FEDCADC: CURVATURE-AWARE DIRECTION CONSISTENCY FOR GLOBALLY FLAT FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite the success of federated learning (FL), improving global generalization under heterogeneous client data remains challenging. To improve the generalization of FL, recent methods seek flatter minima by applying sharpness-aware optimization during local training. However, differences in local curvature and update directions can make locally flat solutions inconsistent after aggregation, such that local flatness does not necessarily translate into global flatness. To address this issue, we propose Federated Learning with Curvature-Aware Direction Consistency (FedCADC). FedCADC uses a curvature-aware constraint to suppress local updates toward perturbation-sensitive regions and a direction-consistency constraint to align local trajectories with the global optimization trend. Their weights are adaptively adjusted according to the optimization geometry of each mini-batch. Our theoretical analysis characterizes the optimization, generalization, and flatness discrepancy of FedCADC. Experiments on CIFAR-10 and CIFAR-100 demonstrate that FedCADC achieves higher accuracy in most heterogeneous settings and generally finds flatter global solutions than recent advanced FL methods.

1 INTRODUCTION

Federated learning (FL) enables multiple clients to collaboratively train a shared model without transmitting their raw data (McMahan et al., 2017; Kairouz & McMahan, 2021). However, client data are often non-independent and identically distributed (non-IID), resulting in heterogeneous local objectives and client drift (Zhao et al., 2018; Hsu et al., 2019). As local training proceeds, clients may update toward conflicting directions, causing direct parameter averaging to produce biased global optimization and weaker generalization (Li et al., 2020; Karimireddy et al., 2020; Wang et al., 2020).

Existing methods mitigate client drift through proximal constraints, control variates, dynamic regularization, or server-side optimization (Li et al., 2020; Karimireddy et al., 2020; Acar et al., 2021; Reddi et al., 2020; Wang et al., 2020; Li et al., 2021). These approaches mainly correct local updates or modify aggregation, but do not explicitly control the geometry of the loss landscape around the global model. Flatness-driven methods address this issue by incorporating sharpness-aware optimization, parameter-scaled perturbations, global perturbation correction, or trajectory averaging (Foret et al., 2020; Kwon et al., 2021; Izmailov et al., 2018; Qu et al., 2022; Caldarola et al., 2022; Sun et al., 2023; Dai et al., 2023; Lee & Yoon, 2024; Liu et al., 2025). Nevertheless, under heterogeneous data, clients exhibit different local minima, curvature responses, and update directions. Consequently, reducing local sharpness does not necessarily yield a flat aggregated model. Moreover, existing methods commonly use fixed correction coefficients and do not fully coordinate local curvature responses with the global optimization trajectory. This raises a key question: how can local flatness-seeking updates be coordinated to improve the flatness of the global model?

To address this challenge, we propose FedCADC (Federated Learning with Curvature-Aware Direction Consistency). FedCADC introduces two complementary constraints into local optimization. The curvature-aware constraint compares the directions and norms of the center and perturbed gradients, thereby capturing the sensitivity of the local loss landscape and suppressing updates toward sharp regions. The global direction-consistency constraint estimates the global optimization trend from consecutive global models and penalizes local updates that deviate excessively from this direction. FedCADC further assigns adaptive weights to these constraints according to the optimization

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.

geometry of each mini-batch, strengthening them when curvature response or directional deviation increases and relaxing them when local optimization is stable. This design coordinates local flatness seeking with the global optimization trend while preserving client-specific learning. Theoretical analysis establishes bounds connecting curvature response and client-update deviation to global flatness. The main contributions of this work are summarized as follows:

- We propose FedCADC for heterogeneous FL. It combines curvature-aware flatness regularization with global direction consistency and adaptively adjusts both constraints using the mini-batch optimization geometry.
- We establish optimization, generalization, and flatness discrepancy bounds for FedCADC. The analysis shows that, under specified geometric alignment conditions, direction consistency reduces the local–global gradient discrepancy and promotes globally aligned flatness seeking.
- We evaluate FedCADC against conventional and flatness-driven FL methods under diverse heterogeneous settings. Extensive experiments and loss-geometry analyses demonstrate its effectiveness in improving generalization and finding flatter global solutions.

2 PRELIMINARIES

2.1 FL PRELIMINARIES: FEDAVG

Consider an FL system with m clients. Client i owns a local dataset $S_i = \{z_{i,j}\}_{j=1}^{n_i}$ and has empirical risk

$$F_{S_i}(w) = \frac{1}{n_i} \sum_{j=1}^{n_i} \ell(w; z_{i,j}), \tag{1}$$

where $w \in \mathbb{R}^d$ denotes the model parameters. The global FL objective is the sample-weighted average

$$\min_w F_S(w), \qquad F_S(w) = \sum_{i=1}^{m} p_i F_{S_i}(w), \qquad p_i = \frac{n_i}{\sum_{j=1}^{m} n_j}. \tag{2}$$

FedAvg alternates between local optimization and server aggregation (McMahan et al., 2017). At round t, the server broadcasts w^t to a selected client set $\mathcal{S}_t$. Each client initializes $w_i^{t,0} = w^t$ and performs K local SGD steps:

$$w_i^{t,k+1} = w_i^{t,k} - \eta_l \nabla F_{\xi_i^{t,k}}(w_i^{t,k}), \qquad k = 0, \dots, K-1, \tag{3}$$

where $\xi_i^{t,k}$ is a mini-batch and η_l is the local learning rate. The server then aggregates the uploaded models as

$$w^{t+1} = \sum_{i \in \mathcal{S}_t} p_i^t w_i^{t,K}, \qquad p_i^t = \frac{n_i}{\sum_{j \in \mathcal{S}_t} n_j}. \tag{4}$$

Under non-IID data, local distributions and objectives differ across clients. Consequently, multiple local updates may follow distinct directions and accumulate client drift, making the aggregated model different from the minimizer of the global objective.

2.2 FLATNESS-SEEKING OPTIMIZATION: SAM

For a loss function $F(w)$, the sharpness within a neighborhood of radius $\rho > 0$ is defined as

$$\mathrm{Sh}_\rho(F, w) = \max_{\|\epsilon\|_2 \le \rho} \left[F(w+\epsilon) - F(w) \right]. \tag{5}$$

SAM seeks parameters with small worst-case perturbed loss (Foret et al., 2020):

$$\min_w \max_{\|\epsilon\|_2 \le \rho} F(w+\epsilon) = \min_w \left[F(w) + \mathrm{Sh}_\rho(F, w) \right]. \tag{6}$$

In practice, SAM approximates the inner maximization using the current mini-batch gradient. Given $F_\xi(w)$, the center gradient, perturbation, and perturbed gradient are

$$g_c = \nabla F_\xi(w), \qquad \epsilon = \rho \frac{g_c}{\|g_c\|_2 + \varepsilon_{\text{num}}}, \qquad g_p = \nabla F_\xi(w + \epsilon), \tag{7}$$

where $\varepsilon_{\text{num}} > 0$ ensures numerical stability. If the Hessian is locally Lipschitz-continuous, then

$$g_p - g_c = \nabla^2 F_\xi(w)\epsilon + \mathcal{O}(\|\epsilon\|_2^2). \tag{8}$$

Thus, the change in gradient direction and norm provides local information about the curvature response to parameter perturbation.

3 Related Work

Federated learning (FL) must reconcile locally optimized models under statistically heterogeneous client data, which can cause client drift and aggregation bias. Existing methods address these issues through local regularization, update correction, and server-side optimization. FedProx and FedDyn regularize local models, while SCAFFOLD and FedDC correct local update bias using control variates or historical information (Li et al., 2020; Acar et al., 2021; Karimireddy et al., 2020; Gao et al., 2022). FedNova, FedAvgM, and FedOpt further modify aggregation or server optimization to reduce the impact of heterogeneous local computation (Wang et al., 2020; Reddi et al., 2020). Other methods improve heterogeneous FL through knowledge distillation, representation alignment, or prototype sharing (Li et al., 2021; Lin et al., 2020; Zhu et al., 2021; Lee et al., 2022; Tan et al., 2022). However, these methods mainly regulate updates, gradients, or representations, without explicitly modeling client-specific loss geometry.

Flatness-based methods seek solutions that are less sensitive to parameter perturbations, which is commonly associated with improved generalization (Hochreiter & Schmidhuber, 1997; Keskar et al., 2016; Li et al., 2018). In centralized learning, SWA, SAM, and ASAM promote flat solutions through trajectory averaging, worst-case neighborhood optimization, and scale-aware perturbations (Izmailov et al., 2018; Foret et al., 2020; Kwon et al., 2021). Their federated variants include FedSAM, FedASAM, and MoFedSAM, which apply sharpness-aware optimization during local training, and FedMut, which uses stochastic gradient mutations (Qu et al., 2022; Hu et al., 2024). Recent approaches additionally exploit server-side or cross-client information. Caldarola et al. (2022) combine stochastic weight averaging with FedAvg, FedSAM, and FedASAM, while FedSMOO, FedGAMMA, and FedGF introduce correction or global perturbation information to reduce local–global flatness mismatch (Sun et al., 2023; Dai et al., 2023; Lee & Yoon, 2024). FedSWA improves generalization through stochastic weight averaging under highly heterogeneous data (Liu et al., 2025). FedGloSS estimates global sharpness using server-side training information (Caldarola et al., 2025), FedVSSAM uses a variance-suppressed direction for perturbation and updating (Xiao et al., 2026), and FedScar corrects heterogeneity-induced geometric bias to improve flatness consistency (Lu et al., 2026).

Despite these advances, existing methods generally emphasize one aspect of the local–global geometric problem, such as local sharpness, global perturbation, model averaging, or directional correction. They do not explicitly separate perturbation-induced curvature response from empirical-risk descent while simultaneously constraining local trajectory orientation. FedCADC addresses this gap with a curvature-aware flatness constraint and a global direction-consistency constraint. Their adaptive coefficients are determined by mini-batch optimization geometry, allowing the method to regulate local sensitivity and directional deviation according to the current optimization state.

4 Methodology

For client i and the global objective, define

$$\text{Sh}_{i,\rho}(w_i) = \text{Sh}_\rho(F_{S_i}, w_i), \qquad \text{Sh}_{S,\rho}(w) = \text{Sh}_\rho(F_S, w). \tag{9}$$

The corresponding local–global flatness discrepancy is

$$\Delta_{\text{flat}} = \left| \text{Sh}_{S,\rho}(w) - \sum_{i=1}^{m} p_i \, \text{Sh}_{i,\rho}(w_i) \right|. \tag{10}$$

Under heterogeneous data, local flat regions occur at different locations and exhibit different curvature structures. Therefore, small local sharpness does not necessarily imply a flat aggregated model. In addition, inconsistent local update directions prevent client trajectories from supporting a common globally flat region. Therefore, to seek flatness in heterogeneous FL, two requirements need to be met: local updates should suppress perturbation-sensitive curvature responses, and their optimization directions should remain consistent with the global optimization trend. FedCADC is designed to satisfy both requirements through its curvature-aware and direction-consistency constraints.

4.1 FEDCADC

4.1.1 CURVATURE-AWARE FLATNESS CONSTRAINT

FedCADC treats the perturbation as a local geometric probe rather than directly using the perturbed gradient as the update gradient. It explicitly regularizes the perturbation-induced loss increase, thereby separating empirical-risk descent from curvature correction and enabling the correction strength to adapt to the current mini-batch geometry.

At local step k of communication round t, client i computes the center gradient, constructs a perturbation along its direction, and evaluates the gradient at the perturbed model:

$$g_{c,i}^{t,k} = \nabla F_{\xi_i^{t,k}}(w_i^{t,k}), \qquad \epsilon_i^{t,k} = \rho \frac{g_{c,i}^{t,k}}{\|g_{c,i}^{t,k}\|_2 + \delta}, \qquad g_{p,i}^{t,k} = \nabla F_{\xi_i^{t,k}}\left(w_i^{t,k} + \epsilon_i^{t,k}\right), \tag{11}$$

where $\rho > 0$ is the perturbation radius and $\delta > 0$ ensures numerical stability. FedCADC then defines

$$R_{\mathrm{curv},i}^{t,k}(w) = \frac{\left[F_{\xi_i^{t,k}}(w + \epsilon_i^{t,k}) - F_{\xi_i^{t,k}}(w)\right]_+}{\rho^2}. \tag{12}$$

This quantity measures the positive loss response along the current gradient direction. As established in Proposition 1, its local expansion contains both the first-order loss increase and the directional Hessian curvature. Dividing by ρ^2 retains the directional curvature term at its natural scale, while $[\cdot]_+$ prevents a loss-decreasing perturbation from being rewarded as negative curvature. Thus, $R_{\mathrm{curv},i}^{t,k}$ characterizes directional geometric sensitivity rather than the unrestricted maximum Hessian eigenvalue.

Unlike SAM, which updates with $g_{p,i}^{t,k}$ to approximate worst-case loss minimization, FedCADC uses the perturbation as a geometric probe while retaining the center loss. It introduces $R_{\mathrm{curv},i}^{t,k}$ as a separate regularizer whose correction, under stop-gradient, is proportional to $(g_{p,i}^{t,k} - g_{c,i}^{t,k})/\rho^2$ on the positive branch (Proposition 2). To account for both gradient rotation and magnitude variation, its adaptive coefficient is

$$\alpha_{\mathrm{curv},i}^{t,k} = \mathrm{clip}\left(\left[1 - \cos(g_{p,i}^{t,k}, g_{c,i}^{t,k})\right] \left| \log \frac{\left\|g_{p,i}^{t,k}\right\|_2 + \delta}{\left\|g_{c,i}^{t,k}\right\|_2 + \delta} \right|, 0, A_{\mathrm{curv}}\right). \tag{13}$$

Thus, curvature suppression is strengthened when the perturbation substantially changes the local gradient geometry and weakened in stable regions.

4.1.2 GLOBAL DIRECTION CONSISTENCY CONSTRAINT

The curvature-aware constraint controls the perturbation sensitivity of each client, but it does not ensure that the resulting flat regions are geometrically compatible. As motivated by equation 10, conflicting local trajectories can disperse client models across different flat regions, such that their aggregation may yield a globally sharp model. FedCADC addresses this inconsistency by using the latest global displacement as a shared optimization reference. Specifically, the server computes

$$g^t = \frac{w^{t-1} - w^t}{\|w^{t-1} - w^t\|_2 + \delta}, \qquad d_i^{t,k} = w^t - w_i^{t,k}, \tag{14}$$

where $d_i^{t,k}$ denotes the cumulative local trajectory from the current global model. FedCADC measures its angular inconsistency with g^t using

$$R_{\mathrm{dir},i}^{t,k}(w) = 1 - \cos\left(w^t - w, g^t\right), \qquad \alpha_{\mathrm{dir},i}^{t,k} = \mathrm{clip}\left(R_{\mathrm{dir},i}^{t,k}(w_i^{t,k}), 0, A_{\mathrm{dir}}\right). \tag{15}$$

Because $R_{\mathrm{dir},i}^{t,k}$ depends on angular deviation rather than trajectory magnitude, it discourages incompatible update components without directly restricting local progress. The adaptive coefficient strengthens this correction only for substantially misaligned clients, encouraging their endpoints to support a common flat region.

Proposition 3 characterizes the alignment effect of this constraint, while Theorem 1 shows that it can contract the mean-squared client-update deviation under the stated geometric condition. Since this deviation contributes to the model-drift term in Theorem 4, direction consistency provides a mechanism for tightening the local–global flatness-discrepancy bound.

The curvature-aware and direction-consistency constraints address complementary sources of global-flatness mismatch. FedCADC combines them with mini-batch-dependent adaptive coefficients:

$$\Phi_i^{t,k}(w) = F_{\xi_i^{t,k}}(w) + \alpha_{\mathrm{curv},i}^{t,k} R_{\mathrm{curv},i}^{t,k}(w) + \alpha_{\mathrm{dir},i}^{t,k} R_{\mathrm{dir},i}^{t,k}(w). \tag{16}$$

The local update is

$$w_i^{t,k+1} = w_i^{t,k} - \eta_\ell \nabla \Phi_i^{t,k}(w_i^{t,k}). \tag{17}$$

During differentiation, $\epsilon_i^{t,k}$ and both adaptive coefficients are treated as fixed. This allows FedCADC to apply stronger corrections only when local curvature or client drift requires them. After local training, the server performs the sample-size-weighted aggregation in equation 4. A detailed comparison between FedCADC and existing methods is provided in Appendix D.

5 Theoretical Analysis

This section summarizes the geometric properties of FedCADC and its guarantees for optimization, generalization, and flatness. The complete assumptions, formal theorem statements, and step-by-step proofs are provided in Appendix A. Our analysis only requires local smoothness along the training trajectory and within its perturbation neighborhood, bounded stochastic variation and data heterogeneity, and finite client-sampling errors. It does not require convexity, strong convexity, or globally bounded gradients.

Geometric effects. Under the stop-gradient treatment, the effective correction induced by the curvature regularizer satisfies

$$r_{\mathrm{curv},i}(w) = \frac{\vartheta_i(w)}{\rho^2}\left[\nabla F_i(w + \epsilon_i(w)) - \nabla F_i(w)\right], \qquad \vartheta_i(w) \in [0, 1], \tag{18}$$

where $\vartheta_i(w) = 1$ when the perturbation strictly increases the loss. This correction therefore directly responds to the gradient variation induced by the perturbation, while its strength is jointly modulated by changes in gradient direction and norm. Moreover, an update induced solely by the direction regularizer does not decrease the cosine similarity between the local trajectory and the historical global direction. If the negative cross term between the direction correction and the original client deviation dominates the additional second-order perturbation, Theorem 1 in Appendix A establishes that

$$D_{\mathrm{dir}}^2 \le \kappa^2 D_{\mathrm{base}}^2 + r, \qquad 0 \le \kappa < 1, \tag{19}$$

where D_{base}^2 and D_{dir}^2 denote the mean-squared client-update deviations before and after introducing the direction regularizer, respectively. The deviation contracts strictly when $r < (1 - \kappa^2) D_{\mathrm{base}}^2$.

Optimization and generalization guarantees. Let $\overline{D}^2$, $\overline{U}^2$, and $\overline{\omega}^2$ denote the average client-update deviation, the conditional mean bias induced by the regularizers and optimizer, and the aggregation error caused by partial client participation, respectively. Under a fixed step size satisfying $\eta^2 K^2 L^2 \le 1/48$, Theorem 2 in Appendix A gives

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(w^t)\|^2 = \mathcal{O}\left[\frac{F(w^0) - F_\star}{\eta K T} + \overline{\omega}^2 + \overline{U}^2 + \eta^2 K^2 L^2 (\overline{D}^2 + \sigma^2)\right]. \tag{20}$$

If $\eta = c_\eta/(L\sqrt{TK})$ is selected before training according to the total training budget and then kept fixed, the leading optimization term scales as $\mathcal{O}(1/\sqrt{TK})$. Based on expected replace-one stability, Theorem 3 in Appendix A further establishes

$$\epsilon_{\text{gen}} = \mathcal{O}\left[\frac{L_\ell M \eta TK B_{\text{eff}}}{N_{\text{tr}}} \exp(L_\Phi \eta TK)\right], \tag{21}$$

where $B_{\text{eff}}/N_{\text{tr}}$ controls the probability of accessing the replaced sample at each step. The constants M and L_Φ control the update magnitude and the propagation of state perturbations, respectively. When $L_\Phi \eta TK = \mathcal{O}(1)$, the exponential factor can be absorbed into a constant.

Flatness discrepancy and theoretical comparison. Let ζ denote gradient heterogeneity, and let w_i and b_i denote the client parameters and inference states, respectively. Theorem 4 in Appendix A shows that

$$\Delta_F \leq \rho\zeta + L\rho^2 + L\rho \sum_i p_i \|w - w_i\| + L_{b,\rho} \sum_i p_i \|b - b_i\|. \tag{22}$$

This bound separates the flatness discrepancy into data heterogeneity, neighborhood approximation, model drift, and inference-state discrepancy. The curvature regularizer limits the feasible-direction loss increase, while the direction regularizer contracts client-update deviation under equation 19. If this contraction offsets the regularization bias without increasing curvature or coverage errors, FedCADC matches or tightens the paired baseline's optimization, stability, and global-flatness bounds, with strict improvement when at least one controlled term decreases.

6 Experiments

We extensively evaluate FedCADC on CIFAR-10 and CIFAR-100 under FL settings following the benchmark protocol of Caldarola et al. (2022). The datasets are partitioned among clients according to Dirichlet distributions with different concentration parameters.

6.1 Experimental Settings

Baselines. We compare FedCADC with the following methods: FedAvg (McMahan et al., 2017), FedProx (Li et al., 2020), SCAFFOLD (Karimireddy et al., 2020), FedMut (Hu et al., 2024), FedSAM (Qu et al., 2022), FedASAM (Caldarola et al., 2022), MoFedSAM (Qu et al., 2022), FedGAMMA (Dai et al., 2023), FedSMOO (Sun et al., 2023), FedGF (Lee & Yoon, 2024), FedSWA (Liu et al., 2025), FedVSSAM (Xiao et al., 2026), FedGloSS (Caldarola et al., 2025), and FedScar (Lu et al., 2026).

Model architecture. We use ResNet-20 as the model architecture for all experiments on CIFAR-10 and CIFAR-100. ResNet-20 follows the standard residual network design for image classification (He et al., 2016). To ensure a fair comparison, all baseline methods and FedCADC use the same model architecture and parameter initialization under each experimental setting.

FL settings. For a given server and 100 clients, we test three different numbers of participating clients per round, i.e., $\{5, 10, 20\}$. For both CIFAR-10 and CIFAR-100, the training data are partitioned among clients according to a Dirichlet distribution with concentration parameter $\alpha_{\text{Dir}} \in \{0.1, 0.5, 10\}$. A smaller α_{Dir} produces more imbalanced class distributions across clients and therefore represents a stronger non-IID setting, whereas a larger α_{Dir} yields more balanced client distributions. All experiments run for 500 communication rounds, with 5 local epochs per round and a batch size of 50. Further details of the benchmarks, hyperparameters, and model architecture are provided in Appendix B.

