FEDTEAS: FEDERATED LEARNING VIA TEACHER DISTILLATION AND SHARPNESS-AWARE MINIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Data heterogeneity remains a central challenge in Federated Learning (FL), inducing client drift and often steering optimization toward sharp solutions with poor generalization. While Sharpness-Aware Minimization (SAM) can improve flatness and robustness, locally computed SAM perturbations may become misaligned with the global objective under heterogeneous data. To address this, we introduce FedTeas, a family of three methods: vanilla FedTeas, FedTeasC, and FedTeasM that use global knowledge distillation (KD) as directional guidance for local sharpness-aware optimization. By aligning local perturbations with global predictive knowledge, FedTeas mitigates client drift and yields more consistent, robust global models across communication rounds. Rather than applying KD as a conventional parameter-wise regularizer, FedTeas injects distillation gradients directly into the SAM perturbation step, aligning local perturbations with global predictive knowledge. We prove that vanilla FedTeas retains the $O(1/\sqrt{T})$ convergence rate of FedSAM (seminal approach adapting SAM to FL), while the remaining family members preserve the convergence guarantees of their corresponding optimization frameworks. We establish a C_Ω-conditioned generalization bound that characterizes the impact of KD-guided perturbation directions on federated generalization. Extensive experiments across heterogeneous federated benchmarks show that FedTeas consistently improves accuracy, robustness, and generalization over state-of-the-art federated optimization and sharpness-aware baselines. Code is available at https://anonymous.4open.science/r/FedTeas-17D1.

1 INTRODUCTION

Federated Learning (FL) enables collaborative model training across decentralized devices while preserving data privacy. The original FedAvg (McMahan et al., 2017) aggregates local updates into a global model but suffers performance degradation under heterogeneous, non-IID data distribution (Diao et al., 2021; Ye et al., 2023). This issue, known as *client drift* (Babendererde et al., 2025), arises when local models diverge toward inconsistent local optima, significantly deviating from the global objective. Existing federated optimization methods that address client drift are largely built upon the Empirical Risk Minimization (ERM) paradigm. From the perspective of optimization, we categorize ERM-based approaches into two orthogonal categories: (i) regularization and alignment in the parameter space (Li et al., 2020a; Reddi et al., 2021; Karimireddy et al., 2020; Xu et al., 2021; Acar et al., 2021), and (ii) alignment in the output space via Knowledge Distillation (KD). Instead of directly regularizing model parameters, these methods transfer compact global knowledge (Buciluundefined et al., 2006; Hinton et al., 2014). Local models are then guided to 'mimic' the predictive behavior of either the global model (Ni et al., 2022; Yao et al., 2024) or a personalized teacher model (Zhang et al., 2018; Qin et al., 2023), thereby mitigating client drift through improved output-level consistency. However, ERM primarily prioritizes local empirical fit over the topological properties of the objective function. By neglecting the curvature of the loss surface, ERM-based models tend to be trapped in sharp regions, leading to a significant generalization gap (Dinh et al., 2017).

Sharpness-Aware Minimization (SAM) addresses this limitation by extending the ERM objective to minimize the worst-case loss within a local parameter neighborhood (Foret et al., 2021). This promotes robustness to adversarial perturbations and convergence towards flatter regions of the loss landscape. (Qu et al., 2022) first integrates SAM into FL to facilitate the convergence to a flatter loss

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

region (as illustrated in Fig. 1). However, SAM is inherently designed for centralized settings, and its direct application in FL faces fundamental challenges under data heterogeneity. Minimizing local sharpness on biased client distributions reshapes local exploration, but the mismatch with the global objective remains, leaving client drift unresolved. Subsequent methods (Dai et al., 2024; Sun et al., 2023a;b) introduce additional parameter-space corrections that require extra information exchange, thus increasing communication overhead. Alternatively, (Fan et al., 2024) simplifies the process by adopting a unified global perturbation. However, this reintroduces the trade-off between global consistency and personalized local adaptation.

Motivated by the geometric directionality of SAM and output-level consistency of KD, we propose the **FedTeas** family: a **Fed**erated framework that integrates **Tea**cher knowledge distillation with **S**harpness-aware minimization. FedTeas differs from prior distillation-based federated methods by using the global teacher signal, evaluated on local client data, to guide perturbation generation rather than directly regularizing the student predictions. This repositions KD as a directional constraint on local optimization, promoting smoother client-side landscapes and better alignment with global knowledge without requiring additional communication. Our contributions are threefold:

- We introduce the **FedTeas family**, comprising vanilla FedTeas, FedTeasC (variance-corrected), and FedTeasM (dual-corrected), which reformulates perturbation generation as KD-guided directional optimization to improve sharpness-aware learning under heterogeneous clients.
- We provide a rigorous theoretical analysis showing that vanilla FedTeas preserves the $O(1/\sqrt{T})$ convergence rate of FedSAM, while FedTeasC and FedTeasM retain the convergence guarantees of their optimization frameworks. We establish a C_Ω-conditioned generalization bound that characterizes the system-level effect of KD-guided directional complexity.
- Extensive evaluations on CIFAR-10, CIFAR-100, and Tiny-ImageNet demonstrate the effectiveness of KD-guided perturbation through the vanilla FedTeas member. The two corrected FedTeas members further demonstrate its benefits in predictive performance, optimization stability, and generalization against state-of-the-art federated optimization and sharpness-aware methods, without introducing additional communication overhead.

2 Related Work

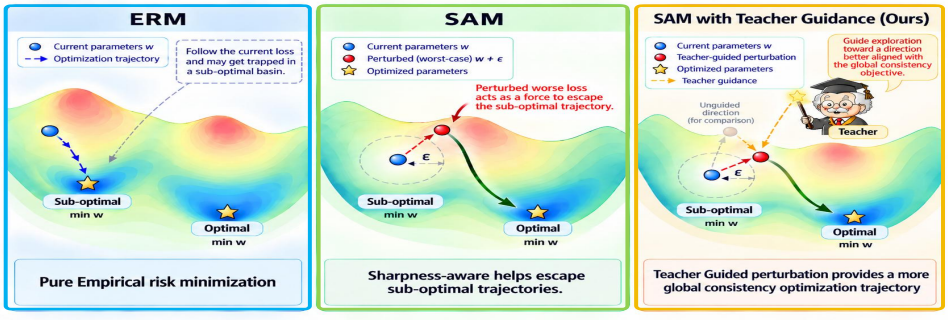


Figure 1: ERM vs. SAM and SAM with teacher guidance (FedTeas family).

McMahan et al. (2017) define a FL process as a central server that coordinates N clients to solve:

$$\min_w f(w) := \frac{1}{N}\sum_{i=1}^{N} f_i(w), \quad \text{with} \quad f_i(w) \triangleq \mathbb{E}_{\xi\sim\mathcal{D}_i}[L_{\text{CE}}(w;\xi)] \tag{1}$$

where $\xi \in \mathcal{D}_i$ represents a mini-batch drawn from local data distribution $\mathcal{D}_i$ and L_{CE} the cross-entropy loss. Due to data heterogeneity across clients, i.e., $\mathcal{D}_i \neq \mathcal{D}_j$, minimizing local empirical risk leads to *client drift* (Karimireddy et al., 2020; Acar et al., 2021). To mitigate this, representative approaches rectify the local objective as $f_i(w) = L_{\text{CE}}(w) + \phi_i(w)$. Traditional correction terms $\phi_i(w)$ focus on

the parameter space, such as proximity constraints (Li et al., 2020a), control variables (Karimireddy et al., 2020), or dual regularization (Acar et al., 2021). Recently, output-space corrections leveraged KD (Hinton et al., 2014; Zhang et al., 2018; Yao et al., 2024; Qin et al., 2023; Ni et al., 2022) to align local output logit distributions with a global teacher model. While parameter and output corrections stabilize the optimization trajectory (Zhu et al., 2021; Li et al., 2020b), they overlook the geometric properties of the loss landscape, and thus converge to sharp minima that generalize poorly under data shift (Dinh et al., 2017; Li et al., 2018). To mitigate this, an alternative paradigm optimizes the *geometry* of the loss landscape. Following SAM (Foret et al., 2021), the *sharpness* at parameter w is defined as the maximum loss increase under a perturbation ϵ within a ρ-radius neighborhood:

$$\mathcal{S}(w, \mathcal{D}) \triangleq \max_{\|\epsilon\|_2 \leq \rho} \mathbb{E}_{\xi \sim \mathcal{D}}[f(w+\epsilon, \xi) - f(w, \xi)]. \tag{2}$$

FedSAM (Qu et al., 2022) extends SAM to the federated setting, yielding the following minimax objective: $\min_w \frac{1}{N} \sum_{i=1}^{N} \max_{\|\epsilon_i\| \leq \rho} f_i(w+\epsilon_i)$. Since solving the inner maximization is intractable, SAM employs a first-order Taylor approximation to derive the optimal local perturbation:

$$\epsilon_i = \arg \max_{\|\epsilon_i\| \leq \rho} f_i(w+\epsilon_i) \approx \rho \frac{\nabla f_i(w)}{\|\nabla f_i(w)\|}, \tag{3}$$

The local update is then performed as $w \leftarrow w - \eta \nabla f_i(w+\epsilon_i)$, where the perturbation encourages exploration of flatter regions which therefore reduces sensitivity to local data variations. Building upon FedSAM, subsequent research has explored more nuanced sharpness-aware strategies. For instance, ASAM (Kwon et al., 2021) and Sparse-SAM (Mi et al., 2022) introduce adaptivity and sparsity to the perturbation calculation to enhance efficiency and robustness. However, a critical challenge remains: in heterogeneous FL, minimizing *local* sharpness on biased local data distribution $\mathcal{D}_i$ does not necessarily align with *global* flatness, potentially exacerbating client drift. To address this mismatch, subsequent research, FedGAMMA (Dai et al., 2024) adapts the control variables from SCAFFOLD (Karimireddy et al., 2020) to rectify the local sharpness gradient, ensuring the perturbed update aligns with the global objective. FedSMOO (Sun et al., 2023a) integrates the Alternating Direction Method of Multipliers (ADMM) framework with the dynamic regularization of FedDyn (Acar et al., 2021) to introduce a dual-correction mechanism on parameter and perturbation spaces. While these methods effectively address the sharpness mismatch, the maintenance of multiple auxiliary states and complex hyper-parameter tuning incur significant overhead. To alleviate this, FedLESAM (Fan et al., 2024) proposes a consistency-based approach using global perturbations to reduce computational complexity; yet, this gain in efficiency comes at a performance cost, where the personalization of local sharpness is significantly weakened. This motivates us to introduce FedTeas, which employs output-space KD to align local predictive behaviors with global knowledge, constraining the optimization trajectory toward mutually flat regions.

3 The FedTeas Family

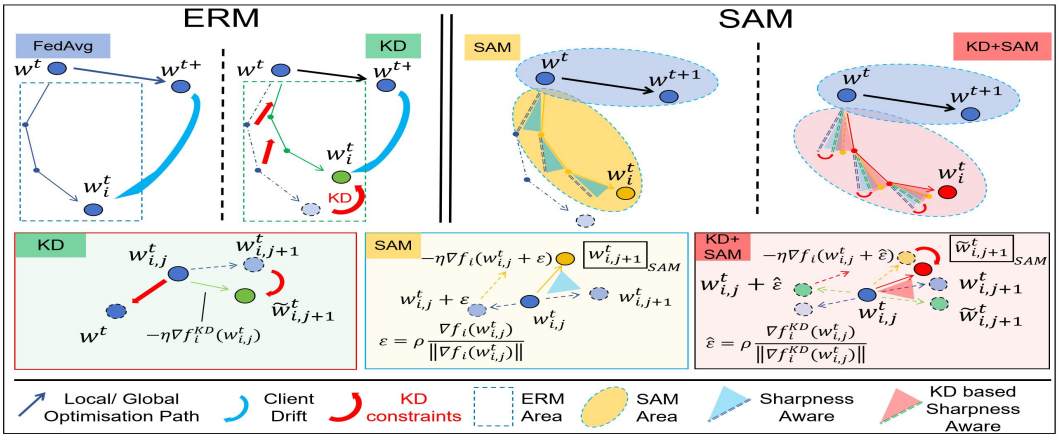


Figure 2: Detailed Comparison of Client-Side Optimization Overviews.

FedTeas comprises three methods: vanilla FedTeas, FedTeasC, and FedTeasM. From a geometric perspective, the perturbation in sharpness-aware optimization can be interpreted as an adversarial

direction that maximizes the local loss within a neighborhood. As illustrated in Fig. 2, for each local update step j over a mini-batch $\xi \in \mathcal{D}_i$, the perturbation drives the client model parameters toward the steepest ascent direction of the gradient $g(w_{i,j}^t) = \nabla f_i(w_{i,j}^t)$, and the update is performed at a perturbed point with radius ρ. Under data heterogeneity, such locally-defined worst-case directions gradually deviate from the global loss landscape without any constraints, leading to accumulated client drift. To mitigate this effect, we incorporate KD as a constraint on the optimization geometry. Specifically, KD provides an additional geometric constraint on the optimization trajectory by transferring 'soft knowledge' from a teacher model:

$$f^{\text{KD}} = \alpha_1 L_{\text{CE}}(p, y) + \alpha_2 D_{\text{KL}}(p^{\text{teacher}} \parallel p), \tag{4}$$

where $L_{\text{CE}}(p, y) = -\sum_{r=1}^{R} y_r \log p_r$ is the cross-entropy loss under local ground-truth labels (hard knowledge) with $p = \text{softmax}(z)$, and the distillation term is defined via the Kullback–Leibler (KL) divergence (Le et al., 2026):

$$D_{\text{KL}}(p^{\text{teacher}} \parallel p) = \tau^2 \sum_{r=1}^{R} \tilde{p}_r^{\text{teacher}} \log \frac{\tilde{p}_r^{\text{teacher}}}{\tilde{p}_r}, \tag{5}$$

where $\tilde{p}_r = \text{softmax}(z_r/\tau)$ and $\tilde{p}_r^{\text{teacher}} = \text{softmax}(z_r^{\text{teacher}}/\tau)$ are temperature-scaled output distributions. Following standard practice (Ni et al., 2022; Yao et al., 2024), the weights are assigned as $\alpha_1 = 1 - \alpha_2$ for balanced KD, and $\alpha_1 = 1$ and $\alpha_2 > 0$ for incremental KD.

From an optimization perspective, the KD objective reshapes the gradient $\nabla f_i^{\text{KD}}(w) = \alpha_1 \nabla L_{\text{CE}}(w) + \alpha_2 \nabla D_{\text{KL}}(w)$, which can be interpreted as a superposition of a data-dependent term and a teacher-aligned correction term. As a result, the descent direction is biased toward regions that simultaneously reduce empirical risk and maintain predictive consistency with the global model. Integrating KD into perturbation generation yields (see the direction in the right red region of Fig. 2)

$$\hat{\epsilon}_i = \rho \frac{\nabla f_i^{\text{KD}}(w)}{\|\nabla f_i^{\text{KD}}(w)\|}, \tag{6}$$

where ∇f_i^{KD} jointly reflects local supervision and teacher alignment. Unlike standard SAM, the resulting perturbation no longer corresponds to the worst-case direction of a single objective, but to a 'teacher-guided worst-case direction' that balances local sharpness and global consistency. Since the influence of KD is inherently dependent on the teacher model and varies across training stages, we adopt a *directional decoupling* strategy to disentangle the contributions of L_{CE} and D_{KL}. We compute the perturbation directions generated by each gradient separately, recombine them, and confine them to the ρ neighborhood. This avoids implicit coupling between local supervision and teacher alignment during the optimization process. We detail the implementation of the decoupled variant in Appendix A.1 and evaluate its convergence and sharpness-aware effect based on objective differences in Appendix C.3. We assess its efficacy through our ablation studies in Section 5.

Compared to directly applying KD as a parameter-update regularizer (Ni et al., 2022; Yao et al., 2024), FedTeas incorporates KD into the perturbation mechanism, thereby constraining the model implicitly via its adversarial neighborhood. This approach offers a less intrusive regularization that aligns the optimization trajectory with the loss landscape's geometry. Unless otherwise specified, the FedTeas family uses $\alpha_1 = 1.0$ and $\alpha_2 = 0.1$ as the default KD weights. The pseudo-code for the FedTeas family is provided in Appendix A.1

Remark 1. *We distinguish the objectives for perturbation and parameter updates: the perturbation $\hat{\epsilon}_i$ is computed based on the distillation objective f^{KD} whereas the parameters are updated via the standard CE Loss objective f. This separation ensures that KD only guides the sharpness-aware perturbation toward the local worst-case neighborhood, rather than directly altering the optimization target. We further analyze the impact of this separation on the FL convergence in Section 4.*

Addressing Heterogeneous Optimization. To further mitigate *client drift* under extreme data heterogeneity, we introduce the FedTeasC and FedTeasM that integrate variance correction and dual-smoothing techniques, respectively. **FedTeasC** adopts the *control variates* mechanism from SCAFFOLD (Karimireddy et al., 2020) to rectify local update directions, ensuring the KD-guided SAM trajectory aligns with the global objective. From a primal-dual perspective, informed by FedSMOO (Sun et al., 2023a). **FedTeasM** utilizes a *dual variable* to penalize the divergence between local updates and the global model. These configurations serve as robust counterparts for evaluation in highly non-IID scenarios. We provide detailed implementations in Appendix A.1 and convergence analysis in Appendix C.4.

4 THEORETICAL ANALYSIS

4.1 CONVERGENCE ANALYSIS

Assumption 1 (Smoothness)**.** $\forall i \in \{1, 2, ..., N\}$*, local loss function* f_i *is L-Lipschitz smooth satisfying* $\forall w_1, w_2$*:* $f_i(w_1) - f_i(w_2) - \langle \nabla f_i(w_2), w_1 - w_2 \rangle \leq \frac{L}{2}\|w_1 - w_2\|^2$.

Assumption 2 (Bounded Gradient Heterogeneity)**.** *The variability of local gradients is bounded:* $\|\nabla f_i(w) - \nabla f(w)\|^2 \leq \sigma_g^2$.

Assumption 1 provides the Lipschitz continuity of gradients and Assumption 2 characterizes the degree of data heterogeneity across clients by bounding the deviation between local gradients and the global gradient. The two general assumptions are widely used in the analysis of the FL framework. Following (Qu et al., 2022) and (Fan et al., 2024), we provide the local gradient boundedness conditions at unit vector scale:

Assumption 3 (Bounded variance of Stochastic gradient)**.** *The gradient* $g_i = \nabla f_i(w, \xi_i)$ *of client* i *under batch* $\xi_i \in \mathbb{R}$ *is an unbiased estimator of* $\nabla f_i(w)$ *i.e.,* $\mathbb{E}[g_i] = \nabla f_i(w)$ *with bounded variance:* $\mathbb{E}_{\xi_i} \left\| \frac{g_i}{\|g_i\|} - \frac{\nabla f_i(w)}{\|\nabla f_i(w)\|} \right\|^2 \leq \sigma_l^2$.

Assumption 3 imposes a tighter bound by constraining the difference between normalized gradients, which can be interpreted as the arc length on a unit circle and is then upper bounded by π^2 (Qu et al., 2022) .

Theorem 1. *Under Assumptions 1-3, an independent* ρ *under full client participation, and learning rates for local and global as* $\eta_l = \frac{1}{\sqrt{T}EL}$ *and* $\eta_g = \sqrt{EN}$*, respectively, FedTeas satisfies:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(w^{t+1})\|] \leq \frac{L(f(\tilde{w}^0) - f^*)}{C\sqrt{ENT}} + \frac{1}{C}\left(\sigma_g^2\left(\frac{90}{T}\right) + \sigma_l^2\left(\frac{10L^2}{ET^2} + \frac{L^2}{2\sqrt{EN}T^{3/2}}\right) + (\alpha_1 + \alpha_2)^2\left(\frac{2L^2}{T^2} + \frac{120}{E^2T^4}\right) + \frac{180L^2}{T^2}\right),$$

where $(\frac{1}{2} - 30E^2L^2\eta_l^2) \geq C > 0$.

Proof. Refer to Appendix B.3 for a proof. □

Theorem 1 establishes a non-asymptotic convergence bound for FedTeas under full client participation. By optimizing the learning rate schedules and characterizing the dominant terms as $T \to \infty$, we derive the following convergence rates for both full and partial participation regimes. For partial participation, we provide the relevant proof in Appendix B Theorem 3.

Corollary 1 (Convergence rate of FedTeas)**.** *Under the hyperparameter settings specified in Theorem 1 (full participation) and Theorem 3 (partial participation, see Appendix B), the convergence rate of FedTeas under full participation is* $\mathcal{O}(\frac{L}{\sqrt{ENT}})$*. For partial participation with* K *active clients per round, the rate becomes* $\mathcal{O}(\frac{\sqrt{E}}{\sqrt{KT}} + \frac{L}{\sqrt{EKT}})$.

The rate in Corollary 1 highlights the impact of partial participation. The term $\mathcal{O}\left(\frac{\sqrt{E}}{\sqrt{KT}}\right)$ characterizes the residual client drift, which is amplified by the number of local updates E when only a subset of clients participates in each round.

Effect of Distillation on Convergence. Based on Theorem 1 and Corollary 1, the effect of knowledge distillation is captured by the dependent term $(\alpha_1 + \alpha_2)^2$, which appears only in the higher-order components of the bound. This term represents the **perturbation bias** induced by the deviation between the joint distillation direction and the original local objective. Importantly, this bias remains *sub-dominant* and does not affect the leading convergence rate $\mathcal{O}(1/\sqrt{T})$, implying that FedTeas preserves the overall convergence behavior comparable to the baseline model (Qu et al., 2022). Under balanced perturbation ($\alpha_1 + \alpha_2 = 1$), the bound reduces to the same order as FedSAM, indicating consistent optimization dynamics. For **incremental distillation** ($\alpha_1 = 1, \alpha_2 \in [0, 1)$), the bias scales with $(1 + \alpha_2)^2$ but only affects higher-order terms (e.g., $\mathcal{O}(1/T^2)$), and thus vanishes as T increases. When $\alpha_2 = 0$, the result recovers the standard FedSAM convergence.

