] is a rigorous privacy notion for measuring privacy risk. In this paper, we consider a relaxed version: Rényi DP (RDP) [39].

Definition 1. (Rényi DP, [39]). Given a real number $\alpha \in (1, \infty)$ and privacy parameter $\rho \geq 0$, a randomized mechanism $\mathcal{M}$ satisfies (α, ρ)-RDP if for any two neighboring datasets $\mathcal{D}$, $\mathcal{D}'$ that differ in a single record, the Rényi α-divergence between $\mathcal{M}(\mathcal{D})$ and $\mathcal{M}(\mathcal{D}')$ satisfies:

$$D_\alpha \left[\mathcal{M}(\mathcal{D}) \| \mathcal{M}\left(\mathcal{D}'\right)\right] := \frac{1}{\alpha - 1} \log \mathbb{E}\left[\left(\frac{\mathcal{M}(\mathcal{D})}{\mathcal{M}\left(\mathcal{D}'\right)}\right)^\alpha\right] \leq \rho, \tag{3}$$

where the expectation is taken over the output of $\mathcal{M}(\mathcal{D}')$.

Rényi DP is a useful analytical tool to measure the privacy and accurately represent guarantees on the tails of the privacy loss, which is strictly stronger than (ϵ, δ)-DP for $\delta > 0$. Thus, we provide the privacy analysis based on this tool for each user's privacy loss.

Definition 2. (l_2 Sensitivity). [9, Definition 2] Let $\mathcal{F}$ be a function, the L_2-sensitivity of $\mathcal{F}$ is defined as $\mathcal{S} = \max_{D \simeq D'} \|\mathcal{F}(D) - \mathcal{F}(D')\|_2$, where the maximization is taken over all pairs of adjacent datasets.

The sensitivity of a function $\mathcal{F}$ captures the magnitude which a single individual's data can change the function $\mathcal{F}$ in the worst case. Therefore, it plays a crucial role in determining the magnitude of noise required to ensure DP.

Definition 3. (Client-level DP) [34, Definition 1]. A randomized algorithm $\mathcal{M}$ is (ϵ, δ)-DP if for any two adjacent datasets U, U' constructed by adding or removing all records of any client, and every possible subset of outputs O satisfy the following inequality:

$$\Pr[\mathcal{M}(U) \in O] \leq e^\epsilon \Pr\left[\mathcal{M}\left(U'\right) \in O\right] + \delta. \tag{4}$$

In client-level DP, we aim to ensure participation information for any clients. Therefore, we need to make local updates similar whether one client participates or not.

4. Methodology

To revisit the severe performance degradation challenge in DPFL, we observe the loss landscapes and surface contours of FedAvg and DP-FedAvg in Figure 1 (a) and (b), respectively. And we find that the DPFL method has a sharper landscape with both poorer generalization ability and weight perturbation robustness than the FL method. It means that the DPFL method may result in poor flatness and make model sensitivity to noise perturbation. However, the study focusing on this issue remains unexplored. Therefore, we plan to face this challenge from the optimizer perspective by adopting a SAM optimizer in each client, dubbed DP-FedSAM, whose local loss function is defined as:

$$f_i(\mathbf{w}) = \mathbb{E}_{\xi \sim \mathcal{D}_i} \max_{\|\delta_i\|_2 \leq \rho} F_i(\mathbf{w}^{t,k}(i) + \delta_i; \xi_i), \quad i \in \mathcal{N}, \tag{5}$$

where $\mathbf{w}^{t,k}(i) + \delta_i$ is viewed as the perturbed model, ρ is a predefined constant controlling the radius of the perturbation and $\|\cdot\|_2$ is a l_2-norm. Instead of searching for a solution via SGD [5, 6], SAM [13] aims to seek a solution in a flat region by adding a small perturbation, i.e., $w + \delta$ with more robust performance. Specifically, for each client $i \in \{1, 2, ..., M\}$, each local iteration $k \in \{0, 1, ..., K-1\}$ in each communication round $t \in \{0, 1, ..., T-1\}$, the k-th

inner iteration in client i is performed as:

$$\mathbf{w}^{t,k+1}(i) = \mathbf{w}^{t,k}(i) - \eta\tilde{\mathbf{g}}^{t,k}(i), \tag{6}$$

$$\tilde{\mathbf{g}}^{t,k}(i) = \nabla F_i(\mathbf{w}^{t,k} + \delta(\mathbf{w}^{t,k}); \xi), \delta(\mathbf{w}^{t,k}) = \frac{\rho \mathbf{g}^{t,k}}{\|\mathbf{g}^{t,k}\|_2}. \tag{7}$$

Where $\delta(\mathbf{w}^{t,k})$ is calculated by using first-order Taylor expansion around $\mathbf{w}^{t,k}$ [13]. After that, we adopt a sampling mechanism, gradient clipping, and Gaussian noise adding to ensure client-level DP. Note that this sampling can amplify the privacy guarantee since it decreases the chances of leaking information about a particular individual [1]. Specifically, after sampling m clients with probability $q = m/M$ at each communication round, which is important for measuring privacy loss. Then, we clip the local updates first:

$$\tilde{\Delta}_i^t = \Delta_i^t \cdot \min\left(1, \frac{C}{\|\Delta_i^t\|_2}\right). \tag{8}$$

After clipping, we add Gaussian noise to the local update for ensuring client-level DP as follows:

$$\hat{\Delta}_i^t = \tilde{\Delta}_i^t + \mathcal{N}(0, \sigma^2 C^2 \cdot \mathbf{I}_d/m). \tag{9}$$

With a given noise variance σ^2, the accumulative privacy budget ϵ can be calculated based on the sampled Gaussian mechanism [54]. Finally, we summarize the training procedures of DP-FedSAM in Algorithm 1.