6.2 Performance Evaluation

6.2.1 Performance under Data Heterogeneity

Table 1 shows that FedCADC achieves the highest accuracy in 12 of the 18 settings. Averaged over nine configurations, it outperforms FedAvg by 20.64 and 16.65 percentage points on CIFAR-10 and

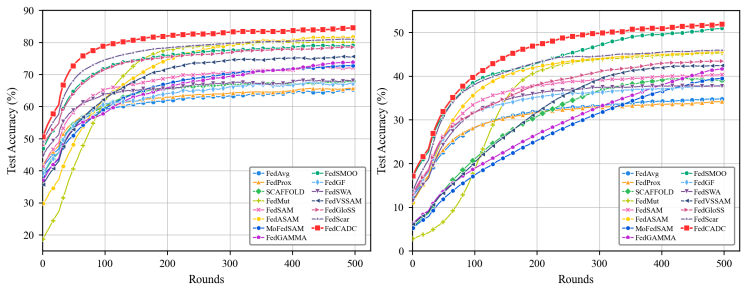
CIFAR-100, respectively, and FedSAM by 11.04 and 11.28 points. Its advantage is particularly evident under moderate heterogeneity and broader client participation: with $\alpha_{\mathrm{Dir}} = 0.5$, FedCADC ranks first in five of the six settings and performs best on both datasets with 20 clients. Under stronger heterogeneity ($\alpha_{\mathrm{Dir}} = 0.1$), FedSMOO and FedScar perform better with five clients on CIFAR-10 and CIFAR-100, respectively. With 20 clients, FedCADC instead performs best on both datasets, reaching 77.52% and 49.86%, respectively.

This pattern is consistent with the direction-consistency mechanism. The reference g^t in equation 14 is estimated from consecutive global models. Severe heterogeneity and limited participation make the sampled update less representative, increasing the partial-participation error $\overline{\omega}^2$ in Theorem 2 and reducing the reliability of g^t. Theorem 1 guarantees strict contraction only when $r < (1-\kappa^2)D_{\mathrm{base}}^2$. An inaccurate reference can weaken the favorable cross term and increase r, making this condition harder to satisfy. Corollary 2 similarly requires the contraction benefit to exceed the bias introduced by direction regularization.

Table 1: Test accuracy (%) of FL algorithms on CIFAR-10 and CIFAR-100 under different Dirichlet distributions and client participation settings.

Task	**Algorithms**	**Dirichlet distribution parameter** α_{Dir}								
		$\alpha_{\mathrm{Dir}} = 0.1$			$\alpha_{\mathrm{Dir}} = 0.5$			$\alpha_{\mathrm{Dir}} = 10$		
		Number of participating clients per round								
		5	10	20	5	10	20	5	10	20
CIFAR-10	**FedCADC**	63.94	72.28	**77.52**	78.85	**84.29**	**84.63**	84.34	**87.22**	**86.57**
	FedAvg	44.20	41.41	40.98	63.72	64.90	63.38	73.68	71.69	69.91
	FedProx	43.08	44.41	44.58	65.43	65.43	64.31	73.90	72.23	70.05
	SCAFFOLD	48.16	52.90	54.43	66.07	67.62	68.66	69.90	68.71	68.34
	FedMut	57.07	63.99	68.33	78.18	80.00	79.16	80.68	82.21	81.56
	FedSAM	54.41	57.81	61.40	71.69	72.01	72.11	77.98	77.26	75.57
	FedASAM	57.46	65.73	70.48	78.67	81.58	82.80	**84.44**	84.56	84.73
	MoFedSAM	53.12	60.55	62.74	71.90	72.65	73.41	76.48	75.29	75.80
	FedGAMMA	51.57	54.79	56.60	74.06	73.67	74.52	78.91	78.73	78.88
	FedSMOO	**68.16**	73.01	75.13	78.74	78.94	78.75	80.65	81.81	81.38
	FedGF	46.46	51.56	50.75	66.64	66.98	66.62	76.48	75.57	73.02
	FedSWA	50.20	44.34	48.26	69.80	67.84	67.47	74.01	73.31	72.28
	FedVSSAM	57.62	63.25	68.18	74.50	75.48	75.42	78.97	77.84	77.26
	FedGloSS	57.99	70.43	69.09	78.52	78.42	76.55	80.28	80.48	79.15
	FedScar	65.98	**73.44**	76.68	**79.66**	81.00	79.57	82.78	81.33	80.56
CIFAR-100	**FedCADC**	40.75	**50.44**	**49.86**	**49.01**	**51.68**	**51.59**	**50.27**	**53.92**	51.23
	FedAvg	28.39	29.12	29.40	35.19	34.73	33.91	37.49	36.38	34.25
	FedProx	29.96	31.08	30.32	36.04	34.03	34.45	38.05	36.09	35.07
	SCAFFOLD	29.12	32.66	37.87	39.28	39.09	41.21	40.60	42.21	41.82
	FedMut	39.42	44.42	42.81	45.27	45.25	45.66	47.22	47.02	44.47
	FedSAM	33.54	35.40	36.55	39.39	40.22	39.76	41.95	39.98	40.43
	FedASAM	38.39	40.58	43.84	45.51	45.48	46.27	46.97	46.38	46.22
	MoFedSAM	32.64	35.92	37.36	38.60	39.19	39.86	38.60	39.21	39.84
	FedGAMMA	36.64	38.75	38.35	41.79	41.24	40.32	41.50	40.76	40.31
	FedSMOO	42.82	46.35	48.48	48.62	50.67	51.34	48.08	51.06	**51.55**
	FedGF	30.70	32.08	33.18	37.34	37.75	36.84	39.37	38.97	37.05
	FedSWA	32.24	39.70	32.36	38.25	37.70	37.47	41.65	39.70	38.46
	FedVSSAM	34.25	38.31	39.88	41.56	42.36	42.43	43.49	42.56	42.11
	FedGloSS	40.81	40.93	40.23	42.50	43.40	41.74	41.70	42.43	42.13
	FedScar	**43.28**	44.68	46.68	48.05	45.82	47.17	47.72	45.50	45.55

The reported values are test accuracies averaged over the last 50 communication rounds. Bold values indicate the highest test accuracy in each column.



(a) CIFAR-10 ($\alpha_{\text{Dir}} = 0.5$) (b) CIFAR-100 ($\alpha_{\text{Dir}} = 0.5$)

Figure 1: Convergence behavior under non-IID data distributions.

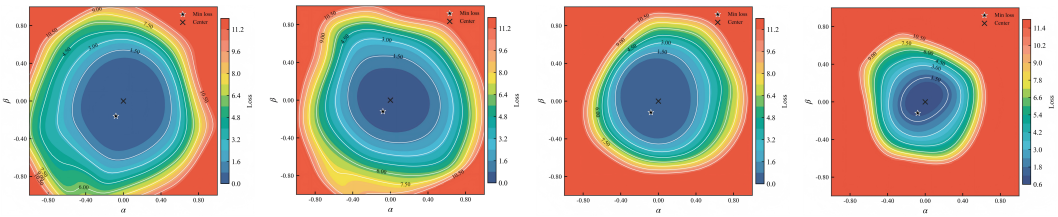
Table 2: Flatness metrics of different FL algorithms on CIFAR-10 under different Dirichlet distribution parameters with 10 participating clients per round.

Algorithm	$\alpha_{\text{Dir}} = 0.1$			$\alpha_{\text{Dir}} = 0.5$			$\alpha_{\text{Dir}} = 10$		
	MALV ↓	$\lambda_{\max}$ ↓	Δ_F ↓	MALV ↓	$\lambda_{\max}$ ↓	Δ_F ↓	MALV ↓	$\lambda_{\max}$ ↓	Δ_F ↓
FedSAM	0.512	395.554	0.618	0.450	336.908	0.578	0.143	227.832	0.404
FedASAM	0.284	101.068	0.258	0.104	49.669	0.107	0.041	58.748	0.051
MoFedSAM	0.334	293.737	0.505	0.130	256.411	0.400	0.297	871.571	0.324
FedGAMMA	0.492	1342.370	1.125	0.281	572.241	0.349	0.327	363.325	0.232
FedSMOO	0.527	103.270	0.689	**0.038**	155.386	0.163	**0.034**	82.200	0.058
FedGF	0.482	257.515	0.443	0.264	378.207	0.301	0.102	272.136	0.179
FedSWA	0.359	627.418	1.121	0.174	415.484	0.413	0.137	402.534	0.446
FedVSSAM	0.452	1168.519	0.242	0.249	1116.752	0.254	0.089	556.614	0.142
FedGloSS	0.290	340.733	0.610	0.096	441.966	0.495	0.071	296.964	0.157
FedScar	0.231	535.839	0.543	0.082	222.871	0.237	0.068	94.731	0.136
FedCADC	**0.210**	**93.773**	**0.163**	0.072	**32.895**	**0.019**	0.049	**25.721**	**0.007**

The ↓ symbol indicates that lower values are preferred. Bold denotes the lowest value in each column.

6.2.2 Convergence Behavior

The curves are locally smoothed using a window size of 50 communication rounds. Figure 1 shows the smoothed test-accuracy curves on CIFAR-10 and CIFAR-100 with $\alpha_{\text{Dir}} = 0.5$ and 10 participating clients per round. FedCADC achieves faster early-stage convergence and maintains the highest accuracy on both datasets. The advantage is more evident on CIFAR-100, where FedCADC preserves a clear margin over the baselines throughout training. This improvement can be attributed to the complementary regularizers: the curvature-aware term suppresses sharp local responses, while the direction consistency term reduces client drift relative to the global optimization trend. Their adaptive weights allow these constraints to be strengthened only when the local geometry or update direction requires correction.



(a) FedCADC, w_i (b) FedCADC, w (c) FedSAM, w_i (d) FedSAM, w

Figure 2: Loss surfaces of FedCADC and FedSAM on CIFAR-10 with 10 participating clients and $\alpha_{\text{Dir}} = 0.5$. Here, w_i denotes a local client model and w denotes the aggregated global model.

Table 3: Ablation study of the adaptive curvature-aware and direction-consistency weights.

Dataset	α_{Dir}	Fixed α_{curv}					Fixed α_{dir}					Adaptive
		0	0.01	0.03	0.05	0.08	0	0.2	0.4	0.6	0.8	FedCADC
CIFAR-10	0.1	59.00	66.29	52.13	23.47	NC	66.07	72.07	68.25	66.69	66.83	**72.28**
	0.5	78.29	83.21	81.79	72.47	NC	76.15	82.28	83.57	82.59	80.98	**84.29**
	10	82.96	85.81	87.21	80.81	NC	79.06	84.19	86.24	86.16	84.95	**87.22**
CIFAR-100	0.1	43.02	47.27	42.20	NC	NC	36.71	40.90	47.54	46.67	44.62	**50.44**
	0.5	50.60	**52.93**	50.16	26.88	NC	40.60	42.54	50.94	51.91	51.39	51.68
	10	51.60	**53.97**	52.90	39.41	NC	41.47	44.02	50.62	53.90	51.47	53.92

NC indicates that training did not converge and remained near random-guess performance. Bold values denote the highest accuracy in each row.

6.2.3 Flatness Results

Flatness metrics (MALV, $\lambda_{\max}$, and Δ_F). We quantify the flatness of the final global models using the mean absolute loss variation (MALV), the largest Hessian eigenvalue $\lambda_{\max}$, and the local–global flatness discrepancy Δ_F. Lower values indicate smaller loss variation, weaker curvature, and better local–global agreement, respectively. Detailed definitions and computational procedures are provided in Appendix C.

As reported in Table 2, FedCADC exhibits the most favorable overall flatness profile across the three Dirichlet settings, with consistently low $\lambda_{\max}$ and Δ_F values and competitive MALV. It also obtains the lowest MALV under $\alpha_{\mathrm{Dir}} = 0.1$, while FedSMOO yields slightly lower MALV for $\alpha_{\mathrm{Dir}} = 0.5$ and $\alpha_{\mathrm{Dir}} = 10$. These results are consistent with Proposition 2 and Theorem 1: curvature-aware correction controls local sharp responses, whereas direction consistency reduces client drift and consequently tightens the local–global flatness bound in Theorem 4.

Visualization of loss surfaces. In Fig. 2, we visualize the loss surfaces of the local and global models produced by FedCADC and FedSAM on CIFAR-10 with $\alpha_{\mathrm{Dir}} = 0.5$ and 10 participating clients. FedCADC exhibits a relatively broad and smooth local basin, while its aggregated global model forms an even flatter loss surface. Compared with FedSAM, FedCADC produces a more expanded and regular global basin, indicating that the curvature-aware and direction consistency regularizers reduce the mismatch between local and global optimization and guide the global model toward a flatter minimum.

6.2.4 Ablation Study

Table 3 reports the mean test accuracy over the final 50 communication rounds with 10 participating clients per round. Removing either regularizer consistently degrades performance. Compared with fixed coefficients, adaptive FedCADC achieves the best results in most settings, although several fixed values perform slightly better in individual cases. Moreover, overly large curvature coefficients may prevent convergence. These results show that the adaptive mechanisms reduce sensitivity to manual coefficient selection and maintain stable performance across different degrees of data heterogeneity.

7 Conclusion

We study flat-minima seeking in heterogeneous FL by jointly considering curvature variation and the mismatch between local and global optimization directions. Data heterogeneity can make locally flat solutions inconsistent after aggregation, limiting global generalization. To address this issue, we propose FedCADC, which combines a curvature-aware constraint with a global direction-consistency constraint and adaptively adjusts their weights according to the mini-batch geometry. Experiments on CIFAR-10 and CIFAR-100 show that FedCADC achieves higher accuracy in most settings and generally produces flatter global solutions. Ablation results further verify the contributions of both regularizers and their adaptive weighting mechanisms.

AI USE STATEMENT

In this work, we used the generative AI tool ChatGPT (GPT-5.6) to assist with English translation, language polishing, literature discovery, code debugging, and LaTeX formatting. Generative AI was used for retrieval and discovery, including identifying potentially relevant literature and related work. This use was limited to candidate discovery and did not replace manual examination of the original publications. The tool was used only as a writing, retrieval, and programming assistant. It was not used to generate or modify experimental data, determine experimental results, draw conclusions from the experiments, derive theoretical results, or formulate the final scientific claims of the paper.

All AI-assisted outputs were reviewed and revised by the authors. In particular, references identified with AI assistance were checked against the original publications, and the associated bibliographic information was manually verified. AI-assisted code suggestions were inspected, adapted to the experimental codebase, and tested by the research team. The experimental design, implementation decisions, theoretical analysis, interpretation of results, and final manuscript content were completed and validated by the authors. We take full responsibility for the final content of this work, including all text, claims, experimental results, theoretical statements, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, personal or sensitive data, animal experiments, or interactions with vulnerable populations. We use standard publicly available CIFAR-10 and CIFAR-100 benchmarks for FL research and do not release or process private client information. The proposed method is intended for improving optimization and generalization in federated image classification and does not introduce harmful applications, security attacks, or discriminatory decision-making procedures.

The authors declare no conflicts of interest or external sponsorship that could influence the design, analysis, or reporting of this study. All experiments were conducted for research purposes, and the reported methodology, results, and conclusions were prepared and reviewed by the authors in accordance with principles of research integrity.

REPRODUCIBILITY STATEMENT

We provide an anonymous supplementary package containing the complete PyTorch implementation of FedCADC. The code is organized into modules for data preprocessing and Dirichlet partitioning, ResNet-20 construction, local training, server aggregation, flatness evaluation, and loss-landscape visualization. The package also includes dependency specifications, configuration files, fixed random-seed support, execution instructions, and scripts for reproducing all CIFAR-10 and CIFAR-100 settings. Both datasets are publicly available and are downloaded automatically by the provided code. The main paper describes the algorithm and experimental protocol, while the appendix reports the implementation details, hyperparameters, metric definitions, theoretical assumptions, and complete proofs. Reported test accuracies are averaged over the last 50 communication rounds, and standard deviations are computed from three independent runs.

REFERENCES

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021.

Debora Caldarola, Barbara Caputo, and Marco Ciccone. Improving generalization in federated learning by seeking flat minima. In *European Conference on Computer Vision*, pp. 654–672. Springer, 2022.

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

Rong Dai, Xun Yang, Yan Sun, Li Shen, Xinmei Tian, Meng Wang, and Yongdong Zhang. Fedgamma: Federated learning with global sharpness-aware minimization. *IEEE Transactions on Neural Networks and Learning Systems*, 35(12):17479–17492, 2023.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. Feddc: Federated learning with non-iid data via local drift decoupling and correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10112–10121, 2022.

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

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

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

Ming Hu, Yue Cao, Anran Li, Zhiming Li, Chengwei Liu, Tianlin Li, Mingsong Chen, and Yang Liu. Fedmut: Generalized federated learning via stochastic mutation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 12528–12537, 2024.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Peter Kairouz and H Brendan McMahan. Advances and open problems in federated learning. *Foundations and trends in machine learning*, 14(1-2):1–210, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143. PMLR, 2020.

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

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International conference on machine learning*, pp. 5905–5914. PMLR, 2021.

Gihun Lee, Minchan Jeong, Yongjin Shin, Sangmin Bae, and Se-Young Yun. Preservation of the global knowledge by not-true distillation in federated learning. *Advances in Neural Information Processing Systems*, 35:38461–38474, 2022.

Taehwan Lee and Sung Whan Yoon. Rethinking the flat minima searching in federated learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 27109–27129. PMLR, 2024.

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

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10713–10722, 2021.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

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

Junkang Liu, Yuanyuan Liu, Fanhua Shang, Hongying Liu, Jin Liu, and Wei Feng. Improving generalization in federated learning with highly heterogeneous data via momentum-based stochastic controlled weight averaging. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 38894–38939. PMLR, 2025.

Jianfeng Lu, YuZhao Xiang, Yue Chen, Gang Li, Shuqin Cao, and Guanghui Wen. Fedscar: Correcting geometric bias for flatness-consistent federated learning. In *Forty-third International Conference on Machine Learning*, 2026.

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

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

Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

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

Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 8432–8440, 2022.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

Bingnan Xiao, Yuan Gao, Bingcong Li, Wei Ni, Xin Wang, and Tony QS Quek. Fedvssam: Mitigating flatness incompatibility in sharpness-aware federated learning. *arXiv preprint arXiv:2605.09144*, 2026.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

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.

A Detailed Theoretical Derivations

This appendix first presents the complete theoretical setup and formal results underlying the conclusions in the main text, and then provides step-by-step proofs.

A.1 Theoretical Setup

Consider an FL system with m clients. Client i holds a local dataset $S_i = \{z_{i,j}\}_{j=1}^{n_i}$. Its empirical and population risks are, respectively,

$$F_{S_i}(w) = \frac{1}{n_i}\sum_{j=1}^{n_i}\ell(w; z_{i,j}), \qquad \mathcal{R}_i(w) = \mathbb{E}_{z\sim\mathcal{D}_i}\ell(w; z). \tag{23}$$

Let $p_i > 0$ satisfy $\sum_i p_i = 1$, and denote $N_{\mathrm{tr}} = \sum_i n_i$. The corresponding global empirical and population risks are

$$F_S(w) = \sum_{i=1}^{m} p_i F_{S_i}(w), \qquad \mathcal{R}(w) = \sum_{i=1}^{m} p_i \mathcal{R}_i(w). \tag{24}$$

To avoid cumbersome subscripts, we write $F_i := F_{S_i}$ and $F := F_S$ in the optimization and flatness analyses. The population risk $\mathcal{R}$ and empirical risk F_S are distinguished explicitly only in the generalization analysis. In round t, each client performs local updates from $w_i^{t,0} = w^t$. Let $K = \max_{t,i} K_i^t$. For a client that has completed its local training, we assign zero updates to the remaining indices, which allows different K_i^t to be represented by a common local depth K. To accommodate the curvature regularizer, direction regularizer, and practical optimizer states within a unified analysis, define the normalized effective update at local step k as

$$\begin{aligned} v_i^{t,k} &:= \frac{w_i^{t,k} - w_i^{t,k+1}}{\eta} \\ &= g_i(w_i^{t,k}; \xi_i^{t,k}) + \alpha_{\mathrm{curv},i}^{t,k} r_{\mathrm{curv},i}^{t,k} + \alpha_{\mathrm{dir},i}^{t,k} r_{\mathrm{dir},i}^{t,k} + e_{\mathrm{opt},i}^{t,k}. \end{aligned} \tag{25}$$

where $e_{\mathrm{opt},i}^{t,k}$ denotes the normalized residual induced by optimizer mechanisms such as SGD momentum, weight decay, and gradient clipping. Define

$$h_i^{t,k} := \mathbb{E}[v_i^{t,k} \mid \mathcal{F}_{t,k}] = \nabla F_i(w_i^{t,k}) + u_i^{t,k}, \tag{26}$$

where $u_i^{t,k}$ denotes the combined bias in the conditional mean update induced by the geometric regularizers and optimizer residual.

Assumption 1 (Trajectory-local smoothness and lower boundedness)**.** Each F_i has an L-Lipschitz gradient along the finite training trajectory and within its radius-ρ perturbation neighborhood. Moreover, the global objective has a finite lower bound $F_\star > -\infty$.

Assumption 2 (Stochastic variation and data heterogeneity)**.** Conditioned on the history $\mathcal{F}_{t,k}$, the centered effective updates satisfy

$$\sum_i p_i \mathbb{E}[\|v_i^{t,k} - h_i^{t,k}\|^2 \mid \mathcal{F}_{t,k}] \le \sigma^2, \tag{27}$$

and, at every trajectory point considered in the analysis,

$$\sum_i p_i \|\nabla F_i(w) - \nabla F(w)\|^2 \le \zeta^2. \tag{28}$$

Assumption 3 (Finite trajectory moments and sampling error)**.** Over the finite training horizon, the conditional mean bias, inter-client update deviation, and local model drift have finite second moments. The mean-squared error between partial-client aggregation and the full-client conditional mean update is bounded by a finite quantity ω_t^2.

All these conditions are restricted to the realized training trajectory and its perturbation neighborhood; they do not require convexity, strong convexity, or globally bounded gradients.

