Dynamic Regularized Sharpness Aware Minimization in Federated Learning: Approaching Global Consistency and Smooth Landscape

Yan Sun [1] **Li Shen** [2] **Shixiang Chen** [2] **Liang Ding** [2] **Dacheng Tao** [1]

Abstract

In federated learning (FL), a cluster of local clients are chaired under the coordination of the global server and cooperatively train one model with privacy protection. Due to the multiple local updates and the isolated non-iid dataset, clients are prone to overfit into their own optima, which extremely deviates from the global objective and significantly undermines the performance. Most previous works only focus on enhancing the consistency between the local and global objectives to alleviate this prejudicial client drifts from the perspective of the optimization view, whose performance would be prominently deteriorated on the high heterogeneity. In this work, we propose a novel and general algorithm `FedSMOO` by jointly considering the optimization and generalization targets to efficiently improve the performance in FL. Concretely, `FedSMOO` adopts a dynamic regularizer to guarantee the local optima towards the global objective, which is meanwhile revised by the global Sharpness Aware Minimization (SAM) optimizer to search for the consistent flat minima. Our theoretical analysis indicates that `FedSMOO` achieves fast $\mathcal{O}(1/T)$ convergence rate with low generalization bound. Extensive numerical studies are conducted on the real-world dataset to verify its peerless efficiency and excellent generality.

1. Introduction

FL is a horizontally distributed framework that allows a mass of edge devices to collaboratively train a global model on their isolated private dataset (McMahan et al., 2017). To protect the data privatization and localization, each client has no direct access to the other dataset. To enable information exchange during training, a global server is employed to communicate the intermediate variables and parameters with the local clients, and to assist them with performing the joint training. Due to the poor bandwidth throttling especially on the global server, it adopts multiple local training and partial participation to mitigate the communication bottleneck to a greater extent. With extensive studies of FL, theoretical analysis reveals that the major influence on limiting the performance of FL is client drifts, whose essence is that the inconsistent local optima deviate from the global objective on the heterogeneous dataset (Karimireddy et al., 2020; Woodworth et al., 2020; Li et al., 2020b; Kairouz et al., 2021). Yang et al. (2021) theoretically demonstrate that the performance of the classical `FedAvg` method suffers from the length of local updates and the number of partial participation multiplied by the constant upper bound of the variance of the heterogeneous gradient, which contributes as the dominant term of the convergence rate. This divergence would be extremely multiplied by both increasing the local interval and reducing the participation ratio.

To address these problems above, spontaneously, most existing works focus on tackling the global consistent objectives from the perspective of optimization view with Empirical Risk Minimization (ERM) (Malinovskiy et al., 2020; Wang et al., 2020b; Charles & Konečný, 2021; Wang et al., 2021; Zhang et al., 2022). With the process of global training, steady correction to the local objectives based on the regularization term productively reduces the upper bound of theoretical convergence rate (Zhang et al., 2021; Acar et al., 2021). Furthermore, considering the asynchrony, reweighted empirical averages contribute to aggregating more stable global parameters (Wang et al., 2020a; Huang et al., 2022). However, in the face of the challenge of a higher heterogeneous dataset, e.g. for heavy tail, the global solution falling into a sharp landscape may fail to provide solid estimates and yet overfit to an unreliable minimum with poor stability (Brownlees et al., 2015; Foret et al., 2020). Therefore, we need to contemplate the training process comprehensively and completely. More attention must be paid to the generalization ability under ensuring the optimization performance in the federated framework.

To further approach the consistent stationary state with cred-

[1]The University of Sydney, Institute of Engineering, Sydney NSW, Australia [2]JD Explore Academy, Beijing, China. Correspondence to: Li Shen <mathshenli@gmail.com>.

ible stability in FL, in this paper, we jointly take into account ensuring both the optimization and generalization performance on the global target and propose a novel algorithm `FedSMOO` to guide the local clients to search for the consistent flat minima aligned to the global objective which significantly improve the performance, especially on the dataset with high heterogeneity. Specifically, we first adopt the dynamic regularizer to align the global and local objectives. Under the assurance of local consistency, we consider the global objective as searching for a flat stationary state based on the global SAM optimizer, which will correct the regularization term dynamically. As the local multiple updates, applying the general SAM optimizer on the global server cannot get the accurate perturbation when updating locally. To tackle this, we introduce an additional equality constraint on the perturbation in the global optimization objective and employ another regularization term to modify the updates of SAM, which maintains the consistency of the local perturbation terms aligned to the global target. When the active local clients converge, they eventually achieve the global stationary state with a smooth loss landscape.

In contrast with Qu et al. (2022) (`FedSAM`) which firstly incorporates the SAM optimizer with FL and allows each client to train the model with a local SAM optimizer, they focus on improving the *local generality* to inherently align the local flatness closer to the global flatness and expect to enhance the consistency. `FedSMOO` jointly considers the global consistency and *global generality* as the target from the perspective of both the optimization and generalization to approach a consistent smooth minimum.

Theoretically, we relax the general assumption of bounded heterogeneous gradients and demonstrate that our proposed `FedSMOO` could achieve the fast $\mathcal{O}(1/T)$ convergence rate on smooth non-convex scenes. Then we provide the generalization bound of the global function to theoretically guarantee the smoothness of the optimal state. Empirically, we conduct extensive experiments on the CIFAR-10/100 dataset and show that `FedSMOO` achieves the faster convergence rate and higher generalization accuracy in practice, compared with 7 baselines including `FedAvg`, `FedAdam`, `SCAFFOLD`, `FedCM`, `FedDyn`, `FedSAM`, and `MoFedSAM`.

In the end, we summarize our main contributions as:

- We propose a novel and general federated algorithm, `FedSMOO`, to jointly consider both global consistency and global generality as the target, which simultaneously stays fast convergence and high generalization.

- We provide the theoretical analysis of the convergence rate and generalization bound. On the smooth non-convex scene, `FedSMOO` could achieve the fast $\mathcal{O}(1/T)$ convergence rate without the general assumption of the bounded heterogeneous gradients.

- Extensive numerical studies are conducted on the CIFAR-10/100 dataset to verify the excellent performance of `FedSMOO`, which outperforms several classical baselines, especially on the high heterogeneity.

2. Related Work

FL: McMahan et al. (2017) propose the federated framework to corporately train a model among a cluster of edged devices with the isolated private dataset (Hard et al., 2018). To reduce the communication costs, multiple local updates and partial participation are yet adopted, which significantly undermines the performance (Woodworth et al., 2020) and becomes a trade-off in the federated framework in practice (Stich, 2018; Asad et al., 2020; Yang et al., 2021). The essential reason for this is due to the global heterogeneous distribution of isolated private datasets (Konečný et al., 2016; Yang et al., 2019; Li et al., 2020a; Liu et al., 2022). With further studies in FL, this is summarized as client drifts (Karimireddy et al., 2020) and inconsistency objectives among clients (Charles & Konečný, 2021; Malinovskiy et al., 2020; Wang et al., 2020b; 2021). To address these difficulties, a series of algorithms are proposed from the perspective of optimization. Sahu et al. (2018) propose to penalize the quadratic term of the equality constraint on the local objective to limit the local updates, which introduces a balance of local and global optima (Hanzely & Richtárik, 2020). Pathak & Wainwright (2020); Tran Dinh et al. (2021) combine proximal mapping to approach the global consistency. Liang et al. (2019); Karimireddy et al. (2020); Haddadpour et al. (2021) utilize the variance reduction technique to reduce the stochastic variance of gradients caused by the sampling among clients. Reddi et al. (2020) improve the performance by applying an efficient adaptive optimizer on the server side. Zhang et al. (2021) employ the primal-dual method on the server and clients respectively and Wang et al. (2022); Gong et al. (2022) extend the method to the Alternating Direction Method of Multipliers (ADMM). Acar et al. (2021) propose a variant of ADMM via dynamic regularization terms. Wang et al. (2019); Das et al. (2022) introduce the momentum to the local optimizer to improve the performance. Xu et al. (2021); Ozfatura et al. (2021) bring the client-level momentum across the clients and global server which forces local updates to be similar.

SAM: Hochreiter & Schmidhuber (1994) firstly explore the strong relationship between the generalization and its flat loss landscape. Inspired by this, Foret et al. (2020) propose the SAM optimizer to search for a flat minimum with higher generality. Kwon et al. (2021) propose a scale-invariant SAM to improve the training stability. Zhuang et al. (2022) redefine the sharpness from an intuitive and theoretical view for SAM. After that, extensive variants of SAM are proposed Zhao et al. (2022); Sun et al. (2023a); Zhong et al. (2022);

Mi et al. (2022). Existing works (Foret et al., 2020; Kwon et al., 2021) show that the flat loss landscape approached by the SAM optimizer holds higher stability and generality. Besides, Qu et al. (2022) firstly take into account the local generality to inherently enhance the consistency. Caldarola et al. (2022); Sun et al. (2023b); Shi et al. (2023a;b) propose some variants to adopt the SAM optimizer in FL.

3. Rethinking `FedSAM` Algorithm

Below, we introduce the preliminaries of FL and `FedSAM`. Then, we provide a deep understanding for `FedSAM`.

3.1. Preliminaries

Federated Learning (FL)   Taking into account the classical FL framework, we focus on the minimization of the following finite sum problem:

$$\min_{w}\left\{f(w) = \frac{1}{m}\sum_{i\in[m]} f_i(w)\right\},$$
$$f_i(w) \triangleq \mathbb{E}_{\varepsilon_i\sim\mathcal{D}_i} f_i(w, \varepsilon_i), \qquad (1)$$

where $f\colon \mathbb{R}^d \to \mathbb{R}$ is the global objective function; w is the parameters; m is the number of total clients; $[m]$ is the set of total clients; ε_i is the data sample randomly drawn from the distribution $\mathcal{D}_i$ which may differ across the local clients as the heterogeneity; f_i is the empirical loss of the i-th client.

Sharpness Aware Minimization (SAM)   SAM aims to jointly minimize the loss function and smooth the loss landscape by solving the following problem:

$$\min_{w}\left\{f_s(w) \triangleq \max_{\|s\|\le r} f(w+s)\right\}, \qquad (2)$$

where $\|\cdot\|$ is l_2-norm and r is the size of the neighborhood. SAM does not solve the min-max objective and adopts an efficient approximation in practice. Via the first-order Taylor expansion of f, the solution of the inner maximization is:

$$s^*(w) \approx \operatorname*{arg\,max}_{\|s\|\le r}\left\{f(w) + s^\top\nabla f(w)\right\}$$
$$= r \cdot \nabla f(w)/\|\nabla f(w)\|, \qquad (3)$$

where ∇ is the abbreviation for ∇_w on parameters w. Substituting back into equation (2) and differentiating:

$$\nabla f_s(w) \approx \nabla f(w + s^*(w))$$
$$\approx \nabla f(w)|_{w+r\cdot\nabla f(w)/\|\nabla f(w)\|}. \qquad (4)$$

Equation (4) approximates the SAM gradient by the stage-wised calculation of 1) performing a gradient ascent step to reach the state $w + r \cdot \nabla f(w)/\|\nabla f(w)\|$ and calculating its gradient; 2) performing a gradient descent step with this gradient at state w. General SAM applies a standard numerical optimizer such as the Stochastic Gradient Descent (SGD) to approach the SAM objective (Foret et al., 2020).

3.2. Rethinking `FedSAM` Algorithm

`FedSAM` (Qu et al., 2022) focuses on improving *local generality* to achieve efficient training based on SAM optimizer for each client. According to equation (1) (2), by applying the SAM objective to replace the local ERM objective:

$$\min_{w}\left\{f^s(w) = \frac{1}{m}\sum_{i\in[m]} f_i^s(w)\right\},$$
$$f_i^s(w) \triangleq \max_{\|s_i\|\le r} f_i(w+s_i), \qquad (5)$$

where s_i is the local perturbation allocated to f_i.

Its core idea is similar to personalized Federated Learning (pFL) (Tan et al., 2022; Kulkarni et al., 2020), which concerns more about improving the performance on the local dataset. Intuitively, the improvement of each client could promote better performance. Global Aggregation of local states with smoother local loss landscape helps boost the flatness of the global model. According to the formula (4):

$$\nabla f^s(w) \approx \frac{1}{m}\sum_{i\in[m]} \nabla f_i(w)|_{w+r\cdot\nabla f_i(w)/\|\nabla f_i(w)\|}. \qquad (6)$$

The calculation of the global gradient can be easily split to each local client i, which allows `FedSAM` to directly enable the periodic aggregation from local training, like `FedAvg`. However, this indirect promotion also has its own limitations. Since the objective of equation (5) is not for the global function f, the flatness of the global landscape can not be directly optimized, which makes the whole process full of uncertainty. We illustrate a situation with the simple schematic shown in Figure 1. Obviously, when the local consistency maintains a high level, which means the local objectives are close to the global target, `FedSAM` effectively improves the smoothness of the global loss landscape like Figure 1 (a). However, when the consistency drops as shown in Figure 1 (b), though it contributes to reducing the loss value to a certain extent, the global loss surface is still sharp. An important reason is that the effective range of SAM

(a) High consistency.   (b) Low consistency.
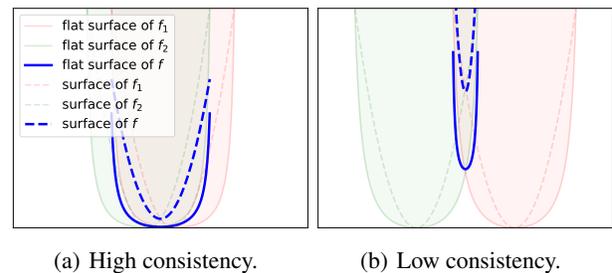
Figure 1. A toy schematic to introduce a bad case of `FedSAM`. We assume $m = 2$ and $f = (f_1 + f_2)/2$. The dotted lines represent the general loss surface trained by SGD, and the solid lines represent the flat loss surface trained by SAM. The red, green, and blue correspond to the local f_1, f_2, and global f, respectively.

usually cannot cover the total surface, and its performance drops significantly far away from the optima. Therefore, simply considering improving the local generality still faces very large limitations, e.g. rugged inconsistent local optimal under the severely heterogeneous dataset in practice.

