Focus on Where You Aggregate: Restricted SAM for Non-IID Federated Learning

Haotong Wen[1*] **Junkang Liu**[2] **Yi Xu**[1†] **Xiao Liu**[3] **Longkun Guo**[4] **Kewen Liao**[3†]
[1]School of Computer Science and Technology, Anhui University
[2]College of Intelligence and Computing, Tianjin University
[3]School of Information Technology, Deakin University
[4]School of Mathematics and Statistics, Fuzhou University

Abstract

Federated learning (FL) under heterogeneous data distributions suffers from client drift, particularly with multiple local updates, which undermines the stability of global model aggregation. Existing sharpness-aware minimization (SAM)-based FL methods mitigate this issue by promoting flatter local loss landscapes. However, their spherical perturbation neighborhoods do not explicitly account for the geometry of global parameter aggregation, limiting the benefits of local flattening for the aggregated model. To address this limitation, we propose Federated Restricted SAM (FedRSAM), which promotes aggregation-aware flatness by restricting SAM perturbations to an anticipated aggregation region. Specifically, FedRSAM constructs a cone-shaped region using the previous global update direction and the variance of client updates to capture where subsequent global aggregation is likely to occur. By restricting perturbations to this region, FedRSAM focuses local sharpness-aware optimization on directions most relevant to global aggregation. Theoretically, we establish tighter aggregation-sensitivity error bounds while retaining the convergence guarantee. Extensive experiments across vision and language benchmarks demonstrate the stability and superior performance of FedRSAM under heterogeneous data distributions. The code is available at `https://github.com/Doctorwngd/FedRSAM-official`.

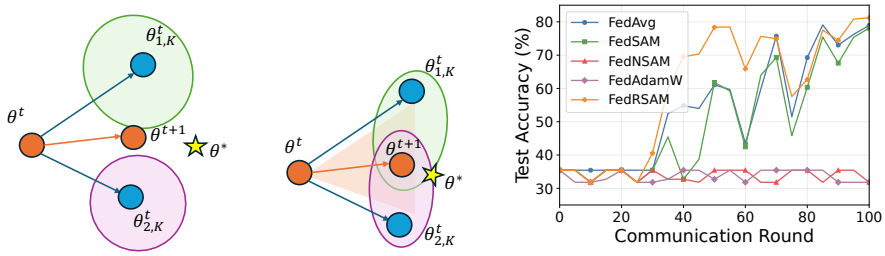
1 Introduction

Federated learning (FL) enables distributed clients to collaboratively train a global model without sharing raw data, thereby preserving privacy and reducing communication costs McMahan et al. [2017], Kairouz et al. [2021]. In each communication round, clients perform multiple steps of local optimization and transmit model updates to a central server for aggregation. However, in practical deployments, client data are often highly heterogeneous (non-IID) Qi et al. [2026], and performing multiple local updates further amplifies the divergence among client models Li et al. [2020b], Karimireddy et al. [2020]. Consequently, local update directions may deviate significantly from the true global descent direction, resulting in client drift, unstable aggregation, and degraded convergence performance Li et al. [2020a].

To mitigate these challenges, recent works Qu et al. [2022], Sun et al. [2023], Liu et al. [2025b] introduce loss landscape flattening into federated optimization. Sharpness-Aware Minimization (SAM) Foret et al. [2021] promotes generalization by minimizing the worst-case loss within a neighborhood of model parameters, thereby favoring flat minima. Building on this idea, FedSAM Qu et al. [2022] applies SAM at the client level to encourage locally flat solutions, reducing sensitivity to

*First author: Haotong Wen (E24101014@stu.ahu.edu.cn).
†Corresponding authors: Kewen Liao (kewen.liao@deakin.edu.au) and Yi Xu (xuyi1023@126.com).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) FedSAM (b) FedRSAM (c) momentum vs momentum-free

Figure 1: (a)(b) Blue circles: local/client model parameters after K gradient update steps at round t; orange circles: global/server parameters; yellow stars: ideal optimum. Green and purple regions denote local loss landscapes, and the orange triangular region denotes the aggregation-relevant region considered in FedRSAM. (c) MNLI dataset performance with RoBERTa-Base + LoRA. FedNSAM and FedAdamW use momentum $\gamma = 0.5$, while FedRSAM uses momentum-free client updates.

data heterogeneity. Subsequent approaches extend this paradigm by incorporating global information to better approximate flat regions shared across clients, such as FedSMOO and FedGloSS Sun et al. [2023], Caldarola et al. [2025]. Other methods, including FedNSAM and FedGAMMA Liu et al. [2025b], Dai et al. [2024], further improve local-global consistency through techniques such as momentum correction and variance-reduction-based control variables.

However, existing SAM-based FL methods still suffer from a key geometric mismatch. FedSAM flattens the local neighborhood of each client, but the server-aggregated parameter usually lies in a region determined by multiple client update directions Liu et al. [2025b], Caldarola et al. [2025]. We call this region the *aggregation-relevant region*. Under strong non-IID data, client-side local minima can be far apart, so locally flattened neighborhoods may not cover this region. As shown in Fig. 1(a), FedSAM may obtain large flat regions around clients, while the aggregated parameter still falls into a high-loss or sharp area between them. FedNSAM partially mitigates this mismatch by using momentum to provide a historical global trend. Yet under extreme heterogeneity, momentum may accumulate biased update directions and push the model toward incorrect or sharp regions. As shown in Fig. 1(c), this can cause unstable convergence and even worse performance than momentum-free methods such as FedSAM and FedAvg McMahan et al. [2017].

Building on the region perspective, we propose a client-side *momentum-free Restricted SAM* (**FedRSAM**) method. Instead of enlarging each client's local flat neighborhood, Restricted SAM directly flattens the region where the aggregated parameter is likely to move. Specifically, we construct an *aggregation-relevant cone* using the client update directions from the previous communication round to approximate the aggregation region of the current round. We then restrict the SAM's worst-case perturbation search to this cone, encouraging clients to optimize the loss landscape along aggregation-relevant directions rather than arbitrary local directions. As a result, FedRSAM promotes flatter connections among client-side low-loss regions, thereby increasing the likelihood that the aggregated global parameter remains within a shared low-loss flat region. As illustrated in Fig. 1(b), FedRSAM may induce a smaller flat region than FedSAM; however, this region is better aligned with the server-side aggregation. Consequently, by avoiding client-side momentum accumulation, FedRSAM mitigates the risk of incorporating biased historical directions under severe data heterogeneity. Moreover, by constraining the perturbation search within the aggregation-relevant region, FedRSAM explicitly reduces the mismatch between local sharpness minimization and global parameter aggregation. Empirical evaluations on both vision and language benchmarks demonstrate that FedRSAM consistently outperforms FedAvg, FedSAM, and a range of recent momentum-based and global-flattening federated optimization methods. On the theoretical side, we establish a tighter aggregation error bound for FedRSAM while retaining the same-order convergence guarantee as FedAvg and FedSAM under standard assumptions Li et al. [2020b], Qu et al. [2022]. The main contributions of this paper are summarized as follows:

- We revisit the client-drift challenge and Sharpness-Aware Minimization (SAM) in federated learning through a geometric lens. We conjecture that, under non-IID heterogeneity, locally flattened client neighborhoods may not align with the region traversed by the aggregated global model, resulting in ineffective local sharpness minimization.

- We propose FedRSAM, a region-constrained sharpness-aware optimizer for federated learning. It constructs an aggregation-relevant cone from the previous global update direction and the variance of client updates, and restricts the SAM perturbation search within this cone.
- We provide empirical and theoretical validation for FedRSAM. Extensive experiments on vision and language benchmarks demonstrate its superior performance and training stability. Theoretically, we bound the aggregation-sensitivity error and show that the restricted cone yields a tighter expected bound than the SAM sphere, while FedRSAM retains the same-order convergence rate as FedSAM.

2 Related Work

Heterogeneity Issues in Federated Learning. Data heterogeneity and multiple local update steps can cause severe client drift in federated learning, where local models gradually deviate from the global optimization direction and destabilize server aggregation. Existing approaches mainly address this issue from the perspective of optimization consistency. For instance, FedAvgM Hsu et al. [2019] stabilizes global updates via server-side momentum; FedProx Li et al. [2020a] introduces a proximal regularization term to constrain local deviation from the global model; SCAFFOLD Karimireddy et al. [2020] employs control variates to correct client drift; and FedACG Kim et al. [2024] improves inter-client consistency through prospective gradient alignment. More recently, FedAdamW Liu et al. [2026] investigates the variance of adaptive optimizer states and local overfitting under non-IID settings, and improves training stability through global correction and moment aggregation. These methods primarily focus on improving the consistency between local updates and global optimization directions, but they do not explicitly consider the geometry of loss landscape.

Generalization and Flatness-Aware Federated Learning. Recently, loss landscape flatness has been widely explored to improve generalization in federated learning. FedSAM Qu et al. [2022] introduces Sharpness-Aware Minimization (SAM) Foret et al. [2021] into local training by searching for the worst-case perturbation within each client's local neighborhood, aiming to improve local flatness and benefit global performance. FedGAMMA Dai et al. [2024] further combines SAM with control variates to reduce the effect of data heterogeneity. FedLESAM Fan et al. [2024] estimates the direction of global perturbation on the client side, and FedSMOO Sun et al. [2023] jointly considers global consistency and global flatness through dynamic regularization and perturbation alignment. FedNSAM Liu et al. [2025b] further points out that local flatness does not necessarily imply global flatness under heterogeneous settings, and improves the consistency between local perturbation and global aggregation through Nesterov momentum correction.

In contrast to prior approaches, FedRSAM explicitly considers the geometry of the loss landscape and does not enforce uniform flatness over the entire local neighborhood. Instead, it restricts perturbation search to a predicted aggregation-relevant region, thereby encouraging robustness to loss variations along the future aggregation trajectory. As a result, FedRSAM directly aligns local sharpness-aware optimization with the geometry of global parameter aggregation.

3 Revisiting SAM in FL

FL Problem Setup Federated learning (FL) aims to optimize model parameters with local clients, i.e., minimizing the following population risk:

$$F(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(F_i(\theta) := \mathbb{E}_{\zeta_i\sim\mathbb{D}_i}\left[F_i(\theta;\zeta_i)\right]\right). \tag{1}$$

The function F_i denotes the loss on client i and model θ, where $\zeta_i \sim \mathbb{D}_i$ is a sample from its local data distribution. The expectation $\mathbb{E}_{\zeta_i}[\cdot]$ is taken over ζ_i. The total number of clients is N.

Sharpness-Aware Minimization Sharpness-Aware Minimization (SAM) aims to jointly minimize the loss function and seek a flat neighborhood around the solution by solving the following problem:

$$\min_{\theta}\left\{f(\tilde{\theta}) \triangleq \max_{\|\delta\|_2\le\rho} f(\theta+\delta)\right\}, \tag{2}$$

where $\|\cdot\|_2$ denotes the ℓ_2-norm and ρ is the radius of the neighborhood. Instead of solving the above min-max problem exactly, SAM adopts an efficient first-order approximation in practice. By the first-order Taylor expansion of f at θ, the inner maximization can be approximated as

$$\delta^*(\theta) \approx \arg\max_{\|\delta\|_2 \le \rho} \left\{ f(\theta) + \delta^\top \nabla f(\theta) \right\} = \rho \cdot \frac{\nabla f(\theta)}{\|\nabla f(\theta)\|_2}. \tag{3}$$

Substituting the perturbation $\delta^*(\theta)$ back into the objective, the SAM gradient can be approximated by

$$\nabla f(\tilde{\theta}) = \nabla f\big(\theta + \delta^*(\theta)\big) \approx \nabla f\left(\theta + \rho \cdot \frac{\nabla f(\theta)}{\|\nabla f(\theta)\|_2}\right). \tag{4}$$

Therefore, SAM can be implemented by a two-step procedure in each iteration: $\tilde{\theta}_k = \theta_k + \rho \cdot \frac{g_k}{\|g_k\|_2}, g_k = \nabla f(\theta_k)$, and $\theta_{k+1} = \theta_k - \eta \nabla f(\tilde{\theta}_k)$, where η is the learning rate and k is a local gradient update step. In other words, SAM first performs a gradient ascent step to reach the worst-case point within the ρ-ball around θ_k, and then updates the parameter using the gradient evaluated at the perturbed point. This procedure encourages the optimizer to seek solutions located in flatter loss regions, which often leads to better generalization.

4 Proposed Algorithm: Restricted SAM in FL

In this section, we first introduce **Restricted SAM**, an aggregation-aware sharpness-aware local optimizer for federated learning. Unlike existing SAM-based federated methods that flatten a broad local neighborhood around each client model, Restricted SAM is designed to flatten the *aggregation-relevant region*, namely, the region where the next aggregated global parameter is likely to lie.

4.1 Method

The global objective of Restricted SAM remains the same as that of FedAvg:

$$\min_{\theta} \left\{ F(\theta) = \frac{1}{N} \sum_{i \in [N]} F_i(\theta) \right\}, \qquad F_i(\theta) \triangleq \mathbb{E}_{\zeta_i \sim \mathbb{D}_i} F_i(\theta; \zeta_i). \tag{5}$$

The key difference lies in the local sharpness-aware objective. Instead of searching for the worst-case perturbation over the entire ρ-ball, Restricted SAM constrains the perturbation inside an aggregation-relevant region $\mathcal{D}_t$. Specifically, at local step k for client i in communication round t, the local objective is defined as

$$F_i^{\mathrm{RSAM}}(\theta_{i,k}^t) \triangleq \max_{\delta \in \mathcal{D}_t,\ \|\delta\|_2 \le \rho} F_i(\theta_{i,k}^t + \delta). \tag{6}$$

Therefore, Restricted SAM does not flatten the entire local neighborhood uniformly. Instead, it only flattens the subset of perturbations that are consistent with the predicted aggregation geometry, so that the local sharpness-aware optimization is better aligned with the future global aggregation.

Region Parameterization. We parameterize the restricted perturbation region $\mathcal{D}_t$ by taking the predicted aggregation direction as the central axis of a cone-like region and modeling the residual uncertainty around this axis. Let $d_t = \theta^t - \theta^{t-1}$ and $\hat{d}_t = \frac{d_t}{\max(\|d_t\|_2, \varepsilon)}$, where $\hat{d}_t$ denotes the stabilized principal aggregation direction at round t. The residual map is $P_{\perp,t} = I - \hat{d}_t \hat{d}_t^\top$, which is an orthogonal projector when $\|d_t\|_2 \ge \varepsilon$. Based on this decomposition, any restricted perturbation is parameterized as

$$\delta(\alpha, z) = \alpha^2 \hat{d}_t + P_{\perp,t} z, \tag{7}$$

where α^2 is a nonnegative reparameterization of the displacement along the principal aggregation direction and z captures the residual deviation. This parameterization restricts the principal component of the perturbation to the forward aggregation direction.