Algorithm 1: DP-FedSAM

Input : Total number of clients M, sampling ratio of clients q, total number of communication rounds T, the clipping threshold C, local learning rate η, and total number of the local iterates are K.
Output: Generate global model $\mathbf{w}^T$.
Initialization: Randomly initialize the global model $\mathbf{w}^0$.
for $t = 0$ **to** $T-1$ **do**
  Sample a set of $m = qM$ clients at random without replacement, denoted by $\mathcal{W}^t$.
  for *client* $i = 1$ **to** m **in parallel do**
    for $k = 0$ **to** $K-1$ **do**
      Update the global parameter as the local parameter $\mathbf{w}^t(i) \leftarrow \mathbf{w}^t$.
      Sample a batch of local data ξ_i and calculate local gradient $\mathbf{g}^{t,k}(i) = \nabla F_i(\mathbf{w}^{t,k}(i); \xi_i)$.
      Gradient perturbation by Equation (7).
      Local iteration update by Equation (6).
    end
    $\Delta_i^t = \mathbf{w}^{t,K}(i) - \mathbf{w}^{t,0}(i)$.
    Clip and add noise for DP by Equation (9).
    Return $\hat{\Delta}^t(i)$.
  end
  $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t + \frac{1}{m}\sum_{i\in\mathcal{W}^t} \hat{\Delta}^t(i)$.
end

Compared with existing DPFL methods [9, 22, 35], the benefits of DP-FedSAM lie in three-fold: (i) We introduce SAM into DPFL to alleviate severe performance degradation via seeking a flatness model in each client, which is caused by the exacerbated inconsistency of local models. Specifically, DP-FedSAM generates both better generalization ability and robustness to DP noise by making the global model flatter. (ii) In addition, we analyze in detail how DP-FedSAM mitigates the negative impacts of DP. Specifically, we theoretically analyze the convergence with the on-average norm of local updates $\overline{\alpha}^t$ and local update consistency among clients $\tilde{\alpha}^t$, and empirically confirm these results via observing the norm distribution and average norm of local updates (see Section 6.3). (iii) Meanwhile, we present the theories unifying the impacts of gradient perturbation ρ in SAM, the on-average norm of local updates $\overline{\alpha}^t$, and local update consistency among clients $\tilde{\alpha}^t$ in clipping operation, and the variance of random noise $\sigma^2 C^2/m$ upon the convergence rate in Section 5.

5. Theoretical Analysis

In this section, we give a rigorous analysis of DP-FedSAM, including its sensitivity, privacy, and convergence rate. The detailed proof is placed in **Appendix** E. Below, we first give several necessary assumptions.

Assumption 1. (Lipschitz smoothness). The function F_i is differentiable and ∇F_i is L-Lipschitz continuous, $\forall i \in \{1, 2, \ldots, M\}$, i.e., $\|\nabla F_i(\mathbf{x}) - \nabla F_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

Assumption 2. (Bounded variance). The gradient of the function f_i have σ_l-bounded variance, i.e., $\mathbb{E}_{\xi_i}\left\|\nabla F_i(\mathbf{w}^k(i); \xi_i) - \nabla F_i(\mathbf{w}(i))\right\|^2 \leq \sigma_l^2$, $\forall i \in \{1, 2, \ldots, M\}, k \in \{1, ..., K-1\}$, the global variance is also bounded, i.e., $\frac{1}{M}\sum_{i=1}^M \|\nabla f_i(\mathbf{w}) - \nabla f(\mathbf{w})\|^2 \leq \sigma_g^2$ for all $\mathbf{w} \in \mathbb{R}^d$. It is not hard to verify that the σ_g is smaller than the homogeneity parameter β, i.e., $\sigma_g^2 \leq \beta^2$.

Assumption 3. (Bounded gradient). For any $i \in \{1, 2, \ldots, M\}$ and $\mathbf{w} \in \mathbb{R}^d$, we have $\|\nabla f_i(\mathbf{w})\| \leq B$.

Assumption 4. (Unbiased Gradient Estimator). For any data sample z from $\mathcal{D}_i$ and $\mathbf{w} \in \mathbb{R}^d$, the local gradient estimator is unbiased, i.e., $\mathbb{E}[\nabla f_i(\mathbf{w}; z)] = \mathbb{E}[\nabla f_i(\mathbf{w})]$.

Note that the above assumptions are mild and commonly used in characterizing the convergence rate of FL [6, 9, 17, 22, 23, 41–43, 49, 53]. Furthermore, gradient clipping operation in DL is often used to prevent the gradient explosion phenomenon, thereby the gradient is bounded. The technical difficulty for DP-FedSAM lies in: (i) how SAM mitigates the impact of DP; (ii) how to analyze in detail the impacts of the consistency among clients and the on-average norm of local updates caused by clipping operation.

5.1. Sensitivity Analysis

At first, we study the sensitivity of local update Δ_i^t from any client $i \in \{1, 2, ..., M\}$ before clipping at t-th communication round in DP-FedSAM. This upper bound of sensitivity can roughly measure the degree of privacy protection. Under Definition 2, the sensitivity can be denoted by $\mathcal{S}_{\Delta_i^t}$ in client i at t-th communication round.

