Rethinking the Flat Minima Searching in Federated Learning

Taehwan Lee [1] Sung Whan Yoon [1 2]

Abstract

Albeit the success of federated learning (FL) in decentralized training, bolstering the generalization of models by overcoming heterogeneity across clients still remains a huge challenge. To aim at improved generalization of FL, a group of recent works pursues flatter minima of models by employing sharpness-aware minimization in the local training at the client side. However, we observe that the global model, i.e., the aggregated model, does not lie on flat minima of the global objective, even with the effort of flatness searching in local training, which we define as *flatness discrepancy.* By rethinking and theoretically analyzing flatness searching in FL through the lens of the discrepancy problem, we propose a method called Federated Learning for Global Flatness (FedGF) that explicitly pursues the flatter minima of the global models, leading to the relieved flatness discrepancy and remarkable performance gains in the heterogeneous FL benchmarks.

1. Introduction

Federated Learning (FL) has drawn great attention as a key framework for enabling decentralized learning across an immense number of distributed clients while preserving data privacy. The essence of FL is keeping local data on the client side and communicating the gradients or model parameters between a server and clients, where the server's direct access to the local data is prohibited (McMahan et al., 2017). Nonetheless, there still exist daunting challenges that remain unsolved. Most importantly, the diversity or heterogeneity of data distribution across clients is shown to hinder the successful aggregation of global model parameters, leading to deteriorated performance and inhibiting model convergence (Li et al., 2020b). To overcome the problem, researchers have dedicated to developing FL algorithms that achieve successful aggregation across heterogeneous clients (Li et al., 2020a; Karimireddy et al., 2020; Acar et al., 2021). Although such intensive efforts have been made to break the hurdle of heterogeneity, the agreed rule of thumb methods and principles have not yet been established.

One of the intriguing recent approaches to enhance the generalization is employing particular optimization methods that find flatter minima on loss surface, which is widely observed to enhance the generalization of deep models against the data distribution shifts (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017; Izmailov et al., 2018; Foret et al., 2021; Cha et al., 2021). The most popular flatness searching method is Sharpness-Aware Minimization (SAM), which incorporates the flatness around the minimum into the cost function (Foret et al., 2021). Building on the promising optimizer, researchers in the FL field have recently confirmed the effectiveness of SAM in strengthening the performance of FL algorithms for heterogeneous settings (Qu et al., 2022; Caldarola et al., 2022). Their approaches basically employ a SAM or SAM-variant optimizer at the local training step at each client to find a flatter local model of the local objective, which indeed yields considerable performance gains for the aggregated global model.

Let us then raise a pivotal question to rethink the flatness searching in FL: *"Does the flatness searching in local training truly imply the flatness of the global model for the global objective?"* The answer is *"Not for the heterogeneous FL cases"*. When the heterogeneity across clients becomes severe, we observed that the existing FL methods with the flat minima searching look effective for finding flatter minima in local training, but the global model does not lie on flatter minima for the global objective. One of the recent works initiated a discussion on this issue, but little intention was paid to elaborate on the formal understanding of it (Sun et al., 2023a). The issue is raised particularly in the regime of decentralized training, and it severely degrades the performance of the flatness searching FL methods. We formally define the issue as *flatness discrepancy*.

In this paper, we empirically and theoretically analyze the relationship between heterogeneity across clients and flatness discrepancy: A strong heterogeneity leads to severe

[1] Graduate School of Artificial Intelligence, Ulsan National Institute of Science and Technology (UNIST), Ulsan, South Korea [2] Department of Electrical Engineering, UNIST, Ulsan, South Korea. Correspondence to: Sung Whan Yoon <shyoon8@unist.ac.kr>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

discrepancy, eventually yielding the degraded performance of the global model. Based on the findings, we propose a method called Federated Learning for Global Flatness (FedGF) that relieves flatness discrepancy, leading to flatter minima of the global model. The key of FedGF is to explicitly consider the sharpness of the global model when running SAM in the local training. Specifically, we utilize the interpolated perturbation for SAM in both views of local and global objectives. We empirically confirm that our method shows remarkable performance gains over prior flatness searching FL methods, ranging up to +5.09% and +10.02% gains in the heterogeneous CIFAR-10 and CIFAR-100 benchmarks, respectively. Also, FedGF shows significantly faster convergence in heterogeneous cases, which is theoretically guaranteed by showing how FedGF suppresses the heterogeneity-related factor in the convergence analysis.

2. Preliminaries and Motivations

We here present the preliminaries for the baseline FL framework called FedAvg (McMahan et al., 2017) and a popular flatness searching FL algorithm, i.e., FedSAM (Qu et al., 2022; Caldarola et al., 2022). Also, we here define the flatness discrepancy, which is the key motivation of our work, and empirically show how it appears to FedSAM.

2.1. Preliminaries of FL: FedAvg

Notations: In the FL setting with N clients and a server, each client contains m_i local data samples, where $i \in [N]$ is the index of the client. $[N]$ indicates the set of integers ranging from 1 to N. A data sample $z_{i,j} = (x_{i,j}, y_{i,j})$ is j-th sample of i-th client with paired input $x_{i,j}$ and its label $y_{i,j}$, and it is drawn from the local data distribution $\mathcal{D}_i$.

The local objective function of client i is $F_i(w) := \frac{1}{m_i}\sum_{j=1}^{m_i} l(w, z_{i,j})$, where w is the model weight and $l(\cdot,\cdot)$ is the loss function. FL basically aims to find the model weight $w^\star$ that minimizes the global objective $F(w)$, i.e.,

$$w^\star = \underset{w}{\operatorname{argmin}} \left\{ F(w) := \sum_{i=1}^{N} \frac{m_i}{m} F_i(w) \right\}, \tag{1}$$

where m is the total number of data samples across clients.

The way to optimize the model weight without accessing the samples on the client side is to adopt a repetitive aggregation process called round, where a round consists of local training of models at each client and aggregation of the locally-trained models at the server.

Local training: At round $r \in [R]$, each client receives the aggregated model w^r from the server and runs local training with K epochs. Specifically, local training is done with empirical risk minimization of the local loss:

$$w^r_{i,k+1} = w^r_{i,k} - \eta_l \nabla F_i(w^r_{i,k-1}), \tag{2}$$

where k is the number of local epochs, η_l is the learning rate and $w^r_{i,0} = w^r$. After K epochs, client i obtains $w^r_{i,K}$.