Accordingly, the aggregation-relevant region is defined as

$$\mathcal{D}_t = \left\{ \delta(\alpha, z) \;\middle|\; \delta(\alpha, z) = \alpha^2 \hat{d}_t + P_{\perp,t} z,\ -\sqrt{v_t} \le z \le \sqrt{v_t} \right\}, \tag{8}$$

where the inequalities are understood element-wise. Here,

$$\Sigma_{t-1} = \frac{1}{\max\{m-1,1\}} \sum_{i=1}^{m} (\bar{\theta}^{t-1} - \theta_i^{t-1})(\bar{\theta}^{t-1} - \theta_i^{t-1})^\top, \quad v_t = \mathrm{diag}(\Sigma_{t-1}), \tag{9}$$

where m is the number of clients selected in round $t-1$, Σ_{t-1} is their covariance around the client mean $\bar{\theta}^{t-1} = \frac{1}{m}\sum_{i=1}^{m} \theta_i^{t-1}$, v_t is the coordinate-wise variance vector, and $V_t = \mathrm{Diag}(v_t)$ is the corresponding diagonal covariance approximation. In practice, FedRSAM uses only V_t rather than the full covariance matrix, since the full $d \times d$ covariance is intractable for large models and may be singular under partial participation. Therefore, the free component z is restricted within the standard-deviation range supported by the aggregation statistics, so that $\mathcal{D}_t$ captures the statistically plausible perturbation region around the principal aggregation direction.

Restricted Sharpness-Aware Minimization. Given the aggregation-relevant region $\mathcal{D}_t$, Restricted SAM keeps the same outer min-max form as standard SAM in Eq. (2), but modifies the admissible perturbation set in the inner maximization. Specifically, instead of searching the worst-case perturbation over the full ρ-ball, Restricted SAM searches only within $\mathcal{D}_t$. For notational simplicity, let θ denote the current local model parameter. The restricted inner maximization is

$$\max_{\delta(\alpha,z)\in\mathcal{D}_t,\ \|\delta(\alpha,z)\|_2\le\rho} f(\theta + \delta(\alpha, z)). \tag{10}$$

Following the same first-order approximation used in standard SAM, we linearize the inner objective around θ as

$$f(\theta + \delta(\alpha, z)) \approx f(\theta) + \langle \delta(\alpha, z), \nabla f(\theta) \rangle. \tag{11}$$

Thus, the restricted perturbation is encouraged to align with the current local gradient $\nabla f(\theta)$, but only through directions that lie in the aggregation-relevant region $\mathcal{D}_t$.

The variance-box constraint controls residual exploration within the statistically plausible aggregation range. We define the restricted inner objective as

$$L_{\mathrm{in}}(\alpha, z) := \langle \delta(\alpha, z), \nabla f(\theta) \rangle, \quad \delta(\alpha, z) \in \mathcal{D}_t,\ \|\delta(\alpha, z)\|_2 \le \rho, \tag{12}$$

In practice, we do not solve (12) exactly. Instead, similar to the one-step first-order perturbation construction in standard SAM, we perform one projected ascent step on the perturbation coordinates while keeping θ fixed. For each mini-batch, we reinitialize the inner variables as $\alpha^{(0)}$ and $z^{(0)}$ (usually 0.1 and 0), compute the coordinate gradients $\left(\frac{\partial L_{\mathrm{in}}}{\partial \alpha}, \frac{\partial L_{\mathrm{in}}}{\partial z}\right)$ at $(\alpha^{(0)}, z^{(0)})$, and update

$$\alpha^{(1)} = \alpha^{(0)} + \eta_\alpha \frac{\partial L_{\mathrm{in}}}{\partial \alpha}, \qquad z^{(1)} = \Pi_{[-\sqrt{v_t}, \sqrt{v_t}]} \left(z^{(0)} + \eta_z \frac{\partial L_{\mathrm{in}}}{\partial z} \right), \tag{13}$$

where η_α and η_z are the coordinate ascent stepsizes and $\Pi_{[-\sqrt{v_t}, \sqrt{v_t}]}(\cdot)$ denotes element-wise projection onto the variance box. This projected step makes the box constraint explicit in the inner solver. After obtaining $(\alpha^{(1)}, z^{(1)})$, we construct the raw restricted ascent direction before normalization

$$\delta^{\mathrm{raw}} = (\alpha^{(1)})^2 \hat{d}_t + P_{\perp,t} z^{(1)}, \tag{14}$$

and the normalized radius is $\epsilon^\star = \delta_{\mathrm{raw}} \min\{1, \rho/(\|\delta_{\mathrm{raw}}\|_2 + \varepsilon)\}$, $\varepsilon = 10^{-12}$.

The final scaling step enforces the SAM radius budget while preserving the restricted ascent direction found inside the aggregation-relevant region. Since this scaling only rescales the already projected raw direction, it does not introduce new residual directions outside the admissible cone structure.

Finally, Restricted SAM follows the standard SAM outer step. We evaluate the sharpness-aware gradient at $\theta^\star = \theta + \epsilon^\star$ and update $\theta \leftarrow \theta - \eta \nabla_\theta f(\theta^\star)$, where η is the local learning rate. Therefore, compared with standard SAM, Restricted SAM modifies only the admissible perturbation directions: standard SAM flattens a broad local neighborhood, whereas Restricted SAM flattens the aggregation-relevant region where the global model is more likely to move after server aggregation. FedRSAM extends Restricted SAM to the FL setting, as summarized in Algorithm 1.

Convergence analysis. The convergence guarantee for FedRSAM under partial client participation, together with its assumptions and complete proof, is provided in Appendix D (Theorem 2). Under the conditions and parameter choices of that theorem, $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(\theta^t)\|^2 = O(T^{-1/2})$.

Algorithm 1 Restricted SAM as Local Optimizer in Federated Learning

Input: global model θ^0, client set $\mathcal{C}$, communication rounds T, local steps K, local learning rate η, perturbation radius ρ, inner stepsizes η_α, η_z, server momentum $\beta_s \in [0, 1)$
Initialize: run one FedSAM warm-up round to obtain θ^1 and v_1
for each round $t = 1, \ldots, T-1$ **do**
  Compute $d_t \leftarrow \theta^t - \theta^{t-1}$, $\hat{d}_t \leftarrow d_t / \max(\|d_t\|_2, \varepsilon)$, $P_{\perp,t} \leftarrow I - \hat{d}_t \hat{d}_t^\top$
  Sample a subset of clients $\mathcal{C}^t \subset \mathcal{C}$
  for each client $i \in \mathcal{C}^t$ in parallel **do**
    Initialize local model $\theta_{i,0} \leftarrow \theta^t$
    for each local epoch and mini-batch $\mathcal{B} \in \mathbb{D}_i$ **do**
      Compute local gradient $g_i \leftarrow \nabla F_i(\theta_{i,k}; \mathcal{B})$ and reinitialize $(\alpha_i^{(0)}, z_i^{(0)})$
      Compute one-step restricted ascent on (α_i, z_i) using L_{in} in Eq. (12):

$$\alpha_i^{(1)} \leftarrow \alpha_i^{(0)} + \eta_\alpha \frac{\partial L_{\text{in}}}{\partial \alpha_i}, \qquad z_i^{(1)} \leftarrow \Pi_{[-\sqrt{v_t}, \sqrt{v_t}]} \left(z_i^{(0)} + \eta_z \frac{\partial L_{\text{in}}}{\partial z_i} \right)$$

      Form raw restricted direction $\delta_i^{\text{raw}} \leftarrow (\alpha_i^{(1)})^2 \hat{d}_t + P_{\perp,t} z_i^{(1)}$ and normalize it to ϵ_i
      Update local model $\theta_{i,k+1} \leftarrow \theta_{i,k} - \eta \nabla F_i(\theta_{i,k} + \epsilon_i; \mathcal{B})$
    end for
    Send $\theta_i^t \leftarrow \theta_{i,K}$ back to server
  end for
  Aggregate $\bar{\theta}^t \leftarrow \frac{1}{|\mathcal{C}^t|} \sum_{i \in \mathcal{C}^t} \theta_i^t$ and update $\theta^{t+1} \leftarrow \bar{\theta}^t + \beta_s(\theta^t - \theta^{t-1})$
  Update $\Sigma_t \leftarrow \frac{1}{\max\{|\mathcal{C}^t|-1, 1\}} \sum_{i \in \mathcal{C}^t} (\bar{\theta}^t - \theta_i^t)(\bar{\theta}^t - \theta_i^t)^\top$ and $v_{t+1} \leftarrow \text{diag}(\Sigma_t)$
end for

4.2 Theoretical Insight

FedRSAM promotes aggregation-relevant flatness by reducing local loss sensitivity along shared aggregation directions. We bound the sensitivity error relative to an oracle-aligned perturbation and show that, under uniform directional sampling and informative historical alignment, the restricted cone yields a tighter expected error bound than the full SAM sphere.

For $\eta, \rho > 0$, let f be a quadratic local objective with symmetric Hessian H, and a a shared realized aggregation displacement. Define the signed aggregation sensitivity $\Gamma_a(x) = f(x + a) - f(x)$, one sharpness-aware update $T_\delta(x) = x - \eta \nabla f(x + \delta)$, and the oracle-aligned perturbation $\delta_a = \rho a / \max(\|a\|_2, \varepsilon)$. A positive $\Gamma_a(x)$ represents an aggregation-induced loss increase.

Theorem 1 (Aggregation-aligned perturbation error)**.** *For any feasible perturbation δ,*

$$|\Gamma_a(T_\delta(x)) - \Gamma_a(T_{\delta_a}(x))| \le \eta \|Ha\|_2 \|H\|_2 \|\delta - \delta_a\|_2. \tag{15}$$

Theorem 1 links aggregation-sensitivity error to the distance from the oracle-aligned perturbation. We aim to bound the aggregation-sensitivity error in Eq (15) between any perturbation δ in a region D and the oracle-aligned perturbation. To extend this bound to the region D, we consider perturbations uniformly sampled from D. For $\|a\|_2 > \varepsilon$, let $u = a/\|a\|_2$, $\delta_D \sim \text{Unif}(D)$. Taking expectations in Theorem 1 and applying Cauchy–Schwarz gives

$$\mathbb{E} \, |\Gamma_a(T_{\delta_D}(x)) - \Gamma_a(T_{\delta_a}(x))| \le B_D, \tag{16}$$

where $B_D = \eta\rho \|Ha\|_2 \|H\|_2 Q_D$ and $Q_D = \rho^{-1} \sqrt{\mathbb{E}\|\delta_D - \delta_a\|_2^2}$. Let $\mathcal{S}_\rho = \{\delta : \|\delta\|_2 = \rho\}$ denote the sphere and $\mathcal{C}_\rho^t = \mathcal{D}_t \cap \mathcal{S}_\rho$ denote the spherical section of the restricted cone. We compare the two regions by substituting $D = \mathcal{S}_\rho$ and $D = \mathcal{C}_\rho^t$ into B_D, where a smaller B_D indicates a tighter aggregation-sensitivity error bound. The following proposition identifies conditions under which the forward cone yields a tighter expected bound.

Proposition 1 (Tighter bound)**.** *Assume $\|d_t\|_2 \ge \varepsilon$ and $\langle \hat{d}_t, u \rangle > 0$. Write each $\delta \in C_t^\rho$ as $s\hat{d}_t + r$, with $s \ge 0$ and $r \perp \hat{d}_t$. If the residual section is centrally symmetric and $\{\delta \in C_t^\rho : s > 0\}$ has positive normalized surface measure, then $Q_{C_t^\rho} < Q_{S_\rho}$, and $B_{C_t^\rho} < B_{S_\rho}$ when $\|Ha\|_2 > 0$.*

Together, the two results show how informative historical alignment can tighten the expected aggregation-sensitivity error bound through the cone restriction. The NCD measurements in Table 4

Table 1: Test accuracy (%) on vision benchmarks under heterogeneous settings ($\alpha_{dir} = 0.01$ for Tiny-ImageNet and 0.1 for rest, 300 rounds for CIFAR datasets and 100 for Tiny-ImageNet, "–" indicates that FedSWA does not converge under the corresponding setting.)

Algorithm	CIFAR100		CIFAR10		Tiny-ImageNet
	LeNet5	ResNet18	LeNet5	ResNet18	ViT-B
FedAvg	$35.97_{\pm.42}$	$34.64_{\pm.43}$	$67.64_{\pm.19}$	$62.63_{\pm.23}$	$74.72_{\pm.38}$
FedAvgM	$36.32_{\pm.17}$	$29.24_{\pm.46}$	$50.64_{\pm.22}$	$53.07_{\pm.22}$	$76.47_{\pm.21}$
SCAFFOLD	$34.13_{\pm.45}$	$36.41_{\pm.44}$	$66.46_{\pm.35}$	$63.03_{\pm.39}$	$70.84_{\pm.52}$
FedSAM	$36.60_{\pm.35}$	$37.90_{\pm.10}$	$68.46_{\pm.47}$	$64.62_{\pm.41}$	$74.35_{\pm.44}$
FedLESAM	$35.77_{\pm.62}$	$32.33_{\pm.46}$	$67.71_{\pm.58}$	$58.99_{\pm.16}$	$75.92_{\pm.49}$
FedGAMMA	$41.32_{\pm.51}$	$39.68_{\pm.30}$	$70.71_{\pm.49}$	$66.89_{\pm.73}$	$77.85_{\pm.13}$
FedSWA	$35.95_{\pm.28}$	$37.31_{\pm.19}$	$68.88_{\pm.11}$	$61.05_{\pm.36}$	–
FedNSAM	$43.86_{\pm.53}$	$45.25_{\pm.64}$	$70.85_{\pm.49}$	$53.21_{\pm.55}$	$77.64_{\pm.57}$
FedRSAM	$\mathbf{44.91}_{\pm.44}$	$\mathbf{52.51}_{\pm.61}$	$\mathbf{76.30}_{\pm.48}$	$\mathbf{75.01}_{\pm.46}$	$\mathbf{81.55}_{\pm.31}$

Table 2: Comparison on GLUE benchmark tasks with RoBERTa-Base with $\alpha_{dir} = 0.1$, Round=100.

Algorithm	QNLI		MNLI		RTE		QQP	
	ACC	Loss	ACC	Loss	ACC	Loss	ACC	Loss
FedAvg	$49.13_{\pm.21}$	0.6941	$78.95_{\pm.34}$	0.5241	$49.82_{\pm.27}$	1.5337	$76.92_{\pm.31}$	0.5678
FedSAM	$49.46_{\pm.24}$	0.8939	$78.13_{\pm.49}$	0.5445	$49.46_{\pm.22}$	0.8516	$78.01_{\pm.26}$	0.4779
FedNSAM	$76.92_{\pm.33}$	0.4916	$78.70_{\pm.28}$	0.5304	$50.54_{\pm.30}$	0.8197	$78.65_{\pm.25}$	0.4568
FedAdamW	$51.58_{\pm.36}$	0.7144	$81.05_{\pm.63}$	0.4622	$47.29_{\pm.32}$	0.7330	$79.55_{\pm.27}$	0.4623
FedRSAM	$\mathbf{81.04}_{\pm.18}$	0.4311	$\mathbf{81.15}_{\pm.20}$	0.4780	$\mathbf{54.23}_{\pm.49}$	1.0954	$\mathbf{80.32}_{\pm.22}$	0.4268

provide corresponding empirical coverage evidence. This comparison uses uniform directional references under the stated assumptions; it does not guarantee faster convergence for optimized perturbations. Proofs and limitations are in Appendix C.

