LAYER-SELECTIVE SHARPNESS-AWARE MINIMIZATION FOR EFFICIENT AND ACCURATE FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning (FL) is a distributed learning paradigm that trains a model on distributed private data while preserving data privacy. To improve the global model's generalization, recent studies have increasingly applied Sharpness-Aware Minimization (SAM), known to guide the model toward a flat region in the parameter space, to local training in FL. However, naive application of SAM to local training doubles the computational cost without noticeably improving the validation accuracy of the global model. Previous works have focused on the discrepancy between global and local loss landscapes, however, our study finds that layer-wise sharpness plays a more important role in determining the impact of local SAM on the global model's generalization. In this paper, we propose a general FL framework, FedLSAM that perturbs only a subset of layers with a significant impact on reducing global sharpness, rather than perturbing the entire model. Experimental results show that the proposed method improves generalization performance compared with perturbing all layers, while having a negligible additional computational cost.

1 INTRODUCTION

Federated Learning (FL) is a distributed learning paradigm in which multiple clients independently train local models on their own data, which are then periodically aggregated by a central server without sharing the private data McMahan et al. (2017). In practical FL environments, clients' private data typically vary in both quantity and distribution, and under such non-IID settings Hsu et al. (2019), each client's local objective pushes model updates toward a different direction. As local training progresses, the global model may overfit to specific client distributions or converge to a sharp minimum, which harms its generalization ability on unseen data or on data from other clients.

Sharpness-Aware Minimization (SAM) is an optimization method which accounts for the geometry of the loss landscape to address the generalization issue Foret et al. (2021). SAM encourages models to converge to a flat minimum, in which the loss does not sharply increase within a small neighborhood around the current parameters. Instead of directly minimizing the original empirical loss, SAM formulates a min-max optimization problem that minimizes the maximum loss obtained after adding an adversarial perturbation to the model parameters. Such flat minima are known to provide better generalization performance Keskar et al. (2016).

Several studies have attempted to employ SAM in FL environments to tackle the poor generalization issue. FedSAM Qu et al. (2022), for example, applies the SAM to the local training of each client, encouraging the models to converge to flatter regions. However, in non-IID settings where local data distributions deviate from the global distribution, the local perturbations computed with clients' private data may not effectively lead the global model to flatter regions, and even amplify the inconsistency among client updates. This limitation suggests that simply applying SAM as a local optimizer is insufficient to improve the generalization performance of the global model.

In this paper, we provide a new perspective on the reason for the poor generalization performance when SAM is applied to FL. Our study finds that *a few network layers exhibit loss landscapes similar to those of the corresponding global model layers while other layers have significantly different loss*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

landscapes. This finding clearly explains why the conventional model perturbation applied during local training degrades the generalization performance of the global model. Furthermore, it provides a practical insight into how to reduce the high model perturbation cost of SAM while still achieving its superior generalization performance.

Building on these insights, we tackle the poor generalization issue of SAM in federated settings by designing an efficient layer-wise adaptive model perturbation scheme. We propose `FedLSAM`, a general FL framework that recovers the superior generalization performance of SAM in federated settings while significantly reducing the model perturbation cost compared to conventional SAM. `FedLSAM` selectively applies SAM to a few critical layers on the client side, which are chosen based on the magnitude of global updates. `FedLSAM`'s superior generalization performance empirically demonstrates that the loss landscape discrepancy does not exist throughout the whole network, and the gap can be mitigated by adaptively perturbing only a few critical layers.

We evaluate our `FedLSAM` framework using representative FL benchmarks and demonstrate the improved generalization performance of the global model with only marginal extra cost. Compared to several recently proposed SAM-based FL methods, such as FedLESAM, FedGF, and FedWMSAM, `FedLSAM` achieves similar or even higher validation accuracy while maintaining training throughput similar to that of FedAvg. For example, `FedLSAM` applies SAM at a subset of ResNet-18 layers and improves generalization performance (4.71% higher validation accuracy than FedSAM and 3.07% higher than FedWMSAM). Our results suggest that layer-wise SAM can be a practical choice as an optimizer to improve generalization in general FL environments.

2 Related Work

2.1 Federated Learning with SAM

FedSAM incorporates SAM into local client training to improve the generalization of the global model, but inherits the two backpropagation steps required by standard SAM Qu et al. (2022). Subsequent federated SAM methods have mainly focused on aligning the perturbation with the global loss landscape while reducing the resulting perturbation overhead. MoFedSAM utilizes global update from previous round Qu et al. (2022). FedSMOO estimates global perturbation with dynamic regularization Sun et al. (2023). FedGF accounts for global flatness by transmitting an additional perturbed model to clients Lee & Yoon (2024), while FedLESAM and FedWMSAM estimate global perturbation directions from consecutive global updates and weighted momentum, respectively Fan et al. (2024); Li et al. (2026). Although the latter approaches reduce client-side gradient computation, they require auxiliary model storage or additional communication. Furthermore, these methods implicitly assume that the whole model parameters are perturbed together, without considering their individual contributions to the flatness of the loss landscape.

2.2 Efficient SAM with Selective Perturbation

In centralized learning, several studies have investigated whether perturbing the entire model is necessary for SAM. ESAM reduces the computational overhead of SAM by applying stochastic perturbations to a subset of model parameters Du et al. (2021). SAM-ON restricts perturbations to the affine parameters of normalization layers and shows that full-model perturbation is not always necessary for effective sharpness-aware training Mueller et al. (2023). Layer-wise SAM further reduces the perturbation cost by identifying and perturbing only a few critical layers while retaining the gradient-norm penalizing effect of SAM Lee (2024). However, these methods are designed for centralized training and do not consider distributed or federated learning environments and how the partial model perturbation affects the global model's behavior.

3 Method

3.1 Federated Learning with Layer-wise Sharpness-Aware Minimization

We first define a few notations. The FL runs for T communication rounds in total, and each worker $i \in [N]$ runs τ local update steps per round, where N denotes the total number of workers. The global model at round t is denoted by $\mathbf{x}_t$ while the locally trained by worker i is denoted by $\mathbf{x}_t^i$.

For convenience, we define a subset of layers $\mathcal{S} \in [L]$, where $[L] = \{1, 2, \cdots, L\}$ is the total set of layers. We then define 'global update' at round t as $\Delta \mathbf{x}_t := \mathbf{x}_t - \mathbf{x}_{t-1}$. The update consists of the layer-wise updates as follows.

$$\Delta \mathbf{x}_t = \left(\Delta \mathbf{x}_t^1, \Delta \mathbf{x}_t^2, \cdots, \Delta \mathbf{x}_t^L\right).$$

We also define gradient subset $\nabla f'(\cdot)$ with respect to $\mathcal{S}$ as follows.

$$\nabla f'(\mathbf{x}_{t,j}^i) \subset \nabla f(\mathbf{x}_{t,j}^i) = \{\nabla_1 f(\mathbf{x}_{t,j}^i), \nabla_2 f(\mathbf{x}_{t,j}^i), \cdots, \nabla_L f(\mathbf{x}_{t,j}^i)\},$$

where $\nabla_l f(\mathbf{x}_{t,j}^i)$ refers to the gradient of $f(\cdot)$ with respect to the parameters of layer l.

We propose `FedLSAM`, a general federated learning framework that applies SAM in a layer-wise manner to efficiently improve the generalization performance of the global model. Before every communication round, `FedLSAM` selects k layers with the largest update norm. At each round, the server selects S_t, a subset of layers with the largest update norm as follows.

$$S_t = \text{Top-k}_{l \in \{1, \cdots, L\}} \|\Delta \mathbf{x}_t^l\|_2$$

Then, it broadcasts $\mathbf{x}_t$ together with S_t to all active clients. Note that S_t is just a few layer indices, and the additional communication cost for broadcasting it is negligible.

Once each active client receives them, it runs local layer-wise SAM for τ steps. First, the layers in S_t are perturbed as follows.

$$\mathbf{x}_{t,j+1/2}^{l,i} = \begin{cases} \mathbf{x}_{t,j}^{l,i} + r \frac{\nabla f'(\mathbf{x}_{t,j}^{l,i})}{\|\nabla f'(\mathbf{x}_{t,j}^i)\|}, & l \in S_t \\ \mathbf{x}_{t,j}^{l,i}, & l \notin S_t \end{cases} \tag{1}$$

where $\mathbf{x}_{t,j}^{l,i}$ denotes the layer l's model parameters of worker i at step j in round t. Then, the model is updated using the gradients computed using $\mathbf{x}_{t,j+1/2}^{l,i}$ as follows.

$$\mathbf{x}_{t,j+1}^{l,i} = \mathbf{x}_{t,j}^{l,i} - \eta \nabla F(\mathbf{x}_{t,j+1/2}^{l,i}), \forall l \in [L]. \tag{2}$$

Algorithm 1 presents pseudo code of `FedLSAM` which updates local model using (2). At the beginning of each communication round, the server selects k layers to perturb based on the global update norm as described in Section 3.2. Then, activated clients receive the global model from the server and then run local SAM. At each local step, the clients perturb only the selected k layers and compute the gradients for local updates. In the following subsections, we will discuss the gradient norm penalizing effect of the layer-wise SAM under federated settings, the computational cost of layer-wise perturbation, and the convergence properties of Algorithm 1.

