ASYMMETRIC EMA FOR FEDERATED LORA: FEED BACK A, NOT B

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated LoRA can suffer from majority-class collapse in the high-heterogeneity regime (Dirichlet $\alpha \leq 0.1$), and whether a run collapses often depends on the random seed. The two LoRA factors multiply, so per-round errors in either factor compound from one round to the next. Applying a time-axis exponential moving average (EMA) over rounds is a natural remedy, but the choice of which factor to feed back is not obvious: feeding a factor back keeps only a $(1-\beta)$ fraction of its per-round displacement, while not feeding it back lets the gap between the EMA and the training iterate accumulate across rounds. We show that both effects are exact server-side identities that hold for any LoRA rank, local optimizer, and number of local steps, and that this common cost acts asymmetrically. Drift is most damaging for the direction factor A, whereas the displacement squeeze is most damaging for the scale factor B. We propose **Stabilized A-Feedback** (SAF), which feeds only the direction factor back through the EMA, keeps the scale factor on the FEDAVG path, and adds one line of code with no new hyperparameters. Under a rank-1, single-step SGD model we quantify four consequences: a residual error that does not vanish as the round budget grows when A-feedback is absent, standard SGD-style convergence of SAF, an optimization-bias penalty when B is fed back that cannot be recovered by retuning the learning rate within the stability constraint, and a heterogeneity crossover below which FEDAVG wins. Across four federated LoRA methods and six tasks (four GLUE and two MedMNIST; 24 five-seed cells), SAF outperforms both FEDAVG and EMA in 22 cells and reduces the MNLI seed standard deviation from up to 23% to at most 8%; ablations that hold the deployed model fixed separate training-time A-feedback from deployment-time averaging. We also report where the method stops: at 355M SAF outperforms FEDAVG on all three seeds but its ordering against EMA is seed-dependent, and stacking with FedSA-LoRA is structurally compatible but not shown to be additive.

1 INTRODUCTION

Large foundation models (Brown et al., 2020; Touvron et al., 2023; Dosovitskiy et al., 2021) are fine-tuned for many downstream tasks. Sensitive data scattered across user devices and institutions cannot be centralized because of privacy, regulatory, and bandwidth constraints, so federated learning (McMahan et al., 2017) is the usual choice. Because these models are large, practitioners use parameter-efficient fine-tuning to control communication and computation costs, and LoRA (Hu et al., 2022) has become the de facto adapter. Research on federated LoRA has explored asymmetric treatment of A and B (Sun et al., 2024; Guo et al., 2025; Yu et al., 2025), aggregation corrections (Singhal et al., 2025; Lee et al., 2025), and asymmetric learning rates (Hayou et al., 2024); a survey appears in Yang et al. (2025). All of these methods operate within a single round. In highly heterogeneous environments the round-by-round dispersion of adapters accumulates over rounds, and FEDAVG converges to the majority class—what we call *heterogeneity collapse* (G. et al., 2024). The collapse is often *stochastic*: the same configuration collapses on some seeds and not on others (MNLI std 18–23%, Table 1). This problem lies in the trajectory across rounds but has rarely been addressed under the LoRA structure. We ask:

How should we apply cross-round smoothing to the two LoRA factors A and B?

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.

Scope. We target controlled label skew with $\alpha \leq 0.1$,[1] which is the headline setting of recent federated-LoRA papers (Sun et al., 2024; Guo et al., 2025; Singhal et al., 2025; Cho et al., 2024; Koo et al., 2025) and lets us vary heterogeneity while holding the dataset, model, and client population fixed. We restrict evaluation to classification because heterogeneity collapse is most clearly defined relative to a majority-class baseline. We do not treat any α as a universal real-world operating point; naturally partitioned multi-device or multi-center data and generative tasks are outside our evidence (§7).

The answer is not obvious. Feeding either factor back through an EMA keeps only a $(1-\beta)$ fraction of that factor's per-round displacement, while not feeding it back lets the gap between the EMA and the training iterate carry over from round to round. Both statements are exact identities of the server update, independent of rank, local optimizer, and number of local steps (Prop. 1). Under the LoRA factorization $\Delta W = BA$ this *common* cost acts asymmetrically. The factor A is a direction term that starts from a random initialization and fixes the subspace in which B acts. It does not require a large learning rate but cannot tolerate drift across rounds, because drift breaks the alignment of BA at deployment. The factor B is a scale term that starts at zero and must accumulate signal magnitude; a squeezed displacement stalls learning. We use "direction" and "scale" operationally, as roles induced by initialization and multiplication rather than basis-invariant properties; the asymmetry vanishes once the BA factorization is removed, which we test directly (App. C.17).

We propose **SAF** (**S**tabilized A-**F**eedback): the only cell of the 2×2 feedback grid satisfying both conditions, with A fed back through the EMA, B kept on the FEDAVG path, and deployment via the EMA-averaged factor endpoint $\bar{B}_T\bar{A}_T$. Our contributions are: (i) we formalize factor-wise cross-round feedback as a design axis orthogonal to round-internal techniques and evaluate all four cells of the grid; (ii) we give exact feedback identities at any rank and for any local optimizer (Prop. 1, Cor. 1), and under a rank-1, single-step SGD model prove $1/(1-\beta)$ bias inflation under B-feedback, SAF convergence, a non-vanishing residual without A-feedback, and a heterogeneity crossover $\sigma^\star$ (§5.2), stating what this layer does not cover; (iii) within $\alpha \leq 0.1$, SAF beats FEDAVG and EMA simultaneously in 22 of 24 five-seed cells (20 of 24 under best-round accuracy) across four federated LoRA methods on four GLUE and two MedMNIST tasks, and ablations with the deployed model held fixed isolate training-time A-feedback from deployment-time averaging; (iv) we report the boundaries (low heterogeneity, HetLoRA, 355M, FedSA-LoRA) and a label-histogram screen, computable before training, for when SAF helps.

2 RELATED WORK

Federated LoRA. The literature on applying LoRA (Hu et al., 2022) to federated learning divides into two broad streams (Yang et al., 2025). The first treats A and B asymmetrically. FFA-LoRA (Sun et al., 2024) freezes A. FedSA-LoRA (Guo et al., 2025) shares and aggregates only A while each client keeps its own B. AltLoRA (Yu et al., 2025), RoLoRA (Chen et al., 2025), and LoRA-A^2 (Koo et al., 2025) alternate training of the two factors. The second stream corrects the aggregation itself. It addresses the inaccuracy of $\bar{B}\,\bar{A} \neq \frac{1}{K}\sum_k B_k A_k$. FedEx-LoRA (Singhal et al., 2025) adds a residual term explicitly to recover the exact mean, and FedSVD (Lee et al., 2025) uses SVD-based orthogonalization to tackle the same issue. FRLoRA (Yan et al., 2025) and SFed-LoRA (Huang et al., 2026) modify this framework through residual accumulation and scale correction, respectively (raFLoRA (Wu et al., 2026)'s "collapse" targets rank-energy concentration, distinct from majority-class collapse here). Including methods that address heterogeneity directly, such as FedProx (Li et al., 2020) and SCAFFOLD (Karimireddy et al., 2020), all of the above operate within a single round; SAF instead decides which server state initializes the next round.

Asymmetry of LoRA A/B and EMA. The asymmetry between the two factors has been reported from several perspectives. LoRA+ (Hayou et al., 2024) analyzes the learning-rate aspect, Zhu et al. (2024a) study representational capacity, iLoRA (Zhu et al., 2024b) examines initialization imbalance, and LoRA-GF (Xu et al., 2025) analyzes gradient flow. All of these analyses are restricted to local learning within a single round. Time-averaging of weights is widely used in centralized training (Polyak & Juditsky, 1992; Izmailov et al., 2018; Morales-Brotons et al., 2024; Busbridge et al., 2023), and adaptive server optimization (FedAdam/FedYogi) (Reddi et al., 2021) also smooths across rounds; these act on the whole model or the aggregated update, whereas

[1] We write α for the Dirichlet concentration α_{Dir} throughout.

SAF acts on a single factor of the LoRA decomposition and feeds the average back into training. SAF borrows the bias-variance intuition of SWA (Izmailov et al., 2018) but we claim no SWA guarantee, since $\bar{B}_T\bar{A}_T \neq \mathrm{EMA}(B_tA_t)$ in general. FedEMA-Distill (Reguieg et al., 2026) applies an EMA to the entire global model and uses it as a distillation target. Our EMA configuration is a controlled full-feedback ablation without the distillation loss, weighting, or temperature; it is *not* an implementation of FedEMA-Distill, and we make no comparison to that method.

Closest related work. Three federated-LoRA methods share the surface mechanism of processing A on the server before the next round but differ in motivation: FedSA-LoRA (Guo et al., 2025) aggregates A selectively within a round while each trained B_k stays client-local, yielding personalized models; FFA-LoRA (Sun et al., 2024) freezes A to block averaging noise; FedSVD (Lee et al., 2025) orthogonalizes A to mitigate DP noise. SAF uses the same channel for *cross-round EMA feedback* and deploys a single global adapter. FedSA-LoRA and SAF are structurally composable; our stacked run composes without failure, but its evaluation is saturated and establishes neither an empirical gain nor superiority in either direction (§6.3, App. C.2). Detailed differentiation in App. B.

3 Problem Analysis

3.1 Preliminaries

We use the standard LoRA setup (Hu et al., 2022): $\Delta W = BA$ with $A \in \mathbb{R}^{r\times d_{\text{in}}}$ random, $B \in \mathbb{R}^{d_{\text{out}}\times r}$ zero-initialized, $r \ll \min(d_{\text{out}}, d_{\text{in}})$. In federated learning (McMahan et al., 2017), at round t the server broadcasts the *input state* $\theta_t^{\text{in}} = (A_t^{\text{in}}, B_t^{\text{in}})$ to a client subset $\mathcal{S}_t$; each client k trains on local data $\mathcal{D}_k$ with any optimizer and any number of local steps, and the server computes the weighted average (the FedAvg *estimate*) and the realized round displacement:

$$\theta_{t+1}^{\text{FA}} = \sum_{k\in\mathcal{S}_t} \frac{n_k}{n}\,\theta_t^{(k)}, \qquad U_t^X := X_{t+1}^{\text{FA}} - X_t^{\text{in}} \quad (X \in \{A, B\}). \tag{1}$$

U_t^X absorbs everything that happens inside the round: local steps, the optimizer, client heterogeneity, and the coupled BA dynamics. Borrowing the time-axis average widely used in centralized training (Polyak & Juditsky, 1992; Izmailov et al., 2018; Busbridge et al., 2023; Morales-Brotons et al., 2024), we introduce cross-round smoothing via the *EMA average* $\bar{\theta}_t = (\bar{A}_t, \bar{B}_t)$, $\bar{\theta}_0 = \theta_0$, updated by

$$\bar{\theta}_{t+1} = \beta\,\bar{\theta}_t + (1-\beta)\,\theta_{t+1}^{\text{FA}}, \qquad \beta \in (0, 1). \tag{2}$$

This creates two *parallel trajectories*: the FedAvg trajectory $\{\theta_t^{\text{FA}}\}$ and the EMA trajectory $\{\bar{\theta}_t\}$. The next input state can assemble arbitrarily from these two trajectories, deciding independently for A and B whether to reflect the EMA (*feed back*) or keep FedAvg. The two choices are the analysis variables in this section. At evaluation or deployment every EMA configuration uses the *EMA-averaged factor endpoint* $W_0 + \bar{B}_T\bar{A}_T$.

3.2 Heterogeneity Collapse and Structural A/B Asymmetry

In highly heterogeneous data, FedAvg accumulates per-round adapter dispersion and converges to the majority class (G. et al., 2024). Round-internal methods address intra-round dispersion but not cross-round trajectory accumulation. The EMA in (2) is a natural cross-round smoother, and feeding it back imposes a *common* trade-off on $X \in \{A, B\}$ (Prop. 1): feeding X back keeps only a $(1-\beta)$ fraction of its displacement U_t^X; not feeding it back lets the gap $\epsilon_t^X := \bar{X}_t - X_t^{\text{FA}}$ follow $\epsilon_{t+1}^X = \beta\,\epsilon_t^X - \beta U_t^X$ and carry over across rounds. The structural roles of A (*direction*; random init, fixes the subspace) vs. B (*scale*; zero init, accumulates magnitude) make these two costs asymmetric in importance.

[A-condition] A requires EMA feedback. Without feedback (EMA-Eval/EMA-B), the noise-driven part of the A gap accumulates to $1/(1-\beta^2)$ times its one-round value (5.3× at $\beta = 0.9$; Cor. 1), and $\bar{B}\bar{A}$'s direction drifts from the path on which B was trained, leaving an irreducible $\Theta(L^2\sigma^2/(1-\beta))$ term in the gradient-norm bound (Prop. 3). A-feedback gives $\epsilon_{t+1}^A = -\beta U_t^A$, which depends on the current round only. The cost is acceptable because A is directional, so a smaller effective step does not stall learning.

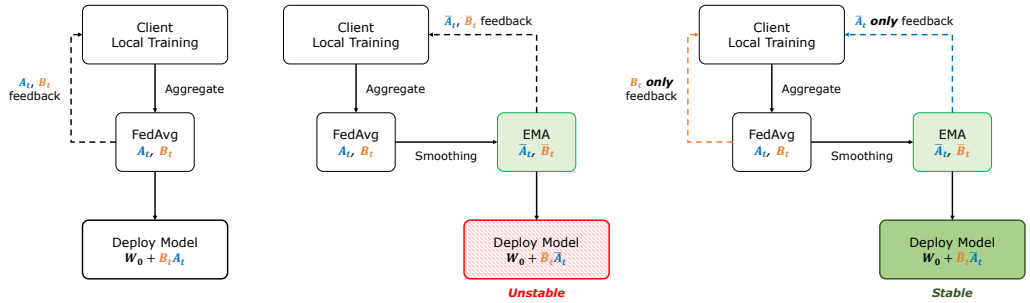

Figure 1: Pipelines of Vanilla, EMA (panel "EMA"), and SAF (definitions in Table 23). (a) Vanilla uses no EMA and deploys the FEDAVG training model directly. (b) EMA feeds back both $\bar{A}_t$ and $\bar{B}_t$ (dotted arrows), keeping only a $(1-\beta)$ fraction of B's per-round displacement (red box; "Unstable"). (c) SAF feeds back only $\bar{A}_t$ (blue) and keeps B_t on the FEDAVG path (orange); both deploy the EMA-averaged factor endpoint $\bar{B}_T\bar{A}_T$.

[B-condition] B rejects EMA feedback. Feeding back B gives $\bar{B}_{t+1} = \bar{B}_t + (1-\beta)U_t^B$; in the SGD model this inflates the bias term by $1/(1-\beta)$ ($10\times$ at $\beta = 0.9$; Thm. 1), which cannot be fully compensated for under the stability constraint $\eta L \leq 1$. The cost matters because B accumulates signal magnitude from zero. We therefore keep B on the FEDAVG path during training and only use $\bar{B}_T$ at deployment.