4.2 Generalization Bound of FedTeas

In line with prior work (Qu et al., 2022; Sun et al., 2023a; Fan et al., 2024), we analyze generalization through a margin-based PAC-Bayesian framework (McAllester, 1999; Neyshabur et al., 2018; Chatterji et al., 2020), which characterizes generalization via both the output sensitivity to perturbations and the complexity of the perturbation space. Detailed derivations are provided in Appendix C.

Definition 1 (Relative directional covering complexity)**.** *For client $i \in \{1, \dots, N\}$, let $P_{i,\boldsymbol{\alpha}}$ denote the induced perturbation-direction law on $\mathbb{S}^{d-1}$, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2) \in (0, 1]^2$ contains the task/KD mixture weights. Let $\Omega_i^{\text{base}} \subseteq \mathbb{S}^{d-1}$ denote the matched baseline direction support, let N be the number of clients, and let $\mathcal{N}(\Omega; \delta, \zeta)$ denote the minimum number of angular centers with radius $\delta > 0$ required to cover at least a $1 - \zeta$ fraction of Ω, where $\zeta \in [0, 1)$. We define*

$$\mathcal{N}_i^{\text{base}} = \mathcal{N}(\Omega_i^{\text{base}}; \delta, \zeta), \quad \mathcal{N}_i^{\text{KD}} = \inf_{\Omega \subseteq \Omega_i^{\text{base}}} \{\mathcal{N}(\Omega; \delta, \zeta) : P_{i,\boldsymbol{\alpha}}(\Omega) \geq 1 - \zeta\}. \quad C_i = \frac{\mathcal{N}_i^{\text{KD}}}{\mathcal{N}_i^{\text{base}}}$$

Here, C_i measures the relative directional covering complexity of the KD-guided perturbation for client i with respect to its matched baseline. A detailed interpretation is provided in Appendix C.1

Theorem 2 (Generalization Bound of FedTeas)**.** *Let $f(w)$ be a depth-d neural network with 1-Lipschitz activations and bounded input $\|\mathbf{X}\| \leq A$. Assume each layer satisfies $\frac{1}{M} \leq \frac{\|W_j\|}{\nu_w} \leq M$, where $\nu_w = (\prod_{j=1}^d \|W_j\|)^{1/d}$. For any margin $\gamma > 0$, dataset size m, and confidence parameter $\zeta > 0$, with probability at least $1 - \zeta$, any parameters for FedTeas under perturbation $\tilde{w} = w + \epsilon$ such that $\max_{\mathbf{x} \in \mathcal{D}_i} \|f_i(w) - f(\tilde{w})\| \leq \frac{\gamma}{8}$, the following holds*

$$\mathcal{L}^{\text{Teas}}(f(w)) \leq \hat{\mathcal{L}}_\gamma^{\text{Teas}}(f(w + \epsilon)) + \mathcal{O}\left(\sqrt{\frac{d^2 A^2 h \log(hd) \prod_{j=1}^d \|W_j\|_2^2 \sum_{j=1}^d \frac{\|W_j\|_F^2}{\|W_j\|_2^2} + d \log \frac{C_\Omega d N m \log(M)}{\zeta}}{m\gamma^2}}\right),$$

where $\mathcal{L}(\cdot)$ and $\hat{\mathcal{L}}(\cdot)$ denote the population risk and the empirical margin loss, $\|\cdot\|_F$ and $\|\cdot\|_2$ denote the Frobenius norm and the spectral norm respectively and $C_\Omega = \frac{1}{N} \sum_{i \in \mathcal{N}} C_i$ denotes the system-level relative directional covering complexity averaged over the participating clients.

Proof. Refer to Appendix C.2 for a proof. □

The Generalization – Optimization Trade-off. Theorem 2 provides a C_Ω-conditioned generalization bound that captures the aggregate effect of KD-guided perturbations across participating clients. At the client level, C_i quantifies the relative directional covering complexity induced by the KD signal under the local heterogeneous objective. This reflects how each client accommodates the global directional guidance. Aggregating these client-level quantities, C_Ω characterizes their system-level effect on federated generalization. In particular, $C_\Omega < 1$ indicates lower average directional complexity relative to the matched baseline and thus a tighter system-level generalization bound. $C_\Omega > 1$ corresponds to a comparatively looser bound. Since C_Ω depends on the relative weighting of the local task and global KD objectives, α_1 and α_2 naturally govern an inherent trade-off between preserving client-specific optimization and following global predictive guidance.Experimentally, this is reflected by the balance between predictive accuracy and the Monte Carlo estimate of C_Ω (Table 5), where the latter captures the strength of global-consistency generalization that may come at the expense of local predictive performance.

5 Experiments

Baseline. We compare the FedTeas family with a wide range of established baselines, including major ERM-based federated learning methods such as FedDyn (Acar et al., 2021), SCAFFOLD (Karimireddy et al., 2020), and FedGKD (Yao et al., 2024). We further compare against state-of-the-art SAM-based FL methods, including FedSAM (Qu et al., 2022), FedGAMMA (Dai et al., 2024), FedSMOO (Sun et al., 2023a), and FedLESAM (Fan et al., 2024). Unless otherwise specified, FedTeas refers to the vanilla version by default. We presented an extended baseline evaluation in Appendix F.1.

Table 1: Test accuracy under Dirichlet splits for CIFAR-10/100 and Tiny-ImageNet. β denotes the Dirichlet coefficient. Each result is calculated 10 times. Higher is better.

Category	Method	CIFAR-10		CIFAR-100		Tiny-ImageNet		Average	
		$\beta = 0.5$	$\beta = 0.1$	$\beta = 0.5$	$\beta = 0.1$	$\beta = 0.5$	$\beta = 0.1$	CIFAR	All
ERM	SCAFFOLD	$81.62_{\pm.64}$	$76.58_{\pm.01}$	**50.74**$_{\pm.08}$	$44.39_{\pm.19}$	$36.25_{\pm.16}$	$30.16_{\pm.07}$	$63.33_{\pm.23}$	$53.29_{\pm.19}$
	FedDyn	$81.83_{\pm.01}$	$77.01_{\pm.65}$	$43.13_{\pm.05}$	$37.16_{\pm.20}$	$32.24_{\pm.34}$	$26.88_{\pm.39}$	$59.78_{\pm.23}$	$49.71_{\pm.27}$
	FedGKD	$79.25_{\pm.53}$	$73.61_{\pm.24}$	$46.51_{\pm.08}$	$42.47_{\pm.36}$	$33.65_{\pm.33}$	$29.93_{\pm.16}$	$60.46_{\pm.30}$	$50.90_{\pm.28}$
SAM	FedSAM	$79.44_{\pm.06}$	$74.39_{\pm.13}$	$47.04_{\pm.14}$	$42.87_{\pm.01}$	$33.69_{\pm.48}$	**29.53**$_{\pm.20}$	$60.94_{\pm.09}$	$51.16_{\pm.17}$
	FedLeSAM	$79.00_{\pm.52}$	$73.76_{\pm.08}$	$46.30_{\pm.40}$	$42.02_{\pm.20}$	$33.82_{\pm.02}$	$29.74_{\pm.36}$	$60.27_{\pm.20}$	$50.77_{\pm.26}$
	FedTeas (vanilla)	$79.65_{\pm.25}$	$74.53_{\pm.31}$	$47.39_{\pm.05}$	$42.13_{\pm.20}$	**33.83**$_{\pm.34}$	$29.44_{\pm.01}$	$60.93_{\pm.20}$	$51.16_{\pm.19}$
+Variance Correction	FedGAMMA	$82.55_{\pm.20}$	$76.42_{\pm.36}$	$\underline{50.44}_{\pm.07}$	$43.60_{\pm.03}$	$35.58_{\pm.42}$	$28.71_{\pm.23}$	$63.25_{\pm.17}$	$52.88_{\pm.22}$
	FedTeasC	$82.33_{\pm.11}$	$76.79_{\pm.04}$	$49.98_{\pm.14}$	$43.75_{\pm.15}$	**35.62**$_{\pm.32}$	$28.88_{\pm.04}$	$63.21_{\pm.11}$	$52.89_{\pm.13}$
+Dual Correction	FedSMOO	$\underline{84.81}_{\pm.19}$	**79.68**$_{\pm.19}$	$50.03_{\pm.55}$	$43.80_{\pm.57}$	$37.18_{\pm.34}$	$31.94_{\pm.90}$	$\underline{64.58}_{\pm.38}$	$\underline{54.57}_{\pm.46}$
	FedTeasM	**84.84**$_{\pm.03}$	$\underline{79.46}_{\pm.52}$	$49.64_{\pm.32}$	**45.57**$_{\pm.44}$	**37.39**$_{\pm.23}$	**32.28**$_{\pm1.05}$	**64.88**$_{\pm.33}$	**54.86**$_{\pm.43}$

Table 2: Test accuracy under Pathological splits for CIFAR-10/100 and Tiny-ImageNet. u denotes the Pathological coefficient. Each result is calculated 10 times. Higher is better.

Category	Method	CIFAR-10		CIFAR-100		Tiny-ImageNet		Average	
		$u = 5$	$u = 3$	$u = 20$	$u = 10$	$u = 20$	$u = 40$	CIFAR	All
ERM	SCAFFOLD	$81.12_{\pm.04}$	$77.05_{\pm.29}$	$45.51_{\pm.12}$	$\underline{40.33}_{\pm.05}$	$27.79_{\pm.25}$	$32.78_{\pm.45}$	$61.00_{\pm.13}$	$50.76_{\pm.20}$
	FedDyn	$81.61_{\pm.30}$	$77.45_{\pm.08}$	$38.84_{\pm.69}$	$33.83_{\pm.11}$	$25.05_{\pm.24}$	$29.05_{\pm.02}$	$57.93_{\pm.30}$	$47.64_{\pm.24}$
	FedGKD	$78.68_{\pm.29}$	$74.71_{\pm.52}$	$44.30_{\pm.46}$	$39.70_{\pm.49}$	$27.83_{\pm.59}$	$31.56_{\pm.14}$	$59.35_{\pm.44}$	$49.46_{\pm.41}$
SAM	FedSAM	$79.16_{\pm.24}$	$75.54_{\pm.95}$	$43.44_{\pm.34}$	$38.79_{\pm.59}$	$27.26_{\pm.19}$	$31.38_{\pm.54}$	$59.23_{\pm.53}$	$49.26_{\pm.48}$
	FedLESAM	$78.44_{\pm.04}$	$74.72_{\pm.01}$	$43.85_{\pm.36}$	$38.78_{\pm.19}$	$27.64_{\pm.51}$	$31.72_{\pm.50}$	$58.95_{\pm.15}$	$49.19_{\pm.27}$
	FedTeas (vanilla)	$79.38_{\pm.05}$	$75.83_{\pm.59}$	$43.26_{\pm.44}$	$38.92_{\pm.46}$	**27.27**$_{\pm.39}$	**31.43**$_{\pm.07}$	$59.35_{\pm.39}$	$49.35_{\pm.33}$
+Variance Correction	FedGAMMA	$81.67_{\pm.51}$	$77.31_{\pm.55}$	$45.37_{\pm.03}$	$39.50_{\pm.31}$	$26.80_{\pm.10}$	$32.27_{\pm.33}$	$60.96_{\pm.35}$	$50.49_{\pm.31}$
	FedTeasC	$81.42_{\pm.02}$	$77.58_{\pm.01}$	$45.01_{\pm.53}$	$39.79_{\pm.31}$	**26.93**$_{\pm.18}$	**32.36**$_{\pm.15}$	$60.95_{\pm.22}$	$50.52_{\pm.20}$
+Dual Correction	FedSMOO	$\underline{83.84}_{\pm.12}$	**80.22**$_{\pm.73}$	$\underline{46.87}_{\pm.05}$	$39.63_{\pm.60}$	$30.02_{\pm.52}$	$35.52_{\pm1.07}$	$\underline{62.64}_{\pm.38}$	$\underline{52.68}_{\pm.52}$
	FedTeasM	**84.20**$_{\pm.01}$	$\underline{80.19}_{\pm.83}$	**46.95**$_{\pm.26}$	**41.09**$_{\pm.27}$	**30.16**$_{\pm.10}$	**35.56**$_{\pm.67}$	**63.11**$_{\pm.34}$	**53.03**$_{\pm.36}$

Datasets and Splitting. Following (Dai et al., 2024), (Sun et al., 2023a) and (Fan et al., 2024), we select CIFAR-10[1], CIFAR-100 (Krizhevsky et al., 2009), and Tiny-ImageNet (Le et al., 2015) as FL benchmark datasets, and employ Dirichlet distribution (Giampiccolo et al., 2007) and Pathological sampling to simulate non-IID data. By default, the FedTeas family uses incremental KD with $\alpha_1 = 1.0$ and $\alpha_2 = 0.1$. Detailed experimental settings are provided in Appendix E Table 11 and 12.

5.1 Experimental Evaluation

Test Accuracy. As shown in Table 1 and 2, we first compare two distillation paradigms: external teacher supervision (FedGKD) and sharpness-aware teacher supervision (FedTeas). In CIFAR-10, FedTeas consistently outperforms FedGKD across all settings, with an average gain of 1.036%, indicating that incorporating teacher signals into sharpness-aware information is more effective than relying on external guidance alone. Compared to FedSAM, FedTeas further achieves an average improvement of 0.28%, demonstrating the benefit of refining sharpness-aware supervision within our framework. FedTeasC improves over FedGAMMA by 0.07%, while FedTeasM surpasses FedSMOO by 0.04%. Given that the KD process is jointly controlled by α and ρ, these gains suggest that the improvement comes mainly from constrained sharpness-aware directional guidance rather than simply stronger supervision. In CIFAR-100, FedTeas and FedTeasC remain competitive with their corresponding baselines, while more notable improvements appear under stronger heterogeneity. In particular, FedTeasM achieves an average accuracy gain of 1.78% over FedSMOO. This trend is consistent with the setting in which the global model aggregates increasingly diverse client information, providing a richer distillation signal for local directional guidance. Tiny-ImageNet (Appendix F, Table 14) further supports this trend, with the FedTeas family achieving an average accuracy gain of approximately 0.35% over the corresponding matched baselines.

Loss and Volatility. We evaluate optimization quality using the average minimum loss and training stability through sliding-window loss volatility. Lower values indicate better optimization outcomes and more stable training dynamics, respectively. Detailed definitions are provided in Appendix E. On CIFAR-10 (Table 3 and Appendix F, Table 15), the FedTeas family improves minimum loss

[1]https://www.cs.toronto.edu/kriz/cifar.html

Table 3: Comparison of minimum loss and loss volatility, quantified by the interquartile range (IQR), under Dirichlet splits for CIFAR-10 and CIFAR-100. β denotes the Dirichlet coefficient.

Category	Method	CIFAR-10				CIFAR-100			
		$\beta = 0.5$		$\beta = 0.1$		$\beta = 0.5$		$\beta = 0.1$	
		Loss ↓	**IQR** ↓	**Loss** ↓	**IQR** ↓	**Loss** ↓	**IQR** ↓	**Loss** ↓	**IQR** ↓
ERM	SCAFFOLD	$0.641_{\pm.005}$	$0.0095_{\pm.0001}$	$0.796_{\pm.021}$	$0.0349_{\pm.0044}$	$1.964_{\pm.006}$	$0.0128_{\pm.0005}$	$2.223_{\pm.010}$	$0.0193_{\pm.0004}$
	FedDyn	$0.597_{\pm.001}$	$0.0128_{\pm.0003}$	$0.740_{\pm.014}$	$0.0552_{\pm.0011}$	$2.254_{\pm.003}$	$0.0192_{\pm.0002}$	$2.517_{\pm.004}$	$0.0235_{\pm.0001}$
	FedGKD	$0.729_{\pm.008}$	$0.0147_{\pm.0012}$	$0.855_{\pm.024}$	$0.0344_{\pm.0088}$	$2.317_{\pm.007}$	$0.0183_{\pm.0004}$	$2.390_{\pm.017}$	$0.0216_{\pm.0003}$
SAM	FedSAM	$0.738_{\pm.004}$	$0.0216_{\pm.0014}$	$0.863_{\pm.005}$	$0.0600_{\pm.0090}$	$2.291_{\pm.008}$	$0.0219_{\pm.0008}$	$2.392_{\pm.002}$	$0.0277_{\pm.0009}$
	FedLESAM	$0.808_{\pm.012}$	$0.0227_{\pm.0006}$	$0.930_{\pm.004}$	$0.0592_{\pm.0091}$	$2.381_{\pm.020}$	$0.0231_{\pm.0004}$	$2.489_{\pm.002}$	$0.0291_{\pm.0003}$
	FedTeas (vanilla)	$0.741_{\pm.008}$	$0.0242_{\pm.0006}$	$0.851_{\pm.001}$	$0.0609_{\pm.0064}$	$2.274_{\pm.005}$	$0.0201_{\pm.0012}$	$2.418_{\pm.006}$	$0.0293_{\pm.0003}$
+Variance Correction	FedGAMMA	$0.596_{\pm.004}$	$0.0097_{\pm.0002}$	$0.741_{\pm.002}$	$0.0402_{\pm.0055}$	$1.973_{\pm.011}$	$0.0126_{\pm.0007}$	$2.233_{\pm.010}$	$0.0198_{\pm.0001}$
	FedTeasC	$0.593_{\pm.004}$	$0.0103_{\pm.0003}$	$0.730_{\pm.003}$	$0.0364_{\pm.0068}$	$1.975_{\pm.002}$	$0.0122_{\pm.0006}$	$2.226_{\pm.003}$	$0.0190_{\pm.0003}$
+Dual Correction	FedSMOO	$0.482_{\pm.002}$	$0.0150_{\pm.0018}$	$0.650_{\pm.001}$	$0.0516_{\pm.0026}$	$2.001_{\pm.004}$	$0.0141_{\pm.0001}$	$2.242_{\pm.025}$	$0.0205_{\pm.0011}$
	FedTeasM	$0.486_{\pm.003}$	$0.0159_{\pm.0021}$	$0.650_{\pm.011}$	$0.0505_{\pm.0038}$	$2.009_{\pm.019}$	$0.0137_{\pm.0012}$	$2.182_{\pm.017}$	$0.0212_{\pm.0023}$

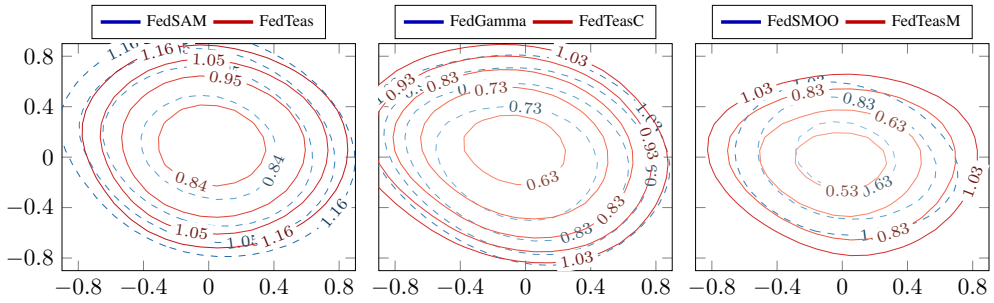


Figure 3: 2-D Loss Landscapes visualization of the global loss surface. The corresponding minimum loss values are: FedSAM (0.7688), FedTeas (0.7891), FedGAMMA (0.6358), FedTeasC (0.5918), FedSMOO (0.5580), and FedTeasM (0.4962).

and/or loss stability over the corresponding SAM-based baselines, with FedTeasC reducing volatility by 5.5% and FedTeasM reducing minimum loss by 1.15%. Similar improvements are observed on CIFAR-100 (Table 3 and Appendix F, Table 15), where FedTeasC reduces volatility by 3.58% and FedTeasM reduces minimum loss and volatility by 1.05% and 1.77%, respectively. On Tiny-ImageNet (Appendix F, Table 16), FedTeas and FedTeasM reduce average volatility by 3.70% and 2.66%, respectively, while FedTeasC reduces the average minimum loss by0.10%. Overall, these results show that the FedTeas family maintains favorable optimization quality and stability across datasets and heterogeneity settings.

Loss Landscapes. Figure 3 visualizes the loss landscapes of vanilla FedTeas, FedTeasC, and FedTeasM in the pathological setting $u = 5$ of CIFAR-10 splitting. Compared to FedSAM, FedTeas reveals a more compact region at the same loss level, indicating that, under distillation constraints, the model converges to a tighter generalization region. The variant, driven by the combined effects of spatial alignment in the distillation direction and parameter alignment in the variance correction, exhibits looser boundaries within the same loss region and demonstrates an improved ability to identify regions with lower loss. FedTeasC achieved 6.9% reduction in minimum loss, while FedTeasM achieved 11.1% reduction. The corresponding 3D landscape figures are found in Appendix Figure 9.

Hessian. Empirical results (Table 4) show that distillation guidance further strengthens the flatness induced by the parameter-alignment mechanism in FedTeasM, substantially reducing local curvature. By contrast, in the variance-reduction setting of FedTeasC, the joint use of distillation and gradient correction increases curvature. These findings suggest that distillation can promote exploration of flatter regions when coupled with effective alignment, but may need to be adaptively decoupled in later stages of training to avoid noise amplification and excessive flattening.

Table 4: Global Hessian Largest Eigenvalue (weighted average of 3 batch-size and 5 seeds).

	FedTeasM	FedSMOO	Decoupled FedTeas	FedTeas	FedSAM	FedGamma	FedGKD	FedTeasC	FedAvg	FedLESAM
Weighted Avg	**78.84**	94.09	159.78	196.55	201.90	219.70	234.68	289.11	321.69	367.27

Directional Alignment and Complexity. Table 5 evaluates the effect of KD-guided perturbations from two complementary perspectives. Across the three matched pairs, the FedTeas family im-

Table 5: Pairwise optimization alignment and directional complexity. Left: local–global gradient/update alignment and inter-client direction variance. Right: Monte Carlo relative covering-number ratios over $\delta \in \{0.6, 0.8, 1.0\}$, C_{next} denotes the ratio following a single KD-guided SAM update.

