Generalized Federated Learning via Sharpness Aware Minimization

Zhe Qu [*1] **Xingyu Li** [*2] **Rui Duan** [1] **Yao Liu** [3] **Bo Tang** [2] **Zhuo Lu** [1]

Abstract

Federated Learning (FL) is a promising framework for performing privacy-preserving, distributed learning with a set of clients. However, the data distribution among clients often exhibits non-IID, i.e., distribution shift, which makes efficient optimization difficult. To tackle this problem, many FL algorithms focus on mitigating the effects of data heterogeneity across clients by increasing the performance of the global model. However, almost all algorithms leverage Empirical Risk Minimization (ERM) to be the local optimizer, which is easy to make the global model fall into a sharp valley and increase a large deviation of parts of local clients. Therefore, in this paper, we revisit the solutions to the distribution shift problem in FL with a focus on local learning generality. To this end, we propose a general, effective algorithm, `FedSAM`, based on Sharpness Aware Minimization (SAM) local optimizer, and develop a momentum FL algorithm to bridge local and global models, `MoFedSAM`. Theoretically, we show the convergence analysis of these two algorithms and demonstrate the generalization bound of `FedSAM`. Empirically, our proposed algorithms substantially outperform existing FL studies and significantly decrease the learning deviation.

1. Introduction

Federated Learning (FL) (McMahan et al., 2017) is a collaborative training framework that enables a large number of clients, which can be phones, network sensors, or alternative local information sources (Kairouz et al., 2019; Mohri et al., 2019). FL trains machine learning models without transmitting client data over the network, and thus it can protect data privacy at some basic levels. Two important settings are introduced in FL (Kairouz et al., 2019): the *cross-device* FL and the *cross-silo* FL. The cross-silo FL is related to a small number of reliable clients, e.g., medical or financial institutions. By contrast, the cross-device FL includes a large number of clients, e.g., billion-scale android phones (Hard et al., 2018). In cross-device FL, clients are usually deployed in various environments. It is unavoidable that the distribution of the local dataset of each client varies considerably and incurs a distribution shift problem, highly degrading the learning performance.

Many existing FL studies focus on the distribution shift problem mainly based on the following three directions: (i) The most popular solution to address this problem is to set the number of local training epochs performed between each communication round (Li et al., 2020b; Yang et al., 2021). (ii) Many algorithmic solutions in (Li et al., 2018b; Karimireddy et al., 2020; Acar et al., 2021) mainly focus on mitigating the influence of heterogeneity across clients via giving a variety of proximal terms to control the local model updates close to the global model. (iii) Knowledge distillation based techniques (Lin et al., 2020; Gong et al., 2021; Zhu et al., 2021) aggregate locally-computed logits for building global models, helping eliminate the need for each local model to follow the same architecture to the global model.

Motivation. In centralized learning, the network generalization technique has been well studied to overcome the overfitting problem (Lakshminarayanan et al., 2017; Woodworth et al., 2020). Even in standard settings where the training and test data are drawn from a similar distribution, models still overfit the training data and the training model will fall into a sharp valley of the loss surface by using Empirical Risk Minimization (ERM) (Chaudhari et al., 2019). This effect is further intensified when the training and test data are of different distributions. Similarly, in FL, overfitting the local training data of each client is detrimental to the performance of the global model, as the distribution shift problem creates conflicting objectives among local models. The main strategy to improve the FL performance is to mitigate the local models to the global model from the average perspective (Karimireddy et al., 2020; Yang et al.,

[*]Equal contribution [1]Department of Electrical Engineering, University of South Florida, Tampa, USA [2]Department of Electrical and Computer Engineering, Mississippi State University, Starkville, USA [3]Department of Computer Science and Engineering, University of South Florida, Tampa, USA. Correspondence to: Zhe Qu <zhequ@usf.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

2021; Li et al., 2018b), which has been demonstrated to accelerate the convergence of FL. However, fewer existing FL studies focus on how to protect the learning performance of the clients with poor performance, and hence parts of clients may lose their unique properties and incur large performance deviation. Therefore, a focus on improving global model generality should be of primary concern in the presence of the distribution shift problem. Improving local training generality would inherently position the objective of the clients closer to the global model objective.

Recently, efficient algorithms Sharpness Aware Minimization (SAM) (Foret et al., 2021) have been developed to make the surface of loss function more smooth and generalized. It does not need to solve the min-max objectives as adversarial learning (Goodfellow et al., 2014; Shafahi et al., 2020); instead, it leverages linear approximation to improve the efficiency. As we discussed previously, applying SAM to be the local optimizer for generalizing the global model in FL should be an effective approach. We first introduce a basic algorithm adopting SAM in FL settings, called `FedSAM`, where each local client trains the local model with the same perturbation bound.

Although `FedSAM` can help to make the global model generalization and improve the training performance, they do not affect the global model directly. In order to bridge the smooth information on both local and global models without accessing others' private data, we develop our second and more important algorithm in our framework, termed Momentum `FedSAM` (`MoFedSAM`) by additionally downloading the global model updates of the previous round, and then letting clients perform local training on both local dataset and global model updates by SAM.

Contributions. We summarize our contributions as follows: (1) We approach one of the most troublesome cross-device FL challenges, i.e., distribution shift caused by data heterogeneity. To generalize the global model, we first propose a simple algorithm `FedSAM` performing SAM to be the local optimizer.

(2) We prove the convergence results $\mathcal{O}(\frac{L}{\sqrt{RKN}})$ and $\mathcal{O}(\frac{\sqrt{K}}{\sqrt{RS}})$ for `FedSAM` algorithm, which matches the best convergence rate of existing FL studies. For the part of local training in the convergence rate, our proposed algorithms show speedup. Moreover, the generalization bound of `FedSAM` is also presented.

(3) To directly smooth the global model, we develop `MoFedSAM` algorithm, which performs local training with both local dataset and global model updates by SAM optimizer. Then, we present the convergence rates are $\mathcal{O}(\frac{\sqrt{\beta L}}{\sqrt{RKN}})$ and $\mathcal{O}(\frac{\sqrt{\beta K}}{\sqrt{RS}})$ on full and partial client participation strategies, which achieves speedup and implies that `MoFedSAM` is a more efficient algorithm to address the distribution shift problem.

Related work. In this paper, we aim to evaluate and distinguish the generalization performance of clients. Throughout this paper, we only focus on the classic cross-device FL setting (McMahan et al., 2017; Li et al., 2018b; Karimireddy et al., 2020) in which a single global model is learned from and served to all clients. In the Personalized FL (PFL) setting (T Dinh et al., 2020; Fallah et al., 2020; Singhal et al., 2021), the goal is to learn and serve different models for different clients. While related, our focus and contribution are orthogonal to personalization. In fact, our proposed algorithms are easy to extend to the PFL setting. For example, by solving a hyperparameter to control the interpolation between local and global models (Deng et al., 2020; Li et al., 2021), the participating clients can be defined as the clients that contribute to the training of the global model. We can use SAM to develop the global model and generate the local models by ERM to improve the performance.

Momentum FL is an effective way to address the distribution shift problem and accelerate the convergence, which is based on injecting the global information into the local training directly. Momentum can be set on the server (Wang et al., 2019; Reddi et al., 2020), client (Karimireddy et al., 2021; Xu et al., 2021) or both (Khanduri et al., 2021). As we introduce previously, while these algorithms accelerate the convergence, the global model will locate in a sharp valley and overfit. As such, the global model may not be efficient for all clients and generate a large deviation.

We propose to train global models using a set of participating clients and examine their performance both on training and validation datasets. In the centralized learning, some studies(Keskar et al., 2016; Lakshminarayanan et al., 2017; Woodworth et al., 2020) consider the out-of-distribution generalization problem, which shows on centrally trained models that even small deviations in the morphology of deployment examples can lead to severe performance degradation. The sharpness minimization is an efficient way to deal with this problem (Foret et al., 2021; Kwon et al., 2021; Zhuang et al., 2022; Du et al., 2021a). The FL setting differs from these other settings in that our problem assumes data is drawn from a distribution of client distributions even if the union of these distributions is stationary. Therefore, in FL settings, we consider the training performance and validation. It incurs more challenges than centralized learning. Although some studies develop algorithms to generalize the global model in FL (Mendieta et al., 2021; Yuan et al., 2021; Yoon et al., 2021), they lack theoretical analysis of how the proposed algorithm can improve the generalization and may incur privacy issues. A recent study (Caldarola et al., 2022) shows via empirical experiments that using SAM to be the local optimizer can improve the generalization of FL.

2. Preliminaries and Proposed Algorithms

2.1. `FedAvg` Algorithm

Consider a FL setting with a network including N clients connected to one aggregator. We assume that for every $i \in [N]$ the i-th client holds m training data samples $\xi_i = (\mathbf{X}_i, Y)$ drawn from distribution $\mathcal{D}_i$. Note that $\mathcal{D}_i$ may differ across different clients, which corresponds to client heterogeneity. Let $F_i(w)$ be the training loss function of the client i, i.e., $F_i(w) \triangleq \mathbb{E}_{\xi_i \sim \mathcal{D}_i}[\mathcal{L}_i(w, \xi_i)]$, where $\mathcal{L}_i(w, \xi_i)$ is the per-data loss function. The classical FL problem (McMahan et al., 2017; Li et al., 2020b; Karimireddy et al., 2020) is to fit the best model w to all samples via solving the following empirical risk minimization (ERM) problem on each client:

$$\min_{w} \left\{ F(w) := \frac{1}{N} \sum_{i \in [N]} F_i(w) \right\}. \tag{1}$$

where $F(w)$ is the loss function of the global model. `FedAvg` (McMahan et al., 2017) is one of the most popular algorithms to address (1). In the communication round r, the server randomly samples $\mathcal{S}^r$ clients with the number of S and downloads the global model w^r to them. After receiving the global model, these sampled clients run K times local Stochastic Gradient Descent (SGD) epochs using their local dataset in parallel, and upload the local model updates $w^r_{i,K}$ to the server. When the server receives all the local model updates, it averages these to obtain the new global model w^{r+1}. The pesudocode of `FedAvg` is shown in Algorithm 1.

2.2. `FedSAM` Algorithm

Statistically heterogeneous local training dataset across the clients is one of the most important problems in FL studies. By capturing the Non-IID nature of local datasets in FL, the common assumption in existing FL studies (Mohri et al., 2019; Li et al., 2020a; Karimireddy et al., 2020; Reisizadeh et al., 2020) considers that the data samples of each client have a local distribution shift from a common unknown mixture distribution $\mathcal{D}$, i.e., $\mathcal{D}_i \neq \mathcal{D}$. While training via minimizing ERM by SGD searches for a single point w with a low loss, which can perfectly fit the distribution $\mathcal{D}$, it often falls into a sharp valley of the loss surface (Chaudhari et al., 2019). As a result, the global model w may be biased to parts of clients (i.e., low heterogeneity compared to the mixture distribution $\mathcal{D}$) and cannot guarantee enough generalization that makes all clients perform well. Moreover, since the training dataset distribution of each client may be different from the validation dataset with high probability, i.e., $\mathcal{D}_i^{\text{tra}} \neq \mathcal{D}_i^{\text{val}}$, and the validation dataset cannot be accessible during the training process, the global model w may not guarantee the learning performance of every client even for the clients working well during the training process.

To address this problem, some FL algorithms with fairness guarantee have been developed (Li et al., 2020a; Du et al., 2021b), but they only consider the learning performance from the average perspective and do not protect the worse clients. In order to focus on the average and deviation for all clients at the same time, it is necessary to create a more general global model to serve all clients.

Instead of searching for a single point solution such as ERM, the state-of-the-art algorithm Sharpness Aware Minimization (SAM) (Foret et al., 2021) aims to seek a region with low loss values via adding a small perturbation to the models, i.e., $w + \delta$ with less performance degradation. Due to the linear property of the FL optimization in (1), it is not difficult to observe that training the perturbed loss via SAM, i.e., $\tilde{w} = w + \delta_i$, on each client should reduce the impact on the distribution shift and improve the generalization of the global model. Based on this observation, we design a more general FL algorithm called `FedSAM` in this paper. The optimization of `FedSAM` is formulated as follows:

$$\min_{w} \max_{\|\delta_i\|_2^2 \leq \rho} \left\{ f(\tilde{w}) := \frac{1}{N} \sum_{i \in [N]} f_i(\tilde{w}) \right\}, \tag{2}$$

where $f(\tilde{w}) \triangleq \max_{\|\delta\| \leq \rho} F(w + \delta)$, $f_i(\tilde{w}) \triangleq \max_{\|\delta_i\| \leq \rho} F_i(w + \delta_i)$, ρ is a predefined constant controlling the radius of the perturbation and $\|\cdot\|_2^2$ is a l_2-norm, which will be simplified to $\|\cdot\|$ in the rest paper. Next, we take a close look at the local perturbed loss function $F_i(w + \delta_i)$ and introduce how to use SAM to approach it. For a small value of ρ, using first order Taylor expansion around w, the inner maximization in (2) turns into the following linear constrained optimization:

$$\begin{aligned} \delta_i &= \operatorname*{argmax}_{\|\delta_i\| \leq \rho} F_i(w + \delta_i) \\ &\approx \operatorname*{argmax}_{\|\delta_i\| \leq \rho} F_i(w) + \delta_i^\top \nabla F_i(w) + O(\rho^2) \\ &= \rho \text{sign}(\nabla F_i(w)) \frac{\nabla F_i(w)}{\|\nabla F_i(w)\|}, \end{aligned} \tag{3}$$

where $\text{sign}(\cdot)$ denotes element-wise signum function. Therefore, the local optimizer of `FedSAM` changes to $\min_w F_i(w) = \min_{\tilde{w}} f_i(\tilde{w})$, where $\tilde{w} \triangleq w + \rho \frac{\nabla F_i(w)}{\|\nabla F_i(w)\|}$. We call $\tilde{w}$ is the perturbed model with the highest loss within the neighborhood. Local SAM optimizer solves the min-max problem by iteratively applying the following two-step procedure for epoch $k = 0, \ldots, K-1$ in communication round r:

$$\begin{cases} \tilde{w}^r_{i,k} = w^r_{i,k} + \rho \dfrac{g^r_{i,k}}{\|g^r_{i,k}\|} \\ w^r_{i,k+1} = w^r_{i,k} - \eta_l \tilde{g}^r_{i,k}, \end{cases} \tag{4}$$

where η_l is the learning rate of local model updates on each client, $g^r_{i,k} = \nabla F_i(w^r_{i,k}, \xi^r_i)$ of $\nabla F_i(w_{i,k})$ and $\tilde{g}^r_{i,k} =$

Algorithm 1 FedAvg and FedSAM

Initialization: w_0, ρ_0 $\Delta^0 = 0$, learning rates η_l, η_g and the number of epochs K.
for $r = 0, \ldots, R-1$ **do**
  Sample subset $\mathcal{S}^r \subseteq [N]$ of clients.
  $w_{i,0}^t = w^r$.
  for each client $i \in \mathcal{S}^r$ in parallel **do**
    for $k = 0, \ldots, K-1$ **do**
      Compute a local training estimate $g_{i,k}^r = \nabla F_i(w_{i,k}^r, \xi_{i,k}^r)$ of $\nabla F_i(w_{i,k}^r)$.
      $w_{i,k}^r = w_{i,k}^r - \eta_l g_{i,k}^r$.
      Compute local model $w_{i,k}^r$ from (4).
    end for
    $\Delta_i^r = w_{i,K}^r - w^r$.
  end for
  $\Delta^{r+1} = \frac{1}{S} \sum_{i \in \mathcal{S}^r} \Delta_i^r$.
  $w^{r+1} = w^r + \eta_g \Delta^r$.
end for

$\nabla f(\tilde{w}_{i,k}^r, \xi_i^r)$ of $f_i(\tilde{w}_{i,k}^r)$. We can see that from (4), local training of each client estimates the point $w_{i,k}^r + \delta_i^r$ at which the local loss is maximized around $w_{i,k}^r$ in a region with a fixed perturbed radius approximately by using gradient ascent, and calculates gradient descent at $w_{i,k}^r$ based on the gradient at the maximum point $w_{i,k}^r + \delta_i^r$.

To present the difference between FedAvg and FedSAM, we summarize the training procedures in Algorithm 1. SAM optimizer comes from the similar idea of adversarial training, and it has been used in FL (Reisizadeh et al., 2020) called FedRobust. It is based on solving min-max objectives, which brings up more computational cost for local training and the worse convergence performance than our proposed algorithms. We will show the comparison both on theoretical and empirical perspectives.

Remark 2.1 Here, we briefly mention the SAM local optimizer can improve the generalization and help convergence from the smoothness perspective. We assume that the local loss function $f(w)$ is L-smooth. Clearly, the loss function f is smoother, when L is smaller. Assume that $f(\tilde{w})$ is G-Lipschitz continuous, and $\delta \sim \mathcal{N}(0, \epsilon^2 I)$, by leveraging (Nesterov & Spokoiny, 2017), we obtain that the perturbed loss function $f(\tilde{w})$ of FedSAM is $\frac{2G}{\epsilon}$-smooth. Based on the analysis in (Lian et al., 2017; Goyal et al., 2017), the best convergence rate should be $\frac{1}{L}$. For SGD based FL with the original loss surface, L can be very high (even close to $+\infty$ due to the non-smooth nature of the ReLU activation). Obviously, L of the perturbed loss $f(\tilde{w})$ in FedSAM should be much smaller due to the loss region. This can explain the intuition why increasing smoothness can significantly improve the convergence of FL.

3. Theoretical Analysis

In what follows, we show the convergence results of FedSAM algorithm for general non-convex FL settings. In order to propose the convergence analysis, we first state our assumptions as follows.

Assumption 1 *(Smoothness). f_i is L-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 *(Bounded variance of global gradient without perturbation). The global variability of the local gradient of the loss function without perturbation δ_i 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 *(Bounded variance of stochastic gradient). The stochastic gradient $\nabla f_i(w, \xi_i)$, computed by the i-th client of model parameter w using mini-batch ξ_i is an unbiased estimator $\nabla F_i(w)$ with variance bounded by σ^2, i.e.,*

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

$\forall i \in [N]$, where the expectation is over all local datasets.

Assumptions 1 and 2 are standard in general non-convex FL studies (Li et al., 2020b; Karimireddy et al., 2020; Reddi et al., 2020; Karimireddy et al., 2021; Yang et al., 2021) in order to assume the loss function continuous and bound the heterogeneity of FL systems. Note that we consider that σ_g^2 mainly depends on the data-heterogeneity, and the perturbation should be calculated. Hence, we only bound it without perturbation. We will present the upper bound of $\|\nabla f_i(\tilde{w}^r) - \nabla f(\tilde{w}^r)\|^2$ in Appendix A. Assumption 3 bounds the variance of stochastic gradient. Although many FL studies use the similar assumption to bound the stochastic gradient variance (Li et al., 2020b; Karimireddy et al., 2020), the definition is $\mathbb{E}_{\xi_i}\|\nabla F_i(w, \xi_i) - \nabla F_i(w)\|^2 \leq \sigma_l^2$, which is not easy to measure the value of σ_l^2, and the upper bound of σ_l^2 may be closed to $+\infty$. In this paper, Assumption 3 is considered as the norm of difference in unit vectors that can be upper bounded by the arc length on a unit circle. Therefore, σ^2 should be less than π^2. Clearly, this assumption is tighter than existing FL studies.

3.1. Convergence Analysis of FedSAM

We now state our convergence results for FedSAM algorithm. The detailed proof is in Appendix B.

