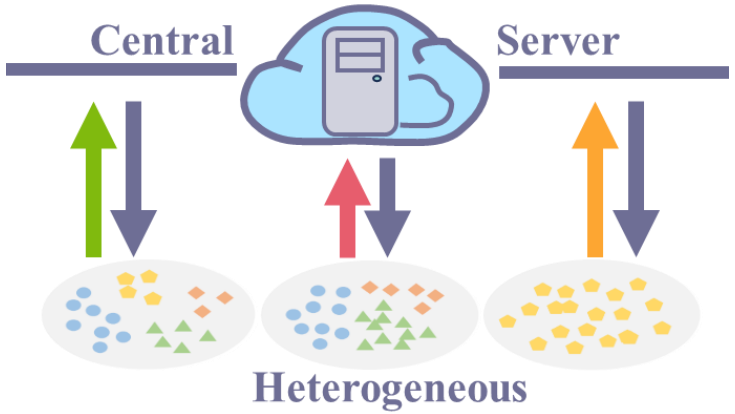
FedWMSAM: Fast and Flat Federated Learning via Weighted Momentum and Sharpness-Aware Minimization

Tianle Li[1*], **Yongzhi Huang[2*],**
Linshan Jiang[3], Qipeng Xie[2], Chang Liu[4], Wenfeng Du[1], Lu Wang[1]
and Kaishun Wu[2]

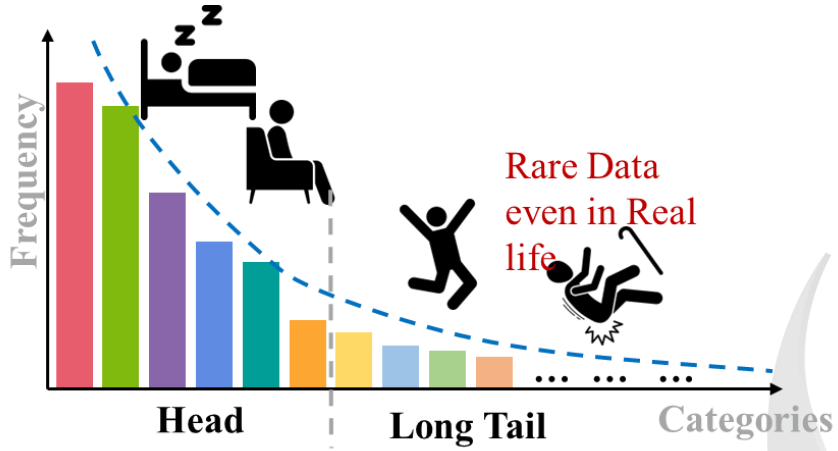
[1]Shenzhen University
[2]The Hong Kong University of Science and Technology (Guangzhou)
[3]National University of Singapore
[4]Nanyang Technological University
* Equal Contribution

Background: Heterogeneous clients & long-tailed data



Heterogeneous clients (non-IID).

- Different clients see different data.



Long-tailed distribution.

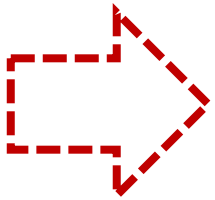
- A few head classes dominate, and many tail classes are rare but essential.

So the Clients learn **conflicting directions** and tail classes **underfit** or vanish.

Background: What these data do to classic FL

Brittle global model.

- Small perturbations tip the model off a narrow ridge.



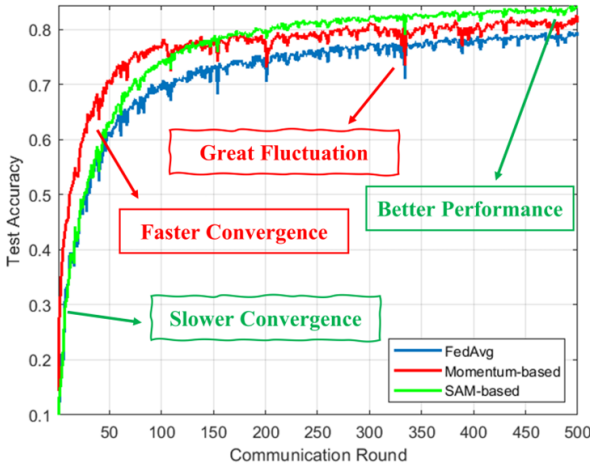


Figure 3: Combine Momentum or SAM with FL.