4. Methodology

To tackle the difficulties and improve the performance of the global function, we propose the `FedSMOO` which jointly considers both consistency and a global flat landscape. Then, we give convergence and generalization analysis.

4.1. `FedSMOO` Algorithm

In contrast to `FedSAM` (Qu et al., 2022), we focus on adopting the SAM objective on the global function as follows:

$$\min_{w} \left\{ \mathcal{F}(w) = \frac{1}{m} \sum_{i \in [m]} \mathcal{F}_i(w) \right\},$$
$$\mathcal{F}_i(w) \triangleq \max_{\|s\| \leq r} f_i(w + s), \tag{7}$$

where s is the global perturbation allocated to f.

In FL, due to the data privatization and localization, gradient calculations can only cover the local dataset of each client i. If directly adopting the equation (7), it requires the global gradient ∇f as mentioned in (4). This is impossible for the local clients. Therefore, we transform the objective (7) into the following form by adding two constraints on w and s:

$$\min_{w_i = w} \left\{ \mathcal{F} = \frac{1}{m} \sum_{i \in [m]} \mathcal{F}_i(w, w_i, s, s_i) \right\},$$
$$\mathcal{F}_i(w, w_i, s, s_i) \triangleq \max_{\substack{\|s_i\| \leq r, \\ s_i = s}} f_i(w_i + s_i). \tag{8}$$

Under the constraints of w_i and s_i, the objective (8) and (7) are equivalent. Meanwhile, the parameters w_i and perturbation s_i could be easily split to each local client, which supports the federated calculations. This target guarantees the global consensus of w_i and s_i, and could search for the consistent flat minima after completing the optimization.

4.1.1. RE-APPROXIMATING THE INNER MAXIMIZATION

Like the general SAM, instead of solving this maximization problem rigorously, we derive an effective approximation of its solution. By taking the first-order Taylor expansion of $f_i(w_i + s_i)$, we define the local augmented Lagrangian function via penalizing the constrain $s_i = s$ as the linear and quadratic regularization terms on the objective as follows:

$$\mathcal{L}_i^s : f_i(w_i) + s_i^\top \nabla f_i(w_i) + \mu_i^\top (s - s_i) + \frac{1}{2\alpha} \|s - s_i\|^2, \tag{9}$$

where the local perturbation satisfies $\|s_i\| \leq r$.
To solve the maximization, we adopt a similar pattern as

ADMM to alternately update the local perturbation s_i, the dual variable μ_i, and global perturbation s. Firstly, we have:

$$\hat{s}_i = \arg\max_{\|s_i\| \leq r} \left\{ s_i^\top (\nabla f_i(w_i) - \mu_i) + \frac{1}{2\alpha} \|s_i - s\|^2 \right\}$$
$$= \arg\max_{\|s_i\| \leq r} \left\{ \frac{1}{2\alpha} \|s_i + \jmath_i\|^2 \right\}, \tag{10}$$

where $\jmath_i = \alpha (\nabla f_i(w_i) - \mu_i) - s$. Thereby, $\hat{s}_i = r\jmath_i / \|\jmath_i\|$. Then we update the dual variable as $\mu_i = \mu_i + \frac{1}{\alpha}(\hat{s}_i - s)$. To maximize the global perturbation s, we have the following:

$$\hat{s} = \arg\max_{\|s\| \leq r} \frac{1}{m} \sum_{i \in [m]} \left\{ s^\top \mu_i + \frac{1}{2\alpha} \|s - \hat{s}_i\|^2 \right\}$$
$$= \arg\max_{\|s\| \leq r} \left\{ \frac{1}{2\alpha} \frac{1}{m} \sum_{i \in [m]} \|s + \alpha \mu_i - \hat{s}_i\|^2 \right\}. \tag{11}$$

Let $\jmath = \frac{1}{m} \sum_{i \in [m]} (\alpha \mu_i - \hat{s}_i)$. Thereby, $\hat{s} = r\jmath / \|\jmath\|$.
With these update rules, we can calculate the local perturbation of approximately satisfying the constraints. Due to the data privatization and localization mentioned above, we can not directly get the $s^\star$ during the local training as its requirement of $s_i^\star$ for $i \in [m]$. Therefore, we iterate the global perturbation on the global server and communicate it to the active clients at the new round. At the local training process, compared to the vanilla `FedSAM`, our proposed `FedSMOO` adopts the new perturbation vector corrected by the global estimation, which gradually approaches the global perturbation during the training. Then, according to the formula (4), we get the new corrected gradient to perform optimization.

4.1.2. GLOBAL CONSISTENCY AND THE MINIMIZATION

As shown in Figure 1, to productively avoid the performance dropping and further improve the consistency, we also adopt a dynamic regularizer (Acar et al., 2021) on each local client via a similar pattern as ADMM, to efficiently minimize the global objective $\mathcal{F}$. Also, we penalize the $w_i = w$ constrain and introduce the global augmented Lagrangian function as:

$$\mathcal{L} : \frac{1}{m} \sum_i \left\{ \mathcal{F}_i + \lambda_i^\top (w^t - w_i) + \frac{1}{2\beta} \|w^t - w_i\|^2 \right\}. \tag{12}$$

We split the finite sum problem to each local client and solve each sub-problem in one communication round for several local updates. In each sub-problem, we firstly minimize the local parameters w_i in the augmented Lagrangian function,

$$w_{i,K}^t = \arg\min_{w_i} \left\{ \mathcal{F}_i - \lambda_i^\top w_i + \frac{1}{2\beta} \|w^t - w_i\|^2 \right\}. \tag{13}$$

In order not to affect the performance of the vanilla SAM, we adopt the SGD to solve this problem. Then we update the dual variable as $\lambda_i = \lambda_i - \frac{1}{\beta}(w_{i,K}^t - w^t)$. At the end of each round, we minimize the global variable w^t from the function (12) and set it as w^{t+1} to start the next round.

Algorithm 1 `FedSMOO` Algorithm

Input: global model w, local model w_i, communication round T, local interval K, local perturbation s_i, global perturbation s, dual variable for parameters λ_i, global dual variable λ, dual variable for perturbation μ_i, penalized coefficient for the quadratic term α, β.

Output: global model w^T.

1: **Initialization** : $w = w^0$, $\lambda_i = \lambda = 0$, $\mu_i = 0$.
2: **for** $t = 0, 1, 2, \cdots, T-1$ **do**
3:      randomly select the active clients set $[n]$ from $[m]$
4:      **for** client $i \in [n]$ **in parallel do**
5:          send the w^t, s^t to the active clients as $w_{i,0}^t, s$
6:          **for** $k = 0, 1, \cdots, K-1$ **do**
7:              get the stochastic gradient $g_{i,k}^t$ at $w_{i,k}^t$
8:              $\hat{s}_{i,k}^t = r s_{i,k}^t / \|s_{i,k}^t\|$ for $s_{i,k}^t = (g_{i,k}^t - \mu_i) - s$
9:              $\mu_i = \mu_i + (\hat{s}_{i,k}^t - s)$
10:         get the stochastic gradient $\hat{g}_{i,k}^t$ at $w_{i,k}^t + \hat{s}_{i,k}^t$
11:              $w_{i,k+1}^t = w_{i,k}^t - \eta \left[\hat{g}_{i,k}^t - \lambda_i + \frac{1}{\beta}(w_{i,k}^t - w^t) \right]$
12:          **end for**
13:          $\widetilde{s}_i = \mu_i - \hat{s}_{i,K}^t$
14:          $\lambda_i = \lambda_i - \frac{1}{\beta}(w_{i,K}^t - w^t)$
15:          send the $w_i^t = w_{i,K}^t, \widetilde{s}_i$ to the global server
16:      **end for**
17:      $s^t = r s^t / \|s^t\|$ for $s^t = \frac{1}{n} \sum_{i \in [n]} \widetilde{s}_i$
18:      $\lambda^{t+1} = \lambda^t - \frac{1}{\beta m} \sum_{i \in [n]} (w_i^t - w^t)$
19:      $w^{t+1} = \frac{1}{n} \sum_{i \in [n]} w_i^t - \beta \lambda^{t+1}$
20: **end for**

4.1.3. OVERVIEW OF FEDSMOO ALGORITHM

Algorithm 1 shows our detailed implementation flow. At round t, we randomly select a sub-set of active clients $[n]$ from the total clients set $[m]$, and send the current global model w^t and global perturbation estimation s^t to all active clients. Line.8 computes the corrected local perturbation according to equation (10) with $\alpha = 1$. Line.9 updates the dual variable for the local perturbation with the same $\alpha = 1$. Line.10 computes the new corrected SAM gradient and Line.11 updates the local model w_i with SGD according to equation (13). Line.14 updates the local dual variable for the local model. After finishing the local training process, we send the required variables w_i^t and $\widetilde{s}_i$ to the global server for aggregation. Line.17 computes the new global perturbation estimation according to the equation (11). Line.18 and Line.19 update the global model w^{t+1} by minimizing the function (12) on the w^t. The global updates repeat T rounds.

4.2. Theoretical Analysis

In this part, we demonstrate the theoretical analysis of our proposed `FedSMOO` Algorithm. We take into account the smooth non-convex case and prove its convergence rate can achieve fast $\mathcal{O}(1/T)$ with a general generalization bound.

4.2.1. CONVERGENCE RATE

Firstly we state some general assumptions in this work.

Assumption 1. Function $f_i(w)$ is L-smooth for all $i \in [m]$, i.e., $\|\nabla f_i(x) - \nabla f_i(y)\| \leq L \|x - y\|$, for all $x, y \in \mathbb{R}^d$.

Assumption 2. Stochastic gradient $g_{i,k}^t = \nabla f_i(w_{i,k}^t, \varepsilon_{i,k}^t)$ with randomly sampled data $\varepsilon_{i,k}^t$ is an unbiased estimator of $\nabla f_i(w_{i,k}^t)$ with bounded variance, i.e., $\mathbb{E}[g_{i,k}^t] = \nabla f_i(w_{i,k}^t)$ and $\mathbb{E}\|g_{i,k}^t - \nabla f_i(w_{i,k}^t)\|^2 \leq \sigma_l^2$, for all $\mathbf{x}_{i,k}^t \in \mathbb{R}^d$.

Assumption 1 assumes the Lipschitz continuity of the gradients and Assumption 2 assumes the bounded stochastic properties of the gradients. The two general assumptions are widely used in the analysis of the FL framework (McMahan et al., 2017; Reddi et al., 2020; Karimireddy et al., 2020; Xu et al., 2021; Acar et al., 2021; Qu et al., 2022). In addition to these, `FedSAM` requires a tighter bounded variance of the stochastic gradient and the assumption of bounded heterogeneity (Qu et al., 2022). In our work, we refer to the approach of Zhang et al. (2021); Acar et al. (2021); Wang et al. (2022); Gong et al. (2022) and do not require the extra assumptions. Proof details can be referred to the *Appendix*.

Theorem 4.1. *Let the assumptions hold, let the size of the active clients' set $|[n]| = n$, and similarly, $|[m]| = m$, let $r \leq \frac{4\kappa_r}{\sqrt{nT}}$ where $\kappa_r \in \mathbb{R}$ is a constant, and let $\beta \leq \frac{\sqrt{n}}{6\sqrt{6m}L}$, the sequence $\left\{ \overline{w}^{t+1} \triangleq \frac{1}{n} \sum_{i \in [n]} w_i^t \right\}_{t \in [T-1]}$ generated by the Algorithm 1 under the non-convex case satisfy:*

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\|\nabla f(\overline{w}^t)\|^2$$
$$\leq \frac{1}{\zeta \beta T} \left[\kappa_f + \frac{1}{n} 3\beta L \kappa_r + \frac{m}{n} 72 \beta^2 L^2 \delta^0 \right], \tag{14}$$

where $\zeta \in \left(0, \frac{1}{2}\right)$ is a constant, $\kappa_f \triangleq f(\overline{w}^1) - f^\star$ for $f^\star$ is the optima of f, $\delta^0 \triangleq \frac{1}{m} \sum_{i \in [m]} \mathbb{E}\|w_i^0 - \overline{w}^1\|^2$ is the inconsistent term at the first round for $\overline{w}^1 \triangleq \frac{1}{n} \sum_{i \in [n]} w_i^0$.

Remark 4.2. Under the participation ratio equal to n/m, our proposed `FedSMOO` achieves fast $\mathcal{O}(1/T)$ convergence rate, which matches the conclusion of existing works (Zhang et al., 2021; Acar et al., 2021; Wang et al., 2022; Gong et al., 2022). κ_f term indicates the impact of the initial state w^0. κ_r term comes from the corrected SAM steps. δ^0 term is the variance of the local parameters w_i optimized in the first round, which indicates the inconsistency level from the heterogeneity. The latter two achieve $n\times$ linear speedup.