Theorem 3.1 *Let the learning rates be chosen as* $\eta_l = \mathcal{O}(\frac{1}{\sqrt{RKL}})$, $\eta_g = \sqrt{KN}$ *and the perturbation amplitude* ρ *proportional to the learning rate, e.g.,* $\rho = \mathcal{O}(\frac{1}{\sqrt{R}})$. *Under Assumptions 1-3 and full client participation, the sequence of iterates generated by* `FedSAM` *in Algorithm 1 satisfies:*

$$\mathcal{O}\bigg(\frac{LF}{\sqrt{RKN}} + \frac{\sigma_g^2}{R} + \frac{L^2\sigma_l^2}{R^{3/2}\sqrt{KN}} + \frac{L^2}{R^2}\bigg),$$

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 and $S \geq K$, *if we choose the learning rates* $\eta_g = \mathcal{O}(\frac{1}{\sqrt{RKL}})$, $\eta_g = \sqrt{KS}$ *and* $\rho = \mathcal{O}(\frac{1}{\sqrt{R}})$, *the sequence of iterates generated by* `FedSAM` *in Algorithm 1 satisfies:*

$$\mathcal{O}\bigg(\frac{LF}{\sqrt{RKS}} + \frac{\sqrt{K}G^2}{\sqrt{RS}} + \frac{L^2\sigma^2}{R^{3/2}K} + \frac{L^2}{R^2}\bigg).$$

Remark 3.2 For the full and partial client participation strategies of `FedSAM` algorithm in this theorem, the dominant terms of the convergence rate are $\mathcal{O}(\frac{L}{\sqrt{RKN}})$ and $\mathcal{O}(\frac{\sqrt{K}}{\sqrt{SR}})$ by properly choosing the learning rates η_l and η_g, which match the best convergence rate in existing general non-convex FL studies (Karimireddy et al., 2020; Yang et al., 2021; Acar et al., 2021). Since the convergence rate structures in this theorem of these two strategies are similar, it indicates that uniformly sampling does not result in fundamental changes of convergence. In addition, both convergence rates include four main terms with an additional term compared to (Karimireddy et al., 2020; Yang et al., 2021; Acar et al., 2021). Note that we only show the dominant part of each term in the main paper. The detailed proof can be found in Appendix.

Remark 3.3 The additional term $\mathcal{O}(\frac{L^2}{R^2})$ comes from the additional SGD step for smoothness via SAM local optimizer in (4). However, this term can be negligible due to its higher order. More specifically, since the smoothness is due to the local training, we can combine it with the local training term, i.e., $\mathcal{O}(\frac{\sigma^2}{R^{3/2}K} + \frac{1}{R^2})$. Clearly, this term also achieves speedup than the existing best rate, i.e., $\mathcal{O}(\frac{1}{R})$. For the partial participation strategy, the dominant term is due to fewer clients participating and random sampling (heterogeneity), i.e., $\mathcal{O}(\frac{\sqrt{K}G^2}{RS})$. The convergence rate improves substantially as the number of clients increases, which matches the results of partial client participation FL (Karimireddy et al., 2020; Yang et al., 2021; Acar et al., 2021). Intuitively, increasing the convergence rate of this term is because SAM optimizer can make the global model more generalization and reduce the distribution shift.

Remark 3.4 The `FedRobust` (Reisizadeh et al., 2020) algorithm is an adversarial learning framework in FL setting, which is based on the similar idea of `FedSAM`. It has the convergence rate of $\mathcal{O}(\frac{Lf}{(RN)^{1/3}} + \frac{L^2}{R^{1/3}N^{2/3}})$, and it does not perform well from the convergence perspective compared with `FedSAM`. Since multiple gradient descents steps should be computed in each local training epoch, `FedRobust` will waste more running time and computational cost to process the local training.

3.2. Generalization Bounds of `FedSAM`

Based on the margin-based generalization bounds in (Neyshabur et al., 2018; Bartlett et al., 2017; Farnia et al., 2018; Reisizadeh et al., 2020), we propose the generalization error of `FedSAM` algorithm with the general neural network as follows:

$$\mathcal{L}_\gamma^{\text{SAM}}(F(w)) := \frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_i\bigg(F_i(w+\delta_i,\mathbf{X})[Y] - \max_{j\neq Y} F_i(w+\delta_i,\mathbf{X})[j] \leq \gamma\bigg). \tag{5}$$

Here, $F_i(w+\delta_i,\mathbf{X})$ is the loss function solving by SAM local optimizer for client i in (2), $\mathbf{X}$ is an input, $\mathbb{P}_i$ is the probability of the underlying distribution of client i, and $F_i(w+\delta_i,\mathbf{X})[j]$ is the output of the last softmax layer for label j about the training neural network. It is worth noting that γ is a constant, and for $\gamma = 0$, (5) can be simplified to the average misclassification rate with the distribution shift, which is denoted by $\mathcal{L}^{\text{SAM}}$. In addition, we use $\hat{\mathcal{L}}_\gamma^{\text{SAM}}(w)$ as the above margin risk to represent the empirical distribution of training samples, and hence we use $\hat{\mathbb{P}}_i$ to replace the underlying $\mathbb{P}_i$ to be the empirical probability, which is calculated by the m training samples on client i.

The following theorem aims to bound the difference of the empirical and the margin-based error defined in (5) under a general deep neural network. We use the spectral norm based generalization bound framework (Neyshabur et al., 2018; Farnia et al., 2018; Chatterji et al., 2019) to prove the next theorem. In order to demonstrate the margin-based error bounds, we assume that the neural network with smooth ReLU activation functions θ are 1-Lipschitz activation functions. The detailed proof is shown in Appendix C.

Theorem 3.5 *Let input* $\mathbf{X}$ *be an* $n \times n$ *image whose norm is bounded by* A, $f(w)$ *be the classification function with* d *hidden-layer neural network with* h *units per hidden-layer, and satisfy* 1*-Lipschitz activation* $\theta(0) = 0$. *We assume the constant* $M \geq 1$ *for each layer* W_j *satisfies* $\frac{1}{M} \leq \frac{\|W_j\|}{\phi_w} \leq M$, *where* $\phi_w := (\prod_{j=1}^{d}\|W_j\|)^{1/d}$ *denotes the geometric mean of* $f(w)$*'s spectral norms across all layers. Then, for any margin value* γ, *size of local training dataset on each client* m, $\zeta > 0$, *with probability* $1-\zeta$ *over the training set, any parameter of SAM local optimizer* $\tilde{w} = w + \delta$ *such*

Algorithm 2 MoFedSAM algorithm.

Initialization: w^0, $\Delta^0 = 0$, ρ^0, momentum parameter β the number of local updates K.
for $r = 0, \ldots, R-1$ **do**
  Sample subset $\mathcal{S}^r \subseteq [N]$ of clients.
  $w_{i,0}^t = w^r$.
  for each client $i \in \mathcal{S}^r$ in parallel **do**
    for $k = 0, \ldots, K-1$ **do**
      Compute a local training estimate $g_{i,k}^r = \nabla F_i(w_{i,k}^r, \xi_{i,k}^r)$ of $\nabla F_i(w_{i,k}^r)$.
      Compute local model $w_{i,k}^r$ from (6).
    end for
    $\Delta_i^r = w_{i,K}^r - w^r$.
  end for
  $\Delta^{r+1} = -\frac{1}{\eta_l KS} \sum_{i \in \mathcal{S}^r} \Delta_i^r$.
  $w^{r+1} = w^r - \eta_g \Delta^{r+1}$.
end for

that $\max_{\mathbf{X} \in \mathcal{D}_i} \|F_i(w) - f(\tilde{w})\| \leq \frac{\gamma}{8}$, *we can obtain the following generalization bound:*

$$\mathcal{L}^{SAM}(F(w)) \leq \hat{\mathcal{L}}_{\gamma}^{SAM}(F(w+\delta)) + \mathcal{O}\left(\frac{32Ad^2h\log(dh)Q(F(w)) + d\log\frac{Nmd\log(M)}{\zeta}}{\gamma^2 m}\right),$$

where $Q(F(w)) := \prod_{j=1}^{d} \|W_j\| \sum_{i=1}^{d} \frac{\|W_j\|_F^2}{\|W_j\|}$ *and* $\|W_j\|_F^2$ *is the Frobenius norm.*

Theorem 3.5 proposes a non-asymptotic bound on the generalization risk of FedSAM for general neural networks. The PAC-Bayesian bounds of SAM (Foret et al., 2021; Kwon et al., 2021; Zhuang et al., 2022; Du et al., 2021a) does not provide the insight about the underlying reason that results in generalization, i.e., how to choose the value of λ in the Gaussian noise $\mathcal{N}(0, \lambda I)$ to be the perturbation. In Theorem 3.5, we present the dependence of the perturbation δ and the different neural network parameters in which we can enforce the loss surface around a point in order to guarantee the smoothness.

4. Momentum FedSAM (MoFedSAM)

4.1. Algorithm of MoFedSAM

Since Δ^r serves as the direction for the global model, while FedSAM algorithm achieves efficient convergence rate theoretically, the influence of local optimizer cannot directly affect the global model, i.e., the term including σ_g^2 in the convergence rate. Note that Δ^r aggregates the global model information of participating clients, and reusing this information should be useful to guide the local training on the participated clients in next communication round, which is similar to momentum FL (Wang et al., 2019; Reddi et al., 2020; Karimireddy et al., 2021; Khanduri et al., 2021; Xu et al., 2021). Inspired by this motivation, we now provide our second algorithm, termed MoFedSAM, which aims to smooth and generalize the global model directly. The training procedure of k-th local training epoch in round r is formulated as follows:

$$\begin{cases} \tilde{w}_{i,k}^r = w_{i,k}^r + \rho \frac{g_{i,k}^r}{\|g_{i,k}^r\|} \\ v_{i,k}^r = \beta \tilde{g}_{i,k}^r + (1-\beta)\Delta^r \\ w_{i,k}^r = w_{i,k}^r - \eta_l v_{i,k}^r, \end{cases} \tag{6}$$

where β is the momentum rate. If $\beta = 1$, MoFedSAM is equivalent to FedSAM. From (6), we can see that the global model information Δ^r directly contributes the local training epoch, since $w_{i,k}^r$ includes $\tilde{g}_{i,k}^r$ and Δ^r at the same time. Therefore, it indicates that MoFedSAM make the local and global models smoothness at the same time. Especially, even if only a subset of clients are sampled in each communication round, the information of gradients of previous local model updates can be still contained in Δ^r. Therefore, MoFedSAM also works well of partial client participation FL. More specifically, the global model information term Δ^r is considered as an approximation to the gradient of the global model $\nabla f(\tilde{w})$, i.e., $\Delta^r \approx \nabla f(\tilde{w}^r)$. One advantage is that MoFedSAM adds a correction term to the local gradient direction, and it also asymptotically aligns with the difference between global and local gradient. It is worth noting that we use (G, B)-BGD in Assumption 2 to prove the convergence rate, which is tighter than $(G, 0)$-BGD in (Xu et al., 2021).

4.2. Convergence Analysis of MoFedSAM

Next theorem is the convergence rate of MoFedSAM algorithm, and the detailed proof is in Appendix D.

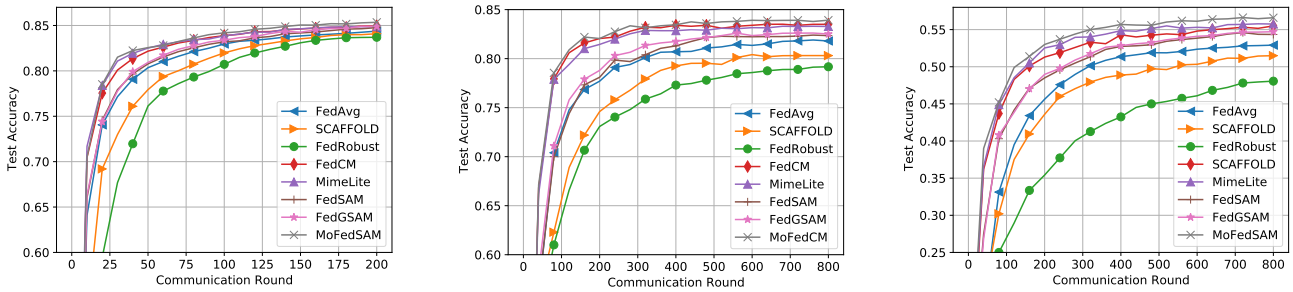
Theorem 4.1 *Let the learning rates be chosen as* $\eta_l = O(\frac{1}{\sqrt{R\beta KL}})$, $\eta_g = O(\frac{\sqrt{KN}}{\sqrt{R\beta L}})$ *and the perturbation amplitude* ρ *proportional to the learning rate, e.g.,* $\rho = \mathcal{O}(\frac{1}{\sqrt{R}})$. *Under the Assumptions 1-3, any momentum parameter* $\beta \leq \frac{1}{2}$ *and the full client participation strategy, the sequence of* $\{\tilde{w}^r\}$ *generated by* MoFedSAM *in Algorithm 2 satisfies:*

$$\mathcal{O}\left(\frac{\beta LF}{\sqrt{RKN}} + \frac{\beta\sigma_g^2}{RL^2} + \frac{L\sigma^2}{R^2\beta} + \frac{\beta L^2}{R^2}\right),$$

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 $\eta_g = \mathcal{O}(\frac{1}{\sqrt{R\beta KL}})$, $\eta_g = \mathcal{O}(\frac{\sqrt{KS}}{\sqrt{R\beta L}})$ *and* $\rho = \mathcal{O}(\frac{1}{\sqrt{R}})$, *the following convergence holds:*

$$\mathcal{O}\left(\frac{\beta LF}{\sqrt{RKS}} + \frac{\beta\sqrt{K}\sigma_g^2}{\sqrt{RS}} + \frac{L^2\sigma^2}{R^{3/2}K} + \frac{\sqrt{K}L^2}{R^{3/2}\sqrt{S}}\right).$$



(a) EMNIST dataset. (b) CIFAR-10 dataset. (c) CIFAR-100 dataset.

Figure 1. Testing accuracy on different datasets.

Table 1. Average (standard deviation) training accuracy and testing accuracy. Communication round to achieve the targeted testing accuracy: EMNIST 80%, CIFAR-10 80% and CIFAR-100 50%.

Algorithm	EMNIST			CIFAR-10			CIFAR-100		
	Train	Validation	Round	Train	Validation	Round	Train	Validation	Round
FedAvg	95.07 (0.94)	84.38 (4.03)	43	93.15 (1.44)	81.87 (5.09)	307	79.57 (1.84)	53.57 (5.40)	302
SCAFFOLD	93.85 (1.31)	84.09 (4.56)	69	91.76 (1.89)	80.61 (5.64)	546	78.49 (2.02)	51.49 (5.87)	551
FedRobust	93.17 (0.62)	83.70 (3.37)	91	90.82 (1.27)	79.63 (4.21)	847	76.80 (1.70)	49.06 (4.75)	893
FedCM	96.16 (1.14)	84.85 (4.11)	28	95.61 (1.50)	83.30 (4.77)	136	82.13 (1.96)	55.50 (5.04)	182
MimeLite	96.22 (1.16)	84.88 (4.22)	25	95.73 (1.56)	83.18 (4.65)	152	82.46 (2.00)	55.73 (5.11)	189
FedSAM	95.73 (0.49)	84.75 (3.04)	38	94.20 (1.08)	83.06 (3.87)	269	81.04 (1.59)	54.69 (4.36)	245
MoFedSAM	96.42 (0.42)	85.07 (2.95)	24	95.67 (1.16)	83.92 (3.65)	124	82.62 (1.53)	56.60 (4.42)	124

Remark 4.2 When T is sufficiently large compared to K, convergence rates under full and partial client participation strategies of MoFedSAM algorithm are $\mathcal{O}(\frac{\sqrt{\beta L}}{\sqrt{RKN}} + \frac{\beta}{RL^2})$ and $\mathcal{O}(\frac{\beta L}{\sqrt{RKS}} + \frac{\beta\sqrt{K}}{\sqrt{RS}})$. The momentum parameter β is small enough, i.e., 0.1 (Karimireddy et al., 2021; Xu et al., 2021), from which the effect is important for convergence, due to the number of local epochs setting less than 20 in usual (Reddi et al., 2020; Yang et al., 2021; Acar et al., 2021). Therefore, our convergence results achieve speedup compared with FedSAM. We also note that the convergence related to the local training is $\mathcal{O}(\frac{L}{R^2\beta} + \frac{\beta L^2}{R^2})$ and $\mathcal{O}(\frac{L^2}{R^{3/2}K} + \frac{\sqrt{K}L^2}{R^{3/2}\sqrt{S}})$, where the second part comes from sharpness, and it can be negligible. From the convergence analysis of FedCM (Xu et al., 2021), i.e., $\mathcal{O}(\frac{\sqrt{KSL}}{\sqrt{R}} + \frac{L}{\beta^{2/3}R^{2/3}})$, we can see that MoFedSAM achieves speedup both on the dominant part and local training part. The analysis indicates the benefit of bridging the sharpness between local and global models.

5. Experiments

We evaluate our proposed algorithms on extensive and representative datasets and learning models to date. To accomplish this, we conduct experiments on three learning models across three datasets comparing to five FL benchmarks with varying different parameters.

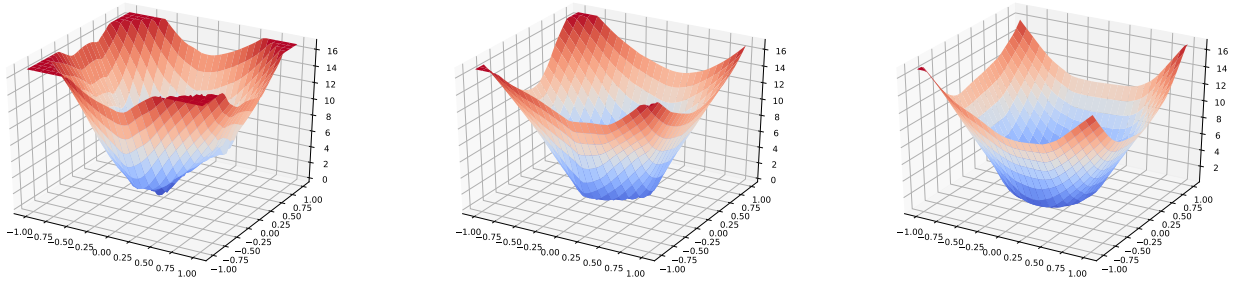
5.1. Experimental Setup

Benchmarks and hyper-parameters. We consider five FL benchmarks: without momentum FL FedAvg (McMahan et al., 2017), SCAFFOLD (Karimireddy et al., 2020), FedRobust (Reisizadeh et al., 2020); momentum FL MimeLite (Karimireddy et al., 2021) and FedCM (Xu et al., 2021). The learning rates are individually tuned and other optimizer hyper-parameters such as $\rho = 0.5$ for SAM and $\beta = 0.1$ for momentum, unless explicitly stated otherwise. We refer to Appendices E-F for detailed experimental setup and additional ablation studies.

Datasets and models. We use three images datasets: EMNIST (Cohen et al., 2017), CIFAR-10, and CIFAR-100 (Krizhevsky et al., 2009). Our cross-device FL setting includes 100 clients in total with participation rate 20%. In each communication round, each client is sampled independently of each other, with probability 0.2. We simulate the data heterogeneity by sampling the label ratios from a Dirchlet distribution with parameter 0.6 (Acar et al., 2021), the number of local epochs is set as $K = 10$ by default. We adopt two learning models on each dataset: (i) CNN on EMNIST with batch 32 and (ii) ResNet-18 (He et al., 2016) on CIFAR-10 and CIFAR-100 with batch 128. The detailed experimental setup and other additional experiments and ablation studies will be shown in Appendices E-F.

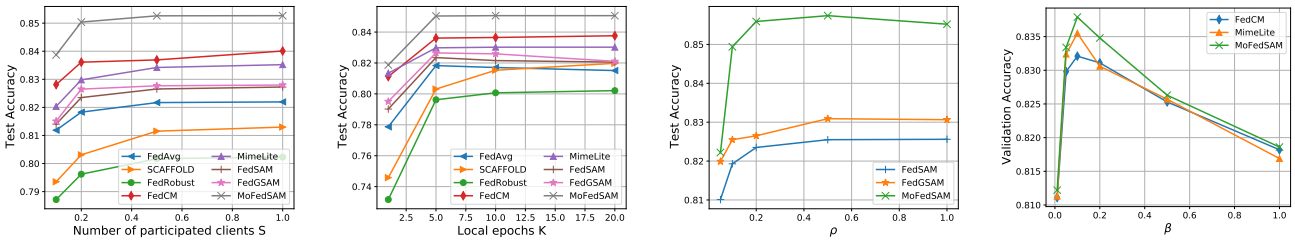
Table 2. Impact of the heterogeneity on CIFAR-10 dataset (IID, Dirichlet 0.6 and Dirichlet 0.3).