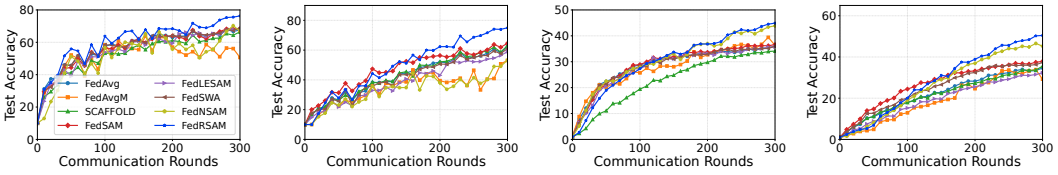
5 Experiments

Datasets. We evaluate the proposed method on both vision and language tasks. (i) For image classification, we use CIFAR-10, CIFAR-100 Krizhevsky [2009], and Tiny-ImageNet Le and Yang [2015]. (ii) For NLP tasks, we adopt benchmark datasets from the GLUE benchmark Wang et al. [2018], including QNLI, MNLI, RTE and QQP. To simulate data heterogeneity across clients, we follow the widely used Dirichlet partitioning scheme Hsu et al. [2019], where smaller Dirichlet concentration parameter α indicates stronger heterogeneity. Specifically, we use $\alpha_{dir} \in \{0.01, 0.1\}$ to represent different non-IID levels (α_{dir} is different from the learnable parameter α).

Model Architectures. We explore a variety of model types: (i) LeNet-5 LeCun et al. [1998] and ResNet-18 He et al. [2016] as representative convolutional neural networks (CNNs), (ii) Vision Transformer (ViT) Dosovitskiy et al. [2021] for Vision Transformers, and (iii) RoBERTa-Base Liu et al. [2019] for large-scale language models.

Baselines. We compare our method against representative federated learning baselines, including FedAvg McMahan et al. [2017], FedAvgM Hsu et al. [2019], SCAFFOLD Karimireddy et al. [2020], FedSAM Qu et al. [2022], FedLESAM Fan et al. [2024], FedGAMMA Dai et al. [2024], FedSWA, Liu et al. [2025a], FedAdamW Liu et al. [2026], and FedNSAM Liu et al. [2025b].

Hyperparameter Settings. For all methods, we use the same client partition, participation ratio, batch size, and model initialization within each benchmark. For the vision benchmarks, we use 100 clients and train all methods for 300 communication rounds with local epochs $E = 5$ on CIFAR-10 and CIFAR-100 ($\rho = 0.1$), and 100 rounds with $E = 1$ on Tiny-ImageNet ($\rho = 0.1$). We set $\alpha_{\text{dir}} = 0.01$ for Tiny-ImageNet and $\alpha_{\text{dir}} = 0.1$ for the other vision datasets. All CNN-based vision experiments use the same normalization setting, i.e., Group Normalization (GN), with server momentum $\beta_s = 0.85$ for FedRSAM. For language experiments, we train RoBERTa-Base with LoRA on GLUE benchmark tasks for 100 communication rounds with $E = 10$, $\alpha_{\text{dir}} = 0.1$, and



(a) CIFAR-10, LeNet5 (b) CIFAR-10, ResNet18 (c) CIFAR-100, LeNet5 (d) CIFAR-100, ResNet18

Figure 2: Accuracy curves of different methods under Dirichlet-0.1 data heterogeneity

Table 3: Component ablation of FedRSAM on CIFAR-10 and CIFAR-100 with ResNet-18. Accuracy (%) is reported.

Variant	**Perturb δ**	**Learnable α**	**Residual z**	**Variance**	**CIFAR-10**	**CIFAR-100**
w/o cone $\mathcal{D}$	✓	–	–	–	$49.79_{\pm 0.41}$	$42.61_{\pm 0.17}$
Cone-axis α only	✓	✓	–	–	$72.04_{\pm 0.22}$	$48.76_{\pm 0.13}$
Residual z only	✓	–	✓	–	$51.33_{\pm 0.39}$	$37.33_{\pm 0.21}$
$\alpha + z$ (w/o variance)	✓	✓	✓	–	$70.80_{\pm 0.45}$	$50.52_{\pm 0.38}$
FedRSAM	✓	✓	✓	✓	$75.01_{\pm 0.46}$	$52.51_{\pm 0.61}$

$\rho = 0.05$. The learning rate is set to 0.1 for SGD-based methods and 3×10^{-4} for FedAdamW, while the momentum coefficient of momentum-based methods is tuned from $\{0.1, 0.7\}$ with a step size of 0.1. For fairness, FedRSAM is not tuned with a full grid over all method-specific hyperparameters. We fix $\eta_\alpha = 0.5$, and tune only the more sensitive $\eta_z \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. Detailed settings, tuning results and sensitivity analysis are provided in Appendices A.1, A.2 and A.3.

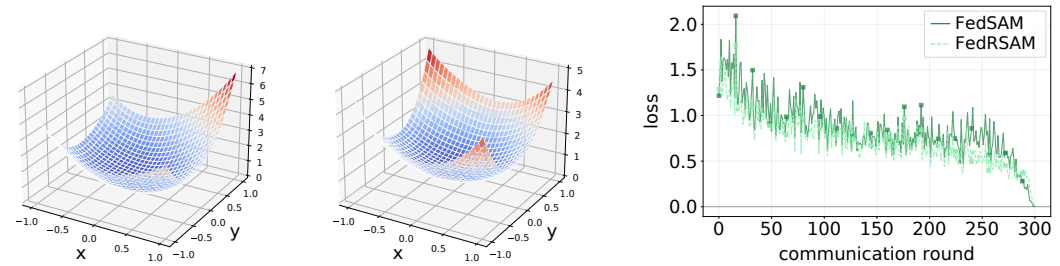
5.1 Main Results

Vision Benchmarks. Table 1 reports the test accuracy on CIFAR100, CIFAR10, and Tiny-ImageNet under heterogeneous settings. FedRSAM consistently achieves the best performance on all five vision benchmark settings, outperforming both classical FL baselines and existing SAM-based federated methods. In particular, on CIFAR100, FedRSAM obtains **44.91%** with LeNet5 and **52.51%** with ResNet18, improving over the strongest baseline FedNSAM by 1.05 and 7.26 points, respectively. On CIFAR10, FedRSAM further reaches **76.30%** on LeNet5 and **75.01%** on ResNet18, outperforming FedNSAM by 5.45 and 21.80 points. On Tiny-ImageNet, FedRSAM also achieves the highest accuracy, i.e., **81.55%** with ViT-B, surpassing FedNSAM by 3.91 points. These results show that FedRSAM is effective across both CNN and Transformer architectures.

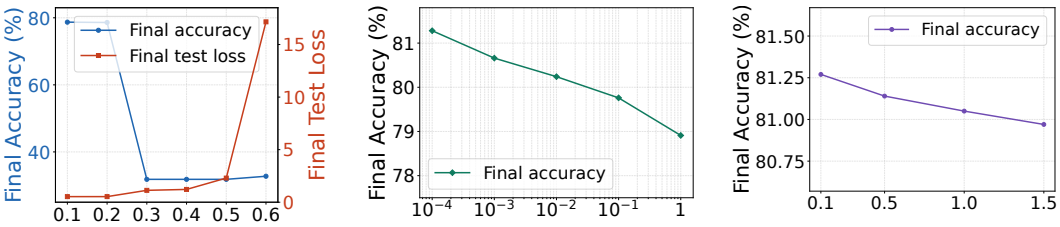
Language Benchmarks. Table 2 reports the results on GLUE benchmark tasks with RoBERTa-Base under heterogeneous settings. FedRSAM achieves the best accuracy on all four tasks, reaching **81.04%**, **81.15%**, **54.23%**, and **80.32%** on QNLI, MNLI, RTE, and QQP, respectively. The improvement is especially significant on QNLI, MNLI, and QQP, where FedRSAM consistently outperforms FedSAM and FedNSAM by a clear margin. Although FedAdamW with a carefully tuned momentum coefficient can achieve competitive results on MNLI and QQP, its performance is highly sensitive to the momentum setting. For example, when the momentum coefficient is increased to 0.2, FedAdamW immediately drops to 77.02% on MNLI, and its performance on QNLI consistently collapses after momentum tuning. Therefore, FedRSAM is more favorable in terms of both stability and average performance. Additional comparison results are provided in Appendix B.1.

5.2 Ablation

Component Analysis. We conduct ablation experiments to evaluate the contributions of the learnable cone-axis coefficient α and the variance-constrained residual z in FedRSAM. Table 3 compares four variants: w/o cone $\mathcal{D}$, which uses unconstrained SAM with server momentum; Cone-axis α only, which retains only the cone-axis perturbation; Residual z only, which restricts perturbations to the orthogonal residual subspace; and $\alpha + z$ (w/o variance), which introduces an unconstrained residual alongside the learnable cone-axis coefficient. Compared with w/o cone, the learnable cone-axis improves mean accuracy from 49.79%/42.61% to 72.04%/48.76% on CIFAR-10/100, confirming the importance of aggregation-aligned perturbations. In contrast, the Residual z only



(a) FedRSAM (b) FedSAM (c) Aggregation loss gap



(d) FedNSAM momentum (e) η_z (f) η_α

Figure 3: Model analysis.

variant achieves 51.33%/37.33%, substantially underperforming the Cone-axis α only variant. The $\alpha + z$ (w/o variance) variant achieves 70.80%/50.52%, showing inconsistent gains over the Cone-axis α only variant (72.04%/48.76%). FedRSAM, combining both components within the restricted cone, further improves accuracy to 75.01%/52.51%, demonstrating the effectiveness of the full cone with a variance-constrained residual.

5.3 Model Analysis

Aggregation Loss Gap. To better understand the difference between global neighborhood flatness and aggregation-aware flatness, we compare the loss geometry and aggregation behavior of FedSAM and FedRSAM on CIFAR-10. As shown in Fig. 3a and Fig. 3b, FedSAM exhibits a flatter overall loss surface than FedRSAM, since it searches the worst-case perturbation in an unconstrained local neighborhood and thus flattens a broader set of directions. In contrast, FedRSAM restricts the perturbation search region and only optimizes a subset of directions. However, Fig. 3c shows an opposite trend from the loss-surface comparison. We define the aggregation gap at round t as $\mathrm{Gap}_t := \mathcal{L}^{(t)}_{\mathrm{global}} - \frac{1}{|S_t|}\sum_{i\in S_t}\mathcal{L}^{(t)}_i$. Although FedSAM achieves a flatter global surface, FedRSAM consistently yields a smaller aggregation gap during training. This suggests that FedRSAM sacrifices part of the global-neighborhood flatness in exchange for more targeted flattening around the aggregation-relevant region, making the aggregated global parameter incur a smaller loss increase relative to the local client losses.

Client-Side Momentum Fed vs. Momentum-free Fed. To investigate the instability of client-side momentum correction under non-IID settings, we conduct sensitivity analyses on MNLI with RoBERTa-Base + LoRA using Dirichlet parameter of 0.1. As shown in Figure 3d, FedNSAM achieves 78.70%/78.60% accuracy with small momentum coefficients ($\gamma = 0.1/0.2$), but collapses to approximately 32% when $\gamma \geq 0.3$. In contrast, FedRSAM uses historical information to constrain SAM perturbations without accumulating client-side momentum. Figures 3e and 3f show that its accuracy ranges from 78.91% to 81.28% across residual stepsizes and from approximately 80.6% to 81.6% across principal-direction stepsizes, without similar performance collapse. These results suggest that FedRSAM provides more stable client-side optimization across the tested hyperparameter ranges. Detailed momentum analysis and computation, communication, and privacy discussions are provided in Appendices A.3 and A.4, respectively.

Reliability Analysis of Aggregation Direction Prediction. We investigate whether the previous global update direction and its cone region can predict the next aggregation direction on CIFAR-100/ResNet-18. Specifically, let $d_t = \theta_t - \theta_{t-1}$. Eq. (17) uses Cosine similarity (Cos) to measure directional prediction and Normalized Cone Distance (NCD) to measure the deviation between the

Table 4: Reliability diagnostics for aggregation direction prediction on CIFAR-100/ResNet-18.

Participation	Cosine ↑	Axis NCD ↓	Full NCD ↓	Shuffled NCD ↓	Hit@10% ↑
10%	0.7464	0.6385	0.0298	0.4667	94.00%
5%	0.7455	0.6457	0.0803	0.5163	72.00%
1%	0.7670	0.6075	0.2090	0.4210	7.36%

Table 5: Stability Test.

Angle	Final acc. (%)	Acc. drop (pp)
0°	75.02	–
14.03°	73.91	1.11
26.56°	73.40	1.62
45.00°	73.07	1.95

Table 6: Efficiency Comparison.

Algorithm	Comm. (GiB)	Time (s/round)
FedSAM	52.571	8.81
FedRSAM	78.768	13.43
FedSMOO	183.997	15.57
FedGloSS	52.571	19.32

realized aggregation direction and the predicted region. Hit@10% denotes the fraction of realized directions with $\mathrm{NCD} \le 0.1$ (The next update d_{t+1} is used only for post-hoc evaluation).

$$\mathrm{Cos}_{t+1} = \frac{\langle d_t, d_{t+1}\rangle}{\|d_t\|_2\|d_{t+1}\|_2}, \qquad \mathrm{NCD}_{t+1} = \frac{\min_{q\in D_t}\|y_{t+1}-q\|_2}{\rho}, \quad y_{t+1} = \frac{\rho d_{t+1}}{\max(\|d_{t+1}\|_2,\varepsilon)}. \tag{17}$$

Table 4 shows Cos of 0.7464/0.7455/0.7670 at 10%/5%/1% participation, indicating consistent directional predictability. However, the historical axis alone yields NCD values of 0.6385/0.6457/0.6075, while the full cone reduces them to 0.0298/0.0803/0.2090. Shuffling the variance coordinates increases NCD to 0.4667/0.5163/0.4210, confirming the importance of coordinate-wise variance. Meanwhile, Hit@10% decreases from 94% to 72% and 7.36%, indicating substantial coverage degradation under extremely sparse participation. The 1% setting uses 500 clients, whereas the others use 100 clients, and therefore serves as an extreme stress test. Under this setting, FedRSAM achieves 9.85% final accuracy, outperforming FedSAM's 7.30%. Overall, these results demonstrate the effectiveness of the cone.

Robustness to Direction Error. To examine whether FedRSAM remains effective under inaccurate historical direction predictions, we inject controlled directional perturbations into the historical axis on CIFAR-10/ResNet-18. As shown in Table 5, angular deviations of 14.03°, 26.56°, and 45° reduce final accuracy from 75.02% by 1.11, 1.62, and 1.95 points, respectively. These results indicate that FedRSAM tolerates a certain degree of directional prediction error, although performance gradually decreases as the error increases. Detailed experimental settings are provided in Appendix B.2.