Remark 4.3. A key property is that local models will converge to the global model after training. With reference to the analysis of Acar et al. (2021); Wang et al. (2022); Gong et al. (2022) and consider the first-order gradient condition of equation (13) as $\nabla \mathcal{F}_i(w_i^t) - \lambda_i + \frac{1}{\beta}(w_i^t - w^t) = 0$. And, considering the update of $\lambda_i = \lambda_i - \frac{1}{\beta}(w_i^t - w^t)$, we infer

$\lambda_i = \nabla \mathscr{F}_i(w_i^t)$ after updating at round t. Thus, rethinking the first-order gradient condition, we have the relationship of $\nabla \mathscr{F}_i(w_i^t) - \nabla \mathscr{F}_i(w_i^{t-1}) + \frac{1}{\beta}(w_i^t - w^t) = 0$. When $t \to \infty$, the local model $w_i^t \to w_i^\infty = w_i^{\infty-1}$ converges to the local state. Thus, we have $w_i^\infty \to w^\infty$. Furthermore, as shown in Theorem 4.1, $\overline{w}^t$ converges with the rate of $\mathcal{O}(1/t)$, which implies $\overline{w}^\infty = \frac{1}{n} \sum_{i \in [n]} w_i^\infty \to w^\infty$. This matches the conclusion in (Acar et al., 2021) and guarantees that all local objectives converge to the consistent global objective.

4.2.2. GENERALIZATION BOUND

Based on the margin generalization bounds in (Neyshabur et al., 2017; Farnia et al., 2018), we consider the generalization error bound on the global function f as following:

$$G_\epsilon^f \triangleq \mathbf{P}\left(f(w+s,\varepsilon)[y] \le \max_{j \ne y} f(w+s,\varepsilon)[j] + \epsilon\right), \quad (15)$$

where ε is the input data and y is its ground truth. If $\epsilon = 0$, we denote the equation (15) as G^f to represent for the average misclassification rate under the global perturbation s. We define the $\widetilde{G}_\epsilon^f$ as the empirical estimate of the above expected margin loss trained on the heterogeneous dataset.

Different from the FedSAM (Qu et al., 2022), which focuses on the *averaged local generality* and define the error bound: $\frac{1}{m} \sum_i \mathbf{P}_i(f_i(w+s_i,\varepsilon)[y] \le \max_{j \ne y} f_i(w+s_i,\varepsilon)[j] + \epsilon)$. This split margin loss cannot describe the detailed characteristics of the global function f. To directly explore the stability of the global function, we adopt equation (15) as the margin $\mathbf{P}(\frac{1}{m} \sum_i f_i(w+s,\varepsilon)[y] \le \max_{j \ne y} \frac{1}{m} \sum_i f_i(w+s,\varepsilon)[j] + \epsilon)$.

Theorem 4.4. *We assume the input data ε as the normalized tensor, e.g. for an image, with the norm of L_n. And we consider a L_l-layer neural network with d parameters per layer. The activation functions hold no biases and bounded 1-Lipschitz property. The input data will be bounded by a positive constant L_w and its total size equal to D. Under the margin value of ϵ and a positive value p, with the probability of at least $1 - p$, the empirical generalization risk on the global parameters w generated by FedSMOO satisfies:*

$$G^f \le \widetilde{G}_\epsilon^f + \mathcal{O}\left(\sqrt{\frac{L_l^2 L_n^2 d \ln(dL_l)\mathcal{V}_L + \ln\frac{L_l D}{p}}{(D-1)\epsilon^2}}\right), \quad (16)$$

where $\mathcal{V}_L = \prod_{l=1}^L \|w_l\|^2 \sum_{l=1}^L \frac{\|w_l\|_F^2}{\|w_l\|^2}$.

FedSMOO focuses on searching for the global consistent flat minima aligned to each local client, which theoretically guarantees the global consistency. Theorem 4.4 indicates the upper bound of the generalization risk on the global parameters w of the global function f under SAM perturbation s, which guarantees the flat landscape of the global objective. Due to the space limitation, detailed proofs can be referred to the *Appendix C*.

5. Experiments

In this section, we firstly introduce our experimental implementation details, including the introduction of the baselines, the backbone, dataset splitting, and hyperparameters selection. Next, we state the evaluation of FedSMOO and study its hyperparameters' sensitivity and ablation test.

5.1. Implementation Details

Baselines: We compare the FedSMOO with several benchmarks: FedAvg (McMahan et al., 2017) firstly introduce the federated framework via partial participation and local multiple training; FedAdam (Reddi et al., 2020) implement the global adaptive optimizer; SCAFFOLD (Karimireddy et al., 2020) utilize the SVRG (Johnson & Zhang, 2013) to diminish the client drift; FedCM (Xu et al., 2021) maintain the consistency of local updates with a momentum term; FedDyn (Acar et al., 2021) guarantee the local collective objectives through the local dynamic regularization; FedSAM and its variant MoFedSAM (Qu et al., 2022) employ the local SAM (Foret et al., 2020) optimizer with momentum to reduce the generalized divergence drift. All methods are evaluated on ResNet-18 (He et al., 2016) implemented in PyTorch (Paszke et al., 2019) official model-zoo. We follow the suggestion of Hsieh et al. (2020) to replace the Batch Normalization with the Group Normalization (Wu & He, 2018) to avoid the non-differentiable parameters.

Hyperparameter selections: In the interests of the fairness, we freeze the general hyperparameters for all baselines on the same setups, including the local learning rate equal to 0.1, the global learning rate equal to 1 expect for FedAdam which adopts 0.1 for the global adaptivity, the perturbation learning rate equal to 0.1 expect for FedSAM and MoFedSAM which adopt 0.01, the weight decay equal to $1e$-3, and the learning rate decreasing by $0.998\times$ exponentially except for FedDyn and FedSMOO which adopt $0.9995\times$ for the proxy term. On the CIFAR-10, we take the batchsize equal to 50, and the local epochs equal to 5. On the CIFAR-100, we adjust the batchsize equal to 20, and the local epochs equal to 2 to avoid the extreme overfitting. More details of the selections can be found in *Appendix B*.

Dataset splitting: We conduct the experiments on CIFAR-10/100 (Krizhevsky et al., 2009) and then follow the Hsu et al. (2019) to partition the dataset via widely used Dirichlet and Pathological sampling. The coefficient u is adopted to control the variance, which measures the heterogeneity.

Additionally, to further enhance the heterogeneity and randomness, we enable the sampling *with replacement* when generating the non-iid dataset. Most of the previous works only consider the local heterogeneity and sample the data *without replacement*, which makes the distribution of the overall samples across the category uniform. In practice,

Table 1. Test accuracy comparison among baselines and our proposed method on the CIFAR-10/100 dataset after 800 rounds. The dataset splitting method is selected from the Dirichlet sampling with replacement and Pathological partition (only a few random categories are enabled for sampling on a local client). The experimental setups are 10%-100 clients (upper part) and 5%-200 clients (lower part) respectively. "u" represents the Dirichlet coefficient which is selected from $[0.1, 0.6]$, and "c" represents the number of active categories on each client which is selected from $[3, 6]$ on CIFAR-10 and $[10, 20]$ on CIFAR-100. Each result is calculated by 2 times.

ALGORITHM	CIFAR-10				CIFAR-100			
	DIRICHLET		PATHOLOGICAL		DIRICHLET		PATHOLOGICAL	
	$u = 0.6$	$u = 0.1$	$c = 6$	$c = 3$	$u = 0.6$	$u = 0.1$	$c = 20$	$c = 10$
FEDAVG	$79.52_{\pm.13}$	$76.00_{\pm.18}$	$79.91_{\pm.17}$	$74.08_{\pm.22}$	$46.35_{\pm.15}$	$42.64_{\pm.22}$	$44.15_{\pm.17}$	$40.23_{\pm.31}$
FEDADAM	$77.08_{\pm.31}$	$73.41_{\pm.33}$	$77.05_{\pm.26}$	$72.44_{\pm.29}$	$48.35_{\pm.17}$	$40.77_{\pm.31}$	$41.26_{\pm.30}$	$32.58_{\pm.22}$
SCAFFOLD	$81.81_{\pm.17}$	$78.57_{\pm.14}$	$83.07_{\pm.10}$	$77.02_{\pm.18}$	$51.98_{\pm.23}$	$44.41_{\pm.15}$	$46.06_{\pm.22}$	$41.08_{\pm.24}$
FEDCM	$82.97_{\pm.21}$	$77.82_{\pm.16}$	$83.44_{\pm.17}$	$77.82_{\pm.19}$	$51.56_{\pm.20}$	$43.03_{\pm.26}$	$44.94_{\pm.14}$	$38.35_{\pm.27}$
FEDDYN	$83.22_{\pm.18}$	$78.08_{\pm.19}$	$83.18_{\pm.17}$	$77.63_{\pm.14}$	$50.82_{\pm.19}$	$42.50_{\pm.28}$	$44.19_{\pm.19}$	$38.68_{\pm.14}$
FEDSAM	$80.10_{\pm.12}$	$76.86_{\pm.16}$	$80.80_{\pm.23}$	$75.51_{\pm.24}$	$47.51_{\pm.26}$	$43.43_{\pm.12}$	$45.46_{\pm.29}$	$40.44_{\pm.23}$
MoFEDSAM	$84.13_{\pm.13}$	$78.71_{\pm.15}$	$84.92_{\pm.14}$	$79.57_{\pm.18}$	$\mathbf{54.38}_{\pm.22}$	$44.85_{\pm.25}$	$47.42_{\pm.26}$	$41.17_{\pm.22}$
OUR	$\mathbf{84.55}_{\pm.14}$	$\mathbf{80.82}_{\pm.17}$	$\mathbf{85.39}_{\pm.21}$	$\mathbf{81.58}_{\pm.16}$	$53.92_{\pm.18}$	$\mathbf{46.48}_{\pm.13}$	$\mathbf{48.87}_{\pm.17}$	$\mathbf{44.10}_{\pm.19}$
FEDAVG	$75.90_{\pm.21}$	$72.93_{\pm.19}$	$77.47_{\pm.34}$	$71.86_{\pm.34}$	$44.70_{\pm.22}$	$40.41_{\pm.33}$	$38.22_{\pm.25}$	$36.79_{\pm.32}$
FEDADAM	$75.55_{\pm.38}$	$69.70_{\pm.32}$	$75.24_{\pm.22}$	$70.49_{\pm.26}$	$44.33_{\pm.26}$	$38.04_{\pm.25}$	$35.14_{\pm.16}$	$30.28_{\pm.28}$
SCAFFOLD	$79.00_{\pm.26}$	$76.15_{\pm.15}$	$80.69_{\pm.21}$	$74.05_{\pm.31}$	$50.70_{\pm.18}$	$41.83_{\pm.29}$	$39.63_{\pm.31}$	$37.98_{\pm.36}$
FEDCM	$80.52_{\pm.29}$	$77.28_{\pm.22}$	$81.76_{\pm.24}$	$76.72_{\pm.25}$	$50.93_{\pm.31}$	$42.33_{\pm.19}$	$42.01_{\pm.17}$	$38.35_{\pm.24}$
FEDDYN	$80.69_{\pm.23}$	$76.82_{\pm.17}$	$82.21_{\pm.18}$	$74.93_{\pm.22}$	$47.32_{\pm.18}$	$41.74_{\pm.21}$	$41.55_{\pm.18}$	$38.09_{\pm.27}$
FEDSAM	$76.32_{\pm.16}$	$73.44_{\pm.14}$	$78.16_{\pm.27}$	$72.41_{\pm.29}$	$45.98_{\pm.27}$	$40.22_{\pm.27}$	$38.71_{\pm.23}$	$36.90_{\pm.29}$
MoFEDSAM	$82.58_{\pm.21}$	$78.43_{\pm.24}$	$84.46_{\pm.20}$	$79.93_{\pm.19}$	$\mathbf{53.51}_{\pm.25}$	$42.22_{\pm.23}$	$42.77_{\pm.27}$	$39.81_{\pm.21}$
OUR	$\mathbf{82.94}_{\pm.19}$	$\mathbf{79.76}_{\pm.19}$	$\mathbf{84.82}_{\pm.18}$	$\mathbf{81.01}_{\pm.19}$	$53.45_{\pm.19}$	$\mathbf{45.83}_{\pm.18}$	$\mathbf{44.70}_{\pm.21}$	$\mathbf{43.41}_{\pm.22}$
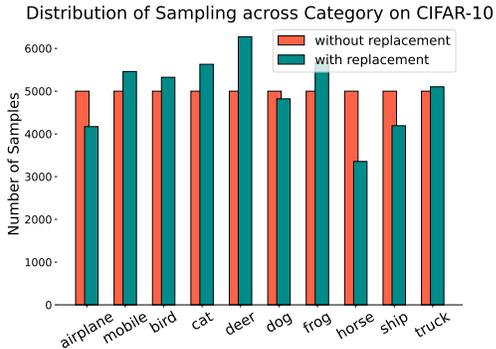
Figure 2. Distribution across category on CIFAR-10 of sampling with/without replacement under the Dirichlet coefficient $u = 0.1$ and the number of total clients $m = 100$. The standard deviation of the samples' number approaches 829.14, which highly increases their imbalance and properly approximates the practical scenes.

due to the privacy protection policy, the distribution is usually disorderly and nonuniform. Some SOTA methods have significant performance degradation unanimously on this rugged dataset. Adopting sampling *with replacement* breaks the balance of the raw data, which approaches more in line with real-world scenarios. Figure 2 demonstrates the difference between these two setups in the experiments.