3.2 Layer-Selection Criterion

Following Foret et al. (2021), we approximate the loss of the perturbed model using a second-order Taylor expansion as follows.

$$F(\mathbf{x} + \epsilon_S) \approx F(\mathbf{x}) + \nabla_S F(\mathbf{x})^\top \epsilon_S + \frac{1}{2} \epsilon_S^\top H_{SS} \epsilon_S, \tag{3}$$

where ϵ_S is the perturbation vector that is nonzero only for the layers in S, $\nabla_S F(\mathbf{x})$ is the full-batch gradient with respect to the parameters of those layers, and H_{SS} is the corresponding block of the Hessian. Since $F(\mathbf{x})$ does not depend on ϵ_S, it can be ignored in the maximization. In addition, since $\|\epsilon_S\| \leq r$ with a small perturbation radius r, the second-order term is $O(r^2)$ and is dominated by the first-order term. Thus, the perturbation vector is obtained by solving

$$\max_{\|\epsilon_S\| \leq r} \nabla_S F(\mathbf{x})^\top \epsilon_S.$$

Based on Cauchy-Schwarz inequality,

$$\nabla_S F(\mathbf{x})^\top \epsilon_S \leq \|\nabla_S F(\mathbf{x})\| \|\epsilon_S\| \leq r \|\nabla_S F(\mathbf{x})\|.$$

Algorithm 1 `FedLSAM` pseudo code.

Input: r: the ascent scaling factor, τ: the number of local update steps, a: the number of active clients for each communication round, η: the learning rate
for $t \in \{0, \cdots, T-1\}$ **do**
  if $t \geq 1$ **then**
    $\Delta x_t \leftarrow x_t - x_{t-1}$.
    $S_t \leftarrow \mathrm{TopK}_{l\in[L]}\{\|\Delta x_t^l\|_2\}$.
  end if
  $\mathcal{A} \leftarrow$ Randomly sample a clients from the total client set.
  Broadcast $\mathbf{x}_t$ to all clients in $\mathcal{A}$.
  for each client $i \in \mathcal{A}$ **do**
    $x_{t,0}^i \leftarrow x_t$.
    for $j \in \{1, \cdots, \tau\}$ **do**
      Compute $f'(\mathbf{x}_{t,j}^i)$
      $\mathbf{x}_{t,j+1/2}^{l,i} \leftarrow$ Layer-wise Perturbation using Equation (1)
      $\mathbf{x}_{t,j+1}^{l,i} \leftarrow \mathbf{x}_{t,j}^{l,i} - \eta \nabla F(\mathbf{x}_{t,j+1/2}^{l,i})$
    end for
    Aggregate $x_{t,\tau}^i$ to the server.
  end for
  $x_{t+1} \leftarrow \frac{1}{a}\sum_{i\in\mathcal{A}} \mathbf{x}_{t,\tau}^i$
end for
Output: w_T

Both equalities hold only when ϵ_S is aligned with $\nabla_S F(\mathbf{x})$ and $\|\epsilon_S\| = r$. Therefore, assuming $\nabla f'(\mathbf{x}) \neq \mathbf{0}$, the unique solution is

$$\epsilon_S^* = r\frac{\nabla_S F(\mathbf{x})}{\|\nabla_S F(\mathbf{x})\|}. \tag{4}$$

By inserting (4) into (3), we have:

$$F(\mathbf{x}+\epsilon_S) \approx F(\mathbf{x}) + r\|\nabla_S F(\mathbf{x})\| + \frac{r^2}{2}\frac{\nabla_S F(\mathbf{x})^\top}{\|\nabla_S F(\mathbf{x})\|} H_{SS} \frac{\nabla_S F(\mathbf{x})}{\|\nabla_S F(\mathbf{x})\|}. \tag{5}$$

For a sufficiently small r, the loss increase along the ascent direction is governed by the first-order term $r\|\nabla_S F(\mathbf{x})\|$. This implies that perturbing layers with a small gradient norm yields only a marginal loss increase, leaving little sharpness for SAM to reduce. Since the global update of each layer reflects the gradient of the global objective, we use the norm of global updates as our layer-selection criterion.

3.3 Gradient Norm Penalizing Effect Analysis

Perturbation of the top-k layers effectively penalizes the gradient norm of the global model. – As known in previous works, SAM is approximately equivalent to penalizing the first order gradient norm of the loss function Andriushchenko & Flammarion (2022); Lee (2024); Zhao et al. (2022). The proposed `FedLSAM` effectively approximates the full perturbation as follows. First, the loss function of the original SAM is written as follows.

$$F(\mathbf{x}_t) + r\sqrt{\sum_{l\in[L]} \|\nabla_{\mathbf{x}_t^l} F(\mathbf{x}_t)\|_2^2}.$$

Likewise, the loss function of layer-wise SAM is written as follows.

$$F(\mathbf{x}_t) + r\sqrt{\sum_{l\in S_t} \|\nabla_{\mathbf{x}_t^l} F(\mathbf{x}_t)\|_2^2}.$$

Then, if the gradient norm of the top-k layers dominates over that of the other layers, these loss functions can be considered as similar:

$$F(\mathbf{x}_t) + r\sqrt{\sum_{l\in S_t} \|\nabla_{\mathbf{x}_t^l} F(\mathbf{x}_t)\|_2^2} \approx F(\mathbf{x}_t) + r\sqrt{\sum_{l\in[L]} \|\nabla_{\mathbf{x}_t^l} F(\mathbf{x}_t)\|_2^2}. \tag{6}$$

Table 1: Per-round computation, communication, and memory cost of federated SAM methods. Costs are expressed in units of M, the cost of a single backpropagation (Comp.) or a single model-sized transfer or buffer (Comm. and Mem.). M_{S_t} denotes the size of the selected top-k layers perturbed by `FedLSAM`.

Algorithm	Comp. Cost	Comm. Cost	Mem. Cost
FedAvg McMahan et al. (2017)	M	$2M$	–
FedGF Lin et al. (2020)	$2M$	$3M$	–
FedSAM Qu et al. (2022)	$2M$	$2M$	–
MoFedSAM Qu et al. (2022)	$2M$	$3M$	–
FedSMOO Sun et al. (2023)	$2M$	$4M$	$2M$
FedLESAM Fan et al. (2024)	M	$2M$	M
FedWMSAM Li et al. (2026)	M	$3M$	–
FedLSAM (proposed)	$M + M_{S_t}$	$2M$	–

Since gradient norms are unlikely to be uniform across layers, perturbing a few layers with the largest norms still yields an effective approximation of the full-layer loss function. In this study, we show that a small value of k suffices to make a practical trade-off between perturbation quality and system-level efficiency.

3.4 Model Perturbation Cost Analysis

Table 1 compares the per-round client-side cost in computation, communication, and auxiliary local memory. The computation cost counts only the dominant backpropagation, omitting minor operations such as perturbation and interpolation. The communication cost is the total per-round download and upload, and '–' denotes no auxiliary local state. As a baseline, FedAvg's computational cost is M since it performs a single gradient descent step per mini-batch. Its communication cost is $2M$, accounting for both uplink and downlink communication. FedAvg requires no additional memory beyond the model parameters and gradients themselves.

FedSAM doubles the computation to $2M$ due to the extra backpropagation for perturbation. FedLESAM estimates the perturbation from the difference between previous and current global models, lowering the computation to M at the cost of storing an M-sized previous model. FedGF keeps the two backpropagations and broadcasts an additional perturbed model, raising the communication to $3M$. FedWMSAM also lowers the computation to M via a momentum-based perturbation but downloads an extra momentum buffer, raising the communication to $3M$.

In contrast, `FedLSAM` perturbs only the top-k layers, so its computational cost is $M + M_{S_t}$, strictly smaller than the $2M$ of FedSAM since $M_{S_t} < M$. It also has the minimal $2M$ communication cost and requires no auxiliary state. Since the selected high-norm layers lie mostly on the output side, the ascent-step backpropagation terminates early and M_{S_t} remains far smaller than M. `FedLSAM` *is thus the only method that reduces the perturbation cost without additional communication or memory overhead.*

3.5 Convergence Analysis

To validate that `FedLSAM` is a practical federated optimizer, we analyze its convergence properties and provide a convergence guarantee for non-convex and smooth optimization problems:

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) := \frac{1}{m}\sum_{i=1}^{m} F_i(\mathbf{x}),$$

where m denotes the total number of clients, and $F_i(\mathbf{x}) := \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[f(\mathbf{x};\xi_i)]$ represents the local objective function associated with the data distribution $\mathcal{D}_i$ of client i. Here, $f(\mathbf{x};\xi_i)$ is the loss function evaluated on a sample ξ_i drawn from $\mathcal{D}_i$.

Our analysis is based on the following assumptions.

Assumption 1 (Lipschitz continuity)**.** *There exists a constant $\mathcal{L} > 0$ such that $\|\nabla F_i(\mathbf{a}) - \nabla F_i(\mathbf{b})\| \le \mathcal{L}\|\mathbf{a}-\mathbf{b}\|, \quad \forall \mathbf{a},\mathbf{b}\in\mathbb{R}^d, \forall i\in[m]$.*

Assumption 2 (Unbiased local gradients)**.** *The stochastic gradient is unbiased such that* $\mathbb{E}[\nabla f(\mathbf{x}, \xi_i)] = \nabla F_i(\mathbf{x}), \quad \forall i \in [m]$.

Assumption 3 (Bounded local and global variances)**.** *There exist constants* $\sigma > 0$ *and* $\zeta > 0$ *such that* $\mathbb{E}_{\xi_i \sim D_i}\left[\|\nabla f(\mathbf{x}, \xi_i) - \nabla F_i(\mathbf{x})\|^2\right] \leq \sigma^2, \quad \forall i \in [m]$, *and* $\mathbb{E}\left[\|\nabla F_i(\mathbf{x}) - \nabla F(\mathbf{x})\|^2\right] \leq \zeta^2, \quad \forall i \in [m]$.