Algorithm	IID			Dirichlet 0.6			Dirichlet 0.3		
	Train	Validation	Round	Train	Validation	Round	Train	Validation	Round
FedAvg	94.95 (1.01)	85.97 (3.53)	238	93.15 (1.44)	81.87 (5.09)	307	91.89 (1.63)	77.39 (5.62)	-
SCAFFOLD	93.04 (1.13)	83.82 (3.72)	290	91.76 (1.89)	80.61 (5.64)	546	90.02 (2.08)	75.67 (5.93)	-
FedRobust	91.63 (0.91)	82.44 (3.15)	361	90.82 (1.27)	79.63 (4.21)	847	89.72 (1.42)	73.11 (5.11)	-
FedCM	97.02 (1.10)	88.14 (3.33)	87	95.61 (1.50)	83.30 (4.77)	136	93.88 (1.67)	81.34 (5.50)	583
MimeLite	97.16 (1.08)	88.53 (3.53)	82	95.73 (1.56)	83.18 (4.65)	152	93.97 (1.72)	81.83 (5.53)	548
FedSAM	95.42 (0.81)	87.36 (2.85)	205	94.20 (1.08)	83.06 (3.87)	269	92.90 (1.26)	79.82 (4.98)	816
MoFedSAM	97.22 (0.88)	88.96 (2.94)	75	95.67 (1.16)	83.92 (3.65)	124	94.12 (1.31)	83.35 (5.06)	490



(a) FedAvg. (b) FedSAM. (c) MoFedSAM.

Figure 2. Loss surface of FedAvg, FedSAM and MoFedSAM algorithm with ResNet-18 on CIFAR-10 dataset.



(a) Impact of S. (b) Impact of K. (c) Impact of ρ. (d) Impact of β.

Figure 3. Impacts of different parameters on CIFAR-10 dataset.

5.2. Performance Evaluation

(1) Performance with compared benchmarks. We first investigate the effect of our proposed algorithms with compared benchmarks on different datasets in Figure 1 and Table 1. From these results, we can clearly see that for the performance of without momentum FL: FedSAM $>$ FedAvg $>$ SCAFFOLD $>$ FedRobust, and the performance momentum FL: MoFedSAM $>$ MimeLite $>$ FedCM. Our proposed algorithms outperform other benchmarks both on accuracy and convergence perspectives. We do not compare the FL algorithms with momentum FL, since momentum FL is required to transmit more information than FL, e.g., Δ^{r+1}. This is the reason why momentum FL outperforms FL benchmarks. More specifically, to present the generalization performance, we show the deviation, i.e., best and worst local accuracy. In addition, the performance improvement on CIFAR-100 dataset is more obvious than others, since SAM optimizers perform more efficiently on more complicated datasets.

(2) Impact of Non-IID levels. In Tables 2, 3, 4 and 5, we can see that our proposed algorithms outperforms the benchmarks across different client distribution levels on the same FL categories. We consider heterogeneous client distributions by varying balanced-unbalanced, number of clients and participation levels settings on various datasets. Client distributions become more non-IID as we go from IID, Dirichlet 0.6 to Dirichlet 0.3 splits which makes global optimization more difficult. For example, as non-IID levels increasing, MoFedSAM achieves a higher test accuracy 0.43%, 1.24% and 1.52% and saving communication round 7, 40, and 59 than MimeLite on CIFAR-10 dataset. In summary, although almost all the algorithms perform well

enough for training dataset, the testing accuracy usually has a significant degradation especially the deviation of local clients. In Table 1, we can see that our proposed algorithms significantly decrease the deviation of local clients, which indicates that our proposed algorithms show enough generalization of the global model.

(3) Loss surface visualization. To visualize the sharpness of the flat minima obtained by `FedAvg`, `FedSAM` and `MoFedSAM`, we show the loss surface, which are trained with ResNet-18 under the CIFAR-10 dataset. We display the loss surfaces in Figure 3, following the plotting algorithm in (Li et al., 2018a). The x- and y-axes are two random sampled orthogonal Gaussian perturbations. We can clearly see that both `FedSAM` and `MoFedSAM` improve the sharpness significantly in comparison to `FedAvg`, which indicates that our proposed algorithms perform more generalization.

(4) Impact of other parameters. Here, we show the impact of different parameters, e.g., number of participated clients S, number of epochs K, perturbation radius ρ for our proposed algorithms and momentum value β in Figures 3, 7, 8 and 9. Our proposed algorithms outperform the same FL categories, i.e., with or without momentum. Similar to existing FL studies, increasing batch size and number of participated clients can improve the learning performance. Increasing the number of epochs K cannot guarantee better accuracy substantially, however, all the benchmarks perform worst when $K = 1$. The best ρ for each dataset is different, the best performance of ρ value is set as 0.2 for EMNIST, 0.5 for CIFAR-10 and 0.6 for CIFAR-100.

6. Conclusion

In this paper, we study the distribution shift coming from the data heterogeneity challenge of cross-device FL from a simple yet unique perspective by making global model generality. To this end, we propose two algorithms `FedSAM` and `MoFedSAM`, which do not generate more communication costs compared with existing FL studies. By deriving the convergence of general non-convex FL settings, these algorithms achieve competitive performance. Furthermore, we also provide the generalization bound of `FedSAM` algorithm. The extensive experiments strongly support that our proposed algorithms decrease the performance deviation among all local clients significantly.

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*, 2021.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 30: 6240–6249, 2017.

Caldarola, D., Caputo, B., and Ciccone, M. Improving generalization in federated learning by seeking flat minima. *arXiv preprint arXiv:2203.11834*, 2022.

Chatterji, N., Neyshabur, B., and Sedghi, H. The intriguing role of module criticality in the generalization of deep networks. In *International Conference on Learning Representations*, 2019.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

Cohen, G., Afshar, S., Tapson, J., and Van Schaik, A. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 2921–2926. IEEE, 2017.

Deng, Y., Kamani, M. M., and Mahdavi, M. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

Dieuleveut, A., Fort, G., Moulines, E., and Robin, G. Federated-em with heterogeneity mitigation and variance reduction. *Advances in Neural Information Processing Systems*, 34, 2021.

Du, J., Yan, H., Feng, J., Zhou, J. T., Zhen, L., Goh, R. S. M., and Tan, V. Y. Efficient sharpness-aware minimization for improved training of neural networks. *arXiv preprint arXiv:2110.03141*, 2021a.

Du, W., Xu, D., Wu, X., and Tong, H. Fairness-aware agnostic federated learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pp. 181–189. SIAM, 2021b.

Fallah, A., Mokhtari, A., and Ozdaglar, A. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33:3557–3568, 2020.

Farnia, F., Zhang, J., and Tse, D. Generalizable adversarial training via spectral normalization. In *International Conference on Learning Representations*, 2018.

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

Gong, X., Sharma, A., Karanam, S., Wu, Z., Chen, T., Doermann, D., and Innanje, A. Ensemble attention distillation for privacy-preserving federated learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15076–15086, 2021.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

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.

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. *arXiv preprint arXiv:1912.04977*, 2019.

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.

Karimireddy, S. P., Jaggi, M., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. Breaking the centralized barrier for cross-device federated learning. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

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. *arXiv preprint arXiv:1609.04836*, 2016.

Khanduri, P., SHARMA, P., Yang, H., Hong, M., Liu, J., Rajawat, K., and Varshney, P. STEM: A stochastic two-sided momentum algorithm achieving near-optimal sample and communication complexities for federated learning. In *Advances in Neural Information Processing Systems*, 2021.

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

Kwon, J., Kim, J., Park, H., and Choi, I. K. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5905–5914. PMLR, 2021.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. *Advances in Neural Information Processing Systems*, 31, 2018a.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018b.

Li, T., Sanjabi, M., Beirami, A., and Smith, V. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020a.

Li, T., Hu, S., Beirami, A., and Smith, V. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pp. 6357–6368. PMLR, 2021.

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*, 2020b.

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *arXiv preprint arXiv:1705.09056*, 2017.

Lin, T., Kong, L., Stich, S. U., and Jaggi, M. Ensemble distillation for robust model fusion in federated learning. In *Advances in Neural Information Processing Systems*, 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.

Mendieta, M., Yang, T., Wang, P., Lee, M., Ding, Z., and Chen, C. Local learning matters: Rethinking data heterogeneity in federated learning. *arXiv preprint arXiv:2111.14213*, 2021.

Mohri, M., Sivek, G., and Suresh, A. T. Agnostic federated learning. In *International Conference on Machine Learning*, pp. 4615–4625. PMLR, 2019.

Nesterov, Y. and Spokoiny, V. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

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*, 32:8026–8037, 2019.

Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečnỳ, J., Kumar, S., and McMahan, H. B. Adaptive federated optimization. In *International Conference on Learning Representations*, 2020.

Reisizadeh, A., Farnia, F., Pedarsani, R., and Jadbabaie, A. Robust federated learning: The case of affine distribution shifts. In *NeurIPS*, 2020.

Shafahi, A., Najibi, M., Xu, Z., Dickerson, J., Davis, L. S., and Goldstein, T. Universal adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5636–5643, 2020.

Singhal, K., Sidahmed, H., Garrett, Z., Wu, S., Rush, J. K., and Prakash, S. Federated reconstruction: Partially local federated learning. In *Advances in Neural Information Processing Systems*, 2021.

T Dinh, C., Tran, N., and Nguyen, T. D. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33, 2020.

Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12 (4):389–434, 2012.

Wang, J., Tantia, V., Ballas, N., and Rabbat, M. Slowmo: Improving communication-efficient distributed sgd with slow momentum. In *International Conference on Learning Representations*, 2019.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020.

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. In *International Conference on Learning Representations*, 2021.

Yoon, T., Shin, S., Hwang, S. J., and Yang, E. Fedmix: Approximation of mixup under mean augmented federated learning. In *International Conference on Learning Representations*, 2021.

Yuan, H., Morningstar, W., Ning, L., and Singhal, K. What do we mean by generalization in federated learning? *arXiv preprint arXiv:2110.14216*, 2021.

Zhu, Z., Hong, J., and Zhou, J. Data-free knowledge distillation for heterogeneous federated learning. *arXiv preprint arXiv:2105.10056*, 2021.

Zhuang, J., Gong, B., Yuan, L., Cui, Y., Adam, H., Dvornek, N. C., sekhar tatikonda, s Duncan, J., and Liu, T. Surrogate gap minimization improves sharpness-aware training. In *International Conference on Learning Representations*, 2022.

A. Preliminary Lemmas

For giving the theoretical analysis of the convergence rate of all proposed algorithms, we firstly state some preliminary lemmas as follows:

Lemma A.1 *(Relaxed triangle inequality). Let $\{v_1, \ldots, v_\tau\}$ be τ vectors in $\mathbb{R}^d$. Then, the following are true: (1) $\|v_i + v_j\|^2 \leq (1+a)\|v_i\|^2 + (1+\frac{1}{a})\|v_j\|^2$ for any $a > 0$, and (2) $\| \sum_{i=1}^{\tau} v_i\|^2 \leq \tau \sum_{i=1}^{\tau} \|v_i\|^2$.*

Lemma A.2 *For random variables $x_1, \ldots, x_n$, we have*

$$\mathbb{E}[\|x_1 + \cdots + x_n\|^2] \leq n\mathbb{E}[\|x_1\|^2 + \cdots + \|x_n\|^2].$$

Lemma A.3 *For independent, mean 0 random variables $x_1, \ldots, x_n$, we have*

$$\mathbb{E}[\|x_1 + \cdots + x_n\|^2] = \mathbb{E}[\|x_1\|^2 + \cdots + \|x_n\|^2].$$

Lemma A.4 *(Separating mean and variance for SAM). The stochastic gradient $\nabla F_i(w, \xi_i)$ computed by the i-th client at model parameter w using minibatch ξ is an unbiased estimator of $\nabla F_i(w)$ with variance bounded by σ^2. The gradient of SAM is formulated by*

$$\mathbb{E}\left[\left\| \sum_{k=0}^{K-1} g_{i,k}^r \right\|^2\right] \leq K \sum_{k=0}^{K-1} \mathbb{E}[\|\nabla F_i(w_{r,k}^i)\|^2] + \frac{KL^2\rho^2}{N}\sigma_l^2,$$
$$\mathbb{E}\left[\left\| \sum_{k=0}^{K-1} g_{i,k}^r \right\|^2\right] \leq K \sum_{k=0}^{K-1} \mathbb{E}[\|\nabla F_i(w_{r,k}^i)\|^2] + KL^2\rho^2\sigma_l^2.$$

Proof. For the first inequality, we can bound as follows

$$\begin{aligned}
\mathbb{E}\left[\left\| \sum_{k=0}^{K-1} g_{i,k}^r \right\|^2\right] &= \mathbb{E}\left[\left\| \sum_{k=0}^{K-1} g_{i,k}^r \right\|^2\right] + \mathbb{E}\left[\left\| \sum_{k=0}^{K-1} (g_{i,k}^r - \nabla F(w_{i,k}^r)) \right\|^2\right] \\
&\overset{(a)}{\leq} K \sum_{k=0}^{K-1} \mathbb{E}[\|g_{i,k}^r\|^2] + L^2 \sum_{k=0}^{K-1} \mathbb{E}\left[\left\| \frac{1}{N} \sum_{i\in[N]} (w_{i,k}^r + \delta_{i,k}^r(\tilde{w}_{i,k}^r; \xi_{i,k}^r) - w_{i,k}^r - \delta_{i,k}^r(\tilde{w}_{i,k}^r)) \right\|^2\right] \\
&\overset{(b)}{\leq} K \sum_{k=0}^{K-1} \mathbb{E}[\|g_{i,k}^r\|^2] + \frac{KL^2\rho^2\sigma_l^2}{N}.
\end{aligned}$$

where (a) is from Assumption 1 and (b) is from Assumption 3 and Lemma A.3. Similarly, we can obtain the second inequality, and hence we omit it here. □

Lemma A.5 *(Bounded global variance of $\|\nabla F_i(w+\delta_i) - \nabla F(w+\delta)\|^2$.) An immediate implication of Assumptions 1 and 2, the variance of local and global gradients with perturbation can be bounded as follows:*

$$\|\nabla F_i(w+\delta_i) - \nabla F(w+\delta)\|^2 \leq 3\sigma_g^2 + 6L^2\rho^2.$$

Proof.

$$\begin{aligned}
\|\nabla f_i(\tilde{w}) - \nabla f(\tilde{w})\|^2 &= \|\nabla F_i(w+\delta_i) - \nabla F(w+\delta)\|^2 \\
&= \|\nabla F_i(w+\delta_i) - \nabla F_i(w) + \nabla F_i(w) - \nabla F(w) + \nabla F(w) - \nabla F(w+\delta)\|^2 \\
&\overset{(a)}{\leq} 3\|\nabla F_i(w+\delta_i) - \nabla F_i(w)\|^2 + 3\|\nabla F_i(w) - \nabla F(w)\|^2 + 3\|\nabla F(w) - \nabla F(w+\delta)\|^2 \\
&\overset{(b)}{\leq} 3\sigma_g^2 + 6L^2\rho^2,
\end{aligned}$$

where (a) is from Lemma A.2 and (b) is from Assumption 1, 2 and the perturbation is bounded by ρ. □

Algorithm 3 FedSAM: Federated Sharpness Aware Minimization

Initialization: w_0, ρ_0, γ the number of local updates K, batch size b, local learning η_l and global learning rate η_g.
for each round $r = 0, \dots, R-1$ **do**
  Sample subset $\mathcal{S}^r \subseteq [N]$ of clients.
  communicate w^r to all clients $i \in \mathcal{S}^r$.
  for each client $i \in \mathcal{S}^r$ in parallel **do**
    initialize local model $w_{i,0}^r \leftarrow w^r$.
    for $k = 0, \dots, K-1$ **do**
      Compute $g_{i,k-1}^r$ by taking an estimation $\nabla F_i(w_{i,k-1}^r, \xi_i^r)$ of $\nabla F_i(w_{i,k-1}^r)$.
      $\tilde{w}_{i,k-1}^r = w_{i,k-1}^r + \rho \frac{g_{i,k-1}^r}{\|g_{i,k-1}^r\|}$.
      Compute $\tilde{g}_{i,k-1}^r$ by taking an estimation $\nabla f_i(\tilde{w}_{i,k-1}^r, \xi_i^r)$ of $\nabla f_i(\tilde{w}_{i,k-1}^r, \xi_i^r)$.
      $w_{i,k}^r = w_{i,k-1}^r - \eta_l \tilde{g}_{i,k-1}^r$.
    end for
    $\Delta_i^r = w_{i,K}^r - w^r$.
  end for
  $\Delta^{r+1} = \frac{1}{S} \sum_{i \in \mathcal{S}^r} \Delta_i^r$.
  $w^{r+1} = w^r + \eta_g \Delta^r$.
end for

B. Convergence Analysis for FedSAM

B.1. Description of FedSAM Algorithm and Key Lemmas

We outline the FedSAM algorithm in Algorithm 3. In round r, we sample $\mathcal{S}^r \subseteq [N]$ clients with $|\mathcal{S}^r| = S$ and then perform the following updates:

- Starting from the shared global parameters $w_{i,0}^r = w^{r-1}$, we update the local parameters for $k \in [K]$

$$\tilde{w}_{i,k}^r = w_{i,k-1}^r + \rho \frac{g_{i,k-1}^r}{\|g_{i,k-1}^r\|}$$
$$w_{i,k}^r = w_{i,k-1}^r - \eta_l \tilde{g}_{i,k-1}^r,$$

- After K times local epochs, we obtain the following

$$\Delta_i^r = w_{i,K}^r - w^r. \tag{7}$$

- Compute the new global parameters using only updates from the clients $i \in \mathcal{S}^r$ and a global step-size η_g:

$$\Delta^{r+1} = \frac{1}{S} \sum_{i \in \mathcal{S}^r} \Delta_i^r$$
$$w^{r+1} = w^r + \eta_g \Delta^r.$$

Lemma B.1 *(Bounded $\mathcal{E}_\delta$ of FedSAM). Suppose our functions satisfies Assumptions 1-2. Then, the updates of FedSAM for any learning rate satisfying $\eta_l \leq \frac{1}{4KL}$ have the drift due to $\delta_{i,k} - \delta$:*

$$\mathcal{E}_\delta = \frac{1}{N} \sum_i \mathbb{E}[\|\delta_{i,k} - \delta\|^2] \leq 2K^2 \beta^2 \eta_l^2 \rho^2.$$

Proof. Recall the definitions of δ and $\delta_{i,k}$ as follows:

$$\delta = \rho \frac{\nabla F(w)}{\|\nabla F(w)\|}, \quad \delta_{i,k} = \rho \frac{\nabla F_i(w_{i,k}, \xi_i)}{\|\nabla F_i(w_{i,k}, \xi_i)\|}.$$