Aggregation: The updated local models are then uploaded to the server and aggregated to obtain global model w^{r+1}:

$$w^{r+1} = \sum_{i \in \mathcal{S}^r} \frac{m_i}{m} w^r_{i,K}, \tag{3}$$

where $\mathcal{S}^r \subseteq [N]$ is the index of participating clients for round r. After the aggregation, the next local training for round $r+1$ follows by broadcasting the global model to the clients. With a sufficient number of rounds up to R, the global model converges to the optimal weight in Eq. (1).

2.2. Preliminaries of Flatness Searching in FL: FedSAM

FedSAM adopts the SAM optimizer (Foret et al., 2021) for flatness searching in the local training of FedAvg (Qu et al., 2022; Caldarola et al., 2022).

SAM optimizer: The SAM optimizer transforms a loss function $f(w)$ into a min-max cost function as follows:

$$\min_{w} \max_{\|\delta\| \le \rho} F(w + \delta), \tag{4}$$

where ρ is a positive real number and $\|\delta\|$ is L2-norm of δ. As a key factor, δ works as the perturbation that maximally raises the loss value so that the SAM optimizer can find flat minima. The perturbation can be simply approximated as the gradient direction, which points to the steepest direction of the loss surface.

FedSAM: By adopting the min-max problem of Eq. (4) in local training, FedSAM perturbs local model $w^r_{i,k}$:

$$\tilde{w}^r_{i,k} = w^r_{i,k} + \delta = w^r_{i,k} + \rho g^r_{i,k}/\|g^r_{i,k}\| \tag{5}$$

$$w^r_{i,k+1} = w^r_{i,k} - \eta_l \tilde{g}^r_{i,k}, \tag{6}$$

where $g^r_{i,k} = \nabla F_i(w^r_{i,k})$ is the gradient computed at $w^r_{i,k}$, $\tilde{w}^r_{i,k}$ is the perturbed model weight, and $\tilde{g}^r_{i,k} = \nabla F_i(\tilde{w}^r_{i,k})$ is the gradient computed at the perturbed model. Thus, FedSAM finds the local model with flatter minima, leading to the improved performance of the aggregated global model.

2.3. Motivations: Flatness Discrepancy

Now, we are ready to discuss the flatness discrepancy issue: *Flatness searching in local training does not imply the flatness of the global model.* First, we formally define the discrepancy, i.e., $\Delta_{\mathcal{F}}$, as the difference gap of the flatness between the global and local models. For simplicity, we here drop the notations of round r and local epoch k.

Definition 1. *Flatness discrepancy $\Delta_{\mathcal{F}}$ of the global model*

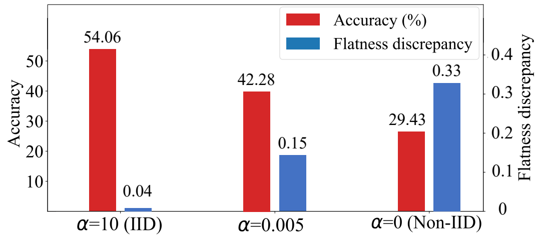


Figure 1: The performance and the flatness discrepancy ($\Delta_{\mathcal{F}}$) of FedSAM for the CIFAR-100 experiment

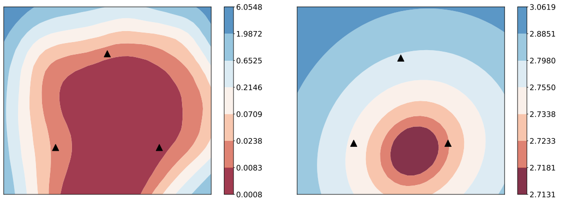
w and the local models $\{w_i\}_{i=1}^N$ is defined as:

$$\Delta_{\mathcal{F}} := \Bigg| \max_{||\delta|| \leq \rho} F(w+\delta) - F(w) - \left[\sum_{i=1}^{N} \frac{m_i}{m} \max_{||\delta_i|| \leq \rho} F_i(w_i+\delta_i) - F_i(w_i) \right] \Bigg|. \quad (7)$$

When $\Delta_{\mathcal{F}}$ is small, it means that the increasing amount of losses of the global and local objectives are similar, i.e., the degree of the flatness is similar. When $\Delta_{\mathcal{F}}$ is large, the increasing amount of losses of the global and local objectives are not the same, i.e., the flatness is discrepant. Herein, we want to provide the motivating empirical results to show how the discrepancy issue arises in flatness searching FL.

As a preliminary experiment, we compute the flatness discrepancy values of FedSAM for the CIFAR-100 FL benchmark. As shown in Fig. 1, FedSAM shows significant performance degradation as the heterogeneity increases (when α decreases to 0, the data distribution across clients becomes heterogeneous, i.e., non-IID (non-independently and identically distributed)). Interestingly, along with the performance degradation, we observe that the flatness discrepancy increases when the heterogeneity gets worse. It empirically reveals that a naive application of SAM to local training of FL does not guarantee the flatness of the global model for the global objective. Also, we visualize how the discrepancy appears on the loss surface. As presented in Fig. 2, FedSAM shows flatter minima around the local model, but the global model does not lie on flatter minima.

We provide an intuition of how the heterogeneity causes the larger flatness discrepancy. With the strong heterogeneity, the local and global objectives, i.e., F_i and F, respectively, trivially deviate from each other due to the data distribution gap, so the flatness searching in the local training does not imply the flatness of the global model. On the other hand, if the heterogeneity is not severe, i.e., close to the IID case, the local objective is ideally the same as the global objective, so the flatness searching for local training directly links to the global model with flatter minima of global objective.



(a) Local model, w_i (b) Global model, w

Figure 2: Visualization of the loss surface of FedSAM for the CIFAR-100 case ($\alpha = 0$).

3. Related Work

3.1. Heterogeneity Issues in FL

In past years, various strategies have been proposed to solve the heterogeneity issues in FL. A main branch of prior approaches focuses on regularizing local training to alleviate the divergence of the local models. As early works, FedAvgM (Hsu et al., 2019) and SCAFFOLD (Karimireddy et al., 2020) utilize the update of the global model as momentum in global and local training, respectively, to regularize the divergence of local gradients. FedProx (Li et al., 2020a) adopts a regularization term, which is called the proximal term, to prevent local models from largely deviating from the global model. FedDyn (Acar et al., 2021) utilizes the dynamic regularization term, which is tailored to each client, so suppressing the discrepancy between the global and local models. FLIX (Gasanov et al., 2021) utilizes the interpolation between global and local model parameters. Our method, FedGF, is based on flat minima searching, which is clearly different from the regularization-based methods.