Assumption 4 (Bounded variance of normalized gradients)**.** *For each layer* $s \in [L]$ *and every subset* $\mathcal{S} \subset [L]$ *with* $s \in \mathcal{S}$*, there exists a constant* $\sigma_{n,s} > 0$ *(independent of* $\mathcal{S}$*) such that* $\mathbb{E}_{\xi_i \sim D_i}\left[\left\| \frac{\nabla_s f(\mathbf{x}, \xi_i)}{\|\nabla f'(\mathbf{x}, \xi_i)\|} - \frac{\nabla_s F_i(\mathbf{x})}{\|\nabla F_i'(\mathbf{x})\|} \right\|^2\right] \leq \sigma_{n,s}^2, \quad \forall i \in [m]$. *Consequently,* $\sigma_n := \|(\sigma_{n,s})_{s \in \mathcal{S}}\|_2$ *satisfies* $\mathbb{E}_{\xi_i \sim D_i}\left[\left\| \frac{\nabla f'(\mathbf{x}, \xi_i)}{\|\nabla f'(\mathbf{x}, \xi_i)\|} - \frac{\nabla F_i'(\mathbf{x})}{\|\nabla F_i'(\mathbf{x})\|} \right\|^2\right] \leq \sigma_n^2$.

Note that Assumption 4 is a simple layer-wise variant of Assumption 3 in Qu et al. (2022). Then, we present our main results as follows. All proofs are provided in the Appendix.

Lemma 1 (model discrepancy)**.** *For any learning rate* $\eta \leq \frac{1}{3\sqrt{2}\tau\mathcal{L}}$*, we have the following result:*

$$\frac{1}{a}\sum_{i \in \mathcal{A}} \mathbb{E}\left[\left\|\mathbf{x}_{t,j+1/2}^i - \mathbf{x}_t\right\|^2\right] \leq 12\tau^2\eta^2\left(\sigma^2 + 3\zeta^2 + 3\mathbb{E}\left[\|\nabla F(\mathbf{x}_t)\|^2\right]\right) + 4\tau r^2\left(\sum_{s \in \mathcal{S}} \sigma_{n,s}^2 + 1\right)\left(36\tau^2\eta^2\mathcal{L}^2 + 1\right). \tag{7}$$

Remark 1 (Model discrepancy)**.** *Lemma 1 clearly shows the impact of layer-wise selective perturbation on the model discrepancy bound. The right-most term on the right-hand side in (7) monotonically increases as* $\mathcal{S}$ *includes more layers. That is, the local updates more strongly encourages the model to diverge within each communication round. This tendency arises because each additional layer in* $\mathcal{S}$ *contributes its own perturbation noise term* $\sigma_{n,s}^2$*, and these contributions accumulate additively inside the same* τ *round amplification factor that governs client drift.*

Theorem 2. *Under assumption 1–4, if* $\eta \leq \frac{1}{6\sqrt{2}\tau\mathcal{L}}$*,* `FedLSAM` *satisfies:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(\mathbf{x}_t)\|^2 \leq \frac{4}{\eta\tau T}\left(F(\mathbf{x}_0) - \mathbb{E}\left[F(\mathbf{x}_T)\right]\right) + \left(\eta\mathcal{M}_1 + \eta^2\mathcal{M}_2\right)\sigma^2 + 3\eta^2\mathcal{M}_2\zeta^2 + \eta^2\mathcal{M}_3\left(\sigma_n^2 + 1\right) + 8\mathcal{L}^2\tau r^2\left(\sigma_n^2 + 1\right),$$

where $\mathcal{M}_1 := \frac{4\mathcal{L}}{a}$*,* $\mathcal{M}_2 := 24\mathcal{L}^2\tau^2$*, and* $\mathcal{M}_3 := 288\mathcal{L}^4\tau^3 r^2$.

Remark 2 (Convergence guarantee)**.** *The last term on the right-hand side in (32),* $8\mathcal{L}^2\tau r^2(\sigma_n^2 + 1)$*, carries an explicit factor of* r^2*. Thus, as* $r \to 0$*,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(\mathbf{x}_t)\|^2 \leq \frac{4}{\eta\tau T}\left(F(\mathbf{x}_0) - \mathbb{E}\left[F(\mathbf{x}_T)\right]\right) + \left(\eta\mathcal{M}_1 + \eta^2\mathcal{M}_2\right)\sigma^2 + 3\eta^2\mathcal{M}_2\zeta^2,$$

which is exactly the standard non-convex convergence bound for FedAvg. As $T \to \infty$*, all the right-hand side terms vanish under the usual learning rate condition* $\eta = \mathcal{O}(1/\sqrt{\tau T})$*. Note that considering the limit* $r \to 0$ *to recover the base optimizers' guarantee is common in the convergence analysis of SAM-type methodsQu et al. (2022); Lee & Yoon (2024).*

Remark 3 (FedLSAM reduces to FedAvg as $r \to 0$)**.** *Setting* $r = 0$ *collapses* $\mathbf{x}_{t,j+1/2}^i = \mathbf{x}_{t,j}^i$*, so FedLSAM reduces to FedAvg exactly, both algorithmically and in its convergence guarantee. Consequently, the bias introduced by partial-model perturbation is governed by* r *and vanishes continuously as* $r \to 0$*, regardless of* σ_n.

Remark 4 (Benefit of partial-model perturbation)**.** *For any* $\mathcal{S} \subset [L]$*,* $\sigma_n = \|(\sigma_{n,s})_{s \in \mathcal{S}}\|$*, so that* $\sigma_n^2 = \sum_{s \in \mathcal{S}} \sigma_{n,s}^2 \leq \sum_{l=1}^{L} \sigma_{n,l}^2$*. The last term on the right-hand side in (32),* $8\mathcal{L}^2\tau r^2(\sigma_n^2 + 1)$*, does not vanish as* $T \to \infty$ *(it is even independent of learning rate), so the model converges only to a neighborhood of a stationary point rather than to it exactly. Since* σ_n^2 *is monotonically non-decreasing in* $\mathcal{S}$*, perturbing fewer layers shrinks this residual term and yields a tighter convergence neighborhood.*

Table 2: The validation accuracy comparison of the FL algorithms on different datasets. `FedLSAM` improves generalization performance across benchmarks.

Algorithm	FEMNIST (CNN)	CIFAR-10 (VGG-11)		CIFAR-100 (ResNet-18)	
Dirichlet coefficient α		0.05	0.5	0.05	0.5
FedAvg	$83.56 \pm 0.2\%$	$48.10 \pm 0.7\%$	$57.34 \pm 0.1\%$	$51.89 \pm 0.1\%$	$58.45 \pm 0.1\%$
FedSAM	$83.93 \pm 0.2\%$	$51.53 \pm 0.1\%$	$59.17 \pm 0.1\%$	$52.22 \pm 0.1\%$	$60.15 \pm 0.1\%$
MoFedSAM	$84.15 \pm 0.1\%$	$52.43 \pm 0.1\%$	$60.05 \pm 0.1\%$	$52.78 \pm 0.1\%$	$60.53 \pm 0.1\%$
FedSMOO	$84.11 \pm 0.1\%$	$54.12 \pm 0.1\%$	$60.95 \pm 0.1\%$	$53.33 \pm 0.1\%$	$63.83 \pm 0.1\%$
FedLESAM	$83.79 \pm 0.2\%$	$53.35 \pm 0.1\%$	$60.62 \pm 0.1\%$	$51.96 \pm 0.1\%$	$63.87 \pm 0.1\%$
FedGF	$84.27 \pm 0.3\%$	$54.75 \pm 0.1\%$	$61.52 \pm 0.1\%$	$52.87 \pm 0.1\%$	$64.42 \pm 0.1\%$
FedWMSAM	$84.50 \pm 0.1\%$	$55.07 \pm 0.1\%$	$61.98 \pm 0.1\%$	$53.86 \pm 0.1\%$	$64.76 \pm 0.1\%$
FedLSAM	$84.34 \pm 0.1\%$	$55.12 \pm 0.1\%$	$62.22 \pm 0.1\%$	$56.93 \pm 0.1\%$	$65.19 \pm 0.1\%$

Table 3: The validation accuracy comparison of the FL algorithms on Tiny-ImageNet with ViT-b16. The non-IID dataset is generated using a label-based Dirichlet distribution with $\alpha = 0.1$.

Algorithm	Tiny-ImageNet (ViT-B/16)
FedAvg	$73.83 \pm 0.2\%$
FedSAM	$74.32 \pm 0.1\%$
FedLESAM	$74.79 \pm 0.2\%$
FedWMSAM	$75.26 \pm 0.1\%$
FedLSAM	$75.55 \pm 0.1\%$

4 Experiments

4.1 Experimental Settings

We evaluate the performance of our proposed method on representative FL benchmarks: CIFAR-10 Krizhevsky et al. (2009) (VGG-11 Simonyan & Zisserman (2014)), CIFAR-100 (ResNet-18 He et al. (2016)) and FEMNIST Caldas et al. (2018) (CNN), and Tiny-ImageNet (ViT-B/16 Dosovitskiy (2020)). We set the total number of clients to 100 and the active ratio to 10%. To generate non-IID CIFAR datasets, we use label-based Dirichlet distributions with $\alpha \in \{0.05, 0.5\}$ which indicates highly non-IID conditions. The experimental settings are shown in Appendix, and the hyper-parameters of the other FL algorithms follow the settings reported in their corresponding papers. Experimental software [1] is developed using PyTorch 2.7.1 and all experiments are conducted on a server with 4 NVIDIA RTX 5090 GPUs.

4.2 Performance Evaluation

We compare `FedLSAM` against FedAvg and four SAM-based FL methods. FedSAM applies SAM to local training and targets to minimize local sharpness. MoFedSAM applies global momentum to FedSAM to minimize global sharpness. FedSMOO estimates global perturbation with dynamic regularization. FedLESAM locally estimates the global perturbation direction from the difference between current and previous global model. FedGF mitigates the flatness discrepancy by interpolating local and global perturbations. FedWMSAM estimates the global perturbation via server-aggregated momentum. Notably, all of these methods perturb the entire model, differing from `FedLSAM` only in how the perturbation direction is estimated, not in which parameters are perturbed.