5.2. Experimental Evaluation

Comparison with the baselines: As shown in Table 1, our proposed FedSMOO performs well with good stability and effectively resists the negative impact of the strong hetero-

geneous dataset. Specifically, on the CIFAR-10 dataset, FedSMOO achieves 80.82% on the Dirichlet-0.1 setups, which is 2.11% higher than the second highest accuracy. On the CIFAR-100 dataset, the improvement achieves 1.63% from the MoFedSAM. Improvements of the vanilla FedSAM on both CIFAR10 / 100 are very limited, which is only about 1% over than FedAvg on average. While its variant MoFedSAM achieves about 5% improvement over the FedSAM on average, which benefits from the momentum updates during the local training. FedSMOO focuses on both the optimization and global generalization performance, and searches for the consistent flat landscape, which efficiently converges to a better minimum and is significantly ahead of the MoFedSAM algorithm, especially on the severely heterogeneous dataset with inconsistent local optima.

Impact of heterogeneity: Via sampling *with replacement*, not only is there a data imbalance between local clients, but also the number of samples among categories in the global dataset is also very different, as shown in Figure 2. With the larger heterogeneity, several baselines are greatly affected. We select the two methods to split the data, the Dirichlet and the Pathological. In the Dirichlet, we select the variance coefficient equal to 0.1 and 0.6 (the results of 0.6 and iid are very close). In detail, on CIFAR-10, when the coefficient decreases from 0.6 to 0.1, MoFedSAM drops from 84.13% to 78.71%, while FedSMOO drops about only 3.73% to 80.82%, which maintains high accuracy. On CIFAR-100, MoFedSAM achieves a little higher accuracy on Dirichlet-0.6 low heterogeneous dataset, while it severely drops

(a) FedAvg $v.s$ FedSMOO   (b) SCAFFOLD $v.s$ FedSMOO   (c) MoFedSAM $v.s$ FedSMOO
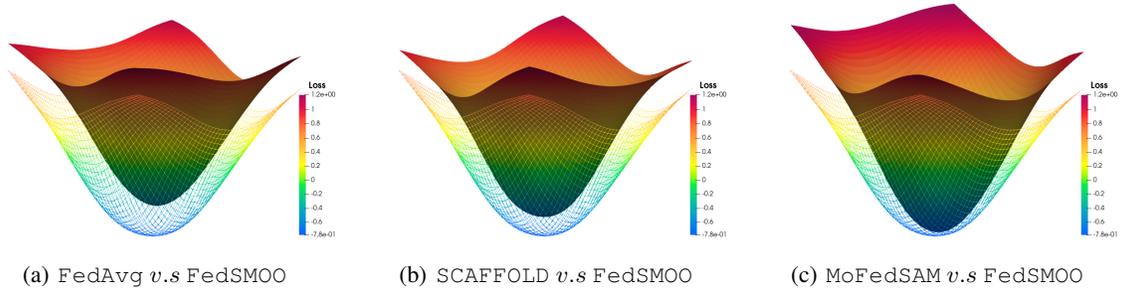
Figure 3. Visualization of the loss landscape of ResNet-18 backbone trained via FedAvg, SCAFFOLD, MoFedSAM and FedSMOO on the CIFAR-10 dataset. For clarity, we use the grid surface on the FedSMOO and compare it with the other three baselines separately. FedSMOO could approach a more general and flat loss landscape which efficiently improves the generalization performance in FL.

(a) Different local interval.   (b) Selection of lr-decay.   (c) Selection of β.   (d) Selection of r.
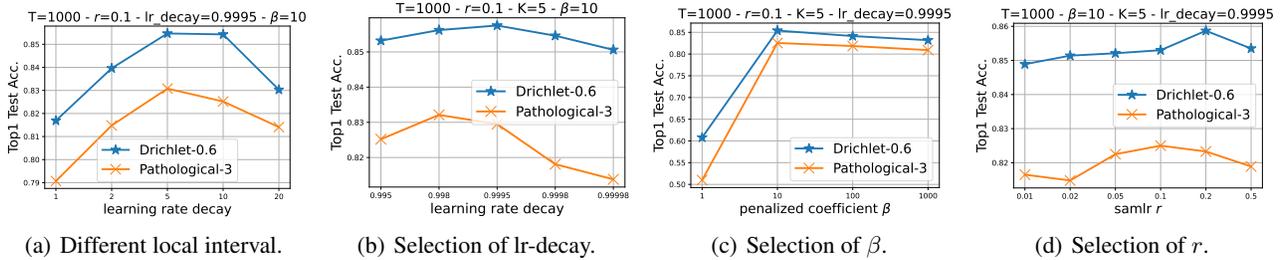
Figure 4. Hyperparameters sensitivity studies of local intervals, learning rate decay, penalized coefficient β and SAM-lr r on CIFAR-10.

about 9.53% when the non-iid level decreases to Dirichlet-0.1. FedSMOO works well on the high heterogeneity and achieves 1.63% ahead of the second-highest accuracy. Furthermore, we test the Pathological split by blocking certain types on a client and only sampling from the total random c categories, which is a higher level of heterogeneity. On the pathological dataset with only 3 categories per client, FedSMOO is 2.01% ahead of the MoFedSAM on the 10% participation and 1.08% on 5% participation. It achieves good performance on the strong heterogeneous dataset with high stability on different setups.

Impact of partial participation: Another important option is the partial participation ratio. To fairly compare with the baselines, we froze the other hyperparameters selections. When the participation ratio decreases from 10% to 5% on CIFAR-10, FedSMOO still achieves excellent performance, which drops from 80.82% to 79.76% on Dirichlet-0.1, and from 81.58% to 81.01% on Pathological-3. On the CIFAR-100 experiments, FedSMOO maintains a high level of generalization on Pathological-10, which achieves 44.10%.

Loss landscape: Figure 3 shows the visualization of the loss landscape. All models are trained on CIFAR-10 with the Dirichlet-0.1 setup. We compare the FedAvg, SCAFFOLD, MoFedSAM and FedSMOO algorithms. For a clear comparison, we superimpose the landscape of the two algorithms together and apply the grid surface on FedSMOO. It can be seen that the global minima approached by our method has not only a lower loss value but also a smoother loss

landscape, than the other three benchmarks.

Hyperparameters Sensitivity: We study the hyperparameters' sensitivity of local interval, learning rate decay, penalized coefficient β, and SAM learning rate r. FedSMOO is stable on the changes in learning rate decay and r. The selection of local intervals needs to be long enough to approximate the solution of Equation (13) as mentioned in Acar et al. (2021). The selection can be fine-tuned appropriately for the gain. The penalized coefficient β can not be selected too small, which will cause the proxy term to severely interfere with the direction of gradient descent.

Hessian: To further illustrate the efficiency of improving the generalization, we test the Hessian matrix of the final models. We test ResNet-18-GN on the CIFAR-10 dataset with the setup of 10%-100 clients under the Dirichlet sampling coefficient equal to 0.1 and 0.6 (the same hyperparameter selection in Table 1. The results are shown in Table 2. It is clear to see that the proposed FedSMOO approaches the flat minimal with a lower Hessian top eigenvalue than the FedSAM method. The Hessian trace also decreases effec-

Table 2. Consistency and Hessian matrix.

method	FedSAM		FedSMOO	
Dirichlet	0.6	0.1	0.6	0.1
$\frac{1}{n}\sum_i \|w_i^t - w^t\|^2$	0.866	1.245	0.821	1.061
Hessian Top Eigenvalue	142.65	177.18	91.46	107.44
Hessian Trace	3104.1	3842.3	1783.3	2689.4

tively. The divergence term $\frac{1}{n}\sum_i \|w_i^t - w^{t+1}\|^2$ maintains similar properties with the flatness in our experiments.

Communication Cost: One of the major concerns in FL is the communication bottleneck. Though some advanced algorithms achieve higher performance, they have to communicate more data to support the training process. This is undoubtedly a very large limitation in the FL paradigm. Therefore, we compare the communication costs for the algorithms in our paper. For convenience, we assume `FedAvg` communicates total V bits in the training process. To achieve 74% on CIFAR-10 of Pathological-3 splitting, the total communication costs are stated in Table 3.

Table 3. Total communication costs.

method	rounds	communication costs
FedAvg	723 (1×)	V
FedAdam	-	-
SCAFFOLD	533 (1.36×)	1.47 V
FedCM	442 (1.64×)	1.22 V
FedDyn	424 (1.71×)	0.59 V
FedSAM	604 (1.20×)	0.84 V
MoFedSAM	298 (2.43×)	0.82 V
FedSMOO	194 (3.73×)	0.53 V

From this table, we clearly see that advanced methods, i.e. `SCAFFOLD` and `FedCM`, suffer from very high communication costs for transferring double vectors per round. Though our proposed `FedSMOO` requires double as well, it focuses on improving the generalization performance in FL and effectively reduces the total communication costs.

6. Conclusion

In this work, we propose a novel and practical federated algorithm `FedSMOO` which jointly considers the optimization and generalization targets via adopting the dynamic regularizer to guarantee the local optima towards the global objective revised by the global SAM optimizer. It efficiently searches for a consistent flat minimum in the FL framework. Theoretical analysis guarantees that `FedSMOO` achieves the fast convergence rate of $\mathcal{O}(1/T)$. Furthermore, we also provide the global generalization bound. We conduct extensive experiments to verify its efficiency and significant performance on the global model under severe heterogeneity.

References

Acar, D. A. E., Zhao, Y., Navarro, R. M., Mattina, M., Whatmough, P. N., and Saligrama, V. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021.

Asad, M., Moustafa, A., and Ito, T. Fedopt: Towards communication efficiency and privacy preservation in federated learning. *Applied Sciences*, 10(8):2864, 2020.

Brownlees, C., Joly, E., and Lugosi, G. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507–2536, 2015.

Caldarola, D., Caputo, B., and Ciccone, M. Improving generalization in federated learning by seeking flat minima. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIII*, pp. 654–672. Springer, 2022.

Charles, Z. and Konečný, J. Convergence and accuracy trade-offs in federated learning and meta-learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2575–2583. PMLR, 2021.

Das, R., Acharya, A., Hashemi, A., Sanghavi, S., Dhillon, I. S., and Topcu, U. Faster non-convex federated learning via global and local momentum. In *Uncertainty in Artificial Intelligence*, pp. 496–506. PMLR, 2022.

Farnia, F., Zhang, J. M., and Tse, D. Generalizable adversarial training via spectral normalization. *arXiv preprint arXiv:1811.07457*, 2018.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Gong, Y., Li, Y., and Freris, N. M. Fedadmm: A robust federated deep learning framework with adaptivity to system heterogeneity. *arXiv preprint arXiv:2204.03529*, 2022.

Haddadpour, F., Kamani, M. M., Mokhtari, A., and Mahdavi, M. Federated learning with compression: Unified analysis and sharp guarantees. In *International Conference on Artificial Intelligence and Statistics*, pp. 2350–2358. PMLR, 2021.

Hanzely, F. and Richtárik, P. Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*, 2020.

Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., and Ramage, D. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hochreiter, S. and Schmidhuber, J. Simplifying neural nets by discovering flat minima. *Advances in neural information processing systems*, 7, 1994.

Hsieh, K., Phanishayee, A., Mutlu, O., and Gibbons, P. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, pp. 4387–4398. PMLR, 2020.

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.

Huang, T., Lin, W., Shen, L., Li, K., and Zomaya, A. Y. Stochastic client selection for federated learning with volatile clients. *IEEE Internet of Things Journal*, 9(20): 20055–20070, 2022.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

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*, pp. 5132–5143. PMLR, 2020.

Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

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

Kulkarni, V., Kulkarni, M., and Pant, A. Survey of personalization techniques for federated learning. In *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, pp. 794–797. IEEE, 2020.

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*, pp. 5905–5914. PMLR, 2021.

Li, L., Fan, Y., Tse, M., and Lin, K.-Y. A review of applications in federated learning. *Computers & Industrial Engineering*, 149:106854, 2020a.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020b.

Liang, X., Shen, S., Liu, J., Pan, Z., Chen, E., and Cheng, Y. Variance reduced local sgd with lower communication complexity. *arXiv preprint arXiv:1912.12844*, 2019.

Liu, J., Huang, J., Zhou, Y., Li, X., Ji, S., Xiong, H., and Dou, D. From distributed machine learning to federated learning: A survey. *Knowledge and Information Systems*, pp. 1–33, 2022.

Malinovskiy, G., Kovalev, D., Gasanov, E., Condat, L., and Richtarik, P. From local sgd to local fixed-point methods for federated learning. In *International Conference on Machine Learning*, pp. 6692–6701. PMLR, 2020.

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*, pp. 1273–1282. PMLR, 2017.

Mi, P., Shen, L., Ren, T., Zhou, Y., Sun, X., Ji, R., and Tao, D. Make sharpness-aware minimization stronger: A sparsified perturbation approach. *arXiv preprint arXiv:2210.05177*, 2022.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.

Ozfatura, E., Ozfatura, K., and Gündüz, D. Fedadc: Accelerated federated learning with drift control. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pp. 467–472. IEEE, 2021.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Pathak, R. and Wainwright, M. J. Fedsplit: An algorithmic framework for fast federated optimization. *Advances in Neural Information Processing Systems*, 33:7057–7066, 2020.

Qu, Z., Li, X., Duan, R., Liu, Y., Tang, B., and Lu, Z. Generalized federated learning via sharpness aware minimization. *arXiv preprint arXiv:2206.02618*, 2022.