The 2×2 feedback grid yields four EMA configurations; only SAF satisfies both [A] and [B]:

	B kept on FEDAVG path	B fed back via EMA
A not fed back	EMA-EVAL ($\times$[A],$\checkmark$[B])	EMA-B ($\times$[A],$\times$[B])
A fed back via EMA	**SAF ($\checkmark$[A],$\checkmark$[B])**	EMA ($\checkmark$[A],$\times$[B])

EMA-EVAL deploys the same endpoint as SAF without training-time feedback, and EMA is the centralized-ML convention of feeding everything back. The asymmetry vanishes for single-matrix training or prefix tuning, where there is no direction/scale split. §4 formalizes SAF.

4 PROPOSED METHOD

Figure 1 contrasts Vanilla, EMA, and SAF. SAF assembles the training model in each round by the rule

$$A_{t+1}^{\text{in}} \leftarrow \bar{A}_{t+1}, \qquad B_{t+1}^{\text{in}} \leftarrow B_{t+1}^{\text{FA}}, \tag{3}$$

and defines the deployment model as $W_0 + \bar{B}_T\bar{A}_T$. The only hyperparameter is the EMA decay β (implementation in Appendix C.18).

A-feedback. SAF sets $A_{t+1}^{\text{in}} \leftarrow \bar{A}_{t+1}$ after each EMA update, ensuring $A_t^{\text{in}} = \bar{A}_t$ for all t. This resets cross-round accumulation: the gap reduces to the current-round term $-\beta U_t^A$ (Prop. 1(i)), and in the rank-1 model its variance shrinks to $\beta^2\sigma^2$ ($1-\beta^2 \approx 0.19\times$ at $\beta = 0.9$) with $O(\sigma\sqrt{L\Delta_0/T})$ convergence (Prop. 2).

B stays on FEDAVG during training; averaging only at deployment. Feeding back B would keep only $(1-\beta)U_t^B$ per round and incur Theorem 1's penalty. SAF keeps B on the FEDAVG path during training and uses $\bar{B}_T$ at deployment. Since $\bar{B}_t$ is never fed back, it does not affect learning dynamics; whether it helps is an empirical question, which we test directly (§6.2: $\bar{B}_T$ beats the raw B_T in 23/24 cells, and EMA-EVAL, which deploys the same endpoint without A-feedback, shows that deployment averaging alone does not explain SAF's gain).

Implementation. SAF adds a single copy of the `lora_A` tensor after the EMA update. Changing the sources of A and B in lines 6–7 of Algorithm 1 reproduces EMA-EVAL/EMA-B/EMA, the ablations of §6.2.

Algorithm 1 Federated LoRA training with SAF

Require: K clients $\{\mathcal{D}_k\}_{k=1}^K$; rounds T; participation fraction p; EMA decay β; LoRA rank r

Initialize A_0 randomly and set $B_0 \leftarrow 0$; set $(\bar{A}_0, \bar{B}_0) \leftarrow (A_0, B_0)$ and $(A_0^{\text{in}}, B_0^{\text{in}}) \leftarrow (A_0, B_0)$
for $t = 0, 1, \ldots, T-1$ **do**
  Sample $\mathcal{S}_t$ with $|\mathcal{S}_t| = pK$, broadcast $(A_t^{\text{in}}, B_t^{\text{in}})$; each $k \in \mathcal{S}_t$ trains locally (any optimizer, any number of steps)
  FedAvg: $A_{t+1}^{\text{FA}} \leftarrow \sum_k \frac{n_k}{n} A_t^{(k)}$, and $B_{t+1}^{\text{FA}} \leftarrow \sum_k \frac{n_k}{n} B_t^{(k)}$
  EMA: $\bar{A}_{t+1} \leftarrow \beta \bar{A}_t + (1-\beta) A_{t+1}^{\text{FA}}$, and $\bar{B}_{t+1} \leftarrow \beta \bar{B}_t + (1-\beta) B_{t+1}^{\text{FA}}$ ▷ $\bar{B}_t$ for deployment only
  SAF assembly: $A_{t+1}^{\text{in}} \leftarrow \bar{A}_{t+1}$ ▷ A-feedback: EMA → training
  $B_{t+1}^{\text{in}} \leftarrow B_{t+1}^{\text{FA}}$ ▷ B remains on the FedAvg path
end for
Deploy $W_0 + \bar{B}_T \bar{A}_T$ ▷ EMA-averaged factor endpoint; no training effect

5 Theoretical Analysis

This section formalizes the [A-condition] and [B-condition] from §3.2 in two layers: exact identities of the server feedback operator for any rank, local optimizer, and step budget (§5.1), and their quantitative consequences under a rank-1, single-step SGD reduction (§5.2). §5.3 states what each layer does and does not cover.

5.1 Exact Feedback Identities (Any Rank, Any Local Optimizer)

Proposition 1 (Feedback-operator identities)**.** *Let U_t^X be the realized displacement of $X \in \{A, B\}$ in* (1)*, produced by any local procedure, and $\epsilon_t^X := \bar{X}_t - X_t^{\text{FA}}$ with $\epsilon_0^X = 0$. For matrices of any shape: (i)* Gap. *With feedback ($X_t^{\text{in}} = \bar{X}_t$), $\epsilon_{t+1}^X = -\beta\, U_t^X$. Without feedback ($X_t^{\text{in}} = X_t^{\text{FA}}$), $\epsilon_{t+1}^X = \beta\, \epsilon_t^X - \beta\, U_t^X = -\sum_{j=0}^t \beta^{t+1-j}\, U_j^X$. (ii)* Kept displacement. *With feedback, $X_{t+1}^{\text{in}} - X_t^{\text{in}} = (1-\beta)\, U_t^X$; without feedback, $X_{t+1}^{\text{in}} - X_t^{\text{in}} = U_t^X$.*

The proof is two lines of EMA algebra (App. A.1). The identities are *symmetric*: feeding back either factor resets its own gap and squeezes its own displacement. The asymmetry lies in which effect matters for which factor—(i) for the direction factor A, (ii) for the scale factor B. The identities hold path-wise and do *not* assume that U_t is the same under different feedback choices, since U_t depends on the broadcast state.

Corollary 1 (Gap statistics at any rank)**.** *Let $\mathcal{F}_t$ be the history up to the broadcast of round t, and decompose $U_t^A = M_t + \Xi_t$ with $M_t = \mathbb{E}[U_t^A \mid \mathcal{F}_t]$ and $\sigma_{\text{lo}}^2 \le \mathbb{E}[\|\Xi_t\|_F^2 \mid \mathcal{F}_t] \le \sigma^2$. Write $\epsilon_{t+1}^A = D_{t+1} + N_{t+1}$ with D_{t+1} $\mathcal{F}_t$-measurable. With A-feedback, $D_{t+1} = -\beta M_t$ and $\mathbb{E}\|N_{t+1}\|_F^2 \in [\beta^2\sigma_{\text{lo}}^2, \beta^2\sigma^2]$. Without A-feedback, $D_{t+1} = -\sum_{j\le t} \beta^{t+1-j} M_j$ and $\mathbb{E}\|N_{t+1}\|_F^2 \in \left[\beta^2\sigma_{\text{lo}}^2, \beta^2\sigma^2\right] \cdot \frac{1-\beta^{2(t+1)}}{1-\beta^2}$.*

At matched per-round statistics, omitting A-feedback thus inflates the noise energy of the gap by up to $1/(1-\beta^2)$ ($5.3\times$ at $\beta = 0.9$) and a constant drift by up to $1/(1-\beta)$; identity (ii) is the exact server-side counterpart of the [B-condition].

5.2 Quantitative Consequences under a Rank-1 Reduction

Assumption 1 (Rank-1 LoRA, single-step SGD, bounded variance)**.** The adapter factors are $a, b \in \mathbb{R}^d$. For A, $a_{t+1}^{\text{FA}} = a_t^{\text{in}} + \mu + \xi_t$, where μ is a deterministic signal and ξ_t is IID heterogeneity noise with $\mathbb{E}\xi_t = 0$, $\mathbb{E}\|\xi_t\|^2 \le \sigma^2$ (σ grows as α decreases); the SGD instantiation is $\mu + \xi_t = -\eta(\nabla_A F + \zeta_t)$. For B, each client takes one SGD step with rate η from its input state and produces an unbiased gradient with variance at most σ_B^2. F is L-smooth in each factor with $\Delta_0 = F(\cdot_0) - F^*$. Each factor is analyzed as a separate block; the other enters only through these noise assumptions.

Intuition. Under B-feedback the EMA keeps only $(1-\beta)$ of B's per-round step (Prop. 1(ii)); since B starts at zero and must accumulate magnitude, this hidden squeeze stalls B's progress and inflates the bias term, as the next theorem makes precise.

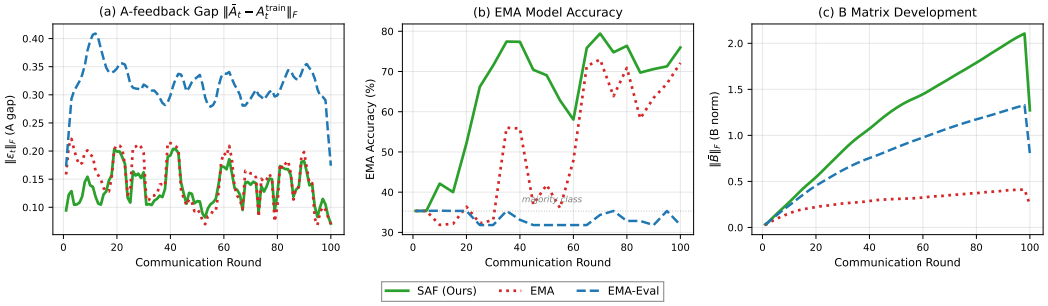

Figure 2: Theory validation (FedIT, MNLI, α=0.05, seed 42; the qualitative pattern is similar across 5 seeds). Panel (a) plots the EMA gap $\|\bar{A}_t - A_t^{\text{train}}\|_F$. EMA-EVAL accumulates to a high steady state (Prop. 1(i), Cor. 1), whereas EMA and **SAF** are nearly identical at ~ 0.13, showing that the difference between the two modes does not lie on the A side. Panel (b) plots the EMA accuracy: **SAF** stays within about 75–82%, whereas EMA fluctuates from 36–73%. Panel (c) shows $\|\bar{B}\|_F$: **SAF** attains 2.13 versus 0.45 for EMA (4.7×), the signature of the B-side squeeze; the $10\times$ of Theorem 1 is an effective-learning-rate ratio, not a predicted norm ratio.

Theorem 1 (Asymmetry in B's effective learning rate)**.** *Under Assumption 1 with the code-level learning rate $\eta \le 1/L$ in both methods,*

$$\begin{aligned} &\text{EMA:} \quad \tfrac{1}{T}\textstyle\sum_t \mathbb{E}\|\nabla_B F(\bar{b}_t)\|^2 \le \tfrac{2\Delta_0}{(1-\beta)\eta T} + L(1-\beta)\eta\sigma_B^2, \\ &\text{SAF:} \quad \tfrac{1}{T}\textstyle\sum_t \mathbb{E}\|\nabla_B F(b_t)\|^2 \le \tfrac{2\Delta_0}{\eta T} + L\eta\sigma_B^2. \end{aligned} \tag{4}$$

(i) *At equal η the bias term of* EMA *is $1/(1-\beta)$ times larger (ten-fold at $\beta = 0.9$).* (ii) *Both bounds attain the same AM–GM minimum, but the optimum for* EMA *is $\eta^\star_{\text{EMA}} = \eta^\star_{\text{SAF}}/(1-\beta)$; keeping it within $\eta L \le 1$ requires $T \gtrsim 2L\Delta_0/((1-\beta)^2\sigma_B^2)$ ($200L\Delta_0/\sigma_B^2$ at $\beta = 0.9$), which typically fails for a limited round budget (we use $T = 100$).*

Proof sketch. In EMA, if each client returns $\bar{b}_t - \eta g_t^B$, the EMA update simplifies to $\bar{b}_{t+1} = \bar{b}_t - (1-\beta)\eta\bar{g}_t^B$ (Prop. 1(ii)); SAF preserves η. Applying the L-smooth descent lemma symmetrically yields the two bounds (App. A). □

Proposition 2 (SAF convergence)**.** *Under Assumption 1 with A-feedback, the A gap reduces to the IID term $-\beta(\mu + \xi_t)$ with variance at most $\beta^2\sigma^2$, and $\frac{1}{T}\sum_t \mathbb{E}\|\nabla_A F(\bar{a}_t)\|^2 \le O(\sigma\sqrt{L\Delta_0/T})$ at the optimal step size.*

Proposition 3 (Accumulation of the A gap)**.** *Without A-feedback, the gap is a stable AR(1) process with stationary variance $\beta^2\sigma^2/(1-\beta^2)$, and it adds a $\Theta(L^2\sigma^2/(1-\beta))$ term to the EMA model's average-gradient-norm upper bound that does not vanish as $T \to \infty$.*

Proposition 4 (Heterogeneity crossover)**.** *Under Assumption 1, compare the upper bounds of* FEDAVG, *$\frac{2\Delta_0}{\eta T} + L\eta\sigma^2$, and of* SAF*'s A path, $\frac{2\Delta_0}{(1-\beta)\eta T} + L(1-\beta)\eta\sigma^2$, at the same η, T. Their difference $\frac{2\Delta_0\beta}{(1-\beta)\eta T} - L\beta\eta\sigma^2$ changes sign at*

$$\sigma^\star = \sqrt{\tfrac{2\Delta_0}{(1-\beta)\,L\,\eta^2\,T}}, \tag{5}$$

so SAF*'s bound is tighter when $\sigma > \sigma^\star$ and* FEDAVG*'s is tighter when $\sigma < \sigma^\star$.*

Proofs in Appendix A. **Bias–variance reading.** Prop. 4 is the classical bias–variance trade-off of iterate averaging on A: variance reduction helps in noisy settings and the slower effective step hurts in clean ones. It compares upper bounds, not losses, and does not model collapse; a heuristic collapse threshold is given in App. A.6 and is not claimed as a theorem.

5.3 SCOPE OF THE THEORY

Exact for any rank, local optimizer, and step budget: Prop. 1 and Cor. 1 (in our implementation the AdamW state is re-created every round, so each round is a fresh map from the broadcast state to U_t). *Modeled only under the rank-1, single-step SGD reduction:* the gradient-norm bounds, their

Table 1: Comparison of Vanilla, EMA, and **SAF** (**ours**). Metric: accuracy at the final round $t = T = 100$ (matched/mismatched average on MNLI), mean±std across five seeds {42–46}; GLUE at $\alpha = 0.05$, MedMNIST at $\alpha = 0.01$. **Bold**: best of the three configurations; underline: second best; *italic*: means at or below the majority-class baseline (collapse). *Bolding convention is reused by all subsequent tables.* Best-round results appear in Appendix C.13.