Bumpy training.

- Inconsistent client updates cause oscillation.

Cause two failure modes.

- Local–global misalignment.
- Momentum echo.

Why failure #1 (misalignment)

Why do these failures happen?

(1) Local–global misalignment of SAM perturbations

- Classic FL computes the perturbation on local data, then updates at $(w + \delta_k)$.
- **Global target is:** $\min_w F(w) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(w)$
- **While SAM objective**: $\min_w F_{SAM}(w) = \min_w \max_{\|\delta\|_2 \le \rho} \mathbb{E}[L(w + \delta) - L(w)]$, and uses a local proxy $\delta_k = \rho \frac{\nabla F_k(w)}{\|\nabla F_k(w)\|}$.
- Under heterogeneity, $\nabla F_k(w) \not\parallel \nabla F(w) \Rightarrow$ clients evaluate gradients at **different** $(w + \delta_k)$, "flattening the wrong places" for the global landscape.

Why failure #2 (momentum echo)

Why do these failures happen?

(2) Momentum echo under inconsistent client directions

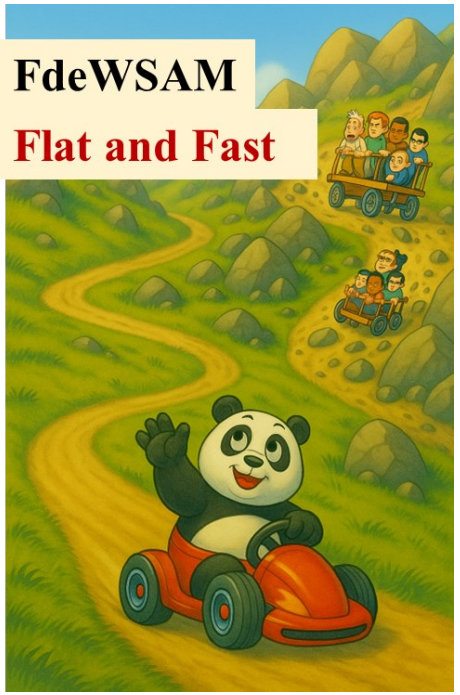
- Server momentum accumulates past updates

$$\Delta_{r+1} = \frac{1}{\eta_l \mid P_r \mid} \sum_{k \in P_r} \Delta_r^k \,, x_{r+1} = x_r - \eta_g \, \Delta_{r+1}$$

- When client directions disagree, Δ_r mixes **stale** signals ⇒ overshoot & late-stage oscillations.
- Perturbations **increase variance**

$$\sigma_\rho^2 = \sigma^2 + (L\rho)^2$$

Design Principles



Idea in one line. Use **server momentum** to encode *global geometry*, **personalize** it per client to correct drift, and **self-adapt** momentum and SAM, implemented with **one backprop per local step.**

What's new here (innovations)

C1: Momentum-guided global perturbation with single backprop.
C2: Personalized momentum to correct client drift.
C3: Self-adaptive schedule (auto momentum ↔ SAM).
C4: Theory for "fast & flat" under heterogeneity.

C1-Momentum-guided global perturbation

- **Align** local updates to the **global landscape** using server-aggregated momentum.
- Keep SAM's flattening effect with **one backprop per local step → FedAvg-like cost**.
- Key update:
- $\delta_{b+1,k}^{r} = \left(x_r + b\,\Delta_r^k\right) - x_{b,k}^{r}$
- $g_{b,k}^{r} = \nabla L\left(x_{b,k}^{r} + \rho\,\delta/\|\,\delta\,\|\right)$
- $x_{b+1,k}^{r} = x_{b,k}^{r} - \eta_l\left[\,\alpha_r g_{b,k}^{r} + (1-\alpha_r)\Delta_r^k\,\right].$