Reddi, S., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečnỳ, J., Kumar, S., and McMahan, H. B. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

Sahu, A. K., Li, T., Sanjabi, M., Zaheer, M., Talwalkar, A., and Smith, V. On the convergence of federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 3:3, 2018.

Shi, Y., Liu, Y., Wei, K., Shen, L., Wang, X., and Tao, D. Make landscape flatter in differentially private federated learning. *arXiv preprint arXiv:2303.11242*, 2023a.

Shi, Y., Shen, L., Wei, K., Sun, Y., Yuan, B., Wang, X., and Tao, D. Improving the model consistency of decentralized federated learning. *arXiv preprint arXiv:2302.04083*, 2023b.

Stich, S. U. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

Sun, H., Shen, L., Zhong, Q., Ding, L., Chen, S., Sun, J., Li, J., Sun, G., and Tao, D. Adasam: Boosting sharpness-aware minimization with adaptive learning rate and momentum for training deep neural networks. *arXiv preprint arXiv:2303.00565*, 2023a.

Sun, Y., Shen, L., Huang, T., Ding, L., and Tao, D. Fedspeed: Larger local interval, less communication round, and higher generalization accuracy. *arXiv preprint arXiv:2302.10429*, 2023b.

Tan, A. Z., Yu, H., Cui, L., and Yang, Q. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Tran Dinh, Q., Pham, N. H., Phan, D., and Nguyen, L. Feddr–randomized douglas-rachford splitting algorithms for nonconvex federated composite optimization. *Advances in Neural Information Processing Systems*, 34: 30326–30338, 2021.

Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, aug 2011. doi: 10.1007/s10208-011-9099-z. URL https://doi.org/10.1007%2Fs10208-011-9099-z.

Wang, H., Marella, S., and Anderson, J. Fedadmm: A federated primal-dual algorithm allowing partial participation. *arXiv preprint arXiv:2203.15104*, 2022.

Wang, J., Tantia, V., Ballas, N., and Rabbat, M. Slowmo: Improving communication-efficient distributed sgd with slow momentum. *arXiv preprint arXiv:1910.00643*, 2019.

Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020a.

Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020b.

Wang, J., Xu, Z., Garrett, Z., Charles, Z., Liu, L., and Joshi, G. Local adaptivity in federated learning: Convergence and consistency. *arXiv preprint arXiv:2106.02305*, 2021.

Woodworth, B., Patel, K. K., Stich, S., Dai, Z., Bullins, B., Mcmahan, B., Shamir, O., and Srebro, N. Is local sgd better than minibatch sgd? In *International Conference on Machine Learning*, pp. 10334–10343. PMLR, 2020.

Wu, Y. and He, K. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Xu, J., Wang, S., Wang, L., and Yao, A. C.-C. Fedcm: Federated learning with client-level momentum. *arXiv preprint arXiv:2106.10874*, 2021.

Yang, H., Fang, M., and Liu, J. Achieving linear speedup with partial worker participation in non-iid federated learning. *arXiv preprint arXiv:2101.11203*, 2021.

Yang, Q., Liu, Y., Cheng, Y., Kang, Y., Chen, T., and Yu, H. Federated learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 13(3):1–207, 2019.

Zhang, L., Shen, L., Ding, L., Tao, D., and Duan, L.-Y. 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*, pp. 10174–10183, 2022.

Zhang, X., Hong, M., Dhople, S., Yin, W., and Liu, Y. Fedpd: A federated learning framework with adaptivity to non-iid data. *IEEE Transactions on Signal Processing*, 69:6055–6070, 2021.

Zhao, Y., Zhang, H., and Hu, X. Penalizing gradient norm for efficiently improving generalization in deep learning. *arXiv preprint arXiv:2202.03599*, 2022.

Zhong, Q., Ding, L., Shen, L., Mi, P., Liu, J., Du, B., and Tao, D. Improving sharpness-aware minimization with fisher mask for better generalization on language models. *arXiv preprint arXiv:2210.05497*, 2022.

Zhuang, J., Gong, B., Yuan, L., Cui, Y., Adam, H., Dvornek, N., Tatikonda, S., Duncan, J., and Liu, T. Surrogate gap minimization improves sharpness-aware training. *arXiv preprint arXiv:2203.08065*, 2022.

A. Appendix

In this part, we provide the appendix supplementary materials including some experimental results and the proof of the main theorem.

B. Experiments

In this part, we present the results of some complementary experiments.

B.1. Introduction of the Dataset

Table 4. Dataset introductions.

Dataset	Training Data	Test Data	Class	Size
CIFAR-10	50,000	10,000	10	$3 \times 32 \times 32$
CIFAR-100	50,000	10,000	100	$3 \times 32 \times 32$

CIFAR-10 / 100 are two basic dataset in the computer version studies. Data samples in both are the colorful images with the small resolution of 32×32. The main reason why FL pays more attention to this kind of small dataset is that FL focuses on privacy-protection training on isolated small dataset, e.g. for medical images. Usually the resolution of such data is not high, and the number of samples per class is very limited due to its expensive labeling cost.

B.2. Distributions of Dirichlet and Pathological Split

(a) Dirichlet-0.6 on CIFAR-100.   (b) Pathological-30 on CIFAR-100.
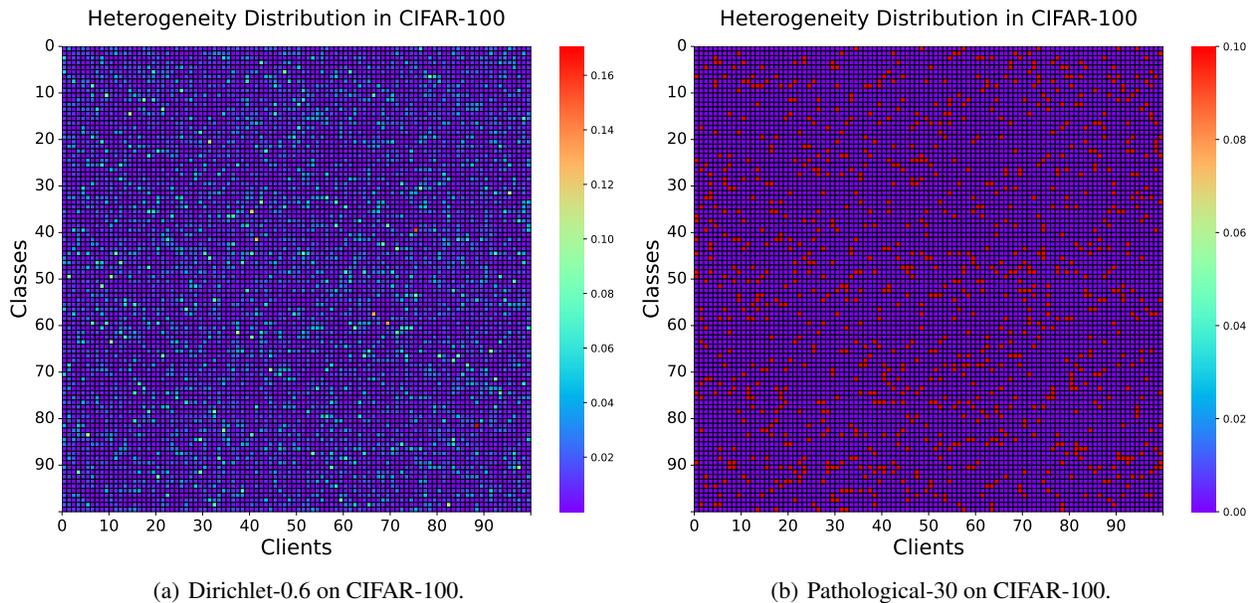
Figure 5. Heat-map of the Dirichlet split and Pathological split.

Above figures show the different distribution of the Dirichlet-0.6 and Pathological-30 on CIFAR-100 dataset. The main differences are:

- **Dirichlet:** Each category can be sampled with a non-zero probability. The local dataset obeys a Dirichlet distribution. As shown in Figure 5 (a), each category has a color.

- **Pathological:** Only selected categories can be sampled with a non-zero probability. The local dataset obeys a uniform distribution of active categories. As shown in Figure 5 (b), the category only has a color of purple or red.

B.3. Sampling *with replacement*

We have shown the difference between the difference between the vanilla process and our process. Adopting sampling *with replacement* will extremely change the distribution of the total dataset. Here we also show their differences in CIFAR-100 dataset.
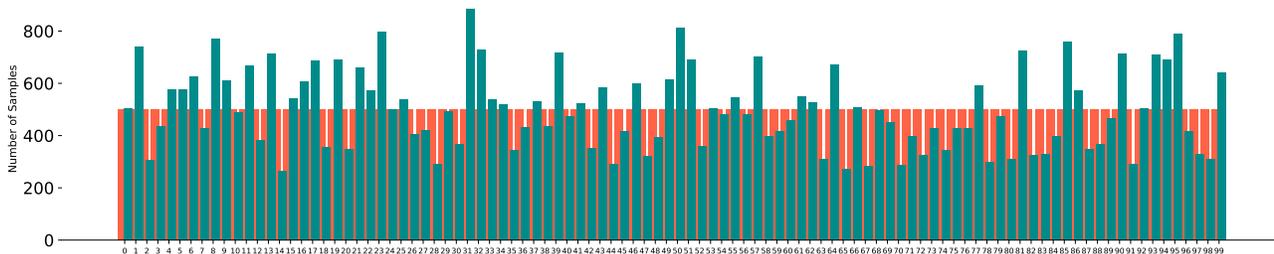
Figure 6. Distribution across category on CIFAR-100 of sampling with (green) /without (red) replacement under the Dirichlet coefficient $u = 0.1$ and the number of total clients $m = 100$.

Why we use this dataset splitting?

Under a true federated learning framework, the distribution between data is strictly forbidden to be accessed, which means we never know the other isolated dataset aligned to other clients. In other words, it is impossible to guarantee an uniform distribution of the categories on the total dataset. Therefore, it is very valuable to explore the real performance of FL on non-uniform dataset, e.g. for the heavy tail. We also find that many algorithms that work well are significantly less effective in the face of categories imbalance of the total dataset. Thus, in our experiments, we select this difficult problem as our dataset splitting method.

B.4. Detailed Hyperparameters Selection

We search for a lot of hyperparameters selection to explore the best performing of the baselines and our method. Our results show that the selection of best-performing hyperparameters is mainly distinguished by two main types, vanilla local SGD and its variants like `FedAvg`, `FedCM`, `SCAFFOLD`, `FedSAM`, and the proxy-based methods like `FedDyn`, `FedSMOO`. In the Table 5 we propose the all selections in our experiments and state the different combinations.

Table 5. Dataset introductions.

Options	`SGD`-type	Best Selection	`proxy`-type	Best Selection
local learning rate	$[0.01, 0.1, 0.5]$	0.1	$[0.01, 0.1, 0.5]$	0.1
global learning rate	$[0.1, 1.0]$	1.0	$[0.1, 1.0]$	1.0
SAM learning rate	$[0.001, 0.01, 0.1]$	0.01	$[0.01, 0.1, 1.0]$	0.1
learning rate decay	$[0.995, 0.998, 0.9995]$	0.998	$[0.998, 0.9995, 0.99995]$	0.9995
penalized coefficient β	-	-	$[1, 10, 100]$	10
client-level momentum α	$[0.05, 0.1, 0.5]$	0.1	-	-

For `FedAdam`, we use a adaptive learning rate which is different from the above due to its global adaptivity. We test a lot of selections to search for the best-performing hyperparameters. Empirically, the global learning rate is 1.0 for the averaged aggregation. The local learning rate usually adopts 0.1. The SAM learning rate is 0.01 for `FedSAM` and 0.1 for `FedSMOO`. The learning rate decay is different from each other. Usually for the `SGD`-type methods, they adopt the 0.998. While for the `proxy`-type methods, they require a larger selection, like 0.9995, 0.9998 for the stability during the training process. About the penalized coefficient β, we select 10 for CIFAR-10 and 100 for CIFAR-100. About the client-level momentum, we follow the (Xu et al., 2021) and set it as 0.1. The other common hyperparameters are selected as introduced in the text of experiments.