Table 2 reports the validation accuracy of the global model. On CIFAR-100 (ResNet-18), `FedLSAM` improves the validation accuracy by 5.04% over FedAvg and by 3.07% over the strongest baseline FedWMSAM, while perturbing only the top-2 layers with the largest global update norm $\|\Delta w_t^l\|_2$. This results verify our hypothesis that the layers with a large update norm dominate the sharpness of the whole model, and applying SAM only to those few layers can effectively lead the global model to a flatter region.

[1] Source code will be released as open source once the paper is published.

Table 4: The validation accuracy comparison of different k settings. This indicates perturbing more layers did not lead to a proportional improvement in generalization performance.

Dataset	Model	k	Accuracy
CIFAR-10	VGG-11	1	$54.87 \pm 0.1\%$
		2	$55.12 \pm 0.1\%$
		3	$55.05 \pm 0.1\%$
		4	$54.87 \pm 0.1\%$
CIFAR-100	ResNet-18	1	$54.75 \pm 0.1\%$
		2	$54.97 \pm 0.1\%$
		4	$55.49 \pm 0.1\%$
		8	$56.93 \pm 0.1\%$
		16	$52.67 \pm 0.1\%$

On CIFAR-10 and FEMNIST , `FedLSAM` remains comparable to the global-sharpness baselines. Since the model has only a few layers, the gap between full and layer-wise perturbation is inherently small, and the benefit of layer selection grows with model depth, as also observed on ResNet-18.

4.3 Ablation Study

Impact of the Number of Perturbed Layers – Table 4 shows how the number of perturbed layers k affects the validation accuracy of `FedLSAM`. On CIFAR-10 with a VGG-11, the accuracy peaks at $k = 2$ and perturbing additional layers brings no further gain. On CIFAR-100 with ResNet-18, the accuracy improves consistently as k grows to 8, but expanding the perturbation to 16 layers sharply degrades the accuracy below the $k = 1$ case. This trend indicates that a small subset of layers with the largest update norm accounts for most of the global sharpness.

Moreover, the degradation at large k shows that perturbing layers with a small update norm is not merely redundant but harmful. Perturbing such low-importance layers injects noise into directions that contribute little to sharpness reduction, which slows the decrease of the training loss while still incurring the full perturbation cost. The optimal k therefore balances sharpness suppression against this noise, and it remains far smaller than the total number of layers in both architectures.

Impact of the Layer Selection Criterion – Table 5 compares the layer selection criteria of `FedLSAM` under the same k, where *None* denotes FedAvg without any perturbation. Selecting the layers with the largest update norm (*Top-*k) achieves the best accuracy on both benchmarks. In contrast, *Bottom-*k, which perturbs the layers with the smallest update norm, yields only a marginal improvement over the unperturbed baseline. *Random* selection achieves accuracy comparable to or slightly lower than *Bottom-*k, recovering only the fraction of *Top-*k's gain proportional to the probability of including high-norm layers.

These empirical results directly support our hypothesis that the update norm is a reliable proxy for the layer-wise contribution to global sharpness. Perturbing high-norm layers flattens the directions that dominate the global sharpness, whereas perturbing low-norm layers incurs additional computation with little flattening effect. The result also confirms that the effectiveness of `FedLSAM` stems from the selection criterion itself rather than from the reduced number of perturbed layers.

4.4 Visualization of Global Loss Surface

We visualize and compare the loss landscapes obtained by `FedLSAM` and other FL algorithms. The parameter space is visualized using the technique proposed in Li et al. (2018). Figure 1 presents the loss surfaces of the global models trained on CIFAR-100 using these methods. FedAvg converges to a relatively sharp region, whereas the SAM-based methods generally exhibit flatter loss surfaces. FedSAM primarily targets sharpness with respect to each client's local objective, while FedLESAM, FedGF and FedWMSAM incorporate global information into the perturbation direction. Despite perturbing only a subset of layers, `FedLSAM` leads the global model to a flatter region than other SAM-based methods. This result demonstrates that concentrating the perturbation on high update

Table 5: Comparison of validation accuracy under different layer-selection criteria. The best accuracy is achieved when the layers with the largest update norm.

Dataset	Perturbation Criterion	Accuracy
CIFAR-10	None	$48.10 \pm 0.7\%$
	Top-k	$55.12 \pm 0.1\%$
	Bottom-k	$51.25 \pm 0.1\%$
	Random	$52.31 \pm 0.4\%$
CIFAR-100	None	$51.89 \pm 0.1\%$
	Top-k	$56.93 \pm 0.1\%$
	Bottom-k	$52.92 \pm 0.2\%$
	Random	$52.99 \pm 0.1\%$

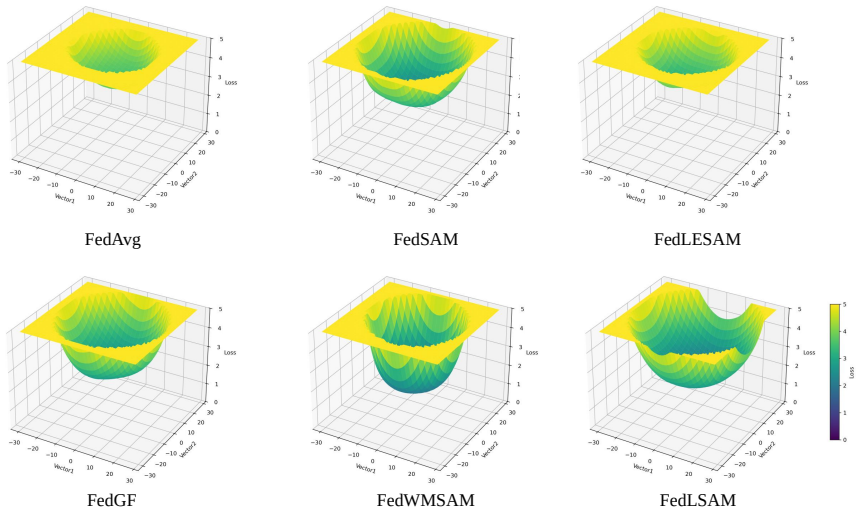

Figure 1: Loss landscape comparison on CIFAR-100 (ResNet-18). The z-axis maximum is fixed to 5. The results show that perturbing only k selected layers effectively guides the global model toward a flatter region.

norm layers is sufficient to suppress global sharpness, which is consistent with the accuracy gains reported in Table 2.

5 Conclusion

In this paper, we proposed `FedLSAM`, a FL framework that applies SAM in a layer-wise manner by perturbing only the top-k layers with the largest update norm $\|\Delta w_t^l\|_2$. Our analysis showed that when the gradient norm of the selected layers dominates that of the remaining layers, this layer-wise perturbation closely approximates the gradient norm penalizing effect of full-model SAM at a fraction of the perturbation cost. Experiments on non-IID benchmarks confirmed this claim, as `FedLSAM` outperformed FedAvg and existing SOTA SAM-based FL methods. The loss surface visualization also demonstrated that `FedLSAM` converges to a flatter minimum than FedAvg and FedSAM despite perturbing only a few layers, saving the additional computational cost. We believe our study can encourage more active utilization of SAM in FL as well as several resource-constrained learning environments.

Limitations and Future Work – `FedLSAM` currently fixes k to a given constant throughout the training. However, the minimal number of layers that maximizes the gradient norm penalizing effect may vary across communication rounds. As future work, we plan to extend our layer selection scheme to reflect layer-wise gradient variance and to adaptively adjust k during training.

AI USE STATEMENT

We used large language models (LLMs) as writing assistants to polish grammar and phrasing and to check consistency between the text, tables, and appendix. We did not use these tools to generate research ideas, derive the theoretical results, or produce experimental results. All content was verified by the authors, who take full responsibility for it.

REPRODUCIBILITY STATEMENT

The complete proofs of convergence analysis and detailed experimental settings are provided in Appendix. Source code will be released once the paper is accepted.

REFERENCES

Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *International conference on machine learning*, pp. 639–668. PMLR, 2022.

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jiawei Du, Hanshu Yan, Jiashi Feng, Joey Tianyi Zhou, Liangli Zhen, Rick Siow Mong Goh, and Vincent YF Tan. Efficient sharpness-aware minimization for improved training of neural networks. *arXiv preprint arXiv:2110.03141*, 2021.

Ziqing Fan, Shengchao Hu, Jiangchao Yao, Gang Niu, Ya Zhang, Masashi Sugiyama, and Yanfeng Wang. Locally estimated global perturbations are better than local perturbations for federated sharpness-aware minimization. *arXiv preprint arXiv:2405.18890*, 2024.

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

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

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

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical Report*, 2009.

Sunwoo Lee. Layer-wise adaptive gradient norm penalizing method for efficient and accurate deep learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1518–1529, 2024.

Taehwan Lee and Sung Whan Yoon. Rethinking the flat minima searching in federated learning. In *Forty-first International Conference on Machine Learning*, 2024.

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

Tianle Li, Yongzhi Huang, Linshan Jiang, Chang Liu, Qipeng Xie, Wenfeng Du, Lu Wang, and Kaishun Wu. Fedwmsam: Fast and flat federated learning via weighted momentum and sharpness-aware minimization. *Advances in Neural Information Processing Systems*, 38:4425–4462, 2026.

Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *Advances in neural information processing systems*, 33:2351–2363, 2020.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. Pmlr, 2017.

Maximilian Mueller, Tiffany Vlaar, David Rolnick, and Matthias Hein. Normalization layers are all that sharpness-aware minimization needs. *Advances in Neural Information Processing Systems*, 36:69228–69252, 2023.