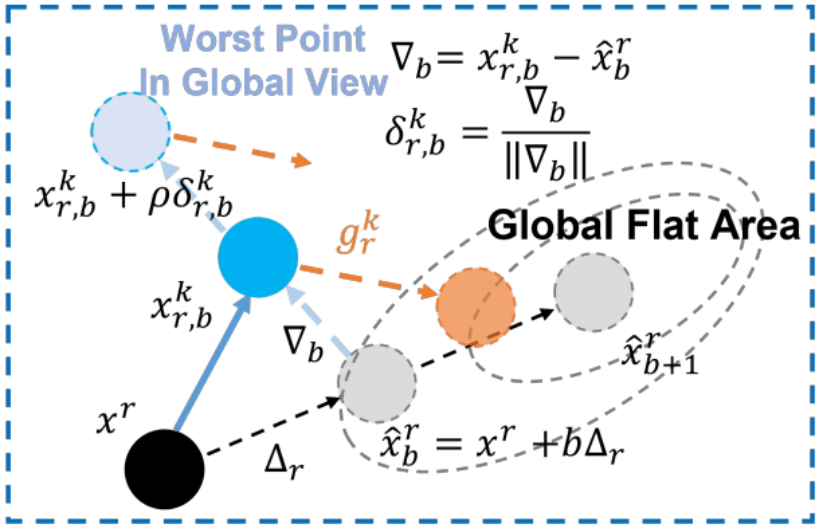


Figure 2 (b): Local model vs. momentum-based model difference guides global perturbation estimation.

C2-Personalized momentum （drift correction）

- **Personalize** the global momentum to each client → **server-equivalent** mixing.
- **Definition** (client-specific momentum).

$$\Delta_r^k = \Delta_r + \frac{\alpha_r}{1-\alpha_r} c_k$$

- **Local velocity** (blend of gradient and momentum).

$$v_{b+1,k} = \alpha_r g_{b,k} + (1-\alpha_r)\Delta_r^k$$

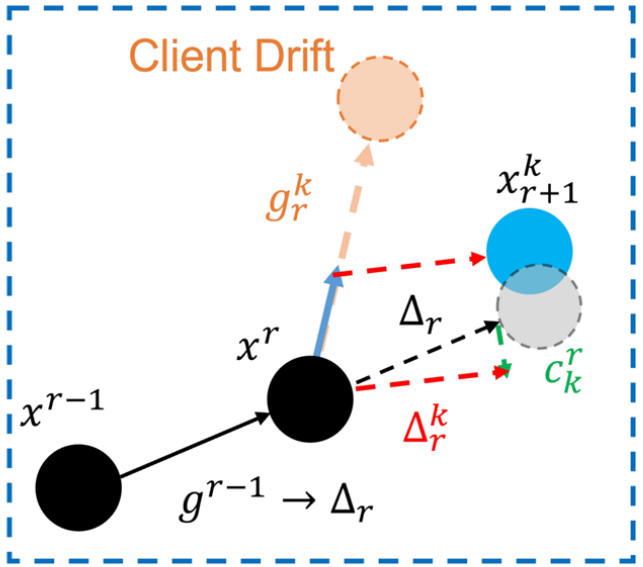


Figure 2 (a): Personalize Momentum will "flatten the wrong hill."

C3-Adaptive training schedule （momentum ↔ SAM）

- **Adapt over time**:
  - **early** momentum for speed → **late** SAM for flatter minima.
- Triggered by direction disagreement and noise rise.
- **Agreement (global ↔ client momenta).**

$$\hat{\alpha}_{r+1} = \frac{1}{|P_r|}\sum_{k\in P_r} \mathrm{sim}\,(\Delta_r, \Delta_r^k), \mathrm{sim}(a, b)$$

$$= \frac{\langle a, b\rangle}{\|a\| \quad \|b\|}$$

- **Smoothed & clipped schedule.**

$$\alpha_{r+1} = (1-\lambda)\,\alpha_r + \lambda\,\mathrm{clip}_{[0.1,\,0.9)}\,(\hat{\alpha}_{r+1})$$