B.5. Evaluation Curves
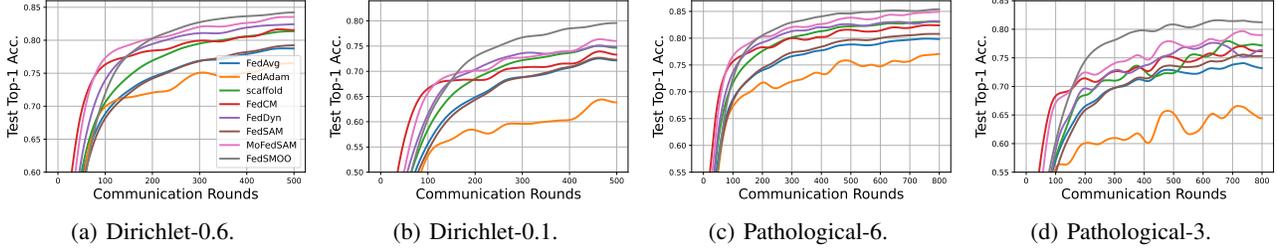
Figure 7. Accuracy on the CIFAR-10 dataset under 10% participation of total 100 clients.
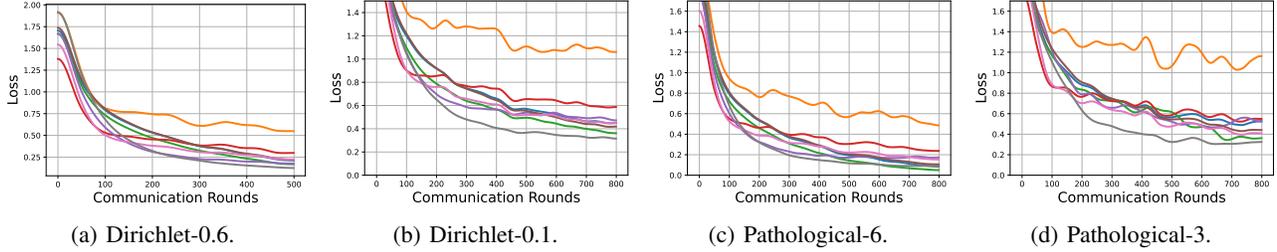
Figure 8. Loss on the CIFAR-10 dataset under 10% participation of total 100 clients.
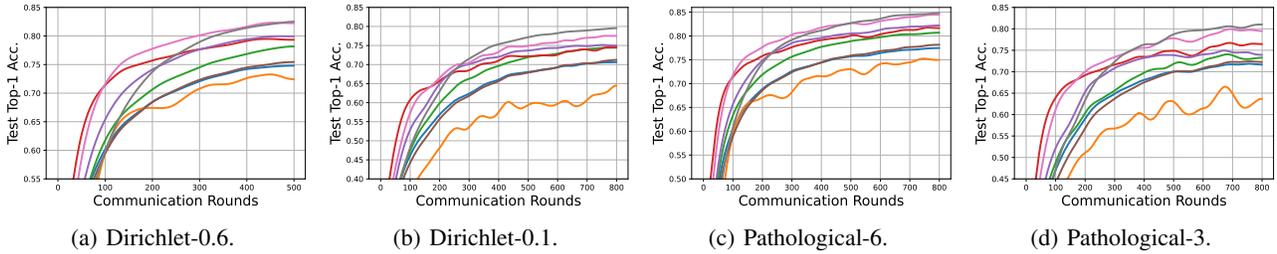
Figure 9. Accuracy on the CIFAR-10 dataset under 5% participation of total 200 clients.
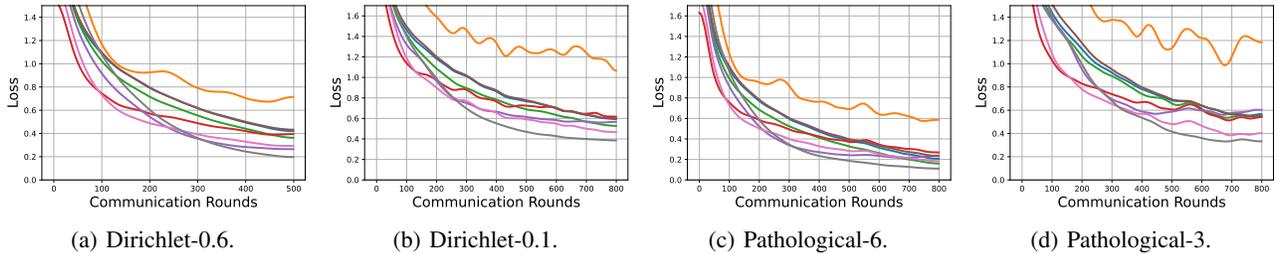
Figure 10. Loss on the CIFAR-10 dataset under 5% participation of total 200 clients.

To smooth the whole curve, we adopt the Hanning window as the filter to adjust them. According to the figures above, we can see that `FedSMOO` significantly outperforms the other algorithms, especially under the large heterogeneity, e.g. for Dirichlet-0.1 and Pathological-3. These results are in line with our expectations. Our original intention is to design the algorithm to effectively find the global flat minimum while strengthening the global consistency to obtain excellent generalization performance. When heterogeneity is low, the gaps are not large, and our algorithm remains ahead. When heterogeneity increases, our algorithm has a significant performance improvement.

B.6. Training Speed

Table 6. Communication rounds required to achieve the target test accuracy. We will note the slowest record as $1\times$ baseline to show the improvement of other algorithms on training rounds on the Dirichlet-0.1 and Pathological-10 setups. The upper part shows the results of the participation ratio equal to 10%-100 clients and the lower part shows the results of the participation ratio equal to 5%-200 clients.

ALGORITHM	CIFAR-10				CIFAR-100			
	DIRICHLET-0.1		PATHOLOGICAL-3		DIRICHLET-0.1		PATHOLOGICAL-10	
	$acc = 68\%$	$acc = 74\%$	$acc = 68\%$	$acc = 74\%$	$acc = 38\%$	$acc = 43\%$	$acc = 35\%$	$acc = 40\%$
FEDAVG	259	-	246	723	339	744	373	732
FEDADAM	445	-	487	-	772	-	788	-
SCAFFOLD	191	419	180	533	323	480	361	617
FEDCM	133	620	96	442	264	603	447	-
FEDDYN	148	410	173	424	239	-	267	-
FEDSAM	264	649	253	604	361	603	380	718
MOFEDSAM	122	404	116	298	271	445	450	599
OUR	132	224	144	194	252	350	258	354
	$acc = 68\%$	$acc = 74\%$	$acc = 68\%$	$acc = 74\%$	$acc = 35\%$	$acc = 40\%$	$acc = 33\%$	$acc = 38\%$
FEDAVG	494	-	397	-	438	747	529	-
FEDADAM	688	-	-	-	700	-	-	-
SCAFFOLD	345	696	352	674	433	609	518	-
FEDCM	251	744	167	438	255	437	374	751
FEDDYN	239	532	234	599	323	600	418	750
FEDSAM	509	-	419	-	478	729	556	-
MOFEDSAM	220	415	170	355	298	432	412	621
OUR	245	355	239	330	352	534	447	541

According to the above Table, we can see that our method performs well on average. To achieve the high accuracy, it is much faster than other benchmarks. SAM optimizers usually make the entire training process slower, see results of `FedAvg` and `FedSAM`. However, enhancing consistency will improve this. `MoFedSAM` force the consistency by adopting a global momentum on each local client weighted by a coefficient α (usually 0.1), which means in the local updates of round t, their gradients share a common direction as 90% and their own local gradients as 10%. `FedSMOO` provides a amended vector on both parameters w and perturbation s, which greatly enhances its consistency. Thus, our method can efficiently train the federated model, especially on the large heterogeneity.

Table 7. Training wall-clock time comparison to achieve 74% accuracy on the CIFAR-10 of Pathological-3 dataset split.

α_1	Times (s/Round)	Rounds	Total Time (s)
FedAvg	13.33 ($1.76\times$)	723 ($1\times$)	9637.59 ($1\times$)
FedAdam	13.77 ($1.70\times$)	-	-
SCAFFOLD	16.97 ($1.38\times$)	533 ($1.36\times$)	9045.01 ($1.06\times$)
FedCM	14.36 ($1.63\times$)	442 ($1.64\times$)	6347.12 ($1.51\times$)
FedDyn	12.81 ($1.83\times$)	424 ($1.71\times$)	5431.44 ($1.77\times$)
FedSAM	19.69 ($1.19\times$)	604 ($1.20\times$)	11892.76 ($0.81\times$)
MoFedSAM	22.08 ($1.06\times$)	298 ($2.43\times$)	6579.84 ($1.46\times$)
OUR	**23.52** ($1\times$)	**194** ($3.73\times$)	**4562.88** ($2.11\times$)

Test Experiments: A100-SXM4-40GB GPU, CUDA Driver 11.7, Driver Version 515.86.01, PyTorch-1.13.1

Table B.6 shows the wall-clock time costs on the CIFAR-10 of the Pathological-3 dataset split. Due to the double calculation of the gradients via SAM optimizer, `FedSAM`, `MoFedSAM`, and `FedSMOO` will take more time in a single round of updates, about $1.5\times$ over the single-calculation methods. However, the communication rounds required are much less than the single-calculation methods. Specifically, to achieve the 74% accuracy, `FedSMOO` can be accelerated by a factor of 3.73 on the communication rounds of the `FedAvg`. Considering the total wall-clock time costs, the acceleration ratio achieves $2.11\times$. An important case is that `FedSAM` can not accelerate the total wall-clock time due to the double-calculation, which must be combined with other methods to further improve its performance, like its variant `MoFedSAM`.

B.7. Ablation studies

Table 8. Ablation studies of different modules.

Dynamic Regularization	SAM	SAM-correction	Accuracy (%)
-	-	-	$74.08_{\pm.22}$
√	-	-	$77.63_{\pm.14}$
√	√	-	$80.13_{\pm.11}$
√	√	√	$81.58_{\pm.16}$

We test the performance of the different modules on the CIFAR-10 dataset of Pathological-3 split, which are named as "Dynamic Regularization", "SAM", and "SAM-correction" modules. The vanilla benchmark is `FedAvg`. After introducing the above three modules, the performance improvements are 3.55%, 6.05%, and 7.5%.

C. Proofs

C.1. Proof of Theorem 4.1

C.1.1. PRELIMINARY LEMMAS

Before proving the theorem, we first introduce some preliminary lemmas used in our proofs.

Lemma C.1. *For random variables $\{x_k\}_{k \in [K]} \in \mathbb{R}^d$, we can bound:*

$$\mathbb{E}\| \sum_k x_k \|^2 \le K \sum_k \mathbb{E}\|x_k\|^2. \tag{17}$$

Proof. It can be proved by Jensen inequality. $\qquad \square$

Lemma C.2. *For two random variables $x, y \in \mathbb{R}^d$, we can bound:*

$$\|x + y\|^2 \le \left(1 + \frac{1}{c}\right)\|x\|^2 + (1 + c)\|y\|^2, \tag{18}$$

where $c > 0$ is a constant.

Proof. It can be proved by triangle inequality. $\qquad \square$

Lemma C.3. *For two random variables $x, y \in \mathbb{R}^d$, we can bound:*

$$\langle x, y \rangle \le \frac{1}{2}\|x\|^2 - \frac{1}{2}\|x - y\|^2. \tag{19}$$

Proof. Applying the product of two vectors:

$$\|x - y\|^2 = \langle x - y, x - y \rangle = \langle x, x \rangle + \langle x, -y \rangle + \langle -y, x \rangle + \langle -y, -y \rangle = \|x\|^2 + \|y\|^2 - 2\langle x, y \rangle.$$

Thus we can bound the product term as:

$$\langle x, y \rangle = \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2 - \frac{1}{2}\|x - y\|^2 \le \frac{1}{2}\|x\|^2 - \frac{1}{2}\|x - y\|^2.$$

$\qquad \square$

C.1.2. PROOFS OF THEOREM 4.1

The paradigm of our proposed method is shown in Algorithm 1. We use a similar proof technique in (Karimireddy et al., 2020; Acar et al., 2021) and define a set of auxiliary variables to match the global update. With the partial participation training, the active clients at each communication round will update the local parameters while the inactive clients will inherit the previous parameters. Thus, we introduce the virtual variables $\widetilde{w}$ as:

$$\widetilde{w}_i^t = \arg\min_w \left\{ \mathscr{F}_i(w) - \langle \lambda_i^t, w - w^t \rangle + \frac{1}{2\beta}\|w - w^t\|^2 \right\}, \quad i \in [m]. \tag{20}$$

The virtual variable is based on partial participation. The active client set is randomly selected with the probability of $\frac{n}{m}$, which means that w_i^t equals to $\widetilde{w}_i^t$ with probability $\frac{n}{m}$ and maintains $\widetilde{w}_i^{t-1}$ otherwise. Thus the first order condition satisfies: $\beta\left(\nabla\mathscr{F}_i(\widetilde{w}_i^t) - \lambda_i^t\right) + \widetilde{w}_i^t - w^t = 0$, which indicates that $\lambda_i^{t+1} = \nabla\mathscr{F}_i(w_i^t)$ among active $i \in [n]$ after the local update of the

dual variable. If the client i is inactive, the $\widetilde{w}_i^t$ maintains the previous parameters w^t. Thus the global variable λ is updated as:

$$
\lambda^{t+1} - \lambda^t = -\frac{1}{\beta m} \sum_{i \in [n]} \left(w_i^t - w^t \right) = -\frac{1}{\beta m} \sum_{i \in [m]} \left(\widetilde{w}_i^t - w^t \right)
$$

$$
= \frac{1}{\beta m} \sum_{i \in [m]} \left(\nabla \mathscr{F}_i(\widetilde{w}_i^t) - \lambda_i^t \right) = \frac{1}{\beta m} \sum_{i \in [m]} (\lambda_i^{t+1} - \lambda_i^t).
$$

According to the update of the λ, we consider that $\lambda^t = \frac{1}{m} \sum_{i \in [m]} \lambda_i^t$ as $\lambda_i^{t+1} = \lambda_i^t$ holds for $i \in [m] - [n]$. In order to distinguish the parameters before and after updating with the dual variable, we define the:

$$
\overline{w}^{t+1} \triangleq \frac{1}{n} \sum_{i \in [n]} w_i^t = w^{t+1} + \beta \lambda^{t+1}. \tag{21}
$$

Lemma C.4. The adjacent averaged parameters generated in Algorithm 1 satisfies the following update:

$$
\mathbb{E}\left[\overline{w}^{t+1} \right] - \mathbb{E}\left[\overline{w}^t \right] = -\frac{\beta}{m} \sum_{i \in [m]} \mathbb{E}\left[\nabla \mathscr{F}_i(\widetilde{w}_i^t) \right]. \tag{22}
$$

Proof. According to the equation (21), we have:

$$
\mathbb{E}\left[\overline{w}^{t+1} \right] - \mathbb{E}\left[\overline{w}^t \right] = \mathbb{E}\left[\frac{1}{n} \sum_{i \in [n]} \left(w_i^t - w^t - \beta \lambda^t \right) \right] = \mathbb{E}\left[\frac{1}{m} \sum_{i \in [m]} \left(\widetilde{w}_i^t - w^t - \beta \lambda^t \right) \right]
$$

$$
= \frac{\beta}{m} \sum_{i \in [m]} \mathbb{E}\left[\lambda_i^t - \lambda^t - \nabla \mathscr{F}_i(\widetilde{w}_i^t) \right] = -\frac{\beta}{m} \sum_{i \in [m]} \mathbb{E}\left[\nabla \mathscr{F}_i(\widetilde{w}_i^t) \right].
$$

$\square$

The second equation is based on the expectation of $\widetilde{w}$ equal to the expectation of clients' sampling. The last equation is based on the relationship between λ_i and λ.