Zhe Qu, Xingyu Li, Rui Duan, Yao Liu, Bo Tang, and Zhuo Lu. Generalized federated learning via sharpness aware minimization. In *International conference on machine learning*, pp. 18250–18280. PMLR, 2022.

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

Yan Sun, Li Shen, Shixiang Chen, Liang Ding, and Dacheng Tao. Dynamic regularized sharpness aware minimization in federated learning: Approaching global consistency and smooth landscape. In *International conference on machine learning*, pp. 32991–33013. PMLR, 2023.

Yang Zhao, Hao Zhang, and Xiuyuan Hu. Penalizing gradient norm for efficiently improving generalization in deep learning. In *International conference on machine learning*, pp. 26982–26992. PMLR, 2022.

A APPENDIX

The appendix is structured as follows:

- Section A.1 provides detailed hyperparameter settings used in our empirical study.
- Section A.2 provides detailed proofs of the proposed lemmas and theorem.

A.1 EXPERIMENTAL SETTINGS

We evaluate propsed method on FEMNIST with a CNN, CIFAR-10 with VGG-11, CIFAR-100 with ResNet-18, and Tiny-ImageNet with ViT-B/16. We consider 100 clients and sample 10 clients in each communication round. For CIFAR-10 and CIFAR-100, we simulate the data heterogeneity by sampling the label ratios from Dirichlet distribution by $\alpha \in \{0.05, 0.5\}$. For Tiny-ImageNet, we use $\alpha = 0.1$. The hyper parameters settings are shown in Table 6. For FedSAM and the other baseline methods, we follow the hyperparameter settings reported in their respective papers. The perturbation radius r is selected from $0.01, 0.1$.

Dataset	Model	Batch Size	Init. Learning Rate	Number of rounds	Number of local steps
FEMNIST	CNN	20	0.01	400	20
CIFAR-10	VGG-11	50	0.1	1000	50
CIFAR-100	ResNet-18	20	0.1	1000	25
Tiny-ImageNet	ViT-B/16	50	0.1	400	50

Table 6: The hyper-parameter settings.

A.2 CONVERGENCE ANALYSIS

We show that `FedLSAM` guarantees convergence for non-convex and smooth optimization problems:

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) := \frac{1}{m}\sum_{i=1}^{m} F_i(\mathbf{x}),$$

where m denotes the total number of clients, and $F_i(\mathbf{x}) := \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[f(\mathbf{x};\xi_i)]$ represents the local objective function associated with the data distribution $\mathcal{D}_i$ of client i. Here, $f(\mathbf{x};\xi_i)$ is the loss function evaluated on a sample ξ_i drawn from $\mathcal{D}_i$.

We first define a subset of layers $\mathcal{S} \in [L]$, where $[L] = \{1, 2, \cdots, L\}$ is the total set of layers. We then define gradient subset $\nabla f'(\cdot)$ with respect to $\mathcal{S}$ as follows.

$$\nabla f'(\mathbf{x}^i_{t,j}) \subset \nabla f(\mathbf{x}^i_{t,j}) = \{\nabla_1 f(\mathbf{x}^i_{t,j}), \nabla_2 f(\mathbf{x}^i_{t,j}), \cdots, \nabla_L f(\mathbf{x}^i_{t,j})\},$$

where $\nabla_l f(\mathbf{x}^i_{t,j})$ refers to the gradient of $f(\cdot)$ with respect to the parameters of layer l. We also define locally accumulated updates and the global update as follows.

$$\begin{aligned}
\mathbf{x}^i_{t,j+1/2} &:= \mathbf{x}^i_{t,j} + r\frac{\nabla f'(\mathbf{x}^i_{t,j})}{\|\nabla f'(\mathbf{x}^i_{t,j})\|} \\
\Delta^i_t &:= \sum_{j=0}^{\tau-1} \nabla f(\mathbf{x}^i_{t,j+1/2}) \\
\Delta_t &:= \frac{1}{a}\sum_{i\in\mathcal{A}} \Delta^i_t,
\end{aligned} \tag{8}$$

where r is the SAM's perturbation coefficient, $\mathcal{A}$ is a random subset of clients, and a is the number of active clients per communication round.

Our analysis relies on the following assumptions.

Assumption 5 (Lipschitz continuity)**.** *There exists a constant $\mathcal{L} > 0$ such that $\|\nabla F_i(\mathbf{a}) - \nabla F_i(\mathbf{b})\| \le \mathcal{L}\|\mathbf{a}-\mathbf{b}\|, \quad \forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^d, \forall i \in [m]$.*

Assumption 6 (Unbiased local gradients)**.** *The stochastic gradient is unbiased such that* $\mathbb{E}[\nabla f(\mathbf{x}, \xi_i)] = \nabla F_i(\mathbf{x}), \quad \forall i \in [m]$.

Assumption 7 (Bounded local and global variances)**.** *There exist constants* $\sigma > 0$ *and* $\zeta > 0$ *such that* $\mathbb{E}_{\xi_i \sim D_i}\left[\|\nabla f(\mathbf{x}, \xi_i) - \nabla F_i(\mathbf{x})\|^2\right] \leq \sigma^2, \quad \forall i \in [m]$, *and* $\mathbb{E}\left[\|\nabla F_i(\mathbf{x}) - \nabla F(\mathbf{x})\|^2\right] \leq \zeta^2, \quad \forall i \in [m]$.

Assumption 8 (Bounded variance of normalized gradients)**.** *For each layer* $s \in [L]$ *and every subset* $\mathcal{S} \subset [L]$ *with* $s \in \mathcal{S}$*, there exists a constant* $\sigma_{n,s} > 0$ *(independent of* $\mathcal{S}$*) such that* $\mathbb{E}_{\xi_i \sim D_i}\left[\left\| \frac{\nabla_s f(\mathbf{x}, \xi_i)}{\|\nabla f'(\mathbf{x}, \xi_i)\|} - \frac{\nabla_s F_i(\mathbf{x})}{\|\nabla F_i'(\mathbf{x})\|} \right\|^2\right] \leq \sigma_{n,s}^2, \quad \forall i \in [m]$. *Consequently,* $\sigma_n := \|(\sigma_{n,s})_{s \in \mathcal{S}}\|_2$ *satisfies* $\mathbb{E}_{\xi_i \sim D_i}\left[\left\| \frac{\nabla f'(\mathbf{x}, \xi_i)}{\|\nabla f'(\mathbf{x}, \xi_i)\|} - \frac{\nabla F_i'(\mathbf{x})}{\|\nabla F_i'(\mathbf{x})\|} \right\|^2\right] \leq \sigma_n^2$.

Note that, in the following proofs, we omit the subscript on the expectation for convenience.

Lemma 3 (framework)**.** *Under assumption* $1 \sim 3$*, if the learning rate satisfies* $\eta \leq \frac{1}{2\mathcal{L}\tau}$*,* `FedLSAM` *guarantees the following bound of the average gradient magnitude:*

$$\sum_{t=0}^{T-1} \|\nabla F(\mathbf{x}_t)\|^2 \leq \frac{2}{\eta\tau}\left(F(\mathbf{x}_0) - \mathbb{E}\left[F(\mathbf{x}_T)\right]\right) + \frac{2\mathcal{L}\eta T}{a}\sigma^2 + \frac{\mathcal{L}^2}{a\tau}\sum_{t=0}^{T-1}\sum_{i \in \mathcal{A}}\sum_{j=0}^{\tau-1}\mathbb{E}\left[\left\|\mathbf{x}_{t,j+1/2}^i - \mathbf{x}_t\right\|^2\right]. \tag{9}$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[F(\mathbf{x}_{t+1})\right] &\leq F(\mathbf{x}_t) + \langle \nabla F(\mathbf{x}_t), \mathbb{E}\left[\mathbf{x}_{t+1} - \mathbf{x}_t\right]\rangle + \frac{\mathcal{L}}{2}\mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2\right] \\
&= F(\mathbf{x}_t) + \langle \nabla F(\mathbf{x}_t), \mathbb{E}\left[-\eta\Delta_t\right]\rangle + \frac{\mathcal{L}}{2}\mathbb{E}\left[\|\eta\Delta_t\|^2\right] \\
&= F(\mathbf{x}_t) + \langle \nabla F(\mathbf{x}_t), \mathbb{E}\left[-\eta\Delta_t + \eta\tau\nabla F(\mathbf{x}_t) - \eta\tau\nabla F(\mathbf{x}_t)\right]\rangle + \frac{\mathcal{L}}{2}\mathbb{E}\left[\|\eta\Delta_t\|^2\right] \\
&= F(\mathbf{x}_t) - \eta\tau\|\nabla F(\mathbf{x}_t)\|^2 + \underbrace{\langle \nabla F(\mathbf{x}_t), \mathbb{E}\left[-\eta\Delta_t + \eta\tau\nabla F(\mathbf{x}_t)\right]\rangle}_{T_1} + \underbrace{\frac{\mathcal{L}\eta^2}{2}\mathbb{E}\left[\|\Delta_t\|^2\right]}_{T_2}
\end{aligned} \tag{10}$$

Now, we will bound T_1 and T_2, separately.