		GLUE (α=0.05)				MedMNIST (α=0.01)	
Method	Config	MNLI	QNLI	QQP	SST-2	OctMNIST	PathMNIST
FedIT	Vanilla	$61.9_{\pm 18.1}$	$55.2_{\pm 11.0}$	*57.2*$_{\pm 17.7}$	$79.9_{\pm 9.8}$	$39.7_{\pm 8.8}$	$63.5_{\pm 11.7}$
	EMA	$74.3_{\pm 6.2}$	$54.1_{\pm 5.9}$	$68.4_{\pm 5.4}$	$82.7_{\pm 16.4}$	$50.3_{\pm 16.1}$	$57.8_{\pm 13.2}$
	SAF	**78.7**$_{\pm 3.1}$	**72.1**$_{\pm 6.4}$	**76.9**$_{\pm 6.6}$	**84.0**$_{\pm 17.4}$	**62.6**$_{\pm 14.8}$	**78.9**$_{\pm 7.0}$
FedEx-LoRA	Vanilla	$57.0_{\pm 23.4}$	$55.4_{\pm 11.4}$	*59.8*$_{\pm 19.8}$	$77.3_{\pm 13.2}$	*36.0*$_{\pm 5.2}$	$62.7_{\pm 11.7}$
	EMA	$74.3_{\pm 5.9}$	$53.5_{\pm 5.9}$	$68.0_{\pm 5.0}$	$82.4_{\pm 16.9}$	$48.7_{\pm 16.4}$	$59.0_{\pm 13.6}$
	SAF	**79.0**$_{\pm 2.8}$	**73.8**$_{\pm 7.4}$	**77.1**$_{\pm 5.8}$	**84.5**$_{\pm 16.6}$	**56.4**$_{\pm 17.4}$	**80.2**$_{\pm 5.6}$
HetLoRA	Vanilla	$70.2_{\pm 7.8}$	$59.7_{\pm 17.4}$	$68.0_{\pm 6.3}$	**70.4**$_{\pm 20.5}$	*35.9*$_{\pm 5.3}$	$49.5_{\pm 12.8}$
	EMA	$71.8_{\pm 6.5}$	$56.5_{\pm 14.5}$	$71.4_{\pm 9.2}$	$67.8_{\pm 22.6}$	**44.9**$_{\pm 18.4}$	$43.1_{\pm 12.2}$
	SAF	**72.1**$_{\pm 7.6}$	**61.0**$_{\pm 15.8}$	**74.2**$_{\pm 10.5}$	$69.7_{\pm 20.2}$	$44.3_{\pm 18.9}$	**56.1**$_{\pm 13.0}$
LoRA-A^2	Vanilla	$65.4_{\pm 12.9}$	$56.1_{\pm 12.6}$	$64.1_{\pm 12.9}$	$78.1_{\pm 14.6}$	$44.8_{\pm 16.6}$	$67.8_{\pm 6.9}$
	EMA	$68.3_{\pm 12.4}$	$57.5_{\pm 10.4}$	$67.5_{\pm 3.9}$	$82.0_{\pm 17.0}$	$44.0_{\pm 14.6}$	$55.1_{\pm 7.6}$
	SAF	**75.8**$_{\pm 5.5}$	**66.9**$_{\pm 8.9}$	**74.9**$_{\pm 6.5}$	**85.3**$_{\pm 15.1}$	**61.2**$_{\pm 13.7}$	**70.0**$_{\pm 8.2}$

constants, and the location of $\sigma^\star$ (Thm. 1, Props. 2–4). *Not covered:* the coupled rank-r dynamics of BA under multi-step AdamW; the block-wise analysis does not track this coupling, so we claim no convergence rate or crossover location for our experimental setting (rank 8, 2 local epochs of AdamW). Our experiments test the predicted *orderings*, not the constants; the orderings persist at rank 16 and with 1 local epoch (§6.3).

6 Experiments

We evaluate SAF on four GLUE tasks (MNLI, QNLI, QQP, SST-2 using RoBERTa-base (Liu et al., 2019)) and two MedMNIST v2 tasks (PathMNIST, OctMNIST using ViT-small (Dosovitskiy et al., 2021; Yang et al., 2023)). The primary comparison is across four SAF-capable methods: FedIT (Zhang et al., 2024), FedEx-LoRA (Singhal et al., 2025), HetLoRA (Cho et al., 2024), and LoRA-A^2 (Koo et al., 2025). Two additional baselines on which SAF is structurally undefined—FFA-LoRA (A frozen) and FedFTG (Mahla et al., 2025) (no A/B factorization)—serve as falsification controls (App. C.17). All experiments use $\beta = 0.9$, $\eta = 2{\times}10^{-4}$, rank 8, $T = 100$ rounds, and 2 local AdamW epochs; main results are at GLUE $\alpha = 0.05$ and MedMNIST $\alpha = 0.01$, with the α sweep reported separately (App. C.18). The metric is final-round accuracy of the deployed model without checkpoint selection. **Evidence policy.** Every main-text claim rests on five seeds {42–46} unless tagged [1 seed], [2 seeds], or [3 seeds]; single-seed differences below the 7% GLUE seed-variance threshold of Dodge et al. (2020) are treated as directional.

6.1 Main Results: Collapse and Recovery

Table 1 reports final-round accuracy for the four SAF-capable methods. We observe:

Collapse and recovery. FedAvg suffers two collapse modes: *(deterministic)* at QNLI $\alpha = 0.05$ all four methods stay at 55–60%, trapped near the majority class (50.5%); *(stochastic)* on MNLI, FedEx-LoRA/FedIT have std 18–23% (seed-dependent collapse). SAF simultaneously beats FedAvg and EMA in 22/24 cells. Per-cell mean improvements are +11.7% over FedAvg and +8.8% over EMA (paired Wilcoxon $p < 10^{-4}$ for both; sign test 23/24). The 2 cells where SAF does not beat both baselines are HetLoRA SST-2 (FedAvg −0.7%) and HetLoRA OctMNIST (EMA −0.6%). On MNLI, SAF cuts the seed std to at most 7.6% ($\leq 3.1\%$ for FedIT/FedEx-LoRA).

Metric robustness. Final-round accuracy could favor stable methods, so we also count wins under best-round accuracy: SAF beats both baselines in 20/24 cells, beats EMA in 24/24, and leads the stronger baseline by 3.8% on average; all 18 non-HetLoRA cells are wins under both criteria

Table 2: Feedback ablation and deployment decomposition on MNLI ($\alpha = 0.05$, five seeds; final-round accuracy, mean±std). All EMA configurations deploy the EMA endpoint; Vanilla deploys the FEDAVG model. "SAF, raw B_T" evaluates SAF's training-path model ($\bar{A}_T$ with the FEDAVG-path B_T and head), i.e., A-feedback without deployment averaging; EMA-EVAL is deployment averaging without A-feedback. Bolding follows Table 1.

Method	Vanilla	EMA	EMA-EVAL	EMA-B	SAF, raw B_T	**SAF**
FedIT	$61.9_{\pm 18.1}$	$74.3_{\pm 6.2}$	$66.9_{\pm 20.7}$	$77.8_{\pm 3.6}$	$60.9_{\pm 19.1}$	**78.7**$_{\pm 3.1}$
FedEx-LoRA	$57.0_{\pm 23.4}$	$74.3_{\pm 5.9}$	$65.2_{\pm 19.4}$	$78.1_{\pm 4.0}$	$61.0_{\pm 19.4}$	**79.0**$_{\pm 2.8}$
HetLoRA	$70.2_{\pm 7.8}$	$71.8_{\pm 6.5}$	$71.2_{\pm 6.9}$	**73.6**$_{\pm 7.4}$	$64.9_{\pm 2.9}$	$72.1_{\pm 7.6}$
LoRA-A^2	$65.4_{\pm 12.9}$	$68.3_{\pm 12.4}$	**76.5**$_{\pm 5.7}$	$71.6_{\pm 9.9}$	$59.0_{\pm 19.6}$	$75.8_{\pm 5.5}$

Table 3: (a) Robustness beyond the theory's setting (MNLI, FedIT, $\alpha = 0.05$, seed 42; $r = 8$, $E = 2$ is the default). (b) RoBERTa-large 355M (MNLI, FedIT, $\alpha = 0.03$, three seeds); †original software environment, not a matched rerun. (c) FedProx combination (MNLI, FedIT, $\alpha = 0.05$). Final-round accuracy.

(a) rank / local epochs

setting	SAF	EMA	$\|\bar{B}\|_F$
r=8, E=2	76.0	72.1	2.13/0.45
r=16, E=2	77.0	74.1	2.64/0.48
r=8, E=1	71.5	*49.5*	1.57/0.33

(b) 355M, α=0.03

seed	Vanilla	EMA	SAF
42	*31.8*	48.3†	55.1
43	*31.8*	70.2	*32.8*
44	41.2	78.7	77.7
mean	34.9	65.7	55.2

(c) FedProx (+ config)

seed	alone	+EMA	+SAF
42	71.2	64.1	77.4
43	60.0	73.9	79.9

(App. C.13). Per-cell t-tests ($n = 5$) have limited power because FEDAVG's within-cell std reaches 23%; they flag SAF ahead in 7/24 cells vs. FEDAVG and 14/24 vs. EMA.

6.2 FEEDBACK ABLATION AND DEPLOYMENT DECOMPOSITION

A-feedback at a fixed deployed model. EMA-EVAL and SAF deploy the same endpoint $W_0 + \bar{B}_T\bar{A}_T$ and differ only in training-time A-feedback. On FedIT/FedEx-LoRA, A-feedback adds $+11.8\%$/$+13.8\%$ and cuts the std from 19.4–20.7% to 2.8–3.1% (Table 2), consistent with the A-gap diagnostics in Figure 2(a).

Deployment averaging, tested directly. Since $\bar{B}_T$ is never fed back, deploying it is a post-hoc choice, and averaging an oscillating B trajectory could smear conflicting client updates. Comparing SAF's deployed endpoint with its training-path model (both with $\bar{A}_T$), the endpoint is better in 23/24 cell means by $+0.7$ to $+26.4\%$ (median $+12.2\%$); the exception is HetLoRA-OctMNIST (44.3 vs. 49.4%), and the raw model is better in 19 of 120 individual runs (App. C.1). Neither ingredient suffices alone: the training-path model ($60.9\pm19.1\%$ on FedIT) is no better than FEDAVG, and averaging without A-feedback (EMA-EVAL, $66.9\pm20.7\%$) remains seed-dependent. We claim no SWA-type guarantee for this endpoint, since $\bar{B}_T\bar{A}_T \neq \mathrm{EMA}(B_tA_t)$.

B-feedback. With A fed back, additionally feeding back B (EMA) loses 4.4%/4.7% on FedIT/FedEx-LoRA and suppresses $\|\bar{B}\|_F$ by $4.7\times$ (Fig. 2c); $3\times$/$10\times$ η_B scaling narrows but does not close the gap [1 seed; App. C.15]. In our implementation EMA also feeds back the classification head, which the other configurations do not. The grid is not additive: EMA-B, which violates both conditions, is within 0.9% of SAF on FedIT/FedEx-LoRA and best on HetLoRA, in line with the symmetry of Prop. 1 (feeding back either factor resets its own gap). The predicted orderings SAF $\geq$ EMA-B $>$ EMA hold on FedIT/FedEx-LoRA but not on HetLoRA/LoRA-A^2; we treat HetLoRA, whose heterogeneous ranks weaken the shared-A premise, as a boundary case with a hypothesized cause (App. C.11). The falsification controls match Theorem 1 sign-wise [1 seed]: with A frozen (FFA-LoRA) EMA-EVAL beats EMA, and without the BA factorization (FedFTG) full EMA helps (App. C.17).

6.3 BEYOND THE ANALYZED REGIME, SCALE, AND COMPLEMENTARITY

Rank and local steps [1 seed]. The theory assumes rank 1 and a single SGD step, whereas the experiments use rank 8 and 2 local AdamW epochs. At rank 16 the SAF–EMA ordering and

the $\|\bar{B}\|_F$ suppression ($5.5\times$) persist (Table 3a). With 1 local epoch—closer to the single-step setting—EMA collapses (49.5%) while SAF reaches 71.5%, in the direction of Theorem 1's bias term, which binds harder when per-round progress is smaller. These are robustness checks, not a rank sweep or a substitute for a coupled rank-r AdamW analysis (§5.3).

Scale [3 seeds]. On RoBERTa-large at $\alpha = 0.03$, SAF outperforms FEDAVG on every seed ($+23.3$, $+1.0$, $+36.5\%$; mean $+20.3\%$), but its ordering against EMA is *not* consistent: SAF collapses on seed 43 and trails by 1.0% on seed 44 (Table 3b). We therefore claim improvement over FEDAVG at this scale, not over EMA; the single-seed $\alpha = 0.05$ run favors SAF (App. C.3).

FedProx [2 seeds]. SAF improves the FedProx endpoint on both seeds ($+6.2\%$, $+19.9\%$), whereas the effect of EMA reverses sign (-7.1%, $+13.9\%$); we therefore make no claim about the EMA–FedProx interaction. On seed 43, FedProx alone falls from 78.0% at its best round to 60.0% at the final round (Table 3c).

FedSA-LoRA [3 seeds]. FedSA-LoRA performs intra-round selective A aggregation with client-local, persistent B_k; SAF applies cross-round EMA feedback to the aggregated A, so the two can be stacked. On QNLI ($\alpha = 0.05$), adding SAF changes personalized accuracy by $+0.04$, -0.27, and -0.04%. A per-client majority-class predictor already scores 94.96–99.60% on the same held-out shards, leaving too little headroom to detect either additivity or over-smoothing of A. FedSA-LoRA's global model is near chance (49.5–49.8%) by construction, since B is never aggregated, so the two methods target different deployments (personalized models vs. one global adapter). The stack composes without failure; it is not a head-to-head comparison, and we claim neither a gain nor superiority (App. C.2).

6.4 HETEROGENEITY SENSITIVITY AND A PRE-TRAINING SCREEN

The α sweep [1 seed; App. C.6] shows two distinct crossovers, consistent with Prop. 4. *Against* FEDAVG, SAF wins or ties on MNLI at every $\alpha \leq 0.3$ ($+8.1\%$ at $\alpha = 0.3$) and loses by 1.1–1.3% at $\alpha \in \{0.5, 1.0\}$, placing the accuracy crossover between 0.3 and 0.5. *Against the strongest baseline* (including passive EMA), the crossover comes earlier: all gains above 10% occur at $\alpha \leq 0.05$, though not at every such α (MNLI $\alpha = 0.03$: -2.3%); for $\alpha \geq 0.07$, SAF is within -3.6 to $+2.6\%$ on GLUE and costs 1.3–1.7% on MNLI at $\alpha \geq 0.3$. Since Eq. (5) is not computable in practice, we test a screen that needs only client label histograms: the mean total-variation distance to the global label distribution has Spearman correlation 0.52 with the SAF$-$FEDAVG gain over the eight MNLI settings (0.82 excluding the degenerate $\alpha = 0.01$ tie), and a threshold near 0.4 separates positive from negative gains. It is fitted in-sample on mostly single-seed points, so it is a preliminary screen, not a validated estimator of $\sigma^\star$ (App. C.4).