Optimization Alignment				Relative Directional Covering Complexity					
Method	Grad. Sim. ↑	Update Sim. ↑	Dir. Var. ↓	Pair	ζ	C_{next}	$C_{next,i} < 1$	C_Ω	$C_i < 1$
FedTeas	$\mathbf{0.1008 \pm 0.026}$	$\mathbf{0.2777 \pm 0.016}$	$\mathbf{0.8966 \pm 0.003}$	Teas/SAM	.05	$1.0897 \pm .1585$	13.33%	$1.0615 \pm .3020$	43.33%
FedSAM	0.0916 ± 0.014	0.2369 ± 0.016	0.8968 ± 0.006		.10	$1.1068 \pm .1729$	26.67%	$1.0752 \pm .3612$	43.33%
FedTeasC	$\mathbf{0.1895 \pm 0.022}$	$\mathbf{0.3082 \pm 0.009}$	$\mathbf{0.8978 \pm 0.005}$	TeasC/Gamma	.05	$\mathbf{0.9947 \pm .0467}$	70.00%	$1.0649 \pm .2663$	40.00%
FedGamma	0.1797 ± 0.030	0.3064 ± 0.009	0.8993 ± 0.003		.10	$\mathbf{0.9940 \pm .0536}$	66.67%	$1.0844 \pm .3253$	40.00%
FedTeasM	$\mathbf{0.3652 \pm 0.033}$	$\mathbf{0.7312 \pm 0.031}$	$\mathbf{0.4593 \pm 0.043}$	TeasM/SMOO	.05	$\mathbf{0.9414 \pm .0842}$	66.67%	$1.0335 \pm .1462$	40.00%
FedSMOO	0.3394 ± 0.026	0.7070 ± 0.027	0.4957 ± 0.037		.10	$\mathbf{0.9345 \pm .0920}$	66.67%	$1.0376 \pm .1638$	40.00%

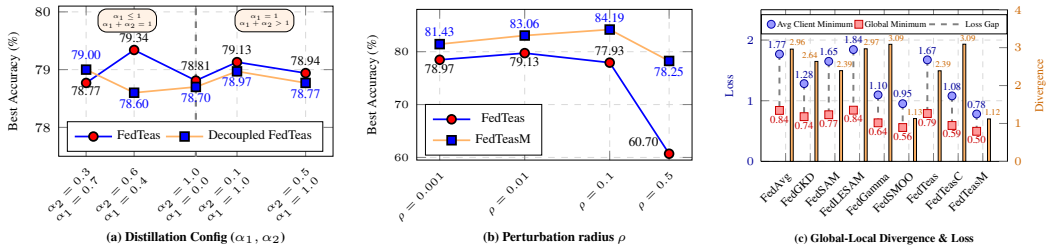


Figure 4: Ablation. (a) FedTeas famillly sensitivity to different α combinations. (b) Sensitivity of FedTeas and dual-corrected FedTeasM to different ρ values. (c) One-dimensional difference between average minimum of client loss landscape (blue circle) and global model minimum (red square) under global test set at round T, and average global-client divergence $\frac{1}{n}\sum_i \|w_i^t - w^t\|^2$ (orange bars).

proves local–global gradient and update similarity by approximately 7.70% and 7.08% on average, respectively, while reducing inter-client direction variance by 2.51%, indicating more consistent optimization directions.The right block further examines directional complexity through Monte Carlo estimates of C_Ω. Under the prescribed δ and ζ, $C_i < 1$ is observed for approximately 40–43% of sampled clients, reflecting client heterogeneity in covering complexity. Moreover, $C_{next,i} < 1$ occurs for approximately 66.67–70.00% of clients under FedTeasC and FedTeasM, indicating that an additional KD-guided SAM step more broadly reduces client-level directional complexity.

Ablation. For an exhaustive comparison in Figure 4 (a) under balanced distillation, a moderate introduction of decoupled distillation directions is more effective. As α_2 increases, incorporating the KD term directly into loss function becomes more effective. During incremental KD, the performance is comparable, but the distillation should not be excessive.(b) The optimal radius for FedTeas is $\rho = 0.01$; after incorporating Dual correction, it is more suitable for $\rho = 0.1$. As shown in Figure 4 (c), FedTeas improves personalized loss landscapes over global perturbation methods like FedLESAM by substantially reducing the gap between the average client minimum and global minimum, indicating better local–global alignment. With correction, particularly FedTeasM, this alignment is further tightened, while the parameter divergence remains stable and consistent with flatter minima demonstrating improved personalization without compromising global coherence. Furthermore, in Appendix D.2, we explore the limitations of FedTeasM by comparing the interaction between the perturbation radius ρ and the ADMM penalty strength ι and analyze FedTeasC's.

6 Conclusion and Limitations

We propose the FedTeas family, coupling KD with SAM to steer sharpness-aware updates: teacher distillation constrains local perturbation search, improving robustness and generalization. We prove convergence and derive a conditional generalization bound under the proposed constraint. Empirically, FedTeas remains effective under severe heterogeneity. However, limitations remain (Appendix D). Since FedTeas relies on distillation, its effectiveness depends on an appropriate dynamic distillation schedule (Appendix D.1). Moreover, although FedTeas introduces no additional communication overhead, distillation inevitably incurs extra local computation (Appendix D.4). Finally, when global and local architectures are identical, distillation mainly acts as a regularizer rather than providing richer cross-model guidance, which may limit its effectiveness.

REFERENCES

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021.

Niklas Babendererde, Moritz Fuchs, Camila Gonzalez, Yuri Tolkach, and Anirban Mukhopadhyay. Jointly exploring client drift and catastrophic forgetting in dynamic learning. *Scientific Reports*, 15 (1):5857, February 2025. ISSN 2045-2322.

Cristian Buciluundefined, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pp. 535–541, 2006. doi: 10.1145/1150402.1150464. URL https://doi.org/10.1145/1150402.1150464.

Niladri Chatterji, Behnam Neyshabur, and Hanie Sedghi. The intriguing role of module criticality in the generalization of deep networks. In *International Conference on Learning Representations*, 2020.

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. doi: 10.1109/TNNLS.2023.3304453.

Enmao Diao, Jie Ding, and Vahid Tarokh. Heterofl: Computation and communication efficient federated learning for heterogeneous clients. *Proceedings of International Conference on Learning Representations (ICLR)*, 2021.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1019–1028, 2017.

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*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12858–12881. PMLR, 21–27 Jul 2024.

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

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third pascal recognizing textual entailment challenge. In *ACL PASCAL*, pp. 1–9, 2007.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *NIPS*, 2014.

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*, ICML'20. JMLR.org, 2020.

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

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5905–5914. PMLR, 18–24 Jul 2021.

Duy-Dong Le, Tuong-Nguyen Huynh, and Nhat Tung Le. Fedklentropy: Enhancing federated learning via model entropy and kullback-leibler divergence. In *Proceedings of the International Conference on Research in Adaptive and Convergent Systems*, RACS '25, New York, NY, USA, 2026. Association for Computing Machinery.

Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 6391–6401, 2018.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In I. Dhillon, D. Papailiopoulos, and V. Sze (eds.), *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020a.

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

Ruixuan Liu, Fangzhao Wu, Chuhan Wu, Yanlin Wang, Lingjuan Lyu, Hong Chen, and Xing Xie. No one left behind: Inclusive federated learning over heterogeneous devices. In *28th ACM SIGKDD*, pp. 3398–3406, 2022.

David A. McAllester. Pac-bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, COLT '99, pp. 164–170, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 1581131674.

H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282, 2017.

Peng Mi, Li Shen, Tianhe Ren, Yiyi Zhou, Xiaoshuai Sun, Rongrong Ji, and Dacheng Tao. Make sharpness-aware minimization stronger: a sparsified perturbation approach. NIPS '22. Curran Associates Inc., 2022. ISBN 9781713871088.

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

Xuanming Ni, Xinyuan Shen, and Huimin Zhao. Federated optimization via knowledge codistillation. *Expert Systems with Applications*, 191:116310, 2022. ISSN 0957-4174.

Adam Paszke, Sam Gross, Francisco Massa, et al. *PyTorch: an imperative style, high-performance deep learning library*. Red Hook, NY, USA, 2019.

Zhen Qin, Shuiguang Deng, Mingyu Zhao, and Xueqiang Yan. Fedapen: Personalized cross-silo federated learning with adaptability to statistical heterogeneity. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 1954–1964, 2023.

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*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18250–18280. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/qu22a.html`.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

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*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32991–33013. PMLR, 23–29 Jul 2023a.

Yan Sun, Li Shen, Tiansheng Huang, Liang Ding, and Dacheng Tao. Fedspeed: Larger local interval, less communication round, and higher generalization accuracy. *ArXiv*, abs/2302.10429, 2023b.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4): 389–434, August 2012. ISSN 1615-3375.

Yuxin Wu and Kaiming He. Group normalization. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*, pp. 3–19, 2018.

Jing Xu, Sen Wang, Liwei Wang, and Andrew Chi-Chih Yao. Fedcm: Federated learning with client-level momentum. *arXiv preprint arXiv:2106.10874*, 2021.

Dezhong Yao, Wanning Pan, Yutong Dai, Yao Wan, Xiaofeng Ding, Chen Yu, Hai Jin, Zheng Xu, and Lichao Sun. Fedgkd: Toward heterogeneous federated learning via global knowledge distillation. *IEEE Transactions on Computers*, 73(1):3–17, 2024. doi: 10.1109/TC.2023.3315066.

Mang Ye, Xiuwen Fang, Bo Du, Pong C. Yuen, and Dacheng Tao. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Comput. Surv.*, 56(3), October 2023. ISSN 0360-0300.

Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4320–4328, 2018.

Mengdan Zhu, Raasikh Kanjiani, Jiahui Lu, Andrew Choi, Qirui Ye, and Liang Zhao. Latentexplainer: Explaining latent representations in deep generative models with multimodal large language models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, pp. 4529–4538, 2025.

Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *International conference on machine learning*, pp. 12878–12889. PMLR, 2021.

7 AI USE STATEMENT

In this work, we used generative AI tools to assist with the formulation and refinement of theoretical arguments, provide feedback on research methodology and experimental design, assist with the writing and presentation of mathematical proofs, support the interpretation of experimental results, and assist with translation. We did not use generative AI tools to generate synthetic datasets, clean or reformat datasets, or conduct qualitative or thematic data analysis, and the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to suggest experimental parameters, create or modify scientific figures (such as figure 1), revise parts of the manuscript, improve the readability of the manuscript, summarize and analyse related literature, support literature and information search, format references, and suggest manuscript structure, titles, and keywords. We have reviewed all AI-assisted work. In particular, all theoretical statements and derivations were independently checked by the authors, AI-assisted code was reviewed and tested for correctness, experimental results and their interpretations were verified against the original outputs, and literature and references were manually checked against the corresponding sources. All AI-assisted text and figures were reviewed and revised by the authors before inclusion in the manuscript. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

A APPENDIX

A.1 PSEUDO CODE

Algorithm 1 FedTeas

Input: number of clients N, Initial global and local model w^0, w_i^0, communication round T, local epochs E, learning rate η_l, KD weight α, SAM scale ρ
for round $t = 0, 1, \ldots, T-1$ **do**
  Sample K clients $\mathcal{N}_k \subset \mathcal{N}$
  for client $i \in \mathcal{N}_k$ in parallel **do**
    Client-side update
    for $j = 0, 1, \ldots, E-1$ epochs **do**
      $g_{i,j}^t \leftarrow \mathbf{KD}(w_{i,j}^t, w^t)$ ▷ using 4
      $\epsilon_{i,j}^t = \rho \frac{g_{i,j}^t}{\|g_{i,j}^t\|}$
      $\tilde{g}_{i,j}^t = \nabla f_i(w_{i,j}^t + \epsilon_{i,j}^t)$
      $w_{i,j+1}^t \leftarrow w_{i,j}^t - \eta_l \tilde{g}_{i,j}^t$
    end for
    Upload $w_i^t = w_{i,E}^t$
  end for
  Server-side update
  $w^{t+1} = \frac{1}{K} \sum_{i \in \mathcal{N}_k} w_i^t$
end for

Algorithm 2 Decoupled FedTeas

Input: number of clients N, Initial global and local model w^0, w_i^0, communication round T, local epochs E, learning rate η_l, KD weight α, SAM scale ρ
for round $t = 0, 1, \ldots, T-1$ **do**
  Sample K clients $\mathcal{N}_k \subset \mathcal{N}$
  for client $i \in \mathcal{N}_k$ in parallel **do**
    Client-side update
    for $j = 0, 1, \ldots, E-1$ epochs **do**
      $\tilde{\epsilon}_{i,j}^t = \mathbf{Combine}(\epsilon_{i,j}^t, \bar{\epsilon}_{i,j}^t)$ ▷ using 7 and 8
      $\tilde{\epsilon}_{i,j}^t = \rho \frac{\tilde{\epsilon}_{i,j}^t}{\|\tilde{\epsilon}_{i,j}^t\|}$
      $\tilde{g}_{i,j}^t = \nabla f_i(w_{i,j}^t + \tilde{\epsilon}_{i,j}^t)$
      $w_{i,j+1}^t \leftarrow w_{i,j}^t - \eta_l \tilde{g}_{i,j}^t$
    end for
    Upload $w_i^t = w_{i,E}^t$
  end for
  Server-side update
  $w^{t+1} = \frac{1}{K} \sum_{i \in \mathcal{N}_k} w_i^t$
end for

Decoupled FedTeas In practice, as training progresses and the model gradually converges, the magnitude of the distillation loss dominates the combined gradient. Consequently, the perturbation direction is increasingly dictated by the KL term (i.e., $\hat{\epsilon}_i \rightarrow \rho \frac{g_{kl}}{\|g_{kl}\|}$). This phenomenon biases the sharpness-aware update toward the teacher distribution, inadvertently discarding the geometric regularization provided by the local hard labels. To alleviate this and maintain a balanced optimization trajectory, we *decouple* the perturbation generation process. The rationale behind this decoupling is to ensure that the geometric regularization derived from the ground-truth labels is not overshadowed

Algorithm 3 FedTeasC

Input: number of clients N, Initial global and local model w^0, w_i^0, communication round T, local epochs E, learning rate η_l, KD weight α, SAM scale ρ, global and local control variant vector c^0, c_i^0.
for round $t = 0, 1, \ldots, T$ **do**
  $\mathcal{N}_k \leftarrow$ random subset of k clients from $\mathcal{N}$
  for client $i \in \mathcal{N}_k$ in parallel **do**
    Client-side update
    for $j = 0, 1, ..., E-1$ epoch **do**
      $g_{i,j}^t \leftarrow \mathbf{KD}(w_{i,j}^t, w^t)$ ▷ using 4
      $\epsilon_{i,j}^t \leftarrow \rho \frac{g_{i,j}^t}{||g_{i,j}^t||}$
      $\tilde{g}_{i,j}^t \leftarrow$ stochastic gradient at $w_{i,j}^t + \epsilon_{i,j}^t$
      $w_{i,j+1}^t \leftarrow w_{i,j}^t - \eta_l \left[\tilde{g}_{i,j}^t - c_i^t + c^t\right]$
    end for
    $\Delta c_i^t \leftarrow \frac{1}{\eta_l E}(w_{i,0}^t - w_{i,E}^t) - c^t$
    $c_i^t \leftarrow c_i^t + \Delta c_i^t$
  end for
  Server-side update
  $c^{t+1} \leftarrow c^t + \frac{1}{N} \sum_{i \in \mathcal{N}_k} \Delta c_i^t$
  $w^{t+1} \leftarrow \frac{1}{K} \sum_{i \in \mathcal{N}_k} w_i^t$
end for

Algorithm 4 FedTeasM

Input: Initial global model w^0, KD weights α, local perturbation ϵ_i^0, global perturbation ϵ^0 SAM perturbation scale ρ, local perturbation dual $\hat{\epsilon}_i^0$, global perturbation dual $\hat{\epsilon}^0$, local parameter dual λ_i^0, global parameter dual λ^0, penalty value ι
for round $t = 0, 1, \ldots, T-1$ **do**
  Select K clients $\mathcal{N}_k \subset \mathcal{N}$
  for client $i \in \mathcal{N}_k$ in parallel **do**
    Client-side update
    for $j = 0, 1, \ldots, E-1$ local epochs **do**
      $g_{i,j}^t \leftarrow \mathbf{KD}(w_{i,j}^t, w^t)$ ▷ using 4
      ▷ Compute KD gradient
      $\epsilon_{i,j}^t \leftarrow \rho \frac{\delta_{i,j}^t}{\|\delta_{i,j}^t\|}$ For $\delta_{i,j}^t = g_{i,j}^t - \hat{\epsilon}_i^t - \epsilon^t$
      ▷ Compute local perturbation
      $\hat{\epsilon}_i^t \leftarrow \hat{\epsilon}_i^t + (\epsilon_{i,j}^t - \epsilon^t)$
      ▷ Update local perturbation dual
      $\tilde{g}_{i,j}^t \leftarrow \nabla \ell(w_{i,j}^t + \epsilon_{i,j}^t)$
      ▷ Stochastic gradient at perturbed weights
      $w_{i,j+1}^t \leftarrow w_{i,j}^t - \eta \left[\tilde{g}_{i,j}^t - \lambda_i^t + \iota(w_{i,j}^t - w^t)\right]$
    end for
    $w_i^t \leftarrow w_{i,E}^t$
    $\bar{\epsilon}_i \leftarrow \hat{\epsilon}_i^t - \epsilon_{i,E}^t$
    $\lambda_i^t \leftarrow \lambda_i^t - \frac{1}{\beta}(w_{i,E}^t - w^t)$
  end for
  Server-side update
  $\epsilon^{t+1} \leftarrow \rho \frac{\delta^t}{||\delta^t||}$ for $\delta^t = \frac{1}{K} \sum_{i \in \mathcal{N}_k} \bar{\epsilon}_i$
  ▷ Aggregate global perturbation
  $\lambda^{t+1} \leftarrow \lambda^t - \frac{1}{\beta K} \sum_{i \in \mathcal{N}_k} (w_i^t - w^t)$
  ▷ Update global dual
  $w^{t+1} = \frac{1}{K} \sum_{i \in \mathcal{N}_k} w_i^t - \beta \lambda^{t+1}$
end for

by the numerical dominance of the teacher's soft supervision. First, we compute the perturbation directions induced by the supervised loss and the distillation loss independently:

$$\epsilon_{i,j}^t = \frac{g_{i,j}^{CE}}{\|g_{i,j}^{CE}\|} \text{ and } \bar{\epsilon}_{i,j}^t = \frac{g_{i,j}^{KD}}{\|g_{i,j}^{KD}\|} \tag{7}$$

Where $g_{i,j}^{CE} = \nabla L_{\text{CE}}(\cdot)$ and $g_{i,j}^{KD} = \nabla D_{\text{KL}}(\cdot)$.Subsequently, rather than a naive gradient summation, we combine the two normalized perturbations corresponding to KD pattern in Eq.4:

$$\tilde{\epsilon}_{i,j}^t := \mathbf{Combine}(\epsilon_{i,j}^t, \bar{\epsilon}_{i,j}^t) = \alpha_1 \epsilon_{i,j}^t + \alpha_2 \bar{\epsilon}_{i,j}^t \tag{8}$$

Similarly, following the constraints of SAM, we then project the combined perturbation onto the ρ-radius hypersphere:

$$\tilde{\epsilon}_{i,j}^t \leftarrow \rho \frac{\tilde{\epsilon}_{i,j}^t}{\|\tilde{\epsilon}_{i,j}^t\|} \tag{9}$$

Despite its generalization benefits, computing perturbations from biased local gradients in heterogeneous settings can misalign the perturbed gradients across clients, thereby amplifying client drift. To address this, we incorporate correction mechanisms and introduce two additional members of the FedTeas family, **FedTeasC** and **FedTeasM**, to correct the minimax objective and stabilize the optimization trajectory. A unified form of the j-th local epoch update can be written as:

$$w_{i,j+1}^t \leftarrow w_{i,j}^t - \eta_l \big(\nabla f_i(w_{i,j}^t + \tilde{\epsilon}_{i,j}^t) - \Phi\big), \tag{10}$$

where Φ is a method-specific correction term.

FedTeasC (Control Variate-based Correction). Inspired by SCAFFOLD (Karimireddy et al., 2020) and FedGAMMA (Dai et al., 2024), FedTeasC employs control variates to correct the perturbed gradients by defining $\Phi = c_i^t - c^t$. The local control variate c_i^t estimates the client-specific drift and is updated after E local epochs:

$$\begin{aligned} \Delta c_i^t &\leftarrow \frac{1}{\eta_l E}(w_{i,0}^t - w_{i,E}^t) - c^t, \\ c_i^t &\leftarrow c_i^t + \Delta c_i^t. \end{aligned} \tag{11}$$

The server then synchronizes the global control variate via

$$c^{t+1} \leftarrow c^t + \frac{1}{K} \sum_{i \in \mathcal{N}_k} \Delta c_i^t \tag{12}$$

FedTeasM (Margin/Dual-based Correction). Alternatively, inspired by FedDyn (Acar et al., 2021) and FedSMOO (Sun et al., 2023a), FedTeasM introduces an ADMM-style dual correction mechanism to jointly regulate model parameters and perturbations. It maintains dual variables for both perturbations ($\hat{\epsilon}_i^t$) and model parameters (λ_i^t). After E local epochs, the local dual variables are updated as:

$$\bar{\epsilon}_i \leftarrow \hat{\epsilon}_i^t - \epsilon_{i,E}^t, \tag{13}$$

$$\lambda_i^t \leftarrow \lambda_i^t - \frac{1}{\beta}(w_{i,E}^t - w^t). \tag{14}$$

The server aggregates these to form the global dual variables:

$$\epsilon^{t+1} \leftarrow \rho \frac{\delta^t}{\|\delta^t\|_2}, \quad \text{where} \quad \delta^t = \frac{1}{K} \sum_{i \in \mathcal{N}_k} \bar{\epsilon}_i \tag{15}$$

$$\lambda^{t+1} \leftarrow \lambda^t - \frac{1}{\beta K} \sum_{i \in \mathcal{N}_k} (w_i^t - w^t). \tag{16}$$

Finally, the global model is updated via the dual-corrected aggregation rule: $w^{t+1} = \frac{1}{K} \sum_{i \in \mathcal{N}_k} w_i^t - \beta \lambda^{t+1}$. By explicitly correcting the optimization path, FedTeasC and FedTeasM effectively neutralize the variance induced by data heterogeneity while preserving the generalization benefits of sharpness-aware distillation.