If the local learning rate η_l is small, the gradient of one epoch $\nabla F_i(w_{i,k}, \xi_i)$ is small. Based on the first order Hessian approximation, the expected gradient is

$$\nabla F_i(w_{i,k}) = \nabla F_i(w_{i,k-1} + g_{i,k-1}) = \nabla F_i(w_{i,k-1}) + H\eta_l g_{i,k-1} + O(\|\eta_l g_{i,k-1}\|^2),$$

where H is the Hessian at $w_{i,k-1}$. Therefore, we have

$$\mathbb{E}[\|\delta_{i,k} - \delta\|^2] = \rho^2 \mathbb{E}\left[\left\| \frac{\nabla F_i(w_{i,k})}{\|\nabla F_i(w_{i,k})\|} - \frac{\nabla F_i(w)}{\|\nabla F_i(w)\|} \right\|^2\right] \leq \rho^2 \phi_{i,k}, \tag{8}$$

where $\phi_{i,k}$ is the square of the angle between the unit vector in the direction of $\nabla F_i(w_{i,k})$ and $\nabla F_i(w_{i,0})$. The inequality follows from that (1) $\left\| \frac{\nabla F_i(\cdot)}{\|\nabla F_i(\cdot)\|} \right\|^2 < 1$, and hence we replace δ with a unit vector in corresponding directions multiplied by ρ^2 and obtain the upper bound, (2) the norm of difference in unit vectors can be upper bounded by the square of the arc length on a unit circle. When the learning rate η_l and the local model update of one epoch $\nabla F_i(w_{i,k})$ are small, $\phi_{i,k}$ is also small. Based on the first order Taylor series, i.e., $\tan x = x + O(x^2)$, we have

$$\begin{aligned}
\tan \phi_{i,k} &= \frac{\|\nabla F_i(w_{i,k}) - \nabla F_i(w_{i,0})\|^2}{\|\nabla F_i(w_{i,0})\|^2} + O(\phi_{i,k}^2) \\
&= \frac{\|\nabla F_i(w_{i,k-1}) - H\eta_l g_{i,k-1} - O(\|\eta_l g_{i,k-1}\|^2) - \nabla F_i(w_{i,0})\|^2}{\|\nabla F_i(w_{i,0})\|^2} + O(\phi_{i,k}^2) \\
&\overset{(a)}{\leq} \left(1 + \frac{1}{K-1}\right) \frac{\|\nabla F_i(w_{i,k-1}) - \nabla F_i(w_{i,0})\|^2}{\|\nabla F_i(w_{i,0})\|^2} + \frac{K\|H\eta_l g_{i,k-1} + O(\|\eta_l g_{i,k-1}\|^2)\|^2}{\|\nabla F_i(w_{i,0})\|^2} + O(\phi_{i,k}^2) \\
&\overset{(b)}{\leq} \left(1 + \frac{1}{K-1}\right) \frac{\|\nabla F_i(w_{i,k-1}) - \nabla F_i(w_{i,0})\|}{\|\nabla F_i(w_{i,0})\|} + KL^2\eta_l^2,
\end{aligned}$$

where (a) is from Lemma A.1 with $a = \frac{1}{K-1}$ and (b) is due to maximum eigenvalue of H is bounded by L because F function is L-smooth. Unrolling the recursion above, we have

$$\frac{1}{N}\sum_{i\in[N]} \frac{\|\nabla F_i(w_{i,k}) - \nabla F_i(w_{i,0})\|^2}{\|\nabla F_i(w_{i,0})\|^2} + O(\phi_{i,k}^2) \leq \sum_{\tau=1}^{k-1} \left(1 + \frac{1}{K-1}\right)^{\tau} KL^2\eta_l^2 \leq 2K^2L^2\eta_l^2. \tag{9}$$

Plugging (9) into (8), we have

$$\mathcal{E}_\delta = \frac{1}{N}\sum_{i\in[N]} \mathbb{E}[\|\delta_{i,k} - \delta\|^2] \leq 2K^2L^2\eta_l^2\rho^2.$$

This completes the proof. □

Lemma B.2 *(Bounded $\mathcal{E}_w$ of `FedSAM`). Suppose our functions satisfies Assumptions 1-2. Then, the updates of `FedSAM` for any learning rate satisfying $\eta_l \leq \frac{1}{10KL}$ have the drift due to $w_{i,k} - w$:*

$$\mathcal{E}_w = \frac{1}{N}\sum_i \mathbb{E}[\|w_{i,k} - w\|^2] \leq 5K\eta_l^2(2L^2\rho^2\sigma_l^2 + 6K(3\sigma_g^2 + 6L^2\rho^2) + 6K\|\nabla f(\tilde{w})\|^2) + 24K^3\eta_l^4L^4\rho^2.$$

Proof. Recall that the local update on client i is $w_{i,k} = w_{i,k-1} - \eta_l \tilde{g}_{i,k-1}$. Then,

$$
\begin{aligned}
&\mathbb{E}\|w_{i,k} - w\|^2 = \mathbb{E}\|w_{i,k-1} - w - \eta_l \tilde{g}_{i,k-1}\|^2 \\
&\overset{(a)}{\leq} \mathbb{E}\|w_{i,k-1} - w - \eta_l(\tilde{g}_{i,k-1} - \nabla f_i(\tilde{w}_{i,k-1}) + \nabla f_i(\tilde{w}_{i,k-1}) - \nabla f_i(\tilde{w}) + \nabla f_i(\tilde{w})) - \nabla f(\tilde{w}) + \nabla f(\tilde{w})\|^2 \\
&\overset{(b)}{\leq} \left(1 + \frac{1}{2K-1}\right)\mathbb{E}\|w_{i,k-1} - w\|^2 + \mathbb{E}\|\eta_l(\tilde{g}_{i,k-1} - \nabla f_i(\tilde{w}_{i,k-1}))\|^2 \\
&+ 6K\mathbb{E}\|\eta_l(\nabla f_i(\tilde{w}_{i,k-1}) - \nabla f_i(\tilde{w}))\|^2 + 6K\mathbb{E}\|\eta_l(\nabla f_i(\tilde{w}) - \nabla f(\tilde{w}))\|^2 + 6K\|\eta_l \nabla f(\tilde{w})\|^2 \\
&\overset{(c)}{\leq} \left(1 + \frac{1}{2K-1} + 2L^2\eta_l^2\right)\mathbb{E}\|w_{i,k-1} - w\|^2 + 2\eta_l^2 L^2 \rho^2 \sigma_l^2 + 12K\eta_l^2 L^2 \mathbb{E}\|w_{i,k-1} - w\|^2 \\
&+ 12KL^2\eta_l^2\mathbb{E}\|\delta_{i,k-1} - \delta\|^2 + 6K\eta_l^2\mathbb{E}\|\nabla f_i(\tilde{w}) - \nabla f(\tilde{w})\|^2 + 6K\|\nabla f(\tilde{w})\|^2 \\
&\overset{(d)}{\leq} \left(1 + \frac{1}{2K-1} + 12K\eta_l^2 L^2 + 2L^2\eta_l^2\right)\mathbb{E}\|w_{i,k-1} - w\|^2 + 2\eta_l^2 L^2 \rho^2 \sigma_l^2 + 12KL^2\eta_l^2\mathbb{E}\|\delta_{i,k} - \delta\|^2 \\
&+ 6K\eta_l^2(3\sigma_g^2 + 6L^2\rho^2) + 6K\|\nabla f(\tilde{w})\|^2,
\end{aligned}
$$

where (a) follows from the fact that $\tilde{g}_{i,k-1}$ is an unbiased estimator of $\nabla f_i(\tilde{w}_{i,k-1})$ and Lemma A.3; (b) is from Lemma A.2; (c) is from Assumption 3 and Lemma A.2 and (d) is from Lemma A.5.

Averaging over the clients i and learning rate satisfies $\eta_l \leq \frac{1}{10KL}$, we have

$$
\begin{aligned}
\frac{1}{N}\sum_{i\in[N]}\mathbb{E}\|w_{i,k} - w\|^2 &\leq \left(1 + \frac{1}{2K-1} + 12K\eta_l^2 L^2 + 2L^2\eta_l^2\right)\frac{1}{N}\sum_{i\in[N]}\mathbb{E}\|w_{i,k-1} - w\|^2 \\
&+ 2\eta_l^2 L^2\rho^2\sigma_l^2 + 12KL^2\eta_l^2\frac{1}{N}\sum_{i\in[N]}\mathbb{E}\|\delta_{i,k} - \delta\|^2 + 6K\eta_l^2(3\sigma_g^2 + 6L^2\rho^2) + 6K\|\nabla f(\tilde{w})\|^2 \\
&\overset{(a)}{\leq} \left(1 + \frac{1}{K-1}\right)\frac{1}{N}\sum_{i\in[N]}\mathbb{E}\|w_{i,k-1} - w\|^2 + \eta_l^2 L^2\rho^2\sigma_l^2 \\
&+ 12KL^2\eta_l^2\frac{1}{N}\sum_{i\in[N]}\mathbb{E}\|\delta_{i,k} - \delta\|^2 + 6K\eta_l^2(3\sigma_g^2 + 6L^2\rho^2) + 6K\|\nabla f(\tilde{w})\|^2 \\
&\leq \sum_{\tau=0}^{k-1}\left(1 + \frac{1}{K-1}\right)^{\tau}[2\eta_l^2 L^2\rho^2\sigma_l^2 + 6K\eta_l^2(3\sigma_g^2 + 6L^2\rho^2) + 6K\|\nabla f(\tilde{w})\|^2] + 12KL^2\eta_l^2\frac{1}{N}\sum_{i\in[N]}\mathbb{E}\|\delta_{i,k} - \delta\|^2 \\
&\overset{(b)}{\leq} 5K\eta_l^2(2L^2\rho^2\sigma_l^2 + 6K(3\sigma_g^2 + 6L^2\rho^2) + 6K\|\nabla f(\tilde{w})\|^2) + 24K^3\eta_l^4 L^4\rho^2,
\end{aligned}
$$

where (a) is due to the fact that $\eta_l \leq \frac{1}{10KL}$ and (b) is from Lemma B.1. □

B.2. Convergence Analysis of Full client participation `FedSAM`

Lemma B.3

$$
\langle \nabla f(\tilde{w}^r), \mathbb{E}_r[\Delta^r + \eta_l K \nabla f(\tilde{w}^r)]\rangle \leq \frac{\eta_l K}{2}\|\nabla f(\tilde{w}^r))\|^2 + K\eta_l L^2\mathcal{E}_w + K\eta_l L^2\mathcal{E}_\delta - \frac{\eta_l}{2KN^2}\mathbb{E}_r\left\|\sum_{i,k}\nabla f_i(\tilde{w}_{i,k})\right\|^2.
$$

Proof.

$$
\begin{aligned}
&\langle \nabla f(\tilde{w}^r), \mathbb{E}_r[\Delta^r + \eta_l K \nabla f(\tilde{w}^r)] \rangle \\
&\overset{\text{(a)}}{=} \frac{\eta_l K}{2}\|\nabla f(\tilde{w}^r))\|^2 + \frac{\eta_l}{2KN^2}\mathbb{E}_r\left\|\sum_{i,k}\nabla f_i(\tilde{w}^r_{i,k}) - \nabla f_i(\tilde{w}^r)\right\|^2 - \frac{\eta_l}{2KN^2}\mathbb{E}_r\left\|\sum_{i,k}\nabla f_i(\tilde{w}^r_{i,k})\right\|^2 \\
&\overset{\text{(b)}}{\leq} \frac{\eta_l K}{2}\|\nabla f(\tilde{w}^r))\|^2 + \frac{\eta_l}{2N}\sum_{i,k}\mathbb{E}_r\|\nabla f_i(\tilde{w}^r_{i,k}) - \nabla f_i(\tilde{w}^r)\|^2 - \frac{\eta_l}{2KN^2}\mathbb{E}_r\left\|\sum_{i,k}\nabla f_i(\tilde{w}^r_{i,k})\right\|^2 \\
&\overset{\text{(c)}}{\leq} \frac{\eta_l K}{2}\|\nabla f(\tilde{w}^r))\|^2 + \frac{\eta_l \beta^2}{2N}\sum_{i,k}\mathbb{E}_r\|\tilde{w}^r_{i,k} - \tilde{w}^r\|^2 - \frac{\eta_l}{2KN^2}\mathbb{E}_r\left\|\sum_{i,k}\nabla f_i(\tilde{w}^r_{i,k})\right\|^2 \\
&\overset{\text{(d)}}{\leq} \frac{\eta_l K}{2}\|\nabla f(\tilde{w}^r))\|^2 + \frac{\eta_l L^2}{N}\sum_{i,k}\mathbb{E}_r\|w^r_{i,k} - w^r\|^2 + \frac{\eta_l L^2}{N}\sum_{i,k}\mathbb{E}_r\|\delta^r_{i,k} - \delta^r\|^2 - \frac{\eta_l}{2KN^2}\mathbb{E}_r\left\|\sum_{i,k}\nabla f_i(\tilde{w}^r_{i,k})\right\|^2 \\
&= \frac{\eta_l K}{2}\|\nabla f(\tilde{w}^r))\|^2 + K\eta_l L^2\mathcal{E}_w + K\eta_l L^2\mathcal{E}_\delta - \frac{\eta_l}{2KN^2}\mathbb{E}_r\left\|\sum_{i,k}\nabla f_i(\tilde{w}^r_{i,k})\right\|^2,
\end{aligned}
\tag{10}
$$

where (a) is from that $\langle a, b\rangle = \frac{1}{2}(\|a\|^2 + \|b\|^2 - \|a-b\|^2)$ with $a = \sqrt{\eta_l K}\nabla f(\tilde{w}^r)$ and $b = -\frac{\sqrt{\eta_l}}{N\sqrt{K}}\sum_{i,k}(\nabla f_i(\tilde{w}^r_{i,k}) - \nabla f_i(\tilde{w}^r))$; (b) is from Lemma A.2; (c) is from Assumption 1 and (d) is from Lemma A.2.

Lemma B.4 *For the full client participation scheme, we can bound $\mathbb{E}[\|\Delta^r\|^2]$ as follows:*

$$
\mathbb{E}_r[\|\Delta^r\|^2] \leq \frac{K\eta_l^2 L^2\rho^2}{N}\sigma_l^2 + \frac{\eta_l^2}{N^2}\left[\left\|\sum_{i,k}\nabla f_i(\tilde{w}^r_{i,k})\right\|^2\right].
$$

Proof. For the full client participation scheme, we have:

$$
\begin{aligned}
\mathbb{E}_r[\|\Delta^r\|^2] &\overset{\text{(a)}}{\leq} \frac{\eta_l^2}{N^2}\mathbb{E}_r\left[\left\|\sum_{i,k}\tilde{g}^r_{i,k}\right\|^2\right] \overset{\text{(b)}}{=} \frac{\eta_l^2}{N^2}\mathbb{E}_r\left[\left\|\sum_{i,k}(\tilde{g}^r_{i,k} - \nabla f_i(\tilde{w}^r_{i,k}))\right\|^2\right] + \frac{\eta_l^2}{N^2}\mathbb{E}_r\left[\left\|\sum_{i,k}\nabla f_i(\tilde{w}^r_{i,k})\right\|^2\right] \\
&\overset{\text{(c)}}{\leq} \frac{K\eta_l^2 L^2\rho^2}{N}\sigma_l^2 + \frac{\eta_l^2}{N^2}\left[\left\|\sum_{i,k}\nabla f_i(\tilde{w}^r_{i,k})\right\|^2\right],
\end{aligned}
$$

where (a) is from Lemma A.2; (b) is from Lemma A.3 and (c) is from Lemma A.4. □

Lemma B.5 *(Descent Lemma). For all $r \in R-1$ and $i \in \mathcal{S}^r$, with the choice of learning rate , the iterates generated by* `FedSAM` *in Algorithm 3 satisfy:*

$$
\begin{aligned}
\mathbb{E}_r[f(\tilde{w}^{r+1})] \leq f(\tilde{w}^r) &- K\eta_g\eta_l\left(\frac{1}{2} - 30K^2L^2\eta_l^2\right)\|\nabla f(\tilde{w}^r)\|^2 + K\eta_g\eta_l(10KL^4\eta_l^2\rho^2\sigma_l^2 + 90K^2L^2\eta_l^2\sigma_g^2 + 180K^2L^4\eta_l^2\rho^2 \\
&+ 120K^4L^6\eta_l^6\rho^2 + 16K^3\eta_l^4L^6\rho^2 + \frac{\eta_g\eta_l L^3\rho^2}{N}\sigma_l^2)
\end{aligned}
$$

where the expectation is w.r.t. the stochasticity of the algorithm.

Proof. We firstly propose the proof of full client participation scheme. Due to the smoothness in Assumption 1, taking

expectation of $f(\tilde{w}^{r+1})$ over the randomness at communication round r, we have:

$$\begin{aligned}
&\mathbb{E}_r[F(w^{r+1})] = \mathbb{E}_r[f(\tilde{w}^{r+1})] \leq f(\tilde{w}^r) + \mathbb{E}_r\langle \nabla f(\tilde{w}^r), \tilde{w}^{r+1} - \tilde{w}^r]\rangle + \frac{L}{2}\mathbb{E}_r[\|\tilde{w}^{r+1} - \tilde{w}^r\|^2] \\
&\overset{\text{(a)}}{=} f(\tilde{w}^r) + \mathbb{E}_r\langle \nabla f(\tilde{w}^r), -\Delta^r + K\eta_g\eta_l\nabla f(\tilde{w}^r) - K\eta_g\eta_l\nabla f(\tilde{w}^r)\rangle + \frac{L}{2}\eta_g^2\mathbb{E}_r[\|\Delta^r\|^2] \\
&\overset{\text{(b)}}{=} f(\tilde{w}^r) - K\eta_g\eta_l\|\nabla f(\tilde{w}^r)\|^2 + \eta_g\langle \nabla f(\tilde{w}^r), \mathbb{E}_r[-\Delta^r + K\eta_l\nabla f(\tilde{w}^r)]\rangle + \frac{L}{2}\eta_g^2\mathbb{E}_r[\|\Delta^r\|^2] \\
&\overset{\text{(c)}}{\leq} f(\tilde{w}^r) - \frac{K\eta_g\eta_l}{2}\|\nabla f(\tilde{w}^r)\|^2 + K\eta_g\eta_l L^2\mathcal{E}_w + K\eta_g\eta_l L^2\mathcal{E}_\delta - \frac{\eta_g\eta_l}{2KN}\mathbb{E}_r\left[\left\|\sum_{i,k}\nabla f_i(\tilde{w}_{i,k}^r)\right\|^2\right] + \frac{L}{2}\eta_g^2\mathbb{E}_r[\|\Delta^r\|^2] \\
&\overset{\text{(d)}}{\leq} f(\tilde{w}^r) - \frac{K\eta_g\eta_l}{2}\|\nabla f(\tilde{w}^r)\|^2 + K\eta_g\eta_l L^2\mathcal{E}_w + K\eta_g\eta_l L^2\mathcal{E}_\delta + \frac{K\eta_g^2\eta_l^2L^3\rho^2}{N}\sigma_l^2 \\
&\overset{\text{(e)}}{\leq} f(\tilde{w}^r) - K\eta_g\eta_l\left(\frac{1}{2} - 30K^2L^2\eta_l^2\right)\|\nabla f(\tilde{w}^r)\|^2 + K\eta_g\eta_l(10KL^4\eta_l^2\rho^2\sigma_l^2 + 90K^2L^2\eta_l^2\sigma_g^2 + 180K^2L^4\eta_l^2\rho^2 \\
&+ 120K^4L^6\eta_l^6\rho^2 + 16K^3\eta_l^4L^6\rho^2 + \frac{\eta_g\eta_l L^3\rho^2}{N}\sigma_l^2),
\end{aligned}$$

where (a) is from the iterate update given in Algorithm 3; (b) results from the unbiased estimators; (c) is from Lemma B.3; (d) is from Lemma B.4 and due to the fact that $\eta_g\eta_l \leq \frac{1}{KL}$ and (e) is from Lemmas B.1 and B.2. □

Theorem B.6 *Let constant local and global learning rates η_l and η_g be chosen as such that $\eta_l \leq \frac{1}{10KL}$, $\eta_g\eta_l \leq \frac{1}{KL}$. Under Assumption 1-2 and with full client participation, the sequence of outputs $\{w^r\}$ generated by `FedSAM` satisfies:*

$$\min_{r\in[R]}\mathbb{E}\|\nabla F(w^r)\|^2 \leq \frac{F^0 - F^*}{CK\eta_g\eta_l} + \Phi,$$

where $\Phi = \frac{1}{C}[10KL^4\eta_l^2\rho^2\sigma_l^2 + 90K^2L^2\eta_l^2\sigma_g^2 + 180K^2L^4\eta_l^2\rho^2 + 120K^4L^6\eta_l^6\rho^2 + 16K^3\eta_l^4L^6\rho^2 + \frac{\eta_g\eta_l L^3\rho^2}{N}\sigma_l^2]$. If we choose the learning rates $\eta_l = \frac{1}{\sqrt{R}KL}$, $\eta_g = \sqrt{KN}$ and perturbation amplitude ρ proportional to the learning rate, e.g., $\rho = \frac{1}{\sqrt{R}}$, we have

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[\|F(w^{r+1})\|] = \mathcal{O}\left(\frac{FL}{\sqrt{RKN}} + \frac{\sigma_g^2}{R} + \frac{L^2\sigma^2}{R^{3/2}\sqrt{KN}} + \frac{L^2}{R^{3/2}}\right).$$