A.2 Geometric Properties of FedCADC

For any nonzero local empirical gradient $g_i(w) = \nabla F_i(w)$, define

$$\epsilon_i(w) = \rho \frac{g_i(w)}{\|g_i(w)\| + \delta}, \qquad R_{\mathrm{curv},i}(w) = \frac{[F_i(w + \epsilon_i(w)) - F_i(w)]_+}{\rho^2}. \tag{29}$$

When computing the gradient of the curvature regularizer, the perturbation $\epsilon_i(w)$ constructed at the current step is treated as a constant through the stop-gradient operation:

$$r_{\mathrm{curv},i}(w) = \nabla_x \frac{[F_i(x + \bar{\epsilon}) - F_i(x)]_+}{\rho^2}\bigg|_{x=w,\ \bar{\epsilon}=\epsilon_i(w)}. \tag{30}$$

Proposition 1 (Local expansion of the curvature-aware regularizer)**.** If F_i is twice continuously differentiable on the line segment connecting w and $w + \epsilon_i(w)$, then there exists $\theta \in (0, 1)$ such that

$$\begin{aligned} R_{\mathrm{curv},i}(w) = \Bigg[&\frac{\|g_i(w)\|^2}{\rho(\|g_i(w)\| + \delta)} + \frac{1}{2}\left(\frac{\|g_i(w)\|}{\|g_i(w)\| + \delta}\right)^2 \\ &\cdot \widehat{g}_i(w)^\top \nabla^2 F_i(w + \theta\epsilon_i)\widehat{g}_i(w) \Bigg]_+, \end{aligned} \tag{31}$$

where $\widehat{g}_i(w) = g_i(w)/\|g_i(w)\|$. When $g_i(w) = 0$, $R_{\mathrm{curv},i}(w) = 0$.

Furthermore, define

$$g_{p,i}(w) = \nabla F_i(w + \epsilon_i(w)), \qquad c_i(w) = \frac{\langle g_{p,i}(w), g_i(w)\rangle}{(\|g_{p,i}(w)\| + \delta)(\|g_i(w)\| + \delta)},$$
$$q_i(w) = \frac{\|g_{p,i}(w)\| + \delta}{\|g_i(w)\| + \delta}, \qquad \alpha_{\mathrm{curv},i}(w) = \mathrm{clip}\left([1 - c_i(w)]|\log q_i(w)|,\ 0,\ A_{\mathrm{curv}}\right). \tag{32}$$

Proposition 2 (Exact form of the curvature correction)**.** Under the stop-gradient implementation for both the perturbation and adaptive coefficient, let $\Delta_i(w) = F_i(w + \epsilon_i(w)) - F_i(w)$. If $\Delta_i(w) > 0$, then

$$r_{\mathrm{curv},i}(w) = \frac{g_{p,i}(w) - g_i(w)}{\rho^2}; \tag{33}$$

The correction is zero when $\Delta_i(w) < 0$. At $\Delta_i(w) = 0$, the right-hand side of equation 33 is multiplied by some $\vartheta \in [0, 1]$. If δ, which is introduced solely for numerical stability, is ignored, then

$$\|g_{p,i} - g_i\|^2 = \|g_i\|^2 \left[(q_i - 1)^2 + 2q_i(1 - c_i)\right]. \tag{34}$$

Proposition 1 shows that $R_{\mathrm{curv},i}$ contains both the loss increment along the current gradient direction and the directional Hessian curvature. In the effective numerical regime $\|g_i(w)\| \gg \delta$, these two terms are approximated by $\|g_i(w)\|/\rho$ and $\widehat{g}_i^\top \nabla^2 F_i \widehat{g}_i/2$, respectively. Both terms decay smoothly when the gradient norm approaches the numerical stabilization scale. Proposition 2 further shows that the actual curvature correction is directly determined by the gradient difference before and after perturbation, while $\alpha_{\mathrm{curv},i}$ responds jointly to directional rotation and relative scale variation. The mechanism therefore characterizes local geometric sensitivity along the current update direction, rather than an unrestricted maximum Hessian eigenvalue.

Let $d = w^t - w \neq 0$, let $\widehat{g}^t$ denote the normalized historical global direction, and define $R_{\mathrm{dir}}(w) = 1 - \cos(d, \widehat{g}^t)$.

Proposition 3 (One-step alignment induced by the direction regularizer)**.** Fix w^t and a nonzero direction $\widehat{g}^t$. Suppose that an update is performed using only the direction regularizer:

$$w^+ = w - \tau \nabla_w [1 - \cos(w^t - w, \widehat{g}^t)] \tag{35}$$

If $0 \leq \tau/\|d\|^2 \leq 2$, then

$$\cos(w^t - w^+, \widehat{g}^t) \geq \cos(w^t - w, \widehat{g}^t). \tag{36}$$

Proposition 3 characterizes the isolated geometric effect of the direction regularizer. For a joint update containing the empirical-risk and curvature terms, its net effect must also account for feedback through the parameter trajectory. We therefore pair a process without direction regularization and a process with direction regularization under the same initialization and randomness. Let $b_i = h_{i,\mathrm{base}} - \sum_j p_j h_{j,\mathrm{base}}$ denote the centered deviation of the baseline conditional mean update, let $q_i = h_{i,\mathrm{dir}} - h_{i,\mathrm{base}}$ denote the effective-update difference induced by direction regularization, and define $\widetilde{q}_i = q_i - \sum_j p_j q_j$. Then

$$D^2_{\mathrm{base}} = \sum_i p_i \|b_i\|^2, \qquad D^2_{\mathrm{dir}} = \sum_i p_i \|b_i + \widetilde{q}_i\|^2. \tag{37}$$

Theorem 1 (Sufficient condition for directional-deviation contraction)**.** Suppose there exist $\kappa \in [0, 1]$ and $r \geq 0$ such that

$$2\sum_i p_i \langle b_i, \widetilde{q}_i\rangle + \sum_i p_i \|\widetilde{q}_i\|^2 \leq -(1 - \kappa^2) D^2_{\mathrm{base}} + r, \tag{38}$$

Then

$$D^2_{\mathrm{dir}} \leq \kappa^2 D^2_{\mathrm{base}} + r. \tag{39}$$

If $\kappa < 1$ and $r < (1 - \kappa^2) D^2_{\mathrm{base}}$, the direction-regularized process strictly contracts the mean-squared client-update deviation.

This result does not presume that direction regularization necessarily improves the joint update. Instead, it provides a verifiable sufficient condition: the negative cross effect between the effective-update difference and the original client deviation must dominate the additional second-order perturbation.

A.3 Optimization and Generalization Guarantees

Define

$$C_t = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\sum_i p_i\|w_i^{t,k}-w^t\|^2,$$
$$U_t^2 = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_i p_i u_i^{t,k}\right\|^2, \tag{40}$$
$$D_t^2 = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\sum_i p_i\left\|h_i^{t,k}-\sum_j p_j h_j^{t,k}\right\|^2,$$

and let $\overline{U}^2$, $\overline{D}^2$, and $\overline{\omega}^2$ denote their respective averages over T rounds. Define $\widehat{h}_t = (w^t - w^{t+1})/(\eta K)$. Then $\omega_t^2 = \mathbb{E}\|\widehat{h}_t - K^{-1}\sum_k\sum_i p_i h_i^{t,k}\|^2$.

Theorem 2 (Non-convex optimization error of FedCADC)**.** Under Assumptions 1–3, suppose that a fixed local step size η is used throughout training and that $a = \eta^2K^2L^2 \le 1/48$. Then

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(w^t)\|^2 \le \frac{4[F(w^0)-F_\star]}{\eta KT} + 4\overline{\omega}^2 + (8+48a)\overline{U}^2 + 16a(\overline{D}^2+\sigma^2). \tag{41}$$

Corollary 1 (Order-wise optimization error)**.** Let $\mathcal{F}_{\mathrm{C}} = F(w^0) - F_\star$. Theorem 2 implies

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(w^t)\|^2 = \mathcal{O}\left[\frac{\mathcal{F}_{\mathrm{C}}}{\eta KT} + \overline{\omega}^2 + \overline{U}^2 + \eta^2K^2L^2(\overline{D}^2+\sigma^2)\right]. \tag{42}$$

For a prescribed training budget T, suppose that the step size is selected before training as

$$\eta = \frac{c_\eta}{L\sqrt{TK}}, \qquad T \ge 48c_\eta^2 K, \tag{43}$$

and kept fixed throughout all communication rounds. Then

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(w^t)\|^2 = \mathcal{O}\left[\frac{L\mathcal{F}_{\mathrm{C}}}{\sqrt{TK}} + \overline{\omega}^2 + \overline{U}^2 + \frac{K}{T}(\overline{D}^2+\sigma^2)\right]. \tag{44}$$

equation 42 decomposes the error into the initial optimality gap, aggregation sampling error, conditional mean bias from the regularizers and optimizer, client-update deviation, and stochastic variation. equation 43 specifies a fixed step size calibrated to the prescribed training budget.

Assumption 4 (Shared sampling envelope for paired processes)**.** For the processes without and with direction regularization that share the same initialization, client sampling, and mini-batch randomness, suppose there exist common finite quantities $\omega_{\mathrm{stoch},t}^2$, $b_{s,t}^2$, and $\chi_s \ge 0$ such that

$$\omega_{t,b}^2 \le \omega_{\mathrm{stoch},t}^2 + b_{s,t}^2 + \chi_s D_{t,b}^2, \qquad b \in \{\mathrm{base}, \mathrm{dir}\}. \tag{45}$$

Corollary 2 (Condition for tightening the optimization bound)**.** Suppose Theorem 1 yields, after averaging over rounds, $\overline{D}_{\mathrm{dir}}^2 \le \kappa^2\overline{D}_{\mathrm{base}}^2 + \overline{r}$. If

$$(4\chi_s + 16a)[(1-\kappa^2)\overline{D}_{\mathrm{base}}^2 - \overline{r}] > (8+48a)(\overline{U}_{\mathrm{dir}}^2 - \overline{U}_{\mathrm{base}}^2) \tag{46}$$

holds, then the optimization-error upper bound for the direction-regularized process is strictly smaller than that of the paired baseline process.

Corollary 2 shows that the provable optimization benefit of FedCADC arises from the improvement in the aggregation and drift terms induced by deviation contraction. This benefit must exceed the conditional mean bias that may be introduced by direction regularization. The comparison is therefore an upper-bound comparison under a sufficient condition.

We next analyze the generalization error using expected replace-one stability. We unfold T rounds, each containing K local depths, into $Q = TK$ synchronous joint-state steps. The global and all local parameters, historical direction, optimizer states, and inference statistics are combined into an augmented state W_q. Here, $q = tK + k$ corresponds to local depth k in round t, and H_q simultaneously applies the parallel updates of all active clients at that depth. When $k = K - 1$, the state transition additionally includes server aggregation, initialization for the next round, and the historical-direction update. Thus, Q counts joint-state transitions rather than the total number of stochastic-gradient evaluations across all clients. The update is written uniformly as

$$W_{q+1} = W_q - \eta_q H_q(W_q; Z_q), \qquad q = 0, \dots, Q-1. \tag{47}$$

Assumption 5 (Expected replace-one stability along coupled trajectories)**.** Sample a training-example index according to $\Pr(J = (i,j)) = p_i/n_i$ and replace it with an independent example drawn from the same client distribution, yielding a neighboring dataset S'. The two training runs share all randomness except for the replaced example. The per-example loss is L_ℓ-Lipschitz with respect to the final global inference state, and the state map satisfies

$$\|H_q(W; Z) - H_q(W'; Z)\| \le L_{\Phi,q}\|W - W'\|. \tag{48}$$

Let $\mathcal{A}_q$ denote the event that the replaced example is accessed at step q, and let $\pi_q = \Pr(\mathcal{A}_q)$. Assume that

$$\mathbb{E}\left[\frac{\|H_q\| + \|H_q'\|}{2}\,\middle|\,\mathcal{A}_q\right] \le M_q. \tag{49}$$

Theorem 3 (Expected generalization error of FedCADC)**.** Under Assumption 5,

$$\epsilon_{\text{gen}} := |\mathbb{E}[\mathcal{R}(W_Q) - F_S(W_Q)]| \le 2L_\ell \sum_{q=0}^{Q-1} \eta_q \pi_q M_q \prod_{r=q+1}^{Q-1} (1 + \eta_r L_{\Phi,r}). \tag{50}$$

If $p_i = n_i/N_{\text{tr}}$, the sampling of example indices is independent of their values, and the set of distinct training-example indices accessed at each synchronous joint-state step, $\mathcal{U}_q$, satisfies $|\mathcal{U}_q| \le B_{\text{eff}}$ almost surely, then $\pi_q \le B_{\text{eff}}/N_{\text{tr}}$. Furthermore, if $\eta_q = \eta$, $M_q \le M$, and $L_{\Phi,q} \le L_\Phi$, then

$$\epsilon_{\text{gen}} \le \frac{2L_\ell M \eta T K B_{\text{eff}}}{N_{\text{tr}}} \exp(L_\Phi \eta T K). \tag{51}$$

Corollary 3 (Order-wise generalization error)**.** Under the fixed-step-size specialization of Theorem 3,

$$\epsilon_{\text{gen}} = \mathcal{O}\left[\frac{L_\ell M \eta T K B_{\text{eff}}}{N_{\text{tr}}} \exp(L_\Phi \eta T K)\right]. \tag{52}$$

When the cumulative propagation strength satisfies $L_\Phi \eta T K = \mathcal{O}(1)$,

$$\epsilon_{\text{gen}} = \mathcal{O}\left[\frac{L_\ell M \eta T K B_{\text{eff}}}{N_{\text{tr}}}\right]. \tag{53}$$

The generalization bound is jointly determined by the probability of accessing the replaced example, the update magnitude, and the propagation of state perturbations. For a baseline process under the same coupling, if $M_{\text{dir},q} \le M_{\text{base},q}$ and $L_{\Phi,\text{dir},q} \le L_{\Phi,\text{base},q}$ hold at every step, then the bound in equation 50 for FedCADC is no larger than that for the baseline process. The bound is strictly tighter if at least one accessed step is strictly improved. Appendix A.8 further provides a sufficient condition under which contraction of the client-update deviation reduces the update-magnitude envelope.

A.4 Flatness Discrepancy

Flatness and generalization are not equivalent concepts. We therefore analyze flatness discrepancy as a separate geometric quantity. Let b denote non-trainable states such as BatchNorm inference statistics, which are held fixed under parameter perturbations. Define

$$\begin{aligned} \mathcal{S}_i(w, b; \rho) &= \max_{\|\delta\| \le \rho} [F_i(w + \delta; b) - F_i(w; b)], \\ \mathcal{S}_F(w, b; \rho) &= \max_{\|\delta\| \le \rho} [F(w + \delta; b) - F(w; b)], \end{aligned} \tag{54}$$

and the flatness discrepancy between the global and local models as

$$\Delta_F = \left| \mathcal{S}_F(w, b; \rho) - \sum_i p_i \mathcal{S}_i(w_i, b_i; \rho) \right|. \tag{55}$$

Assumption 6 (Trajectory-local sensitivity of inference statistics)**.** Along the finite training trajectory, there exists $L_{b,\rho} \geq 0$ such that

$$|\mathcal{S}_i(w, b; \rho) - \mathcal{S}_i(w, b'; \rho)| \leq L_{b,\rho}\|b - b'\|. \tag{56}$$

This term is zero if all sharpness quantities are evaluated using the same fixed inference statistics.

Theorem 4 (Flatness-discrepancy upper bound for FedCADC)**.** Under Assumptions 1, 2, and 6,

$$\Delta_F \leq \rho\zeta + L\rho^2 + L\rho \sum_i p_i\|w - w_i\| + L_{b,\rho} \sum_i p_i\|b - b_i\|. \tag{57}$$

Furthermore, define the coverage error

$$e_{\mathrm{cov},i}(w, b; \rho) = \mathcal{S}_i(w, b; \rho) - \rho^2 R_{\mathrm{curv},i}(w; b) \geq 0, \tag{58}$$

Then

$$\begin{aligned}\mathcal{S}_F(w, b; \rho) \leq \rho^2 \sum_i p_i R_{\mathrm{curv},i}(w_i; b_i) + \sum_i p_i e_{\mathrm{cov},i}(w_i, b_i; \rho) \\ + L\rho \sum_i p_i\|w - w_i\| + L_{b,\rho} \sum_i p_i\|b - b_i\|.\end{aligned} \tag{59}$$

At the communication endpoint of round t, define

$$\mathcal{M}_t = \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \sum_i p_i\|v_i^{t,k}\|^2, \qquad \nu_t^2 = \mathbb{E} \left\| w^{t+1} - \sum_i p_i w_i^{t,K} \right\|^2, \tag{60}$$

and let $\beta_t = \sum_i p_i \mathbb{E}\|b_g^{t+1} - b_i^{t,K}\|$. Theorem 4 further gives

$$\mathbb{E}\Delta_F \leq \rho\zeta + L\rho^2 + L\rho\sqrt{\nu_t^2} + 2L\rho\eta K\sqrt{\mathcal{M}_t} + L_{b,\rho}\beta_t, \tag{61}$$

where

$$\mathcal{M}_t \leq 3\mathbb{E}\|\nabla F(w^t)\|^2 + 3L^2 C_t + 3U_t^2 + D_t^2 + \sigma^2. \tag{62}$$

equation 59 shows that the curvature regularizer directly controls a sharpness component along a feasible direction, while equation 61 and equation 62 further relate the global flatness discrepancy to heterogeneity, client-update deviation, partial-participation error, and differences in inference statistics. If the sum of the directional curvature quantity and coverage error decreases, and the contraction of directional deviation is not offset by other error terms, the flatness bound for FedCADC becomes tighter than that of the paired baseline process.

A.5 Notation and Basic Inequalities

Define

$$\overline{h}^{t,k} = \sum_{i=1}^{m} p_i h_i^{t,k}, \qquad \overline{h}_t = \frac{1}{K} \sum_{k=0}^{K-1} \overline{h}^{t,k}. \tag{63}$$

The effective server update can be written as

$$w^{t+1} = w^t - \gamma\widehat{h}_t, \qquad \gamma = \eta K, \qquad \widehat{h}_t = \frac{w^t - w^{t+1}}{\eta K}. \tag{64}$$

The aggregation error used in the main text therefore satisfies

$$\omega_t^2 = \mathbb{E}\|\widehat{h}_t - \overline{h}_t\|^2. \tag{65}$$

We also define the round-averaged quantities

$$\overline{U}^2 = \frac{1}{T} \sum_{t=0}^{T-1} U_t^2, \qquad \overline{D}^2 = \frac{1}{T} \sum_{t=0}^{T-1} D_t^2, \qquad \overline{\omega}^2 = \frac{1}{T} \sum_{t=0}^{T-1} \omega_t^2. \tag{66}$$

The following basic inequalities will be used repeatedly. For arbitrary vectors $x_1, \dots, x_s$,

$$\left\| \sum_{j=1}^{s} x_j \right\|^2 \le s \sum_{j=1}^{s} \|x_j\|^2, \tag{67}$$

which follows directly from the Cauchy–Schwarz inequality. If $\{p_i\}$ are probability weights, Jensen's inequality gives

$$\left\| \sum_i p_i x_i \right\|^2 \le \sum_i p_i \|x_i\|^2. \tag{68}$$

For arbitrary x, y, we also have

$$\|x+y\|^2 \le 2\|x\|^2 + 2\|y\|^2, \qquad -\langle x, y\rangle = -\frac{1}{2}\|x\|^2 - \frac{1}{2}\|y\|^2 + \frac{1}{2}\|x-y\|^2. \tag{69}$$

A.6 Proofs of the Geometric Properties

Proof of Proposition 1. For notational simplicity, let $g_i = g_i(w)$ and $\epsilon_i = \epsilon_i(w)$. By the second-order Taylor expansion with remainder in mean-value form, there exists $\theta \in (0,1)$ such that

$$\begin{aligned} F_i(w+\epsilon_i) - F_i(w) &= \nabla F_i(w)^\top \epsilon_i \\ &\quad + \frac{1}{2}\epsilon_i^\top \nabla^2 F_i(w+\theta\epsilon_i)\epsilon_i. \end{aligned} \tag{70}$$

When $g_i \neq 0$, substituting $\epsilon_i = \rho g_i/(\|g_i\| + \delta)$ gives the first-order term step by step as

$$\begin{aligned} \nabla F_i(w)^\top \epsilon_i &= g_i^\top \left(\rho \frac{g_i}{\|g_i\| + \delta} \right) \\ &= \rho \frac{g_i^\top g_i}{\|g_i\| + \delta} = \rho \frac{\|g_i\|^2}{\|g_i\| + \delta}. \end{aligned} \tag{71}$$

Since $g_i = \|g_i\| \widehat{g}_i$, the second-order term satisfies

$$\begin{aligned} &\epsilon_i^\top \nabla^2 F_i(w+\theta\epsilon_i)\epsilon_i \\ &= \left(\rho \frac{\|g_i\|}{\|g_i\| + \delta} \widehat{g}_i \right)^\top \nabla^2 F_i(w+\theta\epsilon_i) \left(\rho \frac{\|g_i\|}{\|g_i\| + \delta} \widehat{g}_i \right) \\ &= \rho^2 \left(\frac{\|g_i\|}{\|g_i\| + \delta} \right)^2 \widehat{g}_i^\top \nabla^2 F_i(w+\theta\epsilon_i) \widehat{g}_i. \end{aligned} \tag{72}$$

Substituting equation 71 and equation 72 into equation 70, dividing by ρ^2, and taking the positive part yields equation 31. When $g_i = 0$, $\epsilon_i = 0$, so both the loss increment and $R_{\mathrm{curv},i}$ are zero. □

Proof of Proposition 2. Fix the current point w, and abbreviate $\bar{\epsilon} = \epsilon_i(w)$ and $\Delta_i = F_i(w+\bar{\epsilon}) - F_i(w)$. Because both the perturbation and adaptive coefficient are treated as constants in the current backward pass, the curvature regularizer as a function of the parameter variable x is

$$\varphi_i(x) = \frac{[F_i(x+\bar{\epsilon}) - F_i(x)]_+}{\rho^2}. \tag{73}$$

First consider $\Delta_i > 0$. The derivative of the positive-part function at this point is 1, so the chain rule gives

$$\begin{aligned} r_{\mathrm{curv},i}(w) &= \nabla_x \varphi_i(x)\big|_{x=w} \\ &= \frac{1}{\rho^2} \nabla_x [F_i(x+\bar{\epsilon}) - F_i(x)]\big|_{x=w} \\ &= \frac{1}{\rho^2} [\nabla F_i(w+\bar{\epsilon}) - \nabla F_i(w)] \\ &= \frac{g_{p,i}(w) - g_i(w)}{\rho^2}. \end{aligned} \tag{74}$$

When $\Delta_i < 0$, the derivative of the positive-part function at this point is 0, and hence $r_{\mathrm{curv},i}(w) = 0$. When $\Delta_i = 0$, the subdifferential of $[x]_+$ is the interval $[0, 1]$. Thus, there exists $\vartheta \in [0, 1]$ such that

$$r_{\mathrm{curv},i}(w) = \frac{\vartheta}{\rho^2}[g_{p,i}(w) - g_i(w)]. \tag{75}$$