Bounding T_1**.**

$$\begin{aligned}
T_1 &= \langle \nabla F(\mathbf{x}_t), \mathbb{E}\left[-\eta\Delta_t + \eta\tau\nabla F(\mathbf{x}_t)\right]\rangle \\
&= \langle \sqrt{\eta\tau}\nabla F(\mathbf{x}_t), \mathbb{E}\left[-\frac{\sqrt{\eta}}{\sqrt{\tau}}\Delta_t + \sqrt{\eta\tau}\nabla F(\mathbf{x}_t)\right]\rangle \\
&= \langle \sqrt{\eta\tau}\nabla F(\mathbf{x}_t), \mathbb{E}\left[-\frac{\sqrt{\eta}}{\sqrt{\tau}a}\sum_{i \in \mathcal{A}}\sum_{j=0}^{\tau-1}\nabla f(\mathbf{x}_{t,j+1/2}^i) + \sqrt{\eta\tau}\nabla F(\mathbf{x}_t)\right]\rangle \\
&= \langle \sqrt{\eta\tau}\nabla F(\mathbf{x}_t), \mathbb{E}\left[-\frac{\sqrt{\eta}}{\sqrt{\tau}a}\sum_{i \in \mathcal{A}}\sum_{j=0}^{\tau-1}\nabla F_i(\mathbf{x}_{t,j+1/2}^i) + \sqrt{\eta\tau}\nabla F(\mathbf{x}_t)\right]\rangle \\
&= \langle \sqrt{\eta\tau}\nabla F(\mathbf{x}_t), \mathbb{E}\left[-\frac{\sqrt{\eta}}{\sqrt{\tau}a}\sum_{i \in \mathcal{A}}\sum_{j=0}^{\tau-1}\left(\nabla F_i(\mathbf{x}_{t,j+1/2}^i) - \nabla F_i(\mathbf{x}_t)\right)\right]\rangle.
\end{aligned}$$

Then, due to the fact that $ab = \frac{a^2}{2} + \frac{b^2}{2} - \frac{(a-b)^2}{2}$, T_1 can be bounded as follows:

$$
\begin{aligned}
T_1 &= \frac{\eta\tau}{2}\left\|\nabla F(\mathbf{x}_t)\right\|^2 + \frac{\eta}{2\tau a^2}\mathbb{E}\left[\left\|\sum_{i\in\mathcal{A}}\sum_{j=0}^{\tau-1}\left(\nabla F_i(\mathbf{x}^i_{t,j+1/2}) - \nabla F_i(\mathbf{x}_t)\right)\right\|^2\right] \\
&\quad - \frac{\eta}{2\tau a^2}\mathbb{E}\left[\left\|\sum_{i\in\mathcal{A}}\sum_{j=0}^{\tau-1}\nabla F_i(\mathbf{x}^i_{t,j+1/2})\right\|^2\right] \qquad (11) \\
&\le \frac{\eta\tau}{2}\left\|\nabla F(\mathbf{x}_t)\right\|^2 + \frac{\eta\mathcal{L}^2}{2\tau a^2}\mathbb{E}\left[\left\|\sum_{i\in\mathcal{A}}\sum_{j=0}^{\tau-1}\left(\mathbf{x}^i_{t,j+1/2} - \mathbf{x}_t\right)\right\|^2\right] \\
&\quad - \frac{\eta}{2\tau a^2}\mathbb{E}\left[\left\|\sum_{i\in\mathcal{A}}\sum_{j=0}^{\tau-1}\nabla F_i(\mathbf{x}^i_{t,j+1/2})\right\|^2\right] \qquad (12) \\
&\le \frac{\eta\tau}{2}\left\|\nabla F(\mathbf{x}_t)\right\|^2 + \frac{\eta\mathcal{L}^2}{2a}\sum_{i\in\mathcal{A}}\sum_{j=0}^{\tau-1}\mathbb{E}\left[\left\|\mathbf{x}^i_{t,j+1/2} - \mathbf{x}_t\right\|^2\right] \\
&\quad - \frac{\eta}{2\tau a^2}\mathbb{E}\left[\left\|\sum_{i\in\mathcal{A}}\sum_{j=0}^{\tau-1}\nabla F_i(\mathbf{x}^i_{t,j+1/2})\right\|^2\right], \qquad (13)
\end{aligned}
$$

where (12) holds based on Assumption 1.

Bounding T_2.

$$
\begin{aligned}
T_2 &= \frac{\mathcal{L}\eta^2}{2}\mathbb{E}\left[\left\|\Delta_t\right\|^2\right] = \frac{\mathcal{L}\eta^2}{2a^2}\mathbb{E}\left[\left\|\sum_{i\in\mathcal{A}}\sum_{j=0}^{\tau-1}\nabla f(\mathbf{x}^i_{t,j+1/2})\right\|^2\right] \\
&= \frac{\mathcal{L}\eta^2}{2a^2}\mathbb{E}\left[\left\|\sum_{i\in\mathcal{A}}\sum_{j=0}^{\tau-1}\left(\nabla f(\mathbf{x}^i_{t,j+1/2}) - \nabla F_i(\mathbf{x}^i_{t,j+1/2})\right) + \sum_{i\in\mathcal{A}}\sum_{j=0}^{\tau-1}\nabla F_i(\mathbf{x}^i_{t,j+1/2})\right\|^2\right] \\
&\le \frac{\mathcal{L}\eta^2}{a^2}\mathbb{E}\left[\left\|\sum_{i\in\mathcal{A}}\sum_{j=0}^{\tau-1}\left(\nabla f(\mathbf{x}^i_{t,j+1/2}) - \nabla F_i(\mathbf{x}^i_{t,j+1/2})\right)\right\|^2\right] \\
&\quad + \frac{\mathcal{L}\eta^2}{a^2}\mathbb{E}\left[\left\|\sum_{i\in\mathcal{A}}\sum_{j=0}^{\tau-1}\nabla F_i(\mathbf{x}^i_{t,j+1/2})\right\|^2\right] \\
&\le \frac{\mathcal{L}\eta^2\tau}{a}\sigma^2 + \frac{\mathcal{L}\eta^2}{a^2}\mathbb{E}\left[\left\|\sum_{i\in\mathcal{A}}\sum_{j=0}^{\tau-1}\nabla F_i(\mathbf{x}^i_{t,j+1/2})\right\|^2\right], \qquad (14)
\end{aligned}
$$

where (14) is based on Assumption 3.

By plugging in (13) and (14) into (10), we have

$$\mathbb{E}\left[F(\mathbf{x}_{t+1})\right] \leq F(\mathbf{x}_t) - \frac{\eta\tau}{2}\left\|\nabla F(\mathbf{x}_t)\right\|^2 + \frac{\mathcal{L}\eta^2\tau}{a}\sigma^2 + \frac{\eta\mathcal{L}^2}{2a}\sum_{i\in\mathcal{A}}\sum_{j=0}^{\tau-1}\mathbb{E}\left[\left\|\mathbf{x}_{t,j+1/2}^i - \mathbf{x}_t\right\|^2\right]$$
$$\left(\frac{\mathcal{L}\eta^2}{a^2} - \frac{\eta}{2\tau a^2}\right)\mathbb{E}\left[\left\|\sum_{i\in\mathcal{A}}\sum_{j=0}^{\tau-1}\nabla F_i(\mathbf{x}_{t,j+1/2}^i)\right\|^2\right]$$
$$\leq F(\mathbf{x}_t) - \frac{\eta\tau}{2}\left\|\nabla F(\mathbf{x}_t)\right\|^2 + \frac{\mathcal{L}\eta^2\tau}{a}\sigma^2 + \frac{\eta\mathcal{L}^2}{2a}\sum_{i\in\mathcal{A}}\sum_{j=0}^{\tau-1}\mathbb{E}\left[\left\|\mathbf{x}_{t,j+1/2}^i - \mathbf{x}_t\right\|^2\right], \quad (15)$$

where (15) holds if $\eta \leq \frac{1}{2\mathcal{L}\tau}$. Then, by rearranging the terms,

$$\left\|\nabla F(\mathbf{x}_t)\right\|^2 \leq \frac{2}{\eta\tau}\left(F(\mathbf{x}_t) - \mathbb{E}\left[F(\mathbf{x}_{t+1})\right]\right) + \frac{2\mathcal{L}\eta}{a}\sigma^2 + \frac{\mathcal{L}^2}{a\tau}\sum_{i\in\mathcal{A}}\sum_{j=0}^{\tau-1}\mathbb{E}\left[\left\|\mathbf{x}_{t,j+1/2}^i - \mathbf{x}_t\right\|^2\right]. \quad (16)$$

Finally, by summing up across T communication rounds and unrolling the telescoping sum, we have:

$$\sum_{t=0}^{T-1}\left\|\nabla F(\mathbf{x}_t)\right\|^2 \leq \frac{2}{\eta\tau}\left(F(\mathbf{x}_0) - \mathbb{E}\left[F(\mathbf{x}_T)\right]\right) + \frac{2\mathcal{L}\eta T}{a}\sigma^2 + \frac{\mathcal{L}^2}{a\tau}\sum_{t=0}^{T-1}\sum_{i\in\mathcal{A}}\sum_{j=0}^{\tau-1}\mathbb{E}\left[\left\|\mathbf{x}_{t,j+1/2}^i - \mathbf{x}_t\right\|^2\right]. \quad (17)$$

□

Lemma 4 (model discrepancy)**.** *For any learning rate $\eta \leq \frac{1}{3\sqrt{2}\tau\mathcal{L}}$, we have the following result:*

$$\frac{1}{a}\sum_{i\in\mathcal{A}}\mathbb{E}\left[\left\|\mathbf{x}_{t,j+1/2}^i - \mathbf{x}_t\right\|^2\right] \leq 12\tau^2\eta^2\left(\sigma^2 + 3\zeta^2 + 3\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t)\right\|^2\right]\right) + 4\tau r^2\left(\sum_{s\in\mathcal{S}}\sigma_{n,s}^2 + 1\right)\left(36\tau^2\eta^2\mathcal{L}^2 + 1\right).$$

Proof.