B CONVERGENCE ANALYSIS FOR FEDTEAS

B.1 ESSENTIAL UPDATE RULES

We first outline the update rules for both local and global.In round t, we sample K clients $\mathcal{N}_k \subset \mathcal{N}$, then the fundamental E epochs local training rules of client i are as follows:

$$\begin{aligned}
\epsilon_{i,j}^t &= \rho \frac{g_{i,j}^t}{\|g_{i,j}^t\|} && \textbf{(Perturbation)} \\
\tilde{w}_{i,j}^t &= w_{i,j}^t + \epsilon_{i,j}^t && \textbf{(Maximise to the perturbed position)} \\
w_{i,j+1}^t &= w_{i,j}^t - \eta_l \tilde{g}_{i,j}^t && \textbf{(Minimise at the perturbed position)}
\end{aligned}$$

where the gradient for perturbation $g_{i,j}^t = \nabla f_i^{kd}(w_{i,j}^t, w^t, \xi_{i,j})$ and the gradient for update $\tilde{g}_{i,j}^t = \nabla f_i(\tilde{w}_{i,j}^t, w^t, \xi_{i,j})$ and $j = 0, 1, \ldots, E-1$. The initial model $w_{i,0}^t = w^t$ denotes the standard consistency update with global model w^t. After E local epochs,the server updates global model with a global learning rate η_g:

$$w^{t+1} = w^t + \eta_g \frac{1}{K} \sum_{i \in \mathcal{N}_k} (w_{i,E}^t - w^t)$$

B.2 PRELIMINARY ASSUMPTION AND LEMMAS

To convey the proof clearly, without loss of generality, we first state some preliminary assumption and lemmas under full participation schemes as follows:

Lemma B.1 (Bounded gradient difference)**.** *Suppose the local loss functions f_i satisfies Assumptions1-2. The gradient of SAM is bounded by:*

$$\mathbb{E}\left[\left\|\sum_{j=0}^{E-1} \nabla f_i(w_{i,j}^t)\right\|^2\right] \leq E \sum_{j=0}^{E-1} \mathbb{E}\left[\|\nabla f_i(w_{i,j}^t)\|^2\right] + EL^2\rho^2\sigma_l^2 \tag{17}$$

Proof. Please refer to (Qu et al., 2022) for a proof. □

Then we can obtain the global updated boundedness:

Lemma B.2 (Bounded update difference)**.** *Let Assumption 1 and 2 hold, the squared norm of the averaged update difference can be bounded as*

$$\mathbb{E}_t\left[\left\|\frac{1}{N}\sum_{i=0}^{N}(w_{i,E}^t - w^t)\right\|^2\right] \leq \frac{\eta_l^2}{N^2}\left[\left\|\sum_{i,j}\nabla f_i(\tilde{w}_{i,j}^t)\right\|^2\right] + \frac{EL^2\eta_l^2\rho^2}{N}\sigma_l^2 \tag{18}$$

Proof. Under full participation schemes, we have:

$$\begin{aligned}
\mathbb{E}_t\left[\left\|\frac{1}{N}\sum_{i=0}^{N}(w_{i,E}^t - w^t)\right\|^2\right] &\leq \frac{\eta_l^2}{N^2}\mathbb{E}_t\left[\left\|\sum_{i,j}\tilde{g}_{i,j}^t\right\|^2\right] \\
&= \frac{\eta_l^2}{N^2}\mathbb{E}_t\left[\left\|\sum_{i,j}(\tilde{g}_{i,j}^t - \nabla f_i(\tilde{w}_{i,j}^t))\right\|^2\right] + \frac{\eta_l^2}{N^2}\mathbb{E}_t\left[\left\|\sum_{i,j}\nabla f_i(\tilde{w}_{i,j}^t)\right\|^2\right] \\
&\leq \frac{\eta_l^2}{N^2}\left[\left\|\sum_{i,j}\nabla f_i(\tilde{w}_{i,j}^t)\right\|^2\right] + \frac{EL^2\eta_l^2\rho^2}{N}\sigma_l^2
\end{aligned}$$

where the first inequality follows from **Jensen's inequality**, specifically using the property that $\| \sum_{j=1}^{n} x_j \|^2 \leq n \sum_{j=1}^{n} \|x_j\|^2$ for any n vectors in $\mathbb{R}^d$, and the second inequality is from Lemma B.1 □

Lemma B.3 (Bounded global variance of perturbed gradient)**.** *Let Assumptions 1 and 2 hold, the variance of local and global gradients with perturbation can be bounded as:*

$$\|\nabla f_i(w+\epsilon_i) - \nabla F(w+\epsilon)\|^2 \leq 3\sigma_g^2 + 6L^2\rho^2 \tag{19}$$

Proof.

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

where the first inequality (a) follows from Jensen's inequality similar to Lemma B.2, and the second inequality (b) invokes the L-smoothness of both local and global loss functions (Assumption 1) and the perturbation is bounded by ρ i.e., $\|\epsilon\| \leq \rho$. □

Lemma B.4 (Bounded iteration difference)**.** *Suppose local update functions satisfy Assumption 1-3, if the learning rate satisfy $\eta_l \leq \frac{1}{10EL}$,the update difference at any iterations within a round can be bounded as*

$$\begin{aligned}
\frac{1}{N}\sum_i \mathbb{E}\|w_{i,j}^t - w^t\|^2 \leq & 5E\eta_l^2\big[2L^2\rho^2\sigma_l^2 + 6E(3\sigma_g^2 + 6L^2\rho^2) + 6E\|\nabla f(\tilde{w}^t)\|^2 \\
& + 12E\eta_l^2 L^2 \frac{1}{N}\sum_{i\in\mathcal{N}} \mathbb{E}\|\epsilon_{i,j}^t - \epsilon_{i,0}^t\|^2\big]
\end{aligned} \tag{20}$$

Proof. A proof is provided in (Qu et al., 2022), Note: the outcome here corrects the error occurred in the recursive equation: $A_{i,j} \leq (1 + \frac{1}{E-1})A_{i,j-1} + C$ in (Qu et al., 2022). □

Lemma B.5 (Bounded Perturbation Drift under KD)**.** *Let Assumption 1-3 hold,the variance of difference between the local perturbations $\epsilon_{i,j}^t (j = 0, 1, \ldots, E-1)$ at any step and initial local perturbation $\epsilon_{i,0}^t$ can be bounded as:*

$$\frac{1}{N}\sum_i \mathbb{E}[||\epsilon_{i,j}^t - \epsilon_{i,0}^t||^2] \leq 2E^2(\alpha_1+\alpha_2)^2 L^2 \eta_l^2 \rho^2. \tag{21}$$

We refer to the approach outlined in FedSAM, but given the difference between the objective function for perturbation and that for local updates, we modified the original perturbation margin of error. The proof of Lemma B.5 is as follows

Proof. If the local learning rate η_l is small, the gradient of KD $\nabla f_i^{KD}(w_{i,j}, \xi_i)$ for a single epoch is also small, the first-order Hessian approximation of the expected gradient under KD can be written as

$$\nabla f_i^{KD}(w_{i,j}^t) = \nabla f_i^{KD}(w_{i,j-1}^t + g_{i,j-1}^t) = \nabla f_i^{KD}(w_{i,j-1}^t) + H\eta_l g_{i,j-1}^t + O(\|\eta_l g_{i,j-1}^t\|^2) \tag{22}$$

where H is the Hessian at $w_{i,j-1}^t$ under functions $f_i^{KD}(*) = \alpha_1 L_{CE}(*) + \alpha_2 D_{KL}(*)$. We can bound the perturbation difference using $\phi_{i,j}$, which is the square of the angle between the unit vector in the direction of $\nabla f_i^{KD}(w_{i,j})$ and the unit vector in the direction of the initial position $\nabla f_i^{KD}(w_{i,0}^t)$:

$$\mathbb{E}[\|\epsilon_{i,j}^t - \epsilon_{i,0}^t\|^2] = \rho^2 \mathbb{E}\left[\left\| \frac{\nabla f_i^{KD}(w_{i,j}^t)}{\|\nabla f_i^{KD}(w_{i,j}^t)\|} - \frac{\nabla f_i^{KD}(w_{i,0}^t)}{\|\nabla f_i^{KD}(w_{i,0}^t)\|} \right\|^2\right] \leq \rho^2 \phi_{i,j} \tag{23}$$

Intuitively, the above boundary measures the arc length between two vectors on the unit circle. When $\phi_{i,j}$ is also small, given the first-order Taylor series expansion of the tangent function, we obtain:

$$\begin{aligned}\tan\phi_{i,j} &= \frac{\|\nabla f_i^{KD}(w_{i,j}^t) - \nabla f_i^{KD}(w_{i,0}^t)\|^2}{\|\nabla f_i^{KD}(w_{i,0}^t)\|^2} + O(\phi_{i,j}^2) \\ &= \frac{\|\nabla f_i^{KD}(w_{i,j-1}^t) - H\eta_l g_{i,j-1}^t - O(\|\eta_l g_{i,j-1}^t\|^2) - \nabla f_i^{KD}(w_{i,0}^t)\|^2}{\|\nabla f_i^{KD}(w_{i,0}^t)\|^2} + O(\phi_{i,j}^2)\end{aligned}$$

In contrast to the loose smoothness assumption, here we assume that $f_i(*)$ satisfies L-smoothness, i.e. the loss $L_{CE}(*)$ satisfies L-smoothness; whereas for $f_i^{KD}(*)$, the loss $L_{CE}(*)$ satisfies $\alpha_1 L$-smoothness and $D_{KL}(*)$ satisfies $\alpha_2 L$-smoothness, then the corresponding hybrid objective function $f_i^{KD}(*)$ inherently satisfies $(\alpha_1+\alpha_2)L$-smoothness, then the maximum eigenvalue of H is bounded by $(\alpha_1+\alpha_2)L$. And based on triangle inequality: for any $a>0$ vectors $u, v \in \mathbb{R}$ satisfy $\|u+v\|^2 \leq (1+a)\|u\|^2 + (1+\frac{1}{a})\|v\|^2$. With $a = \frac{1}{E-1}$ we obtain:

$$\begin{aligned}&\tan\phi_{i,j} \\ &\leq \left(1+\frac{1}{E-1}\right)\frac{\|\nabla f_i(w_{i,j-1}^t) - \nabla f_i(w_{i,0}^t)\|^2}{\|\nabla f_i(w_{i,0}^t)\|^2} + \frac{E\|H\eta_l g_{i,j-1}^t + O(\|\eta_l g_{i,j-1}^t\|^2)\|^2}{\|\nabla f_i(w_{i,0}^t)\|^2} + O(\phi_{i,j}^2) \\ &\leq \left(1+\frac{1}{E-1}\right)\frac{\|\nabla f_i(w_{i,j-1}^t) - \nabla f_i(w_{i,0}^t)\|^2}{\|\nabla f_i(w_{i,0}^t)\|^2} + E(\alpha_1+\alpha_2)^2 L^2 \eta_l^2\end{aligned}$$

Let $A_{i,j} = \frac{\|\nabla f_i(w_{i,j}^t) - \nabla f_i(w_{i,0}^t)\|^2}{\|\nabla f_i(w_{i,0}^t)\|^2}$ and $C = E(\alpha_1+\alpha_2)^2 L^2 \eta_l^2$. Given the recursive equation $A_{i,j} \leq (1+\frac{1}{E-1})A_{i,j-1} + C$, together with equation.23, we obtain

$$\begin{aligned}\frac{1}{N}\sum_{i\in\mathcal{N}}\mathbb{E}[||\epsilon_{i,j}^t - \epsilon_{i,0}^t||^2] &\leq \rho^2 E(\alpha_1+\alpha_2)^2 L^2 \eta_l^2 \sum_{j=1}^{E-1}\left(1+\frac{1}{E-1}\right)^j \\ &\leq 2\rho^2 E^2(\alpha_1+\alpha_2)^2 L^2 \eta_l^2 \end{aligned} \tag{24}$$

This completes the proof. □

Lemma B.6. *Let Assumption 1-3 hold, the inner product can be bounded as*

$$\begin{aligned}&\left\langle \nabla f(\tilde{w}^t), \mathbb{E}_t[\frac{1}{N}\sum_{i\in\mathcal{N}}(w_{i,E}^t - w^t) + \eta_l E\nabla f(\tilde{w}^t)] \right\rangle \\ &\leq E\eta_l(\frac{1}{2} + 30E^2\eta_l^2 L^2)\|\nabla f(\tilde{w}^t))\|^2 + Q - \frac{\eta_l}{2EN^2}\mathbb{E}_t\left\|\sum_{i,j}\nabla f_i(\tilde{w}_{i,j}^t)\right\|^2\end{aligned}$$

Where $Q = 10E^2\eta_l^3 L^4\rho^2\sigma_l^2 + 90E^3\eta_l^3 L^2\sigma_g^2 + 180E^3\eta_l^3 L^4\rho^2 + 120E^5\eta_l^7 L^6(\alpha_1+\alpha_2)^2\rho^2 + 2E^3\eta_l^3 L^4(\alpha_1+\alpha_2)^2\rho^2$.

Proof.

$$
\begin{aligned}
&\left\langle \nabla f(\tilde{w}^t), \mathbb{E}_t\left[\frac{1}{N}\sum_i (w_{i,E}^t - w^t) + \eta_l E \nabla f(\tilde{w}^t)\right]\right\rangle \\
&\overset{(a)}{\leq} \frac{\eta_l E}{2}\|\nabla f(\tilde{w}^t)\|^2 + \frac{\eta_l}{2EN^2}\mathbb{E}_t\left\|\sum_{i,j}\nabla f_i(\tilde{w}_{i,j}^t) - \nabla f_i(\tilde{w}^t)\right\|^2 - \frac{\eta_l}{2EN^2}\mathbb{E}_t\left\|\sum_{i,j}\nabla f_i(\tilde{w}_{i,j}^t)\right\|^2 \\
&\overset{(b)}{\leq} \frac{\eta_l E}{2}\|\nabla f(\tilde{w}^t)\|^2 + \frac{\eta_l}{2N}\sum_{i,j}\mathbb{E}_t\|\nabla f_i(\tilde{w}_{i,j}^t) - \nabla f_i(\tilde{w}^t)\|^2 - \frac{\eta_l}{2EN^2}\mathbb{E}_t\left\|\sum_{i,j}\nabla f_i(\tilde{w}_{i,j}^t)\right\|^2 \\
&\overset{(c)}{\leq} \frac{\eta_l E}{2}\|\nabla f(\tilde{w}^t)\|^2 + \frac{\eta_l L^2}{2N}\sum_{i,j}\mathbb{E}_t\|\tilde{w}_{i,j}^t - \tilde{w}^t\|^2 - \frac{\eta_l}{2EN^2}\mathbb{E}_t\left\|\sum_{i,j}\nabla f_i(\tilde{w}_{i,j}^t)\right\|^2 \\
&\overset{(d)}{\leq} \frac{\eta_l E}{2}\|\nabla f(\tilde{w}^t)\|^2 + \frac{\eta_l L^2}{N}\sum_{i,j}\mathbb{E}_t\|w_{i,j}^t - w^t\|^2 + \frac{\eta_l L^2}{N}\sum_{i,j}\mathbb{E}_t\|\epsilon_{i,j}^t - \epsilon_{i,0}^t\|^2 \\
&\qquad - \frac{\eta_l}{2EN^2}\mathbb{E}_t\left\|\sum_{i,j}\nabla f_i(\tilde{w}_{i,j}^t)\right\|^2 \\
&= \frac{\eta_l E}{2}\|\nabla f(\tilde{w}^t)\|^2 + E\eta_l L^2\frac{1}{N}\sum_i \mathbb{E}\|w_{i,j}^t - w^t\|^2 + E\eta_l L^2\frac{1}{N}\sum_{i\in\mathcal{N}}\mathbb{E}[||\epsilon_{i,j}^t - \epsilon_{i,0}^t||^2] \\
&\qquad - \frac{\eta_l}{2EN^2}\mathbb{E}_t\left\|\sum_{i,j}\nabla f_i(\tilde{w}_{i,j}^t)\right\|^2
\end{aligned}
\tag{25}
$$

where inequality (a) is from that $\langle a, b\rangle = \frac{1}{2}(\|a\|^2 + \|b\|^2 - \|a-b\|^2)$ with $a = \sqrt{\eta_l E}\nabla f(\tilde{w}^t)$ and $b = -\frac{\sqrt{\eta_l}}{N\sqrt{E}}\sum_{i,j}(\nabla f_i(\tilde{w}_{i,j}^t) - \nabla f_i(\tilde{w}^t))$; inequality (b) and (d) follow from Jensen's inequality ; inequality (c) is from Assumption 1. Combined with Lemma B.5 and B.4. we obtain the following boundary

$$
\langle * \rangle \leq E\eta_l(\frac{1}{2} + 30E^2\eta_l^2 L^2)\|\nabla f(\tilde{w}^t))\|^2 + Q - \frac{\eta_l}{2EN^2}\mathbb{E}_t\left\|\sum_{i,j}\nabla f_i(\tilde{w}_{i,j}^t)\right\|^2 \tag{26}
$$

with $Q = 10E^2\eta_l^3 L^4\rho^2\sigma_l^2 + 90E^3\eta_l^3 L^2\sigma_g^2 + 180E^3\eta_l^3 L^4\rho^2 + 120E^5\eta_l^7 L^6(\alpha_1+\alpha_2)^2\rho^2 + 2E^3\eta_l^3 L^4(\alpha_1+\alpha_2)^2\rho^2$. This completes the proof. □

Lemma B.7 (Descent Lemma)**.** *Let Assumption 1 - 3 hold, at any iteration round t, the update loss function of FedTeas satisfy:*

$$
\mathbb{E}_t[f(\tilde{w}^{t+1})] \leq f(\tilde{w}^t) - E\eta_g\eta_l\left(\frac{1}{2} - 30E^2L^2\eta_l^2\right)\|\nabla f(\tilde{w}^t)\|^2 + \eta_g Q + \frac{E\eta_g^2\eta_l^2 L^3\rho^2}{2N}\sigma_l^2
$$

Proof.

$$\begin{aligned}
\mathbb{E}_t[f(w^{t+1})] &= \mathbb{E}_t[f(\tilde{w}^{t+1})] \leq f(\tilde{w}^t) + \mathbb{E}_t[\langle \nabla f(\tilde{w}^t), \tilde{w}^{t+1} - \tilde{w}^t \rangle] + \frac{L}{2}\mathbb{E}_t[\|\tilde{w}^{t+1} - \tilde{w}^t\|^2] \\
&= f(\tilde{w}^t) + \mathbb{E}_t \left\langle \nabla f(\tilde{w}^t), -\frac{1}{N}\sum_i (w_{i,E}^t - w^t) + E\eta_g\eta_l \nabla f(\tilde{w}^t) - E\eta_g\eta_l \nabla f(\tilde{w}^t) \right\rangle \\
&\quad + \frac{L}{2}\eta_g^2 \mathbb{E}_t[\|\frac{1}{N}\sum_i (w_{i,E}^t - w^t)\|^2] \\
&= f(\tilde{w}^t) - E\eta_g\eta_l \|\nabla f(\tilde{w}^t)\|^2 + \eta_g \left\langle \nabla f(\tilde{w}^t), \mathbb{E}_t \left[-\frac{1}{N}\sum_i (w_{i,E}^t - w^t) + E\eta_l \nabla f(\tilde{w}^t) \right] \right\rangle \\
&\quad + \frac{L}{2}\eta_g^2 \mathbb{E}_t[\|\frac{1}{N}\sum_i (w_{i,E}^t - w^t)\|^2]
\end{aligned}$$

The second equality is from the client update define in Algorithm 1 and the last equality follows from the unbiasedness of the stochastic gradients and the client sampling process, specifically that $\mathbb{E}_t[\nabla f_i(w;\xi)] = \nabla f_i(w)$. By linearity of the expectation and the inner product, we can decompose the update term into the deterministic global gradient descent step and the cumulative estimation error.Then Substituting Lemma B.6 gives:

$$\begin{aligned}
&\mathbb{E}_t[f(w^{t+1})] \\
&\leq f(\tilde{w}^t) - E\eta_g\eta_l(\frac{1}{2} - 30E^2\eta_l^2 L^2)\|\nabla f(\tilde{w}^t))\|^2 + \eta_g Q - \frac{\eta_g\eta_l}{2EN^2}\mathbb{E}_t \left[\left\| \sum_{i,j} \nabla f_i(\tilde{w}_{i,j}^t) \right\|^2 \right] \\
&\quad + \frac{L}{2}\eta_g^2 \mathbb{E}_t[\|\frac{1}{N}\sum_i (w_{i,E}^t - w^t)\|^2] \\
&\leq f(\tilde{w}^t) - E\eta_g\eta_l(\frac{1}{2} - 30E^2\eta_l^2 L^2)\|\nabla f(\tilde{w}^t))\|^2 + \eta_g Q - \frac{\eta_g\eta_l}{2EN^2}\mathbb{E}_t \left[\left\| \sum_{i,j} \nabla f_i(\tilde{w}_{i,j}^t) \right\|^2 \right] \\
&\quad + \frac{E\eta_g^2\eta_l^2 L^3\rho^2}{2N}\sigma_l^2 + \frac{L\eta_g^2\eta_l^2}{2N^2} \left[\left\| \sum_{i,j} \nabla f_i(\tilde{w}_{i,j}^t) \right\|^2 \right]
\end{aligned} \tag{27}$$

The final inequality follows from Lemma B.2. If we choose $\eta_g\eta_l \leq \frac{1}{EL}$, the inequality can be further simplified to

$$\begin{aligned}
&\mathbb{E}_t[f(w^{t+1})] \\
&\quad \leq f(\tilde{w}^t) - E\eta_g\eta_l \left(\frac{1}{2} - 30E^2L^2\eta_l^2\right) \|\nabla f(\tilde{w}^t)\|^2 + \eta_g Q + \frac{E\eta_g^2\eta_l^2 L^3\rho^2}{2N}\sigma_l^2
\end{aligned} \tag{28}$$