3.2. Flat Minima Searching

In centralized learning: From an early finding of the benefits of flat minima over sharp minima of the model parameters on loss surface (Keskar et al., 2017), the potential of flat minima for enhancing the generalization ability of deep models is largely investigated in both empirical and theoretical ways. A group of works with Stochastic Weight Averaging (SWA) has been suggested as a simple heuristic method for finding flatter minima (Izmailov et al., 2018; Cha et al., 2021). As another branch of tools, Sharpness-Aware Minimization (SAM) embeds the flatness term into the cost function of the optimizer for seeking flatter minima (Foret et al., 2021; Kwon et al., 2021).

Correlation to the generalization: Some researchers have raised doubts about the correlation between the generalization and the flatness. Sharp minima are shown to be able to generalize (Dinh et al., 2017), and large models, e.g., transformers, empirically seem not to well correlate its flatness to the generalization capability (Andriushchenko et al., 2023;

Kim et al., 2023). Also, a flatness metric should be chosen carefully to show the positive correlation to the generalization (Bisla et al., 2022). The recent active debate argues that flatter minima do not directly imply better generalization, but it does not contradict the shown advantages of flat minima searching in standard training (Izmailov et al., 2018; Foret et al., 2021; Kwon et al., 2021), FL (Qu et al., 2022; Caldarola et al., 2022; Dai et al., 2023; Sun et al., 2023a;b), and out-of-distribution generalization (Cha et al., 2021).

In the FL setting: The existing FL methods that pursue flatter minima are based on applying SAM or SAM-variants to FL, e.g., FedSAM, FedASAM, MoFedSAM, FedGAMMA, FedSMOO, and FedSpeed (Caldarola et al., 2022; Qu et al., 2022; Dai et al., 2023; Sun et al., 2023a;b). FedSAM simply applies the SAM optimizer to local training. FedASAM controls the size of the perturbation along with the magnitude of model parameters. On the other hand, MoFedSAM utilizes the momentum, which is the update of the global model, when running SAM on the client side. Its strategy is analogous to that of SCAFFOLD (Karimireddy et al., 2020), which utilizes the update of the global model as momentum in the local updates of FedAvg. Moreover, both two algorithms, FedSpeed (Sun et al., 2023b) and FedGAMMA (Dai et al., 2023), utilize the gradient computed at the SAM-based perturbed weight. Specifically, FedSpeed tunes the local perturbed gradient with the proximal term, and FedGAMMA tunes it with the local perturbation of other clients. FedSMOO (Sun et al., 2023a) further tunes the local perturbation by leveraging the global perturbation approximated by Taylor expansion. Our FedGF is closely related to the methods that tune the local perturbations to pursue better performance. However, FedGF is unique in interpolating the local and global perturbations and managing the interpolation by observing the divergence between local and global models, yielding remarkable gains over others.

4. Proposed Method: FedGF

4.1. Training Process of FedGF

Based on the preliminaries in Section 2, we here focus on the local training and the aggregation steps of FedGF.

Local training: At the beginning of round r, client i receives the aggregated global model w^r, and runs K local epochs. For local epoch k, client i computes the two perturbed models, $\tilde{w}^r_{i,k}$ and $\tilde{w}^r$ as follows:

$$g^r_{i,k} = \nabla F_i(w^r_{i,k}, \zeta_{i,k}) \tag{8}$$

$$\tilde{w}^r_{i,k} = w^r_{i,k} + \rho g^r_{i,k}/\|g^r_{i,k}\| \quad \text{(perturbed local model)} \tag{9}$$

$$\triangle^r = w^{r-1} - w^r \tag{10}$$

$$\tilde{w}^r = w^r + \rho\triangle^r/\|\triangle^r\| \quad \text{(perturbed global model)} \tag{11}$$

As FedSAM works, the perturbation of the local model is computed (referring to Eq. (9)). When only using the perturbed local model, we already found that the aggregated global model is not located on flatter loss surface. To pursue the flatness of the global model for the global objective, we should consider the perturbation in the view of the global model and objective. However, the perturbation of the global model for the global objective cannot be tractable because each client cannot access the globally aggregated data samples across clients. To detour the hardship, we utilize the difference between the previous and the current global model, as formulated by Eq. (10), which is an approximated direction of the gradients for the global objective, i.e., $\nabla F(w^r)$. Based on the global perturbation, FedGF computes the perturbed global model in Eq. (11), which can be understood as the perturbed model with the maximally raised global objective loss. FedGF then interpolates the perturbed global and local models to compute $\tilde{w}^r_{i,k,c}$:

$$\tilde{w}^r_{i,k,c} = c\tilde{w}^r + (1-c)\tilde{w}^r_{i,k}, \tag{12}$$

where $0 \leq c \leq 1$ is the interpolation coefficient to control the position of $\tilde{w}^r_{i,k,c}$ between the global and local models. When c is close to 0, it indicates that FedGF becomes FedSAM. When c is close to 1, FedGF leans toward the global model to find flat minima around the global model. FedGF then computes the gradient based on the local epoch $\zeta_{i,k} \sim \mathcal{D}_i$ at position $\tilde{w}^r_{i,k,c}$ to update the local model:

$$w^r_{i,k+1} = w^r_{i,k} - \eta_l \nabla F_i(\tilde{w}^r_{i,k,c}, \zeta_{i,k}). \tag{13}$$

After K epochs, local training for round r is terminated.

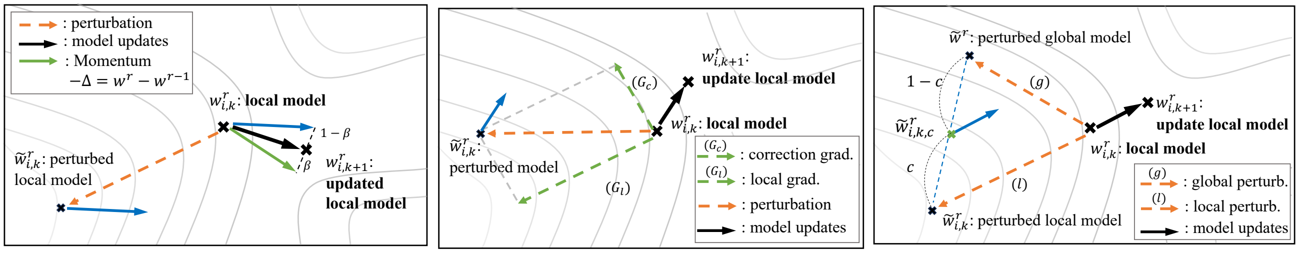
Aggregation: The updated local model is then aggregated at the server to newly update the global model for the next round, i.e., w^{r+1}. Here, we adopt the global learning rate, η_g, suggested in (Reddi et al., 2021) (referring to Eq. (14)). When we set a global learning rate $\eta_g = 1$, it becomes the basis form of aggregation in Eq. (3).

$$w^{r+1} = w^r - \eta_g \sum_{i \in \mathcal{S}^r} \frac{m_i}{m} \triangle^r_i, \tag{14}$$

where $\triangle^r_i = w^r - w^r_{i,K}$.