$$\mathbb{E}\left[\left\|\mathbf{x}_{t,j+1}^i - \mathbf{x}_t\right\|^2\right] = \mathbb{E}\left[\left\|\mathbf{x}_{t,j}^i - \mathbf{x}_t - \eta\nabla f(\mathbf{x}_{t,j+1/2}^i)\right\|^2\right] \leq \left(1 + \frac{1}{2\tau - 1}\right)\mathbb{E}\left[\left\|\mathbf{x}_{t,j}^i - \mathbf{x}_t\right\|^2\right] + 2\tau\eta^2\underbrace{\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_{t,j+1/2}^i)\right\|^2\right]}_{B} \tag{18}$$

The B term can be bounded as follows:

$$\begin{aligned}
B &= \mathbb{E}\left[\left\|\nabla f(\mathbf{x}_{t,j+1/2}^i)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\nabla f(\mathbf{x}_{t,j+1/2}^i) - \nabla F_i(\mathbf{x}_{t,j+1/2}^i) + \nabla F_i(\mathbf{x}_{t,j+1/2}^i)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\nabla f(\mathbf{x}_{t,j+1/2}^i) - \nabla F_i(\mathbf{x}_{t,j+1/2}^i)\right\|^2\right] + \mathbb{E}\left[\left\|\nabla F_i(\mathbf{x}_{t,j+1/2}^i)\right\|^2\right] \\
&\quad + 2\mathbb{E}\left[\langle\nabla f(\mathbf{x}_{t,j+1/2}^i) - \nabla F_i(\mathbf{x}_{t,j+1/2}^i), \nabla F_i(\mathbf{x}_{t,j+1/2}^i)\rangle\right] \\
&= \mathbb{E}\left[\left\|\nabla f(\mathbf{x}_{t,j+1/2}^i) - \nabla F_i(\mathbf{x}_{t,j+1/2}^i)\right\|^2\right] + \mathbb{E}\left[\left\|\nabla F_i(\mathbf{x}_{t,j+1/2}^i)\right\|^2\right], \qquad (19)
\end{aligned}$$

where (19) holds since $\mathbb{E}[\nabla f(\mathbf{x}_{t,j+1/2}^i)] = \nabla F_i(\mathbf{x}_{t,j+1/2}^i)$.

We can further bound the right-hand side of (19) based on Assumption 2 as follows:

$$\begin{aligned}
B &\leq \sigma^2 + \mathbb{E}\left[\left\|\nabla F_i(\mathbf{x}_{t,j+1/2}^i)\right\|^2\right] \\
&= \sigma^2 + \mathbb{E}\left[\left\|\nabla F_i(\mathbf{x}_{t,j+1/2}^i) - \nabla F_i(\mathbf{x}_t) + \nabla F_i(\mathbf{x}_t) - \nabla F(\mathbf{x}_t) + \nabla F(\mathbf{x}_t)\right\|^2\right] \\
&\leq \sigma^2 + 3\mathbb{E}\left[\left\|\nabla F_i(\mathbf{x}_{t,j+1/2}^i) - \nabla F_i(\mathbf{x}_t)\right\|^2\right] \\
&\quad + 3\mathbb{E}\left[\left\|\nabla F_i(\mathbf{x}_t) - \nabla F(\mathbf{x}_t)\right\|^2\right] + 3\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t)\right\|^2\right] \qquad (20) \\
&\leq \sigma^2 + 3\mathcal{L}^2\mathbb{E}\left[\left\|\mathbf{x}_{t,j+1/2}^i - \mathbf{x}_t\right\|^2\right] + 3\zeta^2 + 3\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t)\right\|^2\right], \qquad (21)
\end{aligned}$$

where (20) holds based on the convexity of ℓ_2 norm and Jensen's inequality.

By plugging in (21) into (18) and rearranging the terms, we have:

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{x}_{t,j+1}^i - \mathbf{x}_t\right\|^2\right] &\leq \left(1 + \frac{1}{2\tau - 1}\right)\mathbb{E}\left[\left\|\mathbf{x}_{t,j}^i - \mathbf{x}_t\right\|^2\right] \\
&\quad + 2\tau\eta^2\sigma^2 + 6\tau\eta^2\zeta^2 + 6\tau\eta^2\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t)\right\|^2\right] \\
&\quad + 6\tau\eta^2\mathcal{L}^2\underbrace{\mathbb{E}\left[\left\|\mathbf{x}_{t,j+1/2}^i - \mathbf{x}_t\right\|^2\right]}_{C}. \qquad (22)
\end{aligned}$$

Bounding C – Now we focus on bounding C term as follows.

$$\mathbb{E}\left[\left\|\mathbf{x}_{t,j+1/2}^i - \mathbf{x}_t\right\|^2\right] = \mathbb{E}\left[\left\|\mathbf{x}_{t,j}^i - \mathbf{x}_t + r\frac{\nabla f'(\mathbf{x}_{t,j}^i)}{\|\nabla f'(\mathbf{x}_{t,j}^i)\|}\right\|^2\right]$$
$$\leq \left(1 + \frac{1}{2\tau - 1}\right)\mathbb{E}\left[\left\|\mathbf{x}_{t,j}^i - \mathbf{x}_t\right\|^2\right] + 2\tau r^2 \underbrace{\mathbb{E}\left[\left\|\frac{\nabla f'(\mathbf{x}_{t,j}^i)}{\|\nabla f'(\mathbf{x}_{t,j}^i)\|}\right\|^2\right]}_{A} \tag{23}$$

The A term can be bounded as follows:

$$A = \mathbb{E}\left[\left\|\frac{\nabla f'(\mathbf{x}_{t,j}^i)}{\|\nabla f'(\mathbf{x}_{t,j}^i)\|}\right\|^2\right]$$
$$= \mathbb{E}\left[\left\|\frac{\nabla f'(\mathbf{x}_{t,j}^i)}{\|\nabla f'(\mathbf{x}_{t,j}^i)\|} - \frac{\nabla F_i'(\mathbf{x}_{t,j}^i)}{\|\nabla F_i'(\mathbf{x}_{t,j}^i)\|} + \frac{\nabla F_i'(\mathbf{x}_{t,j}^i)}{\|\nabla F_i'(\mathbf{x}_{t,j}^i)\|}\right\|^2\right]$$
$$\leq 2\mathbb{E}\left[\left\|\frac{\nabla f'(\mathbf{x}_{t,j}^i)}{\|\nabla f'(\mathbf{x}_{t,j}^i)\|} - \frac{\nabla F_i'(\mathbf{x}_{t,j}^i)}{\|\nabla F_i'(\mathbf{x}_{t,j}^i)\|}\right\|^2\right] + 2\mathbb{E}\left[\left\|\frac{\nabla F_i'(\mathbf{x}_{t,j}^i)}{\|\nabla F_i'(\mathbf{x}_{t,j}^i)\|}\right\|^2\right]$$
$$\leq 2\sigma_n^2 + 2, \tag{24}$$

where (24) holds based on Assumption 4.

By plugging in (24) into (23) and rearranging the terms, we have:

$$\mathbb{E}\left[\left\|\mathbf{x}_{t,j+1/2}^i - \mathbf{x}_t\right\|^2\right] \leq \left(1 + \frac{1}{2\tau - 1}\right)\mathbb{E}\left[\left\|\mathbf{x}_{t,j}^i - \mathbf{x}_t\right\|^2\right] + 4\tau r^2\sigma_n^2 + 4\tau r^2. \tag{25}$$

Combining sub-results – Now, we plug-in (25) into (22) and have:

$$\mathbb{E}\left[\left\|\mathbf{x}_{t,j+1}^i - \mathbf{x}_t\right\|^2\right] \leq \left(1 + \frac{1}{2\tau - 1}\right)\mathbb{E}\left[\left\|\mathbf{x}_{t,j}^i - \mathbf{x}_t\right\|^2\right]$$
$$+ 2\tau\eta^2\sigma^2 + 6\tau\eta^2\zeta^2 + 6\tau\eta^2\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t)\right\|^2\right]$$
$$+ 6\tau\eta^2\mathcal{L}^2\left(\left(1 + \frac{1}{2\tau - 1}\right)\mathbb{E}\left[\left\|\mathbf{x}_{t,j}^i - \mathbf{x}_t\right\|^2\right] + 4\tau r^2\sigma_n^2 + 4\tau r^2\right).$$

By re-arranging the terms, we have:

$$\mathbb{E}\left[\left\|\mathbf{x}_{t,j+1}^i - \mathbf{x}_t\right\|^2\right] \leq \underbrace{\left(6\tau\eta^2\mathcal{L}^2\left(1 + \frac{1}{2\tau - 1}\right) + 1 + \frac{1}{2\tau - 1}\right)}_{D}\mathbb{E}\left[\left\|\mathbf{x}_{t,j}^i - \mathbf{x}_t\right\|^2\right] \tag{26}$$
$$+ 2\tau\eta^2\left(\sigma^2 + 3\zeta^2 + 12\tau r^2\mathcal{L}^2\left(\sigma_n^2 + 1\right)\right) + 6\tau\eta^2\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t)\right\|^2\right].$$

Simplifying D – If $\eta \leq \frac{1}{3\sqrt{2}\tau\mathcal{L}}$, it can be further simplified as follows:

$$18\tau\eta^2\mathcal{L}^2 + 1 + \frac{1}{2\tau - 1} \leq \frac{1}{\tau} + 1 + \frac{1}{2\tau - 1} \leq 1 + \frac{1}{\tau - 1}. \tag{27}$$

Therefore, (26) can be simplified as follows:

$$\mathbb{E}\left[\left\|\mathbf{x}_{t,j+1}^i - \mathbf{x}_t\right\|^2\right] \leq \left(1 + \frac{1}{\tau - 1}\right)\mathbb{E}\left[\left\|\mathbf{x}_{t,j}^i - \mathbf{x}_t\right\|^2\right]$$
$$+ 2\tau\eta^2\left(\sigma^2 + 3\zeta^2 + 12\tau r^2\mathcal{L}^2\left(\sigma_n^2 + 1\right)\right) + 6\tau\eta^2\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t)\right\|^2\right].$$