Theorem 1. *(Sensitivity). Denote $\Delta_i^t(\mathbf{x})$ and $\Delta_i^t(\mathbf{y})$ as the local update at t-th communication round, the model $\mathbf{x}(i)$ and $\mathbf{y}(i)$ is conducted on two sets which differ at only one sample. Assume the initial model parameter $\mathbf{w}^t(i) = \mathbf{x}^{t,0}(i) = \mathbf{y}^{t,0}(i)$. With the above assumptions, the expected squared sensitivity $\mathcal{S}^2_{\Delta_i^t}$ of local update is upper bounded,*

$$\mathbb{E}\mathcal{S}^2_{\Delta_i^t} \leq \frac{6\eta^2\rho^2 K L^2(12K^2L^2\eta^2 + 10)}{1 - 2\eta^2 L^2 K} \tag{10}$$

When the local adaptive learning rate satisfies $\eta = \mathcal{O}(1/L\sqrt{KT})$ and the perturbation amplitude ρ proportional to the learning rate, e.g., $\rho = \mathcal{O}(\frac{1}{\sqrt{T}})$, we have

$$\mathbb{E}\mathcal{S}^2_{\Delta_i^t} \leq \mathcal{O}\left(\frac{1}{T^2}\right). \tag{11}$$

For comparison, we also present the expected squared sensitivity of local update with SGD in DPFL, that is $\mathbb{E}\mathcal{S}^2_{\Delta_i^t,SGD} \leq \frac{6\eta^2\sigma_l^2 K}{1-3\eta^2KL^2}$. Thus $\mathbb{E}\mathcal{S}^2_{\Delta_i^t,SGD} \leq \mathcal{O}(\frac{\sigma_l^2}{KL^2T})$ when $\eta = \mathcal{O}(1/L\sqrt{KT})$.

Remark 1. It is clearly seen that the upper bound in $\mathbb{E}\mathcal{S}^2_{\Delta_i^t,SAM}$ is tighter than that in $\mathbb{E}\mathcal{S}^2_{\Delta_i^t,SGD}$. From the perspective of privacy protection, it means DP-FedSAM has a better privacy guarantee than DP-FedAvg. Another perspective from local iteration, which means both better model consistency among clients and training stability.

5.2. Privacy Analysis

To achieve client-level privacy protection, we derive the sensitivity of the aggregation process at first after clipping the local updates.

Lemma 1. *The sensitivity of client-level DP in DP-FedSAM can be expressed as C/m.*

Proof. Given two adjacent batches $\mathcal{W}^t$ and $\mathcal{W}^{t,\text{adj}}$, that $\mathcal{W}^{t,\text{adj}}$ has one more or less client, we have

$$\left\| \frac{1}{m} \sum_{i \in \mathcal{W}^t} \Delta_i^t - \frac{1}{m} \sum_{j \in \mathcal{W}^{t,\text{adj}}} \Delta_j^t \right\|_2 = \frac{1}{m} \left\| \Delta_{j'}^t \right\|_2 \leq \frac{C}{m}, \tag{12}$$

where $\Delta_{j'}^t$ is the local update of one more or less client. □

Remark 2. The value of this sensitivity can determine the amount of variance for adding random noise.

After adding Gaussian noise, we calculate the accumulative privacy budget [54] along with training as follows.

Theorem 2. *After T communication rounds, the accumulative privacy budget is calculated by:*

$$\epsilon = \bar{\epsilon} + \frac{(\alpha - 1)\log(1 - \frac{1}{\alpha}) - \log(\alpha) - \log(\delta)}{\alpha - 1}, \tag{13}$$

where

$$\bar{\epsilon} = \frac{T}{\alpha - 1} \ln \mathbb{E}_{z \sim \mu_0(z)} \left[\left(1 - q + \frac{q\mu_1(z)}{\mu_0(z)} \right)^{\alpha} \right], \tag{14}$$

and q is sample rate for client selection, $\mu_0(z)$ and $\mu_1(z)$ denote the Gaussian probability density function (PDF) of $\mathcal{N}(0, \sigma)$ and the mixture of two Gaussian distributions $q\mathcal{N}(1, \sigma) + (1-q)\mathcal{N}(0, \sigma)$, respectively, σ is the noise STD, α is a selectable variable.

Remark 3. It can be seen that a small sampling rate q can enhance the privacy guarantee by decreasing the privacy budget, but it may also degrade the training performance due to the number of participating clients being reduced in each communication round. So a better trade-off is needed.

5.3. Convergence Analysis

Below, we give a convergence analysis of how DP-FedSAM mitigates the negative impacts of DP. The technical contribution also is combining the impacts of the on-average norm of local updates $\overline{\alpha}^t$ and local update consistency among clients $\tilde{\alpha}^t$ on the rate. Moreover, we also empirically confirm these results in Section 6.3.

Theorem 3. *Under assumptions 1-4, local learning rate satisfies $\eta = \mathcal{O}(1/L\sqrt{KT})$ and f^* is denoted as the minimal value of f, i.e., $f(x) \geq f(x^*) = f^*$ for all $x \in \mathbb{R}^d$. When the perturbation amplitude ρ is proportional to the learning rate, e.g., $\rho = \mathcal{O}(1/\sqrt{T})$, the sequence of outputs $\{\mathbf{w}^t\}$ generated by Alg. 1, we have:*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\overline{\alpha}^t \left\|\nabla f\left(\mathbf{w}^t\right)\right\|^2\right] \leq \underbrace{\mathcal{O}\left(\frac{2L(f(\mathbf{w}^1) - f^*)}{\sqrt{KT}} + \frac{L^2\sigma_l^2}{KT^2}\right)}_{\text{From FedSAM}} + \underbrace{\underbrace{\mathcal{O}\left(\frac{\sum_{t=1}^{T}(\overline{\alpha}^t\sigma_g^2 + \tilde{\alpha}^t L^2)}{T^2}\right)}_{\text{Clipping}} + \underbrace{\mathcal{O}\left(\frac{L^2\sqrt{T}\sigma^2 C^2 d}{m^2\sqrt{K}}\right)}_{\text{Adding noise}}}_{\text{From operations for DP}}$$