4.2. Interpolation Coefficient c

Interpolation coefficient c controls the perturbed model in-between the local and global perturbations. Our key strategy to determine c is based on the following wisdom: When the local model largely deviates from the global model, i.e., a non-IID case, a larger c is preferred for focusing on the global model flatness; otherwise, i.e., an IID case, a smaller c is preferred. However, it is non-trivial to control c based on the non-IIDness because the server cannot access the local data distribution, which makes it difficult to measure the heterogeneity of the given setting. Thus, FedGF determines


: perturbation
: model updates
: Momentum
$-\Delta = w^r - w^{r-1}$
$w^r_{i,k}$: local model
$1-\beta$
β
$\tilde{w}^r_{i,k}$: perturbed local model
$w^r_{i,k+1}$: updated local model
$w^r_{i,k+1}$: update local model
(G_c)
$\tilde{w}^r_{i,k}$: perturbed model
(G_l)
$w^r_{i,k}$: local model
(G_c) : correction grad.
(G_l) : local grad.
: perturbation
: model updates
$\tilde{w}^r$: perturbed global model
$1-c$
(g)
$\tilde{w}^r_{i,k,c}$
c
(l)
$w^r_{i,k+1}$: update local model
$w^r_{i,k}$: local model
$\tilde{w}^r_{i,k}$: perturbed local model
(g) : global perturb.
(l) : local perturb.
: model updates

(a) MoFedSAM (Qu et al., 2022)   (b) FedSMOO (Sun et al., 2023a)   (c) **FedGF** (ours)

Figure 3: Schematic of MoFedSAM, FedSMOO, and FedGF. The gray line illustrates the loss landscape of local distribution.

c based on the divergence metric D^r, which represents how local models deviate from global model:

$$D^r = \frac{1}{|S^r|} \sum_{i \in \mathcal{S}^r} \|w^r - w^r_{i,K}\|_2. \tag{15}$$

For mapping D^r to a value between 0 and 1, we adopt the thresholding: $I^r = \mathbf{I}[D^r > T_D]$, where $\mathbf{I}[\cdot]$ is an indicator function and $T_D > 0$ is a hyperparameter. For stability, we use the averaged I^r across recent W rounds in computing c:

$$c = \frac{1}{W} \sum_{i=r-W+1}^{r} I^i. \tag{16}$$

In the non-IID cases, as the round goes on, we empirically observe that c starts to increase at a relatively earlier round, which means that non-zero c is preferred to pursue the global flatness. Otherwise, c stays near zero for the IID case. Also, we further theoretically and empirically analyze that c is strongly related to the faster convergence of FedGF.

A pseudocode of FedGF is provided in Appendix C.

4.3. In-depth Comparison to the Existing Methods

Fig. 3 shows how FedGF is differentiated from the related methods. We here illustrate the schematics of our FedGF and two state-of-the-art methods, i.e., MoFedSAM and FedSMOO, while omitting the figure of FedSAM which is quite straightforward. Black-colored arrows indicate the model update at each local epoch. FedSAM and MoFedSAM only utilize the perturbation computed in view of the local objective. However, FedGF additionally considers the global model perturbation (see the orange-colored dotted arrows). Blue-colored arrows represent the gradient computed at the perturbed model. MoFedSAM uses the momentum for finding the model with a lower global objective (see the green-colored solid line), but FedGF utilizes the reverse direction of the momentum, as the global perturbation (the orange-colored arrow with a marker (g)), aiming to find flatter minima. FedSMOO adjusts the local perturbations by introducing a correction gradient (G_c), and it is closely related to FedGF in adjusting perturbation vectors. The key differences are that FedGF explicitly utilizes the local and global perturbations and adaptively controls the interpolation between two perturbations by observing the divergence between the local and global models. Also, we found that FedSMOO strongly relies on auxiliary regularizations for restricting the local model to allocate near to the global model. FedGF solely works without the additional regularizations.

5. Theoretical Analysis

We provide the theoretical analysis of FedGF, including the convergence behavior and the flatness discrepancy. Before that, we introduce the following assumptions:

Assumption 1. *(Smoothness of loss function) F_i is Lipsichz-smooth for all $i \in [N]$, i.e.,*

$$\|\nabla F_i(w) - \nabla F_i(v)\| \leq L\|w - v\|$$

for all w, v in its domain and $i \in [N]$.

Assumption 2. *(Bounds of gradients) The global variability of the local gradient is bounded by σ_g^2, i.e.,*

$$\|\nabla F_i(w^r) - \nabla F(w^r)\|^2 \leq \sigma_g^2,$$

for all $i \in [N]$ and r.

Assumption 3. *(Bounds of the stochastic gradients) The stochastic gradient $\nabla F_i(w, \zeta_i)$, computed by client i with model parameter w using mini-batch ζ_i is an unbiased estimator of $\nabla F_i(w)$ with variance bounded by σ_l^2, i.e.,*

$$\mathbb{E}_{\zeta_i} \left\| \frac{\nabla F_i(w, \zeta_i)}{\|\nabla F_i(w, \zeta_i)\|} - \frac{\nabla F_i(w)}{\|\nabla F_i(w)\|} \right\|^2 \leq \sigma_l^2,$$

for all $i \in [N]$.

Assumption 1 and 2 are largely accepted by the prior nonconvex FL convergence studies to assume the smoothness of loss function and the bounded heterogeneity (McMahan et al., 2017; Karimireddy et al., 2020; Reddi et al., 2021). Assumption 3, which bounds the variance of stochastic gradients, is from the work of FedSAM (Qu et al., 2022).

5.1. Convergence Analysis of FedGF

We present the convergence analysis of FedGF. Here, ϵ is the error in estimating the direction of the global perturbation of FedGF: $\epsilon := \|\Delta^r/\|\Delta^r\| - \nabla F(w^r)/\|F(w^r)\|\|$.