This completes the proof. □

B.3 Proof of Theorem 1

Proof. Now, if we take the expectation of Lemma B.7 and rearrange the terms, we obtain:

$$E\eta_g\eta_l \left(\frac{1}{2} - 30E^2L^2\eta_l^2\right) \mathbb{E}[\|\nabla f(\tilde{w}^t)\|^2] \leq \mathbb{E}[f(\tilde{w}^t)] - \mathbb{E}[f(\tilde{w}^{t+1})] + \eta_g Q + \frac{E\eta_g^2\eta_l^2 L^3\rho^2}{2N}\sigma_l^2$$

where $\mathbb{E}[\mathbb{E}_t[f(\tilde{w}^{t+1})]] = \mathbb{E}[f(\tilde{w}^{t+1})]$.By summing both sides of the inequality from $t = 0$ to $T - 1$ and multiplying by $\frac{1}{CE\eta_g\eta_l T}$ with $(\frac{1}{2} - 30E^2L^2\eta_l^2) \geq C > 0$ if $\eta_l < \frac{1}{\sqrt{60}EL}$, we have:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\tilde{w}^t)\|^2] \leq \frac{\sum_{t=0}^{T-1}(\mathbb{E}[f(\tilde{w}^t)] - \mathbb{E}[f(\tilde{w}^{t+1})])}{CE\eta_g\eta_l T} + \frac{Q}{CE\eta_l} + \frac{\eta_g\eta_l L^3\rho^2}{2CEN}\sigma_l^2$$
$$= \frac{\mathbb{E}[f(\tilde{w}^0)] - \mathbb{E}[f(\tilde{w}^T)]}{CE\eta_g\eta_l T} + \frac{Q}{CE\eta_l} + \frac{\eta_g\eta_l L^3\rho^2}{2CN}\sigma_l^2$$
$$\leq \frac{f(\tilde{w}^0) - f^*}{CE\eta_g\eta_l T} + \frac{Q}{CE\eta_l} + \frac{\eta_g\eta_l L^3\rho^2}{2CN}\sigma_l^2$$

If we choose the learning rates $\eta_l = \frac{1}{\sqrt{TEL}}$, $\eta_g = \sqrt{EN}$ and perturbation amplitude ρ proportional to the learning rate, e.g., $\rho = \frac{1}{\sqrt{T}}$. Substituting $Q = 10E^2\eta_l^3L^4\rho^2\sigma_l^2 + 90E^3\eta_l^3L^2\sigma_g^2 + 180E^3\eta_l^3L^4\rho^2 + 120E^5\eta_l^7L^6(\alpha_1+\alpha_2)^2\rho^2 + 2E^3\eta_l^3L^4(\alpha_1+\alpha_2)^2\rho^2$. and organising, we obtain:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(w^{t+1})\|] \leq \frac{L(f(\tilde{w}^0) - f^*)}{C\sqrt{ENT}} + \frac{1}{C}\left(\sigma_g^2\left(\frac{90}{T}\right) + \sigma_l^2\left(\frac{10L^2}{ET^2} + \frac{L^2}{2\sqrt{ENT^{3/2}}}\right)\right.$$
$$\left. + (\alpha_1+\alpha_2)^2\left(\frac{2L^2}{T^2} + \frac{120}{E^2T^4}\right) + \frac{180L^2}{T^2}\right).$$

This completes the proof. □

B.4 Convergence Analysis of Partial Client Participation FedTeas

Lemma B.8 (Variance of Sum under Sampling without Replacement). *Let $\{x_1, x_2, \ldots, x_N\} \subset \mathbb{R}^d$ be a fixed set of vectors, and let $\mathcal{N}_k$ be a random subset of indices of size K ($1 \leq K \leq N$) drawn uniformly without replacement from $\{1, \ldots, N\}$. Denote by $\mathbb{E}_{\mathcal{N}_k}$ the expectation with respect to this sampling scheme. Then*

$$\mathbb{E}_{\mathcal{N}_k}\left[\left\|\sum_{i\in\mathcal{N}_k} x_i\right\|^2\right] = \frac{K}{N}\sum_{i=1}^{N}\|x_i\|^2 + \frac{K(K-1)}{N(N-1)}\sum_{i\neq j} x_i^\top x_j. \tag{29}$$

Equivalently,

$$\mathbb{E}_{\mathcal{N}_k}\left[\left\|\sum_{i\in\mathcal{N}_k} x_i\right\|^2\right] = \frac{K(N-K)}{N(N-1)}\sum_{i=1}^{N}\|x_i\|^2 + \frac{K(K-1)}{N(N-1)}\left\|\sum_{i=1}^{N} x_i\right\|^2. \tag{30}$$

For large N (or when N is sufficiently large relative to K), a frequently used approximation (or exact equality when sampling is with replacement) is

$$\mathbb{E}_{\mathcal{N}_k}\left[\left\|\sum_{i\in\mathcal{N}_k} x_i\right\|^2\right] \approx \frac{K}{N}\sum_{i=1}^{N}\|x_i\|^2 + \frac{K(K-1)}{N^2}\left\|\sum_{i=1}^{N} x_i\right\|^2. \tag{31}$$

In federated learning analysis, the approximation (31) is commonly employed because it simplifies the subsequent derivation and yields the same leading-order terms.

Proof. Introduce indicator variables $I_i = \mathbf{1}\{i \in \mathcal{N}_k\}$. Under uniform sampling without replacement, $\mathbb{E}[I_i] = \frac{K}{N}$, and for $i \neq j$, $\mathbb{E}[I_iI_j] = \frac{K(K-1)}{N(N-1)}$. Then

$$\mathbb{E}\left[\left\|\sum_i I_i x_i\right\|^2\right] = \sum_i \mathbb{E}[I_i^2]\|x_i\|^2 + \sum_{i\neq j}\mathbb{E}[I_iI_j]x_i^\top x_j. \tag{32}$$

Since $I_i^2 = I_i$, $\mathbb{E}[I_i^2] = \frac{K}{N}$. Substituting the expectations yields equation (29). Identity (30) follows from the relation $\sum_{i\neq j} x_i^\top x_j = \left\|\sum_i x_i\right\|^2 - \sum_i \|x_i\|^2$ and rearranging. The approximation (31) is obtained by replacing $N-1$ with N in the denominator of the second term, which is asymptotically accurate for large N. □

Lemma B.9. *For the partial client participation, we can bound* $\mathbb{E}_t[\|\frac{1}{K}\sum_{i\in\mathcal{N}_k}(w_{i,E}^t - w^t)\|^2]$*:*

$$\mathbb{E}_t[\|\frac{1}{K}\sum_{i\in\mathcal{N}_k}(w_{i,E}^t - w^t)\|^2] \leq \frac{E\eta_l^2L^2\rho^2}{K}\sigma_l^2 + \frac{\eta_l^2}{KN}\sum_i\left\|\sum_{j=0}^{E-1}\nabla f_i(\tilde{w}_{i,j}^t)\right\|^2 + \frac{(K-1)\eta_l^2}{KN^2}\left\|\sum_{j=0}^{E-1}\nabla f_i(\tilde{w}_{i,j}^t)\right\|^2.$$

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

$$\begin{aligned}
&\mathbb{E}_t[\|\frac{1}{K}\sum_{i\in\mathcal{N}_k}(w_{i,E}^t - w^t)\|^2] \\
&\overset{(a)}{\leq} \frac{\eta_l^2}{K^2}\mathbb{E}_t\left[\left\|\sum_{i\in\mathcal{N}_k^t}\sum_j\tilde{g}_{i,j}^t\right\|^2\right] = \frac{\eta_l^2}{K^2}\mathbb{E}_t\left[\left\|\sum_i\mathbb{I}\{i\in\mathcal{N}_k\}\sum_j\tilde{g}_{i,j}^t\right\|^2\right] \\
&= \frac{\eta_l^2}{KN}\mathbb{E}_t\left[\left\|\sum_{i=0}^{E-1}(\tilde{g}_{i,j}^t - \nabla f_i(\tilde{w}_{i,j}^t))\right\|^2\right] + \frac{\eta_l^2}{K^2}\mathbb{E}_r\left[\left\|\sum_i\mathbb{I}\{i\in\mathcal{N}_k^t\}\sum_{j=0}^{E-1}\nabla f_i(\tilde{w}_{i,j}^t)\right\|^2\right] \\
&\overset{(b)}{\leq} \frac{E\eta_l^2L^2\rho^2}{K}\sigma_l^2 + \frac{\eta_i^2}{K^2}\mathbb{E}_t\left[\left\|\sum_{i=1}^{K}\sum_{j=0}^{E-1}\nabla f_i(\tilde{w}_{i,j}^t)\right\|^2\right] \\
&\overset{(c)}{=} \frac{E\eta_l^2L^2\rho^2}{K}\sigma_l^2 + \frac{\eta_l^2}{NK}\left\|\sum_{i=1}^{E-1}\nabla f_i(\tilde{w}_{i,j}^t)\right\|^2 + \frac{(K-1)\eta_l^2}{KN^2}\left\|\sum_{j=0}^{E-1}\nabla f_i(\tilde{w}_{i,j}^t)\right\|^2,
\end{aligned} \tag{33}$$

where the first inequality (a) follows from Jensen's inequality ; inequality (b) is from Lemma B.2 and (c) is from Lemma B.8. Note: the output of Lemma B.9 corrects the minor coefficient error in (Qu et al., 2022) □

Lemma B.10. *For* $\mathbb{E}[\|\sum_j\nabla f_i(\tilde{w}_{i,j})\|^2]$*, where* $\nabla f_i(\tilde{w}_{i,j})^2$ *for all* $j\in[E]$ *and* $i\in\mathcal{N}$ *is chosen according to FedTeas, we have:*

$$\begin{aligned}
\sum_i\mathbb{E}\left[\left\|\sum_j\nabla f_i(\tilde{w}_{i,j})\right\|^2\right] &\leq 30NE^2L^2\eta_l^2[2L^2\rho^2\sigma_l^2 + 6E(3\sigma_g^2 + 6L^2\rho^2) + 6E\|\nabla f(\tilde{w})\|^2 \\
&\quad + 12E\eta_l^2L^2\frac{1}{N}\sum_{i\in\mathcal{N}}\mathbb{E}\|\epsilon_{i,j}^t - \epsilon_{i,0}^t\|^2] + 6EL^2\sum_{i,j}\mathbb{E}[\|\epsilon_{i,j}^t - \epsilon_{i,0}^t\|^2] \\
&\quad + 3NE^2(3\sigma_g^2 + 6L^2\rho^2) + 3NE^2\|\nabla f(\tilde{w})\|^2.
\end{aligned}$$

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

Proof.

$$
\begin{aligned}
&\sum_i \mathbb{E}\left[\left\|\sum_j \nabla f_i(\tilde{w}_{i,j})\right\|^2\right] \\
&= \sum_i \mathbb{E}\left[\left\|\sum_j \nabla f_i(\tilde{w}_{i,j}) - \nabla f_i(\tilde{w}) + \nabla f_i(\tilde{w}) - \nabla f(\tilde{w}) + \nabla f(\tilde{w})\right\|^2\right] \\
&\overset{(a)}{\leq} 6EL^2 \sum_{i,j} \mathbb{E}[\|w_{i,j}^t - w^t\|^2] + 6EL^2 \sum_{i,j} \mathbb{E}[\|\epsilon_{i,j}^t - \epsilon_{i,0}^t\|^2] + 3NE^2(3\sigma_g^2 + 6L^2\rho^2) \\
&\quad + 3NE^2\|\nabla f(\tilde{w})\|^2 \\
&\overset{(b)}{\leq} 30NE^2L^2\eta_l^2[2L^2\rho^2\sigma_l^2 + 6E(3\sigma_g^2 + 6L^2\rho^2) + 6E\|\nabla f(\tilde{w})\|^2 \\
&\quad + 12E\eta_l^2L^2\frac{1}{N}\sum_{i\in\mathcal{N},j} \mathbb{E}\|\epsilon_{i,j}^t - \epsilon_{i,0}^t\|^2] + 6EL^2 \sum_{i,j} \mathbb{E}[\|\epsilon_{i,j}^t - \epsilon_{i,0}^t\|^2] \\
&\quad + 3NE^2(3\sigma_g^2 + 6L^2\rho^2) + 3NE^2\|\nabla f(\tilde{w})\|^2.
\end{aligned}
$$

where inequality (a) is from Assumption 1, Jensen's inequality and Lemma B.3; inequality (b) is from Lemmas B.4. □

Theorem 3. *Under Assumption 1-3, with an independent ρ under partial participation if the learning rate for both local and global satisfy $\eta_l = \frac{1}{\sqrt{TEL}}$ and $\eta_g = \sqrt{EK}$ Fedteas satisfies:*

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(w^{t+1})\|] \leq \frac{L(f(\tilde{w}^0) - f^*)}{C\sqrt{EKT}} + \frac{1}{C}\Bigg(&\sigma_g^2\left(\frac{90}{T} + \frac{9\sqrt{E}}{2\sqrt{KT}} + \frac{540}{2\sqrt{KE}\,T^{3/2}}\right) \\
&+ (\alpha_1+\alpha_2)^2\left(\frac{120}{E^2T^4} + \frac{12L^2}{2\sqrt{KE}\,T^{5/2}} + \frac{720L^2}{2\sqrt{KE}\,T^{7/2}} + \frac{2L^2}{T^2}\right) \\
&+ \sigma_l^2\left(\frac{10L^2}{ET^2} + \frac{L^2}{2\sqrt{KE}\,T^{3/2}} + \frac{60L^2}{2\sqrt{K}E^{3/2}T^{5/2}}\right) \\
&+ \left(\frac{180L^2}{T^2} + \frac{9\sqrt{E}L^2}{\sqrt{K}T^{3/2}} + \frac{1080L^2}{2\sqrt{KE}T^{5/2}}\right)\Bigg)
\end{aligned}
$$

Where $(\frac{1}{2} - 30E^2L^2\eta_l^2 - \frac{L\eta_g\eta_l}{2K}(3E + 180E^2L^2\eta_l^2)) \geq C > 0$.

Proof. We have the following descent:

$$
\begin{aligned}
&\mathbb{E}[f(\tilde{w}^{t+1})] \qquad (34)\\
&\leq f(\tilde{w}^t) - \frac{E\eta_g\eta_l}{2}\|\nabla f(\tilde{w}^t)\|^2 + E\eta_g\eta_l L^2\frac{1}{N}\sum_i \mathbb{E}\|w_{i,j}^t - w^t\|^2 + E\eta_g\eta_l L^2\frac{1}{N}\sum_i \mathbb{E}\|\epsilon_{i,j}^t - \epsilon_{i,0}^t\|^2 \\
&\quad - \frac{\eta_g\eta_l}{2EN^2}\mathbb{E}_t\|\sum_{i,j}\nabla f_i(\tilde{w}_{i,j}^t)\|^2 + \frac{L\eta_g^2}{2}\mathbb{E}_t[\|\frac{1}{K}\sum_{i\in\mathcal{N}_k}(w_{i,E}^t - w^t)\|^2] \\
&\leq f(\tilde{w}^t) - \frac{E\eta_g\eta_l}{2}\|\nabla f(\tilde{w}^t)\|^2 + E\eta_g\eta_l L^2\frac{1}{N}\sum_i \mathbb{E}\|w_{i,j}^t - w^t\|^2 + E\eta_g\eta_l L^2\frac{1}{N}\sum_i \mathbb{E}\|\epsilon_{i,j}^t - \epsilon_{i,0}^t\|^2 \\
&\quad + \frac{E\eta_g^2\eta_l^2L^3\rho^2}{2K}\sigma_l^2 - \frac{\eta_g\eta_l}{2EN^2}\mathbb{E}_t\|\sum_{i,j}\nabla f_i(\tilde{w}_{i,j}^t)\|^2 + \frac{\eta_g^2\eta_l^2L(K-1)}{2KN^2}\|\sum_{j=0}^{E-1}\nabla f_i(\tilde{w}_{i,j}^t)\|^2 \\
&\quad + \frac{\eta_g^2\eta_l^2L}{2KN}\sum_i\|\sum_{j=1}^{E-1}\nabla f_i(\tilde{w}_{i,j}^t)\|^2
\end{aligned}
$$

Taking the expectation of the t-th round. And when there are $K > 1$ clients participating in each round (note that when there is only one client, i.e. $K = 1$, FL degenerates into standard mini-batch SGD), if we choose $\eta_g\eta_l \leq \frac{K}{(K-1)EL}$, the inequality can be written as

$$
\begin{aligned}
&\leq f(\tilde{w}^t) - \frac{E\eta_g\eta_l}{2}\|\nabla f(\tilde{w}^t)\|^2 + E\eta_g\eta_l L^2 \frac{1}{N}\sum_i \mathbb{E}\|w_{i,j}^t - w^t\|^2 + E\eta_g\eta_l L^2 \frac{1}{N}\sum_i \mathbb{E}\|\epsilon_{i,j}^t - \epsilon_{i,0}^t\|^2 \\
&\quad + \frac{E\eta_g^2\eta_l^2 L^3\rho^2}{2K}\sigma_l^2 + \frac{\eta_g^2\eta_l^2 L}{2NK}\sum_i \|\sum_{j=1}^{E-1}\nabla f_i(\tilde{w}_{i,j}^t)\|^2 \\
&\leq f(\tilde{w}^t) - E\eta_g\eta_l\left(\frac{1}{2} - 30E^2L^2\eta_l^2 - \frac{L\eta_g\eta_l}{2K}(3E + 180E^2L^2\eta_l^2)\right)\|\nabla f(\tilde{w}^t)\|^2 \\
&\quad + E\eta_g\eta_l(10EL^4\eta_l^2\rho^2\sigma_l^2 + 90E^2L^2\eta_l^2\sigma_g^2 + 180E^2L^4\eta_l^2\rho^2 + 120E^4L^6\eta_l^6(\alpha_1+\alpha_2)^2\rho^2 \\
&\qquad + 2E^2\eta_l^2L^4(\alpha_1+\alpha_2)^2\rho^2 + \frac{\eta_g\eta_l L^3\rho^2}{2K}\sigma_l^2) \\
&\quad + \frac{E\eta_g^2\eta_l^2}{2K}(60E\eta_l^2L^5\rho^2\sigma_l^2 + 540E^2\eta_l^2L^3\sigma_g^2 + 1080E^2\eta_l^2L^5\rho^2 + 720E^4\eta_l^4(\alpha_1+\alpha_2)^2L^7\rho^2 \\
&\qquad + 12E^2\eta_l^2(\alpha_1+\alpha_2)^2L^5\rho^2 + 9EL\sigma_g^2 + 18EL^3\rho^2)
\end{aligned}
$$

where the inequality use Lemmas B.4,B.5 B.10. Similar to the proof of Theorem 1, There exists $C > 0$ such that $(\frac{1}{2} - 30E^2L^2\eta_l^2 - \frac{L\eta_g\eta_l}{2K}(3E + 180E^2L^2\eta_l^2)) \geq C > 0$. Summing over t and multiplying by $\frac{1}{CE\eta_g\eta_l T}$ yields the desired bound. Then substituting the chosen learning rates $\eta_l = \frac{1}{\sqrt{T}EL}$, $\eta_g = \sqrt{EK}$ and perturbation amplitude ρ proportional to the learning rate, e.g., $\rho = \frac{1}{\sqrt{T}}$ gives the final rate.

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(w^{t+1})\|] \leq \frac{L(f(\tilde{w}^0) - f^*)}{C\sqrt{EKT}} + \frac{1}{C}\Bigg(&\sigma_g^2\left(\frac{90}{T} + \frac{9\sqrt{E}}{2\sqrt{KT}} + \frac{540}{2\sqrt{KE}\,T^{3/2}}\right) \\
&+ (\alpha_1+\alpha_2)^2\left(\frac{120}{E^2T^4} + \frac{12L^2}{2\sqrt{KE}\,T^{5/2}} + \frac{720L^2}{2\sqrt{KE}\,T^{7/2}} + \frac{2L^2}{T^2}\right) \\
&+ \sigma_l^2\left(\frac{10L^2}{ET^2} + \frac{L^2}{2\sqrt{KE}\,T^{3/2}} + \frac{60L^2}{2\sqrt{K}E^{3/2}T^{5/2}}\right) \\
&+ \left(\frac{180L^2}{T^2} + \frac{9\sqrt{E}L^2}{\sqrt{K}T^{3/2}} + \frac{1080L^2}{2\sqrt{KE}T^{5/2}}\right)\Bigg)
\end{aligned}
$$

This completes the proof

□

C Generalization Bounds

Our generalization analysis builds on the margin-based PAC-Bayesian framework (McAllester, 1999). By combining a data-dependent KL divergence bound (Lemma C.1) with a spectral-norm-based output perturbation analysis (Lemma C.2), we derive a formal generalization guarantee for our sharpness-aware objective. The detailed roles of these lemmas and the resulting theorem are provided below.

Lemma C.1 ((Neyshabur et al., 2018)). *Let $\mathcal{L}_\gamma(f(w))$ be the expected margin risk and $\hat{\mathcal{L}}_\gamma^{\mathrm{SAM}}(f(w))$ the empirical margin risk. For any $\gamma, m, \zeta > 0$, with probability $1-\zeta$ over the training set $\mathcal{M}$ of size m, for any model w and any perturbation distribution$\mathcal{Q}$ over parameters such that $\mathbb{P}_{\epsilon\sim\mathcal{Q}}[\max_{x\sim\mathcal{D}}|f_{w+\epsilon}(x) - f_w(x)| \leq \frac{\gamma}{4}] \geq \frac{1}{2}$, the following holds:*

$$
\mathcal{L}(f(w)) \leq \hat{\mathcal{L}}_\gamma(f(w+\epsilon)) + \sqrt{\frac{2\mathrm{KL}(w+\epsilon\|P) + \log\frac{m}{\zeta}}{2(m-1)}}. \tag{35}
$$

Lemma C.2 (Perturbation Bound). *Let norm of input $\mathbf{X}$ be bounded by A. For any $A > 0$, let $f(w)$ be a neural network with ReLU activations and depth d with h units per hidden-layer. Then for any*

$w, \mathbf{X} \in \mathcal{X}$, and any perturbation ϵ s.t. $\|\epsilon_j\| \leq \|W_j\|$, where ϵ_j is the size of layer j, the change in the output of the network can be bounded as follows:

$$\|f(w+\epsilon, \mathbf{X}) - f(w, \mathbf{X})\|_2 \leq eA \prod_{j=1}^{d} \|W_j\| \sum_{j=1}^{d} \frac{\|\epsilon_j\|_2}{\|W_j\|_2}. \tag{36}$$

Lemma C.1 provides a PAC-Bayesian guarantee, bounding the expected margin loss by the empirical margin loss depending on the maximum change of the output function over the domain after a random perturbation. Lemma C.2 gives a deterministic bound on the output deviation of a neural network when its weights are perturbed, and this bound depends only on the size of the perturbation.