Proof. For full client participation, summing the result of Lemma B.5 for $r = [R]$ and multiplying both sides by $\frac{1}{CK\eta_g\eta_l R}$ with $(\frac{1}{2} - 30K^2L^2\eta_l^2) > C > 0$ if $\eta_l < \frac{1}{\sqrt{30}KL}$, we have

$$\begin{aligned}
&\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[\|F(w^{r+1})\|^2] = \frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[\|f(\tilde{w}^{r+1})\|^2] \\
&\leq \frac{f(\tilde{w}^r) - f(\tilde{w}^{r+1})}{CK\eta_g\eta_l R} + \frac{1}{C}(10KL^4\eta_l^2\rho^2\sigma_l^2 + 90K^2L^2\eta_l^2\sigma_g^2 + 180K^2L^4\eta_l^2\rho^2 + 120K^4L^6\eta_l^6\rho^2 + 16K^3\eta_l^4L^6\rho^2 + \frac{\eta_g\eta_l L^3\rho^2}{N}\sigma_l^2) \\
&\leq \frac{f(\tilde{w}^0) - f^*}{CK\eta_g\eta_l R} + \frac{1}{C}(10KL^4\eta_l^2\rho^2\sigma_l^2 + 90K^2L^2\eta_l^2\sigma_g^2 + 180K^2L^4\eta_l^2\rho^2 + 120K^4L^6\eta_l^6\rho^2 + 16K^3\eta_l^4L^6\rho^2 + \frac{\eta_g\eta_l L^3\rho^2}{N}\sigma_l^2),
\end{aligned}$$

where the second inequality uses $f(\tilde{w}^{r+1}) \geq f^*$ and $f(\tilde{w}^0) \geq f(\tilde{w}^r)$. If we choose the learning rates $\eta_l = \frac{1}{\sqrt{R}KL}$, $\eta_g = \sqrt{KN}$ and perturbation amplitude ρ proportional to the learning rate, e.g., $\rho = \frac{1}{\sqrt{R}}$, we have

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[\|F(w^{r+1})\|] = \mathcal{O}\left(\frac{FL}{\sqrt{RKN}} + \frac{\sigma_g^2}{R} + \frac{L^2\sigma_l^2}{R^2K} + \frac{L^2\sigma_l^2}{R^{3/2}\sqrt{KN}} + \frac{L^2}{R^2K} + \frac{L^2}{R^{3/2}} + \frac{L^2}{R^3K}\right).$$

Note that the term $\frac{G^2}{R}$ is due to the heterogeneity between each client, $(\frac{L^2}{R^2K} + \frac{L^2}{R^{3/2}\sqrt{KN}})\sigma^2$ is due to the local SGD and $\frac{1}{R^{3/2}} + \frac{1}{R^3K}$ is due to the local SAM. We can see that $\frac{L^2}{R^{3/2}} + \frac{L^2}{R^3K}$ only obtains higher order, and hence SAM part does not

take large influence of convergence. After omitting the higher order, we have

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[\|\|F(w^{r+1})\|\|] = \mathcal{O}\bigg(\frac{FL}{\sqrt{RKN}} + \frac{\sigma_g^2}{R} + \frac{L^2\sigma^2}{R^{3/2}\sqrt{KN}} + \frac{L^2}{R^{3/2}}\bigg).$$

This completes the proof. □

B.3. Convergence Analysis of Partial Client Participation `FedSAM`

Lemma B.7 *For the partial client participation, we can bound* $\mathbb{E}_r[\|\Delta^r\|^2]$*:*

$$\mathbb{E}_r[\|\Delta^r\|^2] \leq \frac{K\eta_l^2L^2\rho^2}{S}\sigma_l^2 + \frac{S}{N}\sum_i\bigg\|\sum_{j=1}^{K-1}\nabla f_i(\tilde{w}_{i,k}^r)\bigg\|^2 + \frac{S(S-1)}{N^2}\bigg\|\sum_{j=0}^{K-1}\nabla f_i(\tilde{w}_{i,j}^r)\bigg\|^2.$$

For the partial client participation scheme w/o replacement, we have:

$$\begin{aligned}
\mathbb{E}_r[\|\Delta^r\|^2] &\overset{(a)}{\leq} \frac{\eta_l^2}{S^2}\mathbb{E}_r\bigg[\bigg\|\sum_{i\in\mathcal{S}^r}\sum_k\tilde{g}_{i,k}\bigg\|^2\bigg] = \frac{\eta_l^2}{S^2}\mathbb{E}_r\bigg[\bigg\|\sum_i\mathbb{I}\{i\in\mathcal{S}^r\}\sum_k\tilde{g}_{i,k}\bigg\|^2\bigg] \\
&\overset{(b)}{=} \frac{\eta_l^2}{SN}\mathbb{E}_r\bigg[\bigg\|\sum_i\sum_{j=0}^{K-1}(\tilde{g}_{i,j}^r - \nabla f_i(\tilde{w}_{i,j}^r))\bigg\|^2\bigg] + \frac{\eta_l^2}{S^2}\mathbb{E}_r\bigg[\bigg\|\sum_i\mathbb{I}\{i\in\mathcal{S}^r\}\sum_{j=0}^{K-1}\nabla f_i(\tilde{w}_{i,j}^r)\bigg\|^2\bigg] \\
&\overset{(c)}{\leq} \frac{K\eta_l^2L^2\rho^2}{S}\sigma_l^2 + \frac{\eta_l^2}{S^2}\mathbb{E}_r\bigg[\bigg\|\sum_{i=1}^{S}\sum_{j=0}^{K-1}\nabla f_i(\tilde{w}_{i,j}^r)\bigg\|^2\bigg] \\
&= \frac{K\eta_l^2L^2\rho^2}{S}\sigma_l^2 + \frac{\eta_l^2}{NS}\sum_i\bigg\|\sum_{j=1}^{K-1}\nabla f_i(\tilde{w}_{i,k}^r)\bigg\|^2 + \frac{(S-1)\eta_l^2}{SN^2}\bigg\|\sum_{j=0}^{K-1}\nabla f_i(\tilde{w}_{i,j}^r)\bigg\|^2,
\end{aligned}$$

where (a) is from Lemma A.2; (b) is from Lemma A.3 and (c) is from Lemma A.4. □

Lemma B.8 *For* $\mathbb{E}[\|\sum_k\nabla f_i(\tilde{w}_{i,k})\|^2]$*, where* $\nabla f_i(\tilde{w}_{i,k})^2$ *for all* $k\in[K]$ *and* $i\in[N]$ *is chosen according to* `FedSAM`*, we have:*

$$\begin{aligned}
\sum_i\mathbb{E}\bigg[\bigg\|\sum_k\nabla f_i(\tilde{w}_{i,k})\bigg\|^2\bigg] &\leq 30NK^2L^2\eta_l^2(2L^2\rho^2\sigma_l^2 + 6K(3\sigma_g^2 + 6L^2\rho^2) + 6K\|\nabla f(\tilde{w})\|^2) + 144K^4L^6\eta_l^4\rho^2 \\
&\quad + 12NK^4L^2\eta_l^2\rho^2 + 3NK^2(3\sigma_g^2 + 6L^2\rho^2) + 3NK^2\|\nabla f(\tilde{w})\|^2,
\end{aligned}$$

where the expectation is w.r.t the stochasticity of the algorithm.

Proof.

$$\begin{aligned}
&\sum_i\mathbb{E}\bigg[\bigg\|\sum_k\nabla f_i(\tilde{w}_{i,k})\bigg\|^2\bigg] = \sum_i\mathbb{E}\bigg[\bigg\|\sum_k\nabla f_i(\tilde{w}_{i,k}) - \nabla f_i(\tilde{w}) + \nabla f_i(\tilde{w}) - \nabla f(\tilde{w}) + \nabla f(\tilde{w})\bigg\|^2\bigg] \\
&\overset{(a)}{\leq} 6KL^2\sum_{i,k}\mathbb{E}[\|w_{i,k} - w\|^2] + 6KL^2\sum_{i,k}\mathbb{E}[\|\delta_{i,k} - \delta\|^2] + 3NK^2(3\sigma_g^2 + 6L^2\rho^2) + 3NK^2\|\nabla f(\tilde{w})\|^2 \\
&\overset{(b)}{\leq} 30NK^2L^2\eta_l^2(2L^2\rho^2\sigma_l^2 + 6K(3\sigma_g^2 + 6L^2\rho^2) + 6K\|\nabla f(\tilde{w})\|^2) + 144K^4L^6\eta_l^4\rho^2 \\
&+ 12NK^4L^2\eta_l^2\rho^2 + 3NK^2(3\sigma_g^2 + 6L^2\rho^2) + 3NK^2\|\nabla f(\tilde{w})\|^2.
\end{aligned}$$

where (a) is from Assumption 1, Lemmas A.2 and A.5; (b) is from Lemmas B.1 and B.2. □

Theorem B.9 *Let constant local and global learning rates* η_l *and* η_g *be chosen as such that* $\eta_l \leq \frac{1}{10KL}$*,* $\eta_g\eta_l \leq \frac{1}{KL}$ *and the condition* $(\frac{1}{2} - 30K^2L^2\eta_l^2 - \frac{L\eta_g\eta_l}{2S}(3K + 180K^3L^2\eta_l^2)) > 0$ *holds. Under Assumption 1-3 and with partial client*

participation, the sequence of outputs $\{w^r\}$ *generated by* `FedSAM` *satisfies:*

$$\min_{r\in[R]} \mathbb{E}\|\nabla F(w^r)\|^2 \leq \frac{F^0 - F^*}{CK\eta_g\eta_l} + \Phi,$$

where $\Phi = \frac{1}{C}[10KL^4\eta_l^2\rho^2\sigma_l^2 + 90K^2L^2\eta_l^2\sigma_g^2 + 180K^2L^4\eta_l^2\rho^2 + 120K^4L^6\eta_l^6\rho^2 + 16K^3\eta_l^4L^6\rho^2 + \frac{L^3\eta_g\eta_l\rho^2}{2S}\sigma^2 + \frac{\eta_g\eta_l}{S}(30KL^5\eta_l^2\rho^2\sigma_l^2 + 180K^2L^3\eta_l^2\sigma_g^2 + 360KL^5\eta_l^2\rho^2 + 72K^3L^7\eta_l^4\rho^2 + 6K^3L^3\eta_l^2\rho^2 + 6KL\sigma_g^2 + 6KL^3\rho^2)]$. *If we choose the learning rates* $\eta_l = \frac{1}{\sqrt{RKL}}$, $\eta_g = \sqrt{KS}$ *and perturbation amplitude* ρ *proportional to the learning rate, e.g.,* $\rho = \frac{1}{\sqrt{R}}$, *we have:*

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[\|F(w^{r+1})\|] = \mathcal{O}\Big(\frac{FL}{\sqrt{RKS}} + \frac{\sqrt{K}G^2}{\sqrt{RS}} + \frac{L^2\sigma^2}{R^{3/2}K} + \frac{\sqrt{K}L^2}{R^{3/2}\sqrt{S}}\Big).$$

$$\begin{aligned}
&\mathbb{E}[\|f(\tilde{w}^{r+1})\|]\\
&\overset{(a)}{\leq} f(\tilde{w}^r) - \frac{K\eta_g\eta_l}{2}\|\nabla f(\tilde{w}^r)\|^2 + K\eta_g\eta_l L^2\mathcal{E}_w + K\eta_g\eta_l L^2\mathcal{E}_\delta - \frac{\eta_g\eta_l}{2KN}\mathbb{E}_r\Big[\Big\|\sum_{i,k}\nabla f_i(\tilde{w}_{i,k}^r)\Big\|^2\Big] + \frac{L}{2}\eta_g^2\mathbb{E}_r[\|\Delta^r\|^2]\\
&\overset{(b)}{\leq} f(\tilde{w}^r) - \frac{K\eta_g\eta_l}{2}\|\nabla f(\tilde{w}^r)\|^2 + K\eta_g\eta_l L^2\mathcal{E}_w + K\eta_g\eta_l L^2\mathcal{E}_\delta + \frac{K\eta_g^2\eta_l^2L^3\rho^2}{2S}\sigma_l^2\\
&- \frac{\eta_g\eta_l}{2KN}\mathbb{E}_r\Big[\Big\|\sum_{i,k}\nabla f_i(\tilde{w}_{i,k}^r)\Big\|^2\Big] + \frac{\eta_g^2LS}{2N}\sum_i\Big\|\sum_{j=1}^{K-1}\nabla f_i(\tilde{w}_{i,k}^r)\Big\|^2 + \frac{\eta_g^2LS(S-1)}{2N^2}\Big\|\sum_{j=0}^{K-1}\nabla f_i(\tilde{w}_{i,j}^r)\Big\|^2\\
&\overset{(c)}{\leq} f(\tilde{w}^r) - \frac{K\eta_g\eta_l}{2}\|\nabla f(\tilde{w}^r)\|^2 + K\eta_g\eta_l L^2\mathcal{E}_w + K\eta_g\eta_l L^2\mathcal{E}_\delta + \frac{K\eta_g^2\eta_l^2L^3\rho^2}{2S}\sigma_l^2 + \frac{L\eta_g^2\eta_l^2}{2NS}\sum_i\|\sum_k\nabla f_i(\tilde{w}_{i,k}^r)\|^2\\
&\overset{(d)}{\leq} f(\tilde{w}^r) - K\eta_g\eta_l\Big(\frac{1}{2} - 30K^2L^2\eta_l^2 - \frac{L\eta_g\eta_l}{2S}(3K + 180K^3L^2\eta_l^2)\Big)\|\nabla f(\tilde{w}^r)\|^2\\
&+ K\eta_g\eta_l\Big(10KL^4\eta_l^2\rho^2\sigma_l^2 + 90K^2L^2\eta_l^2\sigma_g^2 + 180K^2L^4\eta_l^2\rho^2 + 120K^4L^6\eta_l^6\rho^2 + 16K^3\eta_l^4L^6\rho^2 + \frac{L^3\eta_g\eta_l\rho^2}{2S}\sigma^2\Big)\\
&+ \frac{K\eta_g^2\eta_l^2}{S}(30KL^5\eta_l^2\rho^2\sigma_l^2 + 180K^2L^3\eta_l^2\sigma_g^2 + 360KL^5\eta_l^2\rho^2 + 72K^3L^7\eta_l^4\rho^2 + 6K^3L^3\eta_l^2\rho^2 + 6KL\sigma_g^2 + 6KL^3\rho^2)\\
&\overset{(e)}{\leq} f(\tilde{w}^r) - CK\eta_g\eta_l\|\nabla f(\tilde{w}^r)\|^2\\
&+ K\eta_g\eta_l\Big(10KL^4\eta_l^2\rho^2\sigma_l^2 + 90K^2L^2\eta_l^2\sigma_g^2 + 180K^2L^4\eta_l^2\rho^2 + 120K^4L^6\eta_l^6\rho^2 + 16K^3\eta_l^4L^6\rho^2 + \frac{L^3\eta_g\eta_l\rho^2}{2S}\sigma^2\Big)\\
&+ \frac{K\eta_g^2\eta_l^2}{S}(30KL^5\eta_l^2\rho^2\sigma_l^2 + 180K^2L^3\eta_l^2\sigma_g^2 + 360KL^5\eta_l^2\rho^2 + 72K^3L^7\eta_l^4\rho^2 + 6K^3L^3\eta_l^2\rho^2 + 6KL\sigma_g^2 + 6KL^3\rho^2),
\end{aligned}$$

where (a) is from Lemma B.5; (b) is from B.4; (c) is based on taking the expectation of r-th round and if the learning rates satisfy that $KL\eta_g\eta_l \leq \frac{S-1}{S}$; (d) is from Lemmas B.1, B.2 and B.8 and (e) holds because there exists a constant $C > 0$ satisfying $(\frac{1}{2} - 30K^2L^2\eta_l^2 - \frac{L\eta_g\eta_l}{2S}(3K + 180K^3L^2\eta_l^2)) > C > 0$.

Summing the above result for $r = [R]$ and multiplying both sides by $\frac{1}{CK\eta_g\eta_l R}$, we have

$$\begin{aligned}
&\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[\|F(w^{r+1})\|] \leq \frac{f(\tilde{w}^r) - f(\tilde{w}^{r+1})}{CK\eta_g\eta_l R} \\
&+ \frac{1}{C}\bigg(10KL^4\eta_l^2\rho^2\sigma_l^2 + 90K^2L^2\eta_l^2\sigma_g^2 + 180K^2L^4\eta_l^2\rho^2 + 120K^4L^6\eta_l^6\rho^2 + 16K^3\eta_l^4L^6\rho^2 + \frac{L^3\eta_g\eta_l\rho^2}{2S}\sigma^2 \\
&+ \frac{\eta_g\eta_l}{S}(30KL^5\eta_l^2\rho^2\sigma_l^2 + 180K^2L^3\eta_l^2\sigma_g^2 + 360KL^5\eta_l^2\rho^2 + 72K^3L^7\eta_l^4\rho^2 + 6K^3L^3\eta_l^2\rho^2 + 6KL\sigma_g^2 + 6KL^3\rho^2)\bigg) \\
&\leq \frac{F}{CK\eta_g\eta_l R} \\
&+ \frac{1}{C}\bigg(10KL^4\eta_l^2\rho^2\sigma_l^2 + 90K^2L^2\eta_l^2\sigma_g^2 + 180K^2L^4\eta_l^2\rho^2 + 120K^4L^6\eta_l^6\rho^2 + 16K^3\eta_l^4L^6\rho^2 + \frac{L^3\eta_g\eta_l\rho^2}{2S}\sigma^2 \\
&+ \frac{\eta_g\eta_l}{S}(30KL^5\eta_l^2\rho^2\sigma_l^2 + 180K^2L^3\eta_l^2\sigma_g^2 + 360KL^5\eta_l^2\rho^2 + 72K^3L^7\eta_l^4\rho^2 + 6K^3L^3\eta_l^2\rho^2 + 6KL\sigma_g^2 + 6KL^3\rho^2)\bigg),
\end{aligned}$$

where the second inequality uses $F = f(\tilde{w}^0) - f^* \leq f(\tilde{w}^r) - f(\tilde{w}^{r+1})$. If we choose the learning rates $\eta_l = \frac{1}{\sqrt{RKL}}$, $\eta_g = \sqrt{KS}$ and perturbation amplitude ρ proportional to the learning rate, e.g., $\rho = \frac{1}{\sqrt{R}}$, we have:

$$\begin{aligned}
\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[\|F(w^{r+1})\|] = \mathcal{O}\bigg(&\frac{FL}{\sqrt{RKS}} + \frac{\sigma_g^2}{R} + \frac{\sqrt{K}\sigma_g^2}{\sqrt{RS}} + \frac{\sqrt{KS}\sigma_g^2}{R^{3/2}} + \frac{L^2\sigma_l^2}{R^{3/2}K} + \frac{L^2\sigma_l^2}{R^{3/2}\sqrt{KS}} \\
&+ \frac{L^2\sigma_l^2}{R^{5/2}\sqrt{KS}} + \frac{L^2}{R^2} + \frac{1}{R^4K^2} + \frac{L^2}{R^3K} + \frac{\sqrt{KS}}{R^{5/2}SK^2} + \frac{\sqrt{KS}}{R^{7/2}SK^2} + \frac{\sqrt{K}}{R^{5/2}\sqrt{S}} + \frac{\sqrt{K}L^2}{R^{3/2}\sqrt{S}}\bigg),
\end{aligned}$$

If the number of sampling clients are larger than the number of epochs, i.e., $S \geq K$, and omitting the larger order of each part, we have:

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[\|F(w^{r+1})\|] = \mathcal{O}\bigg(\frac{FL}{\sqrt{RKS}} + \frac{\sqrt{K}G^2}{\sqrt{RS}} + \frac{L^2\sigma^2}{R^{3/2}K} + \frac{\sqrt{K}L^2}{R^{3/2}\sqrt{S}}\bigg).$$

This completes the proof. □

C. Generalization Bounds

The generalization bound of `FedSAM` follows the margin-based generalization bounds in (Neyshabur et al., 2018; Bartlett et al., 2017; Farnia et al., 2018). We consider the margin-based error for analyzing the generalization error in `FedSAM` with general neural network as follows:

$$L_\gamma^{\text{SAM}}(F(w)) := \frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_i\bigg(f_i(w+\delta_i, \mathbf{X})[Y] - \max_{j\neq Y} F_i(w+\delta_i, \mathbf{X})[j] \leq \gamma\bigg). \tag{11}$$

Our generalization bound is based on the two following Lemmas in (Chatterji et al., 2019) and (Neyshabur et al., 2018):

Lemma C.1 *((Chatterji et al., 2019)). Let $F(w)$ be any predictor function with parameters w and $\mathcal{P}$ be a prior distribution on parameters w. Then, for any $\gamma, m, \zeta > 0$, with probability $1-\zeta$ over training set $\mathcal{M}$ of size m, for any parameter w and any perturbation distribution $\mathcal{Q}$ over parameters such that $\mathbb{P}_{\delta\sim\mathcal{Q}}[\max_{\mathbf{X}} |F(w+\delta) - F(w)| \leq \frac{\gamma}{4}] \geq \frac{1}{2}$, we have:*

$$L^{SAM}(F(w)) \leq \hat{L}_\gamma^{SAM}(F(w)) + \sqrt{\frac{2KL(w+\delta\|\mathcal{P}) + \log\frac{m}{\zeta}}{2(m-1)}}.$$

where $KL(\cdot\|P)$ is the KL-divergence.