Then, by unrolling the recursion until j goes to 0, we have:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{x}_{t,j+1}^i - \mathbf{x}_t\right\|^2\right] \\
&\quad\leq \sum_{k=0}^{j}\left(1+\frac{1}{\tau-1}\right)^k\left(2\tau\eta^2\left(\sigma^2+3\zeta^2+12\tau r^2\mathcal{L}^2\left(\sigma_n^2+1\right)\right)+6\tau\eta^2\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t)\right\|^2\right]\right) \\
&\quad\leq (\tau-1)\left(\left(1+\frac{1}{\tau-1}\right)^{j+1}-1\right)\left(2\tau\eta^2\left(\sigma^2+3\zeta^2+12\tau r^2\mathcal{L}^2\left(\sigma_n^2+1\right)\right)+6\tau\eta^2\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t)\right\|^2\right]\right) \\
&\quad\leq (\tau-1)\left(\left(1+\frac{1}{\tau-1}\right)^{\tau}-1\right)\left(2\tau\eta^2\left(\sigma^2+3\zeta^2+12\tau r^2\mathcal{L}^2\left(\sigma_n^2+1\right)\right)+6\tau\eta^2\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t)\right\|^2\right]\right).
\end{aligned} \tag{28}$$

Here, the $(\tau-1)\left((1+\frac{1}{\tau-1})^{\tau}-1\right)$ can be simplified as follows:

$$\begin{aligned}
(\tau-1)\left(\left(1+\frac{1}{\tau-1}\right)^{\tau}-1\right) &= (\tau-1)\left(\left(1+\frac{1}{\tau-1}\right)^{\tau-1}\cdot\left(1+\frac{1}{\tau-1}\right)-1\right) \\
&\leq (\tau-1)\left(e\left(1+\frac{1}{\tau-1}\right)-1\right) \qquad (29)\\
&= \tau(e-1)+1 \\
&\leq 2\tau,
\end{aligned}$$

where (29) holds because $(1+1/\beta)^{\beta}\leq e$. Thus, (28) can be bounded as follows:

$$\mathbb{E}\left[\left\|\mathbf{x}_{t,j+1}^i - \mathbf{x}_t\right\|^2\right] \leq \left(4\tau^2\eta^2\left(\sigma^2+3\zeta^2+12\tau r^2\mathcal{L}^2\left(\sigma_n^2+1\right)\right)+12\tau^2\eta^2\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t)\right\|^2\right]\right). \tag{30}$$

By plugging in (30) into (25), we have:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{x}_{t,j+1/2}^i - \mathbf{x}_t\right\|^2\right] \\
&\quad\leq \left(1+\frac{1}{2\tau-1}\right)\left(4\tau^2\eta^2\left(\sigma^2+3\zeta^2+12\tau r^2\mathcal{L}^2\left(\sigma_n^2+1\right)\right)+12\tau^2\eta^2\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t)\right\|^2\right]\right) \\
&\qquad + 4\tau r^2\sigma_n^2+4\tau r^2 \\
&\quad\leq 12\tau^2\eta^2\left(\sigma^2+3\zeta^2+12\tau r^2\mathcal{L}^2\left(\sigma_n^2+1\right)\right)+36\tau^2\eta^2\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t)\right\|^2\right]+4\tau r^2\sigma_n^2+4\tau r^2 \\
&\quad= 12\tau^2\eta^2\left(\sigma^2+3\zeta^2+3\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t)\right\|^2\right]\right)+4\tau r^2\left(\sigma_n^2+1\right)\left(36\tau^2\eta^2\mathcal{L}^2+1\right).
\end{aligned}$$

Finally, because the right-hand side is independent of a, we have:

$$\begin{aligned}
\frac{1}{a}\sum_{i\in\mathcal{A}}\mathbb{E}\left[\left\|\mathbf{x}_{t,j+1/2}^i - \mathbf{x}_t\right\|^2\right] &\leq 12\tau^2\eta^2\left(\sigma^2+3\zeta^2+3\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t)\right\|^2\right]\right) \\
&\quad + 4\tau r^2\left(\sigma_n^2+1\right)\left(36\tau^2\eta^2\mathcal{L}^2+1\right).
\end{aligned}$$

□

Theorem 5. *Under assumption 1–4, if* $\eta \leq \frac{1}{6\sqrt{2}\tau\mathcal{L}}$, `FedLSAM` *satisfies:*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla F(\mathbf{x}_t)\right\|^2 \leq \frac{4}{\eta\tau T}\left(F(\mathbf{x}_0)-\mathbb{E}\left[F(\mathbf{x}_T)\right]\right)\\ +\left(\eta\mathcal{M}_1+\eta^2\mathcal{M}_2\right)\sigma^2+3\eta^2\mathcal{M}_2\zeta^2+\eta^2\mathcal{M}_3\left(\sigma_n^2+1\right)\\ +8\mathcal{L}^2\tau r^2\left(\sigma_n^2+1\right),\end{aligned}$$

where $\mathcal{M}_1 := \frac{4\mathcal{L}}{a}$, $\mathcal{M}_2 := 24\mathcal{L}^2\tau^2$, *and* $\mathcal{M}_3 := 288\mathcal{L}^4\tau^3r^2$.

Proof. Based on lemma 3 and 1, we begin with:

$$\begin{aligned}&\sum_{t=0}^{T-1}\left\|\nabla F(\mathbf{x}_t)\right\|^2\\ &\leq \frac{2}{\eta\tau}\left(F(\mathbf{x}_0)-\mathbb{E}\left[F(\mathbf{x}_T)\right]\right)+\frac{2\mathcal{L}\eta T}{a}\sigma^2+\frac{\mathcal{L}^2}{a\tau}\sum_{t=0}^{T-1}\sum_{i\in\mathcal{A}}\sum_{j=0}^{\tau-1}\mathbb{E}\left[\left\|\mathbf{x}_{t,j+1/2}^i-\mathbf{x}_t\right\|^2\right]\\ &\leq \frac{2}{\eta\tau}\left(F(\mathbf{x}_0)-\mathbb{E}\left[F(\mathbf{x}_T)\right]\right)+\frac{2\mathcal{L}\eta T}{a}\sigma^2\\ &\quad+\frac{\mathcal{L}^2}{\tau}\sum_{t=0}^{T-1}\sum_{j=0}^{\tau-1}\left(12\tau^2\eta^2\left(\sigma^2+3\zeta^2+3\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t)\right\|^2\right]\right)+4\tau r^2\left(\sigma_n^2+1\right)\left(36\tau^2\eta^2\mathcal{L}^2+1\right)\right)\\ &= \frac{2}{\eta\tau}\left(F(\mathbf{x}_0)-\mathbb{E}\left[F(\mathbf{x}_T)\right]\right)+\frac{2\mathcal{L}\eta T}{a}\sigma^2\\ &\quad+T\mathcal{L}^2\left(12\tau^2\eta^2\left(\sigma^2+3\zeta^2\right)+4\tau r^2\left(\sigma_n^2+1\right)\left(36\tau^2\eta^2\mathcal{L}^2+1\right)\right)\\ &\quad+\underbrace{36\tau^2\eta^2\mathcal{L}^2}_{E}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t)\right\|^2\right].\end{aligned}\tag{31}$$

If $\eta \leq \frac{1}{6\sqrt{2}\tau\mathcal{L}}$, $E \leq \frac{1}{2}$. Therefore, (31) is simplified as follows:

$$\begin{aligned}\sum_{t=0}^{T-1}\left\|\nabla F(\mathbf{x}_t)\right\|^2 \leq \frac{2}{\eta\tau}\left(F(\mathbf{x}_0)-\mathbb{E}\left[F(\mathbf{x}_T)\right]\right)+\frac{2\mathcal{L}\eta T}{a}\sigma^2\\ +T\mathcal{L}^2\left(12\tau^2\eta^2\left(\sigma^2+3\zeta^2\right)+4\tau r^2\left(\sigma_n^2+1\right)\left(36\tau^2\eta^2\mathcal{L}^2+1\right)\right)\\ +\frac{1}{2}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t)\right\|^2\right].\end{aligned}$$

Finally, by rearranging the terms and dividing both sides by T, we have:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla F(\mathbf{x}_t)\right\|^2 &\leq \frac{4}{\eta\tau T}\left(F(\mathbf{x}_0)-\mathbb{E}\left[F(\mathbf{x}_T)\right]\right)+\frac{4\mathcal{L}\eta}{a}\sigma^2\\ &\quad+2\mathcal{L}^2\left(12\tau^2\eta^2\left(\sigma^2+3\zeta^2\right)+4\tau r^2\left(\sigma_n^2+1\right)\left(36\tau^2\eta^2\mathcal{L}^2+1\right)\right)\\ &= \frac{4}{\eta\tau T}\left(F(\mathbf{x}_0)-\mathbb{E}\left[F(\mathbf{x}_T)\right]\right)\\ &\quad+\left(\eta\mathcal{M}_1+\eta^2\mathcal{M}_2\right)\sigma^2+3\eta^2\mathcal{M}_2\zeta^2+\eta^2\mathcal{M}_3\left(\sigma_n^2+1\right)\\ &\quad+8\mathcal{L}^2\tau r^2\left(\sigma_n^2+1\right),\end{aligned}\tag{32}$$

where $\mathcal{M}_1 := \frac{4\mathcal{L}}{a}$, $\mathcal{M}_2 := 24\mathcal{L}^2\tau^2$, and $\mathcal{M}_3 := 288\mathcal{L}^4\tau^3r^2$. □