Performance–Efficiency Trade-off. Runtime and communication use MobileNetV2/CIFAR-100 with 100 clients, 10 clients per round, 300 rounds. As shown in Table 6, FedRSAM takes 13.43 s per round, compared with 8.81 s for FedSAM, 15.57 s for FedSMOO, and 19.32 s for FedGloSS. Its total communication is 78.768 GiB, versus 52.571 GiB for FedSAM/FedGloSS and 183.997 GiB for FedSMOO. Thus, its accuracy gains cost approximately 52% more runtime and 50% more communication than FedSAM; it is faster than both stronger baselines and uses less communication than FedSMOO. Comparisons with these baselines are provided in Appendix B.1. Memory measurements are provided in Appendix A.4, Table 11.

6 Conclusion

We proposed FedRSAM, a restricted sharpness-aware optimization method for federated learning that explicitly flattens the aggregation-relevant region rather than the entire local neighborhood. By constructing this region using the previous global update direction and the variance of client updates, FedRSAM better aligns local sharpness-aware optimization with server-side aggregation while avoiding biased client-side momentum accumulation under severe data heterogeneity. We further theoretically demonstrated the advantage of FedRSAM's cone-shaped region over the conventional spherical region for federated aggregation, while retaining the same-order convergence guarantee as FedSAM. Extensive experiments on both vision and language benchmarks demonstrated that FedRSAM consistently improves training stability and achieves superior performance under heterogeneous settings.

Acknowledgments and Disclosure of Funding

The authors declare no competing interests. This work was supported by the Australian Research Council (ARC) under Discovery Project Grant DP220101420 and the Higher Education Natural Science Foundation of Anhui Province, China (No. 2025AHGXZK30347).

References

Debora Caldarola, Pietro Cagnasso, Barbara Caputo, and Marco Ciccone. Beyond local sharpness: Communication-efficient global sharpness-aware minimization for federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25187–25197, 2025.

Rong Dai, Xun Yang, Yan Sun, Li Shen, Xinmei Tian, Meng Wang, and Yongdong Zhang. FedGAMMA: Federated learning with global sharpness-aware minimization. *IEEE Transactions on Neural Networks and Learning Systems*, 35(12):17479–17492, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 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. In *Proceedings of the 41st International Conference on Machine Learning*, pages 12858–12881, 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*, pages 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.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14 (1–2):1–210, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5132–5143, 2020.

Geeho Kim, Jinkyu Kim, and Bohyung Han. Communication-efficient federated learning with accelerated client gradient. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12385–12394, 2024.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. Cs231n project report, Stanford University, 2015.

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

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450, 2020a.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2020b.

Junkang Liu, Yuanyuan Liu, Fanhua Shang, Hongying Liu, Jin Liu, and Wei Feng. Improving generalization in federated learning with highly heterogeneous data via momentum-based stochastic controlled weight averaging. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 38894–38939, 2025a.

Junkang Liu, Fanhua Shang, Yuxuan Tian, Hongying Liu, and Yuanyuan Liu. Consistency of local and global flatness for federated learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3875–3883, 2025b. doi: 10.1145/3746027.3755226.

Junkang Liu, Fanhua Shang, Hongying Liu, Yuxuan Tian, Yuanyuan Liu, Jin Liu, Kewen Zhu, and Zhouchen Lin. FedAdamW: A communication-efficient optimizer with convergence and generalization guarantees for federated large models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(28):23748–23756, 2026.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 2017.

Zhuang Qi, Ying-Peng Tang, Lei Meng, Guoqing Chao, Lei Wu, Han Yu, and Xiangxu Meng. From selection to scheduling: Federated geometry-aware correction makes exemplar replay work better under continual dynamic heterogeneity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17494–17504, June 2026.

Zhe Qu, Xingyu Li, Rui Duan, Yao Liu, Bo Tang, and Zhuo Lu. Generalized federated learning via sharpness aware minimization. In *Proceedings of the 39th International Conference on Machine Learning*, pages 18250–18280, 2022.

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 *Proceedings of the 40th International Conference on Machine Learning*, pages 32991–33013, 2023.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 353–355, 2018.

A Experimental Fairness and Hyperparameter Settings

We provide detailed experimental settings to clarify the fairness of our comparisons. For each benchmark and backbone, all methods are evaluated under the same federated data partition, client pool, participation rate, local batch size, local update length, communication rounds, and model initialization.

Fairness statement. FedRSAM does not use additional local epochs, additional participating clients, or additional communication rounds. Its principal direction is computed from two consecutive global models and therefore does not require transmitting a full basis. The only additional restriction information is the variance constraint, which keeps the auxiliary communication at the same order as momentum-based methods. As shown in Table 7, for fairness, all methods are evaluated under a controlled training protocol, including the same data partition, client pool, participation rate, local batch size, local update length, communication rounds, and model initialization. Common hyperparameters are fixed across methods whenever applicable. Since exhaustive tuning of all hyperparameters for all baselines is computationally expensive, we only conduct limited searches over method-specific sensitive hyperparameters, including the momentum coefficient for momentum-based baselines and the residual stepsize η_z for FedRSAM. For fair comparison, all CNN-based vision experiments use the same normalization setting, i.e., Group Normalization (GN).

Table 7: Controlled experimental protocol. All methods share the same training budget and common hyperparameters. Only method-specific sensitive hyperparameters are searched.

Category	Setting	Status
Data partition	Same Dirichlet split	Fixed
Client sampling	Same participation ratio and sampling protocol	Fixed
Initialization	Same model initialization	Fixed
Communication rounds	Same within each benchmark	Fixed
Local epochs	Same within each benchmark	Fixed
Batch size	Same within each benchmark	Fixed
Model architecture	Same backbone within each benchmark	Fixed
Base learning rate	Same within each optimizer family	Fixed
Weight decay	Same within each benchmark	Fixed
SAM radius ρ	Same for SAM-based methods where applicable	Fixed
Client-side momentum coefficient	Momentum-based baselines	Searched
Server-side momentum coefficient	Fixed by default	Fixed
Residual stepsize η_z	FedRSAM	Searched
Principal stepsize η_α	Fixed by default	Fixed

A.1 Vision Experiments

Table 8 provides a compact overview of the default training settings for vision experiments, including datasets, backbones, client number, participation rate, communication rounds, batch size, and local epochs.

Federated setup. For vision benchmarks, we use 100 clients and sample a fixed fraction of clients in each communication round. Unless otherwise specified, the participation rate is 10%, corresponding to 10 selected clients per round. Data heterogeneity is simulated by Dirichlet label partitioning. We use $\alpha_{\mathrm{dir}} = 0.1$ for CIFAR-10 and CIFAR-100, and $\alpha_{\mathrm{dir}} = 0.01$ for Tiny-ImageNet. All algorithms are evaluated on the same client partition for a given dataset and heterogeneity level.

Training details. For CIFAR-10 and CIFAR-100, we train LeNet-5 and ResNet-18 using SGD with batch size 50. The default number of local epochs is $E = 5$, and the total number of communication rounds is 300. For Transformer-based vision models on Tiny-ImageNet, including ViT-B, we use ImageNet-pretrained weights, batch size 16, local epoch $E = 1$, and 100 communication rounds with a fixed server-side momentum coefficient of $\beta_s = 0.85$.. All Tiny-ImageNet Transformer experiments are conducted with LoRA fine-tuning. Specifically, we set the LoRA rank to $r = 16$, the LoRA scaling factor to $2r = 32$, and the LoRA dropout to 0.05. For ViT backbones, LoRA is applied to the attention modules, including the query, key, and value projections, and the trainable parameters

Table 8: Default settings for vision experiments.

Dataset	Backbone	Clients	Participation	Rounds	Batch size	Local epoch	Partition	LoRA
CIFAR-10	LeNet-5 / ResNet-18	100	10%	300	50	5	Dirichlet $\alpha_{\text{dir}} = 0.1$	–
CIFAR-100	LeNet-5 / ResNet-18	100	10%	300	50	5	Dirichlet $\alpha_{\text{dir}} = 0.1$	–
Tiny-ImageNet	ViT-B	100	10%	100	16	1	Dirichlet $\alpha_{\text{dir}} = 0.01$	$r = 16$

Table 9: Default settings for language experiments.

Backbone	Clients	Participation	Rounds	Batch size	Local epoch	Partition	LoRA rank
RoBERTa-base	20	20%	100	16	10	Dirichlet $\alpha_{\text{dir}} = 0.1$	16

consist of the LoRA parameters and the classification head. The perturbation radius for SAM-based methods is selected from the same candidate set.

A.2 Language Model Experiments

Table 9 summarizes the default training settings for language experiments, including the backbone, client number, participation rate, communication rounds, batch size, local epochs, data partition, and LoRA rank.

Federated setup. For language tasks, we evaluate on GLUE-style benchmarks using RoBERTa-base. We use 20 clients and sample 20% of clients per round. All methods use the same client partition, communication rounds, local epoch, batch size, and model initialization. Unless otherwise specified, we use 100 communication rounds, local epoch $E = 10$, batch size 16, and Dirichlet partition with $\alpha_{\text{dir}} = 0.1$.

Training details. All language experiments use the same parameter-efficient fine-tuning protocol. When LoRA is enabled, we set rank $r = 16$, LoRA scaling $2r$, dropout 0.05, and apply LoRA to the query, key, value, and dense modules of RoBERTa-base. Thus, FedRSAM and all baselines update the same trainable parameter subset.

A.3 Parameter Analysis

Fine tune on momentum Table 10 reports the sensitivity of FedNSAM to the momentum coefficient on MNLI and QNLI with RoBERTa-Base + LoRA. Small momentum coefficients can be effective: FedNSAM converges when $\gamma = 0.1$ or 0.2, achieving 78.70% / 78.60% on MNLI and 71.35% / 76.92% on QNLI. However, when $\gamma \geq 0.3$, the performance collapses on both tasks, and larger coefficients even lead to NaN on QNLI. This suggests that momentum correction is useful only within a narrow range under heterogeneous language-model fine-tuning, while overly large momentum may amplify biased historical update directions and destabilize the main optimization trajectory.

A.4 Computation, Communication, and Privacy Analysis.

As shown in Table 11, FedRSAM has the same local computation order as other SAM-based methods, requiring approximately two forward/backward passes per local update. Compared with FedSAM, FedRSAM increases client resident/peak memory from 42.80/367.33 MiB to 128.41/481.41 MiB, while remaining below FedSMOO (171.22/584.38 MiB). Its additional cost comes from transmitting the coordinate-wise aggregation variance used to constrain the restricted perturbation, which introduces $1 \times d$ auxiliary communication, comparable to the correction or moment states used by FedNSAM and FedAdamW. The variance is computed from client model updates rather than raw data, and is therefore compatible with the standard FL privacy principle that only model updates or their aggregate statistics are exchanged. Moreover, the variance can be computed after secure aggregation or other privacy-preserving aggregation protocols, so FedRSAM does not require exposing individual client data or local samples.

Table 10: FedNSAM momentum sensitivity on language tasks with RoBERTa-Base + LoRA.

Task	γ	Final Acc.	Train Loss	Test Loss	Converged
MNLI	0.1	78.70	0.51547	0.53039	Yes
MNLI	0.2	78.60	0.51512	0.53126	Yes
MNLI	0.3	31.82	1.14474	1.12870	No
MNLI	0.4	31.82	1.22261	1.21543	No
MNLI	0.5	31.82	2.34137	2.29549	No
MNLI	0.6	32.74	15.39927	17.15720	No
QNLI	0.1	71.35	0.65272	0.63041	Yes
QNLI	0.2	76.92	0.52732	0.49156	Yes
QNLI	0.3	49.46	0.96590	0.99035	No
QNLI	0.4	49.46	0.83885	0.85896	No
QNLI	0.5	NaN	NaN	NaN	No
QNLI	0.6	NaN	NaN	NaN	No

Table 11: Memory on CIFAR-100 with 100 clients, 1% participation, $E = 5$, and batch size 50. Client peak denotes effective peak memory; server units are explicit.

Method	Client resident (MiB)	Client peak (MiB)	Server resident
FedSAM	42.80	367.33	42.80 MiB
FedRSAM	128.41	481.41	85.61 MiB
FedSMOO	171.22	584.38	16.44 GiB
FedGloSS	47.83	355.34	128.41 MiB

A.5 Limitations.

FedRSAM uses coordinate-wise client-update variance to constrain residual perturbations, but its effectiveness depends on the reliability of historical aggregation statistics, particularly the previous global update direction. Since the degree of client data heterogeneity is generally unknown, selecting an appropriate residual stepsize η_z remains challenging. Although historical update variance provides an indirect indicator of heterogeneity, it may be unreliable under sparse client participation. Moreover, FedRSAM achieves improved accuracy at the cost of additional runtime and communication overhead compared with FedSAM, reflecting a performance–efficiency trade-off.

A.6 Computation Resources

All experiments were conducted on a Linux server running Ubuntu 20.04.6 LTS with two NVIDIA A800 80GB PCIe GPUs. The machine has 128 CPU threads from two Hygon C86 7380 32-core processors and 251 GiB system memory. The software environment uses Python 3.8.20, PyTorch 2.4.1 with CUDA 12.1, NVIDIA driver 535.183.01, Ray 1.0.0, PEFT 0.13.2, and Transformers 4.46.3.

Computational overhead. Let d denote the number of trainable parameters and K denote the maximum number of local mini-batch steps per selected client. FedAvg performs one forward/backward update per local mini-batch. SAM-based methods such as FedSAM and FedNSAM require two gradient-evaluation passes per mini-batch: one to construct the adversarial perturbation and one to update the model at the perturbed point. FedRSAM follows the same SAM-style two-stage update, but constrains the perturbation to a client-adaptive subspace composed of a principal direction and a variance-controlled residual component. During the perturbation construction, the model parameters θ are fixed, and FedRSAM only solves for the scalar direction coefficient α and the residual variable z. Therefore, FedRSAM has the same order of gradient-evaluation cost as existing SAM-based baselines, while introducing only lightweight element-wise projection, clipping, and normalization operations for the restricted perturbation.

Implementation details. Client training is parallelized with Ray actors. The argument `num_gpus_per` controls the fractional GPU resource assigned to each actor; typical settings are 0.1–0.2 for vision experiments and 0.5 for larger transformer experiments. For FedRSAM, the default hyperparameters used in the implementation are z learning rate 10^{-3}, and LoRA rank 16 when LoRA is enabled.

Table 12: Additional CIFAR-100 comparisons over 300 rounds. Best and final accuracies (%) are reported separately for the two baselines.

Backbone	Part.	FedRSAM	FedSMOO		FedGloSS	
			Best	Final	Best	Final
ResNet-18	5%	**37.99**	26.00	25.92	18.16	14.09
ResNet-18	10%	**52.51**	30.07	29.28	26.95	26.35
MobileNetV2	10%	**42.25**	35.76	32.54	26.21	24.44

Table 13: Settings for the additional CIFAR-100 comparisons. ResNet-18 uses 5% and 10% participation; MobileNetV2 uses 10%.