C.1 Directional Interpretation and Covering Complexity

Directional effect of KD guidance. FedTeas modifies the perturbation direction used in the inner maximization of SAM by combining the task and KD gradients. Let g_{task} and g_{KD} denote the task and distillation gradients, respectively, and let $\alpha_1, \alpha_2 > 0$ denote their mixture weights. Define

$$a = \frac{g_{\text{task}}}{\|g_{\text{task}}\|_2}, \qquad b = \frac{g_{\text{KD}}}{\|g_{\text{KD}}\|_2}, \qquad c = a^\top b, \tag{37}$$

and

$$u_\lambda = \frac{a + \lambda b}{\|a + \lambda b\|_2}, \qquad \lambda = \frac{\alpha_2 \|g_{\text{KD}}\|_2}{\alpha_1 \|g_{\text{task}}\|_2}. \tag{38}$$

For fixed nonzero gradients and a nonzero mixture,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(u_\lambda^\top b\right) = \frac{1 - c^2}{(1 + \lambda^2 + 2\lambda c)^{3/2}} \geq 0. \tag{39}$$

Hence, increasing the effective KD contribution aligns the perturbation more strongly with the KD-induced direction, inducing a directional concentration effect around the global guidance. Under heterogeneous client objectives, however, the same guidance need not be equally compatible with all clients. We therefore introduce the relative directional covering complexity in Definition 1 to quantify the client-dependent effect of this concentration relative to the matched baseline

Relative directional covering complexity. Definition 1 characterizes the effect of this directional guidance through a relative angular covering number. For the same angular resolution δ and tolerance ζ, $\mathcal{N}_i^{\text{base}}$ measures the directional complexity of the matched baseline, whereas $\mathcal{N}_i^{\text{KD}}$ measures the minimum covering complexity required to represent at least $1 - \zeta$ probability mass under the KD-guided direction law. Their ratio

$$C_i = \frac{\mathcal{N}_i^{\text{KD}}}{\mathcal{N}_i^{\text{base}}} \tag{40}$$

therefore provides a client-level relative complexity measure. In particular, $C_i < 1$ indicates lower directional covering complexity under the current KD signal, whereas $C_i > 1$ indicates comparatively higher directional complexity.

For full client participation, we use the client-averaged quantity

$$C_\Omega = \frac{1}{N} \sum_{i=1}^{N} C_i \tag{41}$$

to characterize the aggregate directional effect across the federated system. Accordingly, $C_\Omega < 1$ indicates lower relative directional complexity on average across participating clients, while $C_\Omega > 1$ indicates a comparatively larger average complexity.

The subsequent generalization analysis is conditional on this realized quantity. In particular, we do not assume that KD universally reduces the directional complexity of every client. Rather, C_Ω captures how the current KD signal, determined jointly by the client gradients and the relative weights α_1 and α_2, changes the effective directional complexity of the federated system.

Substituting the constrained covering number we obtain:

C.2 PROOF OF THEOREM 2

We restate Theorem 2 for completeness.

Theorem 4. *Let input $\mathbf{X}$ be an $n \times n$ image whose norm is bounded by A, $f(w)$ be the classification function with d hidden-layer neural network with h units per hidden-layer, and satisfy 1-Lipschitz activation $\theta(0) = 0$. We assume the constant $M \geq 1$ for each layer W_j satisfies: $\frac{1}{M} \leq \frac{\|W_j\|}{\nu_w} \leq M$, where $\nu_w := (\prod_{j=1}^{d} \|W_j\|)^{1/d}$ denotes the geometric average of spectral norms across all layers. Then, for any margin value γ, size of local training dataset on each client m, $\zeta > 0$, with probability $1 - \zeta$ over the training set, any parameter of FedTeas local optimizer $\tilde{w} = w + \epsilon$ such that $\max_{\mathbf{x} \in \mathcal{D}_i} \|f_i(w) - f(\tilde{w})\| \leq \frac{\gamma}{8}$, we can obtain the following constraint generalization bound under conditional domain $\Omega(\alpha_1, \alpha_2)$ with $\alpha_1 \in (0, 1]$ and $\alpha_2 \in [0, 1)$:*

$$\mathcal{L}^{\text{Teas}}(f(w)) \leq \hat{\mathcal{L}}_{\gamma}^{\text{Teas}}(f(w+\epsilon)) + \mathcal{O}\left(\sqrt{\frac{d^2 A^2 h \log(hd) \prod_{j=1}^{d} \|W_j\|_2^2 \sum_{j=1}^{d} \frac{\|W_j\|_F^2}{\|W_j\|_2^2} + d \log \frac{C_\Omega d N m \log(M)}{\zeta}}{m\gamma^2}}\right). \tag{42}$$

where $\|\cdot\|_F$ and $\|\cdot\|_2$ denote the Frobenius norm and the spectral norm respectively and $C_\Omega := \frac{\text{Vol}(\Omega(\alpha_1, \alpha_2))}{\text{Vol}(S^{d-1})}$ is the constraint constant within the conditional domain $\Omega(\alpha_1, \alpha_2)$.

Proof. Based on (Neyshabur et al., 2018) and (Qu et al., 2022) the proof consists of three steps. In the first step, we determine the maximum tolerable parameter perturbation that satisfies the given margin condition γ,based on Lemma C.2. we then calculates the KL-term boundary in PAC-Bayes bound in Lemma C.1. Finally, we utilise covering theory to approximate the generalization boundary and, with reference to (Qu et al., 2022),extend the result to the FL scenario.

Under the PAC-Bayesian setting from Lemma C.1, we choose the probability distribution of the prior P to be $\mathcal{G}(0, \psi^2 I)$. For the random perturbation ϵ_j of the j^{th} layer, we assume it follows a Gaussian distribution $\epsilon_j \sim \mathcal{G}(0, \psi_j^2 I)$ with zero-mean diagonal covariance matrix, where $\psi_j = \frac{\|\tilde{W}\|}{\nu}\psi$ and $\nu := (\prod_{j=1}^{d} \|W_j\|_2)^{1/d}$ is the geometric average of spectral norms across all layers. For any weight vector of W such that$|\|W\|_2 - \|\tilde{W}\|_2| \leq \frac{1}{d}$ for every j, we have:

$$(1/e)^{\frac{d}{d-1}} \prod_{j=1}^{d} \|\tilde{W}_j\| \leq \prod_{j=1}^{d} \|W_j\| \leq e \prod_{j=1}^{d} \|\tilde{W}_i\|. \tag{43}$$

The inequality is due to $(1 + \frac{1}{d})^{d-1} \leq (1 + \frac{1}{d})^{d} \leq e$ and we can further obtain $(1 + \frac{1}{d-1})^{d-1} \leq e$ and $\frac{1}{e} \leq (1 - \frac{1}{d})^{d-1}$.Since $\epsilon_j \sim \mathcal{G}(0, \psi_j^2 I)$, we can bound the normalized perturbation $\frac{\nu\|\epsilon_j\|}{\|\tilde{W}_j\|}$ from (Tropp, 2012) with h denotes the width of the j^{th} hidden layer:

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

Taking a union bound over the all layers with a maximum probability of $\frac{1}{2}$,we obtain an upper bound $\psi\sqrt{2h\log(4hd)}$ for the normalised perturbation $\frac{\nu\|\epsilon_j\|}{\|\tilde{W}_j\|}$ of each layer j. Substituting this boundary into the Lemma C.2, we have:

$$\|F(W+\epsilon, \mathbf{X}) - F(W, \mathbf{X})\| \leq eA\nu^d \sum_{j=1}^{d} \frac{\|\epsilon_j\|_2}{\|W_j\|_2} \leq 4edA\nu^{d-1}\psi\sqrt{2h\log(4hd)} \leq \frac{\gamma}{8}. \tag{44}$$

where we choose $\psi = \frac{\gamma}{32edA\nu^{d-1}\sqrt{2h\log(4hd)}}$ to satisfy the inequality the assumption of Lemma C.1 Then, the KL-term in Lemma C.1 can be bounded:

$$\begin{aligned}\mathrm{KL}(w+\epsilon\|P) &\leq \sum_{j=1}^{d}\frac{\|W_j\|_F^2}{2\psi_j^2} = \frac{32^2e^2d^2A^2\nu^{2d-2}h\log(4hd)}{\gamma^2}\sum_{j=1}^{d}\frac{\|W_j\|_F^2}{\|W_j\|_2^2} \\ &= \mathcal{O}\left(\frac{d^2A^2h\log(hd)\prod_{j=1}^{d}\|W_j\|_2^2}{\gamma^2}\sum_{j=1}^{d}\frac{\|W_j\|_F^2}{\|W_j\|_2^2}\right). \end{aligned} \tag{45}$$

Substitute the above bounds into Lemma C.1. Given a fixed prior distribution P and for any $\zeta > 0$ with probability $1-\zeta$, for any W we have:

$$\begin{aligned}\mathcal{L}^{\mathrm{SAM}}(f(w)) &\leq \hat{\mathcal{L}}_{\gamma}^{\mathrm{SAM}}(f(w+\epsilon)) \\ &\quad + \mathcal{O}\left(\sqrt{\frac{d^2A^2h\log(hd)\prod_{j=1}^{d}\|W_j\|_2^2\sum_{j=1}^{d}\frac{\|W_j\|_F^2}{\|W_j\|_2^2}+\log\frac{m}{\zeta}}{m\gamma^2}}\right). \end{aligned} \tag{46}$$

Base on (Qu et al., 2022), we apply a union bound to have the bound hold simultaneously for the distribution of N clients with a cover of size $\mathcal{O}(d\log(M)^d d)$ which to control the space complexity in the FL scenario, i.e., $\mathcal{O}(\log\frac{m}{\zeta}) \leq \mathcal{O}(d\log\frac{dNm\log(M)}{\zeta})$. We further investigate the impact of additional constraints on the generalization bound, and treat the distillation perturbation as a feasible domain to further restrict the covering complexity, in conjunction with equation 41, we obtained the constrained generalization bound:

$$\begin{aligned}\mathcal{L}^{\mathrm{Teas}}(f(w)) &\leq \hat{\mathcal{L}}_{\gamma}^{\mathrm{Teas}}(f(w+\epsilon)) \\ &\quad + \mathcal{O}\left(\sqrt{\frac{d^2A^2h\log(hd)\prod_{j=1}^{d}\|W_j\|_2^2\sum_{j=1}^{d}\frac{\|W_j\|_F^2}{\|W_j\|_2^2}+d\log\frac{C_\Omega dNm\log(M)}{\zeta}}{m\gamma^2}}\right). \end{aligned} \tag{47}$$

where $\mathcal{L}(\cdot)$ and $\hat{\mathcal{L}}(\cdot)$ denote the population risk and the empirical margin loss, $\|\cdot\|_F$ and $\|\cdot\|_2$ denote the Frobenius norm and the spectral norm respectively and $C_\Omega = \frac{1}{N}\sum_{i\in\mathcal{N}}C_i$ denotes the system-level relative directional covering complexity averaged over the participating clients.

This completes the proof □

C.3 Decoupled Perturbation Invariance Analysis

Table 6: Comparison of standard SAM, FedTeas, and decoupled FedTeas.

Method	Perturbation Objective	Perturbation	Perturbation Radius ρ	Update Objective
Standard SAM	$f(*) = L_{CE}(*)$	$\epsilon \propto \nabla f$	ρ	$f(*)$
FedTeas	$f^{KD}(*) = \alpha_1 L_{CE}(*) + \alpha_2 D_{KL}(*)$	$\epsilon \propto \nabla f^{KD}$	ρ	$f(*)$
Decoupled FedTeas	$\epsilon_{CE} \propto \nabla L_{CE}$ $\epsilon_{KL} \propto \nabla D_{KL}$	$\epsilon \propto \alpha_1\epsilon_{CE} + \alpha_2\epsilon_{KL}$	ρ	$f(*)$

As illustrated in Table 6, although the perturbation construction is decoupled, the underlying optimization objective remains unchanged. Specifically, from a gradient perspective, the perturbation is still derived from the KD objective:

$$f^{KD}(*) = L_{CE}(*) + D_{KL}(*), \tag{48}$$

where the implicit weights is $\alpha_1 = \alpha_2 = 1$ during perturbation construction. Assuming L_{CE} and D_{KL} are both L-smooth, then the objective $f^{KD}(*)$ inherently satisfies $2L$-smoothness,and the maximum eigenvalue of the Hessian matrix H is bounded by $2L$. Therefore, the convergence guarantees of FedTeas remain valid under the decoupled formulation.

Smoother adversarial perturbation While the decoupled formulation preserves the smoothness and convergence guarantees of the original objective, it may inadvertently compromise the effectiveness of SAM. By normalizing each gradient component independently, the magnitude information is lost, suppressing the dominance of high-curvature directions. Consequently, the resulting perturbation deviates from the true steepest ascent direction of the KD objective toward a more balanced and smoother direction. Furthermore, the introduction of two-stage normalization (Eq. 7 and Eq. 9) renders the sharpness-aware process more stable but less adversarial. This softens the exploration of worst-case sharpness, potentially leading to inferior empirical performance compared to the fully coupled FedTeas.

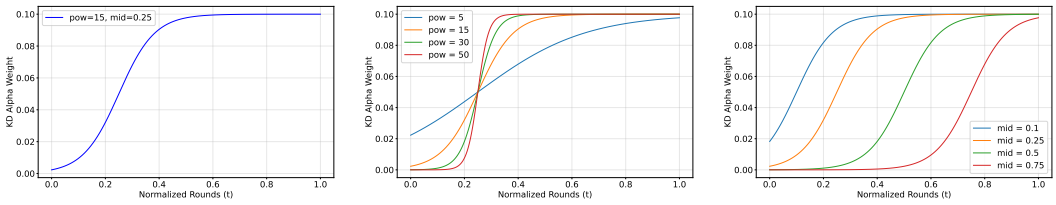
C.4 On the convergence of FedTeas under correction

To comprehensively characterize the influence of perturbations on the optimization trajectory, we follow prior works (Qu et al., 2022; Fan et al., 2024) by explicitly decomposing the perturbation-induced drift along the optimization path over the global time horizon T and local epoch E. This allows us to systematically capture the accumulated effect of perturbations throughout the training process. As demonstrated in Theorem 1 and Theorem 3, the perturbation term occupies a subdominant role in the overall convergence behavior. Therefore, a looser boundary can be introduced to characterize the perturbation-induced heterogeneity. We consider the following standard smoothness-based bound:

$$\frac{1}{N}\sum_{i\in\mathcal{N}}\mathbb{E}_t\left\|\nabla f_i(w_i^t+\epsilon_i^t)-\nabla f_i(w_i^t)\right\|^2 \le \frac{L^2}{N}\sum_{i\in\mathcal{N}}\mathbb{E}_t\|\epsilon_i^t\|^2 = \frac{L^2}{N}\sum_{i\in\mathcal{N}}\mathbb{E}_t\left\|\rho\frac{\nabla f_i}{\|\nabla f_i\|}\right\|^2 = L^2\rho^2.$$

This bound provides a more direct but looser characterization of perturbation effects, as it does not explicitly account for the drift induced by local client iterations. However, since the perturbation is bounded within the neighborhood of ρ and primarily affects the update direction (distillation-related perturbations also exhibit this directionality), the additional local deviation can be bounded by a constant G, following (Dai et al., 2024). Alternatively, under single-step local updates as in (Sun et al., 2023a), the client drift is quantified directly by the deviation between the local and global averages. Following the analytical frameworks in (Dai et al., 2024; Sun et al., 2023a), We can obtain FedTeasC and FedTeasM that maintain the same convergence rates as these works, i.e., $\mathcal{O}(\frac{1}{T})$. In contrast, Vanilla FedTeas attains a convergence rate of $\mathcal{O}(\frac{1}{\sqrt{T}})$. The improvement stems from the correction mechanism, which reduces both the discrepancy between local and global models and the variance across clients. As a result, the client drift is effectively controlled, leading to a tighter convergence bound and improved optimization efficiency.

D Limitations



(a) Dynamic α scheduling. (b) Impact of *pow* on intensity. (c) Impact of *mid* on timing.

Figure 5: The mechanics of global dynamic α scheduling. (a) illustrates the log-based linear growth logic; (b) and (c) compare the effects of growth intensity (*pow*) and midpoint temporal positioning (*mid*), respectively.

D.1 Dynamic Weighting and the Heterogeneity Paradox.

In our implementation, we adopt a log-based linear dynamic scheduling for α, formulated as a global temporal function (shown in Figure. 5a).

$$\alpha^t = \alpha\frac{1}{1+e^{-pow(t_{state}-mid)}}$$

where $t_{state} = \frac{t}{T}$, and the parameters *pow* and *mid* control the growth intensity and the temporal center of the annealing process, respectively (Figure. 5b-5c). While this global strategy ensures training stability, it inherently assumes *client homogeneity* and *teacher superiority*—assumptions that frequently fail in non-IID federated settings.

Due to data heterogeneity, the optimization trajectories of individual clients diverge significantly. This phenomenon is empirically substantiated in Table 19, which provides a global perspective on how local data skews lead to client-specific variations in the loss landscape. For 'advanced' clients whose local updates align more closely with the flat optimal region than the lagging global teacher , a high distillation weight may introduce **alignment noise**, pulling the model away from the true local minima. In such cases, the global teacher acts as a geometric constraint that hinders rather than helps local flatness exploration. Which highlights a fundamental trade-off: while global guidance mitigates drift, it may simultaneously suppress high-performing clients. Thus adaptively calibrates α based on local-to-global gradient alignment, remains a critical direction for future research.

D.2 The Interaction Between Distillation Guidance and Parameter Correction

Our experiments also reveals two critical considerations.

First. We identify a complex interaction between KD and gradient-based variance reduction (Karimireddy et al., 2020; Dai et al., 2024). Both mechanisms function as directional constraints on the gradient: while variance reduction refines the *stochastic gradient trajectory*, KD aligns the *prediction-based objective gradient*. Due to this overlapping directional regularization, the joint exploration of the loss landscape can become overly constrained. As evidenced in Table 4 and detailed results shown in Table 20, 21, 22 and 23, while this synergy effectively minimizes the empirical loss, it induces local *Hessian peaks* within otherwise flat regions. This suggests that dual-gradient constraints may inadvertently increase local curvature (Hessian values), posing a challenge for maintaining consistent landscape flatness.

Second. As evidenced by the divergent trajectories in Figure 6 and 7, there exists a *granularity mismatch* between KD-guided directional exploration and ADMM-based parameter alignment (Acar et al., 2021; Sun et al., 2023a). Under fine-grained exploration (e.g., $\rho = 0.01$), the rigid alignment penalty of ADMM ($\iota = 0.1$) dominates the optimization dynamics, leading to a dynamic instability where the sharpness-aware search is suppressed by parameter-wise constraints. While increasing the perturbation radius ρ restores this balance and yields flatter optima, it inevitably slows initial convergence. This trade-off suggests that a fixed ρ-ι configuration is insufficient. Future research should investigate **intensity matching** mechanisms and adaptive distillation release schedules to harmonize these dual constraints, ensuring that directional guidance and parameter alignment remain synchronized across different stages of federated optimization.

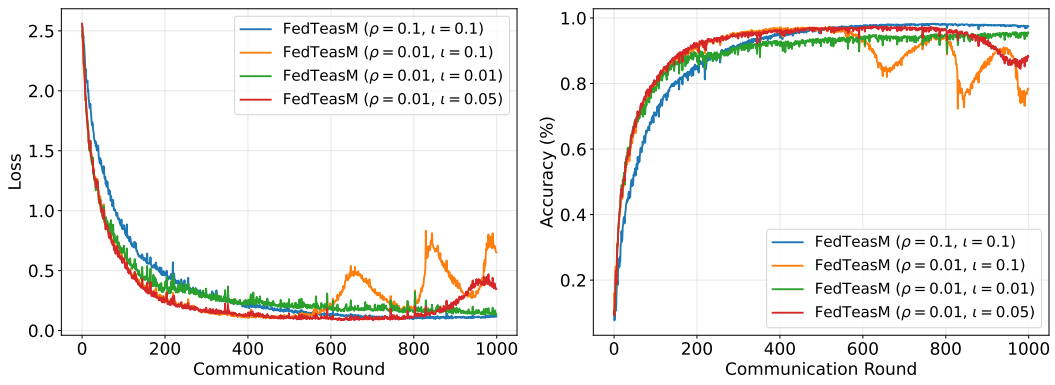


Figure 6: The Interaction Between the Perturbation Radius ρ and the ADMM Penalty Strength ι in Training Sets (CIFAR10, Dirichlet $\beta = 0.5$)

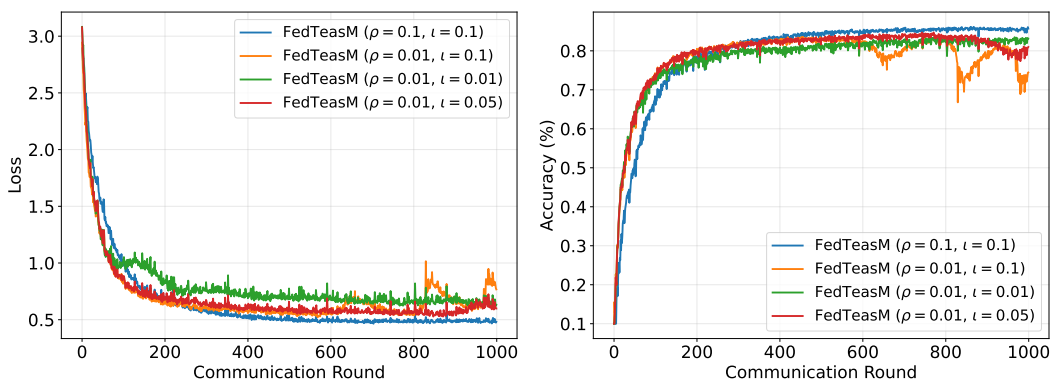


Figure 7: The Interaction Between the Perturbation Radius ρ and the ADMM Penalty Strength ι in Testing Sets (CIFAR10, Dirichlet $\beta = 0.5$)

D.3 Architectural Parity

The adoption of identical architectures across clients is a pervasive assumption in federated distillation, yet it introduces significant intrinsic limitations. **First**, while a synchronized architecture allows the global model to aggregate knowledge efficiently, it inherently imposes a *homogeneity ceiling* on the hypothesis space. In this context, the global teacher functions primarily as a **rigid constraint** (Yao et al., 2024) rather than an **adaptive enhancer**. Although personalized models (Qin et al., 2023) can theoretically boost distillation efficacy by diversifying parameter distributions, they remain fundamentally shackled by the shared structural backbone, limiting the potential for cross-architecture knowledge synergy.