We next prove the gradient-geometry identity. To isolate directional and scale changes, temporarily set the numerical stabilization term $\delta = 0$ and define

$$a = \|g_i\|, \qquad b = \|g_{p,i}\|, \qquad q_i = \frac{b}{a}, \qquad c_i = \frac{\langle g_{p,i}, g_i\rangle}{ab}. \tag{76}$$

When $a, b > 0$, the law of cosines gives

$$\begin{aligned} \|g_{p,i} - g_i\|^2 &= \|g_{p,i}\|^2 + \|g_i\|^2 - 2\langle g_{p,i}, g_i\rangle \\ &= b^2 + a^2 - 2abc_i \\ &= a^2(q_i^2 + 1 - 2q_ic_i). \end{aligned} \tag{77}$$

Rearranging the terms inside the parentheses yields

$$\begin{aligned} q_i^2 + 1 - 2q_ic_i &= q_i^2 - 2q_i + 1 + 2q_i - 2q_ic_i \\ &= (q_i - 1)^2 + 2q_i(1 - c_i). \end{aligned} \tag{78}$$

Substituting equation 78 into equation 77 gives equation 34. If either gradient is zero, the result can be interpreted through the continuous stabilized version with $\delta > 0$. The term δ in the algorithm prevents the corresponding ratio and cosine from becoming undefined at degenerate points. □

Proof of Proposition 3. Let $r = \|d\|$, $\widehat{d} = d/r$, and $c = \langle \widehat{d}, \widehat{g}^t\rangle \in [-1, 1]$. We first differentiate with respect to $d = w^t - w$. From

$$\nabla_d \frac{\langle d, \widehat{g}^t\rangle}{\|d\|} = \frac{\widehat{g}^t}{\|d\|} - \frac{\langle d, \widehat{g}^t\rangle d}{\|d\|^3} = \frac{\widehat{g}^t - c\widehat{d}}{r} \tag{79}$$

and $\nabla_w d = -I$, we obtain

$$\nabla_w[1 - \cos(w^t - w, \widehat{g}^t)] = \frac{\widehat{g}^t - c\widehat{d}}{r}. \tag{80}$$

Let $\lambda = \tau/r^2$. By the update in equation 35,

$$\begin{aligned} d^+ &= w^t - w^+ \\ &= w^t - w + \tau\frac{\widehat{g}^t - c\widehat{d}}{r} \\ &= r[(1 - \lambda c)\widehat{d} + \lambda\widehat{g}^t]. \end{aligned} \tag{81}$$

Let $s = 1 - c^2 \geq 0$. The numerator associated with equation 81 is

$$\langle d^+, \widehat{g}^t\rangle = r[(1 - \lambda c)c + \lambda] = r(c + \lambda s), \tag{82}$$

while its norm satisfies

$$\begin{aligned} \|d^+\|^2 &= r^2\|(1 - \lambda c)\widehat{d} + \lambda\widehat{g}^t\|^2 \\ &= r^2[(1 - \lambda c)^2 + \lambda^2 + 2\lambda c(1 - \lambda c)] \\ &= r^2[1 + \lambda^2(1 - c^2)] = r^2(1 + \lambda^2 s). \end{aligned} \tag{83}$$

Therefore, defining $D_\lambda = \sqrt{1 + \lambda^2 s}$ gives

$$\cos(d^+, \widehat{g}^t) = \frac{c + \lambda s}{D_\lambda}. \tag{84}$$

Subtracting c and using $D_\lambda - 1 = \lambda^2 s/(1 + D_\lambda)$ yields

$$\begin{aligned}\cos(d^+, \widehat{g}^t) - c &= \frac{c + \lambda s - cD_\lambda}{D_\lambda} \\ &= \frac{\lambda s - c(D_\lambda - 1)}{D_\lambda} \\ &= \frac{\lambda s}{D_\lambda}\left(1 - \frac{c\lambda}{1 + D_\lambda}\right).\end{aligned} \tag{85}$$

The first factor is nonnegative. If $c \leq 0$ or $c\lambda \leq 1$, the term in parentheses is nonnegative. It remains to consider $c > 0$ and $c\lambda > 1$. In this case, $c\lambda \leq 1 + D_\lambda$ is equivalent to

$$(c\lambda - 1)^2 \leq D_\lambda^2. \tag{86}$$

Expanding both sides and canceling common terms reduces this condition to

$$\lambda(2c^2 - 1) \leq 2c. \tag{87}$$

When $2c^2 - 1 \leq 0$, this inequality holds automatically. When $2c^2 - 1 > 0$, $0 \leq \lambda \leq 2$ gives

$$\lambda(2c^2 - 1) \leq 2(2c^2 - 1) \leq 2c, \tag{88}$$

where the last inequality is equivalent to $(2c+1)(c-1) \leq 0$, which holds for $c \in (0, 1]$. Therefore, equation 85 is nonnegative, proving the proposition. □

Proof of Theorem 1. By the centering definitions, $\sum_i p_i b_i = 0$ and $\sum_i p_i \widetilde{q}_i = 0$. Directly expanding the squared norm gives

$$\begin{aligned}D_{\text{dir}}^2 &= \sum_i p_i \|b_i + \widetilde{q}_i\|^2 \\ &= \sum_i p_i \left(\|b_i\|^2 + 2\langle b_i, \widetilde{q}_i\rangle + \|\widetilde{q}_i\|^2\right) \\ &= D_{\text{base}}^2 + 2\sum_i p_i \langle b_i, \widetilde{q}_i\rangle + \sum_i p_i \|\widetilde{q}_i\|^2.\end{aligned} \tag{89}$$

Substituting condition equation 38 into the expression above yields

$$D_{\text{dir}}^2 \leq D_{\text{base}}^2 - (1 - \kappa^2) D_{\text{base}}^2 + r = \kappa^2 D_{\text{base}}^2 + r. \tag{90}$$

If $r < (1 - \kappa^2) D_{\text{base}}^2$, the right-hand side is strictly smaller than D_{base}^2, which establishes strict contraction. □

A.7 Detailed Proof of the Optimization Error

Lemma 1 (Local drift and mean-update deviation)**.** Let $a = \eta^2 K^2 L^2$. Under Assumptions 1– 3, if $a \leq 1/48$, then

$$C_t \leq 2\eta^2 K^2 [3\mathbb{E}\|\nabla F(w^t)\|^2 + 3U_t^2 + D_t^2 + \sigma^2], \tag{91}$$

and

$$\begin{aligned}B_t &:= \mathbb{E}\|\overline{h}_t - \nabla F(w^t)\|^2 \\ &\leq 12a\, \mathbb{E}\|\nabla F(w^t)\|^2 + (2 + 12a)U_t^2 + 4a(D_t^2 + \sigma^2).\end{aligned} \tag{92}$$

Proof. From the local recursion $w_i^{t,k+1} = w_i^{t,k} - \eta v_i^{t,k}$ and $w_i^{t,0} = w^t$, for any $k \geq 1$ we have

$$w_i^{t,k} - w^t = -\eta \sum_{j=0}^{k-1} v_i^{t,j}. \tag{93}$$

Applying equation 67 to equation 93 gives

$$\|w_i^{t,k} - w^t\|^2 \leq \eta^2 k \sum_{j=0}^{k-1} \|v_i^{t,j}\|^2 \leq \eta^2 K \sum_{j=0}^{K-1} \|v_i^{t,j}\|^2. \tag{94}$$

Weighting over clients, taking expectations, and averaging over k yields

$$\begin{aligned} C_t &= \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\sum_i p_i\|w_i^{t,k}-w^t\|^2 \\ &\le \eta^2 K\sum_{j=0}^{K-1}\mathbb{E}\sum_i p_i\|v_i^{t,j}\|^2 \\ &= \eta^2K^2\frac{1}{K}\sum_{j=0}^{K-1}\mathbb{E}\sum_i p_i\|v_i^{t,j}\|^2. \end{aligned} \tag{95}$$

We next decompose the second moment of the effective update. Conditioned on $\mathcal{F}_{t,j}$, since $\mathbb{E}[v_i^{t,j}-h_i^{t,j}\mid\mathcal{F}_{t,j}]=0$,

$$\mathbb{E}[\|v_i^{t,j}\|^2\mid\mathcal{F}_{t,j}]=\|h_i^{t,j}\|^2+\mathbb{E}[\|v_i^{t,j}-h_i^{t,j}\|^2\mid\mathcal{F}_{t,j}], \tag{96}$$

because the conditional expectation of the cross term is zero. Using the weighted variance identity

$$\sum_i p_i\|h_i^{t,j}\|^2=\|\overline{h}^{t,j}\|^2+\sum_i p_i\|h_i^{t,j}-\overline{h}^{t,j}\|^2, \tag{97}$$

and applying Assumption 2, we obtain

$$\begin{aligned} \mathbb{E}\sum_i p_i\|v_i^{t,j}\|^2 \le{}& \mathbb{E}\|\overline{h}^{t,j}\|^2 \\ &+\mathbb{E}\sum_i p_i\|h_i^{t,j}-\overline{h}^{t,j}\|^2+\sigma^2. \end{aligned} \tag{98}$$

Since $h_i^{t,j}=\nabla F_i(w_i^{t,j})+u_i^{t,j}$,

$$\begin{aligned} \overline{h}^{t,j} ={}& \nabla F(w^t)+\sum_i p_i[\nabla F_i(w_i^{t,j})-\nabla F_i(w^t)] \\ &+\sum_i p_i u_i^{t,j}. \end{aligned} \tag{99}$$

Applying $\|x+y+z\|^2\le 3(\|x\|^2+\|y\|^2+\|z\|^2)$ to the three terms, followed by Jensen's inequality and L-smoothness for the second term, gives

$$\begin{aligned} &\left\|\sum_i p_i[\nabla F_i(w_i^{t,j})-\nabla F_i(w^t)]\right\|^2 \\ &\le \sum_i p_i\|\nabla F_i(w_i^{t,j})-\nabla F_i(w^t)\|^2 \\ &\le L^2\sum_i p_i\|w_i^{t,j}-w^t\|^2. \end{aligned} \tag{100}$$

Substituting equation 100 into equation 99 and taking expectations yields

$$\begin{aligned} \mathbb{E}\|\overline{h}^{t,j}\|^2 \le{}& 3\mathbb{E}\|\nabla F(w^t)\|^2 \\ &+3L^2\mathbb{E}\sum_i p_i\|w_i^{t,j}-w^t\|^2 \\ &+3\mathbb{E}\left\|\sum_i p_i u_i^{t,j}\right\|^2. \end{aligned} \tag{101}$$

Averaging over j and substituting equation 98 and equation 101 into equation 95 gives the self-consistent inequality

$$C_t\le\eta^2K^2[3\mathbb{E}\|\nabla F(w^t)\|^2+3L^2C_t+3U_t^2+D_t^2+\sigma^2]. \tag{102}$$

Let $a = \eta^2 K^2 L^2$ and move $3aC_t$ to the left-hand side:

$$(1-3a)C_t \le \eta^2 K^2 [3\mathbb{E}\|\nabla F(w^t)\|^2 + 3U_t^2 + D_t^2 + \sigma^2]. \tag{103}$$

When $a \le 1/48$, $1 - 3a \ge 15/16 \ge 1/2$, and hence $(1-3a)^{-1} \le 2$. This gives equation 91.

We next derive the mean-update deviation. By equation 63,

$$\begin{aligned}\overline{h}_t - \nabla F(w^t) = \frac{1}{K}\sum_{k=0}^{K-1}\sum_i p_i[\nabla F_i(w_i^{t,k}) - \nabla F_i(w^t)] \\ + \frac{1}{K}\sum_{k=0}^{K-1}\sum_i p_i u_i^{t,k}.\end{aligned} \tag{104}$$

Applying equation 69 to the two terms and then Jensen's inequality over k gives

$$\begin{aligned}B_t \le \frac{2}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_i p_i[\nabla F_i(w_i^{t,k}) - \nabla F_i(w^t)]\right\|^2 \\ + \frac{2}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_i p_i u_i^{t,k}\right\|^2.\end{aligned} \tag{105}$$

The first term is further controlled by equation 100, while the second is controlled by the definition of U_t^2. Therefore,

$$B_t \le 2L^2 C_t + 2U_t^2. \tag{106}$$

Substituting equation 91 gives

$$\begin{aligned}B_t &\le 4\eta^2 K^2 L^2 [3\mathbb{E}\|\nabla F(w^t)\|^2 + 3U_t^2 + D_t^2 + \sigma^2] + 2U_t^2 \\ &= 12a\,\mathbb{E}\|\nabla F(w^t)\|^2 + (2+12a)U_t^2 + 4a(D_t^2 + \sigma^2),\end{aligned} \tag{107}$$

which proves equation 92. □

Proof of Theorem 2. By the L-smoothness of F and $w^{t+1} - w^t = -\gamma\widehat{h}_t$,

$$\begin{aligned}F(w^{t+1}) &\le F(w^t) + \langle \nabla F(w^t), w^{t+1} - w^t\rangle + \frac{L}{2}\|w^{t+1} - w^t\|^2 \\ &= F(w^t) - \gamma\langle \nabla F(w^t), \widehat{h}_t\rangle + \frac{L\gamma^2}{2}\|\widehat{h}_t\|^2.\end{aligned} \tag{108}$$

Let $g_t = \nabla F(w^t)$. Substituting the inner-product identity in equation 69 gives

$$\begin{aligned}F(w^{t+1}) &\le F(w^t) - \frac{\gamma}{2}\|g_t\|^2 - \frac{\gamma}{2}\|\widehat{h}_t\|^2 \\ &\quad + \frac{\gamma}{2}\|\widehat{h}_t - g_t\|^2 + \frac{L\gamma^2}{2}\|\widehat{h}_t\|^2 \\ &= F(w^t) - \frac{\gamma}{2}\|g_t\|^2 + \frac{\gamma}{2}\|\widehat{h}_t - g_t\|^2 \\ &\quad - \frac{\gamma}{2}(1 - L\gamma)\|\widehat{h}_t\|^2.\end{aligned} \tag{109}$$

Since $L\gamma = \eta K L = \sqrt{a} \le 1/\sqrt{48} < 1$, the last term is nonpositive and can be dropped while preserving the upper bound. Taking expectations then gives

$$\begin{aligned}\mathbb{E}F(w^{t+1}) \le \mathbb{E}F(w^t) - \frac{\gamma}{2}\mathbb{E}\|g_t\|^2 \\ + \frac{\gamma}{2}\mathbb{E}\|\widehat{h}_t - g_t\|^2.\end{aligned} \tag{110}$$

Decompose the update error as

$$\widehat{h}_t - g_t = (\widehat{h}_t - \overline{h}_t) + (\overline{h}_t - g_t), \tag{111}$$

and apply $\|x+y\|^2 \le 2\|x\|^2 + 2\|y\|^2$:

$$\mathbb{E}\|\widehat{h}_t - g_t\|^2 \le 2\omega_t^2 + 2B_t. \tag{112}$$

Substituting equation 112 into equation 110 yields

$$\mathbb{E}F(w^{t+1}) \le \mathbb{E}F(w^t) - \frac{\gamma}{2}\mathbb{E}\|g_t\|^2 + \gamma\omega_t^2 + \gamma B_t. \tag{113}$$

Substituting Lemma 1 further gives

$$\begin{aligned}\mathbb{E}F(w^{t+1}) \le \mathbb{E}F(w^t) - \gamma\left(\frac{1}{2} - 12a\right)\mathbb{E}\|g_t\|^2 \\ + \gamma\omega_t^2 + \gamma(2+12a)U_t^2 + 4a\gamma(D_t^2 + \sigma^2).\end{aligned} \tag{114}$$

Because $a \le 1/48$, we have $12a \le 1/4$, and hence $1/2 - 12a \ge 1/4$. Therefore,

$$\begin{aligned}\frac{\gamma}{4}\mathbb{E}\|g_t\|^2 \le \mathbb{E}F(w^t) - \mathbb{E}F(w^{t+1}) + \gamma\omega_t^2 \\ + \gamma(2+12a)U_t^2 + 4a\gamma(D_t^2 + \sigma^2).\end{aligned} \tag{115}$$

Summing over $t = 0, \dots, T-1$, the function-value terms telescope. Since $F(w^T) \ge F_\star$, we obtain

$$\begin{aligned}\frac{\gamma}{4}\sum_{t=0}^{T-1}\mathbb{E}\|g_t\|^2 \le F(w^0) - F_\star + \gamma\sum_{t=0}^{T-1}\omega_t^2 \\ + \gamma(2+12a)\sum_{t=0}^{T-1}U_t^2 \\ + 4a\gamma\sum_{t=0}^{T-1}(D_t^2 + \sigma^2).\end{aligned} \tag{116}$$

Dividing both sides by $\gamma T/4 = \eta K T/4$ and using equation 66 yields equation 41. □

Proof of Corollary 1. When $a \le 1/48$, $8 + 48a \le 9$, and $16a$ has the same order as $a = \eta^2 K^2 L^2$. Absorbing fixed numerical constants into $\mathcal{O}[\cdot]$, equation 41 directly yields equation 42.

For a prescribed budget T, choose $\eta = c_\eta/(L\sqrt{TK})$ before training. Substituting this value term by term gives

$$\frac{\mathcal{F}_{\mathrm{C}}}{\eta K T} = \frac{L\mathcal{F}_{\mathrm{C}}}{c_\eta\sqrt{TK}}, \qquad \eta^2 K^2 L^2 = \frac{c_\eta^2 K}{T}. \tag{117}$$

Moreover, $T \ge 48c_\eta^2 K$ ensures that $a = c_\eta^2 K/T \le 1/48$. Absorbing factors that depend only on the fixed constant c_η into the order notation yields equation 44. In this fixed-step-size specialization, the same η is used from round 0 through round $T-1$ for each training run, with no within-round or across-round learning-rate decay. □

Proof of Corollary 2. By Assumption 4 and Theorem 2, for $b \in \{\mathrm{base}, \mathrm{dir}\}$ define the shared sampling envelope

$$\begin{aligned}\mathcal{B}_b := \frac{4[F(w^0) - F_\star]}{\eta K T} + 4\overline{\omega}_{\mathrm{stoch}}^2 + 4\overline{b}_s^2 \\ + (4\chi_s + 16a)\overline{D}_b^2 + (8+48a)\overline{U}_b^2 + 16a\sigma^2.\end{aligned} \tag{118}$$

where the overbar denotes averaging over communication rounds. The initial gap, stochastic sampling term, common sampling bias, and noise term are shared by the two processes and cancel when their bounds are subtracted. Therefore,

$$\begin{aligned}\mathcal{B}_{\mathrm{dir}} - \mathcal{B}_{\mathrm{base}} = (4\chi_s + 16a)(\overline{D}_{\mathrm{dir}}^2 - \overline{D}_{\mathrm{base}}^2) \\ + (8+48a)(\overline{U}_{\mathrm{dir}}^2 - \overline{U}_{\mathrm{base}}^2).\end{aligned} \tag{119}$$

Using $\overline{D}_{\mathrm{dir}}^2 \le \kappa^2\overline{D}_{\mathrm{base}}^2 + \overline{r}$,

$$\begin{aligned}\mathcal{B}_{\mathrm{dir}} - \mathcal{B}_{\mathrm{base}} \le -(4\chi_s + 16a)[(1-\kappa^2)\overline{D}_{\mathrm{base}}^2 - \overline{r}] \\ + (8+48a)(\overline{U}_{\mathrm{dir}}^2 - \overline{U}_{\mathrm{base}}^2).\end{aligned} \tag{120}$$

condition equation 46 makes the right-hand side strictly negative. Hence, $\mathcal{B}_{\mathrm{dir}} < \mathcal{B}_{\mathrm{base}}$. □

A.8 Detailed Proof of the Generalization Error

Let $S^{(i,j)}$ denote the dataset obtained by replacing $z_{i,j}$ in S with an independent example $z'_{i,j} \sim \mathcal{D}_i$. The replacement index is sampled according to

$$\Pr(J = (i,j)) = \frac{p_i}{n_i} \tag{121}$$

so that the replacement distribution matches the weights in the empirical risk:

$$F_S = \sum_i \frac{p_i}{n_i} \sum_j \ell(\cdot; z_{i,j}). \tag{122}$$

Represent the complete FL training procedure by the augmented-state recursion

$$W_{q+1} = W_q - \eta_q H_q(W_q; Z_q), \tag{123}$$

and let $\mathsf{P}_g W_q$ denote its global inference state. For brevity, write $H_q = H_q(W_q; Z_q)$ and $H'_q = H_q(W'_q; Z'_q)$.