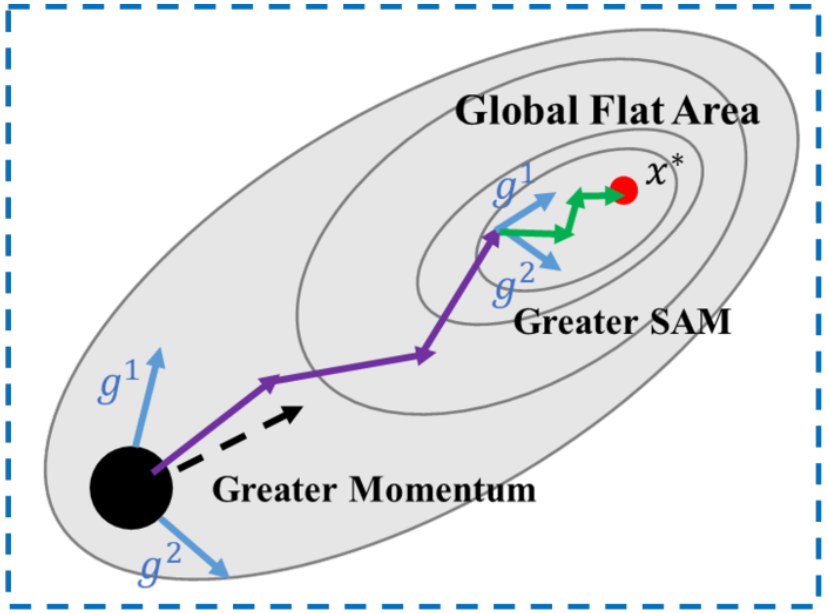


Figure 2 (c): A dynamic weighting adjusts momentum vs. SAM based on gradient–momentum similarity.

Why damping helps（mini-theory）

- **Fact 1**: SAM-style perturbations **increase variance**:

$$\sigma_\rho^2 = \sigma^2 + (L\rho)^2.$$

- **Fact 2**: Our bound **separates noise & heterogeneity**:

$$\frac{1}{R}\sum_r \mathbb{E} \parallel \nabla f(x_r) \parallel^2 \lesssim \frac{L\Delta\, \sigma_\rho^2}{SKR} + \frac{L\Delta}{R}\left(1 + \frac{N^{2/3}}{S}\right).$$

- *Takeaway:* when **noise** or **disagreement** is high → **damp momentum** to avoid oscillation and reach **flatter** minima.

R1: Overall accuracy across datasets

- **Best or 2nd-best** across **3 datasets × 12 heterogeneity settings**.
- Gains grow when **non-IID is stronger** (β=0.1; pathological).

Table 1: Performance comparison of SOTA methods under Dirichlet and Pathological splits after 500 Rounds in different datasets.

Method	**Fashion-MNIST**				**CIFAR-10**				**CIFAR-100**			
#Partition	**Dirichlet**		**Pathological**		**Dirichlet**		**Pathological**		**Dirichlet**		**Pathological**	
#Coefficient	$\beta = 0.6$	$\beta = 0.1$	$\gamma = 6$	$\gamma = 3$	$\beta = 0.6$	$\beta = 0.1$	$\gamma = 6$	$\gamma = 3$	$\beta = 0.6$	$\beta = 0.1$	$\gamma = 20$	$\gamma = 10$
FedAvg	0.8684	0.8226	0.8625	0.8150	0.7886	0.7005	0.7873	0.6426	0.3917	0.3815	0.3968	0.3631
FedCM	0.8283	0.7333	0.8047	0.6630	0.8126	0.7229	0.8167	0.7025	0.4635	0.4290	0.4394	0.3940
SCAFFOLD	0.8789	0.8351	0.8785	0.8311	0.8232	0.7428	0.8179	0.6786	0.4855	0.4437	0.4647	0.4133
FedSAM	0.8683	0.8261	0.8673	0.8045	0.7963	0.6963	0.7908	0.6503	0.4083	0.3790	0.3933	0.3553
MoFedSAM	0.8278	0.7489	0.8141	0.6822	0.8339	0.7386	0.8334	0.7327	0.4859	0.4472	0.4619	0.4279
FedGAMMA	0.8708	0.8298	0.8716	0.8303	0.8292	0.7218	0.8043	0.6105	0.4837	0.4474	0.1739	0.0198
FedSMOO	**0.8846**	0.8337	0.8745	0.8296	**0.8410**	0.7507	0.8382	0.7099	0.3225	0.2987	0.4620	0.3006
FedLESAM(-S/-D)	0.8689	0.8375	0.8732	0.8209	0.8165	0.7284	0.8127	0.6381	0.4260	0.4114	0.4298	0.3914
FedWMSAM (ours)	0.8756	**0.8464**	**0.8805**	**0.8531**	0.8356	**0.7664**	**0.8443**	**0.7446**	**0.4908**	**0.4646**	**0.4786**	**0.4383**