Backbone	Method	Reported settings
MobileNetV2	FedSMOO	$\lambda = 0.01$, client learning rate 0.05, $\rho = 0.05$
	FedGloSS	$\beta = 100$, client learning rate 0.05, server learning rate 1.0
	FedRSAM	$\eta_\alpha = 0.5$, $\eta_z = 10^{-3}$, learning rate 0.1, transferred ρ schedule $0.001 \to 0.2$
ResNet-18	FedSMOO	Fixed $\rho = 0.1$, $\lambda = 0.1$
	FedGloSS	Client/server learning rates $0.05/1.0$, $\rho_l = 0.2$, $\beta = 100$, server-radius schedule $0.001 \to 0.01$ with $T_s = 225$

B Additional Experiments

B.1 Stronger Baselines and MobileNetV2

To further compare FedRSAM with state-of-the-art methods, we evaluate it against FedSMOO Sun et al. [2023] and FedGloSS Caldarola et al. [2025] on CIFAR-100 over 300 rounds, as shown in Table 12. FedRSAM outperforms both methods on ResNet-18 at 5% and 10% participation and on MobileNetV2 at 10%. Table 13 lists the reported method-specific settings. On MobileNetV2, we tune FedSMOO and FedGloSS and report their best configurations; FedRSAM receives no additional CIFAR-100 search and uses a transferred perturbation-radius schedule. All remaining settings follow the corresponding official implementations and the shared experimental protocol.

B.2 Controlled Historical-direction Error

We isolate axis error on CIFAR-10/ResNet-18 with 300 rounds, 10% participation. For a unit vector $u_\perp$ orthogonal to the historical unit direction $\hat{d}_t$, replace only that direction with

$$\widetilde{d}_t = \frac{\hat{d}_t + \sigma u_\perp}{\sqrt{1+\sigma^2}}, \qquad u_\perp^\top \hat{d}_t = 0. \tag{18}$$

The angular deviation is $\arctan(\sigma)$. Variance statistics, perturbation radius, client selection, and the training random trajectory remain fixed. Table 14 reports complete single-seed runs.

Table 14: Controlled historical-direction corruption on CIFAR-10/ResNet-18. Accuracy drops are measured in percentage points relative to $\sigma = 0$.

σ	Angle	Best accuracy (%)	Final accuracy (%)	Final drop
0	0°	75.16	75.02	–
0.25	14.03°	73.91	73.91	1.11
0.5	26.56°	73.58	73.40	1.62
1.0	45.00°	73.07	73.07	1.95

C Proofs of the Aggregation-aligned Results

To analyze the effect of aggregation-aligned perturbations, we consider a quadratic local model motivated by the second-order Taylor approximation of a smooth loss function. Specifically, we assume $f(x) = \frac{1}{2}x^\top Hx + b^\top x + c$, where $H = H^\top$ is a constant Hessian matrix. The following results are exact under this quadratic assumption.

Let a denote a fixed aggregation displacement, and define the signed aggregation sensitivity as $\Gamma_a(x) = f(x+a) - f(x)$. A positive value indicates an aggregation-induced loss increase. For $\eta, \rho > 0$, let $T_\delta(x) = x - \eta\nabla f(x+\delta)$ denote one sharpness-aware update, and define $\delta_a = \rho a / \max(\|a\|_2, \varepsilon)$ as the oracle-aligned reference perturbation. The displacement a is held fixed when comparing perturbations.

Theorem 1 (Restated). For the quadratic objective defined above and any feasible perturbation δ, the aggregation-sensitivity error relative to the oracle-aligned perturbation satisfies

$$|\Gamma_a(T_\delta(x)) - \Gamma_a(T_{\delta_a}(x))| \le \eta\|Ha\|_2\|H\|_2\|\delta - \delta_a\|_2. \tag{19}$$

Proof. By expanding the quadratic objective and using $H = H^\top$, we obtain

$$\begin{aligned}\Gamma_a(x) &= f(x+a) - f(x)\\ &= \frac{1}{2}\left[(x+a)^\top H(x+a) - x^\top Hx\right] + b^\top a\\ &= a^\top Hx + \frac{1}{2}a^\top Ha + b^\top a.\end{aligned} \tag{20}$$

Since a is fixed, the terms independent of x cancel when comparing two parameter points x_1 and x_2. Thus, $\Gamma_a(x_1) - \Gamma_a(x_2) = a^\top H(x_1 - x_2)$. Next, using $\nabla f(x+\delta) = H(x+\delta) + b$, the difference between the two sharpness-aware updates is

$$\begin{aligned}T_\delta(x) - T_{\delta_a}(x) &= -\eta\left[\nabla f(x+\delta) - \nabla f(x+\delta_a)\right]\\ &= \eta H(\delta_a - \delta).\end{aligned} \tag{21}$$

Substituting Eq. (21) into the sensitivity difference and applying the Cauchy–Schwarz and spectral-norm inequalities yields

$$\begin{aligned}&|\Gamma_a(T_\delta(x)) - \Gamma_a(T_{\delta_a}(x))|\\ &= \left|a^\top H\left(T_\delta(x) - T_{\delta_a}(x)\right)\right|\\ &= \eta\left|\langle Ha, H(\delta_a - \delta)\rangle\right|\\ &\le \eta\|Ha\|_2\|H(\delta_a - \delta)\|_2\\ &\le \eta\|Ha\|_2\|H\|_2\|\delta - \delta_a\|_2.\end{aligned} \tag{22}$$

This proves Theorem 1.

Expected Sensitivity Error Bound. To compare different perturbation regions, let $\delta_D \sim \mathrm{Unif}(D)$, where D is a radius-ρ section and uniformity is defined by normalized surface measure. Taking expectations in Eq. (19) and applying the Cauchy–Schwarz inequality yields

$$\begin{aligned}&\mathbb{E}\left|\Gamma_a(T_{\delta_D}(x)) - \Gamma_a(T_{\delta_a}(x))\right|\\ &\le \eta\|Ha\|_2\|H\|_2\mathbb{E}\|\delta_D - \delta_a\|_2\\ &\le \eta\|Ha\|_2\|H\|_2\sqrt{\mathbb{E}\|\delta_D - \delta_a\|_2^2}\\ &= \eta\rho\|Ha\|_2\|H\|_2 Q_D = B_D,\end{aligned} \tag{23}$$

where $Q_D = \frac{\sqrt{\mathbb{E}\|\delta_D - \delta_a\|_2^2}}{\rho}$ is the normalized root-mean-square distance to the oracle-aligned perturbation, and $B_D = \eta\rho\|Ha\|_2\|H\|_2 Q_D$ is the corresponding expected sensitivity error bound. A smaller Q_D therefore yields a tighter upper bound, motivating the comparison between the restricted cone and the full SAM sphere.

Proposition 1 (Restated). Assume $\|a\|_2 > \varepsilon$, $\|d_t\|_2 \geq \varepsilon$, and $\langle \hat{d}_t, u \rangle > 0$, where $u = a/\|a\|_2$ and $\hat{d}_t = d_t/\|d_t\|_2$. Let $S_\rho = \{\delta : \|\delta\|_2 = \rho\}$ denote the full SAM sphere and $C_t^\rho = \mathcal{D}_t \cap S_\rho$ denote the restricted cone section. For each $\delta \in C_t^\rho$, write $\delta = s\hat{d}_t + r$, where $s \geq 0$ and $r \perp \hat{d}_t$. If the residual section is centrally symmetric and the subset with $s > 0$ has positive normalized surface measure, then

$$Q_{C_t^\rho} < Q_{S_\rho}. \tag{24}$$

Furthermore, if $\|Ha\|_2 > 0$, then $B_{C_t^\rho} < B_{S_\rho}$.

Proof. We compare the normalized expected distances Q_{S_ρ} and $Q_{C_t^\rho}$ introduced above. Since $\|a\|_2 > \varepsilon$, we have $\delta_a = \rho u$ and $\|u\|_2 = 1$. Moreover, every perturbation sampled from either S_ρ or C_t^ρ has norm ρ. Expanding the squared distance gives

$$\begin{aligned} Q_D^2 &= \frac{1}{\rho^2}\mathbb{E}\|\delta_D - \delta_a\|_2^2 \\ &= \frac{1}{\rho^2}\mathbb{E}\left[\|\delta_D\|_2^2 + \|\delta_a\|_2^2 - 2\langle \delta_D, \delta_a \rangle\right] \\ &= 2 - \frac{2}{\rho}\langle \mathbb{E}\delta_D, u \rangle. \end{aligned} \tag{25}$$

For the full sphere S_ρ, symmetry about the origin implies $\mathbb{E}\delta_{S_\rho} = 0$. Consequently, Eq. (25) gives $Q_{S_\rho}^2 = 2$. For the restricted cone section C_t^ρ, each perturbation admits the decomposition $\delta = s\hat{d}_t + r$, with $s \geq 0$ and $r \perp \hat{d}_t$. By the central symmetry of the residual section, the reflection $s\hat{d}_t + r \mapsto s\hat{d}_t - r$ preserves the feasible section and its surface measure. Thus, the positive and negative residual components cancel in expectation, yielding $\mathbb{E}[r] = 0$. Let $m = \mathbb{E}[s]$. Since $s \geq 0$ and the subset with $s > 0$ has positive normalized surface measure, $m > 0$. Therefore, the mean perturbation in the cone is $\mathbb{E}\delta_{C_t^\rho} = \mathbb{E}[s]\hat{d}_t + \mathbb{E}[r] = m\hat{d}_t$.

Substituting this result into Eq. (25), we obtain

$$Q_{C_t^\rho}^2 = 2 - \frac{2m}{\rho}\langle \hat{d}_t, u \rangle < 2 = Q_{S_\rho}^2, \tag{26}$$

where the strict inequality follows from $m > 0$ and $\langle \hat{d}_t, u \rangle > 0$. Since both distances are nonnegative, $Q_{C_t^\rho} < Q_{S_\rho}$. Finally, let $\kappa = \eta\rho\|Ha\|_2\|H\|_2 > 0$. By the definition of B_D,

$$B_{C_t^\rho} = \kappa Q_{C_t^\rho} < \kappa Q_{S_\rho} = B_{S_\rho}. \tag{27}$$

This completes the proof of Proposition 1.

Remark. Theorem 1 establishes an aggregation-sensitivity error bound, while Proposition 1 shows that the restricted cone provides a tighter expected bound than the full SAM sphere under informative historical alignment. This comparison concerns uniformly sampled reference perturbations rather than the adversarial perturbations generated by the optimizers, and does not directly guarantee improved aggregation performance.

D Convergence Analysis

D.1 Theoretical Analysis

We next provide a convergence guarantee for Restricted SAM with server momentum under partial client participation. Following standard federated optimization analysis, we consider the original global objective $F(\theta) = \frac{1}{N}\sum_{i=1}^N F_i(\theta)$, where N is the total number of clients. At round t, a subset $S_t \subseteq [N]$ with $|S_t| = S$ is sampled uniformly without replacement. Each participating client performs K local Restricted-SAM updates with local stepsize η_l and perturbation radius ρ, and the server applies momentum to the averaged client increment to obtain θ^{t+1}. Throughout the analysis, we assume that each F_i is L-smooth, the data heterogeneity is bounded by σ_g^2, the outer stochastic gradient satisfies the sequential conditional unbiasedness and variance conditions below, and client sampling is independent of the potential local stochastic trajectories given the preceding round history. The assumptions, auxiliary lemmas, and complete proof are provided below.

Theorem 2 (Convergence with server momentum)**.** *Suppose the assumptions below hold, $K \geq 2$, $F^\star > -\infty$, and the fixed server momentum coefficient satisfies $0 \leq \beta_s < 1$. Initialize the momentum buffer at zero before the warm-up round. Under uniform sampling without replacement, let*

$$q = 1 - \beta_s, \qquad 0 < \eta_l \leq \frac{q^2}{60KL}, \qquad B = \sigma_l^2 + 6L^2\rho^2 + 3\sigma_g^2.$$

Then for any $T \geq 1$, the iterates generated by Restricted SAM with server momentum satisfy

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(\theta^t)\|^2 \leq C_{\beta_s}\left[\frac{F(\theta^0) - F^\star}{\eta_l KT} + \frac{L\eta_l}{S}\sigma_l^2 + LK\eta_l\sigma_g^2 + L^2K^2\eta_l^2 B + L^2\rho^2\right],$$

where $C_{\beta_s} = 250/(1-\beta_s)^2$ is a sufficient, non-optimized constant. For fixed $\beta_s < 1$, choosing $\eta_l = c(1-\beta_s)^2/(LK\sqrt{T})$ with $0 < c \leq 1/60$ and $\rho = c_\rho/\sqrt{T}$ yields

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(\theta^t)\|^2 = O(T^{-1/2}),$$

with all other problem parameters fixed.

Remark. Theorem 2 preserves the same-order non-convex convergence guarantee as FedAvg and FedSAM. Server momentum changes the sufficient stepsize condition and the constants; the guarantee is not uniform as $\beta_s \to 1$. The restricted perturbation continues to enter only through its radius bound. The detailed bound and proof are given in Theorem 3 below. The analysis starts before the FedSAM warm-up round: that round also satisfies the bounded-perturbation conditions, and its update initializes $m_1 = \theta^1 - \theta^0$ as in Algorithm 1.

In this section, we study the convergence behavior of the proposed restricted-SAM local updater under partial client participation. Unlike analyses that reinterpret the algorithm as directly optimizing a global sharpness-aware surrogate objective, our analysis keeps the original federated objective

$$F(w) := \frac{1}{N}\sum_{i=1}^{N} F_i(w) \tag{28}$$

unchanged, and treats the restricted perturbation only as a mechanism for constructing the local update direction. Therefore, the main analytical task is to control how the restricted perturbation affects the local model drift, the aggregated update, and the resulting one-round descent of the original global objective. The convergence guarantee applies to the sequential stochastic gradient oracle in Assumption 3. The practical implementation may reuse the same mini-batch for efficiency. Since the perturbation can depend on that mini-batch, this reuse does not automatically satisfy the stated conditional unbiasedness assumption; applying the guarantee to it requires establishing that condition or extending the analysis to control the bias.

Throughout this section, we consider a federated system with N clients. At communication round r, a subset $S_r \subseteq [N]$ of cardinality $|S_r| = S$ is sampled uniformly without replacement. Let $I_i^r := \mathbf{1}\{i \in S_r\}, \qquad i = 1, \ldots, N$, denote the corresponding sampling indicator. Then

$$\mathbb{E}[I_i^r] = \frac{S}{N}, \qquad \mathbb{E}[I_i^r I_j^r] = \frac{S(S-1)}{N(N-1)}, \quad i \neq j. \tag{29}$$

For theoretical convenience, we define the local trajectory of *every* client from the common global model w^r, regardless of whether the client is eventually selected for aggregation. Specifically, for each client i and each local step $k = 0, \ldots, K-1$, we define

$$w_{i,0}^r = w^r, \qquad \widetilde{w}_{i,k}^r = w_{i,k}^r + \delta_{i,k}^r, \qquad w_{i,k+1}^r = w_{i,k}^r - \eta_l g_{i,k}^r, \tag{30}$$

where $g_{i,k}^r := \nabla F_i(\widetilde{w}_{i,k}^r; \xi_{i,k}^r)$, $\delta_{i,k}^r \in D_r$, $\|\delta_{i,k}^r\| \leq \rho$. Here, $\eta_l > 0$ denotes the local stepsize, $\xi_{i,k}^r$ is the local stochastic sample, and D_r is the feasible restricted perturbation set at round r.