Lemma C.2 *((Neyshabur et al., 2018)). Let norm of input $\mathbf{X}$ be bounded by A. For any $A > 0$, let $F(w)$ be a neural network with ReLU activations and depth d with h units per hidden-layer. Then for any $w, \mathbf{X} \in \mathcal{X}$, and any perturbation $\boldsymbol{\delta}$ s.t. $\|\delta_j\| \leq \|W_j\|$, where δ_j is the size of layer j, the change in the output of the network can be bounded as follows:*

$$\|F(w+\delta, \mathbf{X}) - F(w, \mathbf{X})\|_2 \leq eA \prod_{j=1}^{d} \|W_j\| \sum_{j=1}^{d} \frac{\|\delta_j\|_2}{\|W_j\|_2}.$$

Lemma C.1 gives a data-independent deterministic bound which depends on the maximum change of the output function over the domain after a perturbation. Lemma C.2 bounds the change in the output a network based on the magnitude of the perturbation.

Theorem C.3 *Let input $\mathbf{X}$ be an $n \times n$ image whose norm is bounded by A, $f(w)$ be the classification function with d hidden-layer neural network with h units per hidden-layer, and satisfy 1-Lipschitz activation $\theta(0) = 0$. We assume the constant $M \geq 1$ for each layer W_j satisfies:*

$$\frac{1}{M} \leq \frac{\|W_j\|}{\phi_w} \leq M,$$

where $\phi_w := (\prod_{j=1}^{d} \|W_j\|)^{1/d}$ denotes the geometric mean of $f(w)$'s spectral norms across all layers. Then, for any margin value γ, size of local training dataset on each client m, $\zeta > 0$, with probability $1-\zeta$ over the training set, any parameter of SAM local optimizer $\tilde{w} = w + \delta$ such that $\max_{\mathbf{X} \in \mathcal{D}_i} \|F_i(w) - f(\tilde{w})\| \leq \frac{\gamma}{8}$, we can obtain the following generalization bound:

$$\mathcal{L}^{SAM}(F(w)) \leq \hat{\mathcal{L}}_{\gamma}^{SAM}(F(w+\delta)) + \mathcal{O}\left(\frac{32Bd^2h\log(dh)Q(F(w)) + d\log\frac{Nmd\log(M)}{\zeta}}{\gamma^2 m}\right),$$

where $Q(F(w)) := \prod_{j=1}^{d} \|W_j\| \sum_{i=1}^{d} \frac{\|W_j\|_F^2}{\|W_j\|}$ and $\|W_j\|_F^2$ is the Frobenius norm.

Proof. Based on Lemma C.1, we choose the perturbation δ_j of each layer which is a zero-mean multivariate Gaussian distribution with diagonal covariance matrix, i.e., $\mathcal{N}(0, \lambda_j^2 I)$ and $\lambda_j = \frac{\|\tilde{W}_j\|}{\epsilon_{\tilde{W}}}\lambda$, where $\epsilon_{\tilde{W}} := (\prod_{j=1}^{d} \|W_j\|)^{1/d}$ is the geometric average of spectral norms across all layers. We consider $F(\tilde{W})$ with weights $\tilde{W}$. Since $(1+\frac{1}{d})^d \leq e$ and $\frac{1}{e} \leq (1-\frac{1}{d})^{d-1}$, for any weight vector of W_j such that $|\|W_j\|_2 - \|\tilde{W}_j\|_2| \leq \frac{1}{d}$ for every j, we have:

$$(1/e)^{\frac{d}{d-1}} \prod_{j=1}^{d} \|\tilde{W}_j\| \leq \prod_{j=1}^{d} \|W_j\| \leq e \prod_{j=1}^{d} \|\tilde{W}_i\|.$$

Then, for the jth layer's random perturbation vector $\delta_j \sim \mathcal{N}(0, \lambda_j^2 I)$, we have the following bound from (Tropp, 2012) with h representing the width of the jth hidden layer:

$$\mathbb{P}\left(\epsilon_{\tilde{W}} \frac{\|\delta_j\|}{\|\tilde{W}_j\|} > t\right) \leq 2he^{-\frac{t^2}{2h\lambda^2}}.$$

Based on (Farnia et al., 2018), we now use a union bound over all layers for a maximum union probability of $1/2$, which implies the normalized $\epsilon_{\tilde{W}} \frac{\|\delta_j\|}{\|\tilde{W}_j\|}$ for each layer can be upper-bounded by $\lambda\sqrt{2h\log(4hd)}$. Then, for any W satisfying $|\|W_j\| - \|\tilde{W}_j\|| \leq \frac{1}{d}\|\tilde{W}_j\|$ for all layer j's, we obtain the following:

$$\|F(W+\delta, \mathbf{X}) - F(W, \mathbf{X})\| \leq eA\left(\prod_{j=1}^{d} \|w_j\|\right) \sum_{j=1}^{d} \frac{\|\delta_j\|}{\|W_j\|_2} \leq 4edA\epsilon_{\tilde{W}}^{d-1}\lambda\sqrt{2h\log(4hd)} \leq \frac{\gamma}{8}.$$

where the last inequality is from choosing $\lambda = \frac{\gamma}{32edA\epsilon_{\tilde{W}}^{d-1}\sqrt{h\log(4hd)}}$, where the perturbation satisfies the Lemma C.2. Then,

we can bound the KL-divergence in Lemma C.1 as follows:

$$\mathrm{KL}(w+\delta||\mathcal{P}) \leq \sum_{j=1}^{d} \frac{\|W_j\|_F}{2\lambda_j^2} = \frac{32^2 e^2 d^2 A^2 \epsilon_{\tilde{W}}^{2d} h \log(4hd)}{\gamma^2} \sum_{j=1}^{d} \frac{\|W_j\|_F}{\|\tilde{W}_j\|}$$
$$\leq \frac{32^2 e^2 d^2 A^2 \prod_{j=1}^{d} \|W_j\| h \log(4hd)}{\gamma^2} \sum_{j=1}^{d} \frac{\|W_j\|_F}{\|\tilde{W}_j\|} = \mathcal{O}\bigg(\frac{d^2 A^2 h \log(hd) \prod_{j=1}^{d} \|W_j\|}{\gamma^2} \sum_{j=1}^{d} \frac{\|W_j\|_F}{\|W_j\|} \bigg).$$

Based on (Farnia et al., 2018), we have the following result given a fixed underlying distribution $\mathcal{P}$ and any $\zeta > 0$ with probability $1-\zeta$ for any W:

$$\mathcal{L}^{\mathrm{SAM}}(F(w)) \leq \hat{\mathcal{L}}_{\gamma}^{\mathrm{SAM}}(F(w+\delta)) + \mathcal{O}\bigg(\frac{d^2 A^2 h \log(hd) \prod_{j=1}^{d} \|W_j\| \sum_{j=1}^{d} \frac{\|W_j\|_F}{\|W_j\|} + \log \frac{m}{\zeta}}{m\gamma^2} \bigg).$$

Now, we use a cover of size $\mathcal{O}(d \log(M)^d d)$ points, and hence it can demonstrate that for a fixed underlying distribution for any $\zeta > 0$, with probability $1-\zeta$, we have:

$$\mathcal{L}^{\mathrm{SAM}}(F(w)) \leq \hat{\mathcal{L}}_{\gamma}^{\mathrm{SAM}}(F(w+\delta)) + \mathcal{O}\bigg(\frac{d^2 A^2 h \log(hd) \prod_{j=1}^{d} \|W_j\| \sum_{j=1}^{d} \frac{\|W_j\|_F}{\|W_j\|} + d \log \frac{dm \log(M)}{\zeta}}{m\gamma^2} \bigg).$$

To apply the above result to the FL network of N clients, we apply a union bound to have the bound hold simultaneously for the distribution of every client, which proves for every $\zeta > 0$ with probability at least $1-\zeta$, the average SAM loss of the clients satisfies the following margin-based bound:

$$\mathcal{L}^{\mathrm{SAM}}(F(w)) \leq \hat{\mathcal{L}}_{\gamma}^{\mathrm{SAM}}(F(w+\delta)) + \mathcal{O}\bigg(\frac{d^2 A^2 h \log(hd) \prod_{j=1}^{d} \|W_j\| \sum_{j=1}^{d} \frac{\|W_j\|_F}{\|W_j\|} + d \log \frac{dNm \log(M)}{\zeta}}{m\gamma^2} \bigg).$$

This completes the proof. □

D. Convergence Analysis of `MoFedSAM`

D.1. Description of `FedSAM` Algorithm and Key Lemmas

We outline the `MoFedSAM` algorithm in Algorithm 2. In round r, we sample $\mathcal{S}^r \subseteq [N]$ clients with $|\mathcal{S}^r| = S$ and then perform the following updates:

- Starting from the shared global parameters $w_{i,0}^r = w^{r-1}$, we update the local parameters for $k \in [K]$:

$$\tilde{w}_{i,k}^r = w_{i,k-1}^r + \rho \frac{g_{i,k-1}^r}{\|g_{i,k-1}^r\|}$$
$$v_{i,k-1}^r = \beta \tilde{g}_{i,k-1}^r + (1-\beta)\Delta^r$$
$$w_{i,k}^r = w_{i,k-1}^r - \eta_l v_{i,k-1}^r,$$

- After K times local epochs, we obtain the following:

$$\Delta_i^r = w_{i,K}^r - w^r.$$

- Compute the new global parameters using only updates from the clients $i \in \mathcal{S}^r$ and a global step-size η_g:

$$\Delta^{r+1} = \frac{1}{\eta_l K S} \sum_{i \in \mathcal{S}^r} \Delta_i^r$$
$$w^{r+1} = w^r + \eta_g \Delta^r.$$

To prove the convergence of MoFedSAM, we first propose some lemmas for MoFedSAM as follows:

Lemma D.1 *(Bounded $\mathcal{E}_w$ of MoFedSAM). Suppose our functions satisfies Assumptions 1-2. Then, for any $i \in [N]$, $k \in [K]$ and $r \in [R]$ the updates of MoFedSAM for any learning rate satisfying $\eta_l \leq \frac{1}{\sqrt{30}\beta KL}$ have the drift due to $w_{i,k} - w$:*

$$\begin{aligned}\mathcal{E}_w &= \frac{1}{N}\sum_i \mathbb{E}[\|w_{i,k} - w\|^2] \\ &\leq 5K\eta_l^2(2\beta^2L^2\eta_l^2\rho^2\sigma_l^2 + 7K\beta^2\eta_l^2(3\sigma_g^2 + 6L^2\rho^2) + 14K(1-\beta)^2\eta_l^2\|\nabla f(\tilde{w})\|^2) + 28\beta^2K^3L^4\eta_l^4\rho^2.\end{aligned}$$

Proof. Recall that the local update on client i is $w_{i,k} = w_{i,k-1} - \beta\eta_l g_{i,k-1} + (1-\beta)\Delta^r$. Then, we have:

$$\begin{aligned}
&\mathbb{E}\|w_{i,k} - w\|^2 = \mathbb{E}\|w_{i,k-1} - w - \eta_l(\beta\tilde{g}_{i,k-1} + (1-\beta)\Delta)\|^2 \\
&\overset{(a)}{\leq} \mathbb{E}\|w_{i,k-1} - w - \beta\eta_l(\tilde{g}_{i,k-1} - \nabla f_i(\tilde{w}_{i,k-1}) + \nabla f_i(\tilde{w}_{i,k-1}) - \nabla f_i(\tilde{w}) + \nabla f_i(\tilde{w}) - \nabla f(\tilde{w}) + \nabla f(\tilde{w})) + \eta_l(1-\beta)\Delta\|^2 \\
&\overset{(b)}{\leq} \left(1 + \frac{1}{2K-1} + 2\beta^2L^2\eta_l^2\right)\mathbb{E}\|w_{i,k-1} - w\|^2 + 2\beta^2L^2\eta_l^2\rho^2\sigma_l^2 + 7K^2\beta\eta_l^2\mathbb{E}\|\nabla f_i(\tilde{w}_{i,k-1}) - \nabla f_i(\tilde{w})\|^2 \\
&+ 7K\beta^2\eta_l^2(3\sigma_g^2 + 6L^2\rho^2) + 7K\beta^2\eta_l^2\|\nabla f(\tilde{w})\|^2 + 7K\eta_l^2(1-\beta)^2\|\Delta\|^2 \\
&\overset{(c)}{\leq} \left(1 + \frac{1}{2K-1} + 2\beta^2L^2\eta_l^2 + 14K\beta^2L^2\eta_l^2\right)\mathbb{E}\|w_{i,k-1} - w\|^2 + 2\beta^2L^2\eta_l^2\rho^2\sigma_l^2 + 7K(1-\beta)^2\eta_l^2\|\Delta\|^2 \\
&+ 14K\beta^2L^2\eta_l^2\mathbb{E}\|\delta_{i,k-1} - \delta\|^2 + 7K\beta^2\eta_l^2(3\sigma_g^2 + 6L^2\rho^2) + 7\beta^2K\mathbb{E}\|\nabla f(\tilde{w})\|^2 \\
&\overset{(d)}{\leq} \left(1 + \frac{1}{2K-1} + 2\beta^2L^2\eta_l^2 + 14\beta^2KL^2\eta_l^2\right)\mathbb{E}\|w_{i,k-1} - w\|^2 + 2\beta^2L^2\eta_l^2\rho^2\sigma_l^2 + 14K\beta^2L^2\eta_l^2\mathbb{E}\|\delta_{i,k} - \delta\|^2 \\
&+ 7K\beta^2\eta_l^2(3\sigma_g^2 + 6L^2\rho^2) + 7K(1-\beta)^2\eta_l^2\|\Delta\|^2 + 7\beta^2K\mathbb{E}\|\nabla f(\tilde{w})\|^2 \\
&\overset{(e)}{\leq} \left(1 + \frac{1}{2K-1} + 2\beta^2L^2\eta_l^2 + 14\beta^2KL^2\eta_l^2\right)\mathbb{E}\|w_{i,k-1} - w\|^2 + 2\beta^2L^2\eta_l^2\rho^2\sigma_l^2 + 14K\beta^2L^2\eta_l^2\mathbb{E}\|\delta_{i,k} - \delta\|^2 \\
&+ 7K\beta^2\eta_l^2(3\sigma_g^2 + 6L^2\rho^2) + 14K(1-\beta)^2\eta_l^2\|\nabla f(\tilde{w})\|^2,
\end{aligned}$$

where (a) follows from the fact that $\tilde{g}_{i,k-1}$ is an unbiased estimator of $\nabla f_i(\tilde{w}_{i,k-1})$ and Lemma A.3; (b) is from Lemmas A.2 and A.5; (c) is from Assumption 3; Lemma A.2; (d) is from Assumption 2 and (e) is due to the fact that $\Delta \approx \nabla f(\tilde{w})$ and $\beta < \frac{1}{2}$.

Averaging over the clients i and learning rate satisfies $\eta_l \leq \frac{1}{\sqrt{30}\beta KL}$, we have:

$$\begin{aligned}
\mathcal{E}_w &\leq \left(1 + \frac{1}{2K-1} + 2\beta^2L^2\eta_l^2 + 14\beta^2KL^2\eta_l^2\right)\mathbb{E}\|w_{i,k-1} - w\|^2 + 2\beta^2L^2\eta_l^2\rho^2\sigma_l^2 + 14K\beta^2L^2\eta_l^2\mathbb{E}\|\delta_{i,k} - \delta\|^2 \\
&+ 7K\beta^2\eta_l^2(3\sigma_g^2 + 6L^2\rho^2) + 14K(1-\beta)^2\eta_l^2\|\nabla f(\tilde{w})\|^2 \\
&\overset{(a)}{\leq} \left(1 + \frac{1}{K-1}\right)\frac{1}{N}\sum_{i\in[N]}\mathbb{E}\|w_{i,k-1} - w\|^2 + 2\beta^2L^2\eta_l^2\rho^2\sigma_l^2 \\
&+ 14K\beta^2L^2\eta_l^2\frac{1}{N}\sum_{i\in[N]}\mathbb{E}\|\delta_{i,k} - \delta\|^2 + 7K\beta^2\eta_l^2(3\sigma_g^2 + 6L^2\rho^2) + 14K(1-\beta)^2\eta_l^2\|\nabla f(\tilde{w})\|^2 \\
&\leq \sum_{\tau=0}^{k-1}\left(1 + \frac{1}{K-1}\right)^\tau[2\beta^2L^2\eta_l^2\rho^2\sigma_l^2 + 7K\beta^2\eta_l^2(3\sigma_g^2 + 6L^2\rho^2) + 14K(1-\beta)^2\eta_l^2\|\nabla f(\tilde{w})\|^2] + 14K\beta^2L^2\eta_l^2\frac{1}{N}\sum_{i\in[N]}\mathbb{E}\|\delta_{i,k} - \delta\|^2 \\
&\overset{(b)}{\leq} 5K(2\beta^2L^2\eta_l^2\rho^2\sigma_l^2 + 7K\beta^2\eta_l^2(3\sigma_g^2 + 6L^2\rho^2) + 14K(1-\beta)^2\eta_l^2\|\nabla f(\tilde{w})\|^2) + 28\beta^2K^3L^4\eta_l^4\rho^2,
\end{aligned}$$

where (a) is due to the fact that $\eta_l \leq \frac{1}{\sqrt{30}\beta KL}$ and $\beta \leq \frac{1}{2}$ and (b) is from Lemma B.1. □

D.2. Convergence Analysis of Full client participation `MoFedSAM`

Lemma D.2 *For the full client participation scheme, we can bound $\mathbb{E}[\|\Delta^r\|^2]$ as follows:*

$$\mathbb{E}_r[\|\Delta^r\|^2] \leq \frac{2\beta^2 L^2 \rho^2}{KN}\sigma_l^2 + \frac{2}{K^2N^2}\mathbb{E}_r\left[\left\|\sum_{i,k}\beta\nabla f_i(\tilde{w}_{i,k}^r) + (1+\beta)\Delta^r\right\|^2\right].$$

Proof. For the full client participation strategy, we have:

$$\begin{aligned}
\mathbb{E}_r[\|\Delta^{r+1}\|^2] &\overset{(a)}{\leq} \frac{1}{K^2N^2\eta_l^2}\mathbb{E}_r\left[\left\|\sum_{i,k}\beta\eta_l\tilde{g}_{i,k}^r + (1-\beta)\eta_l\Delta^r\right\|^2\right] \\
&\overset{(b)}{\leq} \frac{\beta^2}{K^2N^2}\mathbb{E}_r\left[\left\|\sum_{i,k}(\tilde{g}_{i,k}^r - \nabla f_i(\tilde{w}_{i,k}^r))\right\|^2\right] + \frac{1}{K^2N^2}\mathbb{E}_r\left[\left\|\sum_{i,k}\beta\nabla f_i(\tilde{w}_{i,k}) + (1-\beta)\Delta^r\right\|^2\right] \\
&\overset{(c)}{\leq} \frac{\beta^2 L^2\rho^2}{KN}\sigma_l^2 + \frac{1}{K^2N^2}\mathbb{E}_r\left[\left\|\sum_{i,k}\beta\nabla f_i(\tilde{w}_{i,k}^r) + (1-\beta)\Delta^r\right\|^2\right] \\
&\overset{(d)}{\leq} \frac{\beta^2 L^2\rho^2}{KN}\sigma_l^2 + \frac{2(1-\beta)^2}{KN}\|f(\tilde{w}^r)\|^2 + \frac{\beta^2}{K^2N^2}\mathbb{E}_r\left[\left\|\sum_{i,k}f_i(\tilde{w}_{i,k}^r)\right\|^2\right],
\end{aligned}$$

where (a) is from Lemma A.2; (b) is from Lemma A.3 and (c) is from Lemma A.4. □

Lemma D.3 *(Descent Lemma of full client participation `MoFedSAM`). For all $r \in R-1$ and $i \in \mathcal{S}^r$, with the choice of learning rate, the iterates generated by `MoFedSAM` under full client participation in Algorithm 2 satisfy:*

$$\begin{aligned}
\mathbb{E}_r[f(\tilde{w}^{r+1})] \leq f(\tilde{w}^r) - K\eta_g\eta_l\left(\frac{1}{2} - 20K^2L^2\eta_l^2B^2\right)\|\nabla f(\tilde{w}^r)\|^2 + K\eta_g\eta_l(6K^2\eta_l^2\beta^4\rho^2 + 5K^2\eta_l\beta^4\rho^2\sigma^2 \\
+ 20K^3\eta_l^3\beta^2G^2 + 16K^3\eta_l^4\beta^6\rho^2 + \frac{\eta_g\eta_l\beta^3\rho^2}{N}\sigma^2)
\end{aligned}$$

where the expectation is w.r.t. the stochasticity of the algorithm.