Theorem 1. *Let the learning rate be $\eta_l = \mathcal{O}(\frac{1}{\sqrt{RKL}}), \eta_g = \sqrt{KN}$, and the amplitude of perturbation is proportional to the learning rate, e.g., $\rho = \mathcal{O}(\frac{1}{\sqrt{R}})$. Under Assumptions 1 - 3 and full client participation, the average of the norm of the gradient generated by the iterative rounds of FedGF satisfies:*

$$\mathcal{O}\Bigg(\frac{FL}{\sqrt{RKN}}+\frac{(1-c)^2}{R}\sigma_g^2 + \frac{L^2(1-c)^2}{R^{3/2}\sqrt{KN}}\sigma_l^2 + \frac{L^2c^2\epsilon^2}{R}\Bigg), \tag{17}$$

where $F = F(\tilde{w}^0) - F(\tilde{w}^*)$ and $F(\tilde{w}^*) = \min_{\tilde{w}} F(\tilde{w})$. For the partial client participation strategy, if we choose the learning rates $\eta_l = \mathcal{O}(\frac{1}{\sqrt{RKL}}), \eta_g = \sqrt{KS}$ and $\rho = \mathcal{O}(\frac{1}{\sqrt{R}})$, the average of the norm of the gradient generated by the iterative rounds of FedGF satisfies:

$$\mathcal{O}\Bigg(\frac{FL}{\sqrt{RKS}}+\left(\frac{(1-c)^2}{R}+1\right)\sigma_g^2 + \frac{L^2(1-c)^2}{R^{3/2}\sqrt{KS}}\sigma_l^2 + \frac{L^2(c^2\epsilon^2+1)}{R}\Bigg), \tag{18}$$

where $S = |\mathcal{S}^r|$.

Remark 1. ***(Faster convergence in non-IID cases)*** Let us focus on two variance terms, i.e., σ_g^2 and σ_l^2, which represent the non-IIDness of the given FL setting and the stochastic variance of local gradients. These terms are crucial in the undesired delay of convergence of FL. When FedGF utilizes a larger c, i.e., reaching 1, then the related terms in Eq. (17) and (18) can be sufficiently suppressed; FedGF can accelerate the convergence even with the strong heterogeneity. In the extensive experiments, we confirm that FedGF tends to use larger values of c in the non-IID cases, leading to a significantly faster convergence than other related baselines.

Remark 2. ***(Error from ϵ vanishes as round goes on)*** Because FedGF approximates the gradient of the global model, there exists the error term ϵ, which probably hinders the convergence (referring to ϵ-involved term in Eq. (17) and (18)). As shown in the theorem, the error term diminishes as the round goes on; it does not ruin FedGF's convergence.

Remark 3. ***(FedGF with $c = 0$ becomes FedSAM)*** As aforementioned, with a smaller c, i.e., around 0, FedGF becomes FedSAM (the convergence analysis also becomes identical to that of (Qu et al., 2022)[1]). Then FedGF does not utilize the global model perturbation, so the last terms in the convergence analysis disappear. However, the heterogeneity and variance terms related to σ_g^2 and σ_l^2 exist in the convergence behavior. When the setting becomes IID, it means that σ_g^2 and σ_l^2 are negligibly small, so FedGF tends to use smaller c values around 0 to behave like FedSAM.

5.2. Flatness Discrepancy Analysis

We address the claims on the flatness discrepancy, $\Delta_{\mathcal{F}}$.

Theorem 2. *$\Delta_{\mathcal{F}}$ is upper bounded as follows:*

$$\Delta_{\mathcal{F}} \le \rho\sigma_g^2 + L\rho \sum_{i\in[N]} \frac{m_i}{m}\|w - w_i\| \tag{19}$$

Remark 4. ***(Heterogeneity, model divergence, and loss smoothness bound the flatness discrepancy)*** As shown in Eq. (19), the term σ_g^2, which increases when heterogeneity gets worse, directly determines the upper bound of $\Delta_{\mathcal{F}}$, coinciding with our understanding of the discrepancy. Also, when the loss is not smooth, i.e., with a larger L, then the discrepancy increases. Finally, when the local model w_i largely deviates from the global model w, then the discrepancy gets worse, which agrees with how FedGF determines interpolation coefficient c based on the model divergence.

Theorem 3. *For FedGF, if we choose $\eta_l = \mathcal{O}(\frac{1}{\sqrt{RKL}})$ and $\rho = \mathcal{O}(\frac{1}{\sqrt{R}}), \eta_g = \sqrt{KN}$, $\Delta_{\mathcal{F}}$ is then bounded as follows:*

$$\Delta_{\mathcal{F}} \le \mathcal{O}\Bigg(\frac{\sigma_g^2}{\sqrt{R}} + \frac{L(1-c)^2\sigma_l^2}{R^{5/2}} + \frac{\sigma_g^2}{LR^{3/2}} + \frac{Lc^2\epsilon^2}{R^{5/2}}\Bigg) \tag{20}$$

Remark 5. ***($\Delta_{\mathcal{F}}$ is suppressed as round increases)*** For FedGF, the upper bound of $\Delta_{\mathcal{F}}$ formalized by Eq. (19) is suppressed as the round increases. For a non-IID, FedGF prefers to use large c, reaching 1, to strongly suppress the heterogeneity-related terms (referring to the second term).

The proofs of all claims are fully presented in Appendix A.

6. Experiments

We extensively evaluate the performance of FedGF[2] on the FL classification benchmarks suggested by (Caldarola et al., 2022), where CIFAR-10 and CIFAR-100 datasets are distributed on clients by utilizing Dirichlet distribution.

6.1. Experimental Settings

Baselines: We compare FedGF with the following methods: FedAvg (McMahan et al., 2017), FedAvgM (Hsu et al., 2019), SCAFFOLD (Karimireddy et al., 2020), FedProx (Li

[1] We found that the analysis for partial participation in (Qu et al., 2022) has a mistake, regarding the remained constant from σ_g. The details are in Appendix B.