Proof of Theorem 3. Define the coupled-trajectory distance

$$\Delta_q = \mathbb{E}\|W_q - W'_q\|. \tag{124}$$

If the replaced example is not accessed at step q, i.e., if $\mathcal{A}_q^c$ occurs, then $Z_q = Z'_q$. By equation 48,

$$\begin{aligned} \|W_{q+1} - W'_{q+1}\| &= \|W_q - W'_q - \eta_q(H_q - H'_q)\| \\ &\le \|W_q - W'_q\| + \eta_q\|H_q - H'_q\| \\ &\le (1 + \eta_q L_{\Phi,q})\|W_q - W'_q\|. \end{aligned} \tag{125}$$

If $\mathcal{A}_q$ occurs, the two trajectories may access different example values. The triangle inequality then gives

$$\begin{aligned} \|W_{q+1} - W'_{q+1}\| &\le \|W_q - W'_q\| + \eta_q\|H_q - H'_q\| \\ &\le \|W_q - W'_q\| + \eta_q(\|H_q\| + \|H'_q\|). \end{aligned} \tag{126}$$

Combining the two events and adding the nonnegative term $\eta_q L_{\Phi,q}\|W_q - W'_q\|$ to the right-hand side gives the pathwise upper bound

$$\begin{aligned} \|W_{q+1} - W'_{q+1}\| \le (1 + \eta_q L_{\Phi,q})\|W_q - W'_q\| \\ + \eta_q \mathbf{1}_{\mathcal{A}_q}(\|H_q\| + \|H'_q\|). \end{aligned} \tag{127}$$

Taking the full expectation of equation 127 and applying the law of total expectation yields

$$\begin{aligned} &\mathbb{E}[\mathbf{1}_{\mathcal{A}_q}(\|H_q\| + \|H'_q\|)] \\ &\quad = \pi_q \mathbb{E}[\|H_q\| + \|H'_q\| \mid \mathcal{A}_q] \le 2\pi_q M_q. \end{aligned} \tag{128}$$

Therefore,

$$\Delta_{q+1} \le (1 + \eta_q L_{\Phi,q})\Delta_q + 2\eta_q \pi_q M_q. \tag{129}$$

Because the two training trajectories share the same initial state, $\Delta_0 = 0$. Unrolling the recursion step by step gives

$$\Delta_1 \le 2\eta_0 \pi_0 M_0, \tag{130}$$

$$\begin{aligned} \Delta_2 &\le (1 + \eta_1 L_{\Phi,1})\Delta_1 + 2\eta_1 \pi_1 M_1 \\ &\le 2\eta_0 \pi_0 M_0(1 + \eta_1 L_{\Phi,1}) + 2\eta_1 \pi_1 M_1. \end{aligned} \tag{131}$$

By induction, we obtain

$$\Delta_Q \le 2 \sum_{q=0}^{Q-1} \eta_q \pi_q M_q \prod_{r=q+1}^{Q-1} (1 + \eta_r L_{\Phi,r}). \tag{132}$$

We next explicitly connect replace-one stability to the expected generalization error. Let the training algorithm be $A(S) = \mathsf{P}_g W_Q(S)$. By the definitions of the population and empirical risks,

$$\begin{aligned} &\mathbb{E}_{S,A}[\mathcal{R}(A(S)) - F_S(A(S))] \\ &= \sum_i \frac{p_i}{n_i} \sum_{j=1}^{n_i} \mathbb{E}_{S, z'_{i,j}, A}[\ell(A(S); z'_{i,j}) - \ell(A(S); z_{i,j})]. \end{aligned} \tag{133}$$

Exchanging the identically distributed and independent variables $z_{i,j}$ and $z'_{i,j}$ in the joint distribution gives

$$\mathbb{E}[\ell(A(S); z_{i,j})] = \mathbb{E}[\ell(A(S^{(i,j)}); z'_{i,j})]. \tag{134}$$

Substituting equation 134 into equation 133 yields

$$\begin{aligned} &\mathbb{E}[\mathcal{R}(A(S)) - F_S(A(S))] \\ &= \sum_i \frac{p_i}{n_i} \sum_j \mathbb{E}[\ell(A(S); z'_{i,j}) - \ell(A(S^{(i,j)}); z'_{i,j})]. \end{aligned} \tag{135}$$

By the L_ℓ-Lipschitz continuity of the loss with respect to the global inference state and the non-expansiveness of P_g under the augmented-state norm,

$$\begin{aligned} &|\ell(A(S); z'_{i,j}) - \ell(A(S^{(i,j)}); z'_{i,j})| \\ &\quad \le L_\ell \|A(S) - A(S^{(i,j)})\| \\ &\quad \le L_\ell \|W_Q - W'_Q\|. \end{aligned} \tag{136}$$

Weighting the replacement indices according to equation 121 and taking expectations gives

$$\epsilon_{\text{gen}} \le L_\ell \Delta_Q. \tag{137}$$

Combining this result with equation 132 yields equation 50.

In the fixed-step-size specialization, let $\mathcal{U}_q$ denote the set of distinct training-example indices actually accessed at synchronous joint-state step q. Since $p_i = n_i/N_{\text{tr}}$, equation 121 assigns the same replacement probability to every training-example index:

$$\Pr(J = (i,j)) = \frac{p_i}{n_i} = \frac{1}{N_{\text{tr}}}. \tag{138}$$

When index sampling is independent of example values, conditioning first on $\mathcal{U}_q$ gives the probability of accessing the replaced index as

$$\begin{aligned} \Pr(\mathcal{A}_q \mid \mathcal{U}_q) &= \Pr(J \in \mathcal{U}_q \mid \mathcal{U}_q) \\ &= \sum_{(i,j)\in\mathcal{U}_q} \Pr(J = (i,j)) \\ &= \frac{|\mathcal{U}_q|}{N_{\text{tr}}}. \end{aligned} \tag{139}$$

Taking expectations again and using $|\mathcal{U}_q| \le B_{\text{eff}}$ yields

$$\pi_q = \Pr(\mathcal{A}_q) = \frac{\mathbb{E}|\mathcal{U}_q|}{N_{\text{tr}}} \le \frac{B_{\text{eff}}}{N_{\text{tr}}}. \tag{140}$$

This calculation jointly accounts for the mini-batches of all participating clients at the current depth without counting parallel client updates as additional state steps. Combining it with $M_q \le M$ and $L_{\Phi,q} \le L_\Phi$ gives

$$\begin{aligned} \epsilon_{\text{gen}} &\le \frac{2L_\ell M \eta B_{\text{eff}}}{N_{\text{tr}}} \sum_{q=0}^{Q-1} (1 + \eta L_\Phi)^{Q-1-q} \\ &= \frac{2L_\ell M B_{\text{eff}}}{N_{\text{tr}} L_\Phi} [(1 + \eta L_\Phi)^Q - 1], \end{aligned} \tag{141}$$

where, for $L_\Phi = 0$, the continuous limit is $2L_\ell M \eta Q B_{\text{eff}}/N_{\text{tr}}$. Using

$$(1+x)^Q - 1 \le Qx(1+x)^{Q-1} \le Qx\exp(xQ), \qquad x \ge 0, \tag{142}$$

and setting $Q = TK$ yields equation 51. □

Proof of Corollary 3. Theorem 3 gives

$$\epsilon_{\text{gen}} \le \frac{2L_\ell M \eta T K B_{\text{eff}}}{N_{\text{tr}}} \exp(L_\Phi \eta T K). \tag{143}$$

Absorbing the fixed numerical constant 2 yields equation 52. If $L_\Phi \eta T K = \mathcal{O}(1)$, there exists a constant C_Φ independent of T, K, and N_{tr} such that $\exp(L_\Phi \eta T K) \le C_\Phi$. This factor can therefore be absorbed into the constant in $\mathcal{O}[\cdot]$, yielding equation 53. □

We next provide the update-magnitude envelope used for the conditional generalization comparison in the main text. For $b \in \{\mathrm{base}, \mathrm{dir}\}$, under event $\mathcal{A}_q$, define the equally weighted averages over the two coupled trajectories, $G_{b,q}$, $C_{b,q}$, $U_{b,q}$, $D_{b,q}$, and $\Sigma_{b,q}$. These quantities respectively control the global gradient norm, local drift, conditional mean bias, inter-client conditional mean deviation, and the stochastic residual together with the remaining augmented-state components. After absorbing fixed product-norm constants into these quantities, define

$$\mathcal{M}_{b,q} = G_{b,q} + L\sqrt{C_{b,q}} + U_{b,q} + D_{b,q} + \Sigma_{b,q}. \tag{144}$$

This envelope follows from the decomposition below. For either coupled trajectory,

$$\begin{aligned} \mathbb{E}_i\|v_i^{t,k}\| &\le \mathbb{E}_i\|h_i^{t,k}\| + \mathbb{E}_i\|v_i^{t,k} - h_i^{t,k}\| \\ &\le \mathbb{E}_i\|h_i^{t,k}\| + \left(\sum_i p_i \mathbb{E}\|v_i^{t,k} - h_i^{t,k}\|^2\right)^{1/2}, \end{aligned} \tag{145}$$

where the second step applies the Cauchy–Schwarz inequality. Moreover,

$$\begin{aligned} \mathbb{E}_i\|h_i^{t,k}\| &\le \|\overline{h}^{t,k}\| + \sum_i p_i \|h_i^{t,k} - \overline{h}^{t,k}\| \\ &\le \|\overline{h}^{t,k}\| + \left(\sum_i p_i \|h_i^{t,k} - \overline{h}^{t,k}\|^2\right)^{1/2}, \end{aligned} \tag{146}$$

and

$$\begin{aligned} \|\overline{h}^{t,k}\| &\le \|\nabla F(w^t)\| \\ &\quad + \left\|\sum_i p_i [\nabla F_i(w_i^{t,k}) - \nabla F_i(w^t)]\right\| + \left\|\sum_i p_i u_i^{t,k}\right\| \\ &\le \|\nabla F(w^t)\| + L\left(\sum_i p_i \|w_i^{t,k} - w^t\|^2\right)^{1/2} + \left\|\sum_i p_i u_i^{t,k}\right\|. \end{aligned} \tag{147}$$

where the last step uses Jensen's inequality and L-smoothness. Conditioning on the access event and averaging equally over the two coupled trajectories gives $M_{b,q} \le \mathcal{M}_{b,q}$.

If Theorem 1 gives $D_{\mathrm{dir},q} \le \kappa_q D_{\mathrm{base},q} + \sqrt{r_q}$, then the sufficient condition

$$\begin{aligned} &(1-\kappa_q) D_{\mathrm{base},q} \\ &\quad > (G_{\mathrm{dir},q} - G_{\mathrm{base},q}) + L(\sqrt{C_{\mathrm{dir},q}} - \sqrt{C_{\mathrm{base},q}}) \\ &\qquad + (U_{\mathrm{dir},q} - U_{\mathrm{base},q}) + (\Sigma_{\mathrm{dir},q} - \Sigma_{\mathrm{base},q}) + \sqrt{r_q} \end{aligned} \tag{148}$$

ensures that $\mathcal{M}_{\mathrm{dir},q} < \mathcal{M}_{\mathrm{base},q}$. If, in addition, $L_{\Phi,\mathrm{dir},q} \le L_{\Phi,\mathrm{base},q}$, then neither the update magnitude at this step nor its subsequent propagation factor in the stability bound increases. If these conditions hold for every q, the overall bound in equation 50 does not increase. It becomes strictly tighter if at least one step satisfying $\eta_q \pi_q > 0$ is strictly improved.

A.9 Detailed Proof of the Flatness Discrepancy

Lemma 2 (Global–local sharpness discrepancy at a common center)**.** Under Assumptions 1 and 2, for a common center w and a fixed inference state b,

$$0 \le \sum_i p_i \mathcal{S}_i(w, b; \rho) - \mathcal{S}_F(w, b; \rho) \le \rho\zeta + L\rho^2. \tag{149}$$

Proof. Fix b and suppress it from the notation throughout this proof. For any $\|\delta\| \le \rho$, define

$$\phi_i(\delta) = F_i(w+\delta) - F_i(w), \qquad \phi(\delta) = \sum_i p_i \phi_i(\delta). \tag{150}$$

For each fixed δ, $\phi_i(\delta) \le \max_{\|\delta'\|\le\rho} \phi_i(\delta')$. Taking the weighted sum and then maximizing over δ gives

$$\begin{aligned}\mathcal{S}_F(w;\rho) &= \max_{\|\delta\|\le\rho} \sum_i p_i \phi_i(\delta) \\ &\le \sum_i p_i \max_{\|\delta\|\le\rho} \phi_i(\delta) = \sum_i p_i \mathcal{S}_i(w;\rho).\end{aligned} \tag{151}$$

This proves the left-hand inequality in equation 149.

For any client i and any feasible δ,

$$\phi_i(\delta) = \phi(\delta) + [\phi_i(\delta) - \phi(\delta)] \le \mathcal{S}_F(w;\rho) + |\phi_i(\delta) - \phi(\delta)|. \tag{152}$$

Maximizing over δ yields

$$\mathcal{S}_i(w;\rho) \le \mathcal{S}_F(w;\rho) + \sup_{\|\delta\|\le\rho} |\phi_i(\delta) - \phi(\delta)|. \tag{153}$$

By the fundamental theorem of calculus,

$$\phi_i(\delta) - \phi(\delta) = \int_0^1 \langle \nabla F_i(w + s\delta) - \nabla F(w + s\delta), \delta \rangle ds. \tag{154}$$

We decompose the gradient discrepancy along the path through its value at the center:

$$\begin{aligned}&\nabla F_i(w + s\delta) - \nabla F(w + s\delta) \\ &= [\nabla F_i(w + s\delta) - \nabla F_i(w)] + [\nabla F_i(w) - \nabla F(w)] \\ &\quad + [\nabla F(w) - \nabla F(w + s\delta)].\end{aligned} \tag{155}$$

Because every F_i is L-smooth, their weighted average F is also L-smooth. The triangle inequality therefore gives

$$\begin{aligned}&\|\nabla F_i(w + s\delta) - \nabla F(w + s\delta)\| \\ &\quad \le Ls\|\delta\| + \|\nabla F_i(w) - \nabla F(w)\| + Ls\|\delta\| \\ &\quad = \|\nabla F_i(w) - \nabla F(w)\| + 2Ls\|\delta\|.\end{aligned} \tag{156}$$

Substituting equation 156 into equation 154 and applying the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}|\phi_i(\delta) - \phi(\delta)| &\le \int_0^1 [\|\nabla F_i(w) - \nabla F(w)\| + 2Ls\|\delta\|]\|\delta\| \, ds \\ &= \|\delta\|\|\nabla F_i(w) - \nabla F(w)\| + L\|\delta\|^2 \\ &\le \rho\|\nabla F_i(w) - \nabla F(w)\| + L\rho^2.\end{aligned} \tag{157}$$

Substituting equation 157 into equation 153 and taking the weighted sum over i gives

$$\begin{aligned}\sum_i p_i \mathcal{S}_i(w;\rho) - \mathcal{S}_F(w;\rho) \le \rho \sum_i p_i \|\nabla F_i(w) - \nabla F(w)\| \\ + L\rho^2.\end{aligned} \tag{158}$$

Finally, applying the Cauchy–Schwarz inequality to the weighted first moment yields

$$\begin{aligned}\sum_i p_i \|\nabla F_i(w) - \nabla F(w)\| &\le \left(\sum_i p_i \|\nabla F_i(w) - \nabla F(w)\|^2 \right)^{1/2} \\ &\le \zeta.\end{aligned} \tag{159}$$

This establishes equation 149. □

Lemma 3 (Stability of sharpness with respect to the parameter center and inference state)**.** Under Assumptions 1 and 6,

$$|\mathcal{S}_i(u, b; \rho) - \mathcal{S}_i(v, b'; \rho)| \le L\rho\|u - v\| + L_{b,\rho}\|b - b'\|. \tag{160}$$

Proof. We first fix the inference state b and define

$$\psi_{\delta,b}(w) = F_i(w+\delta; b) - F_i(w; b). \tag{161}$$

For any $\|\delta\| \le \rho$,

$$\begin{aligned} \|\nabla_w \psi_{\delta,b}(w)\| &= \|\nabla F_i(w+\delta; b) - \nabla F_i(w; b)\| \\ &\le L\|\delta\| \le L\rho. \end{aligned} \tag{162}$$

The mean-value inequality therefore implies

$$|\psi_{\delta,b}(u) - \psi_{\delta,b}(v)| \le L\rho\|u - v\|. \tag{163}$$

We next derive the inequality for the difference of maxima. Let $a_\delta = \psi_{\delta,b}(u)$ and $c_\delta = \psi_{\delta,b}(v)$. For any δ,

$$a_\delta \le c_\delta + |a_\delta - c_\delta| \le \max_{\delta'} c_{\delta'} + \max_{\delta'} |a_{\delta'} - c_{\delta'}|. \tag{164}$$

Maximizing the left-hand side gives

$$\max_\delta a_\delta - \max_\delta c_\delta \le \max_\delta |a_\delta - c_\delta|. \tag{165}$$

Interchanging a_δ and c_δ gives the reverse inequality. Hence,

$$\left| \max_\delta a_\delta - \max_\delta c_\delta \right| \le \max_\delta |a_\delta - c_\delta|. \tag{166}$$

Combining equation 163 and equation 166 yields

$$|\mathcal{S}_i(u, b; \rho) - \mathcal{S}_i(v, b; \rho)| \le L\rho\|u - v\|. \tag{167}$$

Finally, inserting the intermediate term $\mathcal{S}_i(v, b; \rho)$ and applying the triangle inequality together with Assumption 6, we obtain

$$\begin{aligned} &|\mathcal{S}_i(u, b; \rho) - \mathcal{S}_i(v, b'; \rho)| \\ &\quad \le |\mathcal{S}_i(u, b; \rho) - \mathcal{S}_i(v, b; \rho)| \\ &\qquad + |\mathcal{S}_i(v, b; \rho) - \mathcal{S}_i(v, b'; \rho)| \\ &\quad \le L\rho\|u - v\| + L_{b,\rho}\|b - b'\|. \end{aligned} \tag{168}$$

□

Proof of Theorem 4. We insert the local sharpness evaluated at the common center and common inference state into Δ_F:

$$\begin{aligned} \Delta_F &= \left| \mathcal{S}_F(w, b; \rho) - \sum_i p_i \mathcal{S}_i(w_i, b_i; \rho) \right| \\ &\le \left| \mathcal{S}_F(w, b; \rho) - \sum_i p_i \mathcal{S}_i(w, b; \rho) \right| \\ &\qquad + \left| \sum_i p_i [\mathcal{S}_i(w, b; \rho) - \mathcal{S}_i(w_i, b_i; \rho)] \right| \\ &\le \left| \mathcal{S}_F(w, b; \rho) - \sum_i p_i \mathcal{S}_i(w, b; \rho) \right| \\ &\qquad + \sum_i p_i |\mathcal{S}_i(w, b; \rho) - \mathcal{S}_i(w_i, b_i; \rho)|. \end{aligned} \tag{169}$$

The first term is controlled by Lemma 2. Applying Lemma 3 to the second term client by client gives

$$\Delta_F \le \rho\zeta + L\rho^2 + L\rho \sum_i p_i \|w - w_i\| + L_{b,\rho} \sum_i p_i \|b - b_i\|, \tag{170}$$

which is equation 57.

We next prove the curvature-coverage form. Since

$$\|\epsilon_i(w_i)\| = \rho \frac{\|\nabla F_i(w_i)\|}{\|\nabla F_i(w_i)\| + \delta} \le \rho, \tag{171}$$

the perturbation is feasible for the sharpness maximization problem. Moreover, because $\delta = 0$ is feasible, $\mathcal{S}_i(w_i, b_i; \rho) \ge 0$. Therefore,

$$0 \le \rho^2 R_{\mathrm{curv},i}(w_i; b_i) \le \mathcal{S}_i(w_i, b_i; \rho). \tag{172}$$

It follows that $e_{\mathrm{cov},i} \ge 0$. By equation 151 and Lemma 3,

$$\begin{aligned} \mathcal{S}_F(w, b; \rho) &\le \sum_i p_i \mathcal{S}_i(w, b; \rho) \\ &\le \sum_i p_i \mathcal{S}_i(w_i, b_i; \rho) + L\rho \sum_i p_i \|w - w_i\| \\ &\quad + L_{b,\rho} \sum_i p_i \|b - b_i\|. \end{aligned} \tag{173}$$

By the definition of the coverage error,

$$\mathcal{S}_i(w_i, b_i; \rho) = \rho^2 R_{\mathrm{curv},i}(w_i; b_i) + e_{\mathrm{cov},i}(w_i, b_i; \rho). \tag{174}$$

Substitution gives equation 59. □

We finally derive the communication-endpoint form. Define the ideal full-participation aggregation endpoint as

$$\widetilde{w}^{t+1} = \sum_i p_i w_i^{t,K}, \qquad \nu_t^2 = \mathbb{E}\|w^{t+1} - \widetilde{w}^{t+1}\|^2. \tag{175}$$

By the triangle inequality,

$$\begin{aligned} \|\widetilde{w}^{t+1} - w_i^{t,K}\| &= \left\| \sum_j p_j (w_j^{t,K} - w_i^{t,K}) \right\| \\ &\le \sum_j p_j \|w_j^{t,K} - w_i^{t,K}\| \\ &\le \sum_j p_j [\|w_j^{t,K} - w^t\| + \|w_i^{t,K} - w^t\|]. \end{aligned} \tag{176}$$

Taking the weighted sum over i yields

$$\sum_i p_i \|\widetilde{w}^{t+1} - w_i^{t,K}\| \le 2 \sum_i p_i \|w_i^{t,K} - w^t\|. \tag{177}$$

On the other hand, the local recursion and equation 67 imply

$$\begin{aligned} \mathbb{E} \sum_i p_i \|w_i^{t,K} - w^t\|^2 &= \eta^2 \mathbb{E} \sum_i p_i \left\| \sum_{k=0}^{K-1} v_i^{t,k} \right\|^2 \\ &\le \eta^2 K \sum_{k=0}^{K-1} \mathbb{E} \sum_i p_i \|v_i^{t,k}\|^2 \\ &= \eta^2 K^2 \mathcal{M}_t. \end{aligned} \tag{178}$$

Applying Jensen's and Cauchy–Schwarz inequalities to the weighted first moment,

$$\begin{aligned} \sum_i p_i \mathbb{E}\|w_i^{t,K} - w^t\| &\le \left(\mathbb{E} \sum_i p_i \|w_i^{t,K} - w^t\|^2 \right)^{1/2} \\ &\le \eta K \sqrt{\mathcal{M}_t}. \end{aligned} \tag{179}$$

Combining this result with equation 177 gives

$$\sum_i p_i \mathbb{E}\|\widetilde{w}^{t+1} - w_i^{t,K}\| \le 2\eta K\sqrt{\mathcal{M}_t}. \tag{180}$$

Another application of the triangle inequality yields

$$\begin{aligned}\sum_i p_i \mathbb{E}\|w^{t+1} - w_i^{t,K}\| \le \mathbb{E}\|w^{t+1} - \widetilde{w}^{t+1}\| \\ + \sum_i p_i \mathbb{E}\|\widetilde{w}^{t+1} - w_i^{t,K}\|.\end{aligned} \tag{181}$$

By Jensen's inequality, the first term satisfies

$$\mathbb{E}\|w^{t+1} - \widetilde{w}^{t+1}\| \le \sqrt{\mathbb{E}\|w^{t+1} - \widetilde{w}^{t+1}\|^2} = \sqrt{\nu_t^2}. \tag{182}$$

Therefore,

$$\sum_i p_i \mathbb{E}\|w^{t+1} - w_i^{t,K}\| \le \sqrt{\nu_t^2} + 2\eta K\sqrt{\mathcal{M}_t}. \tag{183}$$

Define

$$\beta_t = \sum_i p_i \mathbb{E}\|b_g^{t+1} - b_i^{t,K}\|, \tag{184}$$

and substitute equation 183 into equation 57. This gives equation 61.