We define the cumulative local update of client i at round r as $\Delta_i^r := w_{i,K}^r - w^r = -\eta_l \sum_{k=0}^{K-1} g_{i,k}^r$. Under partial participation, the aggregated update is given by $\Delta^r := \frac{1}{S}\sum_{i=1}^N I_i^r \Delta_i^r = -\frac{\eta_l}{S}\sum_{i=1}^N I_i^r \sum_{k=0}^{K-1} g_{i,k}^r$, and the server updates are

$$m_{r+1} = \beta_s m_r + \Delta^r, \qquad w^{r+1} = w^r + m_{r+1}, \qquad m_0 = 0.$$

Since $m_r = w^r - w^{r-1}$ for $r \geq 1$, this is equivalent to the single server update in Algorithm 1.

For any random quantity X, we write $\mathbb{E}_r[X] := \mathbb{E}[X \mid \mathcal{F}_r]$ for the conditional expectation given the history $\mathcal{F}_r$ before client sampling in round r, including the global model, momentum, and previously constructed perturbation region. The sequential information sets below include this round history. We also denote $G_r^2 := \|\nabla F(w^r)\|^2$, $E_w^{r,k} := \frac{1}{N}\sum_{i=1}^N \mathbb{E}_r\|w_{i,k}^r - w^r\|^2$.

We adopt the following assumptions throughout the analysis.

Assumption 1 (Smoothness)**.** *Each local objective F_i is L-smooth, i.e.,*

$$\|\nabla F_i(x) - \nabla F_i(y)\| \leq L\|x - y\|, \qquad \forall x, y \in \mathbb{R}^d,\ \forall i \in [N]. \tag{31}$$

Assumption 2 (Bounded heterogeneity)**.** *There exists $\sigma_g^2 \geq 0$ such that*

$$\|\nabla F_i(w) - \nabla F(w)\|^2 \leq \sigma_g^2, \qquad \forall w \in \mathbb{R}^d,\ \forall i \in [N]. \tag{32}$$

Assumption 3 (Sequential stochastic gradient oracle)**.** *Within each round r, fix an ordering of the potential local gradient evaluations (i, k) that preserves each client's local-step order. Let $\mathcal{F}_{r,i,k}$ denote the information available immediately before drawing $\xi_{i,k}^r$ for the outer model update. These information sets form an increasing filtration along this ordering and contain $\mathcal{F}_r$, all preceding oracle outputs, and the current local iterate $w_{i,k}^r$ and its constructed perturbation $\delta_{i,k}^r$. Define $\zeta_{i,k}^r := g_{i,k}^r - \nabla F_i(\widetilde{w}_{i,k}^r)$. We assume that, for some finite $\sigma_l^2 \geq 0$ and every i, k, r,*

$$\mathbb{E}[\zeta_{i,k}^r \mid \mathcal{F}_{r,i,k}] = 0, \qquad \mathbb{E}[\|\zeta_{i,k}^r\|^2 \mid \mathcal{F}_{r,i,k}] \leq \sigma_l^2. \tag{33}$$

Thus, the stochastic noises form martingale differences over the sequence of local gradient evaluations across clients and local steps.

Assumption 4 (Independence between client sampling and potential local trajectories)**.** *Conditioned on $\mathcal{F}_r$, the random subset S_r is independent of the potential local stochastic trajectories of all clients, including the randomness used to construct their perturbations and outer stochastic gradients.*

The first step of the proof is to control the local model drift $E_w^{r,k}$, which will later be used to bound both the aggregated update and the one-round descent of the global objective.

Lemma 1 (Local model drift bound)**.** *Suppose Assumptions 1–3 hold. If $\eta_l \leq \frac{1}{10KL}$, then for any round r and any local step $k = 0, 1, \ldots, K$,*

$$E_w^{r,k} := \frac{1}{N}\sum_{i=1}^N \mathbb{E}_r\|w_{i,k}^r - w^r\|^2 \leq 20K^2\eta_l^2\left(\sigma_l^2 + 6L^2\rho^2 + 3\sigma_g^2 + 3\|\nabla F(w^r)\|^2\right). \tag{34}$$

Proof. Recall that the local update on client i is $w_{i,k+1}^r = w_{i,k}^r - \eta_l g_{i,k}^r, g_{i,k}^r = \nabla F_i(\widetilde{w}_{i,k}^r; \xi_{i,k}^r), \widetilde{w}_{i,k}^r = w_{i,k}^r + \delta_{i,k}^r$. Let $d_{i,k}^r := w_{i,k}^r - w^r$. Then $d_{i,0}^r = 0$, and

$$\begin{aligned}
\mathbb{E}_r\|d_{i,k+1}^r\|^2 &= \mathbb{E}_r\|d_{i,k}^r - \eta_l g_{i,k}^r\|^2 \\
&\le \left(1 + \frac{1}{K-1}\right)\mathbb{E}_r\|d_{i,k}^r\|^2 + K\eta_l^2\mathbb{E}_r\|g_{i,k}^r\|^2 && \text{(a)} \\
&\le \left(1 + \frac{1}{K-1}\right)\mathbb{E}_r\|d_{i,k}^r\|^2 + K\eta_l^2\Big(2\mathbb{E}_r\|g_{i,k}^r - \nabla F_i(\widetilde{w}_{i,k}^r)\|^2 + 2\mathbb{E}_r\|\nabla F_i(\widetilde{w}_{i,k}^r)\|^2\Big) && \text{(b)} \\
&\le \left(1 + \frac{1}{K-1}\right)\mathbb{E}_r\|d_{i,k}^r\|^2 + 2K\eta_l^2\sigma_l^2 + 2K\eta_l^2\mathbb{E}_r\|\nabla F_i(\widetilde{w}_{i,k}^r)\|^2 && \text{(c)} \\
&\le \left(1 + \frac{1}{K-1}\right)\mathbb{E}_r\|d_{i,k}^r\|^2 + 2K\eta_l^2\sigma_l^2 + 2K\eta_l^2\Big(3\mathbb{E}_r\|\nabla F_i(\widetilde{w}_{i,k}^r) - \nabla F_i(w^r)\|^2 \\
&\quad + 3\|\nabla F_i(w^r) - \nabla F(w^r)\|^2 + 3\|\nabla F(w^r)\|^2\Big) && \text{(d)} \\
&\le \left(1 + \frac{1}{K-1}\right)\mathbb{E}_r\|d_{i,k}^r\|^2 + 2K\eta_l^2\sigma_l^2 \\
&\quad + 2K\eta_l^2\Big(3L^2\mathbb{E}_r\|\widetilde{w}_{i,k}^r - w^r\|^2 + 3\sigma_g^2 + 3\|\nabla F(w^r)\|^2\Big) && \text{(e)} \\
&\le \left(1 + \frac{1}{K-1}\right)\mathbb{E}_r\|d_{i,k}^r\|^2 + 2K\eta_l^2\sigma_l^2 \\
&\quad + 2K\eta_l^2\Big(3L^2\big(2\mathbb{E}_r\|d_{i,k}^r\|^2 + 2\rho^2\big) + 3\sigma_g^2 + 3\|\nabla F(w^r)\|^2\Big) && \text{(f)} \\
&= \left(1 + \frac{1}{K-1} + 12K\eta_l^2L^2\right)\mathbb{E}_r\|d_{i,k}^r\|^2 \\
&\quad + 2K\eta_l^2\left(\sigma_l^2 + 6L^2\rho^2 + 3\sigma_g^2 + 3\|\nabla F(w^r)\|^2\right). && \text{(g)}
\end{aligned}$$

Here, (a) follows from the inequality $\|x-y\|^2 \le (1+\gamma)\|x\|^2 + (1+\gamma^{-1})\|y\|^2$ with $\gamma = \frac{1}{K-1}$. Step (b) is obtained by decomposing $g_{i,k}^r$ into $(g_{i,k}^r - \nabla F_i(\widetilde{w}_{i,k}^r)) + \nabla F_i(\widetilde{w}_{i,k}^r)$ and applying $\|x+y\|^2 \le 2\|x\|^2 + 2\|y\|^2$. Step (c) uses Assumption 3. Step (d) follows from the three-term decomposition of $\nabla F_i(\widetilde{w}_{i,k}^r)$ around $\nabla F_i(w^r)$ and $\nabla F(w^r)$ together with $\|x+y+z\|^2 \le 3\|x\|^2 + 3\|y\|^2 + 3\|z\|^2$. Step (e) uses Assumptions 1 and 2. Step (f) follows from $\widetilde{w}_{i,k}^r - w^r = (w_{i,k}^r - w^r) + \delta_{i,k}^r$, the inequality $\|x+y\|^2 \le 2\|x\|^2 + 2\|y\|^2$, and the perturbation bound $\|\delta_{i,k}^r\| \le \rho$. Finally, (g) is obtained by collecting the terms involving $\mathbb{E}_r\|d_{i,k}^r\|^2$ and the remaining constants.

Averaging (g) over all clients yields

$$\begin{aligned}
E_w^{r,k+1} \le &\left(1 + \frac{1}{K-1} + 12K\eta_l^2L^2\right)E_w^{r,k} \\
&+ 2K\eta_l^2\left(\sigma_l^2 + 6L^2\rho^2 + 3\sigma_g^2 + 3\|\nabla F(w^r)\|^2\right). && \text{(h)}
\end{aligned}$$

That is, the averaged local drift satisfies a linear recursion.

Now define

$$c := \frac{1}{K-1} + 12K\eta_l^2L^2, \qquad B_r := \sigma_l^2 + 6L^2\rho^2 + 3\sigma_g^2 + 3\|\nabla F(w^r)\|^2. \tag{35}$$

Then (h) can be rewritten as

$$E_w^{r,k+1} \le (1+c)E_w^{r,k} + 2K\eta_l^2B_r. \tag{36}$$

Since $E_w^{r,0} = 0$, iterating the above recursion gives

$$E_w^{r,k} \le 2K\eta_l^2B_r\sum_{\tau=0}^{k-1}(1+c)^\tau = 2K\eta_l^2B_r \cdot \frac{(1+c)^k - 1}{c}. \tag{i}$$

Here (i) follows directly from the geometric-series expansion of the recursion with the initial condition $E_w^{r,0} = 0$.

Next, by the stepsize condition $\eta_l \leq \frac{1}{10KL}$, we have $12K\eta_l^2 L^2 \leq \frac{12}{100K} = \frac{3}{25K}$, and hence $c \leq \frac{1}{K-1} + \frac{3}{25K}$. Therefore,

$$cK \leq \frac{K}{K-1} + \frac{3}{25} \leq 2 + \frac{3}{25} < 2.2, \qquad (1+c)^k \leq (1+c)^K \leq e^{cK} < 10. \tag{j}$$

Moreover, $c \geq \frac{1}{K-1}$, $\frac{1}{c} \leq K-1 \leq K$. In other words, (j) provides a uniform upper bound on the geometric growth factor, while (k) gives a convenient bound on the reciprocal $1/c$.

Substituting (j) and (k) into (i), we obtain

$$E_w^{r,k} \leq 2K\eta_l^2 B_r \cdot \frac{10}{c} \leq 20K\eta_l^2 B_r \cdot K = 20K^2\eta_l^2 \left(\sigma_l^2 + 6L^2\rho^2 + 3\sigma_g^2 + 3\|\nabla F(w^r)\|^2\right). \tag{37}$$

This completes the proof. □

Having established the local model drift bound in Lemma 1, we next turn to the additional technical difficulty introduced by partial participation. Under full participation, the aggregated update is simply the average over all clients, whereas under partial participation it is formed from a uniformly sampled subset of clients. Therefore, before bounding the sampled aggregated update itself, we first need a precise second-moment characterization of the sampled average. The following lemma provides such an identity for uniform sampling without replacement, and will serve as a basic tool in the subsequent analysis of the aggregation error and the one-round descent.

Lemma 2 (Second-moment identity for uniform sampling without replacement)**.** *Let $\{a_i\}_{i=1}^N \subset \mathbb{R}^d$ be a collection of deterministic vectors, and let I_i denote the indicator that client i is selected when a subset of size S is sampled uniformly without replacement from $[N]$. Then*

$$\mathbb{E}\left\|\frac{1}{S}\sum_{i=1}^N I_i a_i\right\|^2 = \frac{N-S}{S(N-1)} \cdot \frac{1}{N}\sum_{i=1}^N \|a_i\|^2 + \frac{N(S-1)}{S(N-1)} \cdot \left\|\frac{1}{N}\sum_{i=1}^N a_i\right\|^2. \tag{38}$$

Proof. By definition,

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{S}\sum_{i=1}^N I_i a_i\right\|^2 &= \frac{1}{S^2}\mathbb{E}\left\langle \sum_{i=1}^N I_i a_i, \sum_{j=1}^N I_j a_j \right\rangle \\
&= \frac{1}{S^2}\mathbb{E}\left[\sum_{i=1}^N I_i \|a_i\|^2 + \sum_{i\neq j} I_i I_j \langle a_i, a_j\rangle\right] && \text{(a)} \\
&= \frac{1}{S^2}\left(\sum_{i=1}^N \mathbb{E}[I_i]\|a_i\|^2 + \sum_{i\neq j} \mathbb{E}[I_i I_j]\langle a_i, a_j\rangle\right) && \text{(b)} \\
&= \frac{1}{S^2}\left(\frac{S}{N}\sum_{i=1}^N \|a_i\|^2 + \frac{S(S-1)}{N(N-1)}\sum_{i\neq j}\langle a_i, a_j\rangle\right) && \text{(c)} \\
&= \frac{1}{S^2}\left(\frac{S}{N}\sum_{i=1}^N \|a_i\|^2 + \frac{S(S-1)}{N(N-1)}\left(\left\|\sum_{i=1}^N a_i\right\|^2 - \sum_{i=1}^N \|a_i\|^2\right)\right) && \text{(d)} \\
&= \left(\frac{1}{SN} - \frac{S-1}{SN(N-1)}\right)\sum_{i=1}^N \|a_i\|^2 + \frac{S-1}{SN(N-1)}\left\|\sum_{i=1}^N a_i\right\|^2 && \text{(e)} \\
&= \frac{N-S}{SN(N-1)}\sum_{i=1}^N \|a_i\|^2 + \frac{S-1}{SN(N-1)}\left\|\sum_{i=1}^N a_i\right\|^2 && \text{(f)} \\
&= \frac{N-S}{S(N-1)} \cdot \frac{1}{N}\sum_{i=1}^N \|a_i\|^2 + \frac{N(S-1)}{S(N-1)} \cdot \left\|\frac{1}{N}\sum_{i=1}^N a_i\right\|^2. && \text{(g)}
\end{aligned}$$