7 DISCUSSION AND CONCLUSION

We show that in EMA smoothing for federated LoRA, the trade-off between displacement squeeze and cross-round gap accumulation is asymmetric between A and B, and propose SAF (feed back only A). Exact server-side identities locate both costs at any rank and for any local optimizer. Within $\alpha \leq 0.1$, SAF beats FEDAVG and EMA in 22/24 five-seed cells and cuts MNLI std from 23.4% to at most 7.6%; the gain requires both training-time A-feedback and deployment-time averaging, and Prop. 4's crossover $\sigma^\star$ ties the advantage to heterogeneity.

Limitations. *Theory–practice gap.* The identities are exact for our setting, but the bounds, constants, and $\sigma^\star$ are proved only for a block-wise rank-1 model with single-step SGD; the coupled rank-r BA dynamics under multi-step AdamW are not covered, and our rank-16 and 1-epoch checks are single-seed. *Closest baseline.* FedSA-LoRA yields personalized models and SAF a global adapter; our stacked run shows only compatibility, and a discriminating comparison requires evaluating every client model on a common balanced test set. *Data.* All partitions are synthetic Dirichlet label skew; naturally partitioned multi-device or multi-center data (feature shift, system heterogeneity) and generative/seq2seq tasks are untested. *Seeds and scale.* The α sweep, falsification controls, and η_B compensation are single-seed, and at 355M the ordering against EMA is seed-dependent. *Boundaries.* HetLoRA does not follow the predicted ordering, SAF costs 1–2% above the crossover, and the label-histogram screen is fitted in-sample.

Ethics Statement

This work does not involve human subjects, user studies, or new data collection. All experiments use public benchmarks under their respective licenses: four GLUE tasks (Wang et al., 2019) and two MedMNIST v2 tasks (Yang et al., 2023). PathMNIST and OctMNIST are pre-processed, de-identified subsets of public histology and retinal OCT collections, and we did not access any other patient data. We use them only as benchmarks with controlled label skew. They do not stand in for multi-center clinical data, and we make no claim about clinical readiness or deployment in a hospital federation (§7).

Privacy. SAF does not change what is exchanged between server and clients. Clients still receive a LoRA adapter of the same size and return a locally trained one; the EMA is kept on the server, and A-feedback only changes which server state is broadcast. SAF therefore inherits the privacy properties of the underlying federated protocol, neither weaker nor stronger. Federated learning alone is not a formal privacy guarantee. SAF acts on the server aggregate, so it is algorithmically compatible with secure aggregation and differential privacy, but we evaluate neither, and deployments on sensitive data should add them.

Fairness. Heterogeneity collapse is a fairness concern in its own right: when FedAvg converges to the majority class, clients whose data is dominated by minority classes are served worst. SAF reduces this collapse and its seed dependence in our benchmarks. However, we report aggregate accuracy only, not per-class or per-client accuracy, so we do not claim that SAF improves minority-class or per-client fairness. The method also has a known cost regime: above the heterogeneity crossover it is 1–2% worse than the best baseline (§6.4), so it should not be applied unconditionally.

Research integrity. We report the boundary and negative cases alongside the main results, including HetLoRA, the seed-dependent 355M comparison, and the saturated FedSA-LoRA evaluation. We state the number of seeds for every result and mark single-seed cells as directional. We withdraw earlier single-seed claims that did not replicate (App. C.16). We see no dual-use risk specific to SAF beyond those of federated fine-tuning in general.

Reproducibility Statement

Algorithm 1 gives the full procedure, and App. C.18 gives the one-line change to a standard federated LoRA loop. Datasets, partition generation, per-dataset client configuration (Table 22), hyperparameters, the evaluation protocol, and compute are listed in App. C.18. All assumptions are stated in §5 with complete proofs in App. A; §5.3 separates the results that hold at any rank and for any optimizer from those that hold only under the rank-1 reduction. Every table caption states its number of seeds. Some robustness runs (Table 3) used a later software environment than the main table, and App. C.18 lists the affected cells. The supplementary material contains the code, the exact configuration file of every reported run (759 runs, each with its recorded final accuracy), the client partition indices, and a data-preparation script that rebuilds the training tensors from the public datasets and verifies them against fingerprints of the tensors we used.

AI Use Statement

We used generative AI tools to edit text (grammar, spelling, word choice), to draft and revise sections of the paper, to help visualize results, and to help derive and check proofs. We did not use generative AI tools to generate experimental data. All experiments were run by the authors, and reported numbers are taken from the run logs and metric files. The authors reviewed all AI-assisted derivations and text, and we take responsibility for the final content of this work, including all claims.

REFERENCES

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Dan Busbridge, Jason Ramapuram, Pierre Ablin, Tatiana Likhomanenko, Eeshan Gunesh Dhekane, Xavier Suau, and Russell Webb. How to scale your EMA. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=DkeeXVdQyu.

Shuangyi Chen, Yuanxin Guo, Yue Ju, Hardik Dalal, Zhongwen Zhu, and Ashish J Khisti. Robust federated finetuning of LLMs via alternating optimization of LoRA. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=e8DrPuJekZ.

Yae Jee Cho, Luyang Liu, Zheng Xu, Aldi Fahrezi, and Gauri Joshi. Heterogeneous LoRA for federated fine-tuning of on-device foundation models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12903–12913, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-main.717/.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. 2020. URL https://arxiv.org/abs/2002.06305.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Daniel M. Jimenez G., David Solans, Mikko Heikkila, Andrea Vitaletti, Nicolas Kourtellis, Aris Anagnostopoulos, and Ioannis Chatzigiannakis. Non-iid data in federated learning: A survey with taxonomy, metrics, methods, frameworks and future directions. 2024. URL https://arxiv.org/abs/2411.12377.

Pengxin Guo, Shuang Zeng, Yanran Wang, Huijie Fan, Feifei Wang, and Liangqiong Qu. Selective aggregation for low-rank adaptation in federated learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=iX3uESGdsO.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. LoRA+: Efficient low rank adaptation of large models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17783–17806. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/hayou24a.html.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. 2019. URL https://arxiv.org/abs/1909.06335.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jiayu Huang, Xiaohu Wu, Tiantian He, and Qicheng Lao. Stabilized fine-tuning with LoRA in federated learning: Mitigating the side effect of client size and rank via the scaling factor. 2026. URL https://arxiv.org/abs/2603.08058.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Proceedings of the 2018 Conference on Uncertainty in Artificial Intelligence*, pp. 876–885. AUAI Press, 2018.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5132–5143. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/karimireddy20a.html.

Jabin Koo, Minwoo Jang, and Jungseul Ok. Towards robust and efficient federated low-rank adaptation with heterogeneous clients. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 416–429, Vienna, Austria, July 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.acl-long.19/.

Seanie Lee, Sangwoo Park, Dong Bok Lee, Dominik Wagner, Haebin Seong, Tobias Bocklet, Juho Lee, and Sung Ju Hwang. FedSVD: Adaptive orthogonalization for private federated learning with LoRA. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/forum?id=Qq19n9LZ97. arXiv:2505.12805.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020. URL https://proceedings.mlsys.org/paper_files/paper/2020/file/1f5fe83998a09396ebe6477d9475ba0c-Paper.pdf.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. 2019. URL https://arxiv.org/abs/1907.11692.

Navyansh Mahla, Kshitij Sharad Jadhav, and Ganesh Ramakrishnan. Exploring gradient subspaces: Addressing and overcoming LoRA's limitations in federated fine-tuning of large language models. 2025. URL https://arxiv.org/abs/2410.23111.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Daniel Morales-Brotons, Thijs Vogels, and Hadrien Hendrikx. Exponential moving average of weights in deep learning: Dynamics and benefits. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=2M9CUnYnBA.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. URL https://doi.org/10.1137/0330046.

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

Hamza Reguieg, Mohamed El Kamili, and Essaid Sabir. Fedema-distill: Exponential moving average guided knowledge distillation for robust federated learning. 2026. URL https://arxiv.org/abs/2603.04422.

Raghav Singhal, Kaustubh Ponkshe, and Praneeth Vepakomma. FedEx-LoRA: Exact aggregation for federated and efficient fine-tuning of large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1316–1336, Vienna, Austria, July 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.acl-long.67/.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving LoRA in privacy-preserving federated learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NLPzL6HWNl.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. 2023. URL https://arxiv.org/abs/2302.13971.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJ4km2R5t7.

Fei Wu, Jia Hu, Geyong Min, and Shiqiang Wang. Preventing rank collapse in federated low-rank adaptation with client heterogeneity. 2026. URL https://arxiv.org/abs/2602.13486.

Ziqing Xu, Hancheng Min, Jinqi Luo, Lachlan Ewen MacDonald, Salma Tarmoun, Enrique Mallada, and Rene Vidal. Understanding the learning dynamics of LoRA: A gradient flow perspective on low-rank adaptation in matrix factorization. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://openreview.net/forum?id=hphdX8WlcT.

Yunlu Yan, Chun-Mei Feng, Wangmeng Zuo, Lei Zhu, Rick Siow Mong Goh, and Yong Liu. Federated residual low-rank adaption of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=e0rQRMUhs7.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Yiyuan Yang, Guodong Long, Qinghua Lu, Liming Zhu, Jing Jiang, and Chengqi Zhang. Federated low-rank adaptation for foundation models: A survey. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25, 2025. URL https://doi.org/10.24963/ijcai.2025/1196.

Xin Yu, Yujia Wang, Jinghui Chen, and Lingzhou Xue. AltLoRA: Towards better gradient approximation in low-rank adaptation with alternating projections. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=9YNJ03jYsU.

Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Tong Yu, Guoyin Wang, and Yiran Chen. Towards building the federatedgpt: Federated instruction tuning. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6915–6919, 2024.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: Memory-efficient LLM training by gradient low-rank projection. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=hYHsrKDiX7.

Jiacheng Zhu, Kristjan Greenewald, Kimia Nadjahi, Haitz Sáez De Ocáriz Borde, Rickard Brüel Gabrielsson, Leshem Choshen, Marzyeh Ghassemi, Mikhail Yurochkin, and Justin Solomon. Asymmetry in low-rank adapters of foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62369–62385. PMLR, 21–27 Jul 2024a. URL https://proceedings.mlr.press/v235/zhu24c.html.

Zhenyu Zhu, Yongtao Wu, Quanquan Gu, and Volkan Cevher. Imbalance-regularized LoRA: A plug-and-play method for improving fine-tuning of foundation models. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*, 2024b. URL https://openreview.net/forum?id=hQfJ7nyDeK.

A PROOFS

A.1 PROOF OF PROPOSITION 1 AND COROLLARY 1

Fix a factor $X \in \{A, B\}$. By definition $X^{\mathrm{FA}}_{t+1} = X^{\mathrm{in}}_t + U^X_t$ and $\bar{X}_{t+1} = \beta\bar{X}_t + (1-\beta)X^{\mathrm{FA}}_{t+1}$. No assumption is made on how U^X_t is produced.

(i) With feedback, $X^{\mathrm{in}}_t = \bar{X}_t$. Then $X^{\mathrm{FA}}_{t+1} = \bar{X}_t + U^X_t$ and

$$\epsilon^X_{t+1} = \bar{X}_{t+1} - X^{\mathrm{FA}}_{t+1} = \beta(\bar{X}_t - X^{\mathrm{FA}}_{t+1}) = \beta(\bar{X}_t - \bar{X}_t - U^X_t) = -\beta U^X_t.$$

Without feedback, $X^{\mathrm{in}}_t = X^{\mathrm{FA}}_t$, so $X^{\mathrm{FA}}_{t+1} = X^{\mathrm{FA}}_t + U^X_t$ and

$$\epsilon^X_{t+1} = \beta(\bar{X}_t - X^{\mathrm{FA}}_{t+1}) = \beta(\bar{X}_t - X^{\mathrm{FA}}_t) - \beta U^X_t = \beta\epsilon^X_t - \beta U^X_t.$$

Unrolling from $\epsilon^X_0 = 0$ gives $\epsilon^X_{t+1} = -\sum_{j=0}^{t} \beta^{t+1-j} U^X_j$.

(ii) With feedback, $X^{\mathrm{in}}_{t+1} - X^{\mathrm{in}}_t = \bar{X}_{t+1} - \bar{X}_t = (1-\beta)(X^{\mathrm{FA}}_{t+1} - \bar{X}_t) = (1-\beta)U^X_t$. **Without feedback**, $X^{\mathrm{in}}_{t+1} - X^{\mathrm{in}}_t = X^{\mathrm{FA}}_{t+1} - X^{\mathrm{FA}}_t = U^X_t$. □

Every step uses only the linearity of the EMA and the definition of U^X_t. The identities therefore hold for matrices of any shape, any local optimizer (including AdamW, whose state our implementation re-creates at every round), any number of local steps, and any client-sampling scheme. Because U^X_t depends on the broadcast state, the displacement sequences under different feedback choices are generally different. The identities describe the feedback operator on each path; they do not compare realized trajectories across paths.

Corollary 1. Let $M_t = \mathbb{E}[U^A_t \mid \mathcal{F}_t]$ and $\Xi_t = U^A_t - M_t$. Then $\mathbb{E}[\Xi_t \mid \mathcal{F}_t] = 0$, and Ξ_j is $\mathcal{F}_t$-measurable for $j < t$. With feedback, $\epsilon^A_{t+1} = -\beta M_t - \beta\Xi_t$, so $D_{t+1} = -\beta M_t$ and $\mathbb{E}\|N_{t+1}\|^2_F = \beta^2\mathbb{E}\|\Xi_t\|^2_F \in [\beta^2\sigma^2_{\mathrm{lo}}, \beta^2\sigma^2]$. Without feedback, $D_{t+1} = -\sum_{j\le t}\beta^{t+1-j}M_j$ and $N_{t+1} = -\sum_{j\le t}\beta^{t+1-j}\Xi_j$. The cross terms vanish, since for $i < j$, $\mathbb{E}\langle\Xi_i, \Xi_j\rangle_F = \mathbb{E}\langle\Xi_i, \mathbb{E}[\Xi_j \mid \mathcal{F}_j]\rangle_F = 0$. Hence $\mathbb{E}\|N_{t+1}\|^2_F = \sum_{j\le t}\beta^{2(t+1-j)}\mathbb{E}\|\Xi_j\|^2_F$, which lies between σ^2_{lo} and σ^2 times $\sum_{k=1}^{t+1}\beta^{2k} = \beta^2(1-\beta^{2(t+1)})/(1-\beta^2)$. For a constant drift $M_j \equiv M$, $D_{t+1} = -\beta M(1-\beta^{t+1})/(1-\beta)$ without feedback and $-\beta M$ with it. □

A.2 PROOF OF THEOREM 1

Step 1: effective learning rate of EMA. In EMA, every client starts round t from $(\bar{a}_t, \bar{b}_t)$, and one SGD step gives $b^{(k)}_{t+1} = \bar{b}_t - \eta g^{B,(k)}_t$. Applying FEDAVG and then the EMA,

$$b^{\mathrm{FA}}_{t+1} = \bar{b}_t - \eta\,\bar{g}^B_t, \qquad \bar{b}_{t+1} = \beta\bar{b}_t + (1-\beta)b^{\mathrm{FA}}_{t+1} = \bar{b}_t - (1-\beta)\eta\,\bar{g}^B_t,$$

where $\bar{g}^B_t = \sum_k \frac{n_k}{n} g^{B,(k)}_t$ is unbiased with variance at most σ^2_B. This is Prop. 1(ii) with $U^B_t = -\eta\bar{g}^B_t$. The EMA trajectory is therefore unbiased SGD with effective learning rate $(1-\beta)\eta$.