It remains to prove equation 62. Averaging equation 98 and equation 101 over k gives

$$\begin{aligned}\mathcal{M}_t &= \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\sum_i p_i \|v_i^{t,k}\|^2 \\ &\le \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|\overline{h}^{t,k}\|^2 + D_t^2 + \sigma^2 \\ &\le 3\mathbb{E}\|\nabla F(w^t)\|^2 + 3L^2 C_t + 3U_t^2 + D_t^2 + \sigma^2.\end{aligned} \tag{185}$$

This completes the derivation of the communication-endpoint flatness-discrepancy bound.

B Experimental Setup Details

This appendix provides additional details on the benchmark datasets, data partitioning protocol, model architecture, optimization settings, and FedCADC-specific hyperparameters. All experiments are implemented in PyTorch, and all compared methods use the same data partitions, client participation protocol, local training budget, and evaluation procedure. We also provide the pseudocode of FedCADC in Algorithm 1.

B.1 Benchmarks and Data Partitioning

We evaluate FedCADC on CIFAR-10 and CIFAR-100. Both datasets contain 50,000 training images and 10,000 test images, where each image has a resolution of 32×32 and three color channels. The CIFAR-10 and CIFAR-100 training sets contain 10 and 100 classes, respectively.

To simulate statistical heterogeneity, the training data are distributed among 100 clients according to a Dirichlet distribution. Specifically, for each class, the proportions assigned to different clients are sampled using a concentration parameter $\alpha_{\mathrm{Dir}} \in \{0.1, 0.5, 10\}$. A smaller α_{Dir} produces more imbalanced class distributions and therefore corresponds to a stronger non-IID setting, whereas a larger α_{Dir} results in more balanced client distributions. The same client partition is used by all methods under each experimental configuration.

The input images are converted to tensors and normalized using the dataset-specific statistics. For CIFAR-10, the normalization mean and standard deviation are $(0.4914, 0.4822, 0.4465)$ and $(0.2470, 0.2435, 0.2616)$, respectively. For CIFAR-100, the corresponding values are $(0.5071, 0.4867, 0.4408)$ and $(0.2675, 0.2565, 0.2761)$. No additional random cropping or horizontal flipping is applied.

Algorithm 1 FedCADC

Require: Initial global model w^0, client set $\mathcal{C}$, communication rounds T, local epochs E, batch size B, client-specific mini-batch steps $K_i^t = E\lceil n_i/B \rceil$, round step sizes $\{\eta_t\}_{t=0}^{T-1}$, perturbation radius ρ, constants δ, A_{curv}, A_{dir}.
Ensure: Final global model w^T.

for $t = 0, 1, \ldots, T-1$ **do**
  Server samples clients $\mathcal{S}_t \subseteq \mathcal{C}$.
  if $t = 0$ or $\|w^{t-1} - w^t\| \le \delta$ **then**
    $g^t \leftarrow \mathbf{0}$
  else
    $g^t \leftarrow \dfrac{w^{t-1} - w^t}{\|w^{t-1} - w^t\| + \delta}$
  end if
  Server broadcasts w^t and g^t to all $i \in \mathcal{S}_t$.
  for each client $i \in \mathcal{S}_t$ in parallel **do**
    $w_i^{t,0} \leftarrow w^t$
    for $k = 0, 1, \ldots, K_i^t - 1$ **do**
      Draw the current mini-batch $\mathcal{B}_i^{t,k}$.
      $g_c^{t,k} \leftarrow \nabla F_{\mathcal{B}_i^{t,k}}(w_i^{t,k})$
      $\epsilon_i^{t,k} \leftarrow \rho \dfrac{g_c^{t,k}}{\|g_c^{t,k}\| + \delta}$
      $\widetilde{w}_i^{t,k} \leftarrow w_i^{t,k} + \epsilon_i^{t,k}$
      $g_p^{t,k} \leftarrow \nabla F_{\mathcal{B}_i^{t,k}}(\widetilde{w}_i^{t,k})$
      $\alpha_{\mathrm{curv}}^{t,k} \leftarrow \mathrm{clip}\left((1 - \cos(g_p^{t,k}, g_c^{t,k})) \left| \log \dfrac{\|g_p^{t,k}\| + \delta}{\|g_c^{t,k}\| + \delta} \right|, 0, A_{\mathrm{curv}} \right)$
      $R_{\mathrm{curv}}(w_i^{t,k}) \leftarrow \dfrac{\left[F_{\mathcal{B}_i^{t,k}}(w_i^{t,k} + \epsilon_i^{t,k}) - F_{\mathcal{B}_i^{t,k}}(w_i^{t,k}) \right]_+}{\rho^2}$
      $d_i^{t,k} \leftarrow w^t - w_i^{t,k}$
      if $\|g^t\| \le \delta$ or $\|d_i^{t,k}\| \le \delta$ **then**
        $R_{\mathrm{dir}}(w_i^{t,k}) \leftarrow 0, \quad \alpha_{\mathrm{dir}}^{t,k} \leftarrow 0$
      else
        $R_{\mathrm{dir}}(w_i^{t,k}) \leftarrow 1 - \cos(d_i^{t,k}, g^t)$
        $\alpha_{\mathrm{dir}}^{t,k} \leftarrow \mathrm{clip}\left(R_{\mathrm{dir}}(w_i^{t,k}), 0, A_{\mathrm{dir}} \right)$
      end if
      Treat $\epsilon_i^{t,k}$, $\alpha_{\mathrm{curv}}^{t,k}$, and $\alpha_{\mathrm{dir}}^{t,k}$ as constants in the current backward pass.
      $\Phi_i^{t,k}(w) \leftarrow F_{\mathcal{B}_i^{t,k}}(w) + \alpha_{\mathrm{curv}}^{t,k} R_{\mathrm{curv}}(w) + \alpha_{\mathrm{dir}}^{t,k} R_{\mathrm{dir}}(w)$
      $w_i^{t,k+1} \leftarrow w_i^{t,k} - \eta_t \nabla \Phi_i^{t,k}(w_i^{t,k})$
    end for
    Client i uploads w_i^{t,K_i^t}.
  end for
  $w^{t+1} \leftarrow \sum_{i \in \mathcal{S}_t} p_i^t w_i^{t,K_i^t}, \quad p_i^t = \dfrac{n_i}{\sum_{j \in \mathcal{S}_t} n_j}$
end for
return w^T

B.2 Model Architecture

For both CIFAR-10 and CIFAR-100, we use the CIFAR version of ResNet-20. The network begins with a 3×3 convolution that maps the three input channels to 16 feature channels, followed by batch normalization. The feature extractor contains three residual stages with channel widths 16, 32, and

64. Each stage consists of three BasicBlocks, and each BasicBlock contains two 3×3 convolutions, batch normalization, ReLU activation, and an identity shortcut.

The first block of the second and third stages uses stride 2 for spatial downsampling. When the spatial resolution or the number of channels changes, the shortcut branch uses a 1×1 convolution followed by batch normalization to match the dimensions of the residual branch. After the three residual stages, adaptive average pooling reduces the feature map to a single spatial location. The resulting 64-dimensional feature vector is fed into a linear classifier with 10 output units for CIFAR-10 and 100 output units for CIFAR-100.

The convolutional and linear layers are initialized using Kaiming normal initialization. The scale and bias parameters of batch normalization layers are initialized to 1 and 0, respectively, and the bias of the final linear classifier is initialized to 0.

B.3 Hyperparameters

For FedCADC, local models are optimized using stochastic gradient descent with learning rate 0.01, momentum 0.9, and weight decay 5×10^{-4}. The mini-batch size is set to 50, and the learning rate remains constant throughout training. The local classification objective is the cross-entropy loss. The perturbation radius is set to $\rho = 0.05$. The maximum coefficients of the curvature-aware and global direction-consistency regularizers are set to $A_{\mathrm{curv}} = 0.08$ and $A_{\mathrm{dir}} = 0.8$, respectively. These coefficients serve as upper bounds for the adaptive weights rather than fixed regularization strengths. Their values are computed separately for each mini-batch according to the local curvature response and the deviation between the local trajectory and the global optimization direction.

For the remaining methods, the local learning rate is set to $\eta_\ell = 0.01$, the global learning rate is set to $\eta_g = 1$, the weight decay is set to 5×10^{-4}, and the mini-batch size is set to 50. For FedProx, the proximal coefficient μ is tuned over the interval $[0.1, 1]$. For SCAFFOLD, the global learning rate is tuned over the interval $[0.01, 0.1]$.

We use method-specific perturbation settings following the corresponding implementations. FedSAM, MoFedSAM, FedGAMMA, and FedSMOO use $\rho = 0.05$, with the momentum coefficient of MoFedSAM tuned over $[0.1, 0.9]$ (Caldarola et al., 2022; Sun et al., 2023; Dai et al., 2023). FedASAM instead uses a base perturbation radius $\rho = 0.7$, an adaptive scaling parameter $\eta = 0.2$, and a local learning rate of 0.01. Its perturbation is scaled coordinate-wise according to parameter magnitudes rather than using the isotropic perturbation of FedSAM (Kwon et al., 2021; Caldarola et al., 2022).

For FedGF, the threshold T_D used to determine the value of c is tuned over the interval $[0.1, 0.3]$. We use $T_D = 0.2$ for CIFAR-10 and $T_D = 0.3$ for CIFAR-100. The perturbation radius is fixed to $\rho = 0.05$, consistent with the other perturbation-based methods. The window size W is tuned over the integer candidate set $\{10, 30, 50\}$ (Caldarola et al., 2022; Lee & Yoon, 2024).

For FedMut, the mutation strength is set to 4.0, following the value reported in the original FedMut paper (Hu et al., 2024). FedAvg and the remaining baselines use the common optimization settings and their standard algorithmic updates. FedSWA does not use perturbations during local training. Its local learning rate starts from $2 \times 0.01 = 0.02$ and decreases linearly toward zero within each local training phase, while the momentum and weight decay remain fixed at 0.9 and 5×10^{-4}, respectively (Liu et al., 2025). FedVSSAM uses $\rho = 0.05$, with the local and global direction mixing coefficients set to $\gamma_l = 0.5$ and $\gamma_g = 0.5$, respectively. Its server-side learning rate is set to 1.0 (Xiao et al., 2026). FedGloSS uses $\rho_{\mathrm{client}} = 0.05$, $\rho_{\mathrm{server}} = 0.05$, and the correction coefficient $\beta_{\mathrm{GloSS}} = 10.0$ (Caldarola et al., 2025). FedScar uses $\rho = 0.05$, a bias coefficient $\alpha_{\mathrm{bias}} = 0.1$, a bias decay factor $\gamma_{\mathrm{bias}} = 0.9$, and a regularization coefficient $\beta_{\mathrm{Scar}} = 0.06$ (Lu et al., 2026).

The main training protocol and method-specific hyperparameters are summarized in Table 4.

C Flatness Metric Definitions

This section provides the definitions of the three flatness metrics used in our experiments: the mean absolute loss variation (MALV), the largest Hessian eigenvalue, and the local–global flatness dis-

Table 4: Training protocol and method-specific hyperparameters.

Setting or method	**Value**
Datasets	CIFAR-10 and CIFAR-100
Model	ResNet-20
Number of clients	100
Participating clients per round	$\{5, 10, 20\}$
Dirichlet concentration	$\alpha_{\mathrm{Dir}} \in \{0.1, 0.5, 10\}$
Communication rounds	500
Local epochs	5
FedCADC settings:	
Optimizer	SGD
Mini-batch size	50
Learning rate	0.01
Momentum	0.9
Weight decay	5×10^{-4}
Perturbation radius	$\rho = 0.05$
Curvature coefficient upper bound	$A_{\mathrm{curv}} = 0.08$
Direction coefficient upper bound	$A_{\mathrm{dir}} = 0.8$
Common baseline settings:	
Local learning rate	$\eta_\ell = 0.01$
Global learning rate	$\eta_g = 1$
Mini-batch size	50
Weight decay	5×10^{-4}
FedProx	$\mu \in [0.1, 1]$
SCAFFOLD	$\eta_g \in [0.01, 0.1]$
FedSAM	$\rho = 0.05$
FedASAM	$\rho = 0.7,\ \eta = 0.2$
MoFedSAM	$\beta \in [0.1, 0.9]$, $\rho = 0.05$
FedGAMMA	$\rho = 0.05$
FedSMOO	$\rho = 0.05$
FedGF	$T_D \in [0.1, 0.3]$, $\rho = 0.05$, $W \in \{10, 30, 50\}$
FedMut	mutation strength $= 4.0$
FedSWA	initial learning rate $= 2 \times 0.01 = 0.02$
FedVSSAM	$\rho = 0.05$, $\gamma_l = 0.5$, $\gamma_g = 0.5$, $\eta_g = 1.0$
FedGloSS	$\rho_{\mathrm{client}} = 0.05$, $\rho_{\mathrm{server}} = 0.05$, $\beta_{\mathrm{GloSS}} = 10.0$
FedScar	$\rho = 0.05$, $\alpha_{\mathrm{bias}} = 0.1$, $\gamma_{\mathrm{bias}} = 0.9$, $\beta_{\mathrm{Scar}} = 0.06$

crepancy Δ_F. All metrics are evaluated on the final global model, and smaller values indicate flatter solutions.

C.1 Mean Absolute Loss Variation

Let $F(w)$ denote the empirical loss evaluated on the flatness evaluation set, and let $\sigma > 0$ be the perturbation scale. We draw J random directions from an isotropic Gaussian distribution:

$$u_j \sim \mathcal{N}(\mathbf{0}, I), \qquad d_j = \frac{u_j}{\|u_j\|_2}, \qquad j = 1, \ldots, J. \tag{186}$$

The mean absolute loss variation is computed as

$$\mathrm{MALV}_{\mathrm{raw}}(w) = \frac{1}{J} \sum_{j=1}^{J} \left| F(w + \sigma d_j) - F(w) \right|. \tag{187}$$

This metric approximates the loss variation around the current model under Gaussian low-pass perturbations. A smaller value indicates that the loss changes more smoothly within the perturbation neighborhood. In the reported table, the metric is expressed in percentage points as

$$\mathrm{MALV}_{\mathrm{pp}}(w) = 100\, \mathrm{MALV}_{\mathrm{raw}}(w). \tag{188}$$

The perturbations are applied to the trainable weight parameters, while bias parameters and batch-normalization affine parameters are excluded from the random perturbation directions.

C.2 Largest Hessian Eigenvalue

Let

$$H(w) = \nabla_w^2 F(w) \tag{189}$$

denote the Hessian matrix of the loss with respect to the model parameters. The largest Hessian eigenvalue is defined by the Rayleigh quotient:

$$\lambda_{\max}(w) = \max_{\|v\|_2=1} v^\top H(w) v. \tag{190}$$

In practice, $\lambda_{\max}(w)$ is estimated using power iteration and Hessian–vector products. Starting from a normalized vector $v^{(0)}$, the iteration is given by

$$\widetilde{v}^{(q+1)} = H(w)v^{(q)}, \qquad v^{(q+1)} = \frac{\widetilde{v}^{(q+1)}}{\|\widetilde{v}^{(q+1)}\|_2}. \tag{191}$$

After Q iterations, the estimated largest eigenvalue is computed as

$$\widehat{\lambda}_{\max}(w) = \left(v^{(Q)}\right)^\top H(w) v^{(Q)}. \tag{192}$$

A smaller $\lambda_{\max}$ indicates weaker curvature in the sharpest parameter direction.

C.3 Local–Global Flatness Discrepancy

For a model w and perturbation radius $\rho > 0$, define the global sharpness as

$$\mathcal{S}_F(w;\rho) = \max_{\|\delta\|_2 \le \rho} \left[F(w+\delta) - F(w)\right]. \tag{193}$$

Similarly, for client i, the local sharpness at model w_i is

$$\mathcal{S}_i(w_i;\rho) = \max_{\|\delta_i\|_2 \le \rho} \left[F_i(w_i+\delta_i) - F_i(w_i)\right]. \tag{194}$$

The local–global flatness discrepancy is defined as

$$\Delta_F = \left| \mathcal{S}_F(w;\rho) - \sum_{i\in\mathcal{S}_t} p_i^t \mathcal{S}_i(w_i;\rho) \right|, \qquad p_i^t = \frac{n_i}{\sum_{j\in\mathcal{S}_t} n_j}. \tag{195}$$

Following the SAM approximation, the perturbation direction for a model w is computed as

$$\epsilon(w) = \rho \frac{\nabla F(w)}{\|\nabla F(w)\|_2 + \varepsilon_{\text{num}}}, \tag{196}$$

where ε_{num} is a small numerical stabilization constant. Therefore, the practical sharpness estimator is

$$\widehat{\mathcal{S}}(w;\rho) = \left[F\big(w+\epsilon(w)\big) - F(w)\right]_+, \tag{197}$$

where $[x]_+ = \max(x, 0)$. The empirical estimator used in the experiments is consequently

$$\widehat{\Delta}_F = \left| \widehat{\mathcal{S}}_F(w;\rho) - \sum_{i\in\mathcal{S}_t} p_i^t \widehat{\mathcal{S}}_i(w_i;\rho) \right|. \tag{198}$$

A smaller Δ_F indicates that the flatness of the aggregated global model is closer to the weighted flatness of the participating local models.

D In-depth Comparison to the Existing Methods

D.1 Visual Comparison of Optimization Mechanisms

To further illustrate the differences between FedCADC and existing flatness-driven FL methods, we visually compare the optimization mechanisms of MoFedSAM, FedSMOO, FedGF, and FedCADC. Figure 3 illustrates the three representative baselines, while Figure 4 presents the curvature-aware regularizer, the direction consistency regularizer, and the unified FedCADC framework.

As shown in Figures 3 and 4, MoFedSAM, FedSMOO, FedGF, and FedCADC all seek to mitigate the mismatch between local flatness seeking and global optimization under heterogeneous FL. However, they differ substantially in their underlying mechanisms. MoFedSAM incorporates global momentum into local SAM to guide client updates, while its perturbation remains primarily determined by the local objective (Qu et al., 2022). FedSMOO uses correction information to adjust local perturbations, thereby reducing their deviation from global optimization under non-IID data (Sun et al., 2023). FedGF directly addresses the discrepancy between local and global flatness by interpolating local and global perturbations (Lee & Yoon, 2024). In contrast, FedCADC jointly introduces a curvature-aware flatness constraint and a global direction consistency constraint into the local objective. The former characterizes the local curvature response using the directional and norm differences between the center and perturbed gradients. The latter estimates a global guidance direction from consecutive global models and constrains the local update trajectory accordingly. FedCADC thus coordinates client-side flatness seeking through two complementary aspects: local loss geometry and the global optimization direction.

D.2 From Perturbed-Objective Optimization to Geometric Response Regulation

Most SAM-based FL methods improve local generalization by optimizing a perturbed loss. Although this strategy can reduce local sharpness, it implicitly couples two different effects in the perturbed gradient: the descent direction of the empirical risk and the gradient variation induced by the perturbation. Specifically, let

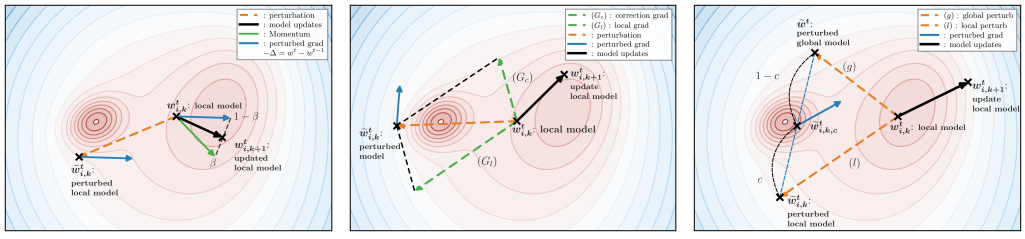
$$g_{c,i}^{t,k} = \nabla F_{\xi_i^{t,k}}(w_i^{t,k}), \qquad g_{p,i}^{t,k} = \nabla F_{\xi_i^{t,k}}\left(w_i^{t,k} + \epsilon_i^{t,k}\right). \tag{199}$$

A SAM-style update directly uses $g_{p,i}^{t,k}$. Under a local second-order approximation,

$$g_{p,i}^{t,k} = g_{c,i}^{t,k} + \nabla^2 F_{\xi_i^{t,k}}(w_i^{t,k})\epsilon_i^{t,k} + O\left(\left\|\epsilon_i^{t,k}\right\|^2\right). \tag{200}$$

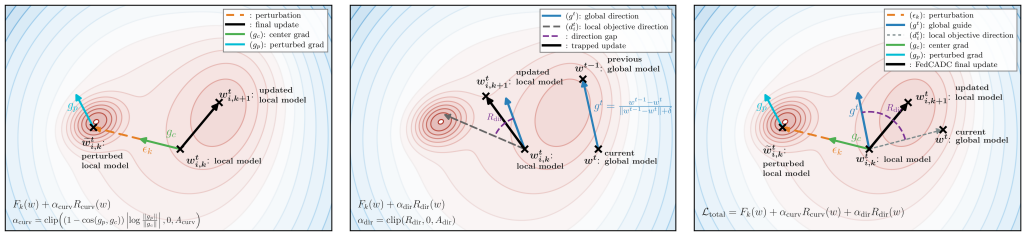
Thus, the perturbed update combines empirical-risk descent with a perturbation-induced geometric response. Under heterogeneous data, however, the gradient scale and local curvature can vary substantially across clients. Consequently, the same SAM-style update rule may produce inconsistent correction magnitudes and does not explicitly distinguish task-related descent from local sharpness correction.

FedCADC adopts a different design principle: the perturbation is used as a geometric probe rather than as a replacement for the empirical-risk gradient. Instead of directly optimizing the perturbed



(a) MoFedSAM (b) FedSMOO (c) FedGF

Figure 3: Schematic of MoFedSAM, FedSMOO, and FedGF. The contour lines illustrate the local loss landscape.



(a) Curvature-aware regularization (b) Direction consistency (c) Unified FedCADC framework

Figure 4: FedCADC curvature regularization, direction consistency, and unified optimization.

gradient, FedCADC measures the change between the center and perturbed losses and introduces it as a separate curvature-aware regularizer. With the perturbation and its coefficient treated as fixed during differentiation, the gradient of this regularizer on the positive branch is

$$\nabla R^{t,k}_{\mathrm{curv},i}(w^{t,k}_i) = \frac{g^{t,k}_{p,i} - g^{t,k}_{c,i}}{\rho^2}. \tag{201}$$

When the adaptive weighting mechanism is omitted and a fixed coefficient λ_{curv} is used, the corresponding local gradient becomes

$$g^{t,k}_{\mathrm{CADC},i} = g^{t,k}_{c,i} + \lambda_{\mathrm{curv}} \frac{g^{t,k}_{p,i} - g^{t,k}_{c,i}}{\rho^2}. \tag{202}$$

This formulation preserves the center gradient as the task-learning direction and applies the perturbation response as an explicitly controlled geometric correction.