Beyond architectural parity, KD faces a more formidable challenge: **asymmetric computational capacity** (Liu et al., 2022). In practical FL scenarios, resource-constrained clients often lack the hardware buffers necessary to host or invoke high-capacity teacher models locally. This necessitates the use of *compressed latent representations* (Zhu et al., 2025) or offline logic transfer. However, such indirect distillation strategies often strip away the **locality-specific features**—the unique inductive biases derived from private local data. Consequently, the student model receives a generalized, "diluted" signal that lacks the personalized guidance essential for navigating heterogeneous data distributions. Addressing this tension between model complexity and local adaptability remains a pivotal frontier for robust federated optimization.

D.4 Local Computational Overhead

System Efficiency. FedTeas families introduce no additional per-round communication payload, since the initialized global model is directly reused as the local distillation teacher. As shown in Table 7, vanilla FedTeas requires a comparable number of communication rounds to FedSAM, while FedTeasC and FedTeasM further reduce the rounds required to reach the target accuracies relative to their matched correction-based baselines. This communication efficiency comes at the cost of additional local computation introduced by KD, as reflected by the wall-clock results in Table 8. Although this computation could in principle be shifted to the server, doing so would limit the use of client-specific samples and local predictive information. FedTeas therefore exhibits a practical computation–communication trade-off while retaining client-side knowledge adaptation.

Table 7: Rounds required to reach 72% accuracy on Dirichlet $\beta = 0.1$ and 77% accuracy on Pathological $u = 5$ splitting under CIFAR10 (FedAvg = 1.0x).

	FedAvg	FedGKD	FedSAM	FedLESAM	FedGamma	FedSMOO	FedTeas	FedTeasC	FedTeasM
Round to 72%	363 (1.00×)	367 (1.01×)	345 (0.95×)	363 (1.00×)	291 (0.80×)	260 (0.72×)	355 (0.98×)	267 (0.74×)	250 (0.69×)
Round to 77%	371 (1.00×)	345 (0.93×)	284 (0.77×)	340 (0.92×)	216 (0.58×)	209 (0.56×)	284 (0.77×)	216 (0.58×)	203 (0.55×)

Table 8: Per-round runtime and relative cost

	FedAVG	Fedprox	SCAFFOLD	FedDyn	FedCM	FedCodl	Fed2Codl	FedGKD
Mean ± Std (s)	9.98 ± 1.94	14.15 ± 2.38	15.08 ± 4.99	16.73 ± 4.57	16.14 ± 6.88	16.00 ± 6.53	25.68 ± 10.82	15.56 ± 4.51
Ratio (FedAVG=1)	1.00	1.42	1.51	1.68	1.62	1.60	2.57	1.56

	Fedsam	Fedlesam	FedGamma	FedSMOO	FedTeas	FedTeasC	FedTeasM
Mean ± Std (s)	15.05 ± 2.03	13.41 ± 1.57	20.53 ± 2.88	21.96 ± 2.58	21.35 ± 1.79	26.38 ± 2.96	27.68 ± 3.80
Ratio (FedAVG=1)	1.51	1.34	2.06	2.20	2.14	2.64	2.77

E EXPERIMENTAL SETUP

Our experiments are implemented in PyTorch (Paszke et al., 2019) and executed on a CPU/GPU cluster, with NVIDIA RTX A5000 GPUs (24GB VRAM) and 16GB of system RAM. Benefiting from (Sun et al., 2023a) architecture design, parameters are only mapped to the model architecture during training and testing, while the aggregation process is performed entirely on flattened vectors to minimize overhead.

Table 9: Hardware utilization during federated training.

Metric	GPU Util. (%)	VRAM (GB)	CPU Util. (%)	RAM (GB)
Total Usage (Including idle)	29.6	0.63	156.0	26.53
Per-algorithm (Active)	40.0 – 43.7	0.9 – 1.0	170.0	30.34
Standard Requested	-	-	200	16.0

Table 9 provides a statistical summary of the hardware utilization during our experiments. The results include both the aggregate GPU usage throughout the entire training cycle (averaging 29.6%) and the active utilization metrics observed during the execution of each specific algorithm (ranging from 40.0% to 43.7%). Notably, the VRAM consumption remains consistently low, between 0.9 and 1.0 GB during active phases, which underscores the **lightweight structure** of our framework. These metrics are obtained under our default experimental configurations; researchers can further increase the **batch size** to achieve higher GPU utilization and accelerate training based on their specific hardware capabilities. Regarding the CPU and system memory usage, the recorded values reflect the execution environment's status under high-concurrency settings, serving as a baseline reference for implementation.

Table 10: Datasets.

Dataset	Task	Total samples	classes
CIFAR-10	Image classification	60,000	10
CIFAR-100	Image classification	60,000	100

Dataset. The detailed datasets information is shown in Table 10,

- **CIFAR-10**[2]: A standard benchmark for evaluating FL models, providing a balanced yet sufficiently diverse set of visual concepts. Its moderate complexity allows for controlled experimentation while retaining real-world relevance.
- **CIFAR-100** (Krizhevsky et al., 2009): Extends CIFAR-10 with finer-grained categorization challenges through hierarchical class structure, where each coarse category contains multiple fine-grained subclasses. This structure tests model discrimination capacity for visually similar concepts.
- **Tiny-ImageNet** (Le et al., 2015): A large-scale image classification benchmark derived from ImageNet, containing 200 classes with 500 training images per class and images

[2]https://www.cs.toronto.edu/kriz/cifar.html

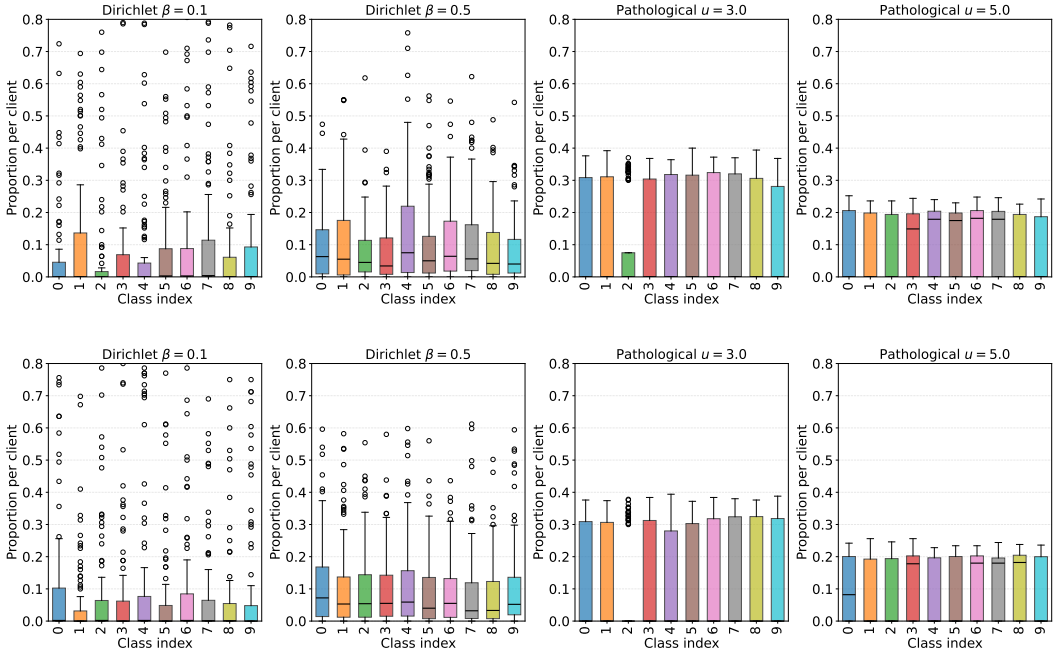


Figure 8: Comparison of Heterogeneous Data Splitting

resized to 64×64 pixels. Its larger class space and increased visual diversity provide a more challenging benchmark for evaluating federated optimization under heterogeneous data.

and Figure 8 shows the comparison of client heterogeneous data distributions under Dirichlet $\beta = \{0.1, 0.5\}$ and pathological $u = \{3, 5\}$ splitting on CIFAR10.

Table 11: Hyperparameter configurations for baseline methods.

Method	local η_l	global η_g	η decay	SAM ρ	proxy coeff.	client-momentum coeff.	KD coeff.
FedAvg	0.1	1.0	0.998	–	–	–	–
FedProx	0.1	1.0	0.998	–	0.1 / 0.01	–	–
SCAFFOLD	0.1	1.0	0.998	–	–	–	–
FedDyn	0.1	1.0	0.9995	–	0.1	–	–
FedCM	0.1	1.0	0.998	–	–	0.1	–
FedCodl	0.1	1.0	0.998	–	–	–	Balanced: 0.5
FedGKD	0.1	1.0	0.998	–	–	–	Incremental: 0.1
FedSAM	0.1	1.0	0.998	0.01	–	–	–
FedLESAM	0.1	1.0	0.998	0.01	–	–	–
FedGAMMA	0.1	1.0	0.998	0.01	–	–	–
FedSMOO	0.1	1.0	0.998	0.1	0.1/0.01	–	–
FedTeas	0.1	1.0	0.998	0.01	–	–	Balanced: 0.6 / Incremental: 0.1
FedTeasC	0.1	1.0	0.998	0.01	–	–	Incremental: 0.1
FedTeasM	0.1	1.0	0.998	0.1	0.1/0.01	–	Incremental: 0.1

Hyperparameter Settings. We distribute each dataset across 100 clients and sample clients at a rate of 0.1 in each communication round. Following prior work, we adopt ResNet-18 (He et al., 2016) with Group Normalization (Wu & He, 2018) as the backbone architecture.Across all methods, we standardize the total number of communication rounds at $T = 500$ and use standard SGD to avoid interference from adaptive optimizers. The global and local learning rates are set to $\eta_g = 1$ and $\eta_l = 0.1$, respectively, with an exponential decay factor of 0.998 for the local learning rate and optimizer weight decay of 10^{-3}. The perturbation radius is set to $\rho = 0.01$ for SAM-based methods, except for FedSMOO and the corresponding FedTeasM, for which $\rho = 0.1$ following the original FedSMOO configuration. The decay factor for the proxy terms in FedDyn and FedSMOO is set to 0.9995. For existing KD baselines, FedCodl and Fed2Codl use balanced distillation with $\alpha_1 = 1 - \alpha_2 = 0.5$, while FedGKD uses incremental KD with $\alpha_1 = 1.0$, $\alpha_2 = 0.1$, and an ensemble

Table 12: Comparison of baseline models.

Model	Corr.	Pert.
ERM		
FedAvg	–	×
FedProx	Proximal term	×
FedDyn	Dynamic regularization	×
SCAFFOLD	Control variates	×
FedCM	Momentum	×
FedCodl	KD / 2-way KD	×
FedGKD	Assembly KD	×
SAM		
FedSAM	–	Local
FedGAMMA	Control variates	Local
FedSMOO	ADMM-style dual	Local with dual
FedLESAM	–	Global
FedTeas Family		
FedTeas	KD	Local
FedTeasC	KD + Control variates	Local
FedTeasM	KD + ADMM-style dual	Local with dual

of two historical global models. Unless otherwise specified, all members of the FedTeas family use $\alpha_1 = 1.0$ and $\alpha_2 = 0.1$ for KD-guided perturbation. For experiments employing dynamic KD scheduling, the relative distillation strength is progressively adjusted during training to balance local empirical optimization and global teacher guidance. For CIFAR-10, each client performs $E = 5$ local epochs with batch size 50, whereas for CIFAR-100 we use $E = 2$ and batch size 20 to reduce overfitting. The complete hyperparameter configurations for all methods are summarized in Table 11.

Loss Volatility. We quantify training stability using the local variation of the loss trajectory. For the loss sequence $\{l_t\}_{t=1}^{T}$, we define a sliding window of size s at round t as

$$W_t = \{l_{t-s}, l_{t-s+1}, \ldots, l_t\}, \tag{49}$$

which is updated at the next round by removing the oldest observation and including the latest loss. The loss volatility at round t is defined as the standard deviation within W_t:

$$\text{Volatility}(W_t) = \left(\mathbb{E}_{k\in[t-s,t]}\left[(l_k - \mathbb{E}(W_t))^2\right]\right)^{1/2}. \tag{50}$$

We further report the interquartile range (IQR) of the resulting volatility sequence to characterize the dispersion of local loss fluctuations throughout training. A lower IQR indicates more consistent optimization dynamics.

F ADDITIONAL EXPERIMENTAL RESULTS

F.1 EXTENDED BASELINE BENCHMARK

Table 13: Test accuracy on CIFAR-10/100 under Dirichlet distribution and Pathological splits. β and u are the Dirichlet and Pathological coefficients, respectively. Each result is calculated 10 times.

Category	Method	CIFAR-10				CIFAR-100				Average
		Dirichlet		Pathological		Dirichlet		Pathological		
		$\beta = 0.5$	$\beta = 0.1$	$u = 5$	$u = 3$	$\beta = 0.5$	$\beta = 0.1$	$u = 20$	$u = 10$	
ERM	FedAvg	$78.92_{\pm.22}$	$73.91_{\pm.08}$	$78.17_{\pm.18}$	$74.82_{\pm.07}$	$46.75_{\pm.04}$	$42.13_{\pm.14}$	$43.50_{\pm.06}$	$38.75_{\pm.11}$	$59.62_{\pm.11}$
	FedProx	$77.56_{\pm.10}$	$72.55_{\pm.23}$	$77.80_{\pm.02}$	$73.90_{\pm.24}$	$36.83_{\pm.33}$	$33.76_{\pm.13}$	$33.95_{\pm.29}$	$30.95_{\pm.11}$	$54.41_{\pm.17}$
	FedDyn	$81.83_{\pm.01}$	$77.01_{\pm.65}$	$81.61_{\pm.30}$	$77.45_{\pm.08}$	$43.13_{\pm.05}$	$37.16_{\pm.20}$	$38.84_{\pm.69}$	$33.83_{\pm.11}$	$58.36_{\pm.26}$
	SCAFFOLD	$81.62_{\pm.64}$	$76.58_{\pm.01}$	$81.12_{\pm.04}$	$77.05_{\pm.29}$	$\mathbf{50.74}_{\pm.08}$	$44.39_{\pm.19}$	$45.51_{\pm.12}$	$\underline{40.33}_{\pm.05}$	$61.42_{\pm.17}$
	FedCM	$82.01_{\pm.24}$	$75.64_{\pm.83}$	$81.34_{\pm.57}$	$78.03_{\pm.79}$	$50.23_{\pm.06}$	$\underline{44.57}_{\pm.35}$	$45.81_{\pm.28}$	$39.18_{\pm.21}$	$61.10_{\pm.41}$
	FedCodl	$78.33_{\pm.32}$	$70.52_{\pm.17}$	$77.79_{\pm.24}$	$72.80_{\pm.59}$	$46.46_{\pm.21}$	$41.21_{\pm.16}$	$43.13_{\pm.67}$	$38.09_{\pm.61}$	$58.17_{\pm.37}$
	Fed2Codl	$78.72_{\pm.30}$	$73.03_{\pm.80}$	$78.52_{\pm.41}$	$74.41_{\pm.31}$	$46.63_{\pm.12}$	$43.93_{\pm.23}$	$43.95_{\pm.31}$	$39.39_{\pm.32}$	$59.32_{\pm.38}$
	FedGKD	$79.25_{\pm.53}$	$73.61_{\pm.24}$	$78.68_{\pm.29}$	$74.71_{\pm.52}$	$46.51_{\pm.08}$	$42.47_{\pm.36}$	$44.30_{\pm.46}$	$39.70_{\pm.49}$	$59.40_{\pm.37}$
SAM	FedSAM	$79.44_{\pm.06}$	$74.39_{\pm.13}$	$79.16_{\pm.24}$	$75.54_{\pm.95}$	$47.04_{\pm.14}$	$42.87_{\pm.01}$	$43.44_{\pm.34}$	$38.79_{\pm.59}$	$60.08_{\pm.31}$
	FedLESAM	$79.00_{\pm.52}$	$73.76_{\pm.08}$	$78.44_{\pm.04}$	$74.72_{\pm.01}$	$46.30_{\pm.40}$	$42.02_{\pm.20}$	$43.85_{\pm.36}$	$38.78_{\pm.19}$	$58.98_{\pm.23}$
	FedTeas	$79.65_{\pm.25}$	$74.53_{\pm.31}$	$79.38_{\pm.05}$	$75.83_{\pm.59}$	$47.39_{\pm.05}$	$42.13_{\pm.20}$	$43.26_{\pm.44}$	$38.92_{\pm.46}$	$60.14_{\pm.29}$
+ Variance Correction	FedGAMMA	$82.55_{\pm.20}$	$76.42_{\pm.36}$	$81.67_{\pm.51}$	$77.31_{\pm.55}$	$\underline{50.44}_{\pm.07}$	$43.60_{\pm.03}$	$45.37_{\pm.03}$	$39.50_{\pm.31}$	$62.11_{\pm.26}$
	FedTeasC	$82.33_{\pm.11}$	$76.79_{\pm.04}$	$81.42_{\pm.02}$	$77.58_{\pm.01}$	$49.98_{\pm.14}$	$43.75_{\pm.15}$	$45.01_{\pm.53}$	$39.79_{\pm.31}$	$62.58_{\pm.16}$
+ Dual Correction	FedSMOO	$\underline{84.81}_{\pm.19}$	$\mathbf{79.68}_{\pm.19}$	$83.84_{\pm.12}$	$\mathbf{80.22}_{\pm.73}$	$50.03_{\pm.55}$	$43.80_{\pm.57}$	$\underline{46.87}_{\pm.05}$	$39.63_{\pm.60}$	$\underline{63.61}_{\pm.38}$
	FedTeasM	$\mathbf{84.84}_{\pm.03}$	$\underline{79.46}_{\pm.52}$	$\mathbf{84.20}_{\pm.01}$	$\underline{80.19}_{\pm.83}$	$49.64_{\pm.32}$	$\mathbf{45.57}_{\pm.44}$	$\mathbf{46.95}_{\pm.26}$	$\mathbf{41.09}_{\pm.27}$	$\mathbf{63.99}_{\pm.33}$

Table 14: Comparison of Test accuracy, on Tiny-ImageNet under Dirichlet and Pathological splitting

Category	Method	Dirichlet		Pathological		Average
		$\beta = 0.5$	$\beta = 0.1$	$u = 20$	$u = 40$	
SAM	FedSAM	33.69 ± 0.48	$\mathbf{29.53 \pm 0.20}$	27.26 ± 0.19	31.38 ± 0.54	30.46 ± 2.55
	FedTeas	$\mathbf{33.83 \pm 0.34}$	29.44 ± 0.01	$\mathbf{27.27 \pm 0.39}$	$\mathbf{31.43 \pm 0.07}$	$\mathbf{30.49 \pm 2.60}$
+ Variance Correction	FedGAMMA	35.58 ± 0.42	28.71 ± 0.23	26.80 ± 0.10	32.27 ± 0.33	30.84 ± 3.61
	FedTeasC	$\mathbf{35.62 \pm 0.32}$	$\mathbf{28.88 \pm 0.04}$	$\mathbf{26.93 \pm 0.18}$	$\mathbf{32.36 \pm 0.15}$	$\mathbf{30.95 \pm 3.56}$
+ Dual Correction	FedSMOO	37.18 ± 0.34	31.94 ± 0.90	30.02 ± 0.52	35.52 ± 1.07	33.66 ± 3.08
	FedTeasM	$\mathbf{37.39 \pm 0.23}$	$\mathbf{32.28 \pm 1.05}$	$\mathbf{30.16 \pm 0.10}$	$\mathbf{35.56 \pm 0.67}$	$\mathbf{33.85 \pm 3.04}$

F.2 MINIMUM LOSS AND LOSS VOLATILITY

Table 15: Comparison of minimum loss and loss volatility, quantified by the interquartile range (IQR), under Pathological splits for CIFAR-10 and CIFAR-100. u denotes the Pathological coefficient.