According to the lemma C.4, we can directly bound the norm term of the update of $\overline{w}$:

$$
\mathbb{E}\|\overline{w}^{t+1} - \overline{w}^t\|^2 = \mathbb{E}\|\frac{1}{n} \sum_{i \in [n]} w_i^t - \overline{w}^t\|^2 \leq \frac{1}{n} \mathbb{E} \sum_{i \in [n]} \|w_i^t - \overline{w}^t\|^2 \leq \frac{1}{m} \sum_{i \in [m]} \mathbb{E}\|\widetilde{w}_i^t - \overline{w}^t\|^2. \tag{23}
$$

The last inequality is based on the expectation of $\widetilde{w}_i$ equals to the expectation of clients' sampling.

In the inequality (23), the term $\widetilde{w}_i$ is the local parameters generated in algorithm 1 at each round t, and the $\overline{w}^t$ is the averaged parameters. Thus the *RHS* term of the inequality (23) represents the average norm expectation of the local changes before dual correction. It bounds that the norm of the global update will not exceed its average local updates. In this proof, we denote $v^t \triangleq \frac{1}{m} \sum_{i \in [m]} \mathbb{E}\|\widetilde{w}_i^t - \overline{w}^t\|^2$ and provide its upper bound in the next part. Furthermore, we need to consider another important term $\frac{1}{m} \sum_{i \in [m]} \mathbb{E}\|w_i^t - \overline{w}^{t+1}\|^2$, the averaged divergence of the heterogeneity during the local training, which is also defined as the 'client drift' in FL. It reflects the volatility caused by the heterogeneity in the local training process. In this proof, we denote δ^t as the divergence term at round t. In the next part, we will bound these two terms to demonstrate the convergence analysis.

Lemma C.5. Based on the assumptions, the averaged local update term could be bounded as:

$$
(1 - 12\beta^2 L^2)v^t \leq 30\beta^2 L^2 r^2 + 36\beta^2 L^2 \delta^{t-1} + 12\beta^2 \mathbb{E}\|\nabla f(\overline{w}^t)\|^2, \tag{24}
$$

where $\beta < \frac{1}{2\sqrt{3}L}$ for $1 - 12\beta^2 L^2 > 0$.

Proof. According to the update rules (21) and the first order gradient condition above, the v^t term could be bounded as:

$$v^t = \frac{1}{m} \sum_{i \in [m]} \mathbb{E}\|\widetilde{w}_i^t - \overline{w}^t\|^2 = \frac{1}{m} \sum_{i \in [m]} \mathbb{E}\|\widetilde{w}_i^t - w^t - \beta\lambda^t\|^2$$

$$= \frac{\beta^2}{m} \sum_{i \in [m]} \mathbb{E}\|\lambda_i^t - \nabla\mathcal{F}_i(\widetilde{w}_i^t) - \lambda^t\|^2 = \frac{\beta^2}{m} \sum_{i \in [m]} \mathbb{E}\|\nabla f_i(w_i^{t-1} + \hat{s}_i^{t-1}) - \nabla f_i(\widetilde{w}_i^t + \hat{s}_i^t) - \lambda^t\|^2$$

$$= \frac{\beta^2}{m} \sum_{i \in [m]} \mathbb{E}\|\nabla f_i(w_i^{t-1} + \hat{s}_i^{t-1}) - \nabla f_i(w_i^{t-1}) + \nabla f_i(w_i^{t-1}) - \nabla f_i(\widetilde{w}_i^t + \hat{s}_i^t) + \nabla f_i(\widetilde{w}_i^t) - \nabla f_i(\widetilde{w}_i^t) - \lambda^t\|^2$$

$$\leq \frac{3\beta^2}{m} \sum_{i \in [m]} \mathbb{E}\|\nabla f_i(w_i^{t-1} + \hat{s}_i^{t-1}) - \nabla f_i(w_i^{t-1})\|^2 + \frac{3\beta^2}{m} \sum_{i \in [m]} \mathbb{E}\|\nabla f_i(\widetilde{w}_i^t + \hat{s}_i^t) - \nabla f_i(\widetilde{w}_i^t)\|^2$$

$$+ \frac{3\beta^2}{m} \sum_{i \in [m]} \mathbb{E}\|\nabla f_i(w_i^{t-1}) - \nabla f_i(\widetilde{w}_i^t) - \lambda^t\|^2$$

$$\leq \frac{3\beta^2 L^2}{m} \sum_{i \in [m]} \|\hat{s}_i^{t-1}\|^2 + \frac{3\beta^2 L^2}{m} \sum_{i \in [m]} \|\hat{s}_i^t\|^2 + \frac{3\beta^2}{m} \sum_{i \in [m]} \mathbb{E}\|\nabla f_i(w_i^{t-1}) - \nabla f_i(\widetilde{w}_i^t) - \lambda^t\|^2$$

$$= \frac{3\beta^2 L^2}{m} \sum_{i \in [m]} \|r \frac{s_i^{t-1}}{\|s_i^{t-1}\|}\|^2 + \frac{3\beta^2 L^2}{m} \sum_{i \in [m]} \|r \frac{s_i^t}{\|s_i^t\|}\|^2 + \frac{3\beta^2}{m} \sum_{i \in [m]} \mathbb{E}\|\nabla f_i(w_i^{t-1}) - \nabla f_i(\widetilde{w}_i^t) - \lambda^t\|^2$$

$$= 6\beta^2 L^2 r^2 + \frac{3\beta^2}{m} \sum_{i \in [m]} \mathbb{E}\|\nabla f_i(w_i^{t-1}) - \nabla f_i(\widetilde{w}_i^t) - \lambda^t\|^2$$

$$= 6\beta^2 L^2 r^2 + \frac{3\beta^2}{m} \sum_{i \in [m]} \mathbb{E}\|\nabla f_i(w_i^{t-1}) - \nabla f_i(\overline{w}^t) + \nabla f_i(\overline{w}^t) - \nabla f_i(\widetilde{w}_i^t) - \nabla f(\overline{w}^t) + \nabla f(\overline{w}^t) - \lambda^t\|^2$$

$$\leq 6\beta^2 L^2 r^2 + \frac{12\beta^2 L^2}{m} \sum_{i \in [m]} \mathbb{E}\|w_i^{t-1} - \overline{w}^t\|^2 + \frac{12\beta^2 L^2}{m} \sum_{i \in [m]} \mathbb{E}\|\overline{w}^t - \widetilde{w}_i^t\|^2 + 12\beta^2 \mathbb{E}\|\nabla f(\overline{w}^t)\|^2$$

$$+ 12\beta^2 \mathbb{E}\|\frac{1}{m} \sum_{i \in [m]} \left(\nabla f_i(\overline{w}^t) - \lambda_i^t\right)\|^2$$

$$\leq 6\beta^2 L^2 r^2 + 12\beta^2 L^2 \delta^{t-1} + 12\beta^2 L^2 v^t + 12\beta^2 \mathbb{E}\|\nabla f(\overline{w}^t)\|^2 + \frac{12\beta^2}{m} \sum_{i \in [m]} \mathbb{E}\|\nabla f_i(\overline{w}^t) - \lambda_i^t\|^2$$

$$\leq 6\beta^2 L^2 r^2 + 12\beta^2 L^2 \delta^{t-1} + 12\beta^2 L^2 v^t + 12\beta^2 \mathbb{E}\|\nabla f(\overline{w}^t)\|^2$$

$$+ \frac{12\beta^2}{m} \sum_{i \in [m]} \mathbb{E}\|\nabla f_i(\overline{w}^t) - \nabla f_i(w_i^{t-1}) + \nabla f_i(w_i^{t-1}) - \nabla f_i(w_i^{t-1} + \hat{s}_i^{t-1})\|^2$$

$$\leq 30\beta^2 L^2 r^2 + 36\beta^2 L^2 \delta^{t-1} + 12\beta^2 L^2 v^t + 12\beta^2 \mathbb{E}\|\nabla f(\overline{w}^t)\|^2.$$

Collecting the like terms of *LHS* and *RHS* above will complete the proofs. □

Lemma C.5 reveal one relationship between the local updates and the divergence drifts. When we provide a constant bounded r, the upper bound of the local updates is always controlled by the global gradient term and divergence drifts, which is consistent with the conclusion of the gradient descent method in FL. In order to further clarify the relationship between these two terms, we introduce the upper bound of divergence drift term δ in the next.

Lemma C.6. Based on the assumptions, the divergence drifts term could be bounded as:

$$\delta^t \leq \frac{2m - 2n}{2m - n}\delta^{t-1} + \frac{6m - 2n}{2m - n}v^t. \tag{25}$$

Proof. According to the definition of $delta$, we have:

$$\delta^t = \frac{1}{m} \sum_{i \in [m]} \mathbb{E}\|w_i^t - \overline{w}^{t+1}\|^2 = \frac{1}{m} \sum_{i \in [m]} \mathbb{E}\|w_i^t - \overline{w}^t + \overline{w}^t - \overline{w}^{t+1}\|^2$$

$$\leq \left(1 + \frac{1}{c}\right) \frac{1}{m} \sum_{i \in [m]} \mathbb{E}\|w_i^t - \overline{w}^t\|^2 + (1 + c)\,\mathbb{E}\|\overline{w}^t - \overline{w}^{t+1}\|^2$$

$$= \frac{n}{m}\left(1 + \frac{1}{c}\right) \frac{1}{m} \sum_{i \in [m]} \mathbb{E}\|\widetilde{w}_i^t - \overline{w}^t\|^2 + \frac{m-n}{m}(1 + c)\frac{1}{m} \sum_{i \in [m]} \mathbb{E}\|w_i^{t-1} - \overline{w}^t\|^2 + (1 + c)\,\mathbb{E}\|\overline{w}^t - \overline{w}^{t+1}\|^2$$

$$= \left[\frac{n}{m}\left(1 + \frac{1}{c}\right) + (1 + c)\right] v^t + \frac{m-n}{m}(1 + c)\,\delta^{t-1}.$$

$\square$

For the convenience of analysis, we select a proper constant c to complete the proofs. Let $c = \frac{n}{2m-n} > 0$ where $n < m$, thus we have $1 + c = \frac{2m}{2m-n}$ and $1 + \frac{1}{c} = \frac{2m}{n}$. Plugging into the last equation and we have: $\left[\frac{n}{m}\left(1 + \frac{1}{c}\right) + (1 + c)\right] = \frac{6m-2n}{2m-n}$ and $\frac{m-n}{m}(1 + c) = \frac{2m-2n}{2m-n}$, which completes the proofs.