The distinction is also evident from the local expansion of the perturbed loss:

$$F(w+\epsilon) = F(w) + \langle \nabla F(w), \epsilon \rangle + \frac{1}{2}\epsilon^\top \nabla^2 F(w)\epsilon + O\left(\|\epsilon\|^3\right). \tag{203}$$

The SAM surrogate contains both the first-order gradient term and the second-order curvature term. In contrast, under the stop-gradient implementation, the gradient of the FedCADC loss-difference regularizer cancels the center-gradient contribution and retains the perturbation-induced variation $g_p - g_c$, which is approximately $\nabla^2 F(w)\epsilon$. FedCADC therefore explicitly decomposes empirical-risk descent and local geometric correction, rather than treating the perturbed loss as a single indivisible optimization target.

This response-based view is particularly important in FL. A client may reach a locally flat region that is geometrically incompatible with the optimization trend of other clients. Merely modifying the local perturbation does not directly specify how the resulting local trajectory should coordinate with the global model. FedCADC consequently introduces a second constraint in the parameter space:

$$R^{t,k}_{\mathrm{dir},i}(w) = 1 - \cos\left(w^t - w, g^t\right), \tag{204}$$

where g^t is estimated from consecutive global models. The curvature constraint controls the sensitivity of a client to local perturbations, whereas the direction-consistency constraint determines whether its local trajectory supports the common global descent direction. These constraints operate on different geometric quantities and are therefore not redundant. The former regulates the magnitude of the local geometric response, while the latter regulates the orientation between local and global updates.

Accordingly, FedCADC goes beyond the conventional perturbation-based update paradigm of SAM-based federated optimization and introduces a geometric-response regulation framework based on probing, decomposition, and coordination. The perturbation first probes the local loss geometry, the center-to-perturbed gradient difference then isolates the curvature response, and the global direction constraint subsequently coordinates the local trajectory with the server-side optimization trend. This perspective shows that FedCADC is not merely a combination of a SAM-style perturbation and an additional regularizer. Instead, it reformulates flatness-aware federated optimization as the joint regulation of local geometric sensitivity and cross-client directional compatibility.

Table 5: Number of model-sized transmissions per round.

Algorithm	Upload	Download	Total
FedAvg	1	1	2
FedProx	1	1	2
SCAFFOLD	2	2	4
FedMut	1	1	2
FedSAM	1	1	2
FedASAM	1	1	2
MoFedSAM	1	2	3
FedGAMMA	2	2	4
FedSMOO	2	2	4
FedGF	1	2	3
FedSWA	1	1	2
FedVSSAM	1	2	3
FedGloSS	1	1	2
FedScar	1	2	3
FedCADC	1	2	**3**

E Additional Experimental Results

E.1 Communication Cost Analysis

We measure communication cost by the number of model-sized parameter vectors transmitted per participating client in each round. One unit corresponds to one transmission of a model-sized parameter vector, while scalar control signals and negligible metadata are excluded. As shown in Table 5, FedAvg, FedMut, FedSAM, and FedASAM require two transmissions. MoFedSAM and FedGF download an additional guidance vector, increasing the cost to three, whereas FedGAMMA and FedSMOO exchange auxiliary information in both directions and require four. FedCADC uploads only the local model but downloads both the global model w^t and the global update direction g^t. Since g^t has the same dimension as w^t, FedCADC requires three transmissions, incurring one additional download over FedAvg while remaining cheaper than FedGAMMA and FedSMOO.

Table 6 evaluates communication efficiency using the first round at which each method reaches FedAvg's final 50-round average accuracy. Across both datasets, FedCADC reaches this target within 12–95 rounds, compared with 41–368 rounds for FedAvg, reducing the required communication rounds by 58.5%–91.6%. It attains the minimum round count in 14 of the 18 settings, including one tie. In particular, FedCADC performs best in all settings with $\alpha_{\mathrm{Dir}} \in \{0.5, 10\}$. Under the strongest heterogeneity with $\alpha_{\mathrm{Dir}} = 0.1$, FedScar, FedSWA, FedSMOO, or FedGloSS converges faster in several settings. Nevertheless, FedCADC requires only 31 rounds on CIFAR-100 with 20 participating clients, achieving the largest reduction of 91.6% relative to FedAvg. These results demonstrate that FedCADC generally improves time-to-target efficiency across different datasets, heterogeneity levels, and participation settings.

E.2 Convergence Behavior Curves under All Dirichlet Distribution Settings for CIFAR-10 and CIFAR-100

Figures 5 and 6 present the locally smoothed test-accuracy curves under all combinations of participating clients and Dirichlet parameters. The rows correspond to 5, 10, and 20 participating clients, while the columns correspond to $\alpha_{\mathrm{Dir}} \in \{0.1, 0.5, 10\}$. Under the highly heterogeneous setting $\alpha_{\mathrm{Dir}} = 0.1$, all methods exhibit larger performance gaps and stronger fluctuations, particularly with only five participating clients. Nevertheless, FedCADC maintains a competitive convergence trajectory on both datasets and reaches a favorable final accuracy in most settings. As α_{Dir} increases to 0.5 and 10, FedCADC converges rapidly and remains close to, or above, the leading baselines during the later training stage.

Increasing the number of participating clients generally improves the stability of the optimization process. This effect is especially visible on CIFAR-100, where FedCADC maintains a clear advantage in most settings with 10 or 20 participating clients. The results suggest that the curvature-aware

Table 6: Communication rounds required by each algorithm to reach FedAvg's final 50-round average test accuracy on CIFAR-10 and CIFAR-100 under different Dirichlet distribution parameters and client participation settings.

Task	Algorithm	Dirichlet distribution parameter α_{Dir}								
		$\alpha_{\mathrm{Dir}} = 0.1$			$\alpha_{\mathrm{Dir}} = 0.5$			$\alpha_{\mathrm{Dir}} = 10$		
		Participating clients per round								
		5	10	20	5	10	20	5	10	20
CIFAR-10	FedAvg	161	52	41	174	239	229	356	280	311
	FedProx	149	64	33	141	229	125	311	295	306
	SCAFFOLD	140	109	140	204	144	118	–	–	–
	FedMut	179	142	90	84	106	87	68	63	57
	FedSAM	81	26	18	79	74	56	126	90	69
	FedASAM	181	95	83	137	101	93	114	79	71
	MoFedSAM	92	54	46	152	144	119	273	215	161
	FedGAMMA	142	139	171	163	165	127	222	185	152
	FedSMOO	55	24	**11**	82	45	31	134	55	34
	FedGF	80	26	24	104	125	77	158	107	107
	FedSWA	101	38	21	61	65	62	269	182	151
	FedVSSAM	96	32	30	108	104	80	168	123	106
	FedGloSS	69	19	**11**	44	48	37	96	72	65
	FedScar	**39**	**12**	**11**	48	31	28	84	54	38
	FedCADC	59	**12**	17	**34**	**24**	**23**	**59**	**42**	**27**
CIFAR-100	FedAvg	278	304	368	341	356	322	340	342	234
	FedProx	243	208	243	221	477	249	299	369	198
	SCAFFOLD	312	289	217	265	256	254	314	281	260
	FedMut	183	122	85	119	147	87	106	139	109
	FedSAM	149	114	89	168	107	99	149	125	82
	FedASAM	137	89	76	106	83	59	101	80	61
	MoFedSAM	331	260	257	376	355	324	431	391	338
	FedGAMMA	241	248	268	324	332	321	381	369	341
	FedSMOO	92	61	48	95	69	54	128	80	48
	FedGF	196	143	117	229	139	118	213	148	114
	FedSWA	169	**53**	109	186	143	102	153	118	92
	FedVSSAM	266	215	189	265	227	205	284	255	212
	FedGloSS	108	89	94	145	144	109	211	182	127
	FedScar	**70**	58	48	92	62	53	98	81	60
	FedCADC	76	68	**31**	**80**	**61**	**39**	**95**	**73**	**42**

Bold values denote the minimum in each column, including ties. An en dash indicates that the threshold was not reached within 500 communication rounds.

and direction-consistency regularizers jointly suppress unstable local updates and better align client optimization with the global training direction.

E.3 Standard Deviation of Test Accuracy for Table 1 of the Main Paper

We report the standard deviation of the final test accuracy across three independent runs, where the final accuracy of each run is computed as the average over the last 50 communication rounds. The results are summarized in Table 7. In general, the standard deviation increases when fewer clients participate or when the Dirichlet distribution becomes more heterogeneous, indicating that limited client participation and severe data heterogeneity make the optimization process less stable. This trend is more pronounced on CIFAR-100 because of its more complex classification task. FedCADC maintains relatively small deviations across most configurations, especially when 10 or 20 clients participate per round. This stability is consistent with its curvature-aware and direction-consistency regularizers, which suppress unstable local responses and reduce deviations from the global optimization direction. Overall, the results indicate that FedCADC provides competitive robustness across a broad range of FL settings.

E.4 Comparison of Additional Flatness Metrics

Tables 8 and 10 report the flatness metrics with 5 and 20 participating clients per round, respectively. Across the two tables, FedCADC obtains the best result in 29 of the 36 metric–setting combinations. With 5 clients, it ranks first in 13 of 18 combinations. For example, on CIFAR-10 with $\alpha_{\mathrm{Dir}} = 10$,

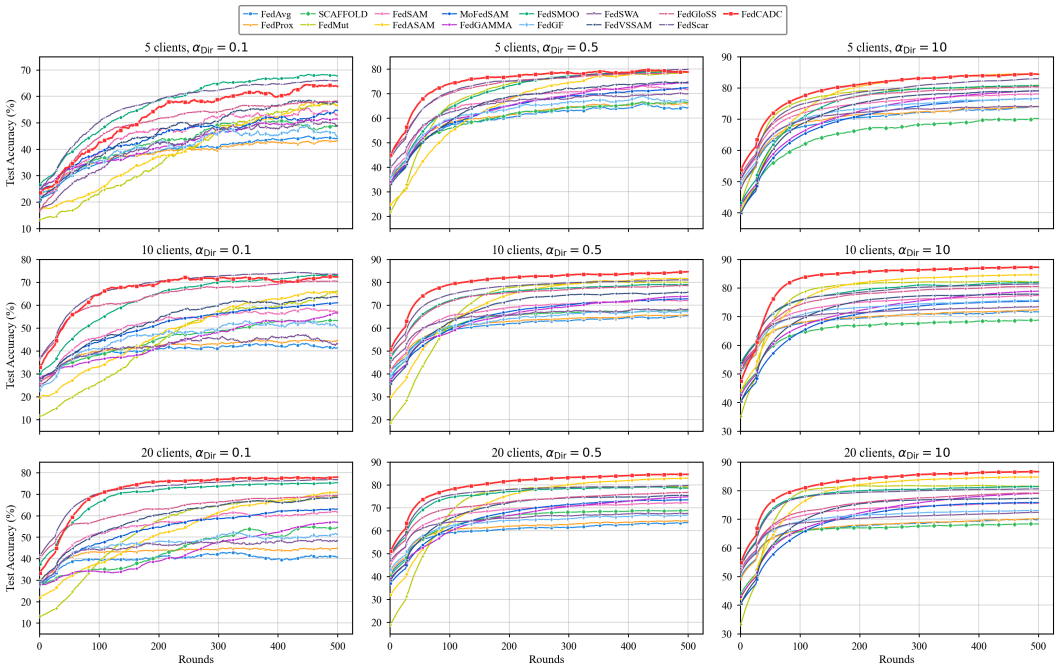


Figure 5: Smoothed test-accuracy convergence curves on CIFAR-10 under all Dirichlet distribution settings $\alpha_{\mathrm{Dir}} \in \{0.1, 0.5, 10\}$ and all client participation settings $\{5, 10, 20\}$. Each curve is locally smoothed using a window size of 50 communication rounds.

FedCADC achieves a MALV of 0.035, $\lambda_{\max}$ of 42.826, and Δ_F of 0.045. These values reduce the corresponding metrics of the strongest competing methods by 38.6%, 25.4%, and 47.1%, respectively. On CIFAR-100, its $\lambda_{\max}$ values of 132.835 and 121.643 under $\alpha_{\mathrm{Dir}} = 0.5$ and 10 represent reductions of 27.8% and 35.5% over the best baselines.

The advantage becomes more consistent with 20 participating clients, where FedCADC achieves 16 of the 18 best results, including the lowest $\lambda_{\max}$ in all six settings and the lowest Δ_F in five. On CIFAR-10 with $\alpha_{\mathrm{Dir}} = 0.5$, it reduces $\lambda_{\max}$ from 62.795 to 26.694 and Δ_F from 0.081 to 0.010, corresponding to reductions of 57.5% and 87.7%. On CIFAR-100 with $\alpha_{\mathrm{Dir}} = 10$, FedCADC decreases MALV, $\lambda_{\max}$, and Δ_F by 46.6%, 51.7%, and 20.0% relative to their strongest baselines. Nevertheless, FedSMOO obtains a lower MALV on CIFAR-10 with $\alpha_{\mathrm{Dir}} = 10$ and a lower Δ_F on CIFAR-100 with $\alpha_{\mathrm{Dir}} = 0.1$. Overall, the substantial reductions in curvature and local–global discrepancy are consistent with the intended roles of the curvature-aware and direction-consistency constraints. The stronger results with 20 clients also suggest that broader participation provides a more representative global direction for coordinating local flatness seeking.

E.5 Loss Landscape Visualization of FedCADC

To visualize the local geometry around the global model, we follow the filter-normalized perturbation procedure of Li et al. (2018). Given a global model with parameters w, we independently sample two random directions d_1 and d_2. For each parameter tensor, the corresponding direction is normalized according to the norm of the parameter tensor, while bias and batch normalization parameters are excluded from the perturbation. The perturbed model is defined as

$$w(\alpha, \beta) = w + \alpha d_1 + \beta d_2, \tag{205}$$

where (α, β) are scalar coordinates in the perturbation plane. We evaluate the training cross-entropy loss $\mathcal{L}_{\mathrm{train}}(w(\alpha, \beta))$ on a two-dimensional grid to obtain the loss surface. In our implementation, the grid contains 51×51 points, with $\alpha, \beta \in [-1, 1]$. The same perturbation directions are used for all points of a given model, and the batch-normalization statistics are kept fixed during evaluation.

We apply this procedure to the global models at different communication rounds and visualize the resulting surfaces using both two-dimensional contour maps and three-dimensional plots. The visu-

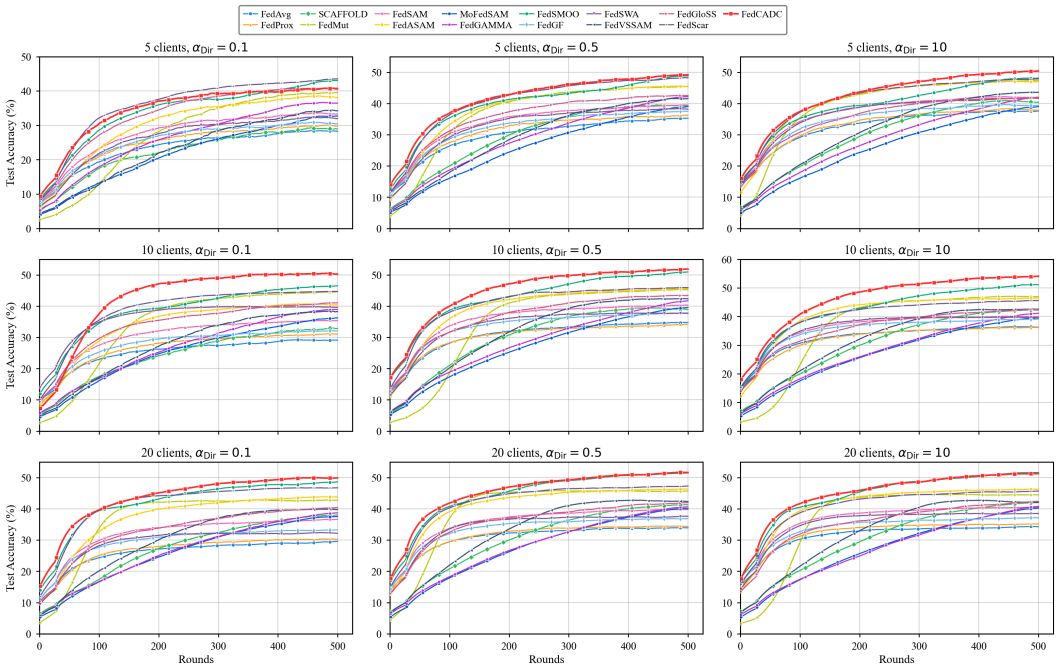


Figure 6: Smoothed test-accuracy convergence curves on CIFAR-100 under all Dirichlet distribution settings $\alpha_{\text{Dir}} \in \{0.1, 0.5, 10\}$ and all client participation settings $\{5, 10, 20\}$. Each curve is locally smoothed using a window size of 50 communication rounds.

alization is conducted only for geometric analysis and does not affect model training or parameter aggregation. Following Li et al. (2018), flatter landscapes are interpreted as exhibiting smaller loss variation within the corresponding parameter neighborhood.

Figures 7, 8, and 9 visualize the loss landscapes of the global FedCADC models at different communication rounds. Across all three Dirichlet settings, the landscape evolves from a narrow and highly irregular basin at Round 1 to a broader and smoother low-loss region as training proceeds. This transition is particularly clear at $\alpha_{\text{Dir}} = 0.5$ and 10, where a relatively broad basin has already emerged by Round 100 and becomes more stable at Rounds 250 and 500. Under the more heterogeneous setting $\alpha_{\text{Dir}} = 0.1$, the early landscapes exhibit stronger asymmetry and sharper boundary variations, but the final global model still converges to a substantially wider low-loss basin.

These observations suggest that FedCADC can rapidly move toward a flat region and further stabilize the global model during later training. The curvature-aware regularizer suppresses sharp loss responses around the current model, while the direction-consistency regularizer reduces deviation between local trajectories and the global optimization trend. Their adaptive weighting enables stronger correction when the local geometry or update direction requires it, thereby helping the aggregated model avoid sharp or geometrically inconsistent regions.

E.6 Adaptive Weight Dynamics in FedCADC

Curvature-aware weight. Figure 10 shows that α_{curv} generally increases during training on both datasets, indicating that perturbation-induced geometric changes become more pronounced as optimization approaches the final solution. Its magnitude also tends to increase with the number of participating clients. In most settings, the weights follow $\alpha_{\text{Dir}} = 10 > 0.5 > 0.1$, although the CIFAR-10 setting with five clients and $\alpha_{\text{Dir}} = 0.1$ remains nearly stationary. According to Eq. equation 13, the weight is jointly determined by the directional rotation and relative norm change between the center and perturbed gradients. Therefore, these differences reflect the local gradient geometry encountered during training rather than the Dirichlet parameter alone. This observation is consistent with Proposition 2, which relates the curvature correction to the perturbation-induced gradient difference.

Table 7: Standard deviations of test accuracy.

Task	Algorithm	Dirichlet distribution parameter α_{Dir}								
		$\alpha_{\mathrm{Dir}} = 0.1$			$\alpha_{\mathrm{Dir}} = 0.5$			$\alpha_{\mathrm{Dir}} = 10$		
		Number of participating clients per round								
		5	10	20	5	10	20	5	10	20
CIFAR-10	FedCADC	1.24	0.76	0.55	0.62	**0.43**	**0.34**	**0.35**	**0.24**	**0.20**
	FedAvg	4.35	3.12	2.10	3.22	2.41	1.82	1.12	0.82	0.61
	FedProx	4.10	2.95	1.96	3.05	2.20	1.70	1.05	0.78	0.59
	SCAFFOLD	3.55	2.50	1.62	2.72	1.90	1.28	0.82	0.57	0.44
	FedMut	2.75	2.08	1.42	2.18	1.57	1.13	0.76	0.55	0.42
	FedSAM	1.35	0.88	0.64	0.58	0.46	0.40	0.41	0.30	0.26
	FedASAM	1.30	0.84	0.61	**0.55**	0.44	0.36	0.40	0.29	0.25
	MoFedSAM	1.50	0.96	0.70	0.72	0.52	0.42	0.43	0.30	0.24
	FedGAMMA	2.45	1.80	1.18	1.92	1.38	0.95	0.62	0.45	0.34
	FedSMOO	1.18	0.79	0.58	0.67	0.49	0.42	0.40	0.28	0.23
	FedGF	2.10	1.55	1.05	1.70	1.20	0.84	0.55	0.40	0.30
	FedSWA	1.30	0.81	0.57	0.69	0.49	0.41	0.40	0.35	0.28
	FedVSSAM	1.14	0.70	0.49	0.66	0.47	0.39	0.38	0.27	0.23
	FedGloSS	1.08	0.67	0.47	0.65	0.46	0.38	0.37	0.30	0.26
	FedScar	**0.96**	**0.63**	**0.45**	0.64	0.45	0.37	0.36	0.25	0.21
CIFAR-100	FedCADC	1.72	**0.99**	**0.76**	0.98	**0.72**	**0.58**	**0.62**	**0.45**	**0.34**
	FedAvg	2.95	2.20	1.55	2.40	1.72	1.25	1.45	1.05	0.78
	FedProx	2.75	2.05	1.48	2.20	1.60	1.18	1.32	0.98	0.74
	SCAFFOLD	2.50	1.85	1.30	1.98	1.43	1.02	1.10	0.82	0.63
	FedMut	2.20	1.66	1.18	1.80	1.30	0.95	1.00	0.74	0.56
	FedSAM	1.95	1.50	1.12	**0.95**	0.78	0.65	0.68	0.51	0.40
	FedASAM	1.84	1.38	1.04	1.02	0.80	0.62	0.67	0.50	0.39
	MoFedSAM	2.40	1.82	1.28	1.95	1.42	1.05	1.18	0.88	0.67
	FedGAMMA	1.88	1.40	0.98	1.55	1.10	0.82	0.86	0.63	0.47
	FedSMOO	1.60	1.30	0.98	0.99	0.82	0.60	0.69	0.52	0.41
	FedGF	1.68	1.27	0.89	1.34	0.96	0.70	0.75	0.55	0.41
	FedSWA	1.79	1.31	0.97	1.07	0.87	0.65	0.67	0.51	0.36
	FedVSSAM	1.55	1.14	0.84	1.03	0.86	0.64	0.64	0.49	0.38
	FedGloSS	1.48	1.09	0.80	1.01	0.84	0.62	0.65	0.48	0.37
	FedScar	**1.35**	1.24	0.92	1.00	0.83	0.61	0.64	0.46	0.36

Bold values indicate the lowest standard deviation in each column.