Here, (a) expands the squared norm into diagonal and off-diagonal terms. Step (b) follows by linearity of expectation. Step (c) uses the standard moments of uniform sampling without replacement, namely $\mathbb{E}[I_i] = \frac{S}{N}$ and $\mathbb{E}[I_i I_j] = \frac{S(S-1)}{N(N-1)}$ for $i \neq j$. Step (d) uses the identity $\sum_{i \neq j} \langle a_i, a_j \rangle = \left\| \sum_{i=1}^N a_i \right\|^2 - \sum_{i=1}^N \|a_i\|^2$. Step (e) is obtained by grouping the coefficients of $\sum_i \|a_i\|^2$ and $\|\sum_i a_i\|^2$, while (f) is a direct simplification of the coefficient

$$\frac{1}{SN} - \frac{S-1}{SN(N-1)} = \frac{N-S}{SN(N-1)}.$$

Finally, (g) rewrites the expression in terms of the empirical second moment $\frac{1}{N}\sum_i \|a_i\|^2$ and the squared norm of the empirical mean $\left\| \frac{1}{N}\sum_i a_i \right\|^2$, which is the desired form. □

Lemma 3 (Second-moment bound of the averaged gradient shift)**.** *Define*

$$A_r := \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbb{E}_r \left[\nabla F_i(\widetilde{w}_{i,k}^r) - \nabla F_i(w^r) \right]. \tag{39}$$

Suppose Assumptions 1–3 hold. If $\eta_l \leq \frac{1}{10KL}$, then

$$\|A_r\|^2 \leq 40K^4 \eta_l^2 L^2 \left(\sigma_l^2 + 6L^2\rho^2 + 3\sigma_g^2 + 3\|\nabla F(w^r)\|^2 \right) + 2K^2L^2\rho^2. \tag{40}$$

Proof. Let

$$a_{i,k}^r := \mathbb{E}_r \left[\nabla F_i(\widetilde{w}_{i,k}^r) - \nabla F_i(w^r) \right]. \tag{41}$$

Then

$$A_r = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} a_{i,k}^r. \tag{42}$$

Hence

$$\begin{aligned}
\|A_r\|^2 &= \left\| \sum_{k=0}^{K-1} \frac{1}{N} \sum_{i=1}^{N} a_{i,k}^r \right\|^2 \\
&\leq K \sum_{k=0}^{K-1} \left\| \frac{1}{N} \sum_{i=1}^{N} a_{i,k}^r \right\|^2 && \text{(a)} \\
&\leq K \sum_{k=0}^{K-1} \frac{1}{N} \sum_{i=1}^{N} \|a_{i,k}^r\|^2 && \text{(b)} \\
&= K \sum_{k=0}^{K-1} \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbb{E}_r \left[\nabla F_i(\widetilde{w}_{i,k}^r) - \nabla F_i(w^r) \right] \right\|^2 \\
&\leq K \sum_{k=0}^{K-1} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_r \left\| \nabla F_i(\widetilde{w}_{i,k}^r) - \nabla F_i(w^r) \right\|^2 && \text{(c)} \\
&\leq K \sum_{k=0}^{K-1} \frac{1}{N} \sum_{i=1}^{N} L^2 \, \mathbb{E}_r \|\widetilde{w}_{i,k}^r - w^r\|^2 && \text{(d)} \\
&= KL^2 \sum_{k=0}^{K-1} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_r \|(w_{i,k}^r - w^r) + \delta_{i,k}^r\|^2 \\
&\leq KL^2 \sum_{k=0}^{K-1} \frac{1}{N} \sum_{i=1}^{N} \left(2\mathbb{E}_r \|w_{i,k}^r - w^r\|^2 + 2\|\delta_{i,k}^r\|^2 \right) && \text{(e)} \\
&\leq 2KL^2 \sum_{k=0}^{K-1} E_w^{r,k} + 2K^2L^2\rho^2. && \text{(f)}
\end{aligned}$$

Now, by Lemma 1,

$$E_w^{r,k} \le 20K^2\eta_l^2 \left(\sigma_l^2 + 6L^2\rho^2 + 3\sigma_g^2 + 3\|\nabla F(w^r)\|^2\right), \qquad \forall k = 0, \dots, K. \tag{43}$$

Summing over $k = 0, \dots, K-1$ gives

$$\sum_{k=0}^{K-1} E_w^{r,k} \le 20K^3\eta_l^2 \left(\sigma_l^2 + 6L^2\rho^2 + 3\sigma_g^2 + 3\|\nabla F(w^r)\|^2\right). \tag{44}$$

Substituting the above into (f) yields

$$\|A_r\|^2 \le 40K^4\eta_l^2 L^2 \left(\sigma_l^2 + 6L^2\rho^2 + 3\sigma_g^2 + 3\|\nabla F(w^r)\|^2\right) + 2K^2L^2\rho^2. \tag{45}$$

Here, (a) uses the inequality $\|\sum_{k=0}^{K-1} u_k\|^2 \le K\sum_{k=0}^{K-1}\|u_k\|^2$. Step (b) follows from Jensen's inequality applied to the client average. Step (c) is again Jensen's inequality, which moves the squared norm inside the conditional expectation. Step (d) uses Assumption 1. Step (e) follows from $\widetilde{w}_{i,k}^r - w^r = (w_{i,k}^r - w^r) + \delta_{i,k}^r$ and $\|x+y\|^2 \le 2\|x\|^2 + 2\|y\|^2$. Step (f) uses the perturbation bound $\|\delta_{i,k}^r\| \le \rho$ and the definition of $E_w^{r,k}$. Finally, the last step substitutes the local drift bound from Lemma 1. This completes the proof. □

Lemma 4 (Sharpened control of the inner product term)**.** *Under the conditions of Lemma 3, we have*

$$\begin{aligned} &-\eta_l \langle \nabla F(w^r), A_r \rangle \\ &\le \frac{\eta_l K}{4}\|\nabla F(w^r)\|^2 + 40K^3\eta_l^3 L^2 \left(\sigma_l^2 + 6L^2\rho^2 + 3\sigma_g^2 + 3\|\nabla F(w^r)\|^2\right) + 2K\eta_l L^2\rho^2. \end{aligned} \tag{46}$$

Proof. By Young's inequality, for any vectors $x, y \in \mathbb{R}^d$ and any $\beta > 0$, $|\langle x, y\rangle| \le \frac{\beta}{2}\|x\|^2 + \frac{1}{2\beta}\|y\|^2$. Applying this with $x = \nabla F(w^r), y = A_r, \beta = \frac{K}{2}$, we obtain

$$-\eta_l\langle \nabla F(w^r), A_r\rangle \le \eta_l |\langle \nabla F(w^r), A_r\rangle| \le \eta_l \left(\frac{K}{4}\|\nabla F(w^r)\|^2 + \frac{1}{K}\|A_r\|^2\right) \tag{a}$$

$$= \frac{\eta_l K}{4}\|\nabla F(w^r)\|^2 + \frac{\eta_l}{K}\|A_r\|^2. \tag{b}$$

Now invoking Lemma 3, we have

$$\frac{\eta_l}{K}\|A_r\|^2 \le \frac{\eta_l}{K}\left[40K^4\eta_l^2 L^2 \left(\sigma_l^2 + 6L^2\rho^2 + 3\sigma_g^2 + 3\|\nabla F(w^r)\|^2\right) + 2K^2L^2\rho^2\right] \tag{c}$$

$$= 40K^3\eta_l^3 L^2 \left(\sigma_l^2 + 6L^2\rho^2 + 3\sigma_g^2 + 3\|\nabla F(w^r)\|^2\right) + 2K\eta_l L^2\rho^2. \tag{d}$$

Substituting (d) into (b) yields

$$\begin{aligned} &-\eta_l \langle \nabla F(w^r), A_r \rangle \\ &\le \frac{\eta_l K}{4}\|\nabla F(w^r)\|^2 + 40K^3\eta_l^3 L^2 \left(\sigma_l^2 + 6L^2\rho^2 + 3\sigma_g^2 + 3\|\nabla F(w^r)\|^2\right) + 2K\eta_l L^2\rho^2. \end{aligned} \tag{47}$$

Here, (a) applies Young's inequality with the specific choice $\beta = \frac{K}{2}$, which is convenient for later matching the dominant descent term $-\eta_l K\|\nabla F(w^r)\|^2$. Step (b) simply rewrites the result by collecting the coefficients. Step (c) invokes the second-moment bound of A_r from Lemma 3, and (d) expands the prefactor $\frac{\eta_l}{K}$ term by term. This completes the proof. □

Lemma 5 (Partial participation aggregated update bound)**.** *Suppose Assumptions 1–3 and 4 hold. Assume $K \ge 2$. If $\eta_l \le \frac{1}{10KL}$, then the aggregated update under uniform sampling without replacement satisfies*

$$\mathbb{E}_r\|\Delta^r\|^2 \le \frac{2K\eta_l^2}{S}\sigma_l^2 + 240K^4\eta_l^4 L^2\Gamma_r + 12K^2\eta_l^2 L^2\rho^2 + 6K^2\eta_l^2\sigma_g^2 + 6K^2\eta_l^2\|\nabla F(w^r)\|^2, \tag{48}$$

where $\Gamma_r := \sigma_l^2 + 6L^2\rho^2 + 3\sigma_g^2 + 3\|\nabla F(w^r)\|^2$.

Proof. Recall that the aggregated update is $\Delta^r = -\frac{\eta_l}{S}\sum_{i=1}^N I_i^r \sum_{k=0}^{K-1} g_{i,k}^r$. For each client i, define $a_i^r := \sum_{k=0}^{K-1} \nabla F_i(\widetilde{w}_{i,k}^r), \zeta_i^r := \sum_{k=0}^{K-1}\left(g_{i,k}^r - \nabla F_i(\widetilde{w}_{i,k}^r)\right)$. Then $\sum_{k=0}^{K-1} g_{i,k}^r = a_i^r + \zeta_i^r, \Delta^r = -\frac{\eta_l}{S}\sum_{i=1}^N I_i^r(a_i^r + \zeta_i^r)$. Therefore, by $\|x+y\|^2 \le 2\|x\|^2 + 2\|y\|^2$,

$$\mathbb{E}_r\|\Delta^r\|^2 \le \frac{2\eta_l^2}{S^2}\mathbb{E}_r\left\|\sum_{i=1}^N I_i^r\zeta_i^r\right\|^2 + \frac{2\eta_l^2}{S^2}\mathbb{E}_r\left\|\sum_{i=1}^N I_i^r a_i^r\right\|^2. \tag{a}$$

We bound the stochastic-noise term and the deterministic-gradient term separately.

First, consider the stochastic-noise term. For any evaluation (i,k) preceding (j,ℓ) in the ordering of Assumption 3, $\zeta_{i,k}^r$ is $\mathcal{F}_{r,j,\ell}$-measurable. The tower property and the martingale-difference condition give

$$\mathbb{E}_r\langle\zeta_{i,k}^r, \zeta_{j,\ell}^r\rangle = \mathbb{E}_r\big\langle\zeta_{i,k}^r, \mathbb{E}[\zeta_{j,\ell}^r \mid \mathcal{F}_{r,j,\ell}]\big\rangle = 0.$$

Assumption 4 allows the sampling indicators to be factored out of these noise moments conditional on $\mathcal{F}_r$. Hence

$$\begin{aligned}
\mathbb{E}_r\left\|\sum_{i=1}^N I_i^r\zeta_i^r\right\|^2 &= \sum_{i=1}^N \mathbb{E}_r[I_i^r]\,\mathbb{E}_r\|\zeta_i^r\|^2 \\
&= \frac{S}{N}\sum_{i=1}^N \mathbb{E}_r\left\|\sum_{k=0}^{K-1}\left(g_{i,k}^r - \nabla F_i(\widetilde{w}_{i,k}^r)\right)\right\|^2 \\
&= \frac{S}{N}\sum_{i=1}^N\sum_{k=0}^{K-1}\mathbb{E}_r\left\|g_{i,k}^r - \nabla F_i(\widetilde{w}_{i,k}^r)\right\|^2 \\
&\le \frac{S}{N}\sum_{i=1}^N K\sigma_l^2 = SK\sigma_l^2.
\end{aligned} \tag{b}$$

Consequently,

$$\frac{2\eta_l^2}{S^2}\mathbb{E}_r\left\|\sum_{i=1}^N I_i^r\zeta_i^r\right\|^2 \le \frac{2K\eta_l^2}{S}\sigma_l^2. \tag{c}$$

Next, consider the deterministic-gradient term. The vectors $\{a_i^r\}_{i=1}^N$ are random because the perturbed local iterates $\{\widetilde{w}_{i,k}^r\}_{i,k}$ depend on the local stochastic trajectory. Therefore, we first condition on their realized values and then apply Lemma 2. Define

$$\chi_{N,S} := \frac{N-S}{S(N-1)}, \qquad \omega_{N,S} := \frac{N(S-1)}{S(N-1)}. \tag{49}$$

Then, by Assumption 4, the tower property, and Lemma 2,

$$\begin{aligned}
\mathbb{E}_r\left\|\frac{1}{S}\sum_{i=1}^N I_i^r a_i^r\right\|^2 &= \mathbb{E}_r\left[\mathbb{E}\left[\left\|\frac{1}{S}\sum_{i=1}^N I_i^r a_i^r\right\|^2 \,\middle|\, \mathcal{F}_r, \{a_i^r\}_{i=1}^N\right]\right] \\
&= \chi_{N,S}\,\frac{1}{N}\sum_{i=1}^N \mathbb{E}_r\|a_i^r\|^2 + \omega_{N,S}\,\mathbb{E}_r\left\|\frac{1}{N}\sum_{i=1}^N a_i^r\right\|^2.
\end{aligned} \tag{d}$$

By Jensen's inequality,

$$\mathbb{E}_r\left\|\frac{1}{N}\sum_{i=1}^N a_i^r\right\|^2 \le \frac{1}{N}\sum_{i=1}^N \mathbb{E}_r\|a_i^r\|^2. \tag{e}$$

Moreover, the sampling coefficients satisfy

$$\chi_{N,S} + \omega_{N,S} = \frac{N-S}{S(N-1)} + \frac{N(S-1)}{S(N-1)} = 1. \tag{f}$$

Combining (d), (e), and (f), we obtain

$$\mathbb{E}_r \left\| \frac{1}{S} \sum_{i=1}^{N} I_i^r a_i^r \right\|^2 \leq \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_r \|a_i^r\|^2. \tag{g}$$

This step avoids equating a random client average with its conditional expectation; instead, it directly controls the second moment of the random average.