[2] Codes are available at github.com/hwan-sig/Official-FedGF

Table 1: Test accuracies with of the FL algorithms on the CIFAR-10 and CIFAR-100 benchmarks

Task	Algorithms	Dirichlet distribution parameter α								
		$Dir.(\alpha=0$, non-IID)			$Dir.(\alpha=0.005)$			$Dir.(\alpha=10$, IID)		
		Number of participating clients per each round								
		5	10	20	5	10	20	5	10	20
CIFAR-10	FedAvg	63.63	65.83	68.33	67.85	71.37	73.03	82.90	82.96	82.93
	FedAvgM	62.73	65.61	68.57	67.56	71.32	75.53	82.72	83.60	83.30
	FedProx	63.13	65.95	67.98	68.06	71.42	72.87	82.72	83.19	82.92
	SCAFFOLD	(✗)	(✗)	(✗)	57.13	56.46	45.27	82.93	83.05	83.39
	FedDyn	66.84	71.01	69.45	70.74	73.78	75.43	83.07	83.58	83.67
	FedSAM	68.11	71.17	72.49	71.87	74.31	76.07	83.78	83.88	83.82
	FedASAM	73.32	74.5	75.49	74.96	75.59	76.57	83.11	83.28	82.89
	MoFedSAM	73.1	71.08	76.66	74.43	77.53	79.27	80.9	81.01	81.02
	FedGAMMA	45.32	47.55	35.07	46.99	48.44	35.58	74.99	66.12	54.85
	FedSMOO	68.82	71.59	72.48	71.9	74.46	75.44	83.72	83.67	83.79
	FedGF	**78.41**	**79.68**	**80.86**	**78.79**	**79.39**	**79.69**	**84.71**	**83.94**	**83.85**
CIFAR-100	FedAvg	29.35	33.79	36.62	38.15	40.58	41.27	50.41	50.20	49.98
	FedAvgM	29.94	30.07	39.35	38.64	40.72	**48.44**	50.37	51.2	50.57
	FedProx	29.19	33.16	36.41	38.54	40.52	40.77	50.10	49.98	49.96
	SCAFFOLD	(✗)	(✗)	(✗)	36.25	(✗)	(✗)	52.28	52.12	52.48
	FedDyn	(✗)	(✗)	(✗)	(✗)	(✗)	(✗)	51.74	52.41	52.59
	FedSAM	29.43	34.32	36.88	42.28	44.57	45.18	54.06	53.75	53.5
	FedASAM	34.43	37.09	38.93	44.36	45.76	46.94	**54.6**	54.42	**54.73**
	MoFedSAM	29.02	35.82	41.26	34.64	42.24	44.92	52.13	52.21	52.07
	FedGAMMA	(✗)	(✗)	(✗)	20.52	14.76	10.33	47.43	38.18	25.06
	FedSMOO	35.35	38.78	40.82	44.39	46.03	47.5	54.31	54.89	54.65
	FedGF	**45.37**	**46.86**	**47.77**	**46.48**	**46.70**	46.08	54.16	**54.62**	54.59

(✗) indicates that the method fails to train, so the results remain at the same level as the random prediction.

et al., 2020a), FedDyn (Acar et al., 2021), FedSAM (Caldarola et al., 2022; Qu et al., 2022), FedASAM (Caldarola et al., 2022), MoFedSAM (Qu et al., 2022), FedGAMMA (Dai et al., 2023), and FedSMOO[3] (Sun et al., 2023a).

Model architecture: We follow the model architecture described in the prior FL works (Hsu et al., 2020; Caldarola et al., 2022), which is a variant of the LeNet architecture by (LeCun et al., 1998). For the larger architecture, such as ResNet-18, we add the results in Appendix D.6.

FL settings: For a given server and 100 clients, we test three different numbers of participating clients per round, i.e., {5, 10, 20}. We distributed 500 data samples per client, and the number of local updates per round is 8, with batch size 64. As done in (Hsu et al., 2020), the prior distribution of local data follows the Dirichlet distribution of α, i.e., $\alpha \in \{0, 0.005, 10\}$ for both CIFAR-10 and CIFAR-100 experiments. When α increases, the setting becomes a IID case. When α goes to zero, it means a non-IID case. The communication round goes up to 10,000 and 20,000 for CIFAR-10 and CIFAR-100, respectively.

Further details of the benchmarks, the hyperparameters, and the model architecture are provided in Appendix C.

6.2. Performance Evaluation

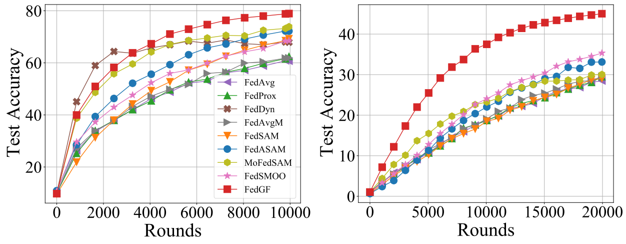
We evaluate the test accuracy of FL algorithms in Table 1 on the CIFAR-10/100 benchmarks. We here provide the following key findings based on the experimental results.

6.2.1. Large Gains for the Non-IID Cases

FedGF significantly outperforms the existing works in the non-IID settings, i.e., $\alpha = 0$. Specifically, it shows the gains ranging from $+4.20\%$ to $+5.30\%$ for CIFAR-10 cases and larger gains ranging from $+6.51\%$ to $+10.94\%$ for CIFAR-100 over the runner-ups. The results verify that FedGF effectively relieves the strong heterogeneity via aggregating a global model with a strong generalization across clients. We believe that the gains directly come from the efforts to search flat minima of global model by FedGF, which is to be confirmed in the following part. As the heterogeneity becomes relieved, i.e., as α increases from 0 to 10, the performance gaps between FedGF and prior works are reduced. This is due to the homogeneity of data distribution in the IID cases, which relieves the discrepancy between the local and global models. Also, we found that the regularization-based FL methods, including SCAFFOLD and FedDyn, is not successful in the non-IID case[4], particularly in CIFAR-100.

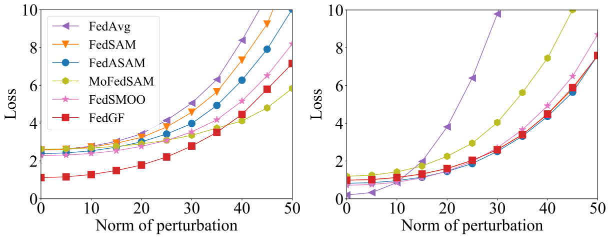
[3] FedSMOO strongly relies on the dynamic regularization. For a fair comparison of the effectiveness on finding flatter global model, we evaluate FedSMOO without the regularizer.

[4] We want remind that the original evaluation in their works are done in the cases with moderate non-IIDness.