where

$$\overline{\alpha}^t := \frac{1}{M}\sum_{i=1}^{M} \alpha_i^t \quad \textit{and} \quad \tilde{\alpha}^t := \frac{1}{M}\sum_{i=1}^{M} |\alpha_i^t - \overline{\alpha_i^t}|, \tag{15}$$

with $\alpha_i^t = \min(1, \frac{C}{\eta \| \sum_{k=0}^{K-1} \tilde{\mathbf{g}}^{t,k}(i) \|})$, respectively. Note that $\overline{\alpha}^t$ and $\tilde{\alpha}^t$ measure the on-average norm of local updates and local update consistency among clients before clipping and adding noise operations in DP-FedSAM, respectively.

Table 1. Averaged training accuracy (%) and testing accuracy (%) on two data in both IID and Non-IID settings for all compared methods.

Task	Algorithm	Dirichlet 0.3		Dirichlet 0.6		IID	
		Train	Validation	Train	Validation	Train	Validation
EMNIST	DP-FedAvg	99.28±0.02	73.10±0.16	99.55±0.02	82.20±0.35	99.66±0.40	81.90±0.86
	Fed-SMP-rand_k	99.24±0.02	73.72±0.53	99.71±0.01	82.18±0.73	99.71±0.61	84.16±0.83
	Fed-SMP-top_k	99.31±0.04	75.75±0.35	99.72±0.02	83.41±0.91	99.73±0.40	83.32±0.52
	DP-FedAvg-blur	99.12±0.02	73.71±0.02	99.66±0.00	83.20±0.01	99.67±0.03	82.92±0.49
	DP-FedAvg-blurs	99.63±0.08	76.25±0.35	99.72±0.02	83.41±0.91	99.74±0.45	82.92±0.49
	DP-FedSAM	96.28±0.64	76.81±0.81	95.07±0.45	84.32±0.19	95.61±0.94	85.90±0.72
	DP-FedSAM-top_k	94.77±0.11	**77.27±0.67**	95.87±1.52	**84.80±0.60**	96.12±0.85	**87.70±0.83**
CIFAR-10	DP-FedAvg	93.65±0.47	47.98±0.24	93.65±0.42	50.05±0.47	93.65±0.15	50.90±0.86
	Fed-SMP-rand_k	95.46±0.43	48.14±0.12	95.36±0.06	51.33±0.36	95.36±0.06	50.61±0.20
	Fed-SMP-top_k	95.49±0.14	49.93±2.29	95.49±0.09	54.11±0.83	95.49±0.10	53.30±0.45
	DP-FedAvg-blur	95.47±0.12	47.66±0.01	99.66±0.42	51.05±0.01	94.50±0.05	52.56±0.47
	DP-FedAvg-blurs	96.79±0.51	51.23±0.66	99.72±0.09	54.11±0.83	96.45±0.30	53.48±0.76
	DP-FedSAM	90.38±0.90	53.92±0.55	90.83±0.15	54.14±0.60	90.83±0.16	55.58±0.50
	DP-FedSAM-top_k	93.25±0.60	**54.85±0.86**	92.60±0.65	**57.00±0.69**	91.52±0.11	**58.82±0.51**

Remark 4. The proposed DP-FedSAM can achieve a tighter bound in general non-convex setting compared with previous work [9, 22] ($\mathcal{O}\left(\frac{1}{\sqrt{KT}}+\frac{6K\sigma_g^2+\sigma_l^2}{T}+\frac{B^2\sum_{t=1}^{T}(\overline{\alpha}^t+\tilde{\alpha}^t)}{T}+\frac{L^2\sqrt{T}\sigma^2C^2d}{m^2\sqrt{K}}\right)$ in [9] and $\mathcal{O}\left(\frac{1}{\sqrt{KT}}+\frac{3\sigma_g^2+2\sigma_l^2}{\sqrt{KT}}+\frac{4L^2\sqrt{T}\sigma^2C^2d}{m^2\sqrt{K}}\right)$ in [22]) and reduce the impacts of the local and global variance σ_l^2, σ_g^2. Meanwhile, we are the first to theoretically analyze these two impacts of both the on-average norm of local updates $\overline{\alpha}^t$ and local update inconsistency among clients $\tilde{\alpha}^t$ on convergence. And the negative impacts of $\overline{\alpha}^t$ and $\tilde{\alpha}^t$ are also significantly mitigated upon convergence compared with previous work [9] due to local SAM optimizer being adopted. It means that we effectively alleviate performance degradation caused by clipping operation in DP and achieve better performance under symmetric noise. This theoretical result has also been empirically verified on several real-world data (see Section 6.2 and 6.3).

6. Experiments

In this section, we conduct extensive experiments to verify the effectiveness of DP-FedSAM.

6.1. Experiment Setup

Dataset and Data Partition. The efficacy of DP-FedSAM is evaluated on three datasets, including **EMNIST** [10], **CIFAR-10** and **CIFAR-100** [28], in both IID and Non-IID settings. Specifically, Dir Partition [19] is used for simulating Non-IID across federated clients, where the local data of each client is partitioned by splitting the total dataset through sampling the label ratios from the Dirichlet distribution Dir(α) with parameters $\alpha = 0.3$ and $\alpha = 0.6$.