Note 1: We report the best accuracy among FedLESAM, FedLESAM-S, and FedLESAM-D in one row.
Note 2: γ represents the number of classes allocated to each client in the pathological distribution.

R2: Real-world heterogeneity

- **Best in 3/4 domains** (Art, Clipart, Product) and **best average**.
- Slightly **trails SCAFFOLD** on Real-World.
- Takeaway: **align-to-global** bias **transfers across domains**.

Table 2: Accuracy on OfficeHome target domains after 500 rounds (10% sample, 100% active).

Method	Art	Clipart	Product	Real World
FedAvg	0.9909	0.9569	0.9725	0.9633
FedCM	0.9316	0.8013	0.8783	0.8411
SCAFFOLD	0.9934	0.9610	0.9745	**0.9749**
FedSAM	0.9851	0.9402	0.9576	0.9685
MoFedSAM	0.9921	0.9458	0.9653	0.9566
FedGamma	0.9934	0.9557	0.9758	0.9605
FedSMOO	0.9868	0.9563	0.9753	0.9629
FedLESAM	0.9930	0.9626	0.9783	0.9713
FedWMSAM	**0.9942**	**0.9650**	**0.9790**	0.9717

R3: Convergence speed & client compute

- **Fast-then-flat** trajectories & **Fewer rounds** to targets.
- **Client time ≈ 15.03 s**, **much lower** than double-backprop SAM (**26.9–29.8 s**).

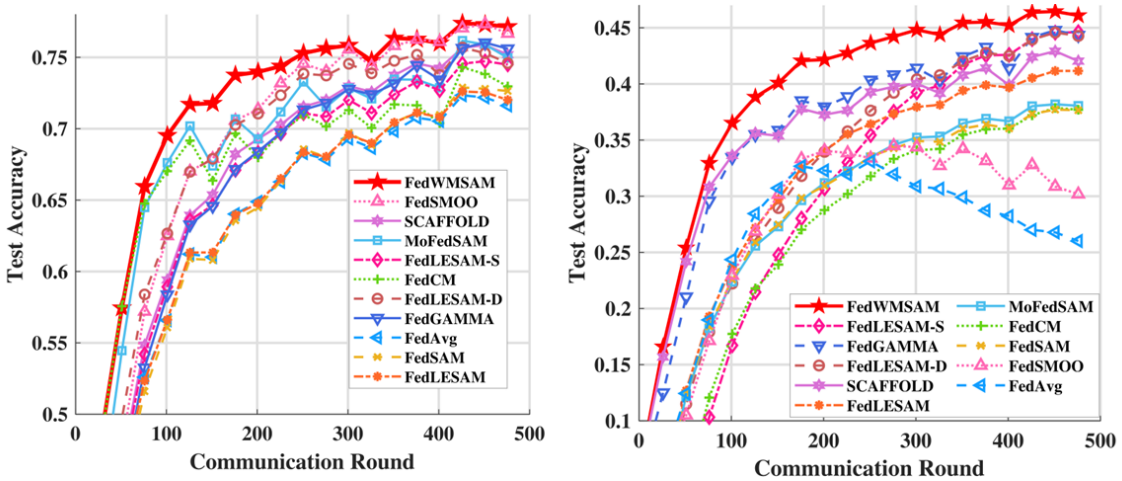


Figure 4: Performance comparison on CIFAR-10/100. FedWMSAM shows fast-then-flat trajectories.

Table 3: Rounds to reach different accuracy levels and client computation time.