Step 2: descent lemma. Under L-smoothness and $(1-\beta)\eta \le \eta \le 1/L$, $\mathbb{E}[F(\bar{b}_{t+1}) \mid \bar{b}_t] \le F(\bar{b}_t) - \frac{\eta_{\mathrm{eff}}}{2}\|\nabla F\|^2 + \frac{L\eta^2_{\mathrm{eff}}}{2}\sigma^2_B$. Telescoping over $t = 0, \dots, T-1$ with $F \ge F^*$ gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla_B F(\bar{b}_t)\|^2 \le \frac{2\Delta_0}{(1-\beta)\eta T} + L(1-\beta)\eta\sigma^2_B. \tag{6}$$

Step 3: the SAF path. In SAF, $b^{\mathrm{FA}}_{t+1} = b_t - \eta\bar{g}^B_t$ and $b_{t+1} = b^{\mathrm{FA}}_{t+1}$ (Eq. 3). This is unbiased SGD with learning rate η, and the same lemma gives $\frac{1}{T}\sum_t \mathbb{E}\|\nabla_B F(b_t)\|^2 \le \frac{2\Delta_0}{\eta T} + L\eta\sigma^2_B$.

Step 4: limits of retuning the learning rate. By AM–GM, both bounds attain the minimum $2\sqrt{2L\Delta_0\sigma^2_B/T}$, at $\eta^\star_{\mathrm{SAF}} = \sqrt{2\Delta_0/(LT\sigma^2_B)}$ and at $\eta^\star_{\mathrm{EMA}} = \eta^\star_{\mathrm{SAF}}/(1-\beta)$, respectively. The client-side stability constraint $\eta L \le 1$ applies to the code-level η in both methods. EMA can therefore reach its minimum only if $\eta^\star_{\mathrm{SAF}} \le (1-\beta)/L$, that is, $\sqrt{2L\Delta_0/(T\sigma^2_B)} \le 1-\beta$, or $T \ge 2L\Delta_0/((1-\beta)^2\sigma^2_B)$. At $\beta = 0.9$ this requires $T \ge 200L\Delta_0/\sigma^2_B$. □

Remark (averaging B at deployment). SAF deploys $\bar{b}_T$, an EMA of the FEDAVG path of B that is never fed back, so it does not change the training dynamics analyzed above. We do not derive a guarantee for the deployed product. In general $\bar{B}_T\bar{A}_T \neq \mathrm{EMA}(B_tA_t)$, and averaging can smear late iterates that oscillate. We therefore evaluate the deployment choice directly (App. C.1).

A.3 Proof of Proposition 2

With A-feedback, $a_t^{\mathrm{in}} = \bar{a}_t$, so $a_{t+1}^{\mathrm{FA}} = \bar{a}_t + \mu + \xi_t$. Then $\bar{a}_{t+1} = \bar{a}_t + (1-\beta)(\mu + \xi_t)$ and $\bar{a}_{t+1} - a_{t+1}^{\mathrm{FA}} = -\beta(\mu + \xi_t)$, which is IID with $\mathbb{E}\|\cdot - \mathbb{E}\cdot\|^2 \le \beta^2\sigma^2$ (this is Prop. 1(i) in this model). With the SGD instantiation $\mu + \xi_t = -\eta(\nabla_A F(\bar{a}_t) + \zeta_t)$ and $\mathbb{E}\|\zeta_t\|^2 \le \sigma_A^2$, the update is $\bar{a}_{t+1} = \bar{a}_t - \eta_{\mathrm{eff}}(\nabla_A F(\bar{a}_t) + \zeta_t)$ with $\eta_{\mathrm{eff}} = (1-\beta)\eta$. The descent lemma and telescoping give

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla_A F(\bar{a}_t)\|^2 \le \frac{2\Delta_0}{\eta_{\mathrm{eff}}T} + L\eta_{\mathrm{eff}}\sigma_A^2,$$

which is minimized at $2\sqrt{2L\Delta_0\sigma_A^2/T} = O(\sigma_A\sqrt{L\Delta_0/T})$. Here σ in the statement plays the role of σ_A. □

A.4 Proof of Proposition 3

Without A-feedback, $a_t^{\mathrm{FA}} = a_{t-1}^{\mathrm{FA}} + \mu + \xi_{t-1}$, and by Prop. 1(i) $\epsilon_t = \beta\epsilon_{t-1} - \beta\mu - \beta\xi_{t-1}$, a stable AR(1) process. At stationarity $\mathbb{E}\epsilon = -\beta\mu/(1-\beta)$ and, because ξ_{t-1} is independent of ϵ_{t-1}, $\mathbb{E}\|\epsilon - \mathbb{E}\epsilon\|^2 \le \beta^2\sigma^2/(1-\beta^2)$. At $\beta = 0.9$ and $\sigma = 0.15$ the standard deviation is ≈ 0.31, close to the EMA-EVAL A-gap of ≈ 0.32 in Figure 2(a). In the descent inequality for the EMA model, the gradient is evaluated at a_t^{FA} rather than at $\bar{a}_t$. Young's inequality with $\|\nabla F(\bar{a}_t) - \nabla F(a_t^{\mathrm{FA}})\| \le L\|\epsilon_t\|$ gives $\langle \nabla F(\bar{a}_t), \nabla F(a_t^{\mathrm{FA}})\rangle \ge \frac{1}{2}\|\nabla F(\bar{a}_t)\|^2 - \frac{L^2}{2}\|\epsilon_t\|^2$, and telescoping yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\bar{a}_t)\|^2 \le \frac{2\Delta_0}{\eta_{\mathrm{eff}}T} + L\eta_{\mathrm{eff}}\sigma_A^2 + L^2\,\mathbb{E}\|\epsilon - \mathbb{E}\epsilon\|^2.$$

The last term is $\Theta(L^2\sigma^2/(1-\beta))$ and does not vanish as $T \to \infty$. It is a residual in an upper bound, not a lower bound on the error. □

A.5 Proof of Proposition 4

The FEDAVG bound $\frac{2\Delta_0}{\eta T} + L\eta\sigma^2$ is the standard SGD bound with learning rate η. The SAF bound on the A path is Prop. 2 with $\eta_{\mathrm{eff}} = (1-\beta)\eta$. Their difference is $\frac{2\Delta_0}{\eta T}(\frac{1}{1-\beta} - 1) - L\eta\sigma^2\beta = \frac{2\Delta_0\beta}{(1-\beta)\eta T} - L\beta\eta\sigma^2$, which is zero at $\sigma^\star$ in (5). □

A.6 Heuristic Collapse Threshold

We model collapse as the event that the A-gap exceeds a threshold τ at some round: $P_T = \Pr(\max_{t\le T}\|\epsilon_t\| > \tau)$. With the stationary standard deviation from Prop. 3, this suggests three regimes as σ varies: safe ($P_T \approx 0$), stochastic ($0 < P_T < 1$), and deterministic ($P_T \approx 1$), with a critical scale $\sigma \sim \tau\sqrt{1-\beta^2}/\beta$. Under SAF the gap standard deviation is smaller by a factor $\sqrt{1-\beta^2}$. The threshold τ is not derived from the objective, and the link between the gap and majority-class prediction is not proved, so we use this only as an organizing heuristic, not as a theorem. The QNLI (α=0.05, all methods near the majority class) and MNLI (α=0.05, seed-dependent) patterns in Table 1 match its deterministic and stochastic regimes.

A.7 What Transfers Beyond Rank 1, and What Does Not

A tempting route from rank 1 to rank r treats each row of A independently. That argument does not track the coupling between the rows of A and the columns of B through $\Delta W = \sum_i B_{:,i}A_{i,:}$, so we do not rely on it. Our position is as follows.

- *Exact at any rank.* Prop. 1 and Cor. 1 hold for matrices of any shape, any local optimizer, and any local step budget. They fully determine how each feedback choice propagates the realized round displacements U_t^A, U_t^B.

- *What the displacements hide.* U_t^A and U_t^B absorb the coupled BA dynamics, the optimizer, and client heterogeneity, and they depend on the feedback path. The identities say nothing about how large they are.

- *Rank-1 quantitative layer.* Thm. 1 and Props. 2–4 add a model for the displacements: one SGD step, block-wise smoothness, and unbiased noise with bounded variance. The algebra of the descent arguments is written in Frobenius norms and does not use $r = 1$ beyond this block-wise model. The block-wise model itself, however, ignores that the objective seen by one factor moves with the other. A formal rank-r result would need a joint descent lemma for (A, B) with explicit coupling terms, which we do not provide. We therefore state these results only for the rank-1 reduction.

- *Empirical bridge.* The predicted orderings (SAF over EMA, suppression of $\|\bar{B}\|_F$ under B-feedback) persist at rank 16 and with one local epoch [1 seed], and at rank 8 with two AdamW epochs [5 seeds] (Table 3a). The measured norm ratios (4.7–$5.5\times$) are signatures of the $(1-\beta)$ shrinkage, not quantitative checks of the $10\times$ effective step-size ratio.

B Detailed Differentiation from Related Methods

vs. FedSA-LoRA (Guo et al., 2025). FedSA-LoRA observes that within a round A tends to capture shared knowledge and B client-specific knowledge. It aggregates A (and the head) at the server and keeps each client's B_k local and persistent. It is an *intra-round* rule that yields *personalized* models. SAF is a *cross-round* rule on the aggregated A that yields one *global* adapter. The two can be implemented together: SAF's EMA and feedback act on whatever A aggregate FedSA-LoRA produces. App. C.2 reports the combination.

vs. FedEMA-Distill (Reguieg et al., 2026). FedEMA-Distill keeps an EMA of the global model and uses it as a distillation target. Our EMA configuration instead feeds the EMA back as the next round's initialization and has no distillation loss, weighting, or temperature. It is a controlled full-feedback ablation for the 2×2 grid, not an implementation of FedEMA-Distill, and our results should not be read as a comparison with that method.

vs. FedSVD (Lee et al., 2025). FedSVD orthogonalizes A at the server to mitigate differential-privacy noise. It uses the same channel, server-side processing of A, for a different failure mode.

vs. FedOpt, FedAdam, and FedYogi (Reddi et al., 2021). These server optimizers smooth the aggregated update $\Delta\theta_t$ with moment estimates and do not act on the factorization. The B-side shrinkage of Prop. 1(ii) arises when an EMA is fed back into a factor that must build up magnitude from zero. Combining the two is algorithmically straightforward (replace line 4 of Algorithm 1 with a FedOpt step), but we have not tested it.

vs. FFA-LoRA (Sun et al., 2024). FFA-LoRA freezes A, which removes A-drift at the cost of half the trainable parameters. It cannot be combined with SAF, since there is no A to feed back. We use it as a falsification control (App. C.17).

Method	Channel	Output	Target failure	Relation to SAF (evidence)
SAF (ours)	cross-round EMA feedback on A	global adapter	cross-round A-gap; B shrinkage	—
FedSA-LoRA	intra-round selective aggregation	personalized models	client-specific B	composable; runs without failure; gain untested (saturated metric)
FedEMA-Distill	EMA of full model as teacher	global model	general drift	not implemented; EMA is not a substitute
FedSVD	SVD re-orthogonalization of A	global adapter	DP noise amplification	compatible; untested
FFA-LoRA	freeze A	global adapter	A drift, by removal	incompatible; used as control
FedOpt family	adaptive server step on $\Delta\theta$	global model	variance of aggregate	compatible; untested

C Additional Experimental Results

C.1 Averaging at Deployment: Training-Path Model vs. Deployed Endpoint

This subsection tests the deployment choice of SAF directly. For every SAF cell in Table 1 we compare the deployed endpoint $W_0 + \bar{B}_T \bar{A}_T$ with the training-path model at the final round; both use $\bar{A}_T$ (since A is fed back), and they differ in B (FedAvg-path B_T vs. $\bar{B}_T$) and in the classification head (FedAvg-path vs. EMA). The deployed endpoint is better in 23 of 24 cell means, by $+0.7$ to $+26.4\%$ (median $+12.2\%$, mean $+12.1\%$); the only negative cell is HetLoRA-OctMNIST, and the training-path model is better in 19 of the 120 individual runs (Table 4). The training-path accuracy of SAF is close to FedAvg's on most tasks (60.9 vs. 61.9% on MNLI/FedIT), with PathMNIST as the exception (54.0 vs. 63.5%), so SAF does not destabilize the training path. Deployment averaging alone, however, does not explain the gain: EMA-Eval, which deploys the same endpoint without A-feedback, reaches $66.9\pm20.7\%$ on MNLI/FedIT and $65.2\pm19.4\%$ on MNLI/FedEx-LoRA, against SAF's $78.7\pm3.1\%$ and $79.0\pm2.8\%$. Deployment-time averaging and training-time A-feedback are therefore both needed.

Table 4: Deployed endpoint vs. training-path model under SAF (five-seed means of final-round accuracy; each cell: deployed / training path / difference). GLUE at α=0.05, MedMNIST at α=0.01.

Method	MNLI	QNLI	QQP	SST-2	OctMNIST	PathMNIST
FedIT	78.7 / 60.9 / +17.8	72.1 / 55.8 / +16.3	76.9 / 63.0 / +13.9	84.0 / 78.1 / +5.9	62.6 / 41.4 / +21.2	78.9 / 54.0 / +24.9
FedEx-LoRA	79.0 / 61.0 / +17.9	73.8 / 55.8 / +18.1	77.1 / 64.4 / +12.7	84.5 / 78.1 / +6.4	56.4 / 41.6 / +14.8	80.2 / 53.8 / +26.4
HetLoRA	72.1 / 64.9 / +7.1	61.0 / 58.4 / +2.6	74.2 / 62.6 / +11.7	69.7 / 68.9 / +0.7	44.3 / 49.4 / −5.1	56.1 / 44.3 / +11.8
LoRA-A^2	75.8 / 59.0 / +16.7	66.9 / 53.8 / +13.1	74.9 / 65.7 / +9.2	85.3 / 74.4 / +10.9	61.2 / 54.8 / +6.4	70.0 / 61.0 / +9.0

C.2 Combination with FedSA-LoRA

Implementation. We follow Guo et al. (2025): A and the classification head are aggregated, and each client's trained B_k stays on the client and is restored at the start of the client's next local round. FedSA-LoRA+SAF applies the EMA and A-feedback of Algorithm 1 to the aggregated A. For personalized evaluation, each client's model (shared A and head plus its own B_k) is evaluated on a held-out 20% of that client's shard. Global evaluation uses the full QNLI dev set.

Metric ceiling. Under α=0.05 most client shards are nearly single-class. A constant predictor that outputs each client's majority label scores 99.60, 94.96, and 99.35% on the same held-out shards for seeds 42–44. FedSA-LoRA sits at this ceiling on two of the three seeds, so its client-local B_k alone saturates the metric.

What this shows and what it does not. Adding SAF changes personalized accuracy by $+0.04$, -0.27, and -0.04%; with the metric at its ceiling, this neither establishes a gain nor rules out over-smoothing of A. The global model is near chance by construction—a global model with $B = 0$ gives $\Delta W = BA = 0$ however good the shared A is—which is a property of FedSA-LoRA's design rather than a failure, so a comparison with SAF's global accuracy would be unfair and we do not

Table 5: FedSA-LoRA ± SAF on QNLI (α=0.05, three seeds). "Global" is on the full dev set (majority class 50.5%); "personalized" is on each client's own held-out shard. The two columns use different models and different test data and should not be compared with each other.

	Personalized		Global		Per-client
Seed	FedSA-LoRA	+SAF	FedSA-LoRA	+SAF	majority baseline
42	99.61	99.65	49.6	49.8	99.60
43	97.99	97.72	49.5	49.5	94.96
44	99.35	99.31	49.5	49.5	99.35

make it. A discriminating experiment would evaluate every client model (shared A plus its own B_k) on the same balanced test set, with and without SAF, alongside SAF's global adapter; this requires persisting the client B_k matrices and is left to future work.

C.3 SCALE: ROBERTA-LARGE

Table 6: RoBERTa-large 355M, MNLI, FedIT. Final-round accuracy (best-round accuracy in parentheses where available). The α=0.03 block has three seeds. The seed-42 EMA run† comes from the software environment of the original runs; the other α=0.03 runs use a rebuilt environment. The seed-42 SAF run was repeated in both environments (56.5 vs. 55.1 final; best 84.0 vs. 84.6). The $\alpha\in\{0.01, 0.05\}$ rows use one seed.

α	seed	Vanilla	EMA	SAF
0.01	42	*31.8*	*31.8*	*31.8*
0.03	42	*31.8* (73.0)	48.3† (70.8)	55.1 (84.6)
	43	*31.8* (35.3)	70.2 (79.4)	*32.8* (35.3)
	44	41.2 (85.2)	78.7 (78.8)	77.7 (85.0)
0.05	42	64.6	78.9	84.5

At 355M, SAF outperforms FEDAVG on every α=0.03 seed (paired mean +20.3%) but does not outperform EMA consistently: SAF collapses on seed 43 and trails by 1.0% on seed 44, and the seed-42 comparison with EMA crosses software environments. The large seed variance in every configuration places this cell in the stochastic regime of App. A.6. We therefore claim improvement over FEDAVG at this scale, but not over EMA.

C.4 LABEL-HISTOGRAM SCREEN FOR THE CROSSOVER

This subsection describes the label-histogram screen used in §6.4. For each α we compute the mean total-variation distance $\mathrm{TV} = \frac{1}{K}\sum_k \frac{1}{2}\|p_k - p\|_1$ between client label distributions p_k and the global distribution p, directly from the MNLI partitions and before any training (Table 7). Its Spearman correlation with the SAF−FEDAVG final-accuracy difference (FedIT, seed 42) is 0.52 over all eight settings and 0.82 excluding α=0.01, where both methods sit at the majority class and the difference is exactly zero; a threshold near $\mathrm{TV} \approx 0.4$ separates positive from negative differences. We also checked early-round variance of client updates as a second candidate, and its correlation is weak (≈ 0.3). The TV screen targets the crossover against FEDAVG, not the earlier crossover against the strongest baseline. Because it is fitted in-sample on one dataset with mostly single-seed outcomes, it needs validation on other datasets and partition schemes before it can be used as an estimator of $\sigma^\star$.

Table 7: Label-histogram statistics of the MNLI partitions (100 clients) and the SAF−FEDAVG difference (FedIT, seed 42, final round). Class coverage is the mean fraction of classes with at least 5% of a client's labels.

α	0.01	0.03	0.05	0.07	0.10	0.30	0.50	1.00
mean TV	0.653	0.634	0.609	0.584	0.566	0.449	0.379	0.284
class coverage	0.36	0.39	0.42	0.47	0.50	0.65	0.79	0.91
SAF−FEDAVG	+0.0	+8.4	+14.7	+38.4	+5.9	+8.1	−1.1	−1.3

Table 8: α=0.05 GLUE, five configurations (FedIT/FedEx-LoRA, seed 42, final round). EMA-B on SST-2 was not measured. *Italic*: at or below the majority-class baseline. Bolding follows Table 1.

Method	Dataset	Vanilla	EMA	EMA-EVAL	EMA-B	**SAF**
FedIT	MNLI	61.3	72.1	*31.8*	75.6	**76.0**
	QNLI	*50.5*	56.7	*50.5*	**81.9**	69.1
	QQP	69.0	77.1	71.8	82.1	**83.3**
	SST-2	84.7	82.5	**89.8**	—	89.6
FedEx-LoRA	MNLI	*31.8*	71.5	*31.8*	**77.7**	76.2
	QNLI	*50.5*	52.7	*50.5*	**81.0**	80.3
	QQP	81.3	76.1	*37.6*	80.0	**82.8**
	SST-2	87.4	82.2	**91.7**	—	89.4

C.5 FIVE-CONFIGURATION GLUE COMPARISON AT α=0.05

C.6 FULL α SWEEP (GLUE)

Tables 9–12 report FedIT/FedEx-LoRA for $\alpha \in [0.01, 0.1]$ (seed 42, final round). The "EMA" column is passive EMA-EVAL, and $\Delta = \text{SAF} - \max(Vanilla, \text{EMA-EVAL})$. MNLI above α=0.1 is in App. C.7. *Italic*: at or below the majority-class baseline. All sweep cells use one seed and are directional.

Table 9: MNLI α sweep (seed 42, final round; majority class 35.3%).

	FedIT				FedEx-LoRA			
α	Vanilla	EMA-EVAL	**SAF**	Δ	Vanilla	EMA-EVAL	**SAF**	Δ
0.01	*31.8*	*31.8*	*31.8*	+0.0	*31.8*	*32.0*	*31.8*	−0.2
0.03	*35.4*	**46.0**	43.7	−2.3	*31.8*	**48.1**	45.4	−2.7
0.05	61.3	*31.8*	**76.0**	+14.7	*31.8*	*31.8*	**76.2**	+44.4
0.07	42.7	**81.6**	81.2	−0.4	41.4	**81.9**	81.3	−0.6
0.10	63.5	**69.4**	**69.4**	−0.0	63.6	**70.6**	70.4	−0.3

Table 10: QNLI α sweep (seed 42, final round; majority class 50.5%).

	FedIT				FedEx-LoRA			
α	Vanilla	EMA-EVAL	**SAF**	Δ	Vanilla	EMA-EVAL	**SAF**	Δ
0.01	*50.5*	*50.5*	*50.5*	+0.0	*50.5*	*48.8*	*50.5*	+0.0
0.03	*49.5*	*50.7*	**74.1**	+23.5	*49.5*	*50.8*	**78.9**	+28.2
0.05	*50.5*	*50.5*	**69.1**	+18.6	*50.5*	*50.5*	**80.3**	+29.7
0.07	80.3	**90.2**	89.4	−0.8	79.4	87.0	**89.6**	+2.6
0.10	86.5	88.4	**88.6**	+0.2	87.5	88.3	**88.4**	+0.1

Table 11: QQP α sweep (seed 42, final round; majority class 63.2%).

	FedIT				FedEx-LoRA			
α	Vanilla	EMA-EVAL	**SAF**	Δ	Vanilla	EMA-EVAL	**SAF**	Δ
0.01	*36.8*	63.2	**77.7**	+14.5	*36.8*	*45.2*	**76.4**	+31.2
0.03	63.2	63.2	63.2	+0.0	63.2	63.2	63.2	+0.0
0.05	69.0	71.8	**83.3**	+11.5	81.3	*37.6*	**82.8**	+1.5
0.07	*57.8*	**86.0**	85.4	−0.5	*57.7*	**85.7**	85.2	−0.5
0.10	**82.1**	80.3	78.5	−3.6	**82.0**	80.8	79.5	−2.6

C.7 ABOVE THE CROSSOVER

Proposition 4 predicts that SAF pays a small optimization-bias cost above the crossover, i.e., in the near-IID regime where client-gradient variance is small. We verify this on MNLI by sweeping $\alpha \in \{0.30, 0.50, 1.00\}$ (seed 42; Table 13). Against FEDAVG, SAF still wins at α=0.3 (+8.1%/+7.9%) and loses by 1.1–1.3% at $\alpha \in \{0.5, 1.0\}$, so the crossover against FEDAVG lies between 0.3 and 0.5. Against passive EMA, SAF underperforms by 1.3–1.7% at all three values. The cost is small and

Table 12: SST-2 α sweep (seed 42, final round; majority class 50.9%).

	FedIT				FedEx-LoRA			
α	Vanilla	EMA-EVAL	**SAF**	Δ	Vanilla	EMA-EVAL	**SAF**	Δ
0.01	*50.9*	*49.1*	**86.5**	$+35.6$	*50.9*	*50.9*	**76.5**	$+25.6$
0.03	*50.9*	*50.9*	*50.9*	$+0.0$	*50.9*	*50.9*	*50.9*	$+0.0$
0.05	84.7	**89.8**	89.6	-0.2	87.4	**91.7**	89.4	-2.3
0.07	*55.7*	**92.5**	91.7	-0.8	*56.8*	**93.0**	91.7	-1.3
0.10	87.7	92.2	**92.4**	$+0.2$	89.4	92.4	**92.7**	$+0.2$

bounded, and this is the regime where Stochastic Weight Averaging (Izmailov et al., 2018) is also known to underperform: averaging hurts when SGD is already converging cleanly.

Table 13: MNLI above the crossover (seed 42, final round). Δ = SAF $-$ $\max(Vanilla, \text{EMA-EVAL})$, computed from unrounded accuracies.

	FedIT				FedEx-LoRA			
α	Vanilla	EMA-EVAL	**SAF**	Δ	Vanilla	EMA-EVAL	**SAF**	Δ
0.30	75.2	**84.8**	83.3	-1.5	75.3	**84.9**	83.2	-1.7
0.50	84.7	**84.9**	83.5	-1.3	84.7	**84.9**	83.6	-1.3
1.00	**85.2**	**85.2**	83.9	-1.4	85.1	**85.2**	83.8	-1.4

C.8 DIRECT COMPARISON WITH FULL EMA FEEDBACK ACROSS α

For the headline regime $\alpha = 0.05$, EMA has five-seed coverage on all four SAF-capable methods (Table 1). For other α values, single-seed EMA runs exist for MNLI on FedIT and FedEx-LoRA across the full sweep, and Table 14 reports every cell where both configurations are measured at seed 42. SAF matches or beats EMA in all 16 cells and is strictly better, by $+3.0$ to $+9.3\%$, in the 14 cells where either configuration escapes the majority class—including near-IID cells where SAF itself trails FEDAVG. This is consistent with the B-side penalty of Theorem 1 being present at every level of heterogeneity. *Caveat: single-seed final-round.*

Table 14: SAF vs. EMA on MNLI across α (seed 42, final round). *Italic*: at or below the majority-class baseline (35.3%).

	FedIT				FedEx-LoRA			
α	Vanilla	EMA	**SAF**	Δ	Vanilla	EMA	**SAF**	Δ
0.01	*31.8*	*31.8*	*31.8*	$+0.0$	*31.8*	*31.8*	*31.8*	$+0.0$
0.03	*35.4*	37.4	43.7	$+6.3$	*31.8*	36.1	45.4	$+9.3$
0.05	61.3	72.1	76.0	$+3.9$	*31.8*	71.5	76.2	$+4.7$
0.07	42.7	76.1	81.2	$+5.1$	41.4	75.7	81.3	$+5.6$
0.10	63.5	62.9	69.4	$+6.5$	63.6	63.2	70.4	$+7.2$
0.30	75.2	79.6	83.3	$+3.7$	75.3	79.6	83.2	$+3.6$
0.50	84.7	80.1	83.5	$+3.4$	84.7	80.1	83.6	$+3.5$
1.00	85.2	80.7	83.9	$+3.2$	85.1	80.8	83.8	$+3.0$

Δ = SAF $-$ EMA.

C.9 PRACTITIONER GUIDANCE: β AND WHERE THE CROSSOVER LIES

How β trades bias for variance. By Prop. 2, a larger β reduces the A-gap noise and slows A's effective step as $(1-\beta)$. By Eq. (5), $\sigma^\star \propto 1/\sqrt{1-\beta}$. On MNLI/FedIT at α=0.05, SAF stays within about $\pm 1\%$ across $\beta \in \{0.7, 0.8, 0.85, 0.9, 0.95\}$ and drops at β=0.99 (Figure 5, one seed). We use β=0.9 throughout and have not swept β at other α values. For strongly skewed settings $\beta \in [0.9, 0.95]$ is a reasonable starting point. Near IID we do not recommend SAF.

Deciding whether to use SAF. Eq. (5) depends on quantities that cannot be observed in deployment. Three practical options follow. (i) The label-histogram screen (App. C.4), which is free but validated only in-sample on MNLI. (ii) A short pilot of FEDAVG and SAF (about 20 rounds), which is the most direct option. (iii) An estimate of the Dirichlet concentration from client label histograms (Hsu

et al., 2019). These options address the crossover against FEDAVG. The crossover against passive EMA comes earlier (§6.4).

Adaptive β. Schedules such as $\beta_t = 1 - 1/(1 + ct)$ (Polyak & Juditsky, 1992), or a β set per layer or by heterogeneity, could reduce the cost above the crossover. We do not study them here.

C.10 FULL MEDMNIST RESULTS

Tables 15 and 16 report $\alpha \in \{0.01, 0.03, 0.05\}$ for the four methods (seed 42, final round). "EMA" is full feedback and "EMA-Eval" is passive EMA. EMA stays at 30–58% on PathMNIST but sometimes beats Vanilla on OctMNIST, so its behavior depends strongly on the task.

Table 15: PathMNIST final-round accuracy (ViT-small, 50 clients, participation 0.3, seed 42). Bolding follows Table 1.

Method	α	Vanilla	EMA	EMA-EVAL	**SAF**
FedIT	0.01	44.0	38.7	70.0	**74.1**
	0.03	61.4	44.2	**84.9**	83.1
	0.05	83.3	57.1	85.1	**85.4**
FedEx-LoRA	0.01	43.1	39.5	70.6	**73.5**
	0.03	60.7	46.0	**85.6**	84.7
	0.05	82.3	58.5	84.7	**85.9**
HetLoRA	0.01	35.1	29.7	**39.7**	36.4
	0.03	72.6	47.9	74.5	**75.6**
	0.05	82.2	50.8	**83.8**	77.4
LoRA-A^2	0.01	59.0	47.1	**66.2**	61.3
	0.03	72.9	36.8	**79.0**	76.2
	0.05	81.9	37.9	**85.6**	83.1

Table 16: OctMNIST final-round accuracy (ViT-small, 50 clients, participation 0.3, seed 42). *Italic*: at or below the majority-class baseline (33.6%). Bolding follows Table 1.

Method	α	Vanilla	EMA	EMA-EVAL	**SAF**
FedIT	0.01	*33.6*	*33.6*	34.2	**44.5**
	0.03	71.7	*38.7*	59.3	**69.2**
	0.05	*41.7*	69.6	**85.1**	82.6
FedEx-LoRA	0.01	*33.6*	*33.6*	*18.2*	**43.3**
	0.03	73.0	*41.6*	**75.7**	68.1
	0.05	*41.1*	71.2	**85.3**	84.0
HetLoRA	0.01	*33.6*	34.3	*33.6*	**39.8**
	0.03	39.3	*36.6*	**43.3**	*41.6*
	0.05	*31.1*	38.7	57.9	**61.8**
LoRA-A^2	0.01	*33.6*	34.1	*9.6*	**59.0**
	0.03	59.9	*34.7*	**74.1**	63.3
	0.05	*33.0*	66.9	**83.4**	81.2

C.11 FEEDBACK ABLATION: BOUNDARY CASES

In Table 2, EMA-B beats SAF on HetLoRA (73.6 ± 7.4 vs. 72.1 ± 7.6), and EMA-EVAL beats SAF on LoRA-A^2 (76.5 vs. 75.8). Both gaps are well within one standard deviation. HetLoRA also contains both of the cells in Table 1 where SAF does not beat both baselines, all of the cells whose win/loss status changes under best-round accuracy, and the only cell where deploying $\bar{B}_T$ is worse than B_T (App. C.1). *Hypothesis:* with client ranks drawn from a pool, the coordinates of the aggregated A are less consistently aligned across clients. This weakens the premise that the fed-back $\bar{A}$ is a shared subspace and makes smoothing B relatively more useful. We have not derived this, so we treat HetLoRA as a boundary case and do not generalize SAF's ordering to methods with heterogeneous ranks.

C.12 Multi-Seed Stochastic Collapse

Table 17: Per-seed MNLI results (α=0.05, final round). EMA-Eval depends strongly on the seed; EMA is intermediate; SAF has standard deviation ≤ 3.1. *Italic*: at or below the majority-class baseline.

Method	Config	s42	s43	s44	s45	s46	Mean±Std
FedIT	EMA-Eval	*31.8*	**82.2**	65.3	74.5	80.6	66.9 ± 20.7
	EMA	72.1	80.5	74.2	65.2	79.3	74.3 ± 6.2
	SAF	**76.0**	**82.2**	79.7	**75.1**	80.6	$\mathbf{78.7 \pm 3.1}$
FedEx-LoRA	EMA-Eval	*31.8*	68.1	68.6	76.7	**80.9**	65.2 ± 19.4
	EMA	71.5	80.4	74.1	66.1	79.5	74.3 ± 5.9
	SAF	**76.2**	**82.6**	78.8	**76.2**	81.0	$\mathbf{79.0 \pm 2.8}$

C.13 Best-Round vs. Final-Round Accuracy

Table 18 reports best-round accuracy, final-round accuracy, and their gap Δ. SAF has the smallest Δ in 17/24 cells (Vanilla in 4 and EMA in 3). The four Vanilla cells (QNLI for FedIT, FedEx-LoRA, and LoRA-A^2; PathMNIST for LoRA-A^2) are runs with low absolute accuracy, where the late drop is mechanically small. Under best-round accuracy, SAF beats both baselines in 20/24 cells, beats EMA in 24/24, and leads the stronger baseline by 3.8% on average. All six cells whose status changes between criteria are HetLoRA cells (four wins become losses and two losses become wins). The best rounds are not early transients: the median best round is 80 of 100 for SAF and EMA (67.5 for FedAvg), and 72–90% of runs peak at round 50 or later. Across the 24 cells, paired Wilcoxon signed-rank tests give $z = -4.26$, $p = 2.1\times10^{-5}$ (vs. FedAvg) and $z = -4.23$, $p = 2.4\times10^{-5}$ (vs. EMA). Paired Cohen's d is 1.62 and 1.28, respectively (Figure 3).

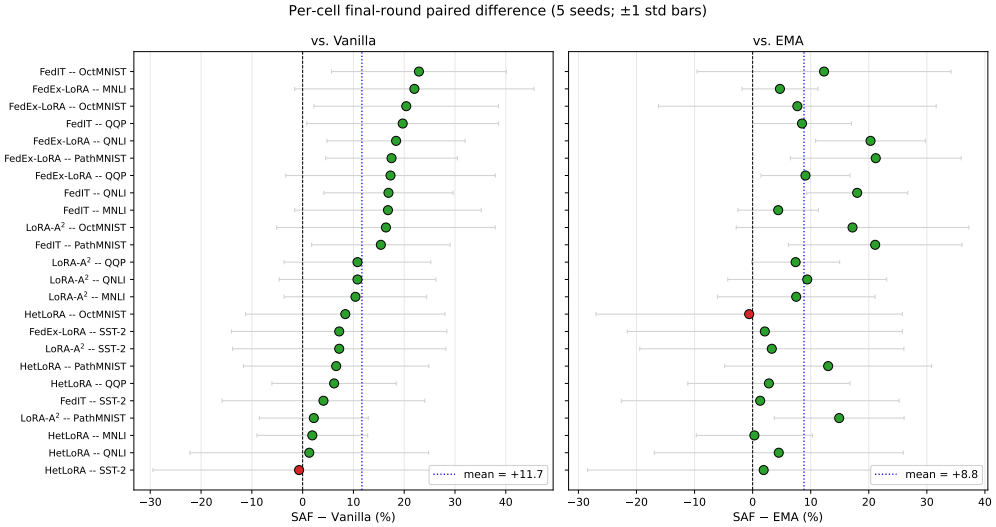

Figure 3: Per-cell paired differences at the final round: SAF − FedAvg (left) and SAF − EMA (right), five seeds, ±1 std. The dotted line is the 24-cell mean (+11.7 / +8.8). The only negative cells are HetLoRA SST-2 (left) and HetLoRA OctMNIST (right).

C.14 Per-Seed Scatter, Decay Sweep, Collapse Map, and A Variation

C.15 Learning-Rate Compensation

C.16 Combination with FedProx

At α=0.01 all three variants stay at the majority class, so cross-round smoothing does not escape deterministic collapse there. At α=0.05, SAF improves the FedProx endpoint on both seeds (+6.2%, +19.9%), whereas the effect of EMA reverses sign (−7.1%, +13.9%). Part of the seed-43 difference

Table 18: Best-round and final-round accuracy and their gap Δ: four methods $\times$ six datasets, five seeds. GLUE at α=0.05, MedMNIST at α=0.01. Bold on Δ rows marks the smallest mean Δ in each block; SAF is smallest in 17/24.

Method	Config	MNLI	QNLI	QQP	SST-2	OctMNIST	PathMNIST
Best-round accuracy:							
FedIT	Vanilla	$72.8_{\pm 10.3}$	$59.0_{\pm 17.8}$	$79.5_{\pm 8.3}$	$88.9_{\pm 8.4}$	$55.3_{\pm 13.4}$	$72.5_{\pm 6.8}$
	EMA	$77.0_{\pm 3.0}$	$73.1_{\pm 14.7}$	$79.9_{\pm 3.1}$	$92.3_{\pm 1.3}$	$67.4_{\pm 7.6}$	$63.1_{\pm 7.7}$
	SAF	$81.2_{\pm 1.0}$	$83.5_{\pm 5.9}$	$83.9_{\pm 1.7}$	$92.7_{\pm 1.2}$	$76.2_{\pm 1.6}$	$81.0_{\pm 4.9}$
FedEx-LoRA	Vanilla	$63.4_{\pm 19.6}$	$58.6_{\pm 18.0}$	$79.7_{\pm 9.3}$	$88.2_{\pm 8.3}$	$55.3_{\pm 13.4}$	$73.3_{\pm 6.2}$
	EMA	$77.1_{\pm 2.8}$	$72.8_{\pm 15.2}$	$79.8_{\pm 3.2}$	$92.2_{\pm 1.3}$	$68.3_{\pm 7.9}$	$64.2_{\pm 7.7}$
	SAF	$81.1_{\pm 1.3}$	$83.7_{\pm 5.4}$	$83.8_{\pm 1.9}$	$92.8_{\pm 1.0}$	$74.7_{\pm 3.7}$	$82.7_{\pm 3.7}$
HetLoRA	Vanilla	$76.0_{\pm 6.0}$	$68.8_{\pm 17.2}$	$81.3_{\pm 3.7}$	$83.0_{\pm 17.5}$	$51.5_{\pm 16.6}$	$63.3_{\pm 7.5}$
	EMA	$76.7_{\pm 2.2}$	$64.2_{\pm 17.4}$	$77.8_{\pm 8.3}$	$75.3_{\pm 17.7}$	$59.4_{\pm 14.9}$	$47.4_{\pm 7.0}$
	SAF	$79.5_{\pm 1.9}$	$67.8_{\pm 19.3}$	$79.7_{\pm 9.3}$	$78.8_{\pm 17.2}$	$62.7_{\pm 13.8}$	$60.1_{\pm 12.4}$
LoRA-A^2	Vanilla	$76.2_{\pm 4.8}$	$66.6_{\pm 16.6}$	$78.1_{\pm 8.6}$	$92.2_{\pm 2.5}$	$59.8_{\pm 14.8}$	$70.8_{\pm 4.1}$
	EMA	$72.6_{\pm 5.6}$	$72.9_{\pm 13.8}$	$76.4_{\pm 5.8}$	$91.7_{\pm 1.8}$	$66.7_{\pm 6.7}$	$58.8_{\pm 4.5}$
	SAF	$79.0_{\pm 2.4}$	$80.4_{\pm 7.3}$	$81.3_{\pm 3.2}$	$92.6_{\pm 1.5}$	$72.8_{\pm 4.4}$	$74.7_{\pm 5.1}$
Final-round accuracy ($t = T = 100$):							
FedIT	Vanilla	$61.9_{\pm 18.1}$	$55.2_{\pm 11.0}$	$57.2_{\pm 17.7}$	$79.9_{\pm 9.8}$	$39.7_{\pm 8.8}$	$63.5_{\pm 11.7}$
	EMA	$74.3_{\pm 6.2}$	$54.1_{\pm 5.9}$	$68.4_{\pm 5.4}$	$82.7_{\pm 16.4}$	$50.3_{\pm 16.1}$	$57.8_{\pm 13.2}$
	SAF	$78.7_{\pm 3.1}$	$72.1_{\pm 6.4}$	$76.9_{\pm 6.6}$	$84.0_{\pm 17.4}$	$62.6_{\pm 14.8}$	$78.9_{\pm 7.0}$
FedEx-LoRA	Vanilla	$57.0_{\pm 23.4}$	$55.4_{\pm 11.4}$	$59.8_{\pm 19.8}$	$77.3_{\pm 13.2}$	$36.0_{\pm 5.2}$	$62.7_{\pm 11.7}$
	EMA	$74.3_{\pm 5.9}$	$53.5_{\pm 5.9}$	$68.0_{\pm 5.0}$	$82.4_{\pm 16.9}$	$48.7_{\pm 16.4}$	$59.0_{\pm 13.6}$
	SAF	$79.0_{\pm 2.8}$	$73.8_{\pm 7.4}$	$77.1_{\pm 5.8}$	$84.5_{\pm 16.6}$	$56.4_{\pm 17.4}$	$80.2_{\pm 5.6}$
HetLoRA	Vanilla	$70.2_{\pm 7.8}$	$59.7_{\pm 17.4}$	$68.0_{\pm 6.3}$	$70.4_{\pm 20.5}$	$35.9_{\pm 5.3}$	$49.5_{\pm 12.8}$
	EMA	$71.8_{\pm 6.5}$	$56.5_{\pm 14.5}$	$71.4_{\pm 9.2}$	$67.8_{\pm 22.6}$	$44.9_{\pm 18.4}$	$43.1_{\pm 12.2}$
	SAF	$72.1_{\pm 7.6}$	$61.0_{\pm 15.8}$	$74.2_{\pm 10.5}$	$69.7_{\pm 20.2}$	$44.3_{\pm 18.9}$	$56.1_{\pm 13.0}$
LoRA-A^2	Vanilla	$65.4_{\pm 12.9}$	$56.1_{\pm 12.6}$	$64.1_{\pm 12.9}$	$78.1_{\pm 14.6}$	$44.8_{\pm 16.6}$	$67.8_{\pm 6.9}$
	EMA	$68.3_{\pm 12.4}$	$57.5_{\pm 10.4}$	$67.5_{\pm 3.9}$	$82.0_{\pm 17.0}$	$44.0_{\pm 14.6}$	$55.1_{\pm 7.6}$
	SAF	$75.8_{\pm 5.5}$	$66.9_{\pm 8.9}$	$74.9_{\pm 6.5}$	$85.3_{\pm 15.1}$	$61.2_{\pm 13.7}$	$70.0_{\pm 8.2}$
Δ = *best* $-$ *final (paired across seeds):*							
FedIT	Vanilla	$+10.9_{\pm 10.6}$	**+3.8**$_{\pm 6.8}$	$+22.2_{\pm 13.9}$	$+9.0_{\pm 10.3}$	$+15.6_{\pm 18.0}$	$+9.0_{\pm 10.3}$
	EMA	$+2.7_{\pm 4.3}$	$+18.9_{\pm 14.1}$	$+11.6_{\pm 6.5}$	$+9.6_{\pm 16.0}$	$+17.1_{\pm 13.9}$	$+5.4_{\pm 6.2}$
	SAF	**+2.5**$_{\pm 2.4}$	$+11.4_{\pm 7.9}$	**+7.0**$_{\pm 7.2}$	**+8.7**$_{\pm 16.8}$	**+13.6**$_{\pm 15.6}$	**+2.1**$_{\pm 2.3}$
FedEx-LoRA	Vanilla	$+6.4_{\pm 7.2}$	**+3.2**$_{\pm 6.7}$	$+20.0_{\pm 15.8}$	$+10.9_{\pm 12.6}$	$+19.2_{\pm 16.3}$	$+10.5_{\pm 10.5}$
	EMA	$+2.8_{\pm 3.6}$	$+19.3_{\pm 13.4}$	$+11.9_{\pm 6.4}$	$+9.8_{\pm 16.6}$	$+19.6_{\pm 14.2}$	$+5.2_{\pm 7.1}$
	SAF	**+2.2**$_{\pm 1.8}$	$+9.9_{\pm 8.9}$	**+6.6**$_{\pm 6.9}$	**+8.4**$_{\pm 16.0}$	**+18.3**$_{\pm 15.9}$	**+2.4**$_{\pm 2.9}$
HetLoRA	Vanilla	$+5.8_{\pm 7.9}$	$+9.0_{\pm 12.5}$	$+13.3_{\pm 7.5}$	$+12.6_{\pm 16.6}$	$+15.6_{\pm 18.4}$	$+13.8_{\pm 6.6}$
	EMA	**+4.9**$_{\pm 5.5}$	$+7.8_{\pm 10.1}$	$+6.3_{\pm 8.3}$	**+7.5**$_{\pm 9.6}$	**+14.5**$_{\pm 15.1}$	$+4.3_{\pm 5.4}$
	SAF	$+7.4_{\pm 6.2}$	**+6.8**$_{\pm 11.3}$	**+5.5**$_{\pm 9.2}$	$+9.1_{\pm 14.6}$	$+18.4_{\pm 18.4}$	**+4.0**$_{\pm 4.8}$
LoRA-A^2	Vanilla	$+10.9_{\pm 10.4}$	**+10.5**$_{\pm 10.3}$	$+14.0_{\pm 15.1}$	$+14.1_{\pm 13.5}$	$+15.0_{\pm 17.4}$	**+3.0**$_{\pm 3.3}$
	EMA	$+4.3_{\pm 8.8}$	$+15.3_{\pm 13.8}$	$+8.9_{\pm 6.1}$	$+9.7_{\pm 16.4}$	$+22.7_{\pm 8.9}$	$+3.6_{\pm 3.7}$
	SAF	**+3.2**$_{\pm 4.7}$	$+13.4_{\pm 6.6}$	**+6.4**$_{\pm 6.5}$	**+7.3**$_{\pm 14.4}$	**+11.5**$_{\pm 13.9}$	$+4.7_{\pm 4.4}$

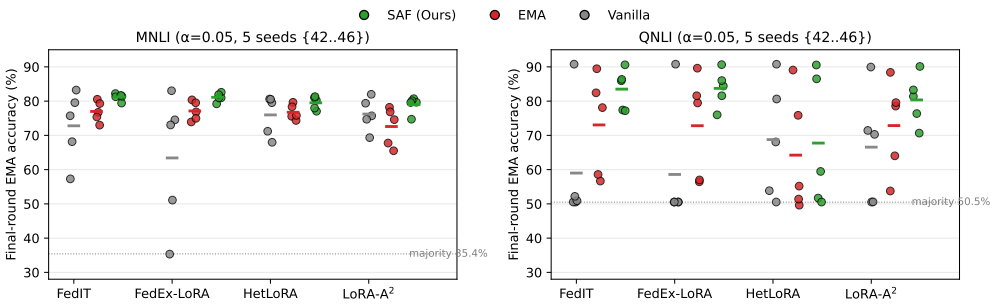

Figure 4: Per-seed best-round accuracy of the deployed model for MNLI and QNLI (α=0.05, seeds 42–46; bars show the mean). Vanilla and EMA vary widely and often bimodally; SAF clusters tightly.

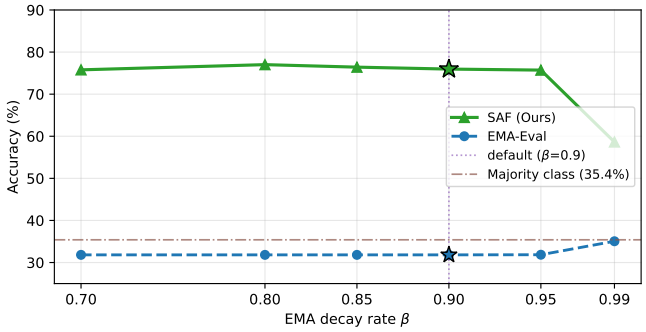

Figure 5: SAF vs. EMA-EVAL across the EMA decay β (FedIT, MNLI, α=0.05, seed 42). EMA-EVAL stays near the majority class at every β. SAF stays in 75.7–77.0% for $\beta \leq 0.95$ and drops to 58.6% at β=0.99, where the effective step $(1-\beta)\eta$ becomes very small.

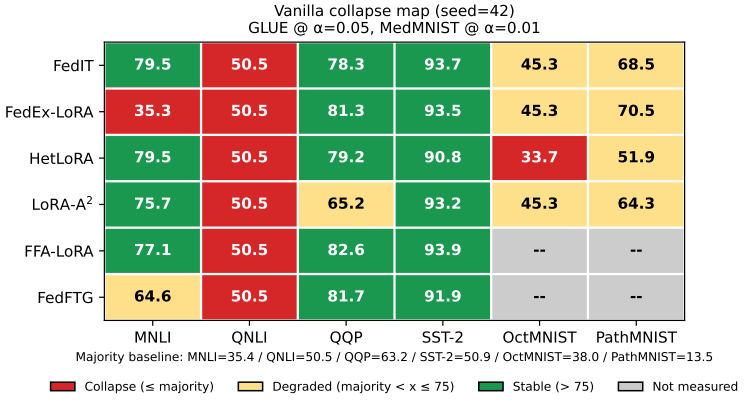

Figure 6: Where Vanilla collapses (seed 42; GLUE at α=0.05, MedMNIST at α=0.01). Red: at or below the majority class; yellow: above the majority class but at most 75%; green: above 75%; gray: not measured. FFA-LoRA and FedFTG are evaluated on GLUE only.

Table 19: Scaling η_B under EMA (MNLI, FedIT, α=0.05, seed 42, final round). The remaining 0.7–0.8% gap may lie within seed variation; we report the direction only.

Configuration	Final-round accuracy	Δ vs. SAF
EMA, $\eta_B \times 1$	72.1	−3.9
EMA, $\eta_B \times 3$	75.2	−0.8
EMA, $\eta_B \times 10$	75.3	−0.7
SAF, $\eta_B \times 1$	**76.0**	—

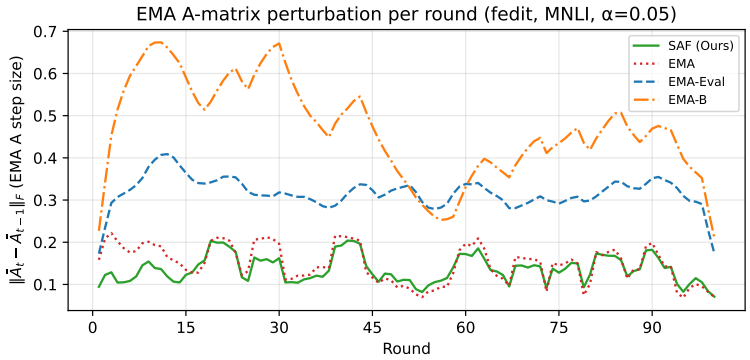

Figure 7: Per-round variation of EMA-A, $\|\bar{A}_t - \bar{A}_{t-1}\|_F$ (FedIT, MNLI, α=0.05, seed 42). This is a different quantity from the gap ϵ_t. Means: SAF 0.138, EMA-EVAL 0.321, EMA-B 0.470, EMA 0.151. EMA and SAF have nearly the same A variation, so the accuracy gap between them is not explained by the A side.

Table 20: FedProx (μ=0.1) on MNLI/FedIT (final round; best round in parentheses). The α=0.01 row uses seed 42 only; α=0.05 uses seeds 42 and 43. The seed-42 runs come from the original software environment and the seed-43 runs from the rebuilt one.

α	seed	FedProx	FedProx+EMA	FedProx+SAF
0.01	42	*31.8*	*31.8*	*31.8*
0.05	42	71.2 (77.9)	64.1 (64.1)	77.4 (77.4)
0.05	43	60.0 (78.0)	73.9 (73.9)	79.9 (79.9)

comes from FedProx alone dropping from 78.0 at its best round to 60.0 at the final round. We therefore make no claim about how EMA and FedProx interact. One explanation, that pulling local B toward a slowly moving EMA anchor compounds the shrinkage, remains an untested hypothesis.

C.17 FALSIFICATION CONTROLS (FFA-LORA, FEDFTG)

Thm. 1 predicts that the B-side penalty comes from feeding back B, whether or not A is trained, and that the A/B asymmetry disappears without the BA factorization. Table 21 checks both predictions (seed 42). On FFA-LoRA (A frozen), EMA-EVAL beats every configuration with B-feedback and beats Vanilla by $+10.5\%$ (MNLI) and $+33.0\%$ (QNLI). On GaLore-based FedFTG (Zhao et al., 2024), which has no BA structure, full EMA feedback gains $+42.3\%$ and $+26.9\%$ over Vanilla. Both results agree in sign with the prediction. Each is a single-seed cell.

Table 21: EMA configurations on the falsification controls (α=0.05, seed 42, final round). Bolding follows Table 1.

Method	Dataset	Vanilla	EMA	EMA-EVAL	EMA-B
FFA-LoRA	MNLI	60.7	61.0	**71.2**	63.9
	QNLI	*50.5*	*50.8*	**83.5**	*53.2*
FedFTG	MNLI	*32.2*	**74.5**	55.5	—
	QNLI	*50.5*	**77.4**	*50.8*	—

C.18 IMPLEMENTATION DETAILS

Datasets, models, methods. We use MNLI, QNLI, QQP, and SST-2 from GLUE (Wang et al., 2019) with RoBERTa-base (125M) and RoBERTa-large (355M, MNLI only) (Liu et al., 2019), and PathMNIST and OctMNIST from MedMNIST v2 (Yang et al., 2023) with ViT-small (22M) (Dosovitskiy et al., 2021). The methods are FedIT, FedEx-LoRA, HetLoRA, and LoRA-A^2, plus the controls FFA-LoRA and FedFTG, and FedSA-LoRA for App. C.2.

Heterogeneity. Client label distributions are drawn from a Dirichlet distribution (Hsu et al., 2019; G. et al., 2024) with $\alpha \in \{0.01, 0.03, 0.05, 0.07, 0.10, 0.30, 0.50, 1.00\}$. The main results use α=0.05 (GLUE) and α=0.01 (MedMNIST).

Hyperparameters. LoRA rank r=8 and LoRA scaling 16 (32 at r=16, keeping the scaling-to-rank ratio fixed). AdamW with η=2×10^{-4} for A, B, and the head, (β_1, β_2)=(0.9, 0.999), ϵ=10^{-8}, weight decay 0.01, linear warmup over the first 10% of rounds, then constant. BF16, T=100 rounds, two local epochs per round (one in the E=1 check). The optimizer is created fresh for each client at every round, so no AdamW state carries across rounds. We do not use LoRA+ learning rates (Hayou et al., 2024). FedProx uses μ=0.1.

EMA configurations. All use β=0.9, and the EMA tracks every trainable parameter. SAF copies the EMA `lora_A` tensors into the training model after each EMA update. EMA-B copies `lora_B`. EMA-EVAL copies nothing. EMA copies all EMA-tracked parameters, which includes the classification head. Every EMA configuration is evaluated with the EMA model. Vanilla is evaluated with the FEDAVG training model.

```
# SAF A-feedback (after the EMA update)
for key in state:
    if 'lora_A' in key:
        state[key] = ema_state[key].clone()
model.load_state_dict(state)
```

Target modules. On RoBERTa, LoRA is applied to `query` and `value`. On ViT-small it is applied to the fused `qkv` projection.

Evaluation. The primary metric is final-round (t=T=100) accuracy, with no validation-based checkpoint selection. MNLI accuracy is the mean of matched and mismatched accuracy. Five-seed cells use seeds 42–46. All other cells are labeled with their seeds.

Table 22: Federated configuration per dataset.

Dataset	Modality	Clients	Fraction	Clients/round
MNLI	Text	100	0.1	10
QQP	Text	100	0.1	10
QNLI	Text	35	0.3	∼10
SST-2	Text	20	0.5	10
PathMNIST	Medical image	50	0.3	15
OctMNIST	Medical image	50	0.3	15

Software environments. The main table, ablations, sweeps, controls, and seed-42 FedProx and RoBERTa-large runs used the original environment. The rank-16, E=1, seed-43 FedProx, FedSA-LoRA, and RoBERTa-large seeds 42 (rerun), 43, and 44 used a rebuilt environment (torch 2.13, transformers 4.57, peft 0.19). The seed-42 SAF run on RoBERTa-large reproduced across environments (56.5/55.1 final, 84.0/84.6 best). Within-table comparisons are within one environment except where marked.

Compute. All runs used a single NVIDIA RTX 3090 (24GB); the rebuilt-environment RoBERTa-large runs used an RTX 5090. Run times: GLUE 1.2–5.4 h, MedMNIST 1.4–2.3 h, RoBERTa-large about 9.7 h per run.

C.19 TERMINOLOGY

D FUTURE WORK

The experiments that would most directly close the gaps stated in §7 are: (i) a FedSA-LoRA comparison on a common balanced test set with persisted client B_k (App. C.2); (ii) a multi-seed rank sweep $r \in \{2, 4, 16, 32\}$ and a plain-SGD control; (iii) naturally partitioned data such as FEMNIST writer splits, and multi-center medical data; (iv) multi-seed α sweeps; and (v) stacking with FedOpt and FedSVD. On the theory side, a joint descent analysis of (A, B) that tracks their rank-r coupling would turn the quantitative layer into a statement about the setting we train in. An adaptive schedule β_t and generative/seq2seq evaluation are further directions.

Table 23: Configurations and notation.

Term	Description
Vanilla	FEDAVG without EMA; the training model is deployed.
EMA	Full feedback: every EMA-tracked parameter (A, B, head) is fed back.
EMA-EVAL	EMA used for deployment only; no feedback.
EMA-B	Only EMA-B is fed back (ablation).
SAF	Only EMA-A is fed back (proposed).
U_t^X	Realized round displacement of factor X, $X_{t+1}^{\mathrm{FA}} - X_t^{\mathrm{in}}$.
ϵ_t^X	Gap $\bar{X}_t - X_t^{\mathrm{FA}}$ between the EMA and the FEDAVG iterate.
β	EMA decay (default 0.9).
σ	Per-round noise scale of the A displacement; grows as α decreases.
α	Dirichlet concentration of client label distributions.