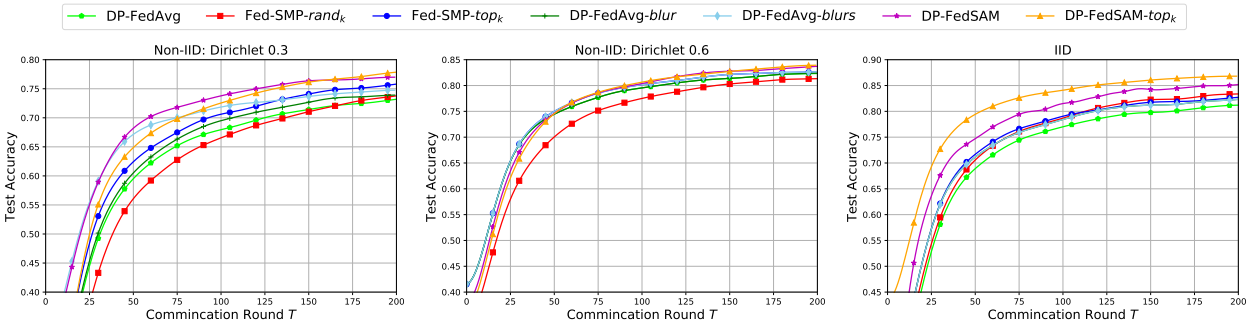
Baselines. We focus on DPFL methods that ensure client-level DP. Thus we compare DP-FedSAM with existing DPFL baselines: **DP-FedAvg** [35] ensures client-level DP guarantee by directly employing Gaussian mechanism to the local updates. **DP-FedAvg-blur** [9] adds regularization method (BLUR) based on DP-FedAvg. **DP-FedAvg-blurs** [9] uses local update sparsification (LUS) and BLUR for improving the performance of DP-FedAvg. **Fed-SMP-**rand_k and **Fed-SMP-**top_k [22] leverage random sparsification and top_k sparsification technique for reducing the impact of DP noise on model accuracy, respectively.

Configuration. For EMNIST, we use a simple CNN model and train 200 communication round. For CIFAR-10 and CIFAR-100 datasets, we use the ResNet-18 [18] backbone and train 300 communication round. For all experiments, we set the number of clients M to 500. The default sample ratio q of the client is 0.1. The local learning rate η is set to 0.1 with a decay rate 0.005 and momentum 0.5, and the number of training epochs is 30. For privacy parameters, noise multiplier σ is set to 0.95 and the privacy failure probability $\delta = \frac{1}{M}$. The clipping threshold C is selected by grid search from set $\{0.1, 0.2, 0.4, 0.6, 0.8\}$, and we find that the gradient explosion phenomenon will occur when $C \geq 0.6$ on EMNIST and $C = 0.2$ performs better on three datasets. The weight perturbation ratio is set to $\rho = 0.5$. We run each experiment 3 trials and report the best-averaged testing accuracy in each experiment.

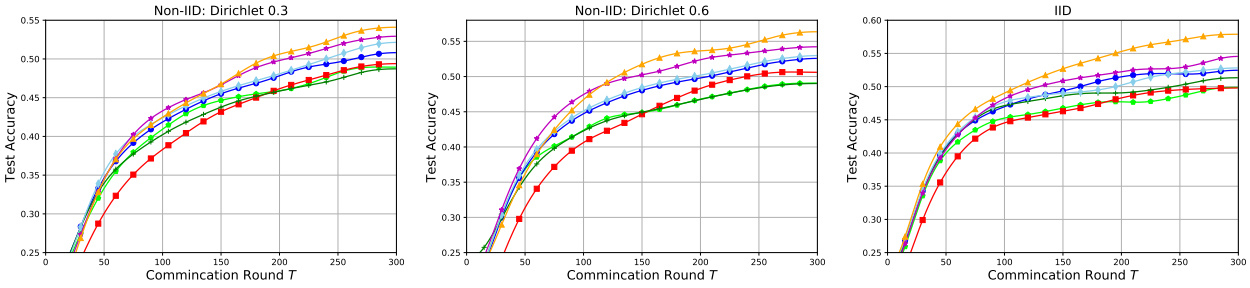
The results on EMNIST and CIFAR-10 datasets are placed in the main paper, while the results on CIFAR-100 are placed in **Appendix** A. For a more comprehensive and fair comparison, we integrate DP-FedSAM with top_k sparsification technique [9, 22], named DP-FedSAM-top_k (More discussion in **Appendix** D), to compare with baselines that also use local sparsification technique, such as Fed-SMP-rand_k, Fed-SMP-top_k, and DP-FedAvg-blurs.

6.2. Experiment Evaluation

Performance with compared baselines. In Table 1 and Figure 2, we evaluate DP-FedSAM and DP-FedSAM-top_k on EMNIST and CIFAR-10 datasets in both settings compared with all baselines from DP-FedAvg to DP-FedAvg-blurs. The baseline methods seem to be overfitting in Table 1, especially on more complex data (e.g., CIFAR-100 in Table 4). The main reasons are the random noise introduced by DP and the over-fitting and inconsistency of



(a) EMNIST



(b) CIFAR-10

Figure 2. The averaged testing accuracy on *EMNIST* and *CIFAR-10* dataset under symmetric noise for all compared methods.

the local models. From all these results, it is seen that our proposed algorithms outperform other baselines under symmetric noise both on accuracy and generalization perspectives. It means that we significantly improve the performance and generate a better trade-off between performance and privacy in DPFL. For instance, the averaged testing accuracy is 85.90% in DP-FedSAM and 87.70% in DP-FedSAM-top$_k$ on EMNIST in the IID setting, which are better than other baselines. Meanwhile, the difference between training accuracy and test accuracy is 9.71% in DP-FedSAM and 8.40% in DP-FedSAM-top$_k$, while that is 17.74% in DP-FedAvg and 16.41% in Fed-SMP-top$_k$, respectively. That means our algorithms significantly mitigate the performance degradation issue caused by DP.