Method	0.7	0.72	0.74	0.76	0.78	Time(s)
FedAvg	254	348	432	-	-	**14.57**
FedCM	**97**	132	255	426	-	14.67
SCAFFOLD	189	241	301	376	-	17.44
FedSAM	247	303	403	-	-	26.90
MoFedSAM	**97**	132	169	255	426	29.73
FedGAMMA	208	241	300	374	-	29.72
FedSMOO	134	167	201	255	382	29.77
FedLESAM	241	303	433	-	-	14.66
FedLESAM-D	149	187	211	255	-	16.96
FedLESAM-S	205	247	313	432	-	16.71
FedWMSAM	**97**	**114**	**153**	**241**	**356**	15.03

R4: Sensitivity & scaling

- **Local epochs**: top in most settings; **no clear drop** as epochs increase.
- **Clients**: best or on-par; consistently leads when **≥75**.
- **Sampling rate**: strong at **sparse participation**;

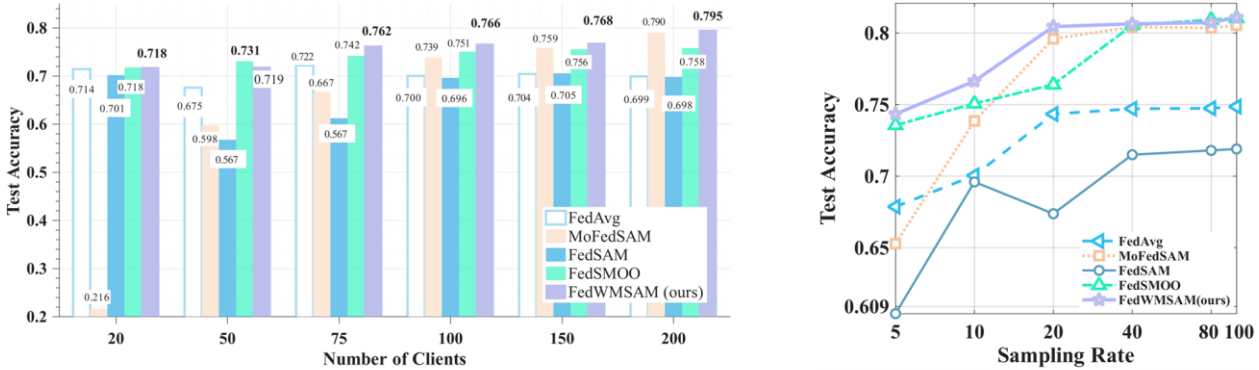


Figure 6: Across different client numbers & sampling rates, FedWMSAM is best or on par across the range.

Table 4: Comparison under different local epochs.

Method	Epoch 1	Epoch 5	Epoch 10	Epoch 20
FedAvg	0.6003	0.7005	0.6988	0.6879
MoFedSAM	0.7237	0.7386	0.6997	0.6776
FedSAM	0.5515	0.6963	0.6862	0.6903
FedSMOO	**0.7888**	0.7507	0.7538	0.7472
FedWMSAM (Ours)	0.7484	**0.7664**	**0.7662**	**0.7515**

R5: Ablations & parameter study

- **Mom** (personalized momentum) gives the **largest single gain** → reduces local–global bias.
- **Adaptive gate** behaves like **data-driven damping**: *early momentum, late SAM.*
- **ρ study**: best near **0.01**; **graceful decay** as ρ↑;

Table 5: Ablation of key modules in FedWMSAM.

Mom.	SAM	Weighted	Acc.	Imp.
✓	✓	✓	0.7664	4.35%
✓	✓	×	0.7556	3.27%
×	✓	✓	0.7265	0.36%
✓	×	✓	0.7326	0.97%
✓	×	×	0.7478	2.49%
×	✓	×	0.7430	2.01%
×	×	×	0.7229	-

Table 6: Ablation study results of ρ.

Method ρ	**0.005**	**0.01**	**0.05**	**0.1**	**0.5**
FedAvg	0.7005	0.7005	0.7005	0.7005	0.7005
MoFedSAM	0.7355	0.7386	0.7102	0.5562	0.1000
FedWMSAM (ours)	**0.7659**	**0.7664**	**0.7563**	**0.7244**	**0.5905**

Thank you!

yhuang849@connect.hkust-gz.edu.cn