According to the lemmas above, we can bound the recursion terms of v^t and δ^t to prove the convergence theorem. From the assumption of smoothness, taking the conditional expectation with t on both side and we have:

$$\mathbb{E}_t\left[f(\overline{w}^{t+1})\right]$$

$$\leq f(\overline{w}^t) + \frac{L}{2}\mathbb{E}_t\|\overline{w}^{t+1} - \overline{w}^t\|^2 + \mathbb{E}_t\langle\nabla f(\overline{w}^t), \overline{w}^{t+1} - \overline{w}^t\rangle$$

$$= f(\overline{w}^t) + \frac{L}{2}\mathbb{E}_t\|\overline{w}^{t+1} - \overline{w}^t\|^2 + \mathbb{E}_t\langle\nabla f(\overline{w}^t), -\frac{\beta}{m}\sum_{i \in [m]}\mathbb{E}_t\left[\nabla \mathscr{F}_i(\widetilde{w}_i^t)\right]\rangle$$

$$\leq f(\overline{w}^t) + \frac{L}{2}\mathbb{E}_t\|\overline{w}^{t+1} - \overline{w}^t\|^2 + \frac{\beta}{2}\mathbb{E}_t\|\frac{1}{m}\sum_{i \in [m]}\left(\nabla f_i(\widetilde{w}_i^t + \hat{s}_i^t) - \nabla f(\overline{w}^t)\right)\|^2 - \frac{\beta}{2}\mathbb{E}_t\|\nabla f(\overline{w}^t)\|^2$$

$$\leq f(\overline{w}^t) + \frac{L}{2}\mathbb{E}_t\|\overline{w}^{t+1} - \overline{w}^t\|^2 + \frac{\beta}{2m}\sum_{i \in [m]}\mathbb{E}_t\|\nabla f_i(\widetilde{w}_i^t + \hat{s}_i^t) - \nabla f_i(\widetilde{w}_i^t) + \nabla f_i(\widetilde{w}_i^t) - \nabla f_i(\overline{w}^t)\|^2$$

$$- \frac{\beta}{2}\mathbb{E}_t\|\nabla f(\overline{w}^t)\|^2$$

$$\leq f(\overline{w}^t) + \frac{L}{2}\mathbb{E}_t\|\overline{w}^{t+1} - \overline{w}^t\|^2 + \frac{\beta}{m}\sum_{i \in [m]}\mathbb{E}_t\|\nabla f_i(\widetilde{w}_i^t + \hat{s}_i^t) - \nabla f_i(\widetilde{w}_i^t)\|^2 + \frac{\beta}{m}\sum_{i \in [m]}\mathbb{E}_t\|\nabla f_i(\widetilde{w}_i^t) - \nabla f_i(\overline{w}^t)\|^2$$

$$- \frac{\beta}{2}\mathbb{E}_t\|\nabla f(\overline{w}^t)\|^2$$

$$\leq f(\overline{w}^t) + \frac{L}{2}\mathbb{E}_t\|\overline{w}^{t+1} - \overline{w}^t\|^2 + \frac{\beta L^2}{m}\sum_{i \in [m]}\mathbb{E}_t\|r\frac{s_i^t}{\|s_i^t\|}\|^2 + \frac{\beta L^2}{m}\sum_{i \in [m]}\mathbb{E}_t\|\widetilde{w}_i^t - \overline{w}^t\|^2 - \frac{\beta}{2}\mathbb{E}_t\|\nabla f(\overline{w}^t)\|^2$$

$$\leq f(\overline{w}^t) + \frac{L}{2}\mathbb{E}_t\|\overline{w}^{t+1} - \overline{w}^t\|^2 + \beta L^2 r^2 + \frac{\beta L^2}{m}\sum_{i \in [m]}\mathbb{E}_t\|\widetilde{w}_i^t - \overline{w}^t\|^2 - \frac{\beta}{2}\mathbb{E}_t\|\nabla f(\overline{w}^t)\|^2.$$

We take full expectation from 0 to t on the above inequality and get:

$$\mathbb{E}\left[f(\overline{w}^{t+1})\right] \leq \mathbb{E}\left[f(\overline{w}^t)\right] + \beta L^2 r^2 + \frac{L(1 + 2\beta L)}{2}v^t - \frac{\beta}{2}\mathbb{E}\|\nabla f(\overline{w}^t)\|^2. \tag{26}$$

Equation (26) indicates the relationship of expected global update and its norm term. Here we consider the lemma C.5 and C.6,

$$(1 - 12\beta^2 L^2)v^t \leq 30\beta^2 L^2 r^2 + 36\beta^2 L^2 \delta^{t-1} + 12\beta^2 \mathbb{E}\|\nabla f(\overline{w}^t)\|^2, \tag{27}$$

$$\delta^t \le \frac{2m-2n}{2m-n}\delta^{t-1} + \frac{6m-2n}{2m-n}v^t. \tag{28}$$

Let formula (27) multiplied by constant p and formula (28) multiplied by constant q, we take the sum of formula (26) (27) (28),

$$
\begin{aligned}
\mathbb{E}\left[f(\overline{w}^{t+1})\right] + p(1-12\beta^2 L^2)v^t + q\delta^t \le {}& \mathbb{E}\left[f(\overline{w}^t)\right] + \beta L^2 r^2 + \frac{L(1+2\beta L)}{2}v^t - \frac{\beta}{2}\mathbb{E}\|\nabla f(\overline{w}^t)\|^2 \\
& + 30p\beta^2 L^2 r^2 + 36p\beta^2 L^2 \delta^{t-1} + 12p\beta^2 \mathbb{E}\|\nabla f(\overline{w}^t)\|^2 \\
& + q\frac{2m-2n}{2m-n}\delta^{t-1} + q\frac{6m-2n}{2m-n}v^t.
\end{aligned}
$$

Collecting the like term of v^t and let the constants p and q satisfy:

$$p(1-12\beta^2 L^2) = \frac{L(1+2\beta L)}{2} + q\frac{6m-2n}{2m-n}. \tag{29}$$

When equation (29) holds, the relationship will be simplified to:

$$
\begin{aligned}
\mathbb{E}\left[\overline{w}^{t+1}\right] + q\delta^t \le {}& \mathbb{E}\left[\overline{w}^t\right] + (1+30p\beta)\beta^2 L^2 r^2 + \left(36p\beta^2 L^2 + q\frac{2m-2n}{2m-n}\right)\delta^{t-1} \\
& - \left(\frac{\beta}{2} - 12p\beta^2\right)\mathbb{E}\|\nabla f(\overline{w}^t)\|^2.
\end{aligned}
$$

Furthermore, considering the coefficient of δ term, we let the constant p and q satisfy:

$$q = 36p\beta^2 L^2 + q\frac{2m-2n}{2m-n}. \tag{30}$$

According to the equation (29) and (30), we can get the solution of p and q as:

$$
\begin{aligned}
p &= \frac{L(1+2\beta L)}{2}\left(1 + 60\beta^2 L^2 - \frac{216m\beta^2 L^2}{n}\right)^{-1}, \\
q &= \frac{(2m-n)36\beta^2 L^2}{n}.
\end{aligned}
$$

This proof requires the constants p and q both to be positive, thus the β satifies $\beta \le \frac{\sqrt{n}}{6\sqrt{6m}L}$. With this selection, the coefficient of the global gradient term maintain a positive value. Let ζ to be the coefficient of the global gradient term, we can rewrite the relationship as:

$$\zeta\mathbb{E}\|\nabla f(\overline{w}^t)\|^2 \le \left(\mathbb{E}\left[f(\overline{w}^t)\right] + q\delta^{t-1}\right) - \left(\mathbb{E}\left[f(\overline{w}^{t+1})\right] + q\delta^t\right) + (1+30p\beta)\beta^2 L^2 r^2.$$

Adding up the above formula from 0 to $T-1$ and applying $q \le \frac{72m\beta^2 L^2}{n}$, let $\beta \le \min\{\frac{\sqrt{n}}{6\sqrt{6m}L}, \frac{1}{2\sqrt{3}L}\}$, we have:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\overline{w}^t)\|^2 \le \frac{1}{\zeta T}\left(\left(f(\overline{w}^1) - f^*\right) + \frac{72\beta^2 L^2 \sum_{i\in[m]}\mathbb{E}\|w_i^0 - \overline{w}^1\|^2}{n}\right) + (1+30p\beta)\beta^2 L^2 r^2. \tag{31}$$

Similar to the FedSAM and MoFedSAM (Qu et al., 2022), we select the perturbation learning rate $r = \mathcal{O}(\frac{1}{\sqrt{T}})$ that the final convergence rate approaches $\mathcal{O}(\frac{1}{T})$, which completes the proofs. Selecting some proper constants makes Theorem 4.1 hold.

C.2. Proof of Theorem 4.4

Based on the margin generalization bounds in (Neyshabur et al., 2017; Farnia et al., 2018), we consider the generalization error bound on the global function f as following:

$$G_\epsilon^f \triangleq \mathbf{P}\Big(f(w + s, \varepsilon)[y] \leq \max_{j \neq y} f(w + s, \varepsilon)[j] + \epsilon\Big). \tag{32}$$

ε is the input data and y is its ground truth. If $\epsilon = 0$, we denote the equation (32) as G^f to represent for the average misclassification rate under the global perturbation s. We define the $\widetilde{G}_\epsilon^f$ as the empirical estimate of the above expected margin loss trained with the heterogeneous dataset.

We assume the input data ε as the normalized tensor, e.g. for an image, with the norm of L_n. And we consider a L-layer neural network with d parameters per layer. The activation functions hold no biases and bounded 1-Lipschitz property. The input of each layer are bounded by a positive constant L_w. Under the margin value of ϵ and total m data samples, with the probability of at least $1 - p$, the global parameters w generated by FedSMOO satisfy:

$$G^f \leq \widetilde{G}_\epsilon^f + \mathcal{O}\left(\sqrt{\frac{L_l^2 L_n^2 d \ln(dL_l)\mathcal{V}_L + \ln \frac{L_l D}{p}}{(D - 1)\epsilon^2}}\right). \tag{33}$$

where $\mathcal{V}_L = \prod_{l=1}^L \|w_l\|^2 \sum_{l=1}^L \frac{\|w_l\|_F^2}{\|w_l\|^2}$.

Lemma C.7. *(Neyshabur et al., 2017) If $f(w, \varepsilon)$ is the predictor with the parameters w. $\mathscr{P}$ is the arbitrary distribution on w. We assume the data ε is sampled from the $\mathbb{D}$ and its size is D. Thus for any $\epsilon, \delta > 0$, with the probability at least $1 - p$, and the arbitrary perturbation s satisfies $\mathbb{P}_s\left(\max_\varepsilon |f(w + s, \varepsilon) - f(w, \varepsilon)|_\infty \leq \frac{\epsilon}{4}\right) \geq \frac{1}{2}$, then:*

$$G^f \leq \widetilde{G}_\epsilon^f + 4\sqrt{\frac{KL(w + s\|\mathscr{P}) + \log \frac{6D}{p}}{D - 1}}. \tag{34}$$

This KL divergence comes from the frozen parameters w and the s is a random variable. It comes from the PAC-Bayes margin bounds for a linear predictors. Then we bound the inference of parameters.

Lemma C.8. *(Neyshabur et al., 2017) We assume the input ε is a tensor with bounded norm of L_n. And we consider a L-layer neural network with d parameters per layer. The activation functions hold no biases and bounded 1-Lipschitz property. Thus, for the arbitrary input sample, let the parameters are perturbed by a bounded noise s with $\|s_l\| \leq \|w_l\|$, the total change of the value of function f is:*

$$\|f(w + s, \varepsilon) - f(w, \varepsilon)\| \leq e L_n \prod_{l=1}^L \|w_l\| \sum_{l=1}^L \frac{\|s_l\|}{\|w_l\|}. \tag{35}$$

This lemma bounds the change from the perturbation s, which judges the stability of the neural network after the training. The perturbation s could be an arbitrary vectors with the same size of the model parameters.

Then with the preliminary lemmas, we prove the generalization bound of our proposed method. Firstly, as the previous works (Farnia et al., 2018; Qu et al., 2022), we assume the perturbation s_l for each layer is a zero-mean Gaussian noise of $\mathcal{N}(0, \sigma_l^2)$. We consider the new parameters w^s which satisfies the $|\|w_l\|_2 - \|w_l^s\|_2| \leq \frac{1}{L}$, thus we have (Tropp, 2011):

$$\mathbb{P}\left(\prod_{l=1}^L \|w_l\|^{\frac{1}{L}} \frac{\|s_l\|}{\|w_l^s\|} > p\right) \leq 2d e^{-\frac{p^2 \prod_{l=1}^L \|w_l\|^{\frac{2}{L}}}{2d\sigma_l^2 \|w_l^s\|^2}}.$$

Thus, we consider the union bound over all layers by at least probability of 0.5, which is:

$$\max_\epsilon \|f(w + s, \varepsilon) - f(w, \varepsilon)\| \leq e L_n \prod_{l=1}^L \|w_l\| \sum_{l=1}^L \frac{\|s_l\|}{\|w_l\|} \leq e^2 L L_n \prod_{l=1}^L \|w_l\|^{\frac{L-1}{L}} \sigma \sqrt{2d \ln(4dL)}.$$

Where we select a constant variance of $\sigma = \dfrac{\epsilon}{4e^2 L L_n \prod_{l=1}^{L} \|w_l\|^{\frac{L-1}{L}} \sqrt{2d \ln(4dL)}}$, we have:

$$\|f(w + s, \varepsilon) - f(w, \varepsilon)\| \le \max \|f(w + s, \varepsilon) - f(w, \varepsilon)\| \le \frac{\epsilon}{4}.$$

Thus we bound the KL divergence on the distribution $\mathscr{P}$ as:

$$
\begin{aligned}
KL(w + s \| \mathscr{P}) &\le \frac{\sum_{l=1}^{L} \|w_l\|_F^2}{2\sigma^2} = \frac{16e^4 L^2 L_n^2 \prod_{l=1}^{L} \|w_l\|^{\frac{2L-2}{L}} d \ln(4dL)}{\epsilon^2} \sum_{l=1}^{L} \|w_l\|_F^2 \\
&= \frac{16e^4 L^2 L_n^2 \prod_{l=1}^{L} \|w_l\|^2 d \ln(4dL)}{\epsilon^2} \sum_{l=1}^{L} \frac{\|w_l\|_F^2}{\prod_{l=1}^{L} \|w_l\|^{\frac{2}{L}}} \\
&\le \frac{16e^4 L^2 L_n^2 \prod_{l=1}^{L} \|w_l\|^2 d \ln(4dL)}{\epsilon^2} \sum_{l=1}^{L} \frac{\|w_l\|_F^2}{\|w_l\|^2}.
\end{aligned}
$$

Hence, when we provide a arbitrary distribution $\mathscr{P}$, with the probability of at least $1 - p$, the optimized parameters w under the cover of the total dataset size $LD^{\frac{1}{2L}}$, we have:

$$
G^f \le \widetilde{G}_\epsilon^f + \mathcal{O}\left(\sqrt{\frac{L^2 L_n^2 d \ln(dL) \prod_{l=1}^{L} \|w_l\|^2 \sum_{l=1}^{L} \frac{\|w_l\|_F^2}{\|w_l\|^2} + \ln \frac{Lm}{p}}{m\epsilon^2}} \right). \tag{36}
$$