Impact of Non-IID levels. In the experiments under different participation cases as shown in Table 1, we further prove the robust generalization of the proposed algorithms. Heterogeneous data distribution of local clients is set to various participation levels from IID, Dirichlet 0.6, and Dirichlet 0.3, which makes the training of the global model more difficult. On EMNIST, as the Non-IID level decreases, DP-FedSAM achieves better generalization than DP-FedAvg, and the difference between training accuracy and test accuracy in DP-FedSAM $\{19.47\%, 10.75\%, 9.71\%\}$ are lower than that in DP-FedAvg $\{26.18\%, 17.35\%, 17.74\%\}$. Similarly, the difference values in DP-FedSAM-top$_k$ $\{17.50\%, 11.07\%, 8.40\%\}$ are also lower than that in Fed-SMP-top$_k$ $\{23.56\%, 16.31\%, 16.41\%\}$. These observations confirm that our algorithms are more robust than baselines in various degrees of heterogeneous data.

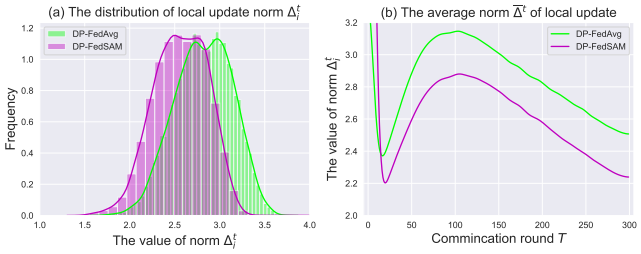


Figure 3. Norm distribution and average norm of local updates.

6.3. Discussion for DP with SAM in FL

In this subsection, we empirically discuss how SAM mitigates the negative impacts of DP from the aspects of the norm of local update and the visualization of loss landscape and contour. Meanwhile, we also observe the training performance with SAM under different privacy budgets ϵ compared with all baselines. These experiments are conducted on CIFAR-10 with ResNet-18 [18] and Dirichlet $\alpha = 0.6$.

The norm of local update. To validate the theoretical results for mitigating the adverse impacts of the norm of local updates, we conduct experiments on DP-FedSAM and DP-FedAvg with clipping threshold $C = 0.2$ as shown in Figure 3. We show the norm Δ_i^t distribution and average norm $\overline{\Delta}^t$ of local updates before clipping during the communication rounds. In contrast to DP-FedAvg, most of the norm is distributed at the smaller value in our scheme in Figure 3 (a), which means that the clipping operation drops less information. Meanwhile, the on-average norm $\overline{\Delta}^t$ is smaller than DP-FedAvg as shown in Figure 3 (b). These observations

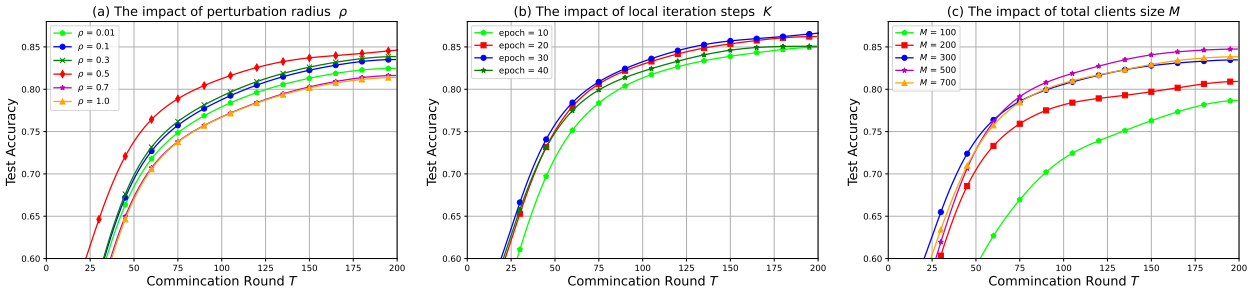


Figure 4. Impact of hyper-parameters: perturbation radius ρ, local iteration steps K, total clients size M.

Table 2. Performance comparison under different privacy budgets.

Algorithm	Averaged test accuracy (%) under different privacy budgets ϵ			
	$\epsilon = 4$	$\epsilon = 6$	$\epsilon = 8$	$\epsilon = 10$
DP-FedAvg	38.23 ± 0.15	43.87 ± 0.62	46.74 ± 0.03	49.06 ± 0.49
Fed-SMP-rand$_k$	33.78 ± 0.92	42.21 ± 0.21	48.20 ± 0.05	50.62 ± 0.14
Fed-SMP-top$_k$	38.99 ± 0.50	46.24 ± 0.80	49.78 ± 0.78	52.51 ± 0.83
DP-FedAvg-blur	38.23 ± 0.70	43.93 ± 0.48	46.74 ± 0.92	49.06 ± 0.13
DP-FedAvg-blurs	39.39 ± 0.43	46.64 ± 0.36	50.18 ± 0.27	52.91 ± 0.57
DP-FedSAM	$\mathbf{39.89 \pm 0.17}$	47.92 ± 0.23	51.30 ± 0.95	53.18 ± 0.40
DP-FedSAM-top$_k$	38.96 ± 0.61	$\mathbf{49.17 \pm 0.15}$	$\mathbf{53.64 \pm 0.12}$	$\mathbf{56.36 \pm 0.36}$

are also consistent with our theoretical results in Section 5.