It remains to bound the right-hand side of (g). By the definition of a_i^r,

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_r \|a_i^r\|^2 = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_r \left\| \sum_{k=0}^{K-1} \nabla F_i(\widetilde{w}_{i,k}^r) \right\|^2 \leq \frac{K}{N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbb{E}_r \|\nabla F_i(\widetilde{w}_{i,k}^r)\|^2. \tag{h}$$

For each i, k, we decompose

$$\nabla F_i(\widetilde{w}_{i,k}^r) = \big(\nabla F_i(\widetilde{w}_{i,k}^r) - \nabla F_i(w^r)\big) + \big(\nabla F_i(w^r) - \nabla F(w^r)\big) + \nabla F(w^r). \tag{50}$$

Using $\|x + y + z\|^2 \leq 3\|x\|^2 + 3\|y\|^2 + 3\|z\|^2$, Assumptions 1 and 2, and the perturbation bound $\|\delta_{i,k}^r\| \leq \rho$, we have

$$\begin{aligned} \mathbb{E}_r \|\nabla F_i(\widetilde{w}_{i,k}^r)\|^2 &\leq 3L^2 \mathbb{E}_r \|\widetilde{w}_{i,k}^r - w^r\|^2 + 3\sigma_g^2 + 3\|\nabla F(w^r)\|^2 \\ &\leq 6L^2 \mathbb{E}_r \|w_{i,k}^r - w^r\|^2 + 6L^2\rho^2 + 3\sigma_g^2 + 3\|\nabla F(w^r)\|^2. \end{aligned} \tag{i}$$

Substituting (i) into (h) gives

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_r \|a_i^r\|^2 \leq 6KL^2 \sum_{k=0}^{K-1} E_w^{r,k} + K^2 \left(6L^2\rho^2 + 3\sigma_g^2 + 3\|\nabla F(w^r)\|^2\right). \tag{j}$$

By Lemma 1, $E_w^{r,k} \leq 20K^2\eta_l^2\Gamma_r, k = 0, \ldots, K$. Therefore,

$$\sum_{k=0}^{K-1} E_w^{r,k} \leq 20K^3\eta_l^2\Gamma_r. \tag{k}$$

Substituting (k) into (j) yields

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_r \|a_i^r\|^2 \leq 120K^4\eta_l^2 L^2\Gamma_r + K^2 \left(6L^2\rho^2 + 3\sigma_g^2 + 3\|\nabla F(w^r)\|^2\right). \tag{l}$$

Combining (g) and (l), we get

$$\begin{aligned} \frac{2\eta_l^2}{S^2} \mathbb{E}_r \left\| \sum_{i=1}^{N} I_i^r a_i^r \right\|^2 &= 2\eta_l^2 \mathbb{E}_r \left\| \frac{1}{S} \sum_{i=1}^{N} I_i^r a_i^r \right\|^2 \\ &\leq 240K^4\eta_l^4 L^2\Gamma_r + 12K^2\eta_l^2 L^2\rho^2 + 6K^2\eta_l^2\sigma_g^2 + 6K^2\eta_l^2\|\nabla F(w^r)\|^2. \end{aligned} \tag{m}$$

Finally, substituting the stochastic-noise bound (c) and the deterministic-gradient bound (m) into (a) gives

$$\mathbb{E}_r \|\Delta^r\|^2 \leq \frac{2K\eta_l^2}{S}\sigma_l^2 + 240K^4\eta_l^4 L^2\Gamma_r + 12K^2\eta_l^2 L^2\rho^2 + 6K^2\eta_l^2\sigma_g^2 + 6K^2\eta_l^2\|\nabla F(w^r)\|^2. \tag{51}$$

This completes the proof. □

Lemma 6 (Momentum energy and auxiliary sequence)**.** *Let* $q = 1 - \beta_s > 0$, $m_0 = 0$, *and* $m_{r+1} = \beta_s m_r + \Delta^r$. *Define*

$$y^r = w^r + \frac{\beta_s}{q} m_r.$$

Then

$$y^{r+1} = y^r + \frac{\Delta^r}{q}, \qquad \sum_{r=0}^{R-1} \mathbb{E}\|m_r\|^2 \leq \frac{1}{q^2} \sum_{r=0}^{R-1} \mathbb{E}\|\Delta^r\|^2. \tag{52}$$

Proof. Since $w^{r+1} = w^r + m_{r+1}$, direct substitution gives

$$y^{r+1} - y^r = \frac{m_{r+1} - \beta_s m_r}{q} = \frac{\Delta^r}{q}.$$

Convexity of the squared norm gives the pathwise inequality

$$\|m_{r+1}\|^2 = \left\|\beta_s m_r + q\frac{\Delta^r}{q}\right\|^2 \le \beta_s \|m_r\|^2 + \frac{\|\Delta^r\|^2}{q}.$$

Summing over $r = 0, \dots, R-1$, using $m_0 = 0$, and discarding $\|m_R\|^2 \ge 0$ proves the second claim. No bound on the gradient norm is required. □

Lemma 7 (Partial participation descent with server momentum)**.** *Suppose Assumptions 1–3 and 4 hold, $K \ge 2$, and $\eta_l \le 1/(10KL)$. Write*

$$a = K\eta_l, \qquad h = La, \qquad q = 1 - \beta_s, \qquad B = \sigma_l^2 + 6L^2\rho^2 + 3\sigma_g^2,$$
$$Q_0 = \frac{2\sigma_l^2}{SK} + 240h^2B + 12L^2\rho^2 + 6\sigma_g^2, \qquad Q_1 = 6 + 720h^2. \tag{53}$$

For the auxiliary sequence in Lemma 6,

$$\mathbb{E}_r F(y^{r+1}) \le F(y^r) - \frac{a}{q}\left(\frac{1}{2} - 180h^2\right) G_r^2 + \frac{a}{q}\left(60h^2B + 3L^2\rho^2\right) + \frac{3aL^2\beta_s^2}{2q^3}\|m_r\|^2 + \frac{La^2}{2q^2}\left(Q_0 + Q_1 G_r^2\right). \tag{54}$$

Proof. Let $g_r = \nabla F(w^r)$ and $e_r = A_r/K$, where A_r is the averaged gradient shift in Lemma 3. Uniform client sampling, Assumption 4, and the tower property applied to Assumption 3 imply

$$\mathbb{E}_r \Delta^r = -a(g_r + e_r).$$

The existing gradient-shift and aggregated-update bounds give

$$\|e_r\|^2 \le 40h^2(B + 3G_r^2) + 2L^2\rho^2,$$
$$\mathbb{E}_r\|\Delta^r\|^2 \le a^2(Q_0 + Q_1 G_r^2). \tag{55}$$

These bounds only concern the current round's local trajectories initialized at w^r; their derivations do not require a particular server update in previous rounds.

Define $b_r = \nabla F(y^r) - \nabla F(w^r)$. Smoothness yields

$$\|b_r\|^2 \le \frac{L^2\beta_s^2}{q^2}\|m_r\|^2.$$

Applying Young's inequality to the three cross terms gives

$$-\langle g_r + b_r, g_r + e_r\rangle \le -G_r^2 + \frac{1}{4}G_r^2 + \|e_r\|^2 + \frac{1}{4}G_r^2 + \|b_r\|^2 + \frac{1}{2}\|b_r\|^2 + \frac{1}{2}\|e_r\|^2$$
$$= -\frac{1}{2}G_r^2 + \frac{3}{2}\|e_r\|^2 + \frac{3}{2}\|b_r\|^2. \tag{56}$$

Using $y^{r+1} = y^r + \Delta^r/q$, the smoothness inequality becomes

$$\mathbb{E}_r F(y^{r+1}) \le F(y^r) + \frac{1}{q}\langle \nabla F(y^r), \mathbb{E}_r \Delta^r\rangle + \frac{L}{2q^2}\mathbb{E}_r\|\Delta^r\|^2.$$

Substituting the preceding bounds proves Eq. (54). □

Theorem 3 (Main convergence theorem with server momentum)**.** *Suppose Assumptions 1–3 and 4 hold, $K \ge 2$, $F^\star := \inf_w F(w) > -\infty$, and $m_0 = 0$. Let $0 \le \beta_s < 1$ be fixed and*

$$0 < \eta_l \le \frac{(1-\beta_s)^2}{60KL}.$$

With the notation of Eq. (53)*, define*

$$\chi = \frac{h}{2q} + \frac{3\beta_s^2 h^2}{2q^4}.$$

Then for every $R \geq 1$*,*

$$\frac{1}{R}\sum_{r=0}^{R-1} \mathbb{E}\|\nabla F(w^r)\|^2 \leq \frac{4q\big(F(w^0) - F^\star\big)}{aR} + 240h^2 B + 12L^2\rho^2 + 4\chi Q_0. \tag{57}$$

In particular, the right-hand side is at most

$$\frac{250}{q^2}\left[\frac{F(w^0) - F^\star}{\eta_l K R} + \frac{L\eta_l}{S}\sigma_l^2 + LK\eta_l\sigma_g^2 + L^2K^2\eta_l^2 B + L^2\rho^2\right]. \tag{58}$$

For fixed $\beta_s < 1$ *and other problem parameters,* $\eta_l = cq^2/(LK\sqrt{R})$ *with* $0 < c \leq 1/60$ *and* $\rho = c_\rho/\sqrt{R}$ *give an* $O(R^{-1/2})$ *average squared-gradient bound.*

Proof. Take total expectations in Eq. (54) and sum over rounds. Lemma 6 and Eq. (55) yield

$$\sum_{r=0}^{R-1} \mathbb{E}\|m_r\|^2 \leq \frac{a^2}{q^2}\left(RQ_0 + Q_1\sum_{r=0}^{R-1}\mathbb{E}G_r^2\right).$$

Since $y^0 = w^0$ and $F(y^R) \geq F^\star$, telescoping the objective terms gives

$$\frac{a}{q}\left(\frac{1}{2} - 180h^2 - \chi Q_1\right)\sum_{r=0}^{R-1}\mathbb{E}G_r^2 \leq F(w^0) - F^\star + \frac{aR}{q}\left(60h^2 B + 3L^2\rho^2 + \chi Q_0\right). \tag{59}$$

Here $h/(2q)$ in χ comes from the smoothness quadratic term, and $3\beta_s^2 h^2/(2q^4)$ comes from the accumulated momentum energy.

The stepsize condition implies $h \leq q^2/60 \leq 1/60$. Therefore,

$$180h^2 \leq \frac{1}{20}, \qquad Q_1 \leq \frac{31}{5}, \qquad \chi \leq \frac{1}{120} + \frac{1}{2400} = \frac{7}{800}.$$

Consequently,

$$\frac{1}{2} - 180h^2 - \chi Q_1 \geq \frac{1}{2} - \frac{1}{20} - \frac{217}{4000} = \frac{1583}{4000} > \frac{1}{4}.$$

Dividing Eq. (59) by $aR/(4q)$ proves Eq. (57).

For the simpler bound, the same stepsize condition gives

$$\chi \leq \left(\frac{1}{2q} + \frac{\beta_s^2}{40q^2}\right)h \leq \frac{21h}{40q^2}.$$

Expanding $4\chi Q_0$ in Eq. (57), the coefficient of $h^2 B$ is at most $240 + 960\chi \leq 1242/5 < 250$, and that of $L^2\rho^2$ is at most $12 + 48\chi \leq 621/50 < 250$. The remaining terms satisfy

$$\frac{8\chi}{SK}\sigma_l^2 \leq \frac{21}{5q^2}\frac{L\eta_l}{S}\sigma_l^2, \qquad 24\chi\sigma_g^2 \leq \frac{63}{5q^2}LK\eta_l\sigma_g^2.$$

Together with $4q \leq 250/q^2$, these inequalities prove Eq. (58), and hence Theorem 2 after relabeling R as T and w^r as θ^t.

Finally, substituting $\eta_l = cq^2/(LK\sqrt{R})$ and $\rho = c_\rho/\sqrt{R}$ gives terms of orders $R^{-1/2}$ from the initial gap, stochastic variance, and heterogeneity, and terms of orders R^{-1} or smaller from local drift and the perturbation radius. This proves the stated same-order rate. Momentum changes the constants and sufficient stepsize, but introduces no nonvanishing residual for fixed $\beta_s < 1$. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction clearly state the main contributions and scope of the paper. The claims about aggregation-aware perturbation search, improved stability under heterogeneous federated learning, and empirical performance are supported by the proposed algorithm, theoretical analysis, and experimental results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: None.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes].

Justification: The main convergence theorem is stated with the required assumptions, including smoothness, bounded heterogeneity, sequential conditional unbiasedness and bounded variance of stochastic gradients with martingale-difference noise, and independence between client sampling and potential local trajectories conditional on the preceding round history. The complete proof, together with auxiliary lemmas for local drift, sampled aggregation, and one-round descent, is provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes].

Justification: The paper specifies the datasets, model architectures, non-IID partitioning scheme, number of clients, partial participation setting, baselines, communication rounds, local steps, and main hyperparameters used in the experiments. These details, together with the released code, provide a reasonable path to reproduce the main experimental results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: The paper uses publicly available datasets and standard model architectures. Code, environment information, and instructions for data preparation and experiment reproduction are available at `https://github.com/Doctorwngd/FedRSAM-official`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: The experimental section describes the datasets, backbones, baselines, Dirichlet heterogeneity levels, number of clients, client participation setting, local training configuration, optimizers, and evaluation metrics. Additional hyperparameter and implementation details are provided in the appendix or supplemental material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

Justification: The main experimental tables report mean accuracy with standard deviation over multiple runs. The reported error bars capture randomness from client sampling, data partitioning, initialization, and stochastic training under the same experimental setting.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: The paper provides information about the GPU-based experimental environment and the compute requirements for the main vision and language experiments. It also discusses the additional computational overhead introduced by the restricted perturbation search compared with FedAvg and FedSAM.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work studies a general optimization algorithm for federated learning using public benchmark datasets and does not involve human-subject experiments, private user data collection, deceptive systems, or unsafe model deployment. We have reviewed the NeurIPS Code of Ethics and believe that the research conforms to it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: The paper discusses that FedRSAM may improve federated training stability and support privacy-preserving collaborative learning. It also acknowledges that stronger federated optimization methods may be used in sensitive applications, where privacy, fairness, security, and deployment risks should be carefully evaluated.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The paper does not release high-risk pretrained generative models, scraped datasets, or systems intended for direct deployment. The experiments use standard public benchmarks and evaluate optimization performance rather than releasing a potentially misuse-prone model.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

    Justification: The paper cites the original sources for the datasets, model architectures, and baseline algorithms used in the experiments, including CIFAR, Tiny-ImageNet, GLUE, ResNet, ViT, Swin Transformer, RoBERTa, FedAvg, SCAFFOLD, FedSAM, and related federated optimization methods. All assets are used for research evaluation under their intended benchmark settings.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes].

    Justification: The paper introduces a new federated optimization algorithm and provides its mathematical formulation, algorithmic procedure, theoretical analysis, and implementation details. The released code documents the training scripts, configurations, and instructions needed to reproduce the reported results.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

Justification: The paper does not involve crowdsourcing, user studies, annotation tasks, or human-subject experiments.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A].

Justification: The paper does not involve human subjects, crowdsourcing, private user data collection, or user-facing interventions, so IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A].

Justification: LLMs are not used as an important, original, or non-standard component of the proposed method or experimental pipeline. Any possible use of LLM tools is limited to writing, editing, or formatting assistance and does not affect the core methodology, results, or scientific claims.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.