(a) CIFAR-10 ($\alpha = 0$) (b) CIFAR-100 ($\alpha = 0$)

Figure 4: Convergence behaviors for non-IID



(a) Non-IID ($\alpha = 0$) (b) IID ($\alpha = 10$)

Figure 5: Loss plots along to perturbation for CIFAR-100

Table 2: LPF, λ_{max}, and $\Delta_{\mathcal{F}}$ results for CIFAR-100

Algorithm	Non-IID ($\alpha = 0$)			IID ($\alpha = 10$)		
	LPF $\downarrow$	$\lambda_{\text{max}} \downarrow$	$\Delta_{\mathcal{F}} \downarrow$	LPF $\downarrow$	$\lambda_{\text{max}} \downarrow$	$\Delta_{\mathcal{F}} \downarrow$
FedAvg	2.67	81.57	0.39	1.24	103.19	0.11
FedSAM	2.67	43.32	0.33	1.10	25.36	**0.04**
FedASAM	2.35	25.46	0.24	0.94	**14.99**	**0.04**
MoFedSAM	2.64	15.11	0.13	1.48	27.28	**0.04**
FedSMOO	2.37	25.83	**0.12**	**0.8**	23.38	0.08
FedGF	**1.36**	**14.07**	0.13	1.08	23.54	**0.04**

$\downarrow$: a lower value is preferred.

6.2.2. Faster Convergence Behavior

We present the convergence behaviors for the non-IID case in Fig. 4a and 4b. We clearly confirm that FedGF shows significantly faster convergence than others, particularly emphasized in the CIFAR-100 case. The results verify **Remark 1** of Theorem 1, which emphasizes that FedGF can suppress the terms of the heterogeneity and the stochastic variance, so it accelerates the speed of convergence by preferring larger c. In the experiments, it indeed pushes c to be 1 for the non-IID cases, leading to faster convergence.

6.2.3. Robustness to Participating Clients

As shown in Table 1, when the heterogeneity gets worse, the FL baselines suffer from severe degradations when the number of participating clients is limited. In contrast, FedGF shows the robust performance for the limited participating clients. We emphasize that MoFedSAM shows a significant drop from 41.26% to 29.02% when the number of participating clients decreases from 20 to 5 for the CIFAR-100 $\alpha = 0$ case. It happens because the momentum of global models used by MoFedSAM largely fluctuates round-by-round when the number of participating clients is limited. FedSMOO suffers from 5.47% drop for the same case. We conjecture that FedSMOO struggles to find the robust perturbation correction term when the number of participations is limited. On the contrary, let us remind **Remark 2** of Theorem 1, which points out that FedGF can suppress the error, ϵ, in estimating the global perturbation as round increases.

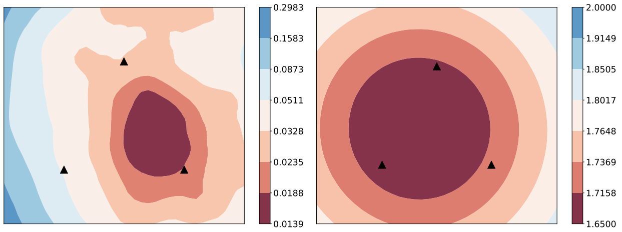
6.2.4. Flatness Results

To confirm the flatness of the global model in both quantitatively and qualitatively, we present various flatness results: i) loss plots, ii) flatness metrics, including flatness discrepancy, and iii) visualization of loss surface.

Loss plots along to perturbations: Fig. 5 shows the plots to confirm how the loss value increases as the perturbation of the model parameter is imposed for the CIFAR-100 experiments with 5 participating clients. In the non-IID case, FedGF shows slightly flatter loss plot than FedAvg, FedSAM, and FedASAM. Although MoFedSAM and FedSMOO show flatter behavior, our FedGF shows much lower loss values in the wide range of perturbation; the gap in loss is more than 1.0, which leads to the large performance gap in accuracies (as confirmed in Table 1). In the IID case, the FL algorithms show similar behavior excepting for FedAvg and MoFedSAM. We conjecture that MoFedSAM suffers from the fluctuation of global momentums caused by the limited number of clients; leading to the unexpected sharp loss plot.

Flatness metrics (LPF, λ_{max}, and $\Delta_{\mathcal{F}}$)**:** The loss plots show a brief understanding of loss surface, but they cannot provide quantitative measurements of flatness. We here compute various flatness metrics for an in-depth analysis: the maximum eigenvalue of the Hessian matrix, i.e., λ_{max}, which is commonly used in prior works, Low-Pass Filter (LPF) based metric, which is recently suggested to show the robust correlation to generalization (Bisla et al., 2022), and the proposed flatness discrepancy $\Delta_{\mathcal{F}}$. While FedAvg showing the worst values, FedGF shows the best flatness for LPF and λ_{max} in the non-IID case. FedGF is the second best for $\Delta_{\mathcal{F}}$ with a minimal gap. As noted in **Remark 4** of Theorem 2, we found that non-IIDness increases $\Delta_{\mathcal{F}}$ for all cases. Also, we confirm that FedGF sufficiently suppresses discrepancy as noted in **Remark 5** of Theorem 3.

Visualization of loss surface: In Fig. 6, we visualize the loss surface of the local and global models of FedGF for F_i and F for the CIFAR-100 cases. It shows that local model shows a moderate flatness, and the global model shows flatter loss surface. When compared with the FedSAM's



(a) Local model, w_i (b) Global model, w

Figure 6: Loss surface of FedGF for CIFAR-100 ($\alpha = 0$).

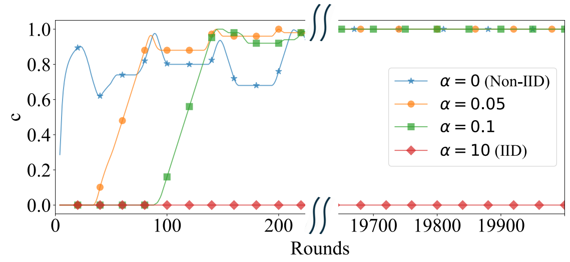


Figure 7: Behavior of c for CIFAR-100

loss surface (as in Fig. 2), we visually confirm that FedGF finds flatter minima of global model for the global objective.

6.2.5. Analysis of communication cost

Herein, we analyze the communication costs of FedGF. In Table 3, the number of model transmissions is measured for the related works focusing on the flatness searching FL methods (‘1’ means a single transmission of model parameters). The baselines, which are FedAvg and FedSAM, upload only the locally trained model and download the averaged global model, leading to a communication cost of 2. FedSMOO and FedGAMMA, state-of-the-art flatness-searching FL methods, require higher costs, i.e., 4 model transmissions, because they transmit both parameters and perturbations between server and clients. In contrast, FedGF transmits the perturbation, which is a pseudo-gradient, from server to client for each round but does not require uploading from the client to the server, leading to the moderate cost of three model transmissions.