Performance under different privacy budgets ϵ. Table 2 shows the test accuracies for the different level privacy guarantees. Note that ϵ is not a hyper-parameter, but can be obtained by the privacy design in each round such as Eq. (13). Our methods consistently outperform the previous SOTA methods under various privacy budgets ϵ. Specifically, DP-FedSAM and DP-FedSAM-top$_k$ significantly improve the accuracy of DP-FedAvg and Fed-SMP-top$_k$ by $1\% \sim 4\%$ and $3\% \sim 4\%$ under the same ϵ, respectively. Furthermore, the test accuracy can improve as the privacy budget ϵ increases, which suggests us a better balance is necessary between training performance and privacy. There is a better trade-off when achieving better accuracy under the same ϵ or a smaller ϵ under approximately the same accuracy.

Loss landscape and contour. The discussion and visualizing results are placed in **Appendix** C due to limited space.

6.4. Ablation Study

We verify the influence of hyper-parameters in DP-FedSAM on EMNIST with Dirichlet partition $\alpha = 0.6$.

Perturbation weight ρ. Perturbation weight ρ has an impact on performance as the adding perturbation is accumulated when the communication round T increases. To select a proper value for our algorithms, we conduct some experiments on various perturbation radius from the set $\{0.01, 0.1, 0.3, 0.5, 0.7, 1.0\}$ in Figure 4 (a). As $\rho = 0.5$, we achieve better convergence and performance.

Local iteration steps K. Large local iteration steps K can help the convergence in previous DPFL work [9] with the theoretical guarantees. To investigate the acceleration on T by adopting a larger K, we fix the total batchsize and change local training epochs. In Figure 4 (b), our algorithm can accelerate the convergence in Theorem 3 as a larger K is adopted, that is, use a larger epoch value. However, the adverse impact of clipping on training increases as K is too large, e.g., epoch = 40. Thus we choose 30 local epoch.

Table 3. The averaged training accuracy and testing accuracy.

Algorithm	Train (%)	Validation (%)	Differential value (%)
DP-FedAvg	99.55±0.02	82.20± 0.35	17.35±0.32
DP-FedSAM	95.07±0.45	84.32±0.19 ↑	10.75±0.26 ↓
Fed-SMP-top$_k$	99.72±0.02	83.41 ± 0.91	16.31± 0.89
DP-FedSAM-top$_k$	95.87±0.52	84.80±0.60 ↑	11.07±0.08 ↓

Client size M. We compare the performance between different numbers of client participation $m = \{100, 200, 300, 500, 700\}$ with the same hyper-parameters in Figure 4 (c). Compared with larger $m = 500$, the smaller m may have worse performance due to large variance $\sigma^2 C^2/m$ in DP noise with the same setting. Meanwhile, when m is too large such as $M = 700$, the performance may degrade as the local data size decreases.

Effect of SAM. As shown in Table 3, it is seen that DP-FedSAM and DP-FedSAM-top$_k$ can achieve performance improvement and better generalization compared with DP-FedAvg and Fed-SMP-top$_k$ as SAM optimizer is adopted with the same setting, respectively.

7. Conclusion

In this paper, we focus on severe performance degradation issue caused by dropped model information and exacerbated model inconsistency. And we are the first to alleviate this issue from the optimizer perspective and propose a novel and effective framework DP-FedSAM with a flatter loss landscape. Meanwhile, we present the analysis in detail of how SAM mitigates the adverse impacts of DP and achieve a tighter bound on convergence. Moreover, it is the first analysis to combine the impacts of the on-average norm of local updates and local update consistency among clients on training, and simultaneously provide the experimental observations. Finally, empirical results also verify the superiority of our approach on several real-world data.

Acknowledgement This work is supported by STI 2030—Major Projects (No. 2021ZD0201405), Shenzhen Philosophy and Social Science Foundation (Grant No. SZ2021B005).

References

[1] Martín Abadi, Andy Chu, Ian J. Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. *CoRR*, 2016.

[2] Momin Abbas, Quan Xiao, Lisha Chen, Pin-Yu Chen, and Tianyi Chen. Sharp-maml: Sharpness-aware model-agnostic meta learning. In *International Conference on Machine Learning, ICML*, pages 10–32, 2022.

[3] Naman Agarwal, Ananda Theertha Suresh, Felix X. Yu, Sanjiv Kumar, and Brendan McMahan. cpSGD: Communication-efficient and differentially-private distributed SGD. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 7575–7586, Dec. 2018.

[4] Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *International Conference on Machine Learning, ICML*, Proceedings of Machine Learning Research, pages 639–668. PMLR, 2022.

[5] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[6] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[7] Debora Caldarola, Barbara Caputo, and Marco Ciccone. Improving generalization in federated learning by seeking flat minima. *CoRR*, abs/2203.11834, 2022.

[8] Anda Cheng, Peisong Wang, Xi Sheryl Zhang, and Jian Cheng. Differentially private federated learning with local regularization and sparsification. *CoRR*, 2022.

[9] Anda Cheng, Peisong Wang, Xi Sheryl Zhang, and Jian Cheng. Differentially private federated learning with local regularization and sparsification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[10] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926. IEEE, 2017.

[11] Jiawei Du, Hanshu Yan, Jiashi Feng, Joey Tianyi Zhou, Liangli Zhen, Rick Siow Mong Goh, and Vincent Tan. Efficient sharpness-aware minimization for improved training of neural networks. In *International Conference on Learning Representations*, 2021.

[12] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, pages 211–407, Aug. 2014.

[13] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

[14] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1322–1333, 2015.

[15] Robin C. Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *CoRR*, Aug. 2017.

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

[17] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[19] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

[20] Rui Hu, Yanmin Gong, and Yuanxiong Guo. Federated learning with sparsification-amplified privacy and adaptive optimization. In *Proc. Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1463–1469, Aug. 2021.