Direction-consistency weight. As shown in Figure 11, α_{dir} adjusts rapidly during the early rounds and then remains within a bounded range. On CIFAR-10 with five clients, the highly heterogeneous setting $\alpha_{\mathrm{Dir}} = 0.1$ produces the largest and most variable weight, suggesting stronger disagreement between local trajectories and the historical global direction. With more participating clients, this curve decreases, whereas the weights for $\alpha_{\mathrm{Dir}} \in \{0.5, 10\}$ gradually increase. On CIFAR-100, the three weights mostly stabilize around 0.45–0.50, with smoother variations under larger participation. By Eq. equation 15, a larger α_{dir} corresponds to a larger angular deviation from the global direction. These dynamics are consistent with Proposition 3; under the sufficient condition of Theorem 1, the resulting correction can contract client-update deviation. Together, the two adaptive mechanisms control the curvature and drift terms appearing in Theorem 4, providing a theoretical explanation for the improved global flatness observed for FedCADC.

E.7 Dynamics of Curvature Penalties

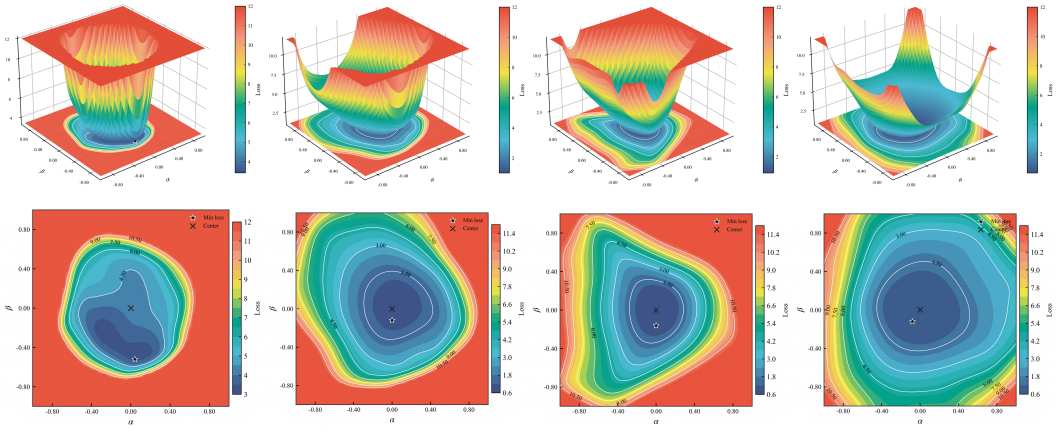
Figure 12 records the average perturbation-normalized loss increment in each communication round. Specifically, the logged curvature penalty is

$$\overline{R}^{t}_{\mathrm{curv}} = \frac{1}{|\mathcal{S}_t|} \sum_{i \in \mathcal{S}_t} \frac{1}{N_i^t} \sum_{k=0}^{N_i^t - 1} \frac{F_{\xi_i^{t,k}}(w_i^{t,k} + \epsilon_i^{t,k}) - F_{\xi_i^{t,k}}(w_i^{t,k})}{\rho^2}, \tag{206}$$

Table 8: Flatness metrics of different FL algorithms on CIFAR-10 and CIFAR-100 with 5 participating clients per round.

Task	Algorithm	Dirichlet parameter α_{Dir}								
		$\alpha_{\text{Dir}} = 0.1$			$\alpha_{\text{Dir}} = 0.5$			$\alpha_{\text{Dir}} = 10$		
		Flatness metrics								
		MALV ↓	$\lambda_{\max}$ ↓	Δ_F ↓	MALV ↓	$\lambda_{\max}$ ↓	Δ_F ↓	MALV ↓	$\lambda_{\max}$ ↓	Δ_F ↓
CIFAR-10	FedSAM	0.646	204.538	0.320	0.177	183.072	0.211	0.086	125.041	0.224
	FedASAM	0.594	130.304	0.280	0.125	**58.171**	0.123	0.057	57.382	0.155
	MoFedSAM	0.528	251.660	0.447	0.283	175.213	0.203	0.268	110.652	0.204
	FedGAMMA	0.972	1435.870	0.659	0.281	480.764	0.305	0.165	371.658	0.085
	FedSMOO	0.916	**103.950**	**0.136**	0.279	493.782	0.107	0.131	372.315	0.095
	FedGF	**0.464**	187.310	0.468	0.132	227.549	0.255	0.103	274.673	0.248
	FedSWA	0.578	159.912	0.247	0.391	123.932	0.162	0.182	74.353	0.173
	FedVSSAM	0.656	696.496	0.504	0.298	880.056	0.234	0.216	526.312	0.152
	FedGloSS	0.564	351.568	0.486	0.122	310.847	0.163	0.056	178.939	0.144
	FedScar	0.779	942.259	0.739	**0.114**	290.623	0.116	0.037	153.285	0.052
	FedCADC	0.548	106.043	0.215	0.124	59.717	**0.093**	**0.035**	**42.826**	**0.045**
CIFAR-100	FedSAM	0.430	370.194	0.443	0.304	353.787	0.203	0.299	322.981	0.160
	FedASAM	0.325	190.172	0.224	0.271	183.893	0.168	0.439	188.499	0.219
	MoFedSAM	0.706	1031.548	0.825	0.579	969.692	0.631	0.451	967.465	0.204
	FedGAMMA	0.698	779.355	0.520	0.432	665.746	0.301	0.346	702.880	0.193
	FedSMOO	**0.212**	251.488	0.217	0.366	243.572	0.185	0.469	239.554	0.123
	FedGF	0.513	483.305	0.302	0.442	464.731	0.327	0.613	455.528	0.221
	FedSWA	0.521	496.411	0.504	0.478	384.880	0.489	0.450	322.785	0.394
	FedVSSAM	0.312	1446.476	0.194	0.683	1274.381	0.193	0.664	1502.864	0.690
	FedGloSS	0.394	564.979	0.369	0.388	478.433	0.262	0.483	489.165	0.092
	FedScar	0.281	353.067	**0.132**	0.280	305.034	0.161	0.263	301.097	0.150
	FedCADC	0.372	**169.619**	0.213	**0.194**	**132.835**	**0.121**	**0.185**	**121.643**	**0.089**

The ↓ symbol indicates that lower values are preferred. Bold values denote the minimum value in each metric column.



(a) Round 1 (b) Round 100 (c) Round 250 (d) Round 500

Figure 7: Two-dimensional and three-dimensional loss landscapes of the global FedCADC models on CIFAR-10 with $\alpha_{\text{Dir}} = 0.1$ and 10 participating clients per round at different communication rounds.

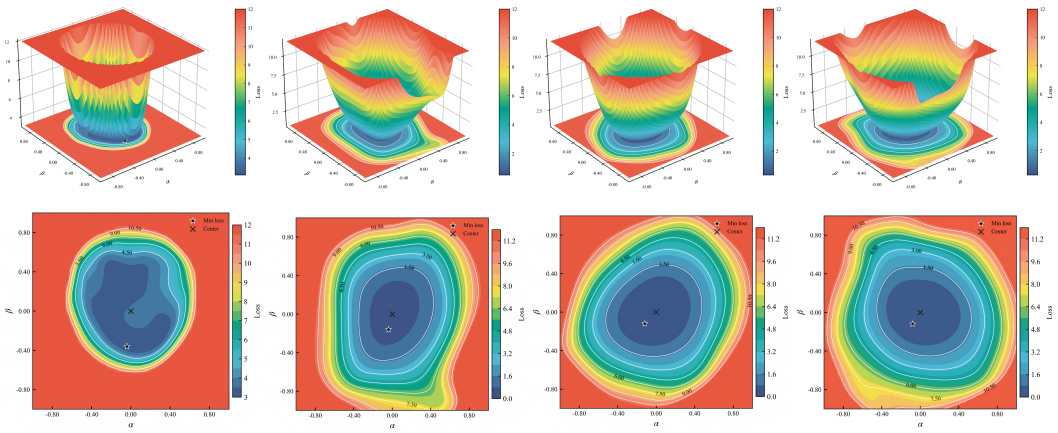
where $\mathcal{S}_t$ is the set of participating clients and N_i^t is the number of local mini-batch updates of client i. The value is first averaged over the local updates of each client and then across the participating clients. It measures the sensitivity of the local loss to the gradient-aligned perturbation, rather than the adaptive coefficient α_{curv} or the weighted regularization term. During optimization, its positive part is used as defined in equation 12.

Across all configurations, the curvature penalty decreases rapidly during the early rounds and subsequently approaches a relatively stable level. Increasing the number of participating clients generally produces smoother trajectories and lower late-stage penalties, particularly on CIFAR-100. Propo-

Table 9: Flatness metrics of different FL algorithms on CIFAR-10 and CIFAR-100 with 10 participating clients per round.

Task	Algorithm	Dirichlet parameter α_{Dir}								
		$\alpha_{\text{Dir}} = 0.1$			$\alpha_{\text{Dir}} = 0.5$			$\alpha_{\text{Dir}} = 10$		
		Flatness metrics								
		MALV ↓	$\lambda_{\max}$ ↓	Δ_F ↓	MALV ↓	$\lambda_{\max}$ ↓	Δ_F ↓	MALV ↓	$\lambda_{\max}$ ↓	Δ_F ↓
CIFAR-10	FedSAM	0.512	395.554	0.618	0.450	336.908	0.578	0.143	227.832	0.404
	FedASAM	0.284	101.068	0.258	0.104	49.669	0.107	0.041	58.748	0.051
	MoFedSAM	0.334	293.737	0.505	0.130	256.411	0.400	0.297	871.571	0.324
	FedGAMMA	0.492	1342.370	1.125	0.281	572.241	0.349	0.327	363.325	0.232
	FedSMOO	0.527	103.270	0.689	**0.038**	155.386	0.163	**0.034**	82.200	0.058
	FedGF	0.482	257.515	0.443	0.264	378.207	0.301	0.102	272.136	0.179
	FedSWA	0.359	627.418	1.121	0.174	415.484	0.413	0.137	402.534	0.446
	FedVSSAM	0.452	1168.519	0.242	0.249	1116.752	0.254	0.089	556.614	0.142
	FedGloSS	0.290	340.733	0.610	0.096	441.966	0.495	0.071	296.964	0.157
	FedScar	0.231	535.839	0.543	0.082	222.871	0.237	0.068	94.731	0.136
	FedCADC	**0.210**	**93.773**	**0.163**	0.072	**32.895**	**0.019**	0.049	**25.721**	**0.007**
CIFAR-100	FedSAM	0.390	491.347	0.591	0.467	419.867	0.236	0.623	417.907	0.221
	FedASAM	0.249	243.826	0.326	0.316	245.762	0.241	0.387	230.162	0.238
	MoFedSAM	0.501	1120.135	0.427	0.592	858.004	0.500	0.528	880.839	0.366
	FedGAMMA	0.273	746.221	0.871	0.271	550.441	0.398	0.403	599.394	0.036
	FedSMOO	0.195	260.531	0.083	0.444	232.151	0.044	0.289	205.126	0.019
	FedGF	0.432	606.104	0.418	0.602	464.144	0.112	0.677	533.835	0.179
	FedSWA	0.253	480.531	0.515	0.305	392.600	0.541	0.657	386.111	0.155
	FedVSSAM	0.584	1678.215	0.175	0.865	1853.359	0.053	0.632	1757.195	0.535
	FedGloSS	0.469	590.511	1.211	0.351	682.242	0.541	0.428	596.671	0.398
	FedScar	0.464	400.935	0.435	0.381	343.590	0.055	0.494	341.433	0.071
	FedCADC	**0.182**	**130.808**	**0.036**	**0.185**	**87.355**	**0.031**	**0.044**	**50.123**	**0.017**

The ↓ symbol indicates that lower values are preferred. Bold values indicate the lowest value in each metric column for each dataset.



(a) Round 1 (b) Round 100 (c) Round 250 (d) Round 500

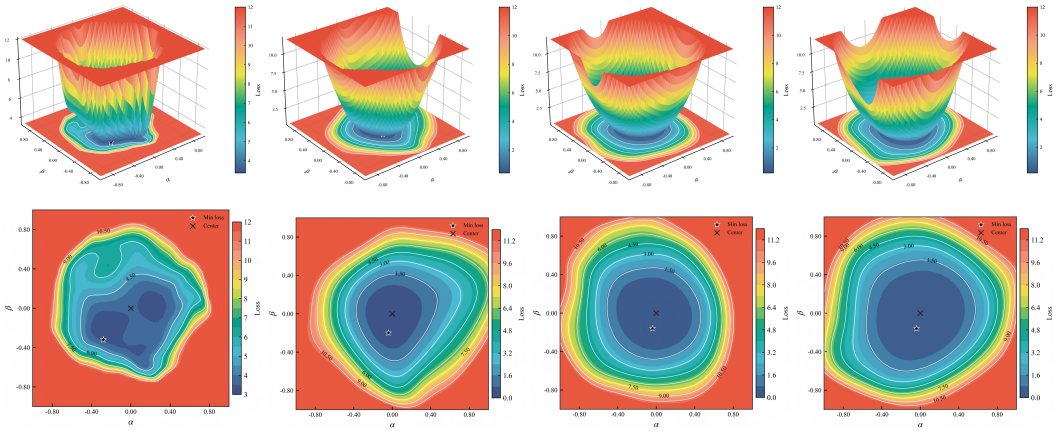
Figure 8: Two-dimensional and three-dimensional loss landscapes of the global FedCADC models on CIFAR-10 with $\alpha_{\text{Dir}} = 0.5$ and 10 participating clients per round at different communication rounds.

sition 1 shows that this penalty contains both a gradient-magnitude term and directional Hessian curvature. Its decrease therefore suggests that FedCADC progressively reduces the local loss response along the current perturbation direction as optimization approaches a stationary and locally flatter region. Moreover, Proposition 2 relates the corresponding correction to the gradient difference before and after perturbation. The large early penalties thus activate stronger geometric correction, whereas their later stabilization indicates reduced correction demand. Consistent with equation 59, the decreasing penalty also reduces a controllable component of the global sharpness bound, provided that the coverage and model-drift terms do not increase.

Table 10: Flatness metrics of different FL algorithms on CIFAR-10 and CIFAR-100 with 20 participating clients per round.

Task	Algorithm	Dirichlet parameter α_{Dir}								
		$\alpha_{\text{Dir}} = 0.1$			$\alpha_{\text{Dir}} = 0.5$			$\alpha_{\text{Dir}} = 10$		
		Flatness metrics								
		MALV ↓	$\lambda_{\max}$ ↓	Δ_F ↓	MALV ↓	$\lambda_{\max}$ ↓	Δ_F ↓	MALV ↓	$\lambda_{\max}$ ↓	Δ_F ↓
CIFAR-10	FedSAM	0.355	240.090	0.348	0.137	348.734	0.269	0.063	249.895	0.370
	FedASAM	0.194	64.603	0.085	0.157	62.795	0.081	0.054	57.752	0.059
	MoFedSAM	1.247	981.149	0.648	0.261	908.000	0.208	0.301	790.269	0.176
	FedGAMMA	0.405	932.143	1.093	0.247	425.682	0.143	0.162	362.203	0.214
	FedSMOO	0.253	224.770	0.167	0.082	137.621	0.338	**0.016**	49.011	0.143
	FedGF	0.273	429.495	0.442	0.191	488.155	0.116	0.118	288.908	0.020
	FedSWA	0.294	811.518	0.753	0.182	550.847	0.377	0.091	307.494	0.774
	FedVSSAM	0.477	1558.562	0.806	0.323	1187.529	0.223	0.189	612.953	0.093
	FedGloSS	0.391	405.400	0.624	0.155	495.178	0.544	0.127	483.369	0.260
	FedScar	0.182	375.770	0.218	0.169	207.338	0.516	0.134	50.669	0.523
	FedCADC	**0.168**	**61.741**	**0.068**	**0.079**	**26.694**	**0.010**	0.023	**15.914**	**0.005**
CIFAR-100	FedSAM	0.445	447.352	0.134	0.475	425.224	0.318	0.629	409.917	0.310
	FedASAM	0.233	215.520	0.169	0.225	208.321	0.158	0.337	201.635	0.221
	MoFedSAM	0.722	1075.546	0.269	0.752	890.017	0.610	0.632	812.709	0.517
	FedGAMMA	0.382	652.240	0.719	0.330	533.066	0.617	0.330	550.879	0.189
	FedSMOO	0.383	240.762	**0.028**	0.343	192.146	0.125	0.305	171.983	0.127
	FedGF	0.484	574.449	0.104	0.447	515.973	0.160	0.665	539.911	0.105
	FedSWA	0.531	458.195	0.547	0.625	537.094	0.694	0.450	485.277	0.752
	FedVSSAM	0.451	1614.634	0.180	0.742	2088.210	0.496	0.775	1946.725	0.569
	FedGloSS	0.366	512.498	0.213	0.579	782.446	0.890	0.455	818.762	0.225
	FedScar	0.527	376.529	0.790	0.340	375.960	0.342	0.634	400.717	0.245
	FedCADC	**0.196**	**107.557**	0.153	**0.213**	**96.453**	**0.124**	**0.163**	**83.136**	**0.084**

The ↓ symbol indicates that lower values are preferred. Bold values indicate the lowest value in each metric column for each dataset.



(a) Round 1 (b) Round 100 (c) Round 250 (d) Round 500

Figure 9: Two-dimensional and three-dimensional loss landscapes of the global FedCADC models on CIFAR-10 with $\alpha_{\text{Dir}} = 10$ and 10 participating clients per round at different communication rounds.

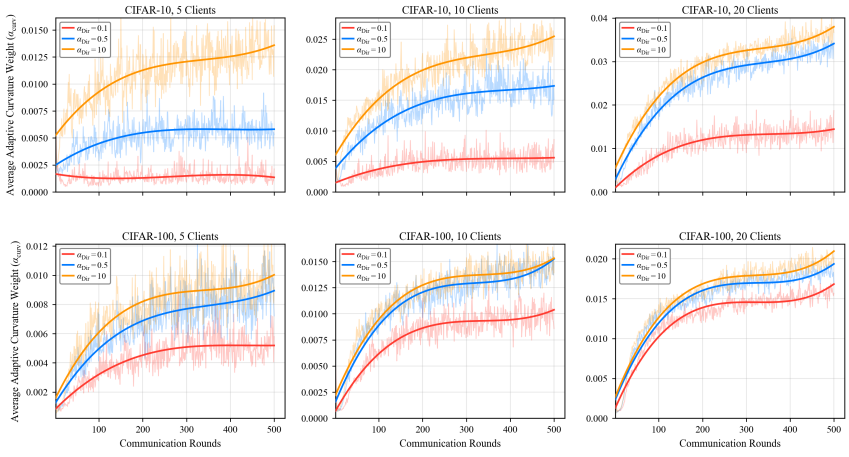


Figure 10: Round-wise evolution of the adaptive curvature-aware weight α_{curv} in FedCADC. Light curves show the raw weights, and solid curves show the fitted trends.

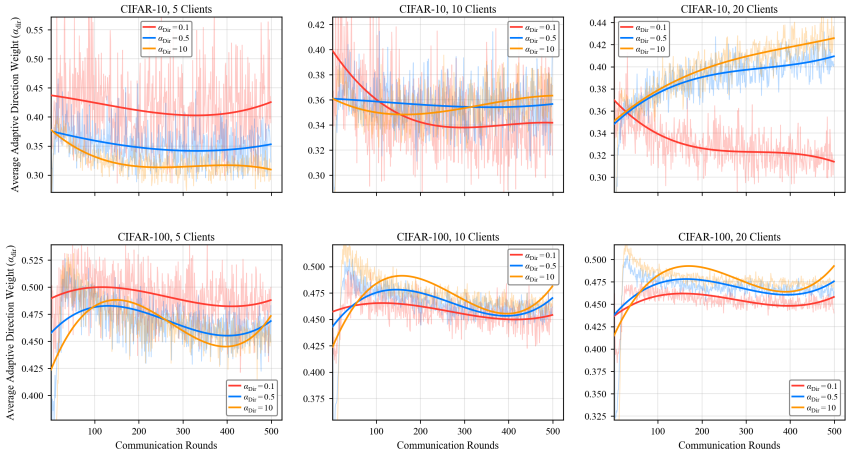


Figure 11: Round-wise evolution of the adaptive direction-consistency weight α_{dir} in FedCADC. Light curves show the raw weights, and solid curves show the fitted trends.

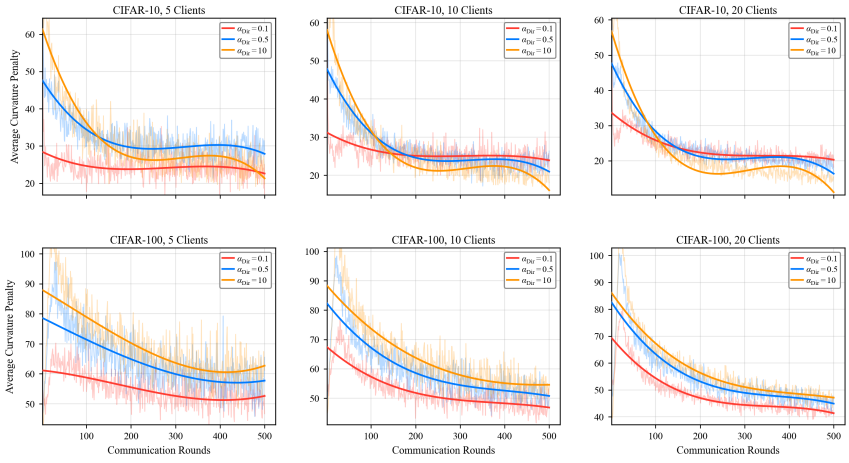


Figure 12: Evolution of the average curvature penalty of FedCADC on CIFAR-10 and CIFAR-100 under different Dirichlet parameters and numbers of participating clients. The transparent curves show the raw values, while the solid curves represent the fitted trends.