Proof.

$$\begin{aligned}
\mathbb{E}_r[F(w^{r+1})] = \mathbb{E}_r[f(\tilde{w}^{r+1})] &\leq f(\tilde{w}^r) + \mathbb{E}_r\langle\nabla f(\tilde{w}^r), \tilde{w}^{r+1} - \tilde{w}^r]\rangle + \frac{L}{2}\mathbb{E}_r[\|\tilde{w}^{r+1} - \tilde{w}^r\|^2] \\
&\overset{(a)}{=} f(\tilde{w}^r) + \eta_g\mathbb{E}_r\langle\nabla f(\tilde{w}^r), -\Delta^{r+1} + \beta\nabla f(\tilde{w}^r) - \beta\nabla f(\tilde{w}^r)\rangle + \frac{L}{2}\eta_g^2\mathbb{E}_r[\|\Delta^{r+1}\|^2] \\
&\overset{(b)}{=} f(\tilde{w}^r) - \beta\eta_g\|\nabla f(\tilde{w}^r)\|^2 + \eta_g\langle\nabla f(\tilde{w}^r), \mathbb{E}_r[-\Delta^{r+1} + \beta\nabla f(\tilde{w}^r)]\rangle + \frac{L}{2}\eta_g^2\mathbb{E}_r[\|\Delta^{r+1}\|^2],
\end{aligned} \tag{12}$$

where (a) is from the iterate update given in Algorithm 3 and (b) results from the unbiased estimators.

For the third term, we bound it as follows:

$$
\begin{aligned}
&\langle \nabla f(\tilde{w}^r), \mathbb{E}_r[-\Delta^{r+1} + \beta \nabla f(\tilde{w}^r)] \rangle \\
&= -(1-\beta)\|\nabla f(\tilde{w}^r)\|^2 + \left\langle \sqrt{\beta}\nabla f(\tilde{w}^r), \mathbb{E}_r\left[-\frac{\sqrt{\beta}}{KN\eta_l} \sum_{i,k} (\eta_l \nabla f_i(\tilde{w}^r_{i,k}) - \eta_l \nabla f_i(\tilde{w}^r)) \right] \right\rangle \\
&\overset{\text{(a)}}{=} \left(\frac{3\beta}{2} - 1\right)\|\nabla f(\tilde{w}^r)\|^2 + \frac{\beta}{2K^2N^2}\mathbb{E}_r\left\| \sum_{i,k} (\nabla f_i(\tilde{w}^r_{i,k}) - \nabla f_i(\tilde{w}^r)) \right\|^2 - \frac{\beta}{2K^2N^2}\mathbb{E}_r\left\| \sum_{i,k} \nabla f_i(\tilde{w}^r_{i,k}) \right\|^2 \\
&\overset{\text{(b)}}{\leq} \left(\frac{3\beta}{2} - 1\right)\|\nabla f(\tilde{w}^r)\|^2 + \frac{\beta}{2KN}\sum_{i,k}\mathbb{E}_r\|\nabla f_i(\tilde{w}^r_{i,k}) - \nabla f_i(\tilde{w}^r)\|^2 - \frac{\beta}{2K^2N^2}\mathbb{E}_r\left\| \sum_{i,k} \nabla f_i(\tilde{w}^r_{i,k}) \right\|^2 \\
&\overset{\text{(c)}}{\leq} \left(\frac{3\beta}{2} - 1\right)\|\nabla f(\tilde{w}^r)\|^2 + \frac{\beta}{2KN}\sum_{i,k}\mathbb{E}_r\|\nabla f_i(\tilde{w}^r_{i,k}) - \nabla f_i(\tilde{w}^r)\|^2 - \frac{\beta}{2K^2N^2}\mathbb{E}_r\left\| \sum_{i,k} \nabla f_i(\tilde{w}^r_{i,k}) \right\|^2 \\
&\overset{\text{(d)}}{\leq} \left(\frac{3\beta}{2} - 1\right)\|\nabla f(\tilde{w}^r)\|^2 + \frac{\beta L^2}{2KN}\sum_{i,k}\mathbb{E}_r\|\tilde{w}^r_{i,k} - \tilde{w}^r\|^2 - \frac{\beta}{2K^2N^2}\mathbb{E}_r\left\| \sum_{i,k} \nabla f_i(\tilde{w}^r_{i,k}) \right\|^2 \\
&\overset{\text{(e)}}{\leq} \left(\frac{3\beta}{2} - 1\right)\|\nabla f(\tilde{w}^r)\|^2 + \beta L^2(\mathcal{E}_w + \mathcal{E}_\delta) - \frac{\beta}{2K^2N^2}\mathbb{E}_r\left\| \sum_{i,k} \nabla f_i(\tilde{w}^r_{i,k}) \right\|^2,
\end{aligned}
\tag{13}
$$

where (a) is from that $\Delta^r = \nabla f(\tilde{w}^r)$ and $\nabla f(\tilde{w}^r) = \sum_i \nabla f_i(\tilde{w}^r)$; (b), (c) and (e) are from Lemma A.2 and (d) is from Assumption 1. Plugging (13) into (12), we have:

$$
\begin{aligned}
&\mathbb{E}_r[f(\tilde{w}^{r+1})] \\
&\leq f(\tilde{w}^r) - \left(\eta_g - \frac{\beta\eta_g}{2}\right)\|\nabla f(\tilde{w}^r)\|^2 + \beta L^2 \eta_g(\mathcal{E}_w + \mathcal{E}_\delta) - \frac{\beta\eta_g}{2K^2N^2}\mathbb{E}_r\left\| \sum_{i,k} \nabla f_i(\tilde{w}^r_{i,k}) \right\|^2 + \frac{L\eta_g^2}{2}\mathbb{E}_r[\|\Delta^{r+1}\|^2] \\
&\overset{\text{(a)}}{\leq} f(\tilde{w}^r) - \left(\frac{3\beta\eta_g}{4} - \frac{2(1-\beta)^2 L\eta_g}{KN}\right)\|\nabla f(\tilde{w}^r)\|^2 + \beta L^2\eta_g(\mathcal{E}_w + \mathcal{E}_\delta) + \frac{\beta^2 L^3\rho^2\eta_g^2}{2KN}\sigma_l^2 \\
&- \frac{\beta\eta_g}{2K^2N^2}\mathbb{E}_r\left\| \sum_{i,k} \nabla f_i(\tilde{w}^r_{i,k}) \right\|^2 + \frac{L\beta^2\eta_g^2}{2K^2N^2}\mathbb{E}_r\left\| \sum_{i,k} \nabla f_i(\tilde{w}^r_{i,k}) \right\|^2 \\
&\overset{\text{(b)}}{\leq} f(\tilde{w}^r) - \beta\eta_g\left(\frac{3}{4} - \frac{2(1-\beta)L}{KN} - 70(1-\beta)K^2L^2\eta_l^2\right)\|\nabla f(\tilde{w}^r)\|^2 \\
&+ \beta\eta_g\left(10\beta^2L^4\eta_l^2\rho^2\sigma_l^2 + 35\beta^2KL^2\eta_l^2(3\sigma_g^2 + 6L^2\rho^2) + 28\beta^2K^3L^6\eta_l^4\rho^2 + 2K^2L^4\eta_l^2\rho^2 + \frac{\beta L^3\eta_g^2\rho^2}{2KN}\sigma_l^2\right) \\
&\overset{\text{(c)}}{\leq} f(\tilde{w}^r) - C\beta\eta_g\|\nabla f(\tilde{w}^r)\|^2 \\
&+ \beta\eta_g\left(10\beta^2L^4\eta_l^2\rho^2\sigma_l^2 + 35\beta^2KL^2\eta_l^2(3\sigma_g^2 + 6L^2\rho^2) + 28\beta^2K^3L^6\eta_l^4\rho^2 + 2K^2L^4\eta_l^2\rho^2 + \frac{\beta L^3\eta_g^2\rho^2}{2KN}\sigma_l^2\right),
\end{aligned}
$$

(a) is from Lemma D.2; (b) is from Lemmas B.1, D.1 and due to the fact that $\eta_g \leq \frac{1}{\beta L}$ and (c) is due to the fact that the condition $\frac{3}{4} - \frac{2(1-\beta)L}{KN} - 70(1-\beta)K^2L^2\eta_l^2 > C > 0$ and $\beta \leq \frac{1}{2}$ hold. □

Theorem D.4 *(Convergence of* `MoFedSAM`*). Let constant local and global learning rates $\eta_l \leq \frac{1}{\sqrt{30}\beta KL}$, $\eta_g \leq \frac{1}{\beta L}$ and $\beta \leq \frac{1}{2}$ and the condition $\frac{3}{4} - \frac{2(1-\beta)L}{KN} - 70(1-\beta)K^2L^2\eta_l^2 > C > 0$ holds. Under Assumptions 1-3 and with full client participation, the sequence of outputs $\{w^r\}$ generated by* `FedGSAM` *satisfies:*

$$
\min_{r\in[R]} \mathbb{E}\|\nabla F(w^r)\|^2 \leq \frac{f^0 - f^*}{C\beta\eta_g} + \Phi,
$$

where $\Phi = \frac{1}{C}(20\beta^3L^4\eta_l^2\rho^2\sigma^2 + 25\beta^2K^2L^2\eta_l^2G^2 + 20\beta^2K^4L^5\eta_l^4\rho^2 + 4\beta^2KL^4\eta_l^2\rho^2 + \frac{\beta L^3\rho^2\eta_g^2}{2KN})$. If we choose the learning rates $\eta_l = O(\frac{1}{\sqrt{RK}\beta L})$, $\eta_g = O(\frac{\sqrt{KN}}{\sqrt{R}\beta L})$ and the perturbation amplitude ρ proportional to the learning rate, e.g., $\rho = \frac{1}{\sqrt{R}}$,

we have:

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[\|F(w^{r+1})\|] = \mathcal{O}\Big(\frac{F\beta L}{\sqrt{RKN}} + \frac{\beta^2\sigma_g^2}{R} + \frac{L\sigma_l^2}{R^2\beta} + \frac{L^2}{R^2\beta^2}\Big).$$

Proof. Summing the result of Lemma D.3 for $r = [R]$ and multiplying both sides by $\frac{1}{C\beta\eta_g R}$, we have:

$$\begin{aligned}\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[\|F(w^{r+1})\|] &\le \frac{F}{C\beta\eta_g R} \\ &+ \frac{1}{C}\Big(10\beta^2L^4\eta_l^2\rho^2\sigma_l^2 + 35\beta^2KL^2\eta_l^2(3\sigma_g^2 + 6L^2\rho^2) + 28\beta^2K^3L^6\eta_l^4\rho^2 + 2K^2L^4\eta_l^2\rho^2 + \frac{\beta L^3\eta_g^2\rho^2}{2KN}\sigma_l^2\Big),\end{aligned}$$

where it is from that $F = f(\tilde{w}^0) - f^* \le f(\tilde{w}^r) - f(\tilde{w}^{r+1})$. If we choose the learning rates $\eta_l = O(\frac{1}{\sqrt{RK\beta L}})$, $\eta_g = O(\frac{\sqrt{KN}}{\sqrt{R}\beta L})$ and the perturbation amplitude ρ proportional to the learning rate, e.g., $\rho = \frac{1}{\sqrt{R}}$, we have

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[\|F(w^{r+1})\|] = \mathcal{O}\Big(\frac{F\beta L}{\sqrt{RKN}} + \frac{\beta^2\sigma_g^2}{R} + \frac{L^2\sigma_l^2}{R^2K} + \frac{L\sigma_l^2}{R^2\beta} + \frac{\beta L^2}{R^2} + \frac{KL^2}{R^3\beta^2} + \frac{L^2}{R^2\beta^2}\Big).$$

If we omit the larger order of each part, we have:

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[\|F(w^{r+1})\|] = \mathcal{O}\Big(\frac{F\beta L}{\sqrt{RKN}} + \frac{\beta^2\sigma_g^2}{R} + \frac{L\sigma_l^2}{R^2\beta} + \frac{L^2}{R^2\beta^2}\Big).$$

This completes the proof. □

D.3. Convergence Analysis of Partial client participation `MoFedSAM`

Lemma D.5 *For the partial client participation, we can bound $\mathbb{E}_r[\|\Delta^r\|^2]$ as follows:*

$$\mathbb{E}_r[\|\Delta^r\|^2] \le \frac{KL^2\eta_l^2\rho^2}{S}\sigma_l^2 + \frac{\eta_l^2}{S^2}\Big[\Big\|\sum_i \mathbb{P}\{i\in\mathcal{S}^r\}\sum_{j=0}^{K-1}\nabla f_i(\tilde{w}_{i,j}^r)\Big\|^2\Big].$$

Proof.

$$\begin{aligned}\mathbb{E}_r[\|\Delta^r\|^2] &\overset{(a)}{\le} \frac{1}{K^2S^2\eta_l^2}\mathbb{E}_r\Big[\Big\|\sum_{i\in\mathcal{S}^r}\sum_k \beta\eta_l\tilde{g}_{i,k}^r + (1-\beta)\eta_l\Delta^r\Big\|^2\Big] \\ &= \frac{1}{K^2S^2\eta_l^2}\mathbb{E}_r\Big[\Big\|\sum_i \mathbb{I}\{i\in\mathcal{S}^r\}\sum_k \beta\eta_l\tilde{g}_{i,k}^r - (1-\beta)\Delta^r\Big\|^2\Big] \\ &\overset{(b)}{=} \frac{\beta^2}{K^2S^2}\mathbb{E}_r\Big[\Big\|\sum_i\sum_{j=0}^{K-1}(\tilde{g}_{i,j}^r - \nabla f_i(\tilde{w}_{i,j}^r))\Big\|^2\Big] + \frac{1}{K^2S^2}\mathbb{E}_r\Big[\Big\|\sum_i \mathbb{I}\{i\in\mathcal{S}^r\}\sum_{j=0}^{K-1}\beta\nabla f_i(\tilde{w}_{i,j}^r) + (1-\beta)\Delta^r\Big\|^2\Big] \\ &\overset{(c)}{\le} \frac{\beta^2L^2\rho^2}{KS}\sigma_l^2 + \frac{2(1-\beta)^2}{KS}\|f(\tilde{w}^r)\|^2 + \frac{2\beta^2}{K^2S^2}\Big[\Big\|\sum_i \mathbb{P}\{i\in\mathcal{S}^r\}\sum_{j=0}^{K-1}\nabla f_i(\tilde{w}_{i,j})\Big\|^2\Big] \\ &= \frac{\beta^2L^2\rho^2}{KS}\sigma_l^2 + \frac{2(1-\beta)^2}{KS}\|f(\tilde{w}^r)\|^2 + \frac{2\beta^2}{K^2SN}\sum_{i=1}^{N}\mathbb{E}_r\Big\|\sum_{j=0}^{K-1}\nabla f_i(\tilde{w}_{i,j}^r)\Big\|^2 + \frac{2\beta^2(S-1)}{K^2SN^2}\mathbb{E}_r\Big\|\sum_{i=1}^{N}\sum_{j=0}^{K-1}\nabla f_i(\tilde{w}_{i,j}^r)\Big\|^2,\end{aligned}$$

where (a) is from Lemma A.2; (b) is from Lemma A.3 and (c) is from Lemma A.4. □

Lemma D.6 *(Descent Lemma of partial client participation* `MoFedSAM`*). For all $r \in R-1$ and $i \in \mathcal{S}^r$, with the choice of learning rate, the iterates generated by* `MoFedSAM` *under partial client participation in Algorithm 2 satisfy:*

$$\mathbb{E}_r[f(\tilde{w}^{r+1})] \leq f(\tilde{w}^r) - K\eta_g\eta_l\left(\frac{1}{2} - 20K^2L^2\eta_l^2B^2\right)\|\nabla f(\tilde{w}^r)\|^2 + K\eta_g\eta_l(6K^2\eta_l^2\beta^4\rho^2 + 5K^2\eta_l\beta^4\rho^2\sigma^2 + 20K^3\eta_l^3\beta^2G^2 + 16K^3\eta_l^4\beta^6\rho^2 + \frac{\eta_g\eta_l\beta^3\rho^2}{N}\sigma^2)$$

where the expectation is w.r.t. the stochasticity of the algorithm.

Proof.

$$\begin{aligned}\mathbb{E}_r[F(w^{r+1})] = \mathbb{E}_r[f(\tilde{w}^{r+1})] &\leq f(\tilde{w}^r) + \mathbb{E}_r\langle\nabla f(\tilde{w}^r), \tilde{w}^{r+1} - \tilde{w}^r]\rangle + \frac{L}{2}\mathbb{E}_r[\|\tilde{w}^{r+1} - \tilde{w}^r\|^2] \\ &= f(\tilde{w}^r) - \beta\eta_g\|\nabla f(\tilde{w}^r)\|^2 + \eta_g\langle\nabla f(\tilde{w}^r), \mathbb{E}_r[-\Delta^{r+1} + \beta\nabla f(\tilde{w}^r)]\rangle + \frac{L}{2}\eta_g^2\mathbb{E}_r[\|\Delta^{r+1}\|^2].\end{aligned} \tag{14}$$

Similar to full client participation strategy, we bound the third term in (14) as follows:

$$\langle\nabla f(\tilde{w}^r), \mathbb{E}_r[-\Delta^{r+1} + \beta\nabla f(\tilde{w}^r)]\rangle \leq \left(\frac{3\beta}{2} - 1\right)\|\nabla f(\tilde{w}^r)\|^2 + \beta L^2(\mathcal{E}_w + \mathcal{E}_\delta) - \frac{\beta}{2K^2N^2}\mathbb{E}_r\left\|\sum_{i,k}\nabla f_i(\tilde{w}_{i,k}^r)\right\|^2, \tag{15}$$