Table 3: Number of model transmissions per round

Algorithms	client→ server (upload)	client← server (download)	total
FedAvg	1	1	2
FedSAM	1	1	2
MoFedSAM	1	2	3
FedGF	1	2	3
FedSMOO	2	2	4
FedGAMMA	2	2	4

In Table 4, we compare the actual communication costs, which are the measured number of model and perturbation transmissions, to reach the saturated accuracies of the most typical baseline, i.e., FedAvg. We compare FedGF with FedSMOO, which mostly follows FedGF as a runner-up for all cases (we consider the cases with 5 participating clients). We found that FedGF shows significantly faster convergence, where FedSMOO requires ×2.40 and ×2.98 times higher costs for the non-IID cases. This result coincides with theories (referring to **Remark 1** of Theorem 1) and the empirical results (referring to the part 6.2.2).

Table 4: Number of model transmissions to reach the FedAvg’s final performance ($\times 10^2$)

	Algorithms	$\alpha = 0$	$\alpha = 0.005$	$\alpha = 10$
CIFAR-10	**FedGF**	75	108	165
	FedSMOO	180 (x2.40)	228 (x2.11)	244 (x1.48)
CIFAR-100	**FedGF**	180	210	240
	FedSMOO	536 (x2.98)	440 (x2.10)	248 (x1.03)

6.2.6. Analysis of c

Behavior: Fig. 7 shows how FedGF utilizes c values according to the rounds. FedGF computes c based on the model divergence in Eq. (15) and (16). For the IID case, FedGF steadily uses $c = 0$ due to the minimal divergence between the local and global models, which coincides with **Remark 3** of Theorem 1. When non-IIDness gets worse, FedGF prefers to use larger c values up to 1, which strongly employs the global perturbation. This exactly agrees with the interpretation of **Remark 1** of Theorem 1, which states that FedGF relieves the heterogeneity by letting c be larger.

Table 5: Static c vs. FedGF (adaptive c)

Dataset	IIDness	$c = 0$	$c = 0.5$	$c = 1$	**FedGF**
CIFAR-10	Non-IID	68.11	71.24	78.05	**78.41**
	IID	83.78	82.95	81.94	**84.71**
CIFAR-100	Non-IID	29.43	26.64	44.39	**45.37**
	IID	54.06	52.47	46.68	**54.16**

Ablation on static c: As shown in Table 5, FedGF, which adaptively computes c, is better than the cases of static c.

7. Conclusion

We rethink flat minima searching in FL with the novel perspective of flatness discrepancy. It gets worse when the heterogeneity becomes severe, leading to the deterioration of the prior flat minima searching FL algorithms. Based on this wisdom, we propose FedGF, which can relieve the discrepancy by utilizing both local and global perturbations in the SAM optimizer. FedGF largely outperforms the existing FL methods, particularly in non-IID cases.

Acknowledgements

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2020-II201336, Artificial Intelligence Graduate School Program (UNIST)), (No. RS-2021-II212201, Federated Learning for Privacy-Preserving Video Caching Networks), and (No. IITP-2024-RS-2022-00156361, Innovative Human Resource Development for Local Intellectualization program).

Impact Statement

With a broader perspective, FL would be a key technology that secures private data while training deep models. We conjecture that our method would not raise an ethical issue.

References

Acar, D. A. E., Zhao, Y., Matas, R., Mattina, M., Whatmough, P., and Saligrama, V. Federated learning based on dynamic regularization. In *International Conference on Learning Representations (ICLR)*, 2021.

Andriushchenko, M., Croce, F., Müller, M., Hein, M., and Flammarion, N. A modern look at the relationship between sharpness and generalization. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Bisla, D., Wang, J., and Choromanska, A. Low-pass filtering sgd for recovering flat optima in the deep learning optimization landscape. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 8299–8339. PMLR, 2022.

Caldarola, D., Caputo, B., and Ciccone, M. Improving generalization in federated learning by seeking flat minima. In *European Conference on Computer Vision (ECCV)*, pp. 654–672. Springer, 2022.

Cha, J., Chun, S., Lee, K., Cho, H.-C., Park, S., Lee, Y., and Park, S. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:22405–22418, 2021.

Dai, R., Yang, X., Sun, Y., Shen, L., Tian, X., Wang, M., and Zhang, Y. Fedgamma: Federated learning with global sharpness-aware minimization. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning (ICML)*, pp. 1019–1028. PMLR, 2017.

Du, Z., Sun, J., Li, A., Chen, P.-Y., Zhang, J., Li, H. H., and Chen, Y. Rethinking normalization methods in federated learning. In *Proceedings of the 3rd International Workshop on Distributed Machine Learning*, pp. 16–22, 2022.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations (ICLR)*, 2021.

Gasanov, E., Khaled, A., Horváth, S., and Richtárik, P. Flix: A simple and communication-efficient alternative to local methods in federated learning. In *International Conference on Machine Learning (ICML)*, pp. 11374–11421. PMLR, 2021.

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural computation*, 9(1):1–42, 1997.

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

Hsu, T.-M. H., Qi, H., and Brown, M. Federated Visual Classification with Real-World Data Distribution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, August 2020.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning (ICML)*, pp. 448–456. pmlr, 2015.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning (ICML)*, pp. 5132–5143. PMLR, 2020.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations (ICLR)*, 2017.

Kim, H., Park, J., Choi, Y., and Lee, J. Fantastic robustness measures: The secrets of robust generalization. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-10 (canadian institute for advanced research).

Kwon, J., Kim, J., Park, H., and Choi, I. K. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning (ICML)*, pp. 5905–5914. PMLR, 2021.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems (MLSys)*, 2:429–450, 2020a.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations (ICLR)*, 2020b.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics (AISTATS)*, pp. 1273–1282. PMLR, 2017.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8024–8035. Curran Associates, Inc., 2019.

Qu, Z., Li, X., Duan, R., Liu, Y., Tang, B., and Lu, Z. Generalized federated learning via sharpness aware minimization. In *International Conference on Machine Learning (ICML)*, pp. 18250–18280. PMLR, 2022.

Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., and McMahan, H. B. Adaptive federated optimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Sun, Y., Shen, L., Chen, S., Ding, L., and Tao, D. Dynamic regularized sharpness aware minimization in federated learning: Approaching global consistency and smooth landscape. In *International Conference on Machine Learning (ICML)*, pp. 32991–33013. PMLR, 2023a.

Sun, Y., Shen, L., Huang, T., Ding, L., and Tao, D. Fedspeed: Larger local interval, less communication round, and higher generalization accuracy. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023b.

Zeng, D., Liang, S., Hu, X., Wang, H., and Xu, Z. Fedlab: A flexible federated learning framework. *Journal of Machine Learning Research*, 24(100):1–7, 2023.