Category	Method	CIFAR-10				CIFAR-100			
		$u = 5$		$u = 3$		$u = 20$		$u = 10$	
		Loss ↓	IQR ↓	Loss ↓	IQR ↓	Loss ↓	IQR ↓	Loss ↓	IQR ↓
ERM	SCAFFOLD	$0.650_{\pm.001}$	$0.0216_{\pm.0005}$	$0.714_{\pm.002}$	$0.0711_{\pm.0063}$	$2.161_{\pm.018}$	$0.0266_{\pm.0021}$	$2.374_{\pm.014}$	$0.0353_{\pm.0006}$
	FedDyn	$0.605_{\pm.001}$	$0.0304_{\pm.0005}$	$0.714_{\pm.002}$	$0.0805_{\pm.0063}$	$2.427_{\pm.014}$	$0.0312_{\pm.0012}$	$2.629_{\pm.020}$	$0.0805_{\pm.0031}$
	FedGKD	$0.716_{\pm.011}$	$0.0298_{\pm.0028}$	$0.780_{\pm.010}$	$0.0361_{\pm.0063}$	$2.318_{\pm.002}$	$0.0254_{\pm.0002}$	$2.462_{\pm.019}$	$0.0252_{\pm.0010}$
SAM	FedSAM	$0.791_{\pm.001}$	$0.0421_{\pm.0014}$	$0.844_{\pm.005}$	$0.0790_{\pm.0034}$	$2.355_{\pm.020}$	$0.0385_{\pm.0006}$	$2.483_{\pm.017}$	$0.0394_{\pm.0012}$
	FedLESAM	$0.704_{\pm.002}$	$0.0365_{\pm.0009}$	$0.839_{\pm.010}$	$0.0826_{\pm.0048}$	$2.444_{\pm.012}$	$0.0374_{\pm.0004}$	$2.538_{\pm.026}$	$0.0430_{\pm.0014}$
	FedTeas (vanilla)	$0.738_{\pm.001}$	$0.0394_{\pm.0027}$	$0.781_{\pm.003}$	$0.0783_{\pm.0008}$	$2.370_{\pm.015}$	$0.0382_{\pm.0008}$	$2.497_{\pm.011}$	$0.0414_{\pm.0064}$
+Variance Correction	FedGAMMA	$0.610_{\pm.008}$	$0.0257_{\pm.0020}$	$0.784_{\pm.006}$	$0.0831_{\pm.0031}$	$2.152_{\pm.002}$	$0.0232_{\pm.0006}$	$2.405_{\pm.017}$	$0.0376_{\pm.0012}$
	FedTeasC	$0.697_{\pm.003}$	$0.0214_{\pm.0013}$	$0.740_{\pm.007}$	$0.0608_{\pm.0034}$	$2.174_{\pm.023}$	$0.0255_{\pm.0012}$	$2.384_{\pm.025}$	$0.0330_{\pm.0015}$
+Dual Correction	FedSMOO	$0.613_{\pm.001}$	$0.0292_{\pm.0002}$	$0.659_{\pm.010}$	$0.0811_{\pm.0019}$	$2.102_{\pm.001}$	$0.0292_{\pm.0001}$	$2.405_{\pm.031}$	$0.0373_{\pm.0007}$
	FedTeasM	$0.686_{\pm.003}$	$0.0379_{\pm.0021}$	$0.630_{\pm.007}$	$0.0708_{\pm.0034}$	$2.107_{\pm.015}$	$0.0283_{\pm.0008}$	$2.355_{\pm.023}$	$0.0356_{\pm.0011}$

Table 16: Comparison of minimum loss and loss volatility, where volatility is quantified by the interquartile range (IQR), on Tiny-ImageNet under Dirichlet and Pathological splitting.

Category	Method	Dirichlet				Pathological			
		$\beta = 0.5$		$\beta = 0.1$		$u = 20$		$u = 40$	
		Loss ↓	**IQR ↓**	**Loss ↓**	**IQR ↓**	**Loss ↓**	**IQR ↓**	**Loss ↓**	**IQR ↓**
ERM	SCAFFOLD	$2.8782_{\pm.0136}$	$0.0141_{\pm.0002}$	$3.1867_{\pm.0158}$	$0.0181_{\pm.0008}$	$3.3243_{\pm.0051}$	$0.0310_{\pm.0008}$	$3.0689_{\pm.0137}$	$0.0230_{\pm.0005}$
	FedDyn	$3.0185_{\pm.0049}$	$0.0183_{\pm.0012}$	$3.2950_{\pm.0233}$	$0.0211_{\pm.0013}$	$3.3651_{\pm.0168}$	$0.0325_{\pm.0006}$	$3.1791_{\pm.0172}$	$0.0286_{\pm.0015}$
	FedGKD	$3.0999_{\pm.0028}$	$0.0179_{\pm.0000}$	$3.2192_{\pm.0081}$	$0.0182_{\pm.0005}$	$3.2849_{\pm.0190}$	$0.0253_{\pm.0003}$	$3.1672_{\pm.0086}$	$0.0250_{\pm.0004}$
SAM	FedSAM	$3.0641_{\pm.0014}$	$0.0200_{\pm.0010}$	$3.2308_{\pm.0094}$	$0.0242_{\pm.0003}$	$3.3190_{\pm.0057}$	$0.0338_{\pm.0004}$	$3.1752_{\pm.0159}$	$0.0354_{\pm.0000}$
	FedLeSAM	$3.1248_{\pm.0035}$	$0.0203_{\pm.0019}$	$3.2537_{\pm.0202}$	$0.0246_{\pm.0001}$	$3.3329_{\pm.0068}$	$0.0359_{\pm.0013}$	$3.2197_{\pm.0139}$	$0.0356_{\pm.0006}$
	FedTeas	$3.0666_{\pm.0076}$	$0.0192_{\pm.0005}$	$3.2444_{\pm.0007}$	$0.0225_{\pm.0012}$	$3.3172_{\pm.0086}$	$0.0362_{\pm.0003}$	$3.1689_{\pm.0142}$	$0.0313_{\pm.0016}$
+Variance Correction	FedGAMMA	$2.9142_{\pm.0202}$	$0.0133_{\pm.0018}$	$3.2431_{\pm.0081}$	$0.0165_{\pm.0003}$	$3.3362_{\pm.0071}$	$0.0327_{\pm.0020}$	$3.0922_{\pm.0217}$	$0.0217_{\pm.0003}$
	FedTeasC	$2.9067_{\pm.0128}$	$0.0141_{\pm.0001}$	$3.2405_{\pm.0031}$	$0.0182_{\pm.0023}$	$3.3451_{\pm.0102}$	$0.0309_{\pm.0007}$	$3.0814_{\pm.0017}$	$0.0218_{\pm.0008}$
+Dual Correction	FedSMOO	$2.9597_{\pm.0213}$	$0.0130_{\pm.0011}$	$3.1348_{\pm.0174}$	$0.0214_{\pm.0008}$	$3.1136_{\pm.0312}$	$0.0241_{\pm.0016}$	$2.8495_{\pm.0312}$	$0.0205_{\pm.0002}$
	FedTeasM	$2.9548_{\pm.0013}$	$0.0136_{\pm.0002}$	$3.1540_{\pm.0446}$	$0.0192_{\pm.0006}$	$3.1027_{\pm.0070}$	$0.0227_{\pm.0002}$	$2.8574_{\pm.0282}$	$0.0214_{\pm.0000}$

Table 17: Comparison of minimum loss and loss volatility, where volatility is quantified by the interquartile range (IQR), on CIFAR-10 under Dirichlet and Pathological splitting.

Model	CIFAR-10							
	Dirichlet				Pathological			
	$\beta = 0.5$		$\beta = 0.1$		$u = 5$		$u = 3$	
	Loss	IQR	Loss	IQR	Loss	IQR	Loss	IQR
FedAvg	$0.814_{\pm.008}$	$0.0212_{\pm.0002}$	$0.919_{\pm.003}$	$0.0644_{\pm.0081}$	$0.809_{\pm.003}$	$0.0421_{\pm.0014}$	$0.844_{\pm.005}$	$0.0790_{\pm.0034}$
FedProx	$0.743_{\pm.001}$	$0.0185_{\pm.0008}$	$0.906_{\pm.001}$	$0.0533_{\pm.0070}$	$0.751_{\pm.004}$	$0.0363_{\pm.0001}$	$0.834_{\pm.002}$	$0.0799_{\pm.0091}$
FedDyn	$0.597_{\pm.001}$	$0.0128_{\pm.0003}$	$0.740_{\pm.014}$	$0.0552_{\pm.0011}$	$0.605_{\pm.001}$	$0.0304_{\pm.0005}$	$0.714_{\pm.004}$	$0.0805_{\pm.0053}$
SCAFFOLD	$0.641_{\pm.005}$	$0.0095_{\pm.0001}$	$0.796_{\pm.021}$	$0.0349_{\pm.0044}$	$0.650_{\pm.012}$	$0.0216_{\pm.0012}$	$0.748_{\pm.002}$	$0.0711_{\pm.0012}$
FedCM	$0.624_{\pm.005}$	$0.0212_{\pm.0002}$	$0.787_{\pm.004}$	$0.0656_{\pm.0025}$	$0.636_{\pm.003}$	$0.0395_{\pm.0016}$	$0.709_{\pm.001}$	$0.0798_{\pm.0068}$
FedCodl	$0.691_{\pm.006}$	$0.0082_{\pm.0001}$	$0.929_{\pm.001}$	$0.0158_{\pm.0055}$	$0.728_{\pm.003}$	$0.0101_{\pm.0001}$	$0.912_{\pm.005}$	$0.0135_{\pm.0010}$
Fed2Codl	$0.702_{\pm.003}$	$0.0111_{\pm.0010}$	$0.846_{\pm.008}$	$0.0283_{\pm.0094}$	$0.708_{\pm.003}$	$0.0161_{\pm.0012}$	$0.815_{\pm.005}$	$0.0218_{\pm.0020}$
FedGKD	$0.729_{\pm.008}$	$0.0147_{\pm.0012}$	$0.855_{\pm.024}$	$0.0344_{\pm.0088}$	$0.716_{\pm.011}$	$0.0208_{\pm.0020}$	$0.780_{\pm.010}$	$0.0361_{\pm.0053}$
FedSAM	$0.738_{\pm.004}$	$0.0216_{\pm.0014}$	$0.863_{\pm.005}$	$0.0600_{\pm.0090}$	$0.732_{\pm.007}$	$0.0398_{\pm.0027}$	$0.784_{\pm.006}$	$0.0834_{\pm.0031}$
FedGAMMA	$0.596_{\pm.004}$	$0.0097_{\pm.0002}$	$0.741_{\pm.002}$	$0.0402_{\pm.0055}$	$0.610_{\pm.008}$	$0.0257_{\pm.0001}$	$0.726_{\pm.012}$	$0.0620_{\pm.0038}$
FedSMOO	$0.482_{\pm.002}$	$0.0150_{\pm.0018}$	$0.650_{\pm.001}$	$0.0516_{\pm.0026}$	$0.513_{\pm.003}$	$0.0301_{\pm.0014}$	$0.650_{\pm.021}$	$0.0810_{\pm.0019}$
FedLESAM	$0.808_{\pm.012}$	$0.0227_{\pm.0006}$	$0.930_{\pm.004}$	$0.0592_{\pm.0091}$	$0.796_{\pm.004}$	$0.0410_{\pm.0054}$	$0.839_{\pm.010}$	$0.0826_{\pm.0048}$
FedTeas	$0.741_{\pm.008}$	$0.0242_{\pm.0006}$	$0.851_{\pm.001}$	$0.0609_{\pm.0064}$	$0.738_{\pm.003}$	$0.0394_{\pm.0057}$	$0.781_{\pm.003}$	$0.0783_{\pm.0008}$
FedTeasC	$0.593_{\pm.004}$	$0.0103_{\pm.0003}$	$0.730_{\pm.003}$	$0.0364_{\pm.0068}$	$0.607_{\pm.003}$	$0.0214_{\pm.0013}$	$0.740_{\pm.007}$	$0.0608_{\pm.0034}$
FedTeasM	$0.486_{\pm.003}$	$0.0159_{\pm.0021}$	$0.650_{\pm.011}$	$0.0505_{\pm.0038}$	$0.501_{\pm.002}$	$0.0373_{\pm.0057}$	$0.630_{\pm.018}$	$0.0708_{\pm.0140}$

Table 18: Comparison of minimum loss and loss volatility, where volatility is quantified by the interquartile range (IQR), on CIFAR-100 under Dirichlet and Pathological splitting.

Model	**CIFAR-100**							
	Dirichlet				Pathological			
	$\beta = 0.5$		$\beta = 0.1$		$u = 20$		$u = 10$	
	Loss	IQR	Loss	IQR	Loss	IQR	Loss	IQR
FedAvg	$2.373_{\pm.022}$	$0.0220_{\pm.0002}$	$2.481_{\pm.001}$	$0.0300_{\pm.0001}$	$2.450_{\pm.008}$	$0.0386_{\pm.0014}$	$2.543_{\pm.018}$	$0.0432_{\pm.0009}$
FedProx	$2.547_{\pm.012}$	$0.0208_{\pm.0003}$	$2.688_{\pm.003}$	$0.0268_{\pm.0001}$	$2.665_{\pm.001}$	$0.0391_{\pm.0027}$	$2.789_{\pm.007}$	$0.0350_{\pm.0009}$
FedDyn	$2.254_{\pm.003}$	$0.0192_{\pm.0002}$	$2.517_{\pm.004}$	$0.0235_{\pm.0001}$	$2.427_{\pm.014}$	$0.0312_{\pm.0012}$	$2.629_{\pm.020}$	$0.0805_{\pm.0031}$
SCAFFOLD	$1.964_{\pm.006}$	$0.0128_{\pm.0005}$	$2.223_{\pm.010}$	$0.0193_{\pm.0004}$	$2.161_{\pm.018}$	$0.0266_{\pm.0021}$	$2.374_{\pm.014}$	$0.0353_{\pm.0006}$
FedCM	$2.057_{\pm.009}$	$0.0184_{\pm.0005}$	$2.273_{\pm.001}$	$0.0343_{\pm.0043}$	$2.252_{\pm.009}$	$0.0407_{\pm.0009}$	$2.539_{\pm.008}$	$0.0627_{\pm.0074}$
FedCodl	$2.167_{\pm.018}$	$0.0103_{\pm.0002}$	$2.375_{\pm.005}$	$0.0123_{\pm.0001}$	$2.312_{\pm.007}$	$0.0143_{\pm.0003}$	$2.528_{\pm.016}$	$0.0148_{\pm.0005}$
Fed2Codl	$2.217_{\pm.002}$	$0.0152_{\pm.0004}$	$2.377_{\pm.001}$	$0.0180_{\pm.0012}$	$2.321_{\pm.006}$	$0.0199_{\pm.0003}$	$2.512_{\pm.019}$	$0.0236_{\pm.0009}$
FedGKD	$2.317_{\pm.007}$	$0.0183_{\pm.0004}$	$2.390_{\pm.017}$	$0.0216_{\pm.0003}$	$2.318_{\pm.002}$	$0.0254_{\pm.0002}$	$2.462_{\pm.019}$	$0.0252_{\pm.0010}$
FedSAM	$2.291_{\pm.008}$	$0.0219_{\pm.0008}$	$2.392_{\pm.002}$	$0.0277_{\pm.0009}$	$2.356_{\pm.020}$	$0.0385_{\pm.0005}$	$2.483_{\pm.017}$	$0.0394_{\pm.0012}$
FedGAMMA	$1.973_{\pm.011}$	$0.0126_{\pm.0007}$	$2.233_{\pm.010}$	$0.0198_{\pm.0001}$	$2.152_{\pm.006}$	$0.0238_{\pm.0020}$	$2.405_{\pm.031}$	$0.0376_{\pm.0073}$
FedSMOO	$2.001_{\pm.004}$	$0.0141_{\pm.0001}$	$2.242_{\pm.025}$	$0.0205_{\pm.0011}$	$2.104_{\pm.001}$	$0.0292_{\pm.0001}$	$2.405_{\pm.011}$	$0.0373_{\pm.0004}$
FedLESAM	$2.381_{\pm.020}$	$0.0231_{\pm.0004}$	$2.489_{\pm.002}$	$0.0291_{\pm.0003}$	$2.444_{\pm.012}$	$0.0374_{\pm.0004}$	$2.538_{\pm.026}$	$0.0430_{\pm.0014}$
FedTeas	$2.274_{\pm.005}$	$0.0201_{\pm.0012}$	$2.418_{\pm.006}$	$0.0293_{\pm.0003}$	$2.370_{\pm.015}$	$0.0382_{\pm.0008}$	$2.497_{\pm.011}$	$0.0414_{\pm.0054}$
FedTeasC	$1.975_{\pm.002}$	$0.0122_{\pm.0006}$	$2.226_{\pm.003}$	$0.0190_{\pm.0003}$	$2.170_{\pm.023}$	$0.0255_{\pm.0012}$	$2.384_{\pm.025}$	$0.0330_{\pm.0015}$
FedTeasM	$2.009_{\pm.019}$	$0.0137_{\pm.0012}$	$2.182_{\pm.017}$	$0.0212_{\pm.0023}$	$2.107_{\pm.015}$	$0.0283_{\pm.0008}$	$2.355_{\pm.023}$	$0.0356_{\pm.0011}$

F.3 Loss Landscape

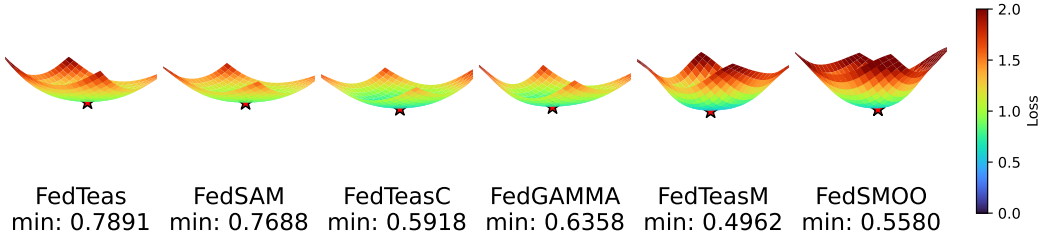


Figure 9: Comparison of 3D Loss Landscape

Table 19: The minimum loss landscape of the client over global test set on cifar 10 under Pathological $u = 5$ splitting

Client	FedAvg	FedGKD	FedSAM	FedLESAM	FedGamma	FedSMOO	FedTeas	FedTeasC	FedTeasM
8	1.7174	1.1294	1.4597	1.9266	1.1297	0.7938	1.3916	1.0598	0.7759
18	1.8417	1.3144	1.8049	2.0611	1.2170	1.0136	1.7026	1.2048	0.8715
29	2.0287	1.4295	1.8847	1.9985	1.0991	1.1023	1.8172	1.1683	0.8945
40	1.5402	1.1570	1.3163	1.6834	0.9284	0.7515	1.5099	0.9139	0.6289
44	1.5871	1.2913	1.4650	1.6743	0.9969	0.9510	1.7109	0.9978	0.7146
47	1.6299	1.1958	1.6659	1.6747	1.0114	0.8507	1.6733	1.1658	0.8721
55	1.8283	1.3671	1.7005	1.8784	1.0715	1.0887	1.8064	1.0501	0.7912
59	1.7576	1.0981	1.6650	1.7461	0.9169	0.8493	1.3694	1.0316	0.6837
69	1.8465	1.3777	1.7138	1.8714	1.2988	1.0771	1.9191	1.0811	0.7275
70	1.9094	1.4584	1.8154	1.9183	1.2861	1.0325	1.8364	1.1208	0.8537
Average	1.7687	1.2819	1.6491	1.8433	1.0956	0.9511	1.6737	1.0794	**0.7814**

F.4 HESSIAN RESULTS

Table 20: Hessian Top Eigenvalue (Batch Size = 64)

Algorithm	Seed 1	Seed 23	Seed 42	Seed 66	Seed 88	Average
FedTeasM	143.27	170.07	143.71	119.97	112.01	137.81
FedSMOO	236.00	118.79	163.85	148.46	190.02	171.42
FedGamma	235.05	232.68	247.23	227.76	246.11	237.77
FedTeasC	297.56	286.44	309.75	313.97	287.55	299.06
Decoupled FedTeas	351.43	315.10	310.58	318.31	323.55	323.80
FedTeas	292.80	472.77	342.52	427.22	311.66	369.39
FedSAM	478.18	397.70	380.83	373.16	402.58	406.49
FedGKD	342.58	487.81	471.51	488.66	378.15	433.74
FedAvg	494.79	532.22	580.89	534.13	709.07	570.22
FedLESAM	628.80	701.60	678.51	826.66	625.75	692.26

Table 21: Hessian Top Eigenvalue (Batch Size = 128)

Algorithm	Seed 1	Seed 23	Seed 42	Seed 66	Seed 88	Average
FedTeasM	86.91	76.28	77.10	80.90	67.90	77.82
FedSMOO	93.75	87.31	113.58	93.53	93.87	96.41
Decoupled FedTeas	213.07	215.27	205.39	193.40	183.45	202.12
FedGamma	230.94	200.01	218.58	203.28	223.07	215.18
FedSAM	218.39	209.64	217.47	213.98	247.54	221.41
FedTeas	315.00	204.01	172.20	198.32	316.35	241.17
FedGKD	299.61	267.28	227.81	260.63	287.12	268.49
FedTeasC	266.97	297.06	296.33	299.28	282.34	288.40
FedAvg	319.86	393.18	408.68	344.58	364.38	366.14
FedLESAM	517.46	386.97	382.60	357.37	372.53	403.39

Table 22: Hessian Top Eigenvalue (Batch Size = 256)

Algorithm	Seed 1	Seed 23	Seed 42	Seed 66	Seed 88	Average
FedTeasM	61.00	63.82	52.03	69.11	68.16	62.82
FedSMOO	66.92	72.70	71.93	74.38	74.78	72.14
Decoupled FedTeas	99.01	101.85	99.25	122.12	135.71	111.59
FedTeas	247.75	108.72	111.14	107.41	115.51	138.11
FedSAM	133.02	135.86	171.81	122.23	135.70	139.73
FedGKD	191.43	191.70	144.91	147.02	212.86	177.59
FedGamma	214.30	200.38	225.35	191.42	249.64	216.22
FedAvg	219.88	216.77	227.88	220.23	273.02	231.56
FedLESAM	243.20	227.89	281.71	221.76	311.02	257.12
FedTeasC	290.07	283.87	286.90	287.94	283.10	286.38

Table 23: Hessian Top Eigenvalue - Batch Size Comparison (Sorted by BS=256 Average)

Algorithm	**Average Hessian**		
	BS=64	**BS=128**	**BS=256**
FedTeasM	137.81	77.82	62.82
FedSMOO	171.42	96.41	72.14
Decoupled FedTeas	323.80	202.12	111.59
FedTeas	369.39	241.17	138.11
FedSAM	406.49	221.41	139.73
FedGKD	433.74	268.49	177.59
FedGamma	237.77	215.18	216.22
FedAvg	570.22	366.14	231.56
FedLESAM	692.26	403.39	257.12
FedTeasC	299.06	288.40	286.38