Plugging (15) into (14), we have:

$$\begin{aligned}&\mathbb{E}_r[f(\tilde{w}^{r+1})] \\ &\leq f(\tilde{w}^r) - \left(\eta_g - \frac{\beta\eta_g}{2}\right)\|\nabla f(\tilde{w}^r)\|^2 + \beta L^2\eta_g(\mathcal{E}_w + \mathcal{E}_\delta) - \frac{\beta\eta_g}{2K^2N^2}\mathbb{E}_r\left\|\sum_{i,k}\nabla f_i(\tilde{w}_{i,k}^r)\right\|^2 + \frac{L\eta_g^2}{2}\mathbb{E}_r[\|\Delta^{r+1}\|^2] \\ &\overset{(a)}{\leq} f(\tilde{w}^r) - \left(\frac{3\beta\eta_g}{4} - \frac{2(1-\beta)^2L\eta_g}{KS}\right)\|\nabla f(\tilde{w}^r)\|^2 + \beta L^2\eta_g(\mathcal{E}_w + \mathcal{E}_\delta) + \frac{\beta^2L^3\rho^2\eta_g^2}{2KS}\sigma_l^2 \\ &- \frac{\beta\eta_g}{2K^2N^2}\mathbb{E}_r\left\|\sum_{i,k}\beta\nabla f_i(\tilde{w}_{i,k}^r)\right\|^2 + \frac{L\beta^2\eta_g^2}{2K^2SN}\sum_i\mathbb{E}_r\left\|\sum_k\nabla f_i(\tilde{w}_{i,k}^r)\right\|^2 + \frac{L\beta^2(S-1)\eta_g^2}{K^2SN^2}\mathbb{E}_r\left\|\sum_{i,k}\nabla f_i(\tilde{w}_{i,k}^r)\right\|^2 \\ &\overset{(b)}{\leq} f(\tilde{w}^r) - \beta\eta_g\left(\frac{3}{4} - \frac{2(1-\beta)L}{KN} - 70(1-\beta)K^2L^2\eta_l^2 - \frac{90\beta L^3\eta_g\eta_l^2}{S} - \frac{3\beta L\eta_g}{2S}\right)\|\nabla f(\tilde{w}^r)\|^2 \\ &+ \beta\eta_g\Bigg(10\beta^2L^4\eta_l^2\rho^2\sigma_l^2 + 35\beta^2KL^2\eta_l^2(3\sigma_g^2 + 6L^2\rho^2) + 28\beta^2K^3L^6\eta_l^4\rho^2 + 2K^2L^4\eta_l^2\rho^2 + \frac{\beta L^3\eta_g^2\rho^2}{2KS}\sigma_l^2 \\ &+ \frac{\beta L\eta_g}{K^2SN}\left(30NK^2L^4\eta_l^2\rho^2\sigma_l^2 + 270NK^3L^2\eta_l^2\sigma_g^2 + 540NK^2L^4\eta_l^2\rho^2 + 72K^4L^6\eta_l^4\rho^2 + 6NK^4L^2\eta_l^2\rho^2 + 4NK^2\sigma_g^2 + 3NK^2L^2\rho^2\right)\Bigg) \\ &\overset{(c)}{\leq} f(\tilde{w}^r) - C\beta\eta_g\|\nabla f(\tilde{w}^r)\|^2 \\ &+ \beta\eta_g\Bigg(10\beta^2L^4\eta_l^2\rho^2\sigma_l^2 + 35\beta^2KL^2\eta_l^2(3\sigma_g^2 + 6L^2\rho^2) + 28\beta^2K^3L^6\eta_l^4\rho^2 + 2K^2L^4\eta_l^2\rho^2 + \frac{\beta L^3\eta_g^2\rho^2}{2KS}\sigma_l^2 \\ &+ \frac{\beta L\eta_g}{K^2SN}\left(30NK^2L^4\eta_l^2\rho^2\sigma_l^2 + 270NK^3L^2\eta_l^2\sigma_g^2 + 540NK^2L^4\eta_l^2\rho^2 + 72K^4L^6\eta_l^4\rho^2 + 6NK^4L^2\eta_l^2\rho^2 + 4NK^2\sigma_g^2 + 3NK^2L^2\rho^2\right)\Bigg),\end{aligned}$$

(a) is from Lemma D.2; (b) is from Lemmas B.1, D.1 and due to the fact that $\eta_g \leq \frac{S}{2\beta L(S-1)}$ and (c) is due to the fact that the condition $\frac{3}{4} - \frac{2(1-\beta)L}{KN} - 70(1-\beta)K^2L^2\eta_l^2 - \frac{90\beta L^3\eta_g\eta_l^2}{S} - \frac{3\beta L\eta_g}{2S} > C > 0$ and $\beta \leq \frac{1}{2}$ hold. □

Theorem D.7 *Let constant local and global learning rates η_l and η_g be chosen as such that $\eta_l \leq \frac{1}{\sqrt{30}\beta KL}$, $\eta_g \leq \frac{S}{2\beta L(S-1)}$ and the condition $\frac{3}{4} - \frac{2(1-\beta)L}{KN} - 70(1-\beta)K^2L^2\eta_l^2 - \frac{90\beta L^3\eta_g\eta_l^2}{S} - \frac{3\beta L\eta_g}{2S} > 0$ holds. Under Assumption 1-3 and with partial client participation, the sequence of outputs $\{w^r\}$ generated by* `MoFedSAM` *satisfies:*

$$\min_{r\in[R]}\mathbb{E}\|\nabla F(w^r)\|^2 \leq \frac{F^0 - F^*}{C\beta\eta_g} + \Phi,$$

where $\Phi = \frac{1}{C}[10\beta^2L^4\eta_l^2\rho^2\sigma_l^2 + 35\beta^2KL^2\eta_l^2(3\sigma_g^2 + 6L^2\rho^2) + 28\beta^2K^3L^6\eta_l^4\rho^2 + 2K^2L^4\eta_l^2\rho^2 + \frac{\beta L^3\eta_g^2\rho^2}{2KS}\sigma_l^2 + \frac{\beta L\eta_g}{K^2SN}\Big(30NK^2L^4\eta_l^2\rho^2\sigma_l^2 + 270NK^3L^2\eta_l^2\sigma_g^2 + 540NK^2L^4\eta_l^2\rho^2 + 72K^4L^6\eta_l^4\rho^2 + 6NK^4L^2\eta_l^2\rho^2 + 4NK^2\sigma_g^2 + 3NK^2L^2\rho^2\Big)]$. *If we choose the learning rates* $\eta_l = \frac{1}{\sqrt{RK\beta L}}$, $\eta_g = \frac{\sqrt{KS}}{\sqrt{R\beta L}}$ *and perturbation amplitude* ρ *proportional to the learning rate, e.g.,* $\rho = \frac{1}{\sqrt{R}}$, *we have:*

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[\|F(w^{r+1})\|] = \mathcal{O}\Big(\frac{FL}{\sqrt{RKS}} + \frac{\sqrt{K}G^2}{\sqrt{RS}} + \frac{L^2\sigma^2}{R^{3/2}K} + \frac{\sqrt{K}L^2}{R^{3/2}\sqrt{S}}\Big).$$

Proof. Summing the above result for $r = [R]$ and multiplying both sides by $\frac{1}{C\beta\eta_g R}$, we have

$$\begin{aligned}
&\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[\|F(w^{r+1})\|] \le \frac{f(\tilde{w}^r) - f(\tilde{w}^{r+1})}{C\beta\eta_g R} \\
&+ \frac{1}{C}\Big(10\beta^2L^4\eta_l^2\rho^2\sigma_l^2 + 35\beta^2KL^2\eta_l^2(3\sigma_g^2 + 6L^2\rho^2) + 28\beta^2K^3L^6\eta_l^4\rho^2 + 2K^2L^4\eta_l^2\rho^2 + \frac{\beta L^3\eta_g^2\rho^2}{2KS}\sigma_l^2 \\
&+ \frac{\beta L\eta_g}{K^2SN}\Big(30NK^2L^4\eta_l^2\rho^2\sigma_l^2 + 270NK^3L^2\eta_l^2\sigma_g^2 + 540NK^2L^4\eta_l^2\rho^2 + 72K^4L^6\eta_l^4\rho^2 + 6NK^4L^2\eta_l^2\rho^2 + 4NK^2\sigma_g^2 + 3NK^2L^2\rho^2\Big)\Big) \\
&\le \frac{F}{C\beta\eta_g R} + \frac{1}{C}\Big(10\beta^2L^4\eta_l^2\rho^2\sigma_l^2 + 35\beta^2KL^2\eta_l^2(3\sigma_g^2 + 6L^2\rho^2) + 28\beta^2K^3L^6\eta_l^4\rho^2 + 2K^2L^4\eta_l^2\rho^2 + \frac{\beta L^3\eta_g^2\rho^2}{2KS}\sigma_l^2 \\
&+ \frac{\beta L\eta_g}{K^2SN}\Big(30NK^2L^4\eta_l^2\rho^2\sigma_l^2 + 270NK^3L^2\eta_l^2\sigma_g^2 + 540NK^2L^4\eta_l^2\rho^2 + 72K^4L^6\eta_l^4\rho^2 + 6NK^4L^2\eta_l^2\rho^2 + 4NK^2\sigma_g^2 + 3NK^2L^2\rho^2\Big)\Big),
\end{aligned}$$

where the second inequality uses $F = f(\tilde{w}^0) - f^* \le f(\tilde{w}^r) - f(\tilde{w}^{r+1})$. If we choose the learning rates $\eta_l = \frac{1}{\sqrt{RK\beta L}}$, $\eta_g = \frac{\sqrt{KS}}{\sqrt{R\beta L}}$ and perturbation amplitude ρ proportional to the learning rate, e.g., $\rho = \frac{1}{\sqrt{R}}$, we have:

$$\begin{aligned}
\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[\|F(w^{r+1})\|] = \mathcal{O}\Big(&\frac{\beta FL}{\sqrt{RKS}} + \frac{\sigma_g^2}{R} + \frac{\beta\sqrt{K}\sigma_g^2}{\sqrt{RS}} + \frac{\sqrt{KS}\sigma_g^2}{R^{3/2}} + \frac{L^2\sigma_l^2}{R^2K} + \frac{L\sigma_l^2}{R^2\beta} + \frac{L^2\sqrt{KS}\sigma_l^2}{R^{5/2}} \\
&+ \frac{\beta L^2}{R^2} + \frac{\beta L^3}{R^2} + \frac{L^2}{R^3\beta^2} + \frac{L^2}{R^2\beta^2} + \frac{L^3}{R^2\beta\sqrt{KS}} + \frac{\sqrt{K}L^2}{R^{7/2}\sqrt{S}\beta^4} + \frac{K^{3/2}L}{R^{5/2}\sqrt{S}}\Big),
\end{aligned}$$

If the number of sampling clients are larger than the number of epochs, i.e., $S \ge K$, and omitting the larger order of each part, we have:

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[\|F(w^{r+1})\|] = \mathcal{O}\Big(\frac{\beta FL}{\sqrt{RKS}} + \frac{\sqrt{K}G^2}{\sqrt{RS}} + \frac{L^2\sigma^2}{R^{3/2}K} + \frac{\sqrt{K}L^2}{R^{3/2}\sqrt{S}}\Big).$$

This completes the proof. □

E. Experimental Setup

Table 3. Datasets and models.

Dataset	Task	Clients	Total samples	Model
EMNIST (Cohen et al., 2017)	Handwritten character recognition	100/50	81,425	2-layer CNN + 2-layer FFN
CIFAR-10 (Krizhevsky et al., 2009)	Image classification	100/50	60,000	ResNet-18 (He et al., 2016)
CIFAR-100 (Krizhevsky et al., 2009)	Image classification	100/50	60,000	ResNet-18 (He et al., 2016)

We ran the experiments on a CPU/GPU cluster, with RTX 2080Ti GPU, and used PyTorch (Paszke et al., 2019) to build and train our models. The description of datasets is introduced in Table 3.

E.1. Dataset Description

EMNIST (Cohen et al., 2017) is a 62-class image classification dataset. In this paper, we use 20% of the dataset, and we divide this dataset to each client based on Dirichlet allocation of parameter 0.6 over 100 client by default. We train the same CNN as in (Reddi et al., 2020; Dieuleveut et al., 2021), which includes two convolutional layers with 3×3 kernels, max pooling, and dropout, followed by a 128 unit dense layer.

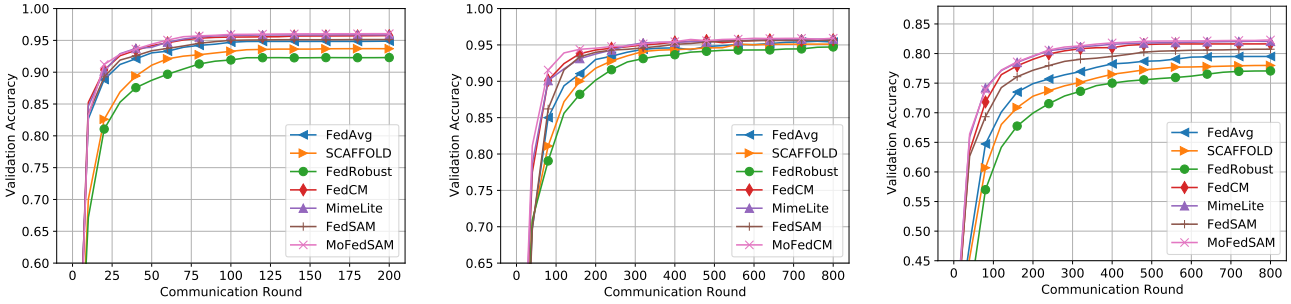
CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009) are labeled subsets of the 80 million images dataset. They both share the same 60,000 input images. CIFAR-100 has a finer labeling, with 100 unique labels, in comparison to CIFAR-10, having 10 unique labels. The Dirichlet allocation of these two datasets are also 0.6. For both of them, we train ResNet-18 (He et al., 2016) architecture.

E.2. Hyperparameters

For each algorithm and each dataset, the learning rate was set via grid search on the set $\{10^{-0.5}, 10^{-1}, 10^{-1.5}, 10^{-2}\}$. `FedCM`, `MimeLite` and `MoFedSAM` momentum term β was tuned via grid search on $\{0.01, 0.1, 0.2, 0.5, 1\}$. The global learning rate $\eta_g = 1$, and local learning rate $\eta_l = 0.1$ by default.

F. Additional Experiments

F.1. Training accuracy on different datasets



(a) EMNIST dataset. (b) CIFAR-10 dataset. (c) CIFAR-100 dataset.

Figure 4. Training accuracy on different datasets.

Table 4. Impact of the heterogeneity on EMNIST dataset (IID, Dirichlet 0.6 and Dirichlet 0.3).

Algorithm	IID			Dirichlet 0.6			Dirichlet 0.3		
	Train	Validation	Round	Train	Validation	Round	Train	Validation	Round
`FedAvg`	96.98(0.73)	89.95(1.95)	32	95.07 (0.94)	84.38 (4.03)	43	93.66(1.27)	82.83(4.42)	61
`SCAFFOLD`	96.04(1.01)	88.79(2.38)	51	93.85 (1.31)	84.09 (4.56)	69	92.85(1.68)	82.01(4.95)	88
`FedRobust`	95.63(0.56)	87.67(1.63)	66	93.17 (0.62)	83.70 (3.37)	91	92.10(1.00)	81.80(3.79)	103
`FedCM`	97.47(0.87)	91.13(2.07)	18	96.22 (1.16)	84.85 (4.22)	25	94.83(1.29)	83.09(4.58)	47
`MimeLite`	97.26(0.85)	91.29(2.11)	16	95.73 (0.49)	84.88 (3.04)	38	94.90(1.33)	83.14(4.55)	46
`FedSAM`	97.42(0.49)	90.22(1.50)	22	96.16 (1.14)	84.75 (4.11)	28	94.32(0.91)	82.97(3.56)	53
`MoFedSAM`	97.58(0.51)	91.52(1.53)	13	96.42 (0.42)	85.07 (2.95)	24	94.98(0.95)	83.28(3.59)	41

Figure 4 shows that the training accuracy on different datasets. Comparing with the validation accuracy results in Figure 1, the performance divergence is not clear. The reason is because the global model is easy to overfit the training dataset. Since the distribution of validation dataset on each client is different from training datasets, compared benchmarks perform less generalization. This indicates that our proposed algorithms benefits. Although `FedSAM` does not show better performance compared to the momentum FL, i.e., `FedCM` and `MimeLite`, it saves more transmission costs, since it does not need to download Δ^r. For example, on CIFAR-100 dataset, FedCM achieves 85.26% training accuracy with 4.41% deviation of local models, however, it obtains 54.09% validation accuracy with 14.38% deviation. For `MoFedSAM` algorithm, it

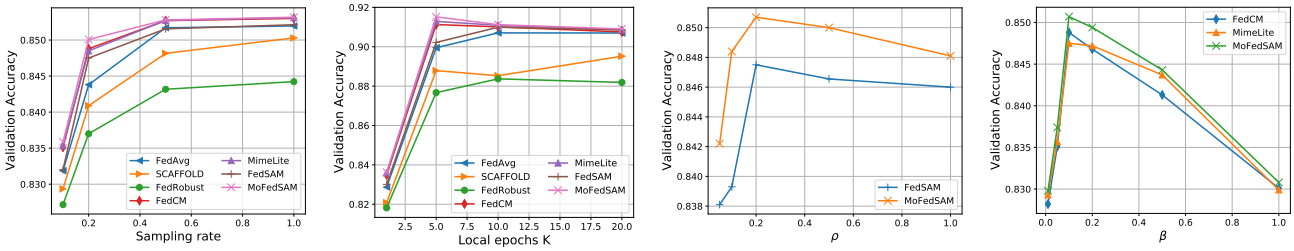
Table 5. Impact of the heterogeneity on CIFAR-100 dataset (IID, Dirichlet 0.6 and Dirichlet 0.3).

Algorithm	IID			Dirichlet 0.6			Dirichlet 0.3		
	Train	Validation	Round	Train	Validation	Round	Train	Validation	Round
FedAvg	84.68 (1.46)	58.97 (3.56)	253	79.57 (1.84)	53.57 (5.40)	302	77.61 (1.99)	51.22 (6.17)	593
SCAFFOLD	83.41 (2.07)	57.16 (4.32)	327	78.49 (2.02)	51.49 (5.87)	551	76.30 (2.67)	48.89 (6.59)	-
FedRobust	82.58 (1.35)	55.87 (3.35)	378	76.80 (1.70)	49.06 (4.75)	893	75.26 (1.87)	47.92 (5.86)	-
FedCM	87.05 (1.48)	59.64 (3.73)	149	82.46 (2.00)	55.73 (5.11)	189	79.91 (2.02)	52.57 (6.28)	410
MimeLite	87.42 (1.56)	59.87 (3.67)	143	82.53 (2.08)	55.82 (5.04)	182	79.96 (2.00)	52.60 (6.31)	397
FedSAM	85.65 (1.27)	59.11 (3.11)	228	81.04 (1.59)	54.69 (4.36)	245	78.05 (1.71)	51.78 (5.43)	561
MoFedSAM	87.82 (1.32)	60.02 (3.20)	129	82.62 (1.53)	56.60 (4.42)	124	80.09 (1.77)	52.90 (5.62)	373

can achieve 86.02% training accuracy with 3.23% deviation of local models, and 55.13% validation accuracy with 3.25% deviation of local models.

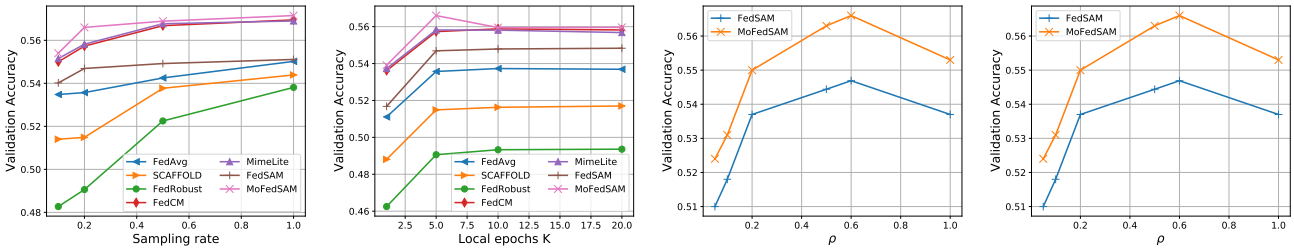
Tables 2, 4 and 5 aim to show the impact of heterogeneous degrees of FL. From these results, we can clearly see that increasing the degree of heterogeneity makes huge degradation of learning performance. However, it does not effect the training accuracy significantly. For example, on CIFAR-10 dataset, FedSAM obtains 95.42%, 94.20%, and 92.90% training accuracy, when heterogeneity is IID, Dirichlet 0.6 and Dirichlet 0.3, and 87.36%, 82.55% and 79.82% for validation accuracy. More specifically, the influence of heterogeneity for our proposed algorithms are less than compared benchmarks, which is due to the fact that the more generalized global model, the less impact of distribution shift.

F.2. Impact of hypeparameters



(a) Impact of S. (b) Impact of K. (c) Impact of ρ. (d) Impact of β.

Figure 5. Impacts of different parameters on EMNIST dataset.



(a) Impact of S. (b) Impact of K. (c) Impact of ρ. (d) Impact of β.

Figure 6. Impacts of different parameters on CIFAR-100 dataset.

Figures 3-6 aim to show the impacts of different hyperparameters, i.e., the number of participated clients S in each communication round, the number of local epochs K, the perturbation control parameter ρ of SAM optimizer, and the momentum parameter β. We can see that increasing S can improve the performance. However, increasing K cannot guarantee increasing the performance. For ρ and β, they depend on the different algorithms and datasets. By grid searching, it is not difficult to find the suitable value to optimize the performance.