[21] Rui Hu, Yanmin Gong, and Yuanxiong Guo. Federated learning with sparsified model perturbation: Improving accuracy under client-level differential privacy. *CoRR*, 2022.

[22] Rui Hu, Yanmin Gong, and Yuanxiong Guo. Federated learning with sparsified model perturbation: Improving accuracy under client-level differential privacy. *arXiv preprint arXiv:2202.07178*, 2022.

[23] Tiansheng Huang, Shiwei Liu, Li Shen, Fengxiang He, Weiwei Lin, and Dacheng Tao. Achieving personalized federated learning with sparse local models. *arXiv preprint arXiv:2201.11380*, 2022.

[24] Zhuo Huang, Xiaobo Xia, Li Shen, Jun Yu, Chen Gong, Bo Han, and Tongliang Liu. Robust generalization against corruptions via worst-case sharpness minimization.

[25] Peter Kairouz, Ziyu Liu, and Thomas Steinke. The distributed discrete gaussian mechanism for federated learning with secure aggregation. In *Proc. International Conference on Machine Learning (ICML)*, pages 5201–5212, Jul. 2021.

[26] P. Kairouz, Ziyu Liu, and T. Steinke. The distributed discrete gaussian mechanism for federated learning with secure aggregation. *ArXiv*, abs/2102.06387, 2021.

[27] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, pages 1–210, 2021.

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

[29] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pages 5905–5914. PMLR, 2021.

[30] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

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

[32] Yong Liu, Siqi Mai, Xiangning Chen, Cho-Jui Hsieh, and Yang You. Towards efficient and scalable sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12360–12370, 2022.

[33] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, pages 1273–1282, April. 2017.

[34] H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*, 2017.

[35] H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *Proc. International Conference on Learning Representations (ICLR)*, Apr. 2018.

[36] H. B. McMahan, D. Ramage, Kunal Talwar, and L. Zhang. Learning differentially private recurrent language models. In *ICLR*, 2018.

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

[38] Peng Mi, Li Shen, Tianhe Ren, Yiyi Zhou, Xiaoshuai Sun, Rongrong Ji, and Dacheng Tao. Make sharpness-aware minimization stronger: A sparsified perturbation approach. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[39] Ilya Mironov. Rényi differential privacy. In *Proc. IEEE computer security foundations symposium (CSF)*, pages 263–275, 2017.

[40] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *Proc. IEEE Symposium on Security and Privacy (SP)*, pages 739–753, 2019.

[41] Zhe Qu, Xingyu Li, Rui Duan, Yao Liu, Bo Tang, and Zhuo Lu. Generalized federated learning via sharpness aware minimization. In *International Conference on Machine Learning, ICML*, pages 18250–18280, 2022.

[42] Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

[43] Yifan Shi, Li Shen, Kang Wei, Yan Sun, Bo Yuan, Xueqian Wang, and Dacheng Tao. Improving the model consistency of decentralized federated learning. *arXiv preprint arXiv:2302.04083*, 2023.

[44] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *Proc. IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017.

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[46] Hao Sun, Li Shen, Qihuang Zhong, Liang Ding, Shixiang Chen, Jingwei Sun, Jing Li, Guangzhong Sun, and Dacheng Tao. Adasam: Boosting sharpness-aware minimization with adaptive learning rate and momentum for training deep neural networks. *arXiv preprint arXiv:2303.00565*, 2023.

[47] Lichao Sun and Lingjuan Lyu. Federated model distillation with noise-free differential privacy. In *Proc. Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1563–1570, Aug. 2021.

[48] Lichao Sun, Jianwei Qian, and Xun Chen. LDP-FL: Practical private aggregation in federated learning with local differential privacy. In *Proc. Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1571–1578, Aug. 2021.

[49] Tao Sun, Dongsheng Li, and Bao Wang. Decentralized federated averaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[50] Yan Sun, Li Shen, Tiansheng Huang, Liang Ding, and Dacheng Tao. Fedspeed: Larger local interval, less communication round, and higher generalization accuracy. In *International Conference on Learning Representations*.

[51] Om Thakkar, Galen Andrew, and H. B. McMahan. Differentially private learning with adaptive clipping. *ArXiv*, abs/1905.03871, 2019.

[52] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Hang Su, Bo Zhang, and H. Vincent Poor. User-level privacy-preserving federated learning: Analysis and performance optimization. *IEEE Transactions on Mobile Computing*, 21(9):3388–3401, 2022.

[53] Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-IID federated learning. In *International Conference on Learning Representations*, 2021.

[54] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Gosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in pytorch. In *Proc. Privacy in Machine Learning (PriML) Workshop, NeurIPS*, Virtual, Dec. 2021.

[55] Lin Zhang, Li Shen, Liang Ding, Dacheng Tao, and Ling-Yu Duan. Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10174–10183, 2022.

[56] Yang Zhao, Hao Zhang, and Xiuyuan Hu. Penalizing gradient norm for efficiently improving generalization in deep learning. In *International Conference on Machine Learning, ICML*, pages 26982–26992. PMLR, 2022.

[57] Qihuang Zhong, Liang Ding, Li Shen, Peng Mi, Juhua Liu, Bo Du, and Dacheng Tao. Improving sharpness-aware minimization with fisher mask for better generalization on language models. *arXiv preprint arXiv:2210.05497*, 2022.

[58] Yuqing Zhu, Xiang Yu, Yi-Hsuan Tsai, F. Pittaluga, M. Faraki, Manmohan Chandraker, and Yu-Xiang Wang. Voting-based approaches for differentially private federated learning. *ArXiv*, abs/2